Jumistik
E-ISSN: 2964-3953 P-ISSN: 2964-3988
Jurnal Manajemen Informatika, Sistem Informasi dan Teknologi Komputer

Vol.4 No.1Juni 2025
Hal 373-383
Https://ojs.amiklps.ac.id

ANALISIS SENTIMEN ULASAN FILM PADA *DATASET IMDB* MENGGUNAKAN ALGORITMA *NAIVE BAYES*

Alya Nur Zhafira[1*], Nurul Afifah[2], Shifa Anditya Putri[3], Viva Marhalatun[4], Dhanar Intan Surya Saputra[5]
[1,2,3,4,5] Program Studi Sistem Informasi,Fakultas Ilmu Komputer, Universitas Amikom Purwokerto
E-mail: alyazhafira83@gmail.com[1*], nafifa140705@gmail.com[2], shifaandityaputri@gmail.com[3], viva19464@gmail.com[4], dhanaraputra@amikompurwokerto.ac.id[5]

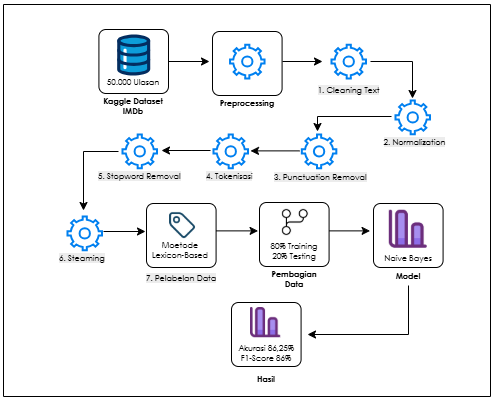
INFO ARTIKEL

Sejarah Artikel
Diterima : 28/05/2025
Direvisi : 30/05/2025
Diterbitkan : 05/06/2025

***Corresponding author**
alyazhafira83@gmail.com

DOI: 10.70247/jumistik.v4i1.139
Https://ojs.amiklps.ac.id

GRAPHICAL ABSTRACT



ABSTRACT

This study aims to analyze sentiment in movie reviews from the IMDb dataset using the Naïve Bayes algorithm. The dataset consists of 50,000 English-language reviews labeled as either positive or negative. The analysis process includes several preprocessing steps such as text cleaning, normalization, tokenization, stopword removal, and stemming. Feature extraction is performed using TF-IDF to convert the textual data into numerical form. The data is split into training (80%) and testing (20%) sets. Model performance is evaluated using a confusion matrix and classification metrics including accuracy, precision, recall, and F1-score. Results show that the Naïve Bayes algorithm achieves an accuracy of 86.25%, with a precision of 87% for positive sentiment and 85% for negative sentiment. The recall scores are 85% and 87%, respectively, with an average F1-score of 86%. These results demonstrate that Naïve Bayes is both effective and computationally efficient for large-scale sentiment classification. Furthermore, the combination of text preprocessing and TF-IDF significantly enhances model performance. This research contributes to the development of sentiment-based movie recommendation systems and offers insights into public opinion analysis from digital text data.

***Keywords**: Sentiment Analysis, Movie Reviews, Naïve Bayes, Kaggle Dataset*

ABSTRAK

Penelitian ini bertujuan untuk menganalisis sentimen pada ulasan film dari dataset IMDb menggunakan algoritma Naïve Bayes. Dataset terdiri dari 50.000 ulasan berbahasa Inggris yang telah diberi label sentimen positif dan negatif. Analisis dilakukan melalui beberapa tahap preprocessing, seperti pembersihan teks, normalisasi, tokenisasi, penghapusan stopword, dan stemming. Ekstraksi fitur menggunakan metode TF-IDF untuk merepresentasikan teks dalam bentuk numerik. Data dibagi menjadi data latih (80%) dan data uji (20%). Evaluasi kinerja model dilakukan menggunakan confusion matrix dan classification report. Hasil menunjukkan bahwa algoritma Naïve Bayes berhasil mencapai akurasi sebesar 86,25%, dengan precision sebesar 87% untuk sentimen positif dan 85% untuk sentimen negatif. Nilai recall masing-masing adalah 85% dan 87%, dengan F1-score rata-rata sebesar 86%. Hasil ini menunjukkan bahwa Naïve Bayes cukup efektif dan efisien dalam menangani data ulasan berskala besar. Selain itu, kombinasi preprocessing dan metode representasi fitur terbukti meningkatkan kualitas klasifikasi. Penelitian ini diharapkan dapat mendukung pengembangan sistem rekomendasi film berbasis sentimen dan menjadi referensi dalam pengolahan opini publik pada data teks digital.

***Kata kunci**: Analisis Sentimen, Ulasan Film, Naïve Bayes, dataset Kaggle*

PENDAHULUAN

Perkembangan teknologi digital telah mengubah cara masyarakat memperoleh dan memproses informasi, terutama dalam konteks pengambilan keputusan. Dalam industri film, ulasan daring menjadi salah satu sumber utama yang digunakan oleh konsumen sebelum memutuskan untuk menonton. Kepercayaan terhadap konten ulasan yang tersedia di situs web dinilai lebih tinggi dibandingkan dengan media sosial, karena dianggap lebih objektif dan informatif [1].

Internet Movie Database (IMDb) dikenal sebagai salah satu situs ulasan film online paling luas dan banyak digunakan secara global, dengan cakupan informasi yang sangat mendetail mengenai film, serial televisi, serta tokoh-tokoh dalam industri hiburan.Platform ini menyediakan data mendalam mengenai film, acara televisi, serta tokoh industri hiburan, dan memungkinkan pengguna memberikan peringkat serta menulis ulasan. Berkat kelengkapan dan popularitasnya, *IMDb* sering dijadikan sumber utama dalam penelitian di bidang pemrosesan bahasa alami dan sistem rekomendasi.

Data ulasan dari *IMDb* sering dijadikan acuan standar atau *benchmark dataset* dalam berbagai penelitian yang fokus pada klasifikasi sentimen. Misalnya, penelitian oleh Islamy et al. menggunakan *dataset IMDb* yang terdiri dari 50.000 ulasan film berlabel positif atau negatif untuk mengembangkan model analisis sentimen berbasis Long Short-Term Memory (LSTM) dan FastText, yang mencapai akurasi sebesar 86,3%.[2] Selain itu, studi oleh Ramadhan et al. membandingkan kinerja algoritma *Decision Tree* dan *Naïve Bayes* dalam menganalisis sentimen pada rating film di *IMDb*, memberikan wawasan berharga bagi pembuat film dan produsen dalam memahami respons penonton [3].

Analisis sentimen merupakan cabang dari pemrosesan bahasa alami (*Natural Language Processing/NLP*) yang berfokus pada identifikasi opini dalam teks dan mengklasifikasikannya menjadi sentimen positif, negatif, atau netral. Dalam konteks ini, pendapat individu di media sosial, forum, atau situs ulasan memiliki nilai signifikan dalam merepresentasikan sikap publik terhadap suatu topik. Penelitian oleh Berliani dan Lestari menunjukkan bahwa metode *Naïve Bayes* efektif dalam mengklasifikasikan sentimen masyarakat terhadap isu-isu terkini di media sosial, dengan performa yang kompetitif dibandingkan algoritma lain seperti Support Vector Machine [4].

Keberadaan ulasan digital dalam jumlah besar menimbulkan kebutuhan akan metode yang efisien dan akurat dalam mengolah data tersebut.Metode pembelajaran mesin telah berkembang menjadi pendekatan utama dalam mengatasi tantangan pengolahan opini publik, terutama dalam hal pengelompokan sentimen.Dari berbagai metode yang tersedia, *Naïve Bayes* merupakan salah satu algoritma yang banyak digunakan karena kesederhanaannya serta kemampuannya dalam memberikan hasil yang kompetitif dalam klasifikasi teks. Studi oleh Damayanti dan Maulidiyah menyoroti bahwa meskipun *Naïve Bayes* memiliki performa yang baik, kombinasi dengan algoritma lain seperti *Support Vector Machine* atau *Random Forest* dapat meningkatkan akurasi dalam klasifikasi sentimen [5].

Selain itu, penelitian oleh Ramadhani dan Fajarianto mengimplementasikan algoritma *Naïve Bayes* dengan teknik *Smoothing Laplace* dalam sistem informasi evaluasi perkuliahan. Hasilnya menunjukkan bahwa pendekatan ini efektif dalam mengklasifikasikan sentimen mahasiswa terhadap perkuliahan, yang dapat digunakan untuk meningkatkan kualitas pengajaran [6].

Sebagai contoh, Razaq et al. [7] melakukan evaluasi terhadap pendekatan *Naïve Bayes* menggunakan teknik TF-IDF pada *dataset IMDb* dan Rotten Tomatoes, dan memperoleh akurasi tinggi yang membuktikan efektivitas metode ini. Sebagai perbandingan, Setiawan et al. [8] melakukan analisis sentimen terhadap topik pemindahan Ibu Kota Nusantara (IKN) dengan membandingkan dua algoritma, yaitu *Support Vector Machine* (SVM) dan *Naïve Bayes*. *dataset* yang digunakan terdiri dari 5.128 tweet yang diklasifikasikan ke dalam sentimen positif dan negatif. Hasilnya menunjukkan bahwa SVM memperoleh akurasi sebesar 84%, sedangkan *Naïve Bayes* hanya mencapai 77%. Penelitian serupa oleh Hidayat et al. [9] menunjukkan bahwa *Naïve Bayes* mampu menangani komentar *YouTube* terkait isu sosial dengan akurasi 79,4%, dan berhasil mengidentifikasi kecenderungan opini publik yang bersifat negatif.studi ini memberikan gambaran bahwa metode *Naïve Bayes* tetap memiliki relevansi yang tinggi dalam berbagai konteks dan domain aplikasi yang melibatkan teks dalam jumlah besar, baik dalam skala internasional maupun nasional.

Selain keakuratannya, *Naïve Bayes* juga memiliki keunggulan dari sisi efisiensi komputasi. Algoritma ini tidak memerlukan waktu pelatihan yang lama atau sumber daya komputasi yang kompleks, sehingga sangat sesuai untuk diterapkan pada sistem yang membutuhkan respons cepat. Proses pelatihan dan prediksi yang cepat menjadikannya pilihan ideal dalam menangani data berskala besar seperti ulasan daring yang terus bertambah seiring waktu. Karakteristik ini sangat penting dalam konteks analisis ulasan film, di mana volume data tinggi dan hasil analisis perlu tersedia secara cepat untuk mendukung proses pengambilan keputusan yang responsif.

Namun demikian, masih terdapat tantangan yang dihadapi peneliti, khususnya dalam membangun model klasifikasi sentimen yang andal, efisien, dan memiliki kemampuan generalisasi yang baik. Permasalahan utama yang ingin dijawab dalam penelitian ini adalah bagaimana memanfaatkan algoritma *Naïve Bayes* untuk mengolah data ulasan

film dalam skala besar secara efektif. Selain itu, diperlukan evaluasi yang komprehensif terhadap performa model dalam mengklasifikasikan sentimen, baik positif maupun negatif, guna memastikan bahwa hasil analisis dapat diandalkan. Meskipun *Naïve Bayes* telah banyak digunakan, studi mengenai penerapannya secara spesifik pada *dataset IMDb* berskala besar, dengan jumlah data mencapai 50.000 ulasan, masih terbatas.

Oleh karena itu, penelitian ini perlu dilakukan untuk mengisi celah tersebut dengan menerapkan algoritma *Naïve Bayes* pada *dataset* ulasan film *IMDb*, serta mengevaluasi kinerjanya menggunakan metrik seperti akurasi, presisi, *recall*, dan *F1-score*. Penelitian ini diharapkan dapat memberikan kontribusi pada pengembangan metode analisis sentimen yang tidak hanya akurat, tetapi juga efisien dan dapat diintegrasikan ke dalam sistem rekomendasi berbasis sentimen. Temuan ini juga diharapkan dapat memperluas pemahaman tentang bagaimana *Naïve Bayes* dapat digunakan secara optimal dalam skenario dunia nyata yang membutuhkan pemrosesan teks dalam jumlah besar secara cepat dan andal.

TINJAUAN PUSTAKA

Islamy et al. [2] mendefinisikan bahwa analisis sentimen merupakan bidang kajian yang digunakan untuk mengidentifikasi dan menganalisis opini manusia terhadap entitas seperti produk, layanan, individu, peristiwa, topik, maupun atribut lainnya. Seiring dengan perkembangan media sosial dan platform digital, analisis sentimen telah menjadi pendekatan yang banyak digunakan untuk menggali opini publik secara otomatis, termasuk dalam konteks ulasan film. Ulasan film menjadi salah satu objek utama dalam penelitian sentimen karena sifatnya yang subjektif dan kaya akan ekspresi emosional.

Dalam bidang pemrosesan bahasa alami *Natural Language Processing (NLP)* dan pembelajaran mesin, ulasan film daring, khususnya dari situs seperti *Internet Movie Database* (*IMDb*), menyediakan sumber data yang sangat besar dan beragam. Data ini mencerminkan persepsi kolektif masyarakat terhadap film tertentu dan memberikan peluang besar bagi pengembangan sistem rekomendasi berbasis opini. Untuk menganalisis ulasan semacam ini, berbagai algoritma telah digunakan, salah satunya adalah algoritma *Naïve Bayes*. Algoritma ini dikenal karena kesederhanaannya, efisiensi komputasi, dan efektivitasnya dalam klasifikasi teks. *Naïve Bayes* bekerja berdasarkan prinsip probabilitas dengan asumsi independensi antar fitur, dan telah terbukti memberikan hasil yang kompetitif dalam banyak studi analisis sentimen.

Beberapa penelitian telah membuktikan efektivitas *Naïve Bayes* dalam konteks ulasan film. Anuar et al. [10] berhasil menerapkan algoritma ini pada ulasan film "*Oppenheimer*" dan mencapai akurasi sebesar 96% dengan presisi rata-rata 98%. Studi tersebut menekankan pentingnya tahap prapemrosesan, seperti pembersihan teks dan ekstraksi fitur menggunakan *TF-IDF*. Sementara itu, Talibzade [11] membandingkan model tradisional seperti *Naïve Bayes* dengan pendekatan berbasis *transformer* seperti *BERT*. Hasilnya menunjukkan bahwa meskipun model *transformer* menawarkan akurasi tinggi, *Naïve Bayes* tetap kompetitif, terutama dalam hal efisiensi dan kecepatan pelatihan.

Perbandingan performa antara algoritma *Naïve Bayes* dan metode lain juga telah dilakukan oleh Azzahra et al. [12] dalam konteks analisis sentimen komentar *YouTube* terkait bencana alam. Penelitian ini menunjukkan bahwa SVM memiliki akurasi lebih tinggi (92%) dibandingkan *Naïve Bayes* (79%), menegaskan pentingnya pemilihan algoritma yang sesuai dengan karakteristik data. Santoso dan Putra [13] juga menyoroti bahwa teknik prapemrosesan seperti penghapusan tanda baca, *stopword removal*, dan *stemming* berkontribusi signifikan terhadap peningkatan performa model, terutama saat dikombinasikan dengan representasi fitur *TF-IDF* dibandingkan metode sederhana seperti *Bag of Words*.

Meskipun banyak studi telah mengkaji penerapan *Naïve Bayes* dalam analisis sentimen, masih jarang penelitian yang secara komprehensif mengevaluasi pengaruh tahapan prapemrosesan terhadap performa algoritma ini dalam skala besar, khususnya pada *dataset IMDb* yang terdiri dari 50.000 ulasan. Sebagian besar penelitian terdahulu berfokus pada skala data yang lebih kecil atau tidak menekankan evaluasi menyeluruh terhadap metrik seperti *precision*, *recall*, dan *F1-score*. Dengan demikian, masih terdapat celah penelitian dalam menguji konsistensi dan efisiensi *Naïve Bayes* pada data ulasan film berskala besar yang kompleks dan variatif.

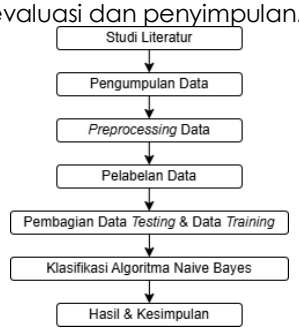
Menjawab tantangan tersebut, penelitian ini menawarkan kebaruan dengan menerapkan algoritma *Naïve Bayes* pada *dataset IMDb* berjumlah 50.000 ulasan dan mengevaluasi performanya secara mendalam menggunakan confusion matrix serta classification report. Evaluasi mencakup metrik akurasi, presisi, *recall*, dan *F1-score*, untuk menilai keseimbangan klasifikasi pada sentimen positif dan negatif. Penelitian ini tidak hanya memvalidasi kembali efektivitas *Naïve Bayes* dalam skenario dunia nyata, tetapi juga memberikan wawasan baru tentang pentingnya kualitas prapemrosesan data dalam membangun model yang andal. Temuan dari penelitian ini diharapkan dapat memberikan kontribusi praktis dan teoretis dalam pengembangan sistem analisis sentimen, khususnya dalam industri film dan sistem rekomendasi berbasis opini pengguna.

METODOLOGI PENELITIAN

Tahapan Penelitian

Penelitian ini dilakukan secara terstruktur dengan mengikuti tahapan yang digambarkan dalam Gambar 1. Setiap langkah dirancang untuk

mendukung keakuratan dan validitas hasil akhir. Tahapan penelitian mencakup studi literatur, pengumpulan dan preprocessing data, pelabelan sentimen, pembagian data, klasifikasi menggunakan algoritma *Naïve Bayes* dengan representasi fitur *TF-IDF*, hingga tahap evaluasi dan penyimpulan.



Gambar 1. Alur Metode Penelitian

a. Tahapan Penelitian

Metodologi penelitian ini dimulai dengan studi literatur untuk mengidentifikasi penelitian-penelitian terkait dan mendalami teknik-teknik terbaru dalam bidang analisis sentimen. Studi literatur dilakukan dengan mengkaji berbagai referensi ilmiah seperti jurnal nasional dan internasional, artikel konferensi, serta buku teks yang relevan. Fokus utama dari studi literatur ini adalah pada teknik analisis sentimen, *preprocessing* data teks, dan penerapan algoritma *Naïve Bayes* dalam klasifikasi teks. Hasil dari studi literatur ini akan memberikan dasar teoritis yang kuat dan memberikan gambaran tentang tren terkini dalam analisis sentimen serta solusi yang telah diterapkan oleh para peneliti sebelumnya. Setelah tahap studi literatur, penelitian ini melanjutkan ke pengumpulan data. Pada tahap ini, data yang digunakan adalah ulasan film yang diambil dari sumber terbuka yang tersedia secara online, seperti situs web film, forum diskusi, atau platform ulasan pengguna lainnya. *dataset* yang digunakan dipilih berdasarkan relevansinya dengan tujuan penelitian, yaitu untuk melakukan analisis sentimen terhadap ulasan film. Data yang telah diperoleh selanjutnya diolah dan dipersiapkan sedemikian rupa sehingga layak untuk dianalisis pada tahap berikutnya.

b. Pengumpulan Data

Pengumpulan data merupakan proses sistematis dalam memperoleh informasi atau fakta dari berbagai sumber yang relevan untuk mendukung kebutuhan penelitian [14].Proses pengumpulan data dilakukan dengan tujuan memperoleh kumpulan ulasan film yang dapat merepresentasikan opini penonton secara menyeluruh. *dataset* ini mencakup berbagai opini dari penonton yang dapat dianalisis untuk mengetahui apakah ulasan tersebut mengandung sentimen positif atau negatif. Proses pengumpulan data melibatkan ekstraksi data dari sumber-sumber terbuka dan relevan secara *online*. Usai proses pengumpulan selesai, data dipersiapkan terlebih dahulu agar memenuhi syarat untuk masuk ke tahap pra-pemrosesan secara optimal.Data yang dikumpulkan pada tahap ini berjumlah 983 ulasan yang akan digunakan untuk analisis lebih lanjut.

c. Preprocessing Data

Tahap preprocessing data merupakan langkah penting dalam penelitian ini untuk memastikan data yang akan digunakan telah bersih dan dalam format yang tepat.Hal ini sejalan dengan tujuan pengolahan data, yaitu menghasilkan informasi yang lebih mudah dipahami oleh penerima untuk menggambarkan peristiwa nyata dan mendukung pengambilan keputusan [15]. Oleh karena itu,Proses preprocessing meliputi penghapusan kata-kata yang tidak relevan (*stopwords*), penghilangan tanda baca yang tidak diperlukan, dan normalisasi kata agar konsisten. Teknik ini dilakukan agar data lebih mudah dipahami oleh algoritma yang akan digunakan dalam tahap selanjutnya. *Preprocessing* juga mencakup konversi teks menjadi bentuk yang lebih sederhana dan representatif, seperti mengubah kata menjadi bentuk dasar (*stemming*) dan penghilangan karakter-karakter yang tidak diperlukan. Setelah tahap-tahap tersebut, data teks kemudian direpresentasikan dalam bentuk numerik menggunakan metode *Term Frequency - Inverse Document Frequency (*TF-IDF). Teknik ini berfungsi untuk memberikan bobot lebih pada kata-kata penting dalam sebuah dokumen dan menurunkan bobot untuk kata yang sering muncul dalam seluruh korpus. TF-IDF sangat efektif dalam klasifikasi teks karena mampu menangkap pentingnya kata spesifik dalam membedakan sentimen. Proses ini penting untuk mempersiapkan data agar dapat dianalisis dengan akurat dan efisien.

d. Pelabelan Data

Setelah data diproses, langkah selanjutnya adalah pelabelan data, yaitu mengkategorikan setiap ulasan dalam *dataset* ke dalam dua kelas utama: sentimen positif dan sentimen negatif. Proses ini penting untuk membangun *dataset* yang akan digunakan untuk melatih algoritma klasifikasi. Dalam penelitian ini, pendekatan yang digunakan adalah *lexicon-based* yang mengandalkan kamus opini untuk mengidentifikasi sentimen dalam ulasan. Setiap kalimat dalam ulasan akan dianalisis untuk menentukan apakah mengandung sentimen positif atau negatif. Selain itu, pendekatan ini mengandalkan analisis kata-kata kunci yang berfungsi untuk menentukan kategori opini yang terkandung dalam ulasan tersebut.

e. Pembagian Data

dataset yang telah diproses dan dilabeli akan dibagi menjadi dua bagian, yaitu data pelatihan (training data) dan data pengujian (testing data). Dalam penelitian ini, pembagian data dilakukan dengan proporsi 80% untuk data pelatihan dan 20% untuk data pengujian. Dari 50.000 ulasan yang ada, 40.000 data digunakan untuk pelatihan, dan 10.000 data lainnya digunakan untuk pengujian model. Pembagian ini dilakukan untuk memastikan bahwa model yang dikembangkan dapat mengklasifikasikan ulasan dengan baik dan memiliki kemampuan generalisasi yang baik saat diuji dengan data yang

belum pernah dilihat sebelumnya. Dengan pembagian yang proporsional ini, diharapkan model yang dihasilkan memiliki tingkat akurasi yang tinggi dalam mengklasifikasikan sentimen ulasan.

f. Klasifikasi Menggunakan Algoritma *Naïve Bayes*

Setelah data dibagi menjadi data pelatihan dan pengujian, tahap selanjutnya adalah klasifikasi menggunakan algoritma Naïve Bayes. Algoritma ini dipilih karena memiliki keunggulan dalam klasifikasi teks, yaitu efisien dalam waktu komputasi, memiliki asumsi sederhana namun efektif (independensi antar fitur), dan telah terbukti mampu memberikan performa yang baik pada berbagai penelitian sebelumnya, khususnya pada analisis sentimen. Selain itu, model ini sangat cocok untuk diterapkan pada data teks yang direpresentasikan menggunakan TF-IDF [7]. Algoritma Naïve Bayes menentukan kelas sentimen (positif atau negatif) dengan menganalisis peluang kemunculan kata-kata tertentu dalam masing-masing kategori.Dalam penelitian ini, model Naïve Bayes digunakan untuk mengklasifikasikan ulasan film menjadi dua kategori sentimen, yaitu positif dan negatif. Pelatihan model dilakukan pada data latih, sedangkan evaluasi performa dilakukan menggunakan data uji guna mengukur tingkat akurasi prediksinya.

g. Evaluasi Model dengan *Confusion Matrix*

Evaluasi model dilakukan dengan menggunakan *confusion matrix* yang bertujuan untuk mengukur seberapa baik model dalam mengklasifikasikan ulasan. Dalam evaluasi ini, sejumlah metrik digunakan, seperti akurasi (*accuracy*), presisi (*precision*), *recall*, dan *f-measure*. Metrik-metrik ini memberikan gambaran yang lebih lengkap mengenai kualitas prediksi yang dihasilkan oleh model. *Confusion matrix* menunjukkan jumlah prediksi yang benar dan salah, baik untuk kategori positif maupun negatif, yang dapat digunakan untuk menghitung nilai-nilai evaluasi tersebut. Dengan menggunakan metrik ini, peneliti dapat menentukan sejauh mana algoritma *Naïve Bayes* efektif dalam mengklasifikasikan sentimen ulasan film dengan akurat.

h. Hasil dan Kesimpulan

Tahap terakhir dari penelitian ini adalah hasil dan kesimpulan. Pada tahap ini, hasil analisis sentimen dari ulasan film akan disajikan berdasarkan evaluasi yang dilakukan pada model *Naïve Bayes*. Peneliti akan menarik kesimpulan mengenai efektivitas penggunaan *Naïve Bayes* dalam klasifikasi sentimen, serta memberikan rekomendasi untuk perbaikan atau pengembangan lebih lanjut. Evaluasi hasil akan didasarkan pada nilai akurasi, presisi, *recall*, dan *f-measure* yang diperoleh dari *confusion matrix*. Hasil penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan teknologi analisis sentimen, terutama untuk aplikasi yang berhubungan dengan ulasan film. Dengan mengikuti tahapan yang sistematis ini, diharapkan penelitian ini dapat menghasilkan kesimpulan yang valid dan dapat dipercaya, serta memberikan kontribusi yang signifikan bagi pengembangan teknik analisis sentimen berbasis *machine learning*.

HASIL DAN PEMBAHASAN

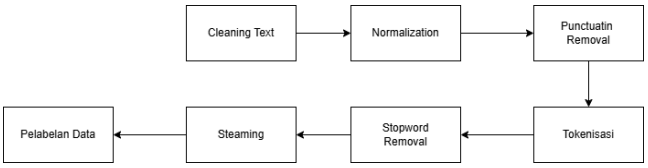
Algoritma yang diterapkan untuk klasifikasi analisis sentimen ulasan film pada *dataset IMDb* adalah algoritma *Naïve Bayes*, karena algoritma ini dikenal efisien dalam memproses teks dalam jumlah besar dengan kecepatan tinggi. Penelitian ini melibatkan beberapa langkah untuk menerapkan model menggunakan algoritma *Naïve Bayes*, yang akan dijelaskan secara rinci seperti berikut:

1. Proses Akuisisi Dataset

dataset yang digunakan dalam penelitian ini diperoleh dari platform *Kaggle*, yang menyediakan kumpulan data ulasan film dari *IMDb*. *dataset* ini dipilih karena memiliki jumlah data yang besar dan telah diberi label sentimen, yaitu positif dan negatif, sehingga cocok digunakan dalam proses pelatihan dan pengujian model klasifikasi. *dataset* terdiri dari 50.000 ulasan film berbahasa Inggris. Data tersebut tersedia dalam format teks dan label yang terstruktur, sehingga memudahkan dalam proses pra-pemrosesan dan implementasi algoritma *Naïve Bayes*. Akuisisi data dilakukan dengan mengunduh langsung file *dataset* dari halaman kompetisi *"IMDb Movie Reviews Dataset"* di *Kaggle* melalui akun pengguna terdaftar [16].

2. *Preprocessing* Data

Tahapan *preprocessing* bertujuan menyiapkan data agar lebih terstruktur dan layak digunakan dalam proses pelatihan maupun pengujian model klasifikasi. Tahapan ini mencakup beberapa langkah penting, seperti pembersihan data dari karakter yang tidak relevan, konversi seluruh teks menjadi huruf kecil (lowercasing), penghapusan tanda baca dan angka, tokenisasi untuk memisahkan kata per kata, serta penghapusan stopword yang tidak memiliki nilai informasi signifikan. Proses ini bertujuan untuk menyederhanakan representasi teks dan meningkatkan kinerja model dalam mengenali pola-pola sentimen dalam ulasan film [17], [18].



Gambar 2. Tahapan Preprocessing

3. *Cleaning* Text

Proses *cleaning* merupakan tahap untuk menghilangkan *noise* dari teks ulasan, seperti karakter spesial, *URL*, *username*, *mention*, *hashtag*, serta spasi berlebih. Tujuan utama tahap ini adalah memastikan bahwa data bersih dari gangguan yang tidak relevan sehingga memudahkan proses analisis dan meningkatkan akurasi model [19]. Hasil data setelah dilakukan proses *cleaning* dapat dilihat pada tabel di bawah.

Tabel 1. *Cleaning Text*

Data Mentah	Hasil *Cleaning*
A wonderful little production. The filming technique is very unassuming- very old-time-BBC …	A wonderful little production. The filming technique is very unassuming- very old-time-BBC fashion …
The cast played Shakespeare. Shakespeare lost. I appreciate that this is trying to bring Shakespeare to the masses …	The cast played Shakespeare.Shakespeare lost.I appreciate that this is trying to bring Shakespeare to the masses, but why ruin something so good …
This movie made it into one of my top 10 most awful movies. Horrible. There wasn't a continuous minute where there wasn't a fight …	This movie made it into one of my top 10 most awful movies. Horrible. There wasn't a continuous minute where there wasn't a fight with …

4. *Normalization Text*

Setelah teks dibersihkan, dilakukan normalisasi, yaitu proses untuk menyamakan format kata agar menjadi bentuk baku. Hal ini sangat penting karena banyak ulasan menggunakan bahasa informal, slang, atau penulisan yang tidak konsisten. Normalisasi dapat membantu model mengenali kata-kata secara konsisten [20], [21]. Oleh karena itu, pada penelitian kali ini dilakukan normalisasi ulasan film yang terdapat pada *dataset*.

Tabel 2. *Normalization Text*

Hasil Cleaning	Hasil *Normalization*
So im not a big fan of Boll's work but then again not many are…	So im not a big fan of Bolls work but then again not many are…
As someone has already mentioned on this board, it's very difficult to make a fake documentary…	As someone has already mentioned on this board, its very difficult to make a fake documentary…
This movie made it into one of my top 10 most awful movies. Horrible. There wasn't a continuous minute where there wasn't a fight with …	this movie made it into one of my top most awful movies. horrible there wasnt a continuous minute where there wasnt a fight with…

5. *Punctuation Removal*

Punctuation Removal berguna menghapus simbol-simbol baca seperti tanda tanya (?), seru (!), titik (.), koma (,) dan tanda baca lain, sehingga teks menjadi bersih dan sistem dapat memproses teks pada *dataset* dengan akurat. Selain itu, semua teks yang terdapat pada *dataset* diubah menjadi huruf kecil, dilakukan agar format seragam dan mudah terbaca. Hasil data setelah dilakukan proses ini ditunjukkan pada tabel 3.

Table 3. *Punctuation Removal*

Hasil *Cleaning*	Hasil *Punctuation removal*
SPOILERS All too, in real life as well as in the movies, familiar story…	spoilers all too in real life as well as in the movies familiar story…
So let's begin!)The movie itself is as original as Cronenberg's movies would usually…	so lets beginthe movie itself is as original as cronenbergs movies would usually…
This movie made it into one of my top 10 most awful movies. Horrible. There wasn't a continuous..,	this movie made it into one of my top most awful movie. horrible there wasnt a continuous…

6. Tokenisasi

Tokenisasi adalah proses pemecahan teks ulasan menjadi unit kata atau token. Hal ini bertujuan agar setiap kata dapat dianalisis secara individu oleh model klasifikasi. Proses ini sangat penting dalam text mining dan *Natural Language Processing* (NLP), termasuk dalam analisis sentimen [22].Hasil data setelah dilakukan Tokenisasi Pada Tabel 4.

Table 4. Tokenisasi

Hasil *Punctuation removal*	Hasil Tokenisasi
a wonderful little production the filming technique is very unassuming very..	['a', 'wonderful', 'little', 'production', 'the', 'filming', 'technique', 'is', 'very', 'unassuming', 'very'..
the cast played shakespeareshakespeare losti appreciate that this is trying to bring shakespeare to the masses but why…	['the', 'cast', 'played', 'shakespeareshakespeare', 'losti', 'appreciate', 'that', 'this', 'is', 'trying', 'to', 'bring', 'shakespeare', 'to', 'the', 'masses', 'but', 'why'...
this movie made it into one of my top most awful movies horrible there wasnt a continuous minute where there wasnt…	['this', 'movie', 'made', 'it', 'into', 'one', 'of', 'my', 'top', 'most', 'awful', 'movies', 'horrible', 'there', 'wasnt', 'a', 'continuous', 'minute', 'where', 'there', 'wasnt'...

7. *Stopword Removal*

Penghapusan *stopword* merupakan tahap penting dalam preprocessing teks, karena *stopword* adalah kata-kata umum yang tidak mengandung makna deskriptif dan sering kali hanya menambah noise dalam analisis teks, terutama pada pendekatan *bag-of-words* [22]. Oleh karena itu, pada tahap preprocessing perlu dilakukan *stopword removal* supaya pada teks kalimat yang diolah tidak ada kata-kata yang tidak diperlukan seperti konjungsi dan determiner.*Stopword* seperti "the", "this", "is", "was", dan "of" umumnya tidak membawa nilai semantik yang kuat terhadap konteks sentimen dalam ulasan film. Dengan menghapus kata-kata tersebut, data menjadi lebih ringkas dan fokus analisis dapat diarahkan pada kata-kata bermakna yang benar-benar menggambarkan opini pengguna. Proses ini juga membantu mempercepat pemrosesan data dan meningkatkan akurasi klasifikasi karena algoritma tidak perlu memproses kata-kata yang bersifat umum dan tidak informatif.Hasil data setelah dilakukan *stopword removal* dapat dilihat padaHasil data setelah dilakukan *stopword removal* paada tabel 5.

Table 5. *Stopword Removal*

Hasil Tokenisasi	Hasil *Stopword Removal*
['a', 'wonderful', 'little', 'production', 'the', 'filming', 'technique', 'is', 'very', 'unassuming', 'very', 'oldtimebbc', 'fashion', 'and', 'gives', …	['wonderful', 'little', 'production', 'filming', 'technique', 'unassuming', 'oldtimebbc', 'fashion', 'gives', …
['the', 'cast', 'played', 'shakespeareshakespeare', 'losti', 'appreciate', 'that', 'this', 'is', 'trying', 'to', 'bring', 'shakespeare', 'to', 'the', 'masses', 'but', 'why', 'ruin', 'something', 'so', 'goodis' …	['cast', 'played', 'shakespeareshakespeare', 'losti', 'appreciate', 'trying', 'bring', 'shakespeare', 'masses', 'ruin', 'something', 'goodis', 'scottish', 'play', 'favorite', …
['this', 'movie', 'made', 'it', 'into', 'one', 'of', 'my', 'top', 'most', 'awful', 'movies', 'horrible', 'there', 'wasnt', 'a', 'continuous', 'minute',…	['movie', 'made', 'one', 'top', 'awful', 'movies', 'horrible', 'continuous', 'minute', 'fight', 'one', 'monster', 'another', …

8. *Stemming*

Tahap terakhir dalam proses *preprocessing* pada penelitian ini adalah *stemming*. *Stemming* merupakan proses pengubahan kata berimbuhan menjadi bentuk dasar tanpa imbuhan seperti awalan, akhiran, atau sisipan. Penggunaan algoritma *stemming* yang tepat sangat penting dalam menyederhanakan representasi kata pada data teks sehingga meningkatkan efektivitas model klasifikasi, termasuk algoritma *Naïve Bayes*. Dengan penerapan *stemming*, ukuran kosakata dapat diminimalkan sehingga model dapat mengenali pola secara lebih optimal [23]..Hasil data setelah dilakukan *Stemming* pada tabel 6.

Table 6. Stemming

Hasil Stopword Removal	Hasil Stemming
['wonderful', 'little', 'production', 'filming', 'technique', 'unassuming', 'oldtimebbc', 'fashion', 'gives', …	['wonder', 'littl', 'product', 'film', 'techniqu', 'unassum', 'oldtimebbc', 'fashion', 'give', …
['cast', 'played', 'shakespeareshakespeare', 'losti', 'appreciate', 'trying', 'bring', 'shakespeare', 'masses', 'ruin', 'something', 'goodis', …	['cast', 'play', 'shakespeareshakespear', 'losti', 'appreci', 'tri', 'bring', 'shakespear', 'mass', 'ruin', 'someth', 'goodi', …
['movie', 'made', 'one', 'top', 'awful', 'movies', 'horrible', 'continuous', 'minute', 'fight', 'one', 'monster', 'another', 'chance', …	['movi', 'made', 'one', 'top', 'aw', 'movi', 'horribl', 'continu', 'minut', 'fight', 'one', 'monster', 'anoth', 'chanc', 'charact', 'develop' …

9. Pelabelan Data

Pada tahap ini, ulasan film diberi label sentimen menggunakan metode *lexicon-based*. Kamus yang dipakai adalah *ID-Opinion Words*, yaitu kumpulan kata-kata yang merepresentasikan sentimen positif dan negatif dalam bahasa Indonesia. Metode ini efektif dalam mengklasifikasikan sentimen teks karena menggunakan daftar kata yang telah ditentukan polarisasinya [24]. Contoh hasil pelabe

lan komentar positif dan negatif disajikan pada tabel 7.

Table 7. Pelabelan Data

Negatif	Positif
basic famili littl boy jake think zombi closet parent fight timebr movi slower soap opera suddenli jake decid becom rambo kill zombiebr …	one review mention watch oz episod youll hook right exactli happen mebr first thing struck oz brutal unflinch scene violenc set right word go …
show amaz fresh innov idea first air first year brilliant thing drop show realli funni anymor continu declin complet …	wonder littl product film techniqu unassum oldtimebbc fashion give comfort sometim discomfort sens realism …
encourag posit comment film look forward watch film bad mistak seen film truli one worst aw almost everi way …	thought wonder way spend time hot summer weekend sit air condit theater watch lightheart comedi plot simplist dialogu …

Implementasi Algoritma *Naïve Bayes*

Implementasi algoritma *Naïve Bayes* dalam penelitian ini dilakukan dengan menggunakan bahasa pemrograman Pyhton. Pyhton memiliki *library Sklearn* yang dapat digunakan untuk mengimplementasikan *Naïve Bayes*. *Sklearn*, merupakan library untuk berbagai metode dan algoritma yang digunakan untuk membuat Machine learning. Baris kode implementasi *Naïve Bayes* dapat dilihat pada gambar dibawah.

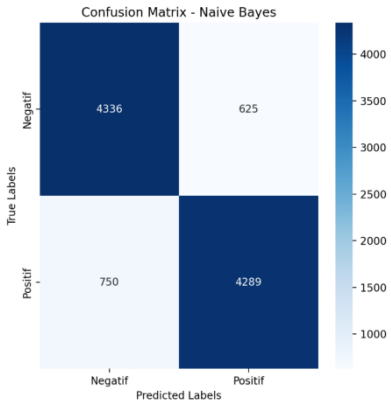
```
from sklearn.naive_bayes import MultinomialNB
nb_model = MultinomialNB()
nb_model.fit(X_train_tfidf, y_train)
y_pred_nb = nb_model.predict(X_test_tfidf)
```

Gambar 3. Kode implementasi naive bayes

Proses implementasi dimulai dengan mengimpor pustaka *Naïve Bayes* dari *Scikit-learn* menggunakan perintah *"from sklearn.naive_bayes import MultinominalNB"* yang digunakan untuk mengimport *library sklearn* untuk model multinominal *Naïve Bayes*, dimana model multinominal ini cocok digunakan untuk data diskrit seperti teks. Baris kode nc_model = MultinominalNB() digunakan untuk membuat model *Naïve Bayes* dan mengubahnya menjadi variabel string Bernama nb_model. Kemudian baris kode nb_model.fit(X_train_tfidf, y_train) digunakan untuk melatih model dengan data latih. Lalu baris kode y_pred_nb = nb_model.predict(X_test_tfidft) digunakan untuk melakukan prediksi dengan data uji menggunakan model yang sebelumnua sudah dilatuh menggunakan data latih.

Evaluasi Model

Evaluasi performa model menggunakan *confusion matrix,* yang merupakan salah satu metode efektif untuk mengukur kinerja metode klasifikasi [23]. Model *confusion matrix* menampilkan hasil salah dan benar pada setiap kelas dengan visualisasi yang detail. Berdasarkan *confusion matrix,* model berhasil mengklasifikasikan ulasan negatif dengan benar sebanyak 4336 data dan ulasan positif dengan benar sebanyak 4289 data. Terdapat kesalahan klasifikasi sebanyak 625 data negatif yang salah diprediksi sebagai positif dan 750 data positif yang salah diprediksi sebagai negatif.Hasil dari *confusion matrix* dapat dilihat padda gambar dibawah.



Gambar 4. *Predicted* Label *Confussion Matrix*

Selain data di atas, hasil dari *confusion matrix* ini juga menghasilkan empat parameter utama yaitu akurasi, *precision, recall, dan F1-score*. Berdasarkan confusion matrix pada model *naive bayes* mencapai akurasi sebesar 0.85 untuk kelas negatif dan 0.87 untuk kelas positif. Nilai recall untuk kelas negatif adalah 0.87 dan untuk kelas positif sebesar 0.85, dengan *F1-score* rata-rata sebesar 0.86 untuk kedua kelas. Secara keseluruhan, akurasi model mencapai 86,25%, atau hasil dapat dilihat pada tabel di bawah ini.

Tabel 8. Evaluasi Model

No	Algortima	Accuracy	Precision	Recall	*F1-score*
		P N	P N	P N	P N
1	Naive Bayes	0.87 0.85	0.87 0.85	0.85 0.87	0.86 0.86

Berdasarkan Tabel 8, model Naïve Bayes mencapai akurasi sebesar 86,25%. Nilai ini menunjukkan bahwa sekitar 86 dari 100 prediksi model sesuai dengan label sentimen yang sebenarnya. Angka ini cukup tinggi, mengingat data yang digunakan bersifat teks alami yang kompleks dan bersifat subjektif.

Dari segi *precision*, nilai tertinggi dicapai pada kelas positif sebesar 0,87 , yang berarti 87% dari prediksi "positif" oleh model memang benar-benar positif. Sebaliknya, precision untuk kelas negatif sedikit lebih rendah, yaitu 0,85, menunjukkan bahwa masih terdapat beberapa prediksi negatif yang keliru. Perbedaan ini bisa disebabkan oleh kemiripan kosakata antara ulasan positif yang lembut dan ulasan negatif ringan, sehingga membingungkan model.

Recall menunjukkan kemampuan model untuk menemukan seluruh data relevan dari masing-masing kelas. Nilai *recall* untuk kelas negatif adalah 0,87 , sementara untuk kelas positif adalah 0,85. Hal ini berarti model lebih mampu mengenali semua ulasan negatif dibandingkan ulasan positif. Ini bisa terjadi karena ulasan negatif sering kali menggunakan kata-kata yang lebih eksplisit dan konsisten yang lebih

mudah dikenali oleh model. Sebaliknya, ulasan positif mungkin lebih bervariasi dan mengandung ambiguitas, seperti penggunaan ironi atau ungkapan netral.

F1-score merupakan metrik gabungan dari precision dan *recall*, dan dalam model ini menghasilkan nilai yang seimbang, yakni 0,86 untuk kedua kelas, baik positif maupun negatif. Nilai ini menunjukkan bahwa model memiliki performa klasifikasi yang stabil dan seimbang dalam menangani kedua jenis sentimen. Meskipun bukan yang tertinggi dibanding algoritma lain dalam beberapa studi, hasil ini tetap menunjukkan bahwa *Naïve Bayes* adalah pendekatan yang efisien dan kompetitif, terutama jika dikombinasikan dengan tahapan prapemrosesan dan ekstraksi fitur yang tepat.

Nilai akurasi 86,25% dihasilkan dari keseluruhan evaluasi model menggunakan *confusion matrix*. Akurasi ini dihitung berdasarkan jumlah prediksi yang benar baik positif maupun negatif dibagi dengan total seluruh prediksi. Dalam penelitian ini, model berhasil mengklasifikasikan 4336 data negatif dengan benar dan 4289 data positif dengan benar. Terdapat kesalahan klasifikasi sebanyak 625 data negatif yang salah diprediksi sebagai positif dan 750 data positif yang salah diprediksi sebagai negatif. Akurasi 86,25% menunjukkan bahwa mayoritas ulasan diklasifikasikan dengan benar. Hasil ini muncul dari kombinasi efektivitas algoritma *Naïve Bayes* dalam klasifikasi teks dan serangkaian tahapan pra-pemrosesan data yang cermat, termasuk pembersihan teks, normalisasi, penghapusan tanda baca, tokenisasi, penghapusan stopword, dan stemming, yang secara signifikan meningkatkan kualitas data masukan untuk model.

KESIMPULAN DAN SARAN

Penelitian ini dilakukan untuk menjawab rumusan masalah mengenai efektivitas algoritma *Naïve Bayes* dalam mengklasifikasikan sentimen positif dan negatif pada ulasan film dari *dataset IMDb* berskala besar. Berdasarkan hasil evaluasi model, algoritma *Naïve Bayes* mampu mengklasifikasikan sentimen dengan akurasi sebesar 86,25%. Nilai *precision* dan *recall* yang seimbang, masing-masing 85%–87%, menunjukkan bahwa model cukup andal dalam menangani kedua kelas sentimen. Proses *preprocessing* yang meliputi pembersihan teks, normalisasi, tokenisasi, penghapusan *stopword*, dan *stemming* turut berkontribusi terhadap peningkatan kinerja model. Dengan demikian, dapat disimpulkan bahwa *Naïve Bayes* merupakan algoritma yang efektif dan efisien untuk klasifikasi sentimen ulasan film.

Namun, model ini memiliki beberapa keterbatasan. *Naïve Bayes* mengasumsikan independensi antar fitur, sehingga kurang mampu menangkap konteks kalimat secara menyeluruh, seperti ironi atau sarkasme. Selain itu, performa model dapat menurun apabila terdapat ketidakseimbangan jumlah data antar kelas, atau ketika berhadapan dengan kosakata yang ambigu dan makna ganda. Kelemahan ini menjadi dasar penting untuk pengembangan penelitian selanjutnya.

Untuk mengatasi keterbatasan model yang ditemukan, disarankan agar penelitian selanjutnya membandingkan performa *Naïve Bayes* dengan algoritma lain seperti *Support Vector Machine* (SVM), Random Forest, atau pendekatan deep learning seperti *LSTM*, yang lebih mampu memahami konteks kalimat. Selain itu, representasi fitur berbasis *TF-IDF* dapat ditingkatkan dengan menggunakan teknik word embedding seperti *Word2Vec* atau *BERT* agar model dapat mengenali makna kata dalam konteks yang lebih dalam.

Selanjutnya, perlu juga dilakukan pengujian terhadap data dengan distribusi sentimen yang tidak seimbang, serta penerapan metode penyeimbangan data seperti *SMOTE*. Penelitian lanjutan juga disarankan untuk memperluas domain pengujian, misalnya pada ulasan produk, layanan publik, atau media sosial, guna mengukur sejauh mana model mampu beradaptasi terhadap variasi bahasa dan konteks di luar domain film.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada semua pihak yang telah memberikan dukungan dan kontribusi dalam pelaksanaan penelitian ini, khususnya kepada institusi dan pembimbing yang telah memberikan arahan dan fasilitas.

DAFTAR PUSTAKA

[1] D. Audia, R. F. Novinta, and F. F. Hidayat, "Pengaruh Situs Web Ulasan Film terhadap Keputusan Menonton Film," *J. Ilmu Komun. APiK*, vol. 7, no. 1, pp. 1–10, 2019, [Online]. Available: https://journal.unpak.ac.id/index.php/apik/article/download/7573/4090

[2] M. A. A. Islamy, I. Indriati, and P. P. Adikara, "Analisis Sentimen *IMDb* Movie Reviews menggunakan Metode Long Short-Term Memory dan FastText," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 9, pp. 4106–4115, 2022, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/11515

[3] N. J. Ramadhan, V. A. P. Putri, and D. A. Riyadi, "Eksplorasi Analisis Sentimen pada Rating Film *IMDb*: Pendekatan Perbandingan menggunakan Decision Tree dan Naive Bayes," *Innov. J. Soc. Sci. Res.*, vol. 4, no. 3, pp. 7273–7286, 2024, [Online]. Available: https://j-innovative.org/index.php/Innovative/article/view/10438

[4] S. Berliani and S. Lestari, "Analisis Sentimen Masyarakat Terhadap Isu Pecat Sri Mulyani Pada Twitter Menggunakan Metode Naive Bayes dan Support Vector Machine," *J. Sains dan Teknol.*, vol. 5, no. 3, pp. 951–960, 2024, [Online]. Available: https://ejournal.sisfokomtek.org/index.php/sain

tek/article/view/2746

[5] P. Nugraha, R. Rasiban, F. M. Sarimole, and Tundo, "Analisis Sentimen Kepuasan Publik Terhadap Masa Kepemimpinan Shin Tae Yong Menggunakan Algoritma *Naïve Bayes*," *J. Teknol. Inf. dan Komun.*, vol. 9, no. 1, pp. 149–158, 2025, [Online]. Available: https://journal.lembagakita.org/index.php/jtik/article/view/3020

[6] N. Ramadhani and N. Fajarianto, "Sistem Informasi Evaluasi Perkuliahan dengan Sentimen Analisis Menggunakan *Naïve Bayes* dan Smoothing Laplace," *J. Sist. Inf. Bisnis*, vol. 10, no. 2, pp. 228–234, 2020, [Online]. Available: https://ejournal.undip.ac.id/index.php/jsinbis/article/view/29799

[7] M. T. Razaq, D. Nurjanah, and H. Nurrahmi, "Analisis Sentimen Review Film Menggunakan Naive Bayes Classifier Dengan Fitur TF-IDF," *eProceedings Eng.*, vol. 8, no. 5, pp. 9007–9016, 2021, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/enginee/article/view/19997

[8] A. Setiawan and R. R. Suryono, "Analisis Sentimen Ibu Kota Nusantara menggunakan Algoritma Support Vector Machine dan Naïve Bayes," *Edumatic J. Pendidik. Inform.*, vol. 8, no. 1, pp. 183–192, 2024, doi: 10.29408/edumatic.v8i1.25667.

[9] D. S. Hidayat, O. Nurdiawan, F. M. Basysyar, and M. Sulaeman, "Peningkatan Model Analisis Sentimen Melalui Algoritma *Naïve Bayes* Berdasarkan Data Komentar YouTube," *J. Manaj. Inform. dan Sist. Inf.*, vol. 8, no. 1, pp. 164–175, 2025, [Online]. Available: https://e-journal.stmiklombok.ac.id/index.php/misi/article/view/1413

[10] F. A. R. Putra, F. F. Fadilah, and U. Enri, "Analisis Sentimen Ulasan Film Oppenheimer Pada Situs *IMDb* Menggunakan Metode *Naïve Bayes*," *Maj. Ilm. UNIKOM*, vol. 21, no. 2, pp. 87–94, 2023, [Online]. Available: https://ojs.unikom.ac.id/index.php/jurnal-unikom/article/view/11338

[11] R. Talibzade, "Sentiment Analysis of *IMDb* Movie Reviews Using Traditional Machine Learning Techniques and Transformers." 2023. [Online]. Available: https://www.researchgate.net/publication/378691218_Sentiment_Analysis_of_IMDb_Movie_Reviews_Using_Traditional_Machine_Learning_Techniques_and_Transformers

[12] T. A. Azzahra *et al.*, "Perbandingan efektivitas *Naïve Bayes* dan SVM dalam menganalisis sentimen kebencanaan di YouTube," *J. Media Inform. Budidarma*, vol. 8, no. 1, pp. 303–311, 2024, [Online]. Available: https://ejurnal.stmik-budidarma.ac.id/index.php/mib/article/view/7186

[13] L. F. Santoso and W. Putra, "The Impact of Text Preprocessing on Sentiment Classification Performance: A Case Study on *IMDb* Dataset," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 4, pp. 11515–11522, 2021, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/download/11515/5107

[14] S. Safitri, D. M. Sari, C. N. Insani, and S. A. Rachmini, "Sistem Kontrol dan Monitoring Pemberi Pakan Ikan Otomatis Berbasis IOT," *J. Manaj. Inform. Sist. Inf. dan Teknol. Komput.*, vol. 1, no. 1, pp. 74–82, 2022, doi: 10.70247/jumistik.v1i1.12.

[15] E. Haerani, A. Rahmatulloh, and K. Layanan, "Analisis Tingkat Kepuasan Mahasiswa Terhadap Simak Universitas," vol. 5, no. 2, pp. 40–46, 2019, doi: 10.70247/jumistik.v1i1.16.

[16] J. Camacho-Collados and M. T. Pilehvar, "On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis," *arXiv Prepr. arXiv1707.01780*, vol. 1707.01780, 2017, [Online]. Available: https://arxiv.org/abs/1707.01780

[17] G. Shalihah, R. Kurniawan, and T. Suprapti, "Pengembangan Sistem Analisis Sentimen Warung Makan Metode *Naïve Bayes* Classifier Berbasis Optimasi Partikel," *J. Teknol. Inf. dan Sist. Inf.*, vol. 5, no. 2, pp. 1–10, 2024, [Online]. Available: https://ojs.stmik-banjarbaru.ac.id/index.php/jutisi/article/viewFile/2058/1194

[18] A. N. Fauziah, "Analisis Sentimen Instagram Vaksinasi Masa Pandemi Menggunakan *Naïve Bayes*," *J. Sist. dan Teknol. Inf. Komun.*, vol. 5, no. 2, pp. 22–26, 2022, [Online]. Available: https://journal.ukmc.ac.id/index.php/jutsi/article/download/787/704/4069

[19] S. Khomsah and A. S. Aribowo, "Text-Preprocessing Model Komentar YouTube dalam Bahasa Indonesia," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 4, pp. 648–654, 2020, [Online]. Available: https://doi.org/10.29207/resti.v4i4.2035

[20] N. A. Putri, R. S. Pratiwi, and A. R. S. Putri, "Analisis Sentimen Terhadap Aplikasi KitaLulus Menggunakan Metode *Naïve Bayes* Classifier," *Smart Comp Jurnalnya Orang Pint. Komput.*, vol. 14, no. 2, pp. 273–280, 2025, [Online]. Available: https://ejournal.poltekharber.ac.id/index.php/smartcomp/article/download/7230/7230

[21] A. N. Hidayat and R. A. Sibarani, "Analisis Sentimen Berdasarkan Hasil Review Lokasi Google Map Menggunakan Metode TextBlob," *J. Akunt. dan Manaj. Indones.*, vol. 5, no. 2, pp. 118–124, 2024, [Online]. Available: https://www.journal.akb.ac.id/index.php/jami/article/download/311/123

[22] A. D. Nugroho, "Analisis Sentimen Terhadap Dampak Inflasi Menggunakan Naive Bayes," *J. BIT*, vol. 5, no. 1, pp. 45–52, 2024, [Online]. Available:

https://journal.fkpt.org/index.php/BIT/article/view/1796

[23] I. M. A. Agastya, "Pengaruh Stemmer Bahasa Indonesia terhadap Performa Analisis Sentimen Terjemahan Ulasan Film," *J. Teknokompak*, vol. 6, no. 2, pp. 45–50, 2018, [Online]. Available: https://ejurnal.teknokrat.ac.id/index.php/teknokompak/article/view/70

[24] Y. Azhar, "Analisa Sentimen Tweet Berbahasa Indonesia Dengan Menggunakan Metode Lexicon Pada Topik Perpindahan Ibu Kota Indonesia," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 1, pp. 15–22, 2020, [Online]. Available: https://www.researchgate.net/publication/343203710_Analisa_Sentimen_Tweet_Berbahasa_Indonesia_Dengan_Menggunakan_Metode_Lexicon_Pada_Topik_Perpindahan_Ibu_Kota_Indonesia