KLASIFIKASI *GENRE* BERDASARKAN DESKRIPSI FILM MENGGUNAKAN TF-IDF DAN SELEKSI FITUR DENGAN *XGBOOST*

Kiven Keita Yosua Damanik, Muhammad Ezar Al Rivan
Fakultas Ilmu Komputer dan Rekayasa, Prodi Informatika, Universitas Multi Data Palembang
Jl Rajawali No.14, 9 Ilir, Kec. Ilir Tim. II, Kota Palembang, Sumatera Selatan
kipenkei@gmail.com

ABSTRAK

Perkembangan industri perfilman yang semakin pesat menyebabkan meningkatnya jumlah film yang diproduksi setiap tahunnya sehingga diperlukan sistem pengelolaan informasi yang efektif, salah satunya melalui pengelompokan film berdasarkan *genre*. Penentuan *genre* film yang masih banyak dilakukan secara manual cenderung bersifat subjektif, memerlukan waktu yang lama, serta kurang efektif dalam menangani film dengan karakteristik *multi-genre*. Penelitian ini bertujuan untuk mengembangkan sistem klasifikasi *genre* film secara otomatis berdasarkan deskripsi film berupa sinopsis. Sinopsis film memuat ringkasan alur cerita dan tema yang dapat merepresentasikan karakteristik suatu *genre*, namun masih berbentuk teks bebas sehingga memerlukan tahapan pengolahan teks agar dapat dianalisis secara komputasional. Metode yang digunakan dalam penelitian ini meliputi *preprocessing* teks, ekstraksi fitur menggunakan *Term Frequency–Inverse Document Frequency* (TF-IDF), seleksi fitur untuk mengurangi dimensi data, serta algoritma *Extreme Gradient Boosting* (*XGBoost*) dengan pendekatan *One-Vs-Rest* untuk melakukan klasifikasi *multi-genre*. Evaluasi model dilakukan menggunakan metrik *accuracy, precision, recall,* dan *f1-score*. Hasil pengujian menggunakan dataset *IMDb Movie Genres* mampu menghasilkan akurasi sebesar 87%, serta menunjukkan kinerja yang baik dalam mengklasifikasikan *genre* film berdasarkan sinopsis.

Kata kunci : *Deskripsi Film; TF-IDF; XGBoost*

1. PENDAHULUAN

Perkembangan teknologi informasi saat ini menyebabkan peningkatan jumlah data digital dalam berbagai bentuk, termasuk data teks. Data teks banyak ditemukan pada berbagai bidang seperti pendidikan, bisnis, media sosial, dan industri hiburan digital. Data teks merupakan salah satu jenis data yang paling banyak dihasilkan dan membutuhkan teknik khusus untuk dapat dianalisis secara otomatis. Industri perfilman merupakan salah satu domain yang menghasilkan data teks dalam jumlah besar, terutama melalui metadata film seperti judul, *genre*, dan sinopsis. Informasi tekstual tersebut merepresentasikan alur cerita serta karakteristik film dan menjadi sumber data utama dalam proses analisis dan klasifikasi *genre* film berbasis teks [1].

Pemanfaatan teks deskriptif film terbukti efektif dalam membantu sistem memahami konten film dan meningkatkan kinerja klasifikasi maupun rekomendasi *genre* secara otomatis. *Genre* film menjadi informasi yang sangat penting karena berfungsi sebagai kategori utama yang menggambarkan tema dan karakteristik film. Menurut [2] [3] *genre* film berperan besar dalam sistem pencarian dan rekomendasi film pada platform digital. Namun, proses pencarian dan pengelompokan *genre* film yang masih dilakukan secara manual sering kali bersifat subjektif, memerlukan waktu yang lama, dan berpotensi menimbulkan ketidakkonsistenan. Oleh karena itu, diperlukan pendekatan otomatis berbasis *machine learning* untuk mengatasi keterbatasan tersebut [4] Selain itu, sebuah film dapat memiliki lebih dari satu *genre*, sehingga proses pengelompokan menjadi semakin kompleks. Oleh karena itu, diperlukan suatu pendekatan otomatis yang mampu mengklasifikasikan *genre* film secara efisien.

Informasi tekstual yang digunakan dalam penelitian ini adalah deskripsi film yang direpresentasikan dalam bentuk sinopsis film. Sinopsis film merupakan ringkasan cerita yang menggambarkan alur cerita, konflik, tokoh, serta gambaran umum isi film. Informasi tersebut mengandung kata kunci dan pola semantik yang merepresentasikan tema dan karakteristik film. Menurut [5] menyatakan bahwa sinopsis film mampu merepresentasikan isi dan konteks film secara ringkas sehingga dapat dimanfaatkan secara efektif dalam proses klasifikasi *genre* film berbasis teks. Selain itu, penelitian lain menunjukkan bahwa informasi tekstual seperti sinopsis dan plot film memiliki hubungan yang kuat dengan kategori *genre*, sehingga banyak digunakan sebagai sumber data utama dalam tugas klasifikasi *genre* film secara multi-*genre*.

Namun, karena deskripsi film berupa sinopsis yang disajikan dalam bentuk teks bebas (*unstructured text*) maka diperlukan tahapan pengolahan teks agar dapat dianalisis secara komputasional. Proses *preprocessing* dilakukan untuk mengurangi *noise*, menyeragamkan bentuk teks, serta mengubah data menjadi representasi numerik yang dapat dipahami oleh model[6] [7].

Salah satu metode yang paling umum digunakan untuk klasifikasi teks adalah *Term Frequency–Inverse Document Frequency* (TF-IDF). Menurut [8], TF-IDF memberikan bobot pada setiap kata berdasarkan frekuensi kemunculannya dalam dokumen dan tingkat kepentingannya dalam keseluruhan kumpulan

dokumen. Penelitian oleh [9], menunjukkan bahwa penerapan TF-IDF pada teks sinopsis film efektif dalam mengekstraksi fitur penting yang merepresentasikan isi film dan mendukung proses klasifikasi *genre* film secara komputasional.

Namun demikian, penggunaan TF-IDF dalam representasi teks sering kali menghasilkan jumlah fitur yang sangat besar, sehingga dapat meningkatkan kompleksitas pada komputasi dan berpotensi menurunkan kinerja model klasifikasi [8]. Oleh karena itu, diperlukan tahap seleksi fitur untuk mengurangi dimensi data dan mempertahankan fitur-fitur yang relevan. Salah satu metode seleksi fitur yang banyak digunakan dalam klasifikasi teks adalah Chi-Square, yang mengukur tingkat ketergantungan antara suatu fitur dan kelas tertentu [10]. Metode ini memungkinkan pemilihan kata-kata yang memiliki kontribusi signifikan terhadap proses klasifikasi dengan cara mengeliminasi fitur yang kurang informatif.

Setelah melakukan seleksi fitur, pemilihan algoritma klasifikasi juga berpengaruh terhadap kualitas hasil klasifikasi yang dihasilkan. Salah satu algoritma yang digunakan dalam penelitian klasifikasi adalah *Extreme Gradient Boosting* (XGBoost). XGBoost merupakan algoritma berbasis ensemble yang mampu menggabungkan sejumlah weak learnes untuk menghasilkan model yang lebih kuat dan stabil. Algoritma ini dikenal efektif dalam menangani data berdimensi tinggi serta mampu memodelkan hubungan non-linear dengan baik [11]. Selain itu, XGBoost menerapkan pendekatan boosting yang dilakukan secara bertahap, di mana setiap model dibangun untuk memperbaiki kesalahan dari model sebelumnya. XGBoost juga dilengkapi dengan teknik regulatization yang berfungsi mengendalikan kompleksitas model sehingga dapat mengurangi risiko overfitting. Dengan dukungan pemrosesan pararel, XGBoost mampu meningkatkan efisiensi komputasi tanpa mengorbankan akurasi, sehingga sesuai digunakan pada klasifikasi teks dengan jumlah fitur yang besar. Sebagai perbandingan, algoritma decision tree merupakan metode klasifikasi yang membangun model prediksi dalam bentuk struktur pohon keputusan tunggal. Decision tree bekerja dengan membagi data berdasarkan nilai fitur tertentu hingga menghasilkan keputusan akhir pada setiap daun pohon [12] Metode ini memiliki keunggulan dalam hal kemudahan interpretasi dan proses pelatihan yang relatif cepat. Namun, decision tree memiliki keterbatasan karena cenderung sering terjadi overfitting terutama pada data berdimensi tinggi dan kompleks seperti data teks hasil ekstraksi TF-IDF. Selain itu, penggunaan satu pohon keputusan saja membuat kemampuan generalisasi model menjadi lebih rendah dibandingkan metode ensemble.

Berbagai penelitian terdahulu klasifikasi *genre* film berbasis teks telah banyak dilakukan dengan memanfaatkan sinopsis atau deskripsi film sebagai sumber data utama. Namun, sebagian besar penelitian tersebut masih menerapkan pendekatan single-label sehingga setiap film hanya dipetakan ke dalam satu *genre* dominan. Pendekatan ini belum sepenuhnya mampu menggambarkan karakteristik film modern yang umumnya memiliki lebih dari satu *genre* secara bersamaan. Sejumlah penelitian terdahulu juga lebih berfokus pada perbandingan performa beberapa algoritma untuk menentukan model terbaik, tanpa menitikberatkan pada pengembangan pendekatan yang sederhana dan mudah diterapkan pada skenario penggunaan nyata. Di sisi lain, penelitian yang menerapkan pendekatan multi-*genre* film berbasis sinopsis masih relatif terbatas. Pendekatan yang digunakan masih memanfaatkan dataset hasil scraping mandiri dari Internet Movie Databases (IMDb) website yang berpotensi memiliki ketidakseimbangan distribusi *genre*. Selain itu, belum ada yang mengeksplorasikan dataset IMDb Movie *Genre*s yang tersedia pada platform HuggingFace yang memiliki struktur data lebih konsisten. Selain itu, metode Chi-Square sebagai teknik seleksi atau ekstraksi fitur teks belum dimanfaatkan dalam penelitian klasifikasi *genre* film dan penggunaan algoritma XGBoost sebagai metode klasifikasi juga belum diterapkan dalam konteks klasifikasi *genre* film berbasis teks. Oleh karena itu, penelitian ini diangkat dengan judul "Klasifikasi *Genre* Berdasarkan Deskripsi Film Menggunakan TF-IDF, dan seleksi fitur dengan XGBoost".

2. TINJAUAN PUSTAKA

Landasan teori yang akan dikemukan pada bab ini berisikan tentang berbagai teori-teori yang menjadi acuan dalam penelitian ini. Teori tersebut berisi tentang berbagai hal yang terkait dengan topik penelitian kali ini. Bagian ini menjelaskan teori-teori tersebut agar mudah dipahami oleh pembaca.

2.1. Penelitian Terdahulu

Penelitian terdahulu, Sebagian besar penelitian terkait klasifikasi genre film berbasis teks masih menunjukkan variasi pada dataset, jenis teks, serta pendekatan pemodelan. Sejumlah studi menggunakan sinopsis/plot dari Kaggle/IMDb dengan ukuran dataset yang relatif terbatas atau berasal dari metadata, sementara penelitian lain berfokus pada pendekatan yang lebih kompleks seperti deep learning (misalnya LSTM dengan word embedding) maupun multimodal (poster, trailer, subtitle) yang membutuhkan sumber daya komputasi dan proses ekstraksi fitur yang lebih rumit. Selain itu, beberapa penelitian lebih menekankan perbandingan banyak algoritma untuk mencari model terbaik, bukan pada rancangan sistem yang sesuai dengan skenario penggunaan nyata. Di sisi lain, karakteristik film yang dapat memiliki lebih dari satu genre menuntut pendekatan multi-label, namun sebagian besar penelitian masih menerapkan skema single-label/multi-class atau belum menekankan keluaran multi-genre. Kondisi ini menunjukkan masih adanya gap pada pendekatan yang ringan, mudah

diimplementasikan, dan mudah direproduksi, tetapi tetap mampu menghasilkan prediksi multi-genre dari sinopsis film. Karena itu, diperlukan penelitian yang terarah pada skenario penggunaan spesifik, yaitu prediksi genre film dari sinopsis yang diketik langsung oleh pengguna dengan memanfaatkan dataset yang lebih terstandarisasi dan mudah diakses seperti jquigl/imdb-genres (Hugging Face).

2.2. *Genre*

Genre berasal dari bahasa Prancis yang secara harfiah berarti "bentuk" atau "jenis". Dalam konteks perfilman, *genre* digunakan sebagai cara untuk mengelompokkan film berdasarkan tipe tertentu. Setiap *genre* memiliki karakteristik yang khas seperti pola alur cerita, latar, karakter sampai tema yang berulang dan mudah dikenali oleh penonton [13]

Secara umum, *genre* dapat dipahami sebagai istilah yang digunakan untuk mengelompokkan teks media ke dalam kategori-kategori tertentu berdasarkan kesamaan karakteristik dan konvensi yang dimilikinya. Dalam kajian perfilman, para ahli telah mengembangkan beragam cara untuk mengklasifikasikan film berdasarkan kesamaan ciri yang dimilikinya. Film-film tertentu menunjukkan pola yang berulang pada aspek naratif maupun sinematik, seperti struktur cerita, gaya visual, serta pendekatan pengisahan. Kesamaan unsur-unsur tersebut kemudian dijadikan dasar dalam membentuk sistem pengelompokan film yang dikenal luas sebagai *genre* dan masih digunakan hingga saat ini.

2.3. Sinopsis Film (Deskripsi Teks)

Sinopsis pada dasarnya merupakan bentuk ringkasan dari sebuah karya, ide, atau gagasan yang disajikan dalam narasi singkat. Melalui sinopsis, penonton dapat memperoleh gambaran umum mengenai jalannya cerita sebuah film, sehingga membantu mereka memahami alur serta karakter yang ditampilkan sebelum menyaksikan karya tersebut secara utuh [14]. Secara lebih spesifik, sinopsis dapat dipahami sebagai uraian singkat atau gambaran garis besar naskah yang menjelaskan isi suatu film atau pertunjukan. Penyajiannya dapat dilakukan secara konkret maupun abstrak, tergantung pada tujuan dan sudut pandang penulis dalam merangkum karya tersebut [15]. Dalam konteks perfilman, sinopsis memiliki peran penting karena memuat informasi utama mengenai alur cerita dan tema. Informasi tersebut dapat dimanfaatkan untuk memperkirakan jenis atau *genre* film yang disajikan, sehingga memberikan gambaran awal kepada penonton mengenai karakteristik film yang akan ditonton [16].

2.4. TF-IDF

Term Frequency–Inverse Document Frequency (TF-IDF) merupakan salah satu metode yang banyak digunakan dalam bidang pengolahan teks dan pemodelan bahasa alami. Metode ini bertujuan untuk menentukan tingkat kepentingan suatu kata dalam sebuah dokumen dengan mempertimbangkan kemunculan kata tersebut di dalam kumpulan dokumen yang lebih luas. TF-IDF mampu menunjukkan kata mana yang paling merepresentasikan isi dokumen [17].

Dalam penerapannya, nilai TF dan IDF dikalikan untuk menghasilkan bobot kata (*term weight*). Bobot ini mencerminkan seberapa penting sebuah kata dalam suatu dokumen jika dibandingkan dengan seluruh dokumen lain dalam koleksi [17]

Pendekatan umum dalam menghitung nilai *Term Frequency* dilakukan dengan membagi jumlah kemunculan suatu kata dengan total kata yang terdapat dalam dokumen. Semakin sering kata tersebut muncul, maka semakin besar pula nilai TF yang dihasilkan sebagaimana ditunjukkan pada persamaan (1).

$$tf_{(t,d)} \frac{n_{(t)}}{n} \tag{1}$$

Keterangan:
$tf(t,d)$ = frekuensi kemunculan *term t* pada dokumen *d*
$n(t)$ = jumlah kemunculan *term* pada dokumen
n = jumlah total kata dalam dokumen

Sebaliknya, kata yang jarang muncul pada keseluruhan kumpulan dokumen cenderung memiliki nilai IDF yang lebih besar. Nilai IDF diperoleh dengan membagi jumlah total dokumen dalam korpus dengan jumlah dokumen yang mengandung term tersebut, kemudian hasilnya dihitung menggunakan fungsi logaritma untuk menstabilkan skala nilai, seperti ditunjukkan pada persamaan (2).

$$idf_{(t)} \log(\frac{D}{df_{(t)}}) \tag{2}$$

Keterangan:
$idf(t)$ = *inverse document frequency*
D = jumlah keseluruhan dokumen
$df(t)$ = jumlah dokumen yang memuat *term t*

Berdasarkan persamaan (1) dan (2), nilai TF-IDF dihitung dengan mengalikan nilai TF dan IDF. Penerapan metode TF-IDF banyak dimanfaatkan dalam proses ekstraksi kata kunci, klasifikasi dokumen, pengelompokan teks berdasarkan topik, serta sistem temu kembali informasi. Perhitungan bobot TF-IDF dirumuskan dalam persamaan (3) sebagai berikut:

$$W_{(t,d)} = tf_{(t,d)} * idf_{(t)} \tag{3}$$

Keterangan:
$tf(t,d)$ = frekuensi *term* dalam dokumen d
$idf(t)$ = nilai *inverse document frequency term*

2.5. *Chi-Square*

Chi-Square merupakan salah satu metode uji statistik untuk data diskrit yang digunakan dalam

pengujian hipotesis, khususnya untuk menilai ada atau tidaknya hubungan antara dua variabel. Metode ini berfungsi untuk mengevaluasi tingkat keterkaitan antar variabel serta menentukan apakah hubungan tersebut bersifat signifikan atau justru tidak ada sama sekali. Dalam konteks pengujian keterkaitan pada suatu populasi, uji *Chi-Square* digunakan untuk memastikan apakah suatu objek atau subjek saling berhubungan atau tidak

Menurut Alam [18], dalam penerapannya pada seleksi fitur perhitungan nilai *Chi-Square* untuk setiap kata t pada setiap kelas c dapat dilakukan dengan bantuan tabel kontingensi seperti yang ditunjukkan pada Tabel 1.

Tabel 1 Tabel Kontigensi Chi-Square

Kata	**Kelas**	
	C	**C***
t	A	B
t*	C	D

Nilai-nilai pada tabel kontingensi tersebut merupakan frekuensi observasi dengan keterangannya sebagai berikut:

A = jumlah dokumen dalam kelas c yang mengandung kata t.
B = jumlah dokumen di luar kelas c yang mengandung kata t.
C = jumlah dokumen dalam kelas c yang tidak mengandung kata t.
D = jumlah dokumen di luar kelas c yang tidak mengandung kata t.

Berdasarkan tabel kontingensi tersebut, nilai *Chi-Square* dapat dihitung menggunakan rumus yang ditunjukkan pada Persamaan 4 berikut :

$$X^2(t,c) = \sum \frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (4)$$

Keterangan:

t : Kata
c : Kelas/Kategori
N : Jumlah dokumen latih
A : Jumlah dokumen pada kategori c yang memuat kata t
B : Jumlah dokumen bukan kategori c yang memuat kata t
C : Jumlah dokumen pada kategori c yang tidak memuat kata t
D : Jumlah dokumen bukan kategori c yang tidak memuat kata t.

Untuk keperluan seleksi fitur, diperlukan satu nilai *Chi-Square* tunggal untuk setiap kata. Nilai ini diperoleh dengan menjumlahkan nilai *Chi-Square* kata tersebut pada seluruh kategori yang ada. Perhitungan nilai *Chi-Square* tunggal dapat dilihat pada Persamaan 5 berikut :

$$X^2(t) = \sum_{c=1}^{k} X^2(t,c) \quad (5)$$

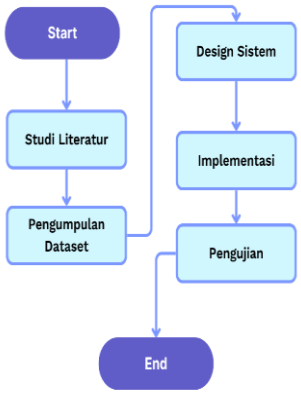
Keterangan :

t : Kata
c : Kelas/Kategori

2.6. *Xtreme Gradient Boosting* (*XGBoost*)

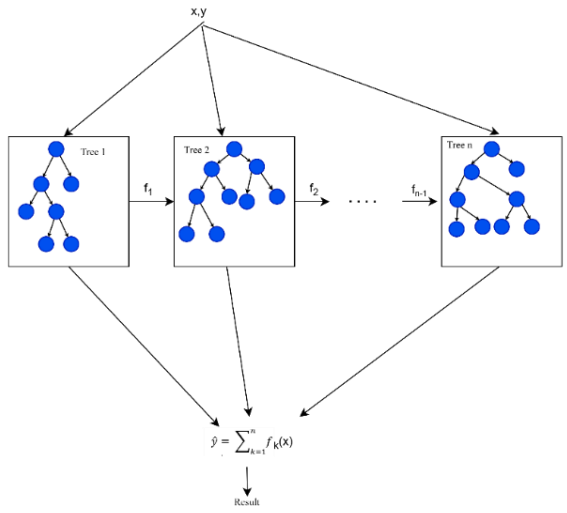
XGBoost adalah salah satu algoritma pembelajaran mesin yang efektif dan sering digunakan dalam berbagai aplikasi seperti klasifikasi, regresi, dan ranking [19]. Algoritma ini menggunakan *ensemble learning*, di mana beberapa model digabungkan untuk menghasilkan prediksi yang lebih akurat. Algoritma ini juga menggunakan metode *gradient boosting*, yaitu memperbaiki kesalahan prediksi dengan menurunkan nilai fungsi kesalahan (*loss function*) secara bertahap menggunakan arah turunan (*gradient*). Keunggulan utama XGBoost adalah kemampuannya untuk menangani data besar dan kompleks dengan waktu yang relatif cepat. Selain itu, algoritma ini mampu mengatasi masalah *overfitting* melalui proses *tuning* parameter, seperti *learning rate*, *maximum depth*, *minimum child weight*, dan *subsample*.

Algortimanya bekerja dengan menggabungkan banyak pohon Keputusan (*decision tree*) kecil secara bertahap. Pada tahap awal, *XGBoost* membuat satu pohon sederhana untuk melakukan prediksi awal terhadap data. Setelah itu, sistem akan menghitung kesalahan dari hasil prediksi tersebut, lalu membuat pohon baru untuk memperbaiki kesalahan tersebut, lalu membuat pohon baru untuk memperbaiki kesalahan tersebut. Proses ini dilakukan secara berulang hingga semua kesalahan semakin kecil dan model menjadi lebih akurat. Secara umum, proses algoritma *XGBoost* ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Metodologi

Proses pembelajaran pada algoritma *XGBoost* didasarkan pada konsep fungsi objektif (*objective function*) yang menggabungkan dua komponen utama, yaitu fungsi *loss* dan fungsi regularisasi.



Gambar 2. Visualisasi XGBoost

3. METODE PENELITIAN

Dalam pelaksanaan penelitian ini, diterapkan perancangan metode terlebih dahulu untuk meminimalisir kesalahan dalam proses penelitian. Tahapan metodologi penelitian dapat dilihat pada Gambar 2.

3.1. Studi Literatur

Tahap pertama yang dilakukan adalah studi literatur yang bertujuan memperoleh pemahaman yang mendalam mengenai teori, konsep, serta hasil penelitian terdahulu yang berkaitan dengan topik klasifikasi *genre* berdasarkan deskripsi teks. Pencarian referensi dilakukan melalui berbagai sumber ilmiah seperti jurnal, prosiding konferensi, buku daring (*online books*), dan artikel penelitian yang relevan.

3.2. Pengumpulan Dataset

Pada tahap ini dilakukan pengumpulan data yang akan digunakan dalam penelitian. Dataset yang digunakan berupa data film yang terdiri dari deskripsi film (sinopsis) dan label *genre* film. Dataset diperoleh dari sumber dataset publik yaitu IMDB *Genre*s Dataset yang tersedia pada platform HuggingFace. Data ini digunakan sebagai bahan utama dalam proses pelatihan dan pengujian model klasifikasi.

3.3. Desain Sistem

`Tahap desain sistem bertujuan untuk menggambarkan alur kerja sistem klasifikasi *genre* film yang dikembangkan dalam penelitian ini menggunakan pendekatan text mining dan *machine learning*. Proses dimulai dengan tahap preprocessing teks yang bertujuan untuk membersihkan serta menormalisasi data teks sebelum diproses lebih lanjut. Tahapan preprocessing meliputi lowercasing, tokenization, stopword removal, serta stemming atau lemmatization untuk mengubah kata menjadi bentuk dasar. Setelah preprocessing dilakukan, dataset dibagi menjadi data latih dan data uji dengan perbandingan 80% dan 20%. Data latih kemudian diubah menjadi representasi numerik menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF). Selanjutnya dilakukan proses seleksi fitur menggunakan metode Chi-Square untuk memilih fitur kata yang paling relevan terhadap kelas *genre*, kemudian fitur yang telah dipilih digunakan dalam proses pelatihan model klasifikasi menggunakan algoritma *Extreme Gradient Boosting* (XGBoost) untuk menghasilkan model yang dapat memprediksi *genre* film berdasarkan deskripsi film.

3.4. Implementasi

Tahap implementasi sistem merupakan tahap realisasi dari rancangan sistem yang telah disusun sebelumnya ke dalam bentuk program yang dapat dijalankan. Pada penelitian ini, bahasa pemrograman Python digunakan sebagai alat utama dalam pengembangan sistem karena memiliki berbagai library yang mendukung pengolahan data teks dan *machine learning* seperti *pandas*, *NumPy*, *scikit-learn*, serta *NLTK* untuk proses *text preprocessing*.

Implementasi sistem dilakukan sesuai dengan alur yang telah dirancang pada flowchart penelitian. Seluruh tahapan implementasi disusun dalam satu alur (*pipeline*) agar proses pelatihan dan pengujian model dapat dilakukan secara terstruktur dan konsisten.

3.5. Pengujian

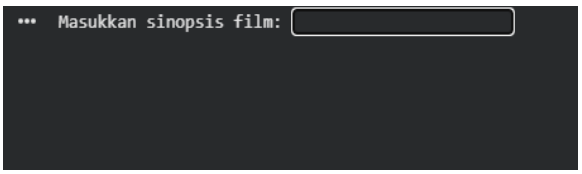
Tahap pengujian dilakukan untuk mengevaluasi performa sistem klasifikasi *genre* film yang telah dibangun. Pengujian dilakukan menggunakan data uji yang sebelumnya telah dipisahkan dari data latih pada tahap pembagian dataset. Data uji terlebih dahulu diproses menggunakan model TF-IDF yang telah dibangun dari data latih untuk menghasilkan representasi fitur numerik yang konsisten dengan fitur yang digunakan pada proses pelatihan model. Selanjutnya data uji dimasukkan ke dalam model XGBoost yang telah dilatih untuk menghasilkan prediksi *genre* film yang beruba *multi-genre*. Di mana satu film dapat memiliki lebih dari satu *genre* secara bersamaan. Oleh karena itu, hasil proses penentuan *label* tersebut dilakukan dengan menggunakan nilai probabilitas dari model yang dibandingkan menggunakan nilai *threshold* untuk menentukan nilai prediksi *genre* termasuk ke dalam prediksi atau tidak. Evaluasi performa model dilakukan menggunakan confusion matrix yang terdiri dari *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN), kemudian dihitung menggunakan beberapa metrik evaluasi yaitu *Accuracy*, *Precision*, *Recall*, dan *F1-Score* untuk mengetahui tingkat akurasi dan efektivitas model dalam mengklasifikasikan *genre* film berdasarkan deskripsi film.

4. HASIL DAN PEMBAHASAN

Tabel 1. Proses Preprocessing

No	Text awal	Teks Hasil Preprocessing
1	A new generation of kids, called Indigo, who have more mental and spiritual abilities than others. A company of such kids gather together in hidden places.	such kids gather together in hidden places. new generation kids called indigo mental spiritual abilities others company kids gather together hidden places
2	A Manny Velazquez underground slasher tribute about a number of serial killers on the loose in Chicago and begins a murder spree.	manny velazquez underground slasher tribute number serial killers loose chicago begins murder spree
3	A look at the scandalous love triangle between Victorian art critic John Ruskin (Greg Wise), his teenage bride Euphemia "Effie" Gray (Dakota Fanning), and Pre-Raphaelite painter John Everett Millais (Tom Sturridge)	look scandalous love triangle victorian art critic john ruskin greg wise teenage bride euphemia effie gray dakota fanning pre raphaelite painter john everett millais tom sturridge
4	Flaming Ears is a pop sci-fi lesbian fantasy feature set in the year 2700 in the fictive burned-out city of Asche. It follows the tangled lives of three women - Volley, Nun and Spy.	flaming ears pop sci lesbian fantasy feature set year fictive burned city asche follows tangled lives three women volley nun spy

Tabel 2 menunjukkan hasil yang telah melalui tahap *preprocessing*, yang mengalami perubahan berupa pengubahan huruf menjadi kecil, menghapus tanda baca, menghapus kata yang tidak penting. Tahapan ini bertujuan untuk menghasilkan data yang konsisten sebelum dilakukan proses TF-IDF dan XGBOOST.



Gambar 3. Tampilan input sistem

Masukkan sinopsis film: Flaming Ears is a pop sci-fi lesbian fantasy feature set in the year 2700 in the fictive burned-out city of Asche. It follows the tangled lives of three women - Volley, Nun and Spy.

=== HASIL PREPROCESSING ===
flaming ears pop sci lesbian fantasy feature set year fictive burned city asche follows tangled lives three women volley nun spy

=== GENRE HASIL PREDIKSI ===
['Fantasy', 'Sci-Fi']

Gambar 3. Menunjukkan tampilan input sinopsis film yang akan digunakan pengguna sebagai awal dalam prediksi film

.

Hasil akhir prediksi ditampilkan pada textbox, yang telah disusun berdasarkan nilai kemiripan yang paling tinggi dari yang diperoleh.

Tabel 2. Hasil Confusion Matrik Per *Genre*

No.	*Genre*	TP	TN	FP	FN
1.	Drama	2381	503	3065	51
2.	Horror	711	4304	635	350
3.	Crime	832	3891	877	400
4.	Action	1487	1855	2447	211
5.	Sci-Fi	220	5383	184	213
6.	Adventure	608	4178	792	422
7.	Romance	576	4267	749	408
8.	War	88	5701	121	90
9.	Animation	131	5541	136	192
10.	Thriller	801	3165	1574	460
11.	Comedy	675	3532	1278	515
12.	Sport	37	5861	52	50
13.	Biography	85	5652	110	153
14.	Mystery	222	4921	420	437
15.	History	70	5656	112	162
16	Fantasy	155	5209	263	373

Tabel 3. Hasil Evaluasi

No.	*Genre*	Prec	Rec	F1	Acc
1.	Drama	43.72	97.90	60.45	48.07
2.	Horror	52.82	67.01	59.08	83.58
3.	Crime	46.68	67.53	56.58	78.72
4.	Action	37.80	87.57	52.81	55.70
5.	Sci-Fi	54.46	50.81	52.57	93.38
6.	Adventure	43.43	59.03	50.04	79.77
7.	Romance	43.47	58.54	49.89	80.72
8.	War	42.11	49.44	45.48	96.49
9.	Animation	49.06	40.56	44.41	94.53
10.	Thriller	33.73	63.52	44.06	66.10
11.	Comedy	34.56	56.72	42.95	70.12
12.	Sport	41.57	42.53	42.05	98.30
13.	Biography	43.59	35.71	39.26	95.62
14.	Mystery	34.58	33.69	34.13	85.72
15.	History	38.46	30.17	33.82	95.43
16	Fantasy	37.08	29.36	32.77	89.40
Rata - Rata		39.82	54.43	46.27	81.98

Berdasarkan dari tabel 3, model menunjukkan hasil yang cukup baik, terutama *genre* drama dan *action*, yang memiliki nilai *True Positive* (TP) tinggi. Selain itu, nilai *True Negative* (TN) beberapa *genre*, seperti *Sci-Fi*, *War*, dan *Sport*, menunjukkan bahwa model mampu dengan baik membedakan data yang

tidak termasuk ke dalam *genre* tertentu. Namun, ada beberapa *genre* seperti drama, *action*, dan *thriller* memiliki nilai *False Positive* (FP) yang cukup besar. Di sisi lain, nilai *False Negative* (FN) pada beberapa *genre* seperti *comedy*, *adventure*, menunjukkan bahwa masih ada data yang tidak dikenali oleh model.

Berdasarkan dari Tabel 4, model menunjukkan performa yang cukup baik dari nilai *accuracy* yang tinggi dari rata-rata *genre* 81.89%. Dan untuk *accuracy* tertinggi pada *genre sport* 98.30%, untuk yang terendah pada *genre* drama 48.07%. Hal ini menunjukkan model mampu mengklasifikasikan data secara keseluruhan dengan cukup akurat dan stabil. Selain itu, terdapat metrik lain yaitu *precision* untuk menunjukkan ketepatan antar prediksi, *recall* menujukkan kemampuan model dalam menemukan data yang sesuai, dan *F1-Score* menunjukkan kemampuan keseimbangan antara keduanya.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa penggunaan metode TF-IDF yang dikombinasikan dengan seleksi fitur *Chi-Square* serta algoritma *Extreme Gradient Boosting* (XGBOOST) mampu membangun model klasifikasi *genre* film berbasis deskripsi sinopsis yang cukup baik dalam mengidentifikasi keterkaitan antara kata kunci dan kategori *genre*. Tahapan *preprocessing* terbukti membantu mengurangi *noise* dan menghasilkan representasi fitur yang lebih relevan, sementara seleksi fitur *Chi-Square* mampu menurunkan dimensi data dengan mempertahankan fitur yang memiliki hubungan kuat terhadap kelas. Hasil pengujian menunjukkan bahwa model memiliki performa yang baik dari sisi akurasi, namun masih terdapat kesalahan klasifikasi pada beberapa data yang memiliki karakteristik *genre* yang mirip atau *multi-genre*. Oleh karena itu, penelitian selanjutnya disarankan untuk melakukan optimasi parameter model, menambahkan teknik penanganan ketidakseimbangan data, serta membandingkan dengan metode representasi teks atau algoritma klasifikasi lain untuk memperoleh performa yang lebih optimal dan hasil klasifikasi yang lebih stabil.

DAFTAR PUSTAKA

[1] 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.

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

[3] A. Zakharia *et al.*, "Sistem Rekomendasi Film Indonesia Menggunakan Metode Content-Based Filtering," *Jurnal Ilmu Komputer dan Pendidikan*, vol. 2, no. 4, pp. 671–678, 2024, [Online]. Available: https://journal.mediapublikasi.id/index.php/logic

[4] A. V Siddharth, P. Rakshitha, S. Mazher, V. S. Reddy, and M. G. Uma Maheshwari, "Movie *Genre* Classification Using *Machine learning*," 2025.

[5] J. Wehrmann, M. A. Lopes, and R. C. Barros, "Self-Attention for Synopsis-Based Multi-Label Movie *Genre* Classification," in *Artificial Intelligence Research Society Conference*, 2023. Accessed: Jan. 11, 2026. [Online]. Available: https://aaai.org/papers/236-flairs-2018-17686/

[6] Y. HaCohen-Kerner, D. Miller, and Y. Yigal, "The influence of preprocessing on text classification using a bag-of-words representation," *PLoS One*, vol. 15, no. 5, May 2020, doi: 10.1371/journal.pone.0232525.

[7] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning Based Text Classification: A Comprehensive Review," in *CEUR Workshop Proceedings*, CEUR-WS, 2020, pp. 1–9. doi: 10.1145/nnnnnnn.nnnnnnn.

[8] M. Das, S. Kamalanathan, and P. Alphonse, "A Comparative Study on TF-IDF feature Weighting Method and its Analysis using Unstructured Dataset," 2020.

[9] A. Anugrah Ma, C. Hamud, J. Zulkaidah, R. Leo Saputra, and M. Faiz Alimuddin, "Penerapan Text Mining menggunakan Metode TF-IDF untuk Menentukan *Genre* Film," *ILKOMNIKA: Journal of Computer Science and Applied Informatics E*, vol. 7, no. 3, pp. 415–422, 2025, doi: 10.28926/ilkomnika.v7i3.734.

[10] P. Rama Bena Putra and R. Setya Perdana, "Klasifikasi Judul Berita Online menggunakan Metode Support Vector Machine (SVM) dengan Seleksi Fitur Chi-square," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 7, no. 5, pp. 2132–2141, 2023, [Online]. Available: http://j-ptiik.ub.ac.id

[11] O. M. Ahmed, S. R. M Zeebaree, and S. Askar, "Comparative Analysis of XGBoost Performance for Text Classification with CPU Parallel and Non-Parallel Processing," *Indonesian Journal of Computer Science Attribution*, vol. 13, no. 2, pp. 2024–1781, 2024.

[12] Habib Hakim Sinaga, "Perbandingan Metode Decision Tree dan Xgboost untuk Klasifikasi Sentimen Vaksin COVID-19 di Twitter," Universitas Islam Negeri Sultan Syarif Kasim Riau Pekanbaru, 2021.

[13] W. Fauzi, "UNIKOM_Wildan Fauzi_12. BAB II," Universitas Komputer Indonesia (UNIKOM), 2020.

[14] A. Chilliya Basuki, M. T. Furqon, and P. P. Adikara, "Temu Kembali Informasi terhadap Sinopsis Film menggunakan Metode BM25F," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 6, no. 3, pp. 1239–1246, 2022, [Online]. Available: http://j-ptiik.ub.ac.id

[15] R. D. Siahaan, "Implementasi Algoritma Zhu Takaoka Pada Aplikasi Sinopsis Film Bioskop Berbasis Mobile," *Terapan Informatika Nusantara*, vol. 1, no. 12, pp. 587–590, 2021.

[16] R. Amelia and D. B. Santoso, "Prediksi *Genre* Film dengan Klasifikasi Multi Kelas Sinopsis Menggunakan Jaringan LSTM," *Journal of Information Technology and Computer Science (INTECOMS)*, vol. 6, no. 2, 2023.

[17] D. Septiani and I. Isabela, "Analisis Term Frequency Inverse Document Frequency (TF-IDF) dalam Temu kembali Informasi Pada Dokumen Teks," *Jurnal Sistem dan Teknologi Informasi Indonesia*, vol. 1, no. 2, 2022.

[18] P. Alam Jusia, R. Pahlevi, and D. Sintong Pardamean Simanjuntak, "Peningkatan Performa Naive Bayes dengan Fitur Chi-Square pada Analisis Sentimen Komentar Pengguna Aplikasi Netflix," *BULLETIN OF COMPUTER SCIENCE RESEARCH*, vol. 5, no. 4, pp. 614–621, 2025, doi: 10.47065/bulletincsr.v5i4.532.

[19] E. H. Yulianti, O. Soesanto, and Y. Sukmawaty, "Penerapan Metode *Extreme Gradient Boosting* (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit," *JOMTA Journal of Mathematics: Theory and Applications*, vol. 4, no. 1, 2022