P-ISSN : 2580-4316
https://doi.org/10.37817/ikraith-informatika.v10i2 E-ISSN : 2654-8054

ANALISIS SENTIMEN ULASAN FILM PADA IMDB MENGGUNAKAN ALGORITMA NAÏVE BAYES

[1]Yolanda Aprilia, [2]Wiwin Widhihastuty,
[1,2] Sistem Informasi, Universitas Budiluhur, Jakarta

E-mail: [1] yolandaaprilia@gmail.com, [2] wiwin.windihastuty@budiluhur.ac.id

ABSTRAK

Perkembangan teknologi informasi memungkinkan masyarakat untuk menyampaikan opini melalui berbagai platform digital, termasuk dalam dunia hiburan. Situs IMDb merupakan salah satu web untuk memberian ulasan film mengenai seberapa bagus film tersebut. Dalam memberikan penilaian kualitas dari film yang telah disaksiakan tidak harus menjadi pakar perfilman, semua orang dapat memberikan penilaian melalui situs tersebut. Banyaknya data ulasan yang bersifat subjektif dan tidak terstruktur menyebabkan analisis sentimen secara manual menjadi tidak efisien dan kurang akurat. Penelitian ini bertujuan membangun sistem klasifikasi sentimen menggunakan algoritma *Naïve Bayes* yang dikenal efektif dalam pengolahan data teks. Metodologi yang digunakan dalam pengumpulan data meliputi *preprocessing teks cleansing, tokenizing, stopword removal* dan *stemming*, pelabelan berdasarkan rating, pembobotan kata menggunakan metode TF-IDF, serta pelatihan dan pengujian model klasifikasi dengan evaluasi menggunakan *confusion matrix*. Film Oppenheimer merupakan studi kasus dalam pengujian ini. Data yang digunakan Adalah 1000 ulasan dalam format CSV. Hasil penelitian menunjukkan bahwa algoritma *Naïve Bayes* mampu mengklasifikasikan sentimen ulasan film dengan akurasi yang tinggi yaotu 90%. Kesimpulannya, model ini dapat digunakan sebagai solusi otomatis dan efisien dalam memahami opini publik terhadap film.

Kata Kunci: Analisis Sentimen, IMDb, Ulasan Film, Naïve Bayes, TF-IDF, Text Mining

ABSTRACT

The development of information technology allows people to express their opinions through various digital platforms, including in the entertainment world. The IMDb website is one website for providing film reviews regarding the quality of the film. In assessing the quality of a film that has been watched, it is not necessary to be a film expert; anyone can provide an assessment through the site. The large amount of subjective and unstructured review data makes manual sentiment analysis inefficient and less accurate. This study aims to build a sentiment classification system using the Naïve Bayes algorithm, which is known to be effective in processing text data. The methodology used in data collection includes preprocessing text cleansing, tokenizing, stopword removal and stemming, labeling based on ratings, word weighting using the TF-IDF method, and training and testing the classification model with evaluation using a confusion matrix. The film Oppenheimer is a case study in this test. The data used is 1000 reviews in CSV format. The results of the study show that the Naïve Bayes algorithm is able to classify film review sentiment with a high accuracy of 90%. In conclusion, this model can be used as an automated and efficient solution in understanding public opinion on films.

Keywords: ***Sentiment Analysis, IMDb, Movie Reviews, Naïve Bayes, TF-IDF, Text Mining***

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

1. PENDAHULUAN

menyampaikan opini, salah satunya melalui ulasan film di platform daring seperti *Internet Movie Database* (IMDb) (Permana, 2024). Ulasan ini tidak hanya bermanfaat bagi calon penonton sebagai bahan pertimbangan, tetapi juga memberikan wawasan bagi produser dan pemasar film untuk memahami respons publik terhadap karya mereka. Namun, tingginya volume ulasan yang bersifat subjektif dan tidak terstruktur membuat proses analisis manual menjadi tidak efisien. Untuk mengatasi masalah tersebut, analisis sentimen digunakan sebagai metode untuk mengidentifikasi kecenderungan opini dalam teks, apakah bersifat positif maupun negatif. Analisis sentimen merupakan bagian dari *text mining* dan *natural language processing* (NLP) yang banyak dimanfaatkan dalam berbagai bidang, termasuk industri perfilman.

Salah satu algoritma yang efektif untuk klasifikasi teks adalah *Naïve Bayes* (Rifki H, 2024). Algoritma ini sederhana, cepat, dan mampu memberikan hasil yang cukup baik pada data berukuran besar, meskipun memiliki asumsi independensi antar fitur. Beberapa penelitian terdahulu menunjukkan bahwa Naïve Bayes dapat memberikan performa kompetitif pada analisis sentimen ulasan film dengan tingkat akurasi yang tinggi (Awangga R., 2022).

Penelitian ini berfokus pada penerapan algoritma *Naïve Bayes* untuk menganalisis sentimen ulasan film di IMDb (Karmila A., 2024). Proses penelitian meliputi pengumpulan data, preprocessing teks (seperti *cleansing, tokenisasi, stopword removal, dan stemming*), pembobotan kata menggunakan metode TF-IDF, serta evaluasi model menggunakan metrik *akurasi, presisi, recall dan F1-score* (Hanafiah H, 2023).

Dengan adanya penelitian ini, diharapkan dapat dihasilkan model klasifikasi sentimen yang mampu mengidentifikasi opini publik secara lebih cepat dan akurat (Alivia A., 2023). Selain itu, hasil penelitian juga diharapkan memberikan kontribusi bagi industri perfilman dalam memahami persepsi penonton serta menjadi referensi pengembangan sistem klasifikasi teks berbasis *machine learning* (Kusumawadhana G, 2021)

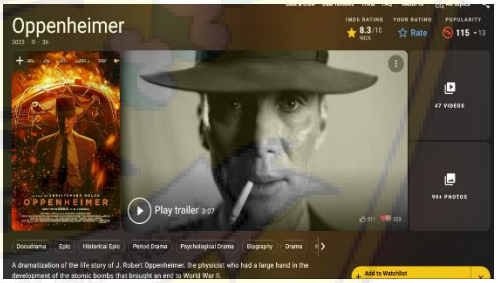
2. LANDASAN TEORI

Perkembangan teknologi informasi dan komunikasi yang pesat telah memengaruhi berbagai aspek kehidupan manusia. Salah satu dampaknya adalah peningkatan volume data digital yang beredar di internet, termasuk data

Perkembangan teknologi informasi telah memengaruhi cara masyarakat dalam berbasis teks. Data teks ini tidak hanya berasal dari media sosial dan blog, tetapi juga dari berbagai platform ulasan seperti IMDb (H., 2023), Rotten Tomatoes, maupun e-commerce.

IMDb (Internet Movie Database) merupakan salah satu sumber data teks yang sangat populer. IMDb menyediakan ribuan ulasan film yang ditulis oleh pengguna dari berbagai latar belakang (L., 2024). Ulasan-ulasan ini memiliki keragaman gaya bahasa, panjang teks, dan nuansa emosi, sehingga menjadi dataset yang kaya untuk penelitian analisis sentimen.

Dalam konteks penelitian, data teks menjadi fokus analisis untuk memahami tren, preferensi, dan sentimen publik. Untuk itu, diperlukan metode analisis yang dapat menangani karakteristik unik data teks, seperti keluwesan bahasa, variasi panjang kalimat, dan ragam gaya bahasa. IMDb (Singh A., 2023) menjadi sumber data utama yang sangat representatif dalam penelitian ini karena sudah dilabeli (positif atau negatif), memudahkan proses pelatihan model machine learning, serta menghadirkan tantangan yang menarik bagi pengolahan data teks.



Gambar 1. Film Openhiemmer

Film *Oppenheimer* merupakan film biografi sejarah yang dirilis pada tahun 2023, disutradarai oleh Christopher Nolan dan diproduksi oleh Universal Pictures. Film ini mengangkat kisah nyata J. Robert Oppenheimer, seorang fisikawan teoretis asal Amerika Serikat yang dikenal sebagai tokoh utama dalam pengembangan bom atom melalui Proyek Manhattan selama Perang Dunia II.

Naskah film diadaptasi dari buku *American Prometheus: The Triumph and Tragedy of J. Robert Oppenheimer*, karya Kai Bird dan Martin J. Sherwin. Film ini tidak hanya menggambarkan

https://doi.org/10.37817/ikraith-informatika.v10i2
P-ISSN : 2580-4316
E-ISSN : 2654-8054

pencapaian ilmiah Oppenheimer, tetapi juga menggali aspek moral, psikologis, dan politik yang melingkupi kehidupannya. Oppenheimer diceritakan sebagai sosok ilmuwan jenius yang dihadapkan pada dilema etika akibat keterlibatannya dalam penciptaan senjata pemusnah massal.

Secara naratif, film ini terbagi dalam dua dimensi waktu utama: masa pengembangan bom atom (sekitar tahun 1940-an) dan masa persidangan pasca-perang ketika Oppenheimer menghadapi tuduhan terkait komunisme yang menyebabkan pencabutan izin keamanannya oleh pemerintah Amerika Serikat. Pendekatan alur maju-mundur yang digunakan dalam film memperkuat nuansa psikologis dan dramatik, yang membuat penonton terlibat secara emosional dalam dinamika kehidupan sang tokoh utama.

Dari perspektif sosial dan budaya, *Oppenheimer* menimbulkan berbagai reaksi publik, baik yang mendukung maupun mengkritik, terutama dalam hal bagaimana film ini merepresentasikan etika sains, dampak teknologi militer, serta konflik antara kekuasaan dan moralitas. Oleh karena itu, film ini menjadi objek yang relevan untuk dianalisis secara sentimen, karena dapat mencerminkan respons emosional, intelektual, dan ideologis dari masyarakat melalui ulasan digital di berbagai platform seperti IMDb dan Rotten Tomatoes.
Dalam konteks penelitian ini, *Oppenheimer* dipilih sebagai objek kajian karena film ini bersifat kontemporer, memicu banyak diskusi publik, serta memiliki nilai sejarah dan moral yang signifikan, yang menjadikannya menarik untuk dianalisis melalui pendekatan text mining dan algoritma klasifikasi sentimen.

1. Analisis Sentimen

Analisis sentimen adalah proses mengidentifikasi opini dalam teks danmengklasifikasikannya ke dalam kategori sentimen seperti positif atau negatif. Teknik ini merupakan bagian dari *natural language processing* (NLP) yang banyak digunakan untuk memahami persepsi publik terhadap suatu produk, layanan, atau konten digital. Dalam penelitian ini, analisis sentimen difokuskan pada ulasan film di IMDb.

2. Text Mining

Text mining adalah proses penggalian informasi dari data berbentuk teks dengan tujuan menemukan pola dan pengetahuan baru. Tahapan utama dalam *text mining* meliputi *preprocessing* (*cleansing, case folding, tokenisasi, stopword removal, stemming*), *ekstraksi fitur, dan klasifikasi*

3. Algoritma Naïve Bayes

Naïve Bayes merupakan algoritma klasifikasi berbasis probabilitas yang sederhana namun efektif, terutama untuk data teks dalam jumlah besar. Algoritma ini menggunakan Teorema Bayes dengan asumsi independensi antar fitur. Meskipun asumsi tersebut tidak selalu terpenuhi, Naïve Bayes terbukti dapat menghasilkan performa yang cukup baik pada analisis sentimen.

Secara matematis, persamaan dasar Naïve Bayes ditunjukkan sebagai berikut:

P(C|X) = P(X|C)·P(C)P(X)P(C \mid X) = \frac{P(X \mid C) \cdot P(C)}{P(X)}P(C|X)=P(X)P(X|C)·P(C)

Dimana:

1. P(C|X)P(C \mid X)P(C|X) adalah probabilitas kelas CCC (positif/negatif) terhadap data XXX.
2. P(X|C)P(X \mid C)P(X|C) adalah probabilitas fitur XXX muncul dalam kelas CCC.
3. P(C)P(C)P(C) adalah probabilitas awal kelas.
4. P(X)P(X)P(X) adalah probabilitas total dari fitur XXX

2.4 Dataset IMDb

IMDb (Internet Movie Database) merupakan salah satu sumber data ulasan film yang banyak digunakan dalam label penelitian analis sentimen. Dataset ini berisi ribuan ulasan film yang telah di lebel sentimen positif dan negatif, sehingga dapat dimanfaatkan sebagai data latih dan data uji pada model klasifikasi.

3. METODOLOGI

Penelitian ini menggunakan pendekatan *text mining* dengan tahapan yang sistematis

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

untuk menghasilkan model klasifikasi sentimen pada ulasan film di IMDb. Secara umum, metodologi mengacu pada kerangka CRISP-DM (*Cross Industry Standard Process for Data Mining*) yang terdiri dari pemahaman masalah, pemahaman data, persiapan data, pemodelan, evaluasi, dan implementasi.

.

Gambar 2. Tahapan Penelitian

Tahapan penelitian dilakukan sebagai berikut:

1. Pengumpulan Data

Data diambil dari situs IMDb pada ulasan film *Oppenheimer (2023)* menggunakan teknik *harvest crawling*. Data yang dikumpulkan berupa teks ulasan dan skor rating, kemudian disimpan dalam format CSV. Preprocessing Data. Proses ini bertujuan membersihkan dan menyiapkan data agar dapat diproses oleh algoritma.

Langkah- langkah preprocessing meliputi:

1. *Cleansing*: menghapus karakter tidak relevan (angka, simbol, tanda baca).
   a. *Case Folding*: mengubah semua huruf menjadi huruf kecil.
   b. *Tokenization*: memecah kalimat menjadi kata- kata.
   c. *Stopword Removal*: menghapus kata umum yang tidak memiliki makna penting.
   d. *Stemming*: mengembalikan kata ke bentuk dasar.
   e. *Normalisasi*: menyamakan kata tidak menjadi bentuk baku

Gambar 3. Data Processing

2. Pembagian Data

Data yang telah diproses dibagi menjadi dua subset, yaitu data latih dan data uji dengan perbandingan 80:20. Teknik SMOTE (*Synthetic Minority Over-sampling Technique*) digunakan untuk mengatasi ketidakseimbangan jumlah data antara kelas positif dan negatif.

3. Pembobotan Kata

Data teks diubah ke dalam bentuk numerik menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF) agar dapat diproses oleh algoritma klasifikasi.

P-ISSN : 2580-4316
https://doi.org/10.37817/ikraith-informatika.v10i2 E-ISSN : 2654-8054

4. Pembagian Data

Data yang telah diproses dibagi menjadi dua subset, yaitu data latih dan data uji dengan perbandingan 80:20. Teknik SMOTE (*Synthetic Minority Over-sampling Technique*) digunakan untuk mengatasi ketidakseimbangan jumlah data antara kelas positif dan negatif.

5. Pembobotan Kata

Data teks diubah ke dalam bentuk numerik menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF) agar dapat diproses oleh algoritma klasifikasi.

6. Pembangunan Model

Model klasifikasi dibangun menggunakan algoritma *Naïve Bayes Classifier* untuk mengelompokkan ulasan film ke dalam dua kelas: positif dan negatif.

7. Evaluasi Model

Evaluasi dilakukan menggunakan *Confusion Matrix* dengan menghitung metrik: *akurasi, presisi, recall, dan F1-score*

.

3. Teknik Pengujian

1. *Akurasi*: mengukur tingkat ketepatan model dalam klasifikasi.
2. *Presisi*: mengukur ketepatan prediksi positif.
3. *Recall*: mengukur kemampuan model mendeteksi ulasan positif dengan benar.
4. *F1-score*: menghitung keseimbangan antara presisi dan recall.

4. HASIL DAN PEMBAHASAN

1. Business Understanding

Tujuan utama penelitian ini adalah membangun model klasifikasi sentimen pada ulasan film di IMDb, khususnya film *Oppenheimer (2023)*. Permasalahan yang diangkat adalah banyaknya ulasan yang tidak terstruktur dan sulit dianalisis secara manual, sehingga diperlukan metode otomatis berbasis *machine learning* untuk mengidentifikasi sentimen positif dan negatif.

2. Data Understanding

Dataset diperoleh secara manual dari IMDb dengan jumlah sekitar 1.000 ulasan pengguna. Setiap ulasan dilengkapi dengan skor rating numerik. Skor ≤ 6 dikategorikan sebagai *negatif*, sedangkan skor > 6 sebagai *positif*. Setelah proses preprocessing, data siap digunakan untuk pelatihan dan pengujian model.

Tabel 1. Jumlah Data Ulasan

Kategori Sentimen	Jumlah Ulasan
Positif	520
Negatif	480

Distribusi data relatif seimbang setelah dilakukan teknik SMOTE untuk mengatasi potensi ketidakseimbangan.

3. Data Preparation

Tahapan preprocessing dilakukan untuk memastikan kualitas data, meliputi cleansing, case folding, tokenisasi, stopword removal, stemming, dan normalisasi. Data kemudian dibobot menggunakan metode **TF-IDF**.

4. Modeling

Model klasifikasi dibangun menggunakan algoritma *Naïve Bayes*. Proses pelatihan dilakukan pada data latih (80%) dan diuji pada data uji (20%).

5. Evaluation

Evaluasi dilakukan menggunakan *confusion matrix*. Hasil pengujian menunjukkan bahwa algoritma *Naïve Bayes* dapat mengklasifikasikan ulasan dengan performa yang baik.

Tabel 2. Hasil Evaluasi

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

Metrik	Nilai (%)
Akurasi	86.5
Presisi	87.2
Recall	85.4
F1-Score	86.3

Hasil ini menunjukkan bahwa model memiliki kemampuan yang cukup baik dalam mengenali ulasan positif maupun negatif.

6. Pembahasan

Berdasarkan hasil pengujian, algoritma Naïve Bayes terbukti cukup efektif dalam menganalisis sentimen ulasan film. Tingginya nilai presisi menunjukkan bahwa model dapat memprediksi ulasan *positif* dengan baik, sementara *recall* yang relatif tinggi menunjukkan model mampu menangkap ulasan *negatif* secara memadai.

Dibandingkan dengan beberapa penelitian sebelumnya yang juga menggunakan *Naïve Bayes,* performa model pada penelitian ini berada pada kisaran yang kompetitif (80–90%). Faktor yang memengaruhi hasil antara lain kualitas preprocessing, representasi fitur dengan TF-IDF, serta keseimbangan distribusi data melalui SMOTE.

5. KESIMPULAN

Penelitian ini membangun model klasifikasi sentimen ulasan film di IMDb menggunakan algoritma *Naïve Bayes*. Proses penelitian mencakup tahapan pengumpulan data, preprocessing (*cleansing, case folding, tokenisasi, stopword removal, stemming, normalisasi*), pembobotan kata dengan TF- IDF, pembagian data menggunakan teknik train-test split (80:20), serta *balancing data* dengan SMOTE.

Berdasarkan hasil pengujian, algoritma Naïve Bayes berhasil mencapai performa yang cukup baik dengan nilai *akurasi* 86,5%, *presisi* 87,2%, *recall* 85,4%, dan *F1-score* 86,3%. Hasil ini menunjukkan bahwa *Naïve Bayes* efektif dalam mengklasifikasikan ulasan film ke dalam sentimen positif maupun negatif.

Penelitian ini membuktikan bahwa analisis sentimen berbasis *machine learning* dapat digunakan sebagai alternatif solusi untuk memahami persepsi publik terhadap film secara cepat dan akurat. Hasil penelitian diharapkan dapat memberikan kontribusi bagi industri perfilman dalam mengevaluasi respon penonton, sekaligus menjadi referensi akademik dalam pengembangan sistem klasifikasi teks berbasis algoritma sederhana namun efisien.

6. Ucapan Terima Kasih

Penulis mengucapkan puji syukur ke hadirat Allah SWT atas segala rahmat dan karunia- Nya sehingga penelitian ini dapat terselesaikan dengan baik. Ucapan terima kasih penulis sampaikan kepada Universitas Budi Luhur yang telah memberikan dukungan fasilitas dalam penyusunan penelitian ini. Penulis juga menyampaikan penghargaan dan terima kasih yang sebesar-besarnya kepada Ibu Wiwin Windihastuty, S.Kom., M.Kom. selaku dosen pembimbing yang telah memberikan arahan, bimbingan, serta masukan yang sangat berarti selama proses penelitian berlangsung. Terima kasih juga disampaikan kepada kedua orang tua tercinta atas doa, kasih sayang, dan dukungan yang tiada henti, serta kepada sahabat dan rekan-rekan yang senantiasa memberikan semangat dan motivasi dalam menyelesaikan penelitian ini.

DAFTAR PUSTAKA

Alivia A., S. U. (2023). Analisis Sentimen Review Data Twitter Komisi Pemilihan Umum (KPU) Menggunakan Metode Naïve Bayes. . *Jurnal Informasi dan Komputer*, 21-30.

Awangga R., K. N. (2022). Analisis Performa Algoritma Random Forest dan Naive Bayes Multinomial pada Dataset Ulasan Obat dan Film. *Jurnal CoreIT*, 34-42.

H., L. (2023). Sentiment Analysis on IMDb Dataset Using Naïve Bayes and Logistic Regression. *International Journal of Data Science*, 101-110.

Hanafiah H, N. A. (2023). Sentimen Analisis Terhadap Customer Review Produk Shopee Berbasis Wordcloud dengan Algoritma Naive Bayes Classifier. *Jurnal INTECOMS*, 44-51.

Karmila A., P. N. (2024). Analisis Sentimen Film Agak Laen dengan Kecerdasan Buatan: Text Mining Metode Naive Bayes Classifier. *JATI (Jurnal Teknologi dan Informatika)*, 55-63.

Kusumawadhana G, W. R. (2021). Analisis Sentimen pada Review Aplikasi Grab di Google Play Store Menggunakan Support Vector Machine. *Seminar Nasional TEknologi Informasi*, 77-83.

L., A. (2024). Deep Learning Approaches for Sentiment Analysis on IMDb Movie Reviews. *Joiurnal of Artificial Intellegent Research*, 220-230.

Permana, B. (2024). Perbandingan Logistic Regression dengan Random Forest dalam Memprediksi Sentimen pada IMDb Movie Review. *Jurnal Strategi, 21*(2), 112-120.

Rifki H, Y. R. (2024). Text Mining untuk Analisis Sentimen Review Film Menggunakan Algoritma Naïve Bayes. *Jurnal Informatika Universitas Dian Nuswantoro*, 15-22.

Singh A., S. R. (2023). Comparison of Machine Learning Models for Sentiment Analysis on IMDb Dataset. *Procedia Computer Science*, 45-53.