Klasifikasi Genre Film Berbasis Sinopsis dengan Pendekatan NLP dan Naïve Bayes

Noviana Riza, Iga Putri Ramadhani, Dini Aprilia Cahyani*, Kiki Mustaqim
Universitas Logistik dan Bisnis Internasional (ULBI)

Received: 2 September 2026 Accepted: 14 October 2026 Published: 1 December 2026

*apriliadini206@gmail.com

Kata Kunci: Klasifikasi Genre Film, Naïve Bayes, Natural Language Processing, TF-IDF, Streamlit

DSI: Jurnal Data Science Indonesia is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

Abstrak: Pesatnya kemajuan industri perfilman mendorong kebutuhan akan sistem otomatis yang dapat mengklasifikasikan genre film secara efisien dan tepat. Penentuan genre secara manual berpotensi menghasilkan penilaian yang subjektif, membutuhkan waktu yang relatif lebih lama, serta memiliki tingkat efisiensi yang rendah, seiring dengan meningkatnya jumlah film yang diproduksi dan dipublikasikan. Dengan demikian, penelitian ini bermaksud untuk melakukan klasifikasi genre film dengan berdasarkan teks sinopsis dengan menerapkan algoritma Naïve Bayes yang didukung oleh pendekatan Natural Language Processing (NLP). Data penelitian diperoleh dari TMDB Movies Dataset yang memuat informasi judul, sinopsis, keywords, dan genre film. Setelah melalui proses seleksi data, preprocessing, serta konversi label multi-label menjadi single-label, diperoleh 173.524 data film yang digunakan dalam penelitian. Klasifikasi difokuskan pada lima genre, yaitu Drama, Documentary, Comedy, Horror, dan Animation. Data sinopsis dan keywords digabungkan, kemudian ditransformasikan menggunakan pembobotan Term Frequency–Inverse Document Frequency (TF-IDF). Algoritma Multinomial Naïve Bayes diterapkan untuk membangun model klasifikasi dengan membagi dataset menjadi 80% data latih dan 20% data uji. Selain itu, metode random undersampling diterapkan pada data latih untuk menyeimbangkan jumlah sampel pada setiap kelas genre sebelum proses pelatihan model. Model yang dikembangkan berhasil memperoleh akurasi sebesar 83% dan memberikan performa terbaik pada genre documentary. Sehingga model yang telah dibangun diintegrasikan ke dalam aplikasi berbasis Streamlit untuk mendukung proses prediksi genre film secara interaktif.

PENDAHULUAN

Industri perfilman yang mengalami perkembangan pesat menyebabkan produksi film terus meningkat dari tahun ke tahun. Dalam konteks ini, genre film memiliki peranan penting karena mempengaruhi preferensi penonton serta strategi pemasaran film[1]. Penentuan genre film umumnya masih dilakukan secara manual berdasarkan interpretasi pembuat film atau tim kurator. Proses tersebut memiliki kelemahan karena bersifat subjektif, membutuhkan waktu yang lama, serta berpotensi menimbulkan ketidakkonsistenan, terutama ketika jumlah film yang diproduksi terus meningkat[2]. Berdasarkan kondisi tersebut, diperlukan sistem otomatis yang mampu mengidentifikasi genre film secara efektif dan presisi. Keunggulan algoritma Naïve Bayes terletak pada kemampuannya dalam menangani dataset berukuran besar dengan tetap mempertahankan tingkat ketepatan yang tinggi [3][4][5].

Sinopsis dapat dimanfaatkan dalam klasifikasi genre film karena memberikan gambaran mengenai alur cerita dan tema yang menjadi ciri suatu genre. Sinopsis tersebut dipandang sebagai bentuk data teks yang mampu merepresentasikan alur, tema, dan karakteristik film[6] sehingga dimanfaatkan secara luas sebagai sumber informasi utama dalam penelitian klasifikasi genre berbasis teks[7][8][9]. Pengolahan data teks seperti sinopsis memerlukan pendekatan khusus agar dapat dianalisis secara

komputasi. Natural Language Processing (NLP) pada penelitian ini membantu mengidentifikasi fitur-fitur serta informasi yang terkandung dalam teks bahasa alami sehingga data dapat diproses oleh sistem komputasi secara efektif [10]. Beberapa penelitian sebelumnya telah memanfaatkan teks sinopsis sebagai sumber utama dalam klasifikasi genre film dengan menerapkan pendekatan Natural Language Processing (NLP) yang dipadukan dengan algoritma pembelajaran mesin untuk meningkatkan akurasi klasifikasi.

Dalam konteks pengolahan data teks perfilman, algoritma Naïve Bayes terbukti efektif dan secara luas diterapkan untuk mengelompokkan kategori narasi berbasis frekuensi kemunculan kata[11]. Cara kerja algoritma ini bertumpu pada prinsip probabilitas Teorema Bayes yang didasarkan pada asumsi ketidakbergantungan antar fitur. Keunggulan metode ini terletak pada kemudahan implementasi, serta tetap efektif dalam mengestimasi parameter meskipun jumlah data pelatihan terbatas, sehingga mampu menghasilkan performa yang baik pada permasalahan dunia nyata yang bersifat kompleks[12][13].

Meskipun berbagai penelitian telah dilakukan terkait klasifikasi genre film berbasis teks, sebagian besar studi sebelumnya masih diuji pada dataset berskala terbatas atau hanya menggunakan satu jenis fitur teks saja. Sebagai contoh, klasifikasi genre berbasis sinopsis oleh[14] pada 1.000 data TMDB menghasilkan akurasi sebesar 59,00%, sementara studi oleh[11] pada 811 data IMDb memperoleh akurasi sebesar 40,49%. Namun demikian, studi-studi tersebut maupun pendekatan berbasis optimasi hyperparameter[5] belum mengintegrasikan fitur keywords secara optimal untuk memperkaya informasi sinopsis pada dataset berskala besar. Tantangan utama dalam penelitian ini terletak pada bagaimana memproses representasi gabungan teks sinopsis dan keywords agar dapat diklasifikasikan secara presisi dan efisien, serta mengatasi permasalahan multi-label dan ketidakseimbangan kelas genre.

Guna menjawab tantangan tersebut, penelitian ini mengintegrasikan fitur sinopsis dan keywords pada TMDB Movies Dataset berskala besar menggunakan pendekatan Natural Language Processing (NLP) dan algoritma Multinomial Naïve Bayes berbasis pembobotan TF-IDF. Pendekatan rule-based keyword matching diterapkan bersamaan dengan teknik penyeimbangan data latih untuk mengoptimalkan performa klasifikasi single-label. Selain menghasilkan model klasifikasi yang presisi, penelitian ini juga menghadirkan antarmuka aplikasi interaktif berbasis Streamlit yang terhubung ke basis data guna memfasilitasi prediksi genre secara real-time. Hasil penelitian ini diharapkan dapat memberikan kontribusi nyata bagi pengembangan bidang pengolahan bahasa alami, sekaligus menjadi solusi yang efektif dan efisien dalam mengklasifikasikan genre film di industri perfilman.

TINJAUAN LITERATUR

Penelitian mengenai klasifikasi genre film berbasis teks sinopsis telah berkembang menggunakan beragam pendekatan pembelajaran mesin konvensional hingga jaringan saraf tiruan[7][8][9]. Penerapan arsitektur deep learning seperti Long Short-Term Memory(LSTM) terbukti mampu memprediksi genre film dengan menangkap ketergantungan urutan kata pada sinopsis[7][8]. Namun demikian, model berbasis LSTM membutuhkan sumber daya komputasi yang besar serta waktu pelatihan yang relatif lama[7][8]. Di sisi lain, evaluasi kinerja algoritma pembelajaran mesin konvensional seperti K-Nearest Neighbors (KNN), Support Vector Machine (SVM), dan Naïve Bayes pada klasifikasi genre berbasis sinopsis menunjukkan bahwa Naïve Bayes memiliki efektivitas yang lebih baik dan efisien dalam menangani data teks dibandingkan algoritma berbasis jarak seperti KNN[9][14]. Naïve Bayes terbukti lebih unggul dibanding KNN pada dataset sinopsis seimbang karena kemampuannya dalam memanfaatkan pola probabilitas kosakata pada masing-masing genre[14]. Penerapan optimasi hyperparameter[5], serta eksplorasi arsitektur lain seperti FastText, BiGRU[3], dan mekanisme attention[4] juga turut dikembangkan untuk mengoptimalkan klasifikasi genre. Kendati metode deep learning menawarkan kapabilitas pemrosesan kompleks, algoritma Multinomial Naïve Bayes tetap menjadi pilihan yang sangat efisien dan andal dalam pengolahan dataset berbasis teks berskala besar[5][11].

Natural Language Processing (NLP) memfasilitasi sistem komputasi untuk memahami dan mengolah bahasa alami manusia[15][16]. Tahapan preprocessing seperti case folding, pembersihan karakter, tokenization, dan stopword removal berperan krusial dalam meminimalkan noise dan meningkatkan kualitas fitur[17][18]. Untuk mentransformasikan teks ke dalam bentuk numerik, teknik

Term Frequency-Inverse Document Frequency (TF-IDF) digunakan untuk memberikan bobot pada setiap kata berdasarkan relevansinya dalam dokumen dan korpus data[19]. Penggabungan elemen fitur teks seperti sinopsis dan keywords mampu memperkaya representasi informasi sehingga meningkatkan keterbedaan antar kelas[6][20].

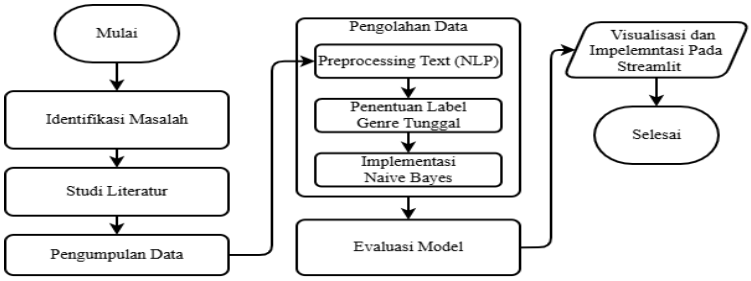
Algoritma Naïve Bayes merupakan metode klasifikasi berbasis probabilitas mengadaptasi Teorema Bayes dengan asumsi ketidakbergantungan antar fitur[12][21]. Varian Multinomial Naïve Bayes dirancang khusus untuk menangani data berbasis frekuensi kemunculan kata atau bobot numerik seperti TF-IDF[13]. Algoritma memiliki keunggulan berupa efisiensi komputasi, kemudahan implementasi, serta performa yang stabil pada dataset berdimensi tinggi[22]. Oleh karena itu, Multinomial Naïve Bayes sangat relevan diterapkan untuk klasifikasi genre film berbasis teks sinopsis dan keywords[5][9]. Algoritma tersebut salah satu algoritma metode klasifikasi dengan memanfaatkan pendekatan probabilitas statistik yang ditemukan oleh Thomas Bayes[21]. Berikut merupakan rumus algoritma Naïve Bayes.

$$P\ (C|X)\ = \frac{(P(X|C)) \times (P(C))}{P(X)} \text{.....(1)}$$

Keterangan:
C = Kelas atau Kategori target
X = Fitur, karakteristik, atau data teks
P(C|X) = Probabilitas posterior, yaitu peluang sebuah kelas C terpilih berdasarkan bukti fitur teks X yang diamati
P(X|C) = Kemungkinan, yaitu peluang munculnya bukti fitur teks X pada kelas tertentu C
P(C) = Probabilitas prior, yaitu peluang awal munculnya kelas C secara umum sebelum ada bukti teks
P(X) = Probabilitas prediktor, yaitu peluang total kemunculan fitur teks X di seluruh dataset

METODE PENELITIAN

Pendekatan yang digunakan pada penelitian ini yaitu kuantitatif, dengan didukung oleh metode klasifikasi machine learning dalam rangka menentukan genre film berdasarkan teks sinopsis. Proses yang dilakukan yaitu pengumpulan dataset, pra pemrosesan teks, ekstraksi fitur, pembangunan model klasifikasi, evaluasi performa model, serta implementasi sistem. Berikut merupakan alur proses penelitian pada gambar 1.



Gambar 1. Flowchart

1. Pengumpulan Data
   Data yang digunakan bersumber dari TMDB Movies Dataset, yang memiliki kolom judul, sinopsis, keywords, dan genre. Data sinopsis dan kolom keywords yang menjadi sumber utama dalam tahap melakukan proses klasifikasi, sedangkan genre dijadikan sebagai label kelas. Karena genre pada dataset bersifat multi-label, dilakukan konversi menjadi label tunggal dengan memilih genre utama agar proses klasifikasi dapat dilakukan secara supervised.
2. Preprocessing Data dengan NLP
   Pada tahap ini, data teks di praproses untuk meningkatkan kualitas serta konsistensinya sebelum dilakukan analisis lebih lanjut [17], [22], khususnya dalam konteks pengolahan teks menggunakan pendekatan NLP [19]. Proses ini meliputi pembersihan teks dari karakter non-alfabet, pengubahan melibatkan tahapan penyeragaman dengan mengkonversi huruf ke lowercase, tokenisasi kata, menghapus kata yang tidak memberikan informasi penting dalam teks (stopwords), serta stemming untuk memperoleh kata dasar [18]. Pada penelitian ini, tahapan preprocessing yang diterapkan

meliputi data cleaning, case folding, pembersihan karakter non-huruf, serta penggabungan teks overview dan keywords. Proses tokenisasi dan penghapusan stopwords dilakukan secara otomatis pada tahap ekstraksi fitur menggunakan TF-IDF Vectorizer.

a. Data Cleaning
   Tahap data cleaning yang dilakukan yakni yang tidak memiliki nilai pada atribut overview dan genres dihapus, sedangkan atribut keywords yang kosong diisi dengan string kosong. Tahap ini bertujuan demi menjamin bahwa data yang digunakan benar dan lengkap.
b. Case folding
   Tahap case folding pada penelitian ini yakni teks sinopsis dan kata kunci dikonversi ke huruf kecil agar arti tidak berubah karena kapitalisasi.
c. Pembersihan karakter Non-Huruf
   Pembersihan yang dilakukan yakni seperti angka dan simbol dihapus menggunakan ekspresi reguler. Spasi berlebih juga dirapikan agar teks menjadi lebih bersih dan konsisten.
d. Penggabungan teks
   Teks sinopsis (overview) dan keywords digabungkan menjadi satu atribut teks agar informasi sehingga representasi data pada proses klasifikasi menjadi lebih optimal.
e. Tokenisasi dan stopword removal otomatis
   TF-IDF vectorizer secara otomatis menyaring kata-kata umum pada proses ekstraksi fitur. Kata "the", "and", dan "a", dihilangkan karena dianggap kurang memberikan informasi yang relevan, sehingga kualitas fitur dapat ditingkatkan.
f. TF-IDF vectorization
   Tahap akhir preprocessing adalah transformasi teks ke dalam bentuk numerik menggunakan metode TF-IDF melalui TfidfVectorizer. Parameter yang digunakan adalah max_features=15000 dan stop_words='english', sedangkan parameter min_df dan max_df tidak ditentukan secara eksplisit. Ekstraksi fitur dilakukan setelah dataset dibagi menjadi data training dan testing. Proses fit_transform diterapkan hanya pada data training yang telah melalui proses penyeimbangan kelas untuk membentuk kosakata dan bobot TF-IDF. Selanjutnya, data testing hanya diproses menggunakan transform berdasarkan kosakata dan bobot yang diperoleh dari data training. Dengan demikian, data testing tidak digunakan dalam proses pembentukan representasi TF-IDF sehingga potensi kebocoran data (data leakage) dapat dihindari.

3. Penentuan Label Genre Tunggal
   Dataset film yang digunakan memiliki atribut genres yang dapat berisi lebih dari satu genre untuk setiap film. Maka dilakukan tahap penyederhanaan pada label dengan menentukan satu genre utama (single genre) untuk setiap film. Penentuan genre tunggal dilakukan dengan rule-based menggunakan daftar kata kunci tertentu yang mewakili masing-masing genre utama, yaitu Drama, Documentary, Comedy, Horror, dan Animation. Genre yang memiliki jumlah kemunculan kata kunci terbanyak dalam teks sinopsis dan keywords akan dipilih sebagai label akhir. Langkah tersebut untuk memudahkan proses klasifikasi algoritma Naïve Bayes yang bekerja secara optimal pada satu label kelas.
4. Implementasi Naïve Bayes
   Pada proses ini, algoritma Multinomial Naïve Bayes diterapkan untuk melakukan prediksi genre film. Dataset dialokasikan menjadi data latih dan data uji dengan proporsi 80:20. Sebelum pelatihan model, dilakukan penyeimbangan data (data balancing) khusus pada data latih dengan cara resampling agar setiap kelas genre memiliki jumlah data yang seimbang. Selanjutnya, data latih tersebut dilatih menggunakan model Multinomial Naïve Bayes dengan fitur TF-IDF dan parameter alpha (Laplace smoothing) $\alpha = 0{,}3$. Berikut merupakan rumus dasar metode Naïve Bayes.

$$P\ (C|X)\ =\ \frac{(P(X|C))\ x\ (P(C))}{P(X)} \text{.....(2)}$$

5. Evaluasi Model
   Evaluasi model dilakukan untuk mengukur performa dan akurasi model Multinomial Naïve Bayes yang telah dilatih dalam memprediksi genre film. Pengujian dilakukan menggunakan data uji (*data testing*) sebanyak 34.705 sampel yang bersifat terisolasi dan tetap dipertahankan distribusi aslinya tanpa melalui proses penyeimbangan (*balancing*). Metrik evaluasi yang digunakan untuk mengukur

kinerja model meliputi Accuracy, Precision, Recall, dan F1-Score (baik macro average maupun weighted average).

6. Visualisasi dan Implementasi

Model Naïve Bayes yang telah dilatih dan dievaluasi kemudian disimpan menggunakan pustaka joblib. Model ini diimplementasikan pada Streamlit. hal tersebut dapat membantu pengguna saat memasukkan teks sinopsis film baru, kemudian sistem akan melakukan preprocessing serupa dan menampilkan hasil prediksi genre film. Implementasi yang dirancang untuk mendukung pengguna dalam memahami serta memanfaatkan hasil penelitian secara interaktif.

HASIL PENELITIAN

Pada bagian ini dipaparkan temuan penelitian dan pembahasan yang dihasilkan melalui penggunaan metode Multinomial *Naïve Bayes* dalam proses klasifikasi genre film berdasarkan teks sinopsis. Pembahasan difokuskan pada hasil pengolahan awal data teks, serta tahapan ekstraksi fitur dengan metode TF-IDF, pelatihan dan pengujian model, serta evaluasi kinerja model klasifikasi. Hasil tersebut selanjutnya kemudian dianalisis untuk mengevaluasi kinerja model dalam mengklasifikasikan genre film serta faktor-faktor yang mempengaruhi performa klasifikasi berdasarkan karakteristik data teks yang digunakan.

Untuk mendukung proses klasifikasi genre film, sumber data yang digunakan adalah TMDB Movies Dataset yang berisi informasi terkait film. Dari dataset tersebut, dilakukan seleksi atribut yang selaras dengan tujuan penelitian, yaitu judul film, teks sinopsis atau *overview, keywords*, dan genre film. Atribut sinopsis atau *overview* dan *keywords* digunakan sebagai sumber data teks utama dalam proses pengolahan dan pembentukan fitur, sedangkan atribut genre digunakan sebagai label kelas pada proses klasifikasi. Sebagai gambaran awal terhadap data yang menjadi objek penelitian, Tabel 1 menyajikan lima contoh data film dari dataset awal berdasarkan atribut terpilih, yaitu judul film, sinopsis (overview), genre, dan keywords. Dataset awal memiliki karakteristik multi-label, yaitu satu film dapat memiliki lebih dari satu genre. Oleh karena itu, genre yang ditampilkan pada Tabel 1 masih mencakup berbagai kategori dan belum terbatas pada lima genre yang digunakan dalam proses klasifikasi.

Tabel 1. Lima Contoh Data Film Teratas Berdasarkan Atribut Terpilih

Title	Overview	genres	keywords
Inception	Cobb, a skilled thief who commits corporate es...	Action, Science Fiction, Adventure	rescue, mission, dream, airplane, paris, franc...
Interstellar	The adventures of a group of explorers who mak...	Adventure, Drama, Science Fiction	rescue, future, spacecraft, race against time,...
The Dark Knight	Batman raises the stakes in his war on crime. ...	Drama, Action, Crime, Thriller	joker, sadism, chaos, secret identity, crime f...
Avatar	In the 22nd century, a paraplegic Marine is di...	Action, Adventure, Fantasy, Science Fiction	future, society, culture clash, space travel, ...
The Avengers	When an unexpected enemy emerges and threatens...	Science Fiction, Action, Adventure	new york city, superhero, shield, based on com...

Selanjutnya akan dilakukan seleksi genre dan konversi dari multi-label menjadi single-label untuk menyesuaikan data dengan kebutuhan klasifikasi. Lima genre yang digunakan dalam penelitian ini adalah Drama, Documentary, Comedy, Horror, dan Animation. Proses penentuan label tunggal dilakukan menggunakan pendekatan rule-based berdasarkan kemunculan kata kunci pada teks gabungan (text_combined) yang terdiri atas sinopsis dan keywords.

1. Hasil *Preprocessing* dan Pembentukan Fitur

Proses preprocessing dilakukan guna meningkatkan kualitas data teks sebelum dimanfaatkan dalam pembangunan model klasifikasi genre film. Data teks yang digunakan pada penelitian ini berupa sinopsis dan keywords film yang diperoleh dari TMDB Movies Dataset. Kedua sumber data teks tersebut digabungkan untuk memperkaya representasi informasi yang digunakan sebagai fitur dalam proses klasifikasi.

Proses preprocessing meliputi data cleaning, seperti pengubahan seluruh teks menjadi huruf kecil (case folding), pembersihan karakter non-huruf, serta penggabungan teks sinopsis (overview) dan keywords. Selain itu, proses tokenisasi dan penghapusan stopword dilakukan secara otomatis pada

tahap ekstraksi fitur menggunakan TF-IDF Vectorizer. Tahapan tersebut dilakukan untuk meminimalkan noise dan menghasilkan teks yang lebih bersih serta konsisten sebelum ditransformasikan ke dalam bentuk numerik. Berikut merupakan Tabel 2 mengenai Informasi Jumlah Nilai Unik Genre.

Tabel 2. Informasi Jumlah Data per Genre

Genres	Count
Documentary	124464
Drama	94951
Comedy	47595
Animation	29964
Horror	23781

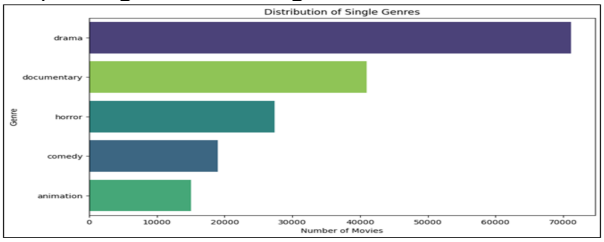
Pada dataset awal, atribut genre memiliki karakteristik multi-label, sehingga satu film dapat memiliki lebih dari satu genre. Tabel 2 menunjukkan distribusi kemunculan genre pada data multi-label sebelum dilakukan konversi menjadi label tunggal. Untuk mendukung proses klasifikasi supervised, genre multi-label kemudian dikonversi menjadi single-label dengan memilih satu genre berdasarkan rule-based scoring pada teks gabungan (text_combined) yang terdiri dari sinopsis dan keywords. Kemudian genre yang memiliki frekuensi kemunculan kata kunci terbanyak dalam gabungan teks sinopsis dan keywords dipilih sebagai label tunggal film tersebut. Hasil dari proses penentuan label tunggal ini menghasilkan distribusi data untuk masing-masing kelas genre seperti yang disajikan pada Tabel 2. Upaya konversi ini bertujuan memangkas kerumitan klasifikasi dan menyesuaikan dengan pendekatan klasifikasi supervised yang dipakai pada studi ini. Hasil informasi dataset secara rinci setelah melalui seluruh tahap preprocessing dapat dilihat pada Gambar 2.

```
<class 'pandas.core.frame.DataFrame'>
Index: 173524 entries, 1 to 633454
Data columns (total 8 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   title           173524 non-null  object
 1   overview        173524 non-null  object
 2   genres          173524 non-null  object
 3   keywords        173524 non-null  object
 4   overview_clean  173524 non-null  object
 5   keywords_clean  173524 non-null  object
 6   text_combined   173524 non-null  object
 7   genre_single    173524 non-null  object
dtypes: object(8)
memory usage: 11.9+ MB
```

Gambar 2. Info Dataset Setelah Preprocessing

Setelah seluruh tahapan preprocessing teks dan proses penentuan atau konversi label genre tunggal, serta seleksi data dilakukan, terlihat pada gambar 2 didapatkan sebuah dataset yang terdiri dari 173524 data film dengan 8 kolom utama. Hasil pemeriksaan menunjukkan bahwa semua kolom berisi nilai non-null, yang menandakan kelengkapan data telah melalui proses penyaringan secara menyeluruh.

Dataset akhir mencakup informasi judul film, sinopsis, genre asli, kata kunci, serta hasil preprocessing teks dalam bentuk kolom overview_clean, keywords_clean, dan text_combined. Kolom text_combined digunakan sebagai representasi utama teks yang menggabungkan sinopsis dan kata kunci film, sedangkan kolom genre_single berfungsi sebagai label kelas tunggal pada proses klasifikasi. Dengan struktur data yang telah bersih, konsisten, dan memiliki label tunggal yang jelas, dataset ini siap dimanfaatkan dalam tahapan ekstraksi fitur berbasis TF-IDF sekaligus perancangan model klasifikasi Naive Bayes. Berikut merupakan gambar 3 mengenai Bar Chart Distribusi Lima Genre.



Gambar 3. Bar Chart Distribusi Lima Genre

Terlihat pada gambar 3, hasil single label atau label tunggal pada genre yang tertinggi distribusinya yaitu genre drama. Setelah dilakukan konversi dari genre multi-label menjadi single-label, distribusi

label mengalami perubahan. Hal ini disebabkan oleh setiap film yang sebelumnya dapat memiliki lebih dari satu genre kemudian hanya diberikan satu label berdasarkan hasil keyword-based scoring. Pada distribusi label tunggal, genre Drama menjadi kategori terbanyak, diikuti Documentary, Horror, Comedy, dan Animation. Dengan demikian, jumlah 124.464 pada Tabel 2 merupakan jumlah kemunculan genre Documentary pada data sebelum dilakukan konversi dari genre multi-label menjadi single-label., sedangkan distribusi pada Gambar 3 merupakan distribusi label tunggal yang telah melewati tahapan konversi genre multi-label menjadi single-label, yang datanya digunakan dalam proses klasifikasi.

Setelah proses preprocessing dan penentuan label genre, dataset dibagi menjadi data training dan data testing dengan proporsi 80:20 menggunakan metode train_test_split. Pembagian dilakukan dengan random_state=42 dan stratifikasi berdasarkan label genre untuk mempertahankan proporsi kelas pada masing-masing subset data. Data testing yang digunakan dalam proses evaluasi berjumlah 34.705 data. Data testing tersebut tidak digunakan dalam proses penyeimbangan maupun pelatihan model, tetapi digunakan secara khusus untuk mengukur performa model setelah proses pelatihan.

Berdasarkan hasil penentuan label genre, distribusi jumlah data pada setiap kelas menunjukkan adanya ketidakseimbangan jumlah data antar genre. Sejumlah genre punya porsi data yang jauh lebih dominan dibanding kategori lain. Hal tersebut dapat berpotensi mempengaruhi performa model klasifikasi, dikarenakan model tampak berat sebelah ke arah kelas yang memiliki jumlah data lebih banyak.

Untuk mengatasi ketidakseimbangan jumlah data antar kelas, dilakukan proses penyeimbangan data pada data training menggunakan metode random undersampling. Proses dilakukan dengan menentukan jumlah data pada kelas yang memiliki jumlah sampel paling sedikit sebagai jumlah acuan, kemudian mengambil sampel secara acak dari setiap kelas hingga jumlah data pada seluruh kelas training menjadi sama. Pengambilan sampel dilakukan dengan random_state=42 untuk menjaga konsistensi hasil. Proses balancing hanya diterapkan pada data training, sedangkan data testing tetap dipertahankan tanpa proses balancing. Dengan demikian, data training yang telah diseimbangkan digunakan dalam proses pembentukan fitur dan pelatihan model, sedangkan data testing tetap digunakan sebagai data uji untuk mengevaluasi performa model.

Hasil dari proses penyeimbangan data menunjukkan bahwa distribusi jumlah data antar kelas genre pada data latih menjadi lebih merata. Dengan kondisi data yang lebih seimbang, sehingga proses pelatihan model diharapkan dapat membuahkan kinerja pengelompokan yang jauh lebih optimal dan adil terhadap seluruh kelas genre.

Setelah tahap preprocessing, dilakukan proses pembentukan fitur lewat penggunaan Term Frequency–Inverse Document Frequency (TF-IDF). Teknik tersebut membantu mengukur bobot urgensi tiap kata mengacu pada intensitas munculnya istilah tersebut dalam sebuah dokumen maupun lintas korpus data. Penerapan teknik TF-IDF memungkinkan kata-kata yang bersifat informatif mendapatkan skor yang lebih besar ketimbang istilah-istilah yang lazim, sehingga mampu meningkatkan kemampuan model dalam membedakan genre film berdasarkan karakteristik teks sinopsis.

2. Hasil Pelatihan dan Pengujian Model

Setelah seluruh tahapan preprocessing dan penentuan label tunggal dilakukan, diperoleh total 173.524 data film. Dataset dibagi menggunakan rasio 80:20, yang menghasilkan 138.819 data latih (training) dan 34.705 data uji (testing). Pada tahap pembentukan fitur, fitur teks diekstraksi menggunakan teknik TF-IDF dari data latih. Penting untuk ditekankan bahwa proses penyeimbangan data (data balancing) hanya diterapkan pada data latih guna mencegah bias model terhadap kelas mayoritas. Sebaliknya, 34.705 data uji (testing data) tetap dipertahankan sesuai distribusi aslinya dan terisolasi sepenuhnya dari proses pelatihan model.

Model Multinomial Naïve Bayes yang telah dilatih kemudian diuji menggunakan 34.705 data uji terisolasi tersebut. Adapun rincian jumlah sampel (support) data uji untuk setiap kategori genre meliputi: Drama (14.230 data), Documentary (8.192 data), Horror (5.477 data), Comedy (3.798 data), dan Animation (3.008 data) sebagaimana tertera pada Gambar 4. Evaluasi yang dilakukan secara

murni pada data uji terisolasi ini bertujuan untuk memastikan bahwa hasil performa klasifikasi yang diperoleh benar-benar mencerminkan kemampuan generalisasi model dalam memprediksi data baru.

3. Evaluasi Performa Model

Pada tabel 3 merupakan hasil evaluasi model yang didapatkan.

Tabel 3 Hasil Evaluasi Model

	precision	recall	f1-score	support
animation	0.75	0.85	0.80	3008
comedy	0.68	0.76	0.72	3798
documentary	0.84	0.88	0.86	8192
drama	0.90	0.79	0.84	14230
horror	0.80	0.87	0.84	5477
accuracy			0.83	34705
macro avg	0.80	0.83	0.81	34705
weighted avg	0.83	0.83	0.83	34705

Berdasarkan Tabel 3, model Naive Bayes dengan representasi fitur TF-IDF yang diterapkan pada data uji menghasilkan nilai accuracy sebesar 83%. Hasil tersebut menunjukkan bahwa model mampu mengklasifikasikan sebagian besar data uji dengan benar. Evaluasi menggunakan classification report menunjukkan bahwa performa model berbeda pada setiap genre film. Pada hasil f1-score, genre documentary mendapatkan hasil terbaik yaitu sebesar 0,86 yang didukung oleh nilai precision dan recall yang sama yaitu memiliki nilai tinggi, menandakan model mampu memprediksi genre ini secara tepat dan konsisten. Kemudian disusul oleh genre drama dan horor yang serta menampilkan kinerja yang cukup baik melalui masing-masing f1-score dengan 0,84 dimana nilai precision yang relatif tinggi membuat kesalahan prediksi relatif rendah, sementara nilai recall yang baik menampilkan kebolehan model dalam mengenali sebagian besar data dalam genre tersebut. Genre animation mempunyai perolehan F1-score di angka 0,80 dengan perbandingan yang selaras antara precision dan recall, sedangkan genre comedy memiliki nilai f1-score sebesar 0,72 yang menampilkan bahwa genre tersebut sulit dengan hasil terkecil dibandingkan genre lainnya, dapat terjadi karena kemiripan kata dengan genre lain.

Selain itu nilai support pada setiap genre menunjukkan jumlah data uji yang berbeda-beda yang dapat mempengaruhi kestabilan hasil evaluasi pada seluruh kelas klasifikasi. Secara umum, skor macro avg F1 sebesar 0,81 serta weighted avg F1 sebesar 0,83 menunjukkan keandalan model yang optimal dalam mengklasifikasikan beragam genre film secara konsisten dan seimbang. Berlandaskan capaian evaluasi model klasifikasi, algoritma Multinomial Naïve Bayes mengindikasikan efektivitas yang optimal dalam mengklasifikasikan genre film dengan tingkat akurasi mencapai 82,69%. Perolehan macro average dan weighted average f1-score yang terlihat tinggi sehingga model memperlihatkan kinerja yang relatif stabil dalam mengelompokkan berbagai genre film yang diuji.

Hasil analisis per genre menunjukkan bahwa genre documentary memperoleh performa terbaik dengan nilai f1-score tertinggi. Hal ini disebabkan oleh karakteristik teks sinopsis genre documentary yang cenderung memiliki kosakata yang lebih spesifik dan konsisten, sehingga memudahkan model dalam mengenali pola kata yang khas pada genre tersebut. Selain itu, struktur narasi pada genre documentary relatif berbeda dibandingkan genre lainnya.

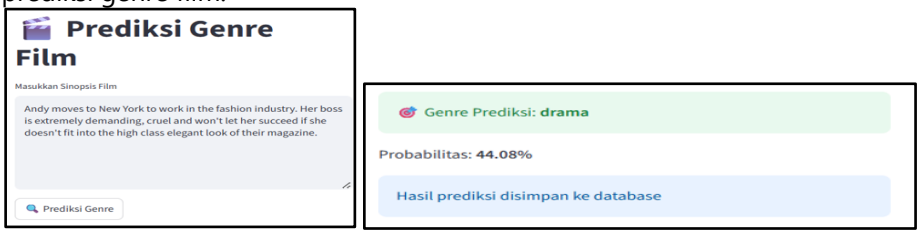
Genre drama dan horor juga menunjukkan performa klasifikasi yang baik. Pada genre drama, teks sinopsis umumnya mengandung kata-kata berkaitan konflik emosional dan hubungan antar karakter, sedangkan genre horror terdapat kosakata khas berkaitan dengan unsur ketegangan dan supranatural. Karakteristik tersebut membantu model membedakan kedua genre, dari genre lainnya. Sementara itu, genre comedy memiliki performa klasifikasi terendah dari genre lainnya. Rendahnya performa genre comedy dikarenakan kemiripan kosakata digunakan pada genre comedy dengan genre lain, seperti drama atau animation, sehingga menyebabkan model kesulitan membedakan batas antar genre. Selain itu, penggunaan pendekatan single-label pada data film pada dasarnya bersifat multi-genre juga berkontribusi terhadap terjadinya kesalahan klasifikasi.

Secara keseluruhan, hasil analisis menampilkan model Multinomial Naïve Bayes terbukti memadai melaksanakan proses klasifikasi genre film berbasis teks sinopsis. Namun demikian, keterbatasan

representasi genre tunggal dan kemiripan kosakata antar genre menjadi tantangan mempengaruhi performa model, sehingga berpotensi menjadi acuan penting dalam penyempurnaan studi lanjutan.

4. Implementasi Streamlit

Sebagai bentuk implementasi dari model klasifikasi genre film yang telah dibangun, penelitian ini dilengkapi dengan pengembangan aplikasi berbasis web menggunakan Streamlit. Aplikasi ini bertujuan membantu pengguna dalam melakukan klasifikasi genre film dengan interaktif berdasarkan teks sinopsis yang dimasukkan. Berikut merupakan gambar 4 yang merupakan tampilan halaman prediksi genre film.



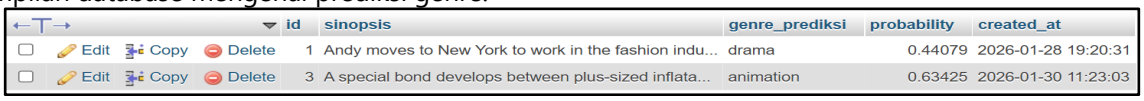
Gambar 5. Halaman Prediksi Genre Film

Aplikasi Streamlit dirancang menerima input berupa teks sinopsis film dari pengguna. Terlihat pada gambar 5 merupakan halaman prediksi genre, dimana teks sinopsis di input kemudian akan diproses menggunakan tahapan preprocessing yang sama dengan data penelitian, termasuk pembersihan teks dan transformasi fitur menggunakan metode TF-IDF. Selanjutnya, fitur teks digunakan sebagai masukan bagi model Multinomial Naïve Bayes yang telah dilatih dan menghasilkan prediksi genre film. Berikut gambar 6 menampilkan halaman riwayat prediksi genre film.



Gambar 6. Halaman Riwayat Prediksi Genre Film

Hasil prediksi genre ditampilkan secara langsung pada antarmuka aplikasi, sehingga pengguna mengetahui genre film yang diprediksi berdasarkan teks sinopsis. Terlihat pada gambar 5 ketika user melakukan input lalu submit untuk prediksi, hasil prediksi akan terlihat langsung. Gambar 6 merupakan halaman riwayat prediksi, merupakan hasil input data film yang telah diprediksi. Gambar 7 merupakan tampilan database mengenai prediksi genre.



Gambar 7. Database Prediksi Genre

Pada gambar 7 merupakan kumpulan data yang sudah diprediksi, data tersimpan di database. Lalu tampil secara interface melalui web streamlit ke user. Implementasi memperlihatkan hasil model klasifikasi yang dikembangkan tidak hanya dapat memberikan hasil evaluasi secara kuantitatif, namun dapat juga diterapkan pada bentuk aplikasi bersifat interaktif dan mudah digunakan. Streamlit dipilih karena memudahkan aplikasi berbasis data dan integrasi model machine learning.

PEMBAHASAN

Tahapan preprocessing mampu menghasilkan data teks lebih terstruktur sebelum digunakan dalam proses klasifikasi. Penggabungan antara sinopsis dan keywords memberikan representasi teks yang lebih lengkap karena kedua atribut tersebut mengandung informasi berkaitan dengan karakteristik film. Setelah melalui proses pembersihan teks dan pembentukan fitur, diperoleh dataset akhir sebanyak 173.524 data film dengan kondisi seluruh kolom tidak memiliki nilai non-null. Hal tersebut menunjukkan bahwa rangkaian seleksi beserta penataan data telah membuahkan dataset siap diimplementasikan dalam proses pengelompokan. Penggunaan TF-IDF memungkinkan teks hasil pemrosesan dikonversikan menjadi representasi angka berdasarkan tingkat kepentingan kata pada dokumen.

Konversi genre dari multi-label menjadi single-label membantu menyesuaikan karakteristik data dengan algoritma Multinomial Naïve Bayes. Namun, proses tersebut menyebabkan informasi mengenai genre lain sebenarnya dimiliki oleh suatu film tidak digunakan sebagai label akhir. Kondisi ini dinilai sebagai salah satu variabel berdampak terhadap hasil klasifikasi karena satu judul film dapat mengkombinasikan lebih dari satu tipologi genre. Dengan hanya menggunakan satu genre utama, model diarahkan untuk mempelajari batas kelas yang lebih sederhana, tetapi pada beberapa film terdapat kemungkinan karakteristik teks yang juga merepresentasikan genre lain. Oleh karena itu, pendekatan single-label menjadi salah satu keterbatasan dalam merepresentasikan karakteristik film secara keseluruhan.

Berdasarkan uji coba, algoritma Multinomial Naïve Bayes terbukti mampu memberikan akurasi sebesar 82,69%, disertai nilai macro average F1-score sebesar 0,81 dan weighted average F1-score sebesar 0,83. Luaran tersebut mengindikasikan model memiliki kemampuan efektif dalam mengenali pola teks berkaitan dengan genre film. Performa tersebut juga menunjukkan bahwa kombinasi representasi teks menggunakan teknik TF-IDF berbantukan model Multinomial Naïve Bayes dapat digunakan untuk melakukan klasifikasi genre berdasarkan informasi yang terdapat pada sinopsis dan keywords. Meskipun demikian, performa model tidak sepenuhnya sama pada setiap genre karena karakteristik kosakata dan pola teks yang digunakan pada masing-masing genre berbeda.

Perbedaan performa antar genre terlihat dari nilai F1-score. Genre documentary memperoleh performa terbaik dengan nilai 0,86. Kondisi tersebut dapat dikaitkan dengan karakteristik teks sinopsis documentary cenderung menggunakan kosakata lebih spesifik dan konsisten sehingga pola kata yang digunakan lebih mudah dikenali oleh model. Genre drama dan horor pun mencatatkan kapabilitas yang relatif optimal dengan raihan F1-score masing-masing di angka 0,84. Hal tersebut berkaitan dengan adanya karakteristik kosakata yang cukup khas, seperti kata-kata menggambarkan konflik emosional dan hubungan antar karakter pada drama serta unsur ketegangan dan supranatural pada horror. Sebaliknya, genre comedy memperoleh F1-score paling rendah, yaitu 0,72. Rendahnya performa tersebut menunjukkan bahwa model lebih sulit membedakan comedy dari genre lain. Kemiripan kosakata antara comedy dengan genre seperti drama atau animation dapat menyebabkan fitur teks yang digunakan model tidak selalu memberikan pembeda yang kuat antar kelas.

Selain kemiripan karakteristik teks, penggunaan label tunggal pada data bersifat multi-genre berkontribusi terhadap kesalahan klasifikasi. Film yang memiliki beberapa genre pada data asli direpresentasikan hanya dengan satu genre utama, sehingga kata-kata dalam sinopsis yang sebenarnya berkaitan dengan genre lain tetap terdapat dalam teks, tetapi tidak direpresentasikan sebagai label. Kondisi tersebut dapat membuat model menemukan pola kata yang beririsan antar kelas dan meningkatkan kemungkinan terjadinya kesalahan prediksi. Dengan demikian, meskipun model memperlihatkan efektivitas yang relatif solid pada performa global, kemiripan kosakata antar genre dan penyederhanaan label menjadi satu genre merupakan faktor yang perlu diperhatikan dalam interpretasi hasil klasifikasi.

Untuk memperkuat kontribusi serta menunjukkan keunggulan pendekatan yang dikembangkan, dilakukan perbandingan performa secara kuantitatif dengan beberapa penelitian terdahulu yang berfokus pada klasifikasi genre film berbasis teks sinopsis[7][8][9]. Pengujian algoritma pembelajaran mesin konvensional pada dataset berskala besar (117 ribu data) menunjukkan penurunan akurasi, di mana Naïve Bayes hanya memperoleh 59,41% dan SVM sebesar 65,26%[9]. Sementara itu, penerapan arsitektur deep learning Long Short-Term Memory (LSTM) pada klasifikasi multi-label menghasilkan F1-score tertinggi sebesar 60,30%[7]. Pengujian yang dilakukan pada skala data terbatas yaitu 1.005 film mencapai akurasi sebesar 98,00% menggunakan LSTM[8]. Penggunaan kombinasi fitur sinopsis dan keywords yang ditransformasikan melalui TF-IDF serta dilatih menggunakan Multinomial Naïve Bayes berhasil memperoleh akurasi sebesar 82,69% pada dataset masif sebanyak 173.524 film[1]. Peningkatan kinerja yang signifikan terlihat lebih unggul jika dibandingkan dengan studi klasifikasi single-label berbasis sinopsis TMDB yang memperoleh akurasi 59,00% pada 1.000 data[14]. Hasil membuktikan integrasi fitur keywords bersama sinopsis terbukti efektif memperkaya representasi konteks teks, serta menjadikan model yang dikembangkan tidak hanya lebih unggul dan stabil secara kuantitatif

dibandingkan metode konvensional[9][14] maupun deep learning[7] pada dataset berskala masif, tetapi juga jauh lebih efisien dalam hal waktu dan daya komputasi pelatihan[1][5][23].

Implementasi model aplikasi Streamlit menunjukkan bahwa hasil klasifikasi bukan hanya dapat diaplikasikan pengujian performa secara kuantitatif, melainkan diterapkan dalam bentuk aplikasi interaktif. Pengguna dapat memasukkan sinopsis film, kemudian teks tersebut diproses dengan tahapan yang sama dan digunakan sebagai masukan bagi model Multinomial Naïve Bayes untuk menghasilkan prediksi genre. Adanya halaman riwayat prediksi dan penyimpanan hasil prediksi ke dalam database juga menunjukkan bahwa model dapat digunakan dalam alur penggunaan yang lebih praktis. Dengan demikian, implementasi Streamlit menjadi bentuk penerapan hasil penelitian yang memungkinkan proses klasifikasi genre film dilakukan secara lebih mudah dan interaktif.

KESIMPULAN

Penelitian mengklasifikasikan genre film berdasarkan narasi sinopsis menggunakan Natural Language Processing dan algoritma Multinomial Naïve Bayes. Tahap pengolahan teks meliputi pembersihan data, normalisasi huruf, dan penghapusan karakter non-alfabet. Sinopsis dan keywords digabungkan, lalu ditransformasi menggunakan metode TF-IDF. Proses menghasilkan representasi teks yang tertata dan meningkatkan kualitas data pembelajaran model. Multinomial Naïve Bayes diterapkan pada fitur TF-IDF setelah data splitting dan penyeimbangan distribusi data. Model mampu mengategorikan genre film secara otomatis dengan akurasi 83%. Genre documentary memiliki kinerja klasifikasi terbaik. Genre comedy paling sulit diklasifikasikan karena kosakatanya mirip dengan genre lain. Pendekatan single-label pada dataset multi-genre membuat setiap film hanya diwakili satu genre. Akibatnya, karakteristik genre lain pada film tidak sepenuhnya terwakili. Kemiripan kosakata antar genre menyulitkan model, terutama pada genre dengan F1-score rendah. Dataset belum merepresentasikan domain di luar film, sehingga penerapan pada domain lain perlu diuji lebih lanjut. Model diimplementasikan dalam aplikasi Streamlit untuk prediksi genre secara interaktif berdasarkan teks sinopsis. Model berpotensi dipakai pada sistem rekomendasi atau analisis konten film berbasis teks. Penelitian selanjutnya disarankan memakai multi-label classification, dataset yang lebih beragam, serta algoritma yang lebih memahami makna dan hubungan antar kata.

REFERENCES

[1] S. Gupta, U. Adhikari, S. Varshney, and T. Choudhury, "Computational Analysis of Genre Effects on Movie Ratings Using MLP Algorithms," *Journal of Computer Science*, vol. 21, no. 4, pp. 904–917, 2025, doi: 10.3844/jcssp.2025.905.917.

[2] F. Shaukat, N. Ejaz, R. Kamal, T. Alkhalifah, S. Aslam, and M. Mu, "Multi-Label Movie Genre Classification with Attention Mechanism on Movie Plots," *Computers, Materials and Continua*, vol. 83, no. 3, pp. 5595–5622, 2025, doi: 10.32604/cmc.2025.061702.

[3] M. Fairuzabadi and T. A. Munandar, "Enhancing Film Genre Classification Using FastText, BiGRU, and Attention Mechanisms," *International Journal of Information Technology and Computer Science Applications*, vol. 02, no. 03, pp. 202–216, 2024, [Online]. Available: https://creativecommons.org/licenses/by/4.0/

[4] F. Shaukat, N. Ejaz, Z. Ashraf, M. M. Alnfiai, N. N. Alotaibi, and S. M. M. Alnefaie, "An interpretable multi-transformer ensemble for text-based movie genre classification," *PeerJ Comput. Sci.*, vol. 11, 2025, doi: 10.7717/peerj-cs.2945.

[5] S. C. Wardoyo and Umniy Salamah, "Hyperparameter Optimization Analysis of MultinomialNB and Logistic Regression in Multi-Feature Text-Based Film Genre Classification," *Indonesian Journal on Computing (Indo-JC)*, vol. 10, no. 1, Aug. 2025, doi: 10.21108/indojc.v10i1.9100.

[6] D. Velamentosa, E. Zuliarso, and J. Raya Tri Lomba Juang, "Sistem Rekomendasi Film Menggunakan Metode Content-Based Filtering", *Jurnal Mahasiswa Teknik Informatika (JATI)*, vol. 9, no. 2, 2025, doi: https://doi.org/10.36040/jati.v9i2.13251.

[7] J. Akbar, H. Fahmi, and W. Murniati, "Multi Label Klasifikasi Genre Film Berdasarkan Sinopsis Menggunakan Metode Long Short-Term Memory (LSTM)," *Jurnal Manajemen Informatika & Sistem Informasi (MISI)*, vol. 8, no. 1, 2025, doi: https://doi.org/10.36595/misi.v8i1.1436.

[8] R. Amelia and D. B. Santoso, "Prediksi Genre Film Dengan Klasifikasi Multi Kelas Sinopsis Menggunakan Jaringan Lstm Film Genre Prediction With Multi-Class Classification Synopsis Using Lstm Network," *Journal of Information Technology and Computer Science (INTECOMS)*, vol. 6, no. 2, 2023.

[9] N. Buslim, L. K. Oh, M. H. Athallah Hardy, and Y. Wijaya, "Comparative Analysis of KNN, Naïve Bayes and SVM Algorithms for Movie Genres Classification Based on Synopsis.," *JURNAL TEKNIK INFORMATIKA*, vol. 15, no. 2, pp. 169–177, Dec. 2022, doi: 10.15408/jti.v15i2.29302.

[10] M. K. Najib, A. Irawan, F. N. Salsabilla, and S. Nurdiati, "Performance Comparison of Gradient-based Optimizer for Classification of Movie Genres," *Indonesian Journal of Mathematics and Applications*, vol. 3, no. 1, pp. 1–18, Apr. 2025, doi: 10.21776/ub.ijma.2025.003.01.1.

[11] M. Alfian Ilyassa *et al.*, "Implementasi Klasifikasi Genre Film Pada Platform Imdb Berdasarkan Deskripsi Sinopsis Film Menggunakan Algoritma Naïve Bayes, SVM, Dan Logistic Regression," *Indonesian Journal of Multidisciplinary on Social and Technology Homepage*, vol. 4, no. 3, pp. 1966–1976, 1966, doi: 10.69693/ijmst.v4i2.12210.

[12] A. Nurian, M. S. Ma'arif, I. N. Amalia, and C. Rozikin, "Analisis Sentimen Pengguna Aplikasi Shopee Pada Situs Google Play Menggunakan Naive Bayes Classifier," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 1, Jan. 2024, doi: 10.23960/jitet.v12i1.3631.

[13] A. Zhafira, N. Afifah, S. Anditya Putri, V. Marhalatun, and D. I. Surya Saputra, "Analisis Sentimen Ulasan Film pada Dataset IMDB menggunakan Algoritma Naive Bayes," *Jurnal Manajemen Informatika, Sistem Informasi dan Teknologi Komputer (JUMISTIK)*, vol. 4, no. 1, pp. 373–383, Jun. 2025, doi: 10.70247/jumistik.v4i1.139.

[14] A. Gimnastiar and M. Ikhsan, "Perbandingan K-Nearest Neighbor dan Naive Bayes dalam Klasifikasi Genre Film Berdasarkan Sinopsis," *(sudo) Jurnal Teknik Informatika*, vol. 5, no. 3, pp. 254–264, Sep. 2026, doi: 10.56211/sudo.v5i3.1974.

[15] M. F. Naufal and S. Ferdiana Kusuma, "Natural Language Processing Untuk Otomatisasi Pengenalan Pronomina Dalam Kalimat Bahasa Indonesia," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 9, pp. 1011–1017, Oct. 2022, doi: 10.25126/jtiik.202296394.

[16] V. R. Prasetyo, N. Benarkah, and V. J. Chrisintha, "Implementasi Natural Language Processing Dalam Pembuatan Chatbot Pada Program Information Technology Universitas Surabaya," *Teknika*, vol. 10, no. 2, pp. 114–121, Jul. 2021, doi: 10.34148/teknika.v10i2.370.

[17] M. H. Humaidi, Sutrisno, and P. W. Laksono, "Implementation of Machine Learning for Text Classification Using the Naive Bayes Algorithm in Academic Information Systems at Sebelas Maret University Indonesia," *E3S Web of Conferences*, vol. 465, Dec. 2023, doi: 10.1051/e3sconf/202346502048.

[18] A. Y. Muaad *et al.*, "Arabic Document Classification: Performance Investigation of Preprocessing and Representation Techniques," *Math. Probl. Eng.*, vol. 2022, 2022, doi: 10.1155/2022/3720358.

[19] F. D. Dermawan and Supriyono, "Implementasi Artificial Intelligence dalam Pembuatan Website Klasifikasi Genre Buku," *bit-Tech*, vol. 7, no. 2, pp. 618–627, Dec. 2024, doi: 10.32877/bt.v7i2.1977.

[20] A. N. Laili, P. P. Adikara, and S. Adinugroho, "Rekomendasi Film Berdasarkan Sinopsis Menggunakan Metode Word2Vec," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 3, no. 6, pp. 6035–6043, Jun. 2019, [Online]. Available: http://j-ptiik.ub.ac.id

[21] T. Dwi, R. Jati, S. Baco, and N. P. Husain, "Penerapan Algoritma Naïve Bayes Pada Analisis Sentimen Pengguna Game Mobile Legends Di Google Play Store," 2025. [Online]. Available: https://jiktif.ft-uim.ac.id

[22] M. Veziroğlu, E. Veziroğlu, and İ. Ö. Bucak, "Performance Comparison between Naive Bayes and Machine Learning Algorithms for News Classification," in *Bayesian Inference-Recent Trends*, IntechOpen, 2024. doi: 10.5772/intechopen.1002778.

[23] D. Subedi, Nabin Lamichhane, and N. Subedi, "Sentiment Analysis of IMDb Movie Reviews Using SVM and Naive Bayes Classifier," *Journal of Engineering and Sciences*, vol. 4, no. 1, pp. 56–68, May 2025, doi: 10.3126/jes2.v4i1.70138.