e-ISSN: 2460-870X | p-ISSN: 2085-1367

Klasifikasi Genre Film Berdasarkan Sinopsis Menggunakan Metode *Support Vector Machine* (SVM) dengan TF-IDF dan N-Gram

Movie Genre Classification Based on Synopsis Using the Support Vector Machine (SVM) Method with TF-IDF and N-Gram

Rita Novita Sari[1*], Dea Alya[2], Dodyk Fahlome[3]
[1,2,3]Universitas Islam Negeri Sumatera Utara, Medan, Medan, 20353, Indonesia
ritanovitasari@uinsu.ac.id*, alya0701232102@uinsu.ac.id[2], dodyk0701232093@uinsu.ac.id[3]

ABSTRAK

Banyaknya variase genre film yang terseda menyebabkan proses identifikasi genre berdasarkan sinopsis menjadi semakin penting untuk membantu pengguna dalam memilih film yang sesuai dengan preferensinya. Penelitian ini bertujuan membangun model klasifikasi genre film menggunakan algoritma Support Vector Machine (SVM) ysng dipadukan dengan ekstraksi fitur TF-IDFdan N-Gram. Data penelitian diperoleh dari Kaggle dan teridiri atas 10.000 data film dengan karakterisitik multi-label yang kemudian di transformasikan menjadi single label. Penelitian difokuskan pada empat genre utama, yaiut action, comedy, drama dan horror sehingga diperoleh 6.000 data yang digunakan dalam proses pemodelan. Tahapan penelitian meliputi persiapan data, prapemrosesan teks, ekstrkasi fitur menggunakan TF-IDF berbasis unigram dan bigram, pembagian data serta evaluasi model. Seluruh proses dilakukan menggunakan bahasa pemrograman Python pada platform Google Colab. Hasil pengujian menunjukan bahwa model SVM dengan linear mampu mencapai akurasi sebesar 83%. Berdasarkan hasil evaluasi, genre Horror memperoleh performa terbaik dengan nilai precision sebesar 91%, recall sebesar 92% dan F1-Score sebesar 92%. Hasil penelitian menunjukan bahwa kombinasi TF-IDF, N-Grm dan SVM efekti digunakan untuk melakukan klasifikasi genre film berdasarkan informasi yang terkantung dalam sinopsis.

Kata Kunci : Support Vector Machine, TF-IDF, N-Gram, Klasifikasi Genre Film, Sinopsis

ABSTRACT

The wide variety of film genres available makes genre identification based on synopses increasingly important in helping users select films that match their preferences. This study aims to develop a film genre classification model using the Support Vector Machine (SVM) algorithm combined with TF-IDF and N-Gram feature extraction. The research data was obtained from Kaggle and consists of 10,000 movie records with multi label characteristics, which were subsequently transformed into single-label data. The study focused on four main genres Action, Comedy, Drama, and Horror resulting in 6,000 data points used in the modeling process. The research stages included data preparation, text preprocessing, feature extraction using unigram- and bigram-based TF-IDF, data splitting, and model evaluation. The entire process was carried out using the Python programming language on the Google Colab platform. The test results showed that the SVM model with a linear kernel achieved an accuracy of 83%. Based on the evaluation results, the Horror genre achieved the best performance, with a precision of 91%, a recall of 92%, and an F1-score of 92%. The results of the study indicate that the combination of TF-IDF, N-Gram, and SVM is effective for classifying movie genres based on the information contained in the synopsis.

Keywords : Support Vector Machine, TF-IDF, N-Gram, Movie Genre Classification, Synopsis.

Info Artikel :
Disubmit: *Direview:* *Diterima :*

Copyright © 2026 – CSRID Journal. All rights reserved.

DOI: https://www.doi.org/10.22303/csrid.18.2.2026.381-394

1. PENDAHULUAN

Film telah menjadi salah satu bentuk hiburan yang banyak diminatin masyarakat, dimana film dapat menyampaikan pesan, nilai, dan emosi kepada penonton. Seiring dengan pertumbuhan industri film dan peningkatan jumlah film yang diproduksi setiap tahunnya, kebutuhan akan sistem klasifikasi genre yang dapat diandalkan semakin meningkat. Film dapat berasal dari berbagai genre, seperti aksi, horor, komedi, dan drama [1].

Keberagaman genre film yang tersedia, banyak penikmat film yang mengalami kesulitan pada pengguna dalam mengidentifikasi jenis film yang sesuai dengan minat mereka hanya berdasarkan sinopsis yang tersedia [2]. Dari konflik yang di ambil, masyarakat memerlukan suatu sistem yang dapat membantu dalam mengklasifikasikan genre film.

Dalam bidang *text mining*, klasifikasi teks digunakan untuk mengidentifikasi kategori suatu dokumen secara otomatis berdasarkan pola informasi yang terkandung di dalamnya [3]. Sinopsis film dapat digunakan sebagai sumber informasi untuk menentukan genre film. Namun, variasi panjang teks, penggunaan bahasa yang beragam, dan adanya kata-kata yang tidak secara langsung menunjukkan genre tertentu adalah semua faktor yang memengaruhi proses klasifikasi sinopsis [4]. Dengan demikian, klasifikasi manual menjadi kurang di minati karena penggunaan yang dilakukan dengan waktu yang cukup lama.

Long Short-Term Memory (LSTM) yang telah di terapkan pada penelitian terdahulu untuk klasifikasi genre film multilabel dan menghasilkan nilai F1-Score sebesar 0,603 [5]. Pada penelitian terdahulu mengenai ekstraksi fitur seperti *Term Frequency Inverse DocumenFrecuency*(TF-IDF) dengan pendekatan *rule based* memperoleh akurasi sebesar 78,5% dalam menentukan genre film berdasarkan sinopsis [6]. Selain itu, penerapan algoritma Support Vector Machine (SVM) yang dipadukan dengan TF-IDF pada klasifikasi berita hoax di media sosial X berhasil memperoleh tingkat akurasi sebesar 82,29% [7].

Meskipun metode TF-IDF dan SVM telah berhasil diterapkan pada berbagai tugas klasifikasi teks, termasuk klasifikasi genre film dan deteksi berita hoax, penerapan kombinasi TF-IDF berbasis unigram dan bigram dengan algoritma SVM Linear pada klasifikasi genre film berbasis sinopsis masih relatif terbatas. Di sisi lain, penelitian berbasis TF-IDF sebelumnya masih banyak menggunakan pendekatan *rule-based* sehingga kemampuan model dalam melakukan generalisasi terhadap variasi teks sinopsis film masih terbatas.

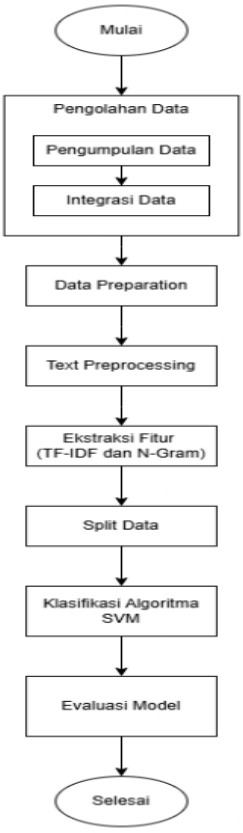
Beberapa penelitian terdahulu telah menunjukkan bahwa metode TF-IDF, N-Gram, dan *Support Vector Machine* (SVM) mampu memberikan hasil yang baik pada berbagai tugas klasifikasi teks. Namun, sebagian besar penelitian tersebut diterapkan pada domain selain klasifikasi genre film atau belum secara khusus mengevaluasi kombinasi fitur TF-IDF berbasis unigram dan bigram dengan algoritma SVM Linear pada dataset sinopsis film yang telah melalui proses penyeimbangan kelas (*class balancing*). Kondisi tersebut menunjukkan masih adanya celah penelitian terkait penerapan kombinasi TF-IDF berbasis unigram dan bigram dengan algoritma SVM Linear untuk klasifikasi genre film berdasarkan sinopsis.

Berdasarkan research gap tersebut, penelitian ini mengusulkan penerapan kombinasi TF-IDF, N-Gram, dan SVM Linear untuk mengklasifikasikan genre film berdasarkan sinopsis. Kontribusi penelitian ini terletak pada penggunaan fitur unigram dan bigram secara bersamaan, penerapan Random *Oversampling* untuk mengatasi ketidakseimbangan kelas, serta evaluasi model menggunakan metrik *accuracy*, *precision*, *recall*, F1-*score*, dan *confusion matrix*. Dengan pendekatan tersebut, diharapkan model mampu menghasilkan performa klasifikasi yang baik serta memberikan alternatif metode yang lebih sederhana dan efisien dibandingkan pendekatan *deep learning*.

e-ISSN: 2460-870X | p-ISSN: 2085-1367

2. METODE

Penelitin ini menggunakan teknik *machine learning* dan *text mining* dengan bantuan tools Goggle Colab berbasis Python, yang mencakup pengolahan data, preprocessing teks, ekstraksi fitur TF-IDF dan N-Gram, dan proses klasifikasi menggunakan algoritma *Support Vector Machine* (SVM). Tahapan penelitian yang dilakukan dilihat pada Gambar 1.



Gambar 1. Tahapan alur penelitian

A. Pengumpulan Data

Pengumpulan data pada penelitian ini diambil dari platform Kaggle [8]. *Dataset* terdiri dari dua *dataset*, *dataset* genre film dan *dataset* film. Data genre film berisi informasi identitas genre berupa id dan nama genre, sedangkan data film berisi *atribut title, overview, dan genre_ids* yang bersifat *multi-label Classification*. Jumlah data yang didapat 10.000 data film dalam proses klasifikasi genre film berdasarkan sinopsis.

B. Integrasi Data

Proses menyatukan data dari berbagai sumber ke dalam database baru dikenal sebagai integrasi data. Ini dapat terjadi pada atribut yang sama atau dengan penambahan atribut baru [9]. Integrasi data dilakukan dengan penggabungan dua *dataset yang terpisah*, yaitu *dataset* data film dan *dataset* genre film. Proses penggabungan dilakukan dengan mencocokkan atribut genre_ids pada data film dengan atribut id pada tabel genre sehingga setiap data film dapat memperoleh informasi nama genre yang sesuai.

DOI: https://www.doi.org/10.22303/csrid.18.2.2026.381-394

C. Data Preparation

Data preparation dilakukan untuk mempersiapkan data karena bertujuan meningkatkan kualitas data sebelum proses ekstraksi fitur dan klasifikasi algoritma SVM [9]. Pada tahap ini dilakukan beberapa proses, yaitu pemilihan atribut, pembersihan data, transformasi label.

1. Pemilihan Atribut

   Pada tahap ini, atribut yang relevan dipilih, yang berdampak pada proses klasifikasi data [10]. Pada penelitian ini, atribut *overview* sebagai teks sinopsis dan genre sebagai label klasifikasi.

2. Pembersihan Data

   Dataa yang tidak valid atau memiliki missing value (nilai kosong) harus di hilangkan pada proses pembersihan data.. Data yang tidak lengkap dapat mempengaruhi kualitas analisis dan mengurangi kemampuan model dalam proses klasifikasi algoritma SVM [11].

3. *Multi-label* menjadi *Single-label*

   Dalam penelitian ini, dataset film memiliki karakteristik klasifikasi multi-label, dimana pada salah satu kolom genre memiliki lebih dari satu genre. Namun, yang dibutuhkan dari algoritma Support Vector Machine (SVM) hanya satu kelas target untuk setiap data yang digunakan [12]. Pada tahap ini dilakukan proses transformasi data *multi-label* menjadi *Single-label* dengan memilih satu genre utama pada setiap data film, karena pada *dataset* IMDb, urutan genre umumnya merepresentasikan genre dominan film. Selanjutnya, penelitian ini hanya menggunakan empat genre dengan frekuensi kemunculan tertinggi, yaitu Action, Drama, Comedy, dan Horror.

4. Penyeimbangan Kelas (*Class balancing*)

   Penyeimbangan kelas atau *Class balancing* merupakan teknik yang digunakan untuk mengatasi ketidakseimbangan distribusi jumlah data pada setiap kelas. Ketidakseimbangan data dapat memperburuk model untu mengenali kelas tertentu [13]. Pada penelitian ini, proses *Class balancing* dilakukan menggunakan teknik *Random Oversampling* pada empat genre dominan, yaitu Action, Comedy, Drama, dan Horror digunakan teknik *Random Oversampling* yang dapat melakukan proses dengan menambahkan sampel data secara acak pada kelas dengan jumlah data yang lebih sedikit, sehingga data pada tiap kelas menjadi seimbang [14].

D. Text preprocessing

Pada tahap ini dilakukan proses pembersihan dan normalisasi data teks. *Text preprocessing* bertujuan untuk mengurangi noise pada *dataset* yang dapat direpresentasikan dengan lebih baik dan mudah diolah [15]. Beberapa tahapan processing terdiri dari:

1. *Cleaning Text* (Pembersihan Teks)

   Proses pembersihan teks, teks dari karakter atau elemen yang tidak perlu akan di bersihkan [16]. Tujuan dari langkah ini adalah untuk mengurangi noise pada *dataset* dengan menghapus angka, simbol, tanda baca, karakter khusus, URL, dan spasi yang berlebihan dalam data sinopsis film.

2. *Case Folding*

   Setelah pembersihan teks selesai, langkah selanjutnya adalah melakukan perubahan pada setiap huruf menjadi huruf kecilkarakter dari teks diumengubah seluruh huruf dalam teks menjadi huruf kecil, sehingga format penulisan kata sebanding dengan data sinopsis film [17].

e-ISSN: 2460-870X | p-ISSN: 2085-1367

3. *Tokenizing*

   Proses ini bertujuan untuk memisahkan kalimat menjadi bagian kata [18]. Pada tahap ini, proses *Tokenizing* dilakukan dengan memisahkan teks sinopsis film berdasarkan spasi dan tanda pemisah tertentu.

4. *Stopword Removal*

   Stopword Removal dilakukan dengan menghapus kata umum yang tidak memiliki makna, seperti "a", "the", "her", dan "in" agar model dapat fokus pada kata yang relavan untuk proses klasifikasi teks [19].

5. *Lemmatization*

   Tahap ini bertujuan untuk mengubah kata ke bentuk dasar atau kata asli berdasarkan struktur morfologi dan konteks kata dalam kalimat [20]. Dalam penelitian ini, proses lematisasi diterapkan pada data sinopsis film berbahasa Inggris. Misalnya, perubahan kata "watching" menjadi "watch" dan "movies" menjadi "movie". Tujuan dari proses ini adalah untuk meningkatkan representasi fitur TF-IDF dan membantu algoritma SVM.

E. Ekstraksi Fitur

Setelah melalui tahap data preparation, dilakukan proses ekstraksi fitur dengan mentransformasi data yang awalmya berbentuk teks menjadi numerik sehingga algoritma SVM dapat memproses dengan baik [21]. Salah satu metode yang dipakai adalah TF-IDF berbasis N-GRam

1. N-Gram

   Metode N-Gram adalah model bahasa yang didasarkan pada hubungan kolinear antar karakter atau kata dalam teks, yaitu mengambil potongan n karakter atau kata dari string secara berurutan [22].
   Rumus N-Gram

$$Ntr - gram = (w1, w2, \ldots, wn) \quad (1)$$

   Dimana:
   $w1, w2, \ldots, wn$ = urutan kata
   nnn = jumlah kata

   Rumus Jumlah N-Gram

$$Jumlah\ N - gram = T - N + 1 \quad (2)$$

   Dimana:
   T = jumlah token
   N = ukuran gram
   Fitur teks yang terbentuk dari hasil penerapan metode N-Gram adalah unigram dan bigram yang akan digunakan pada proses TF-IDF.

2. TF-IDF
   Metode ekstraksi fitur TF-IDF memberikan bobot pada kata berdasarkan tingkat kemuculan kata dalam seluruh dataset [21]. Metode ini juga digunakan untuk memberikan bobot pada fitur teks hasil *preprocessing* dan untuk mengekstraksi N-Gram dari sinopsis film.
   Rumus *Term Frequency* (TF)

$$TF(t, d) = \sum_k^f \frac{t,d}{fk,d} \quad (3)$$

   Dimana:
   Ft, d = jumlah kemunculan term t pada dokumen

DOI: https://www.doi.org/10.22303/csrid.18.2.2026.381-394

$\sum_k fk,d$ = jumlah seluruh kata pada dokumen

Rumus *Inverse Document Frequency* (IDF)

$$IDF(t) = log\left(\frac{N}{df(t)}\right) \quad (4)$$

Dimana:
N = jumlah seluruh dokumen
$df(t)$ = jumlah dokumen yang mengandung term t

Rumus TF-IDF

$$TF - IDF(t,d) = TF(t,d)\backslash timesIDF(t) \quad (5)$$

Metode TF-IDF digunakan untuk meningkatkan fitur teks yang dihasilkan dari *preprocessing* serta ekstraksi N-Gram dari sinopsis film. Library TfidfVectorizer digunakan untuk melakukan proses ekstraksi fitur dalam penelitian ini. Parameter ngram_range=(1,2) memungkinkan unigram dan bigram, dan parameter max_features=5000 membatasi jumlah fitur yang digunakan. Untuk mengurangi kompleksitas data, parameter max_features=5000 digunakan untuk mengambil fitur kata yang dapat mengurangi jumlah kata yang muncul kurang dari dua. Sementara itu, parameter min_df=2 mengabaikan kata yang muncul kurang dari dua, dan parameter max_df=0.8 mengabaikan kata yang terlalu sering muncul di sebagian besar data.

F. Modeling

Pada tahap ini, SVM yang menjadi modeling pada penelitian ini digunakan untuk proses klasifikasi data yang dimasukkan ke dalam proses klasifikasi genre film dari hasil ekstraksi fitur TF-IF. Metode klasifikasi yang mencari hyperplane terbaik untuk memisahkan dat atar kelas berdasarkan fitur yang digunakan adalah SVM [23].

Sebelum proses pelatihan model dimulai, dilakukan pelatihan dan pengujian data dengan membagi data menjadi dua bagian. Data pelatihan digunakan untuk membangun model klasifikasi SVM, sedangkan data pengujian digunakan untuk mengevaluasi kinerja model dalam memprediksi genre film berdasarkan sinopsis. Karena kernel linear memiliki kinerja yang baik dalam menangani data teks berukuran besar, proses klasifikasi SVM menggunakan kernel ini.

Pada tahap pelatihan, model SVM mempelajari pola dari fitur TF-IDF untuk membentuk hyperplane optimal yang mampu memisahkan setiap kelas genre film. Apabila terdapat dua fitur TF-IDF yang direpresentasikan sebagai x_1 dan x_2, maka hyperplane dapat dituliskan dalam bentuk:

$$2x_1 + 3x_2 - 6 = 0 \quad (6)$$

Jika suatu data memiliki nilai fitur $x_1 = 1$ dan $x_2 = 2$, maka proses klasifikasi dilakukan dengan menghitung nilai hyperplane sebagai berikut:

$$2(1) + 3(2) - 6 = 2 \quad (7)$$

Karena hasil perhitungan bernilai positif, maka data diklasifikasikan ke dalam kelas genre tertentu Sebaliknya, apabila hasil perhitungan bernilai negatif, maka data akan diklasifikasikan ke kelas lainnya.

G. Evaluasi

e-ISSN: 2460-870X | p-ISSN: 2085-1367

Untuk mengetahui seberapa baik model klasifikasi yang telah dibagun untuk setiap kelas genre film, evaluasi dilakukan di akhir pemrosesan dengan menggunakan metode confusion matrix. Confusion matrix kan menghasilkan sejumlah metrik pengujian seperti *accuracy, precision, recall* dan *F1-Score.*

Tahap akhir dari proses data mining adalah tahap evaluasi. Tujuan dari tahap ini adalah untuk mengetahui seberapa baik model klasifikasi yang telah dibangun untuk setiap kelas genre film berfungsi. Evaluasi dilakukan dengan menggunakan metode *confusion matrix*, yang menghasilkan sejumlah metrik pengujian, yaitu *accuracy*, *precision*, *recall*, dan *F1-score* [24].

Berda+sarkan nilai *Confusion matrix* tersebut, selanjutnya dilakukan perhitungan *accuracy*, *precision*, *recall*, dan *F1-score* [25].

Berdasarkan nilai *Confusion matrix* tersebut, selanjutnya dilakukan perhitungan *accuracy*, *precision*, *recall*, dan *F1-score* [28].

1. *Accuracy* digunakan untuk menghitung jumlah prediksi yang berhasil diklasifikasikan dengan benar terhadap data pengujian secara keseluruhan.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} X\ 100 \quad (8)$$

2. *Precision* digunakan untuk menentukan seberapa tepat model dalam memberikan prediksi untuk kelas tertentu.

$$Precision = \frac{TP}{TP+FP} X\ 100 \quad (9)$$

3. *Recall* digunakan untuk mengukur kemampuan model untuk menemukan seluruh data yang benar-benar termasuk ke dalam suatu kelas.

$$Recall = \frac{TP}{TP+FN} X\ 100 \quad (10)$$

4. *F1-score* adalah metrik yang menggabungkan nilai *Precision* dan *Recall* untuk memberikan gambaran yang lebih seimbang tentang bagaimana model klasifikasi bekerja.

$$F1\ Score = 2\ X\ \frac{Precision \times Recall}{Precision+Recal} \quad (11)$$

3. HASIL DAN PEMBAHASAN

A. Pembahasan

1) *Dataset*

Dataset yang digunakan dalam penelitian ini diperoleh dari platform Kaggle dengan judul IMDb Movie Genre Classification. *Dataset* yang didapat terdiri dari dua bagian data yang terpisah, kemudian dilakukan proses integrasi data sehingga menjadi satu *dataset* yang utuh.

Dataset berisi data teks berupa sinopsis film (overview) beserta label genre film. Jumlah awal *dataset* yang digunakan sebanyak 10.000 data dengan karakteristik *multi-label*, yaitu satu data film dapat memiliki lebih dari satu genre. Atribut yang digunakan dalam penelitian ini terdiri dari atribut *overview* sebagai data teks sinopsis dan atribut genre sebagai label klasifikasi.

Tabel 2. Contoh *dataset*

Title	Overview	Genre
The Pianist	The actual account of pianist Wałysław Szpilman's experiences during the Nazi occupation of Warsaw. Szpilman gets work playing at a café	Drama, War

DOI: https://www.doi.org/10.22303/csrid.18.2.2026.381-394

Title	Overview	Genre
	when the city's Jews are forced into a ghetto. When his family is deported in 1942, he stays behind, works as a laborer for a while, and eventually hides among the city's ruins.	
Whiplash	A gifted young drummer starts to strive for excellence at any costs, even his humanity, under the guidance of a cruel master.	Drama, Music
Se7en	Two homicide detectives are on a desperate hunt for a serial killer whose crimes are based on the "seven deadly sins" in this dark and haunting film that takes viewers from the tortured remains of one victim to the next. The seasoned Det. Somerset researches each sin in an effort to get inside the killer's mind, while his novice partner, Mills, scoffs at his efforts to unravel the case.	Crime, Mystery, Thriller

2) Data Preparation

Tahap data preparation dilakukan untuk mempersiapkan *dataset* sebelum proses ekstraksi fitur dan klasifikasi menggunakan algoritma SVM. Proses ini meliputi pemilihan atribut, pembersihan data, transformasi *multi-label* menjadi *Single-label*, serta penyeimbangan kelas (*class balancing*). Pada tahap pemilihan atribut digunakan atribut *overview* sebagai data teks sinopsis dan atribut genre sebagai label klasifikasi. Setelah dilakukan pemilihan label, selanjutnya dilakukan pengecekan pada data untuk menghapus data kosong (missing value).

```
overview    1
genre       3
```

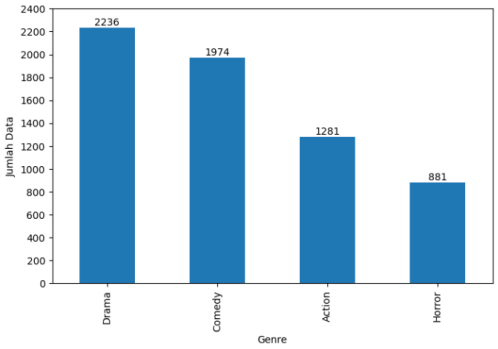
Gambar 2. *Hasil pengecekan missing value dataset*

Berdasarkan hasil pengecekan pada Gambar 2, ditemukan data kosong (missing value) pada atribut *overview* sebanyak 1 data dan atribut genre sebanyak 3 data. Data yang memiliki nilai kosong tersebut kemudian dihapus karena dapat mempengaruhi proses pengolahan dan klasifikasi data. Jumlah data yang digunakan setelah proses pembersihan menjadi sebanyak 9.996 data.

```
overview    0
genre       0
```

Gambar 3. *Hasil pembersihan missing value dataset*

Setelah proses pengecekan missing value dilakukan, tahapan berikutnya adalah dilakukan proses transformasi data *multi-label* menjadi *Single-label* dengan memilih genre pertama pada atribut genre sebagai label utama film. Penelitian ini hanya menggunakan empat genre, yaitu Action, Comedy, Drama, dan Horror. Pemilihan keempat genre tersebut didasarkan pada jumlah frekuensi kemunculan data terbanyak pada *dataset*.

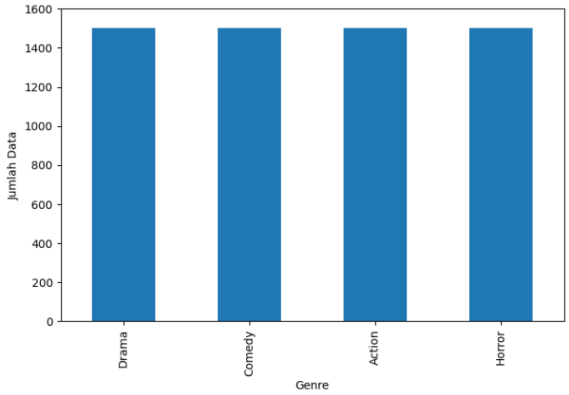


Gambar 4. *Hasil distribusi setelah transformasi Single-label*

e-ISSN: 2460-870X | p-ISSN: 2085-1367

Pada Gambar 4, distribusi jumlah data pada setiap genre masih belum seimbang. Genre *Drama* memiliki jumlah data terbanyak sebanyak 2236 data, sedangkan genre *Horror* memiliki jumlah data paling sedikit sebanyak 881 data. Ketidakseimbangan distribusi data tersebut dapat mempengaruhi performa model klasifikasi sehingga diperlukan proses *class balancing*.

Proses *class balancing* dilakukan menggunakan teknik *Random Oversampling* pada empat genre utama, yaitu *Action*, *Comedy*, *Drama*, dan *Horror*. Teknik ini dilakukan dengan menambahkan data secara acak pada kelas minoritas hingga jumlah data pada setiap kelas menjadi seimbang.



Gambar 5. *Hasil distribusi class balancing dataset*

Berdasarkan Gambar 5, proses *Random Oversampling* berhasil menyeimbangkan jumlah data pada setiap genre sehingga masing-masing kelas memiliki 1500 data.

3) *Text preprocessing*

Tahapan *text preprocessing* yang dilakukan pada penelitian ini terdiri dari *Cleaning Text*, *Case Folding*, *Tokenizing*, *Stopword Removal*, dan *Lemmatization*.

- *Cleaning Text (*Pembersihan Teks)

Proses pembersihan teks melibatkan penghapusan karakter yang tidak diperlukan, seperti tanda baca, angka, simbol, dan spasi berlebihan pada data sinopsis film. Berikut adalah contoh hasil pembersihan teks: pada tabel 3.

Tabel 3. Contoh hasil *Cleaning Text*

Kalimat asli	**Setelah Cleaning Text**
At the height of the Cold War, a mysterious criminal organization plans to use nuclear weapons and technology to upset the fragile balance of power between the United States and Soviet Union. CIA agent Napoleon Solo and KGB agent Illya Kuryakin are forced to put aside their hostilities and work together to stop the evildoers in their tracks. The duo's only lead is the daughter of a missing German scientist, whom they must find soon to prevent a global catastrophe.	At the height of the Cold War a mysterious criminal orgnization plans to use nuclear weapons and technology to upset the fragile blance of power between the United Statess and Soviet Union CIA agent Napoleon Solo and KGB agent Illya Kuryakin are forced to put aside their hostilites and work together to stop the evildoers in their tracks The duos only lead is the daughter of missing German scientist whom they must find soon to prevent a global catastrophe

- *Case Folding*

DOI: https://www.doi.org/10.22303/csrid.18.2.2026.381-394

Tujuan dari teknik ini adalah untuk mengubah teks sinopsis menjadi huruf kecil. Contoh hasil *Case Folding* ditunjukan pada tabel 4.

Tabel 4. Contoh hasil *Case Folding*

Kalimat Case Folding
at the height of the cold war a mysterious criminal organization plans to use nuclear weapons and technology to upset the fragile blance of power betwen the united states and soviet union cia agent napoleon solo and hostilites and work together to stop the evildoers in their the duos only lead is the daughter od missing german scientist whom they must find soon to prevent a global catastrophe

- *Tokenizing*

Tahap ini bertujuan untuk memecah kalimat pada sinopsis menjadi bagian kata atau token. Contoh hasil *Tokenizing* ditunjukan pada tabel 5.

Tabel 5. Contoh hasil *Tokenizing*

Kalimat Tokenizing
'at', 'the', 'height', 'of', 'the', 'cold', 'war', 'a', 'mysterious', 'criminal', 'organization', 'plans', 'to', 'use', 'nuclear', 'weapons', 'and', 'technology', 'to', 'upset', 'the', 'fragile', 'balance', 'of', 'power', 'between', 'the', 'united', 'states', 'and', 'soviet', 'union', 'cia', 'agent', 'napoleon', 'solo', 'and', 'kgb', 'agent', 'illya', 'kuryakin', 'are', 'forced', 'to', 'put', 'aside', 'their', 'hostilities', 'and', 'work', 'together', 'to', 'stop', 'the', 'evildoers', 'in', 'their', 'tracks', 'the', 'duos', 'only', 'lead', 'is', 'the', 'daughter', 'of', 'a', 'missing', 'german', 'scientist', 'whom', 'they', 'must', 'find', 'soon', 'to', 'prevent', 'a', 'global', 'catastrophe'

- *Stopword Removal*

Stopword removal dilakukan dengan menghapus kata umum yang dianggap tidak memiliki informasi apapun. Contoh hasil *Stopword Removal* ditunjukkan pada Tabel 6.

Tabel 6. Contoh hasil *Stopword Removal*

Kalimat Stopword Removal
'height', 'cold', 'war', 'mysterious', 'criminal', 'organization', 'plans', 'use', 'nuclear', 'weapons', 'technology', 'upset', 'fragile', 'balance', 'power', 'united', 'states', 'soviet', 'union', 'cia', 'agent', 'napoleon', 'solo', 'kgb', 'agent', 'illya', 'kuryakin', 'forced', 'put', 'aside', 'hostilities', 'work', 'together', 'stop', 'evildoers', 'tracks', 'duos', 'lead', 'daughter', 'missing', 'german', 'scientist', 'must', 'find', 'soon', 'prevent', 'global', 'catastrophe'

- *Lematization*

Tahap *Lemmatization* dilakukan untuk mengubah kata ke bentuk dasar semula kata. Contoh hasil *Lemmatization* ditunjukkan pada Tabel 7.

Tabel 7. Contoh hasil *Lematization*

Kalimat Lematization
'height', 'cold', 'war', 'mysterious', 'criminal', 'organization', 'plan', 'use', 'nuclear', 'weapons', 'technology', 'upset', 'fragile', 'balance', 'power', 'unite', 'state', 'soviet', 'union', 'cia', 'agent', 'napoleon', 'solo', 'kgb', 'agent', 'illya', 'kuryakin', 'force', 'put', 'aside', 'hostilities', 'work', 'together', 'stop', 'evildoers', 'track', 'duos', 'lead', 'daughter', 'miss', 'german', 'scientist', 'must', 'find', 'soon', 'prevent', 'global', 'catastrophe'

Dari tahapan *text preprocessing* yang telah dilakukan, data sinopsis film berhasil dibersihkan dan dinormalisasi dan siap digunakan pada tahap ekstraksi fitur TF-IDF dan N-Gram.

e-ISSN: 2460-870X | p-ISSN: 2085-1367

4) Ekstraksi Fitur

Setelah tahap *text preprocessing* selesai dilakukan, data teks kemudian diekstraksi menggunakan metode N-Gram dan TF-IDF untuk menghasilkan representasi numerik dari sinopsis film.

Tabel 8 Contoh hasil bobot TF-IDF

Term	TF-IDF Score
friends	0.008518
high school	0.004766
world war	0.002474
serial killer	0.001615

Tabel 8 menunjukkan hasil pembobotan TF-IDF, yang menunjukkan bahwa fitur kata memiliki nilai bobot yang berbeda sesuai dengan tingkat kemunculannya dalam dataset. Kata dengan nilai TF-IDF yang lebih tinggi menunjukkan bahwa kata tersebut lebih spesifik dalam mewakili genre film tertentu. Selanjutnya, hasil pembobotan ini dimasukkan ke dalam proses klasifikasi dengan algoritma SVM.

5) Pembagian Data

Pada tahap modeling, SVM digunakan untuk mengklasifikasikan genre film. Hasil dari ekstraksi fitur TF-IDF berbasis N-Gram dibagi menjadi data pelatihan dan data pengujian dengan rasio 80:20. Model klasifikasi menggunakan kernel *linear* untuk memproses teks yang dihasilkan dari ekstraksi fitur. Setelah pelatihan selesai, model digunakan untuk memprediksi genre film berdasarkan data pengujian.

B. Hasil Evaluasi Model

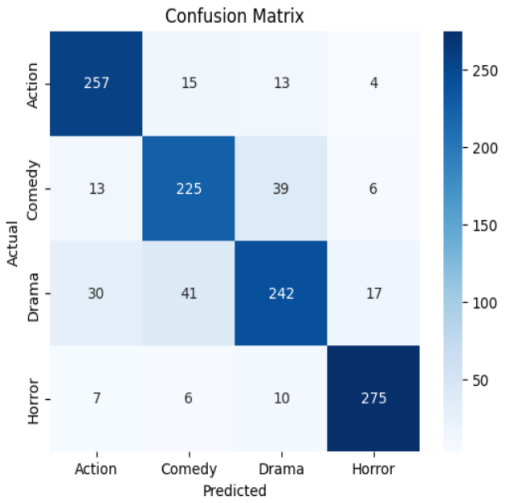
Algoritma *Support Vector Machine* (SVM) digunakan untuk melakukan proses klasifikasi genre film menggunakan fitur yang diperoleh dari metode TF-IDF berbasis N-Gram, dan data pengujian (*testing data*) sebesar 20% dari seluruh dataset digunakan untuk menguji model. Hasil pengujian menunjukkan bahwa keakuratan model sebesar 83% lebih tinggi dibandingkan dengan hasil penelitian sebelumnya yang berbasis aturan TF-IDF sebesar 78,5% [26]. Hasil ini menunjukkan bahwa model dibuat dapat mengklasifikasikan genre film berdasarkan sinopsis dengan tingkat ketepatan yang cukup baik.

Tabel 9 Hasil evaluasi model

	Precision	***Recall***	***F1-score***	**Support**
Action	84%	89%	86%	289
Comedy	78%	80%	79%	283
Drama	80%	73%	76%	330
Horor	91%	92%	92%	298
Accuracy		83%		

Genre horor memiliki performa terbaik, menurut Tabel 9, dengan nilai *precision* sebesar 0,91, *recall* sebesar 0.92, dan nilai F1-sebesar 0.92, menunjukkan bahwa karakteristik kata genre horor lebih mudah dikenali oleh model klasifikasi. Genre Drama memiliki nilai F1-sebesar paling rendah, yaitu 0.76. Situasi ini menunjukkan bahwa genre drama memiliki ciri-ciri teks yang lebih beragam dan lebih mirip dengan genre lain.

DOI: https://www.doi.org/10.22303/csrid.18.2.2026.381-394



Gambar 6. *Hasil confusion matrix*

Berdasarkan hasil confusion matrix pada gambar 6, Pada masing-masing genre data berhasil diklasifikasi dengan benar. Jumlah data yang benar untuk genre Action adalah 257, genre Comedy adalah 225, genre Drama adalah 242, dan genre Horror adalah 275. Namun, masih ada beberapa kesalahan dalam klasifikasi genre, terutama pada genre Drama yang sering dianggap sebagai komedi karena kosakata yang mirip seperti hubungan sosial, keluarga, dan kehidupan sehari-hari, yang menunjukkan adanya kemiripan karakteristik kata pada kedua genre tersebut.

4. KESIMPULAN

Support Vector Machine (SVM) yang digunakan dengan paduan fitur TF-IDF dan N-Gram paa penelitian ini berhasil diterapkan untuk melakukan klasifikasi genre film berdasarkan sinopsis. Melalui tahapan prapemrosesan teks, pembersihan data, normalisasi huruf, pemecahan kata, penghapusan *stopword* dan *lemmatisasi* dapat meningkatkan kualitas representasi teks sehingga proses klasifikasi menjadi lebih efisien. Penggunaan N-Gram juga membantu model menemukan hubungan antar kata yang muncul dalam sinopsis film.

Hasil pengujian menunjukkan bahwa model yang dibangun mampu mencapai tingkat akurasi sebesar 83%. Genre horor memiliki nilai F1 tertinggi sebesar 92%, sedangkan genre drama memiliki nilai F1 yang lebih rendah karena karakteristik teksnya cenderung lebih beragam dan mirip dengan genre lain. Secara keseluruhan, kombinasi TF-IDF, N-Gram, dan SVM terbukti mampu memberikan kinerja yang baik dalam proses klasifikasi genre film berdasarkan sinopsis.

Penelitian ini masih memiliki beberapa keterbatasan. Dataset yang digunakan hanya difokuskan pada empat genre utama, yaitu Action, Comedy, Drama, dan Horror, sehingga belum dapat merepresentasikan seluruh variasi genre film yang tersedia. Selain itu, proses transformasi data multi-label menjadi single-label berpotensi menyebabkan hilangnya informasi genre lain yang juga dimiliki oleh suatu film. Untuk penelitian selanjutnya, pendekatan multi-label classification dapat diterapkan sehingga satu film dapat diklasifikasikan ke dalam lebih dari satu genre secara bersamaan.

REFERENSI

[1] K. Sleman, P. Daerah, and I. Yogyakarta, “SISTEM REKOMENDASI FILM INDONESIA MENGGUNAKAN COLLABORATIVE FILTERING,” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 5, pp. 9070–9076, 2025.

e-ISSN: 2460-870X | p-ISSN: 2085-1367

[2] R. Octavia and A. P. Yuwono, "Analisis Resepsi Penonton terhadap Makna Normalisasi Pelecehan Seksual pada Film Dear David," *J. Ilmu Komun.*, vol. 8, no. 1, pp. 11–23, 2024, [Online]. Available: https://doi.org/10.30596/ji.v8i1.14608 E-ISSN:

[3] A. Firizkiansah, A. Muhammad, I. R. Maulana, U. S. Indonesia, and K. Bekasi, "Optimasi Klasifikasi Data Teks Menggunakan Algoritma Logistic Regression dengan TF-IDF dan SMOTE," *JIKOMTI J. Ilm. Ilmu Komput. dan Teknol. Inf.*, vol. 2, no. 1, pp. 29–36, 2025.

[4] B. Falakhi, I. Cholissodin, and R. S. Perdana, "Klasifikasi Sinopsis Novel berdasarkan Jenis Genre menggunakan Multi- class Support Vector Machine dan Chi-square," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 1, pp. 192–202, 2023.

[5] W. M. Jihadul Akbar, Hairul Fahmi, "MULTI LABEL KLASIFIKASI GENRE FILM BERDASARKAN SINOPSIS MENGGUNAKAN METODE LONG SHORT-TERM MEMORY (LSTM)," *J. Manaj. Inform. Sist. Inf.*, vol. 8, pp. 110–119, 2025.

[6] A. Anugrah, M. Arif, C. Hamud, J. Zulkaidah, and R. Leo, "Penerapan Text Mining menggunakan Metode TF-IDF untuk Menentukan Genre Film," *Ilk. J. Comput. Sci. Appl. Informatics*, vol. 7, no. 3, pp. 415–422, 2025.

[7] Alaysha Rihadatul Aisy, Rahmaddeni, Azfar Huzaifah Siregar, Suandi Daulay, and Nita Ananta, "Implementasi Algoritma Support Vector Machine Dalam Mengklasifikasikan Berita Hoax Di Indonesia Pada Media Sosial X," *J. RESTIKOM Ris. Tek. Inform. dan Komput.*, vol. 7, no. 2, pp. 206–216, 2025, doi: 10.52005/restikom.v7i2.457.

[8] A. Shamim, "IMDb Movie Genre Classification Dataset," kaggle. [Online]. Available: https://www.kaggle.com/datasets/adilshamim8/nlp-task

[9] T. Gori *et al.*, "PREPROCESSING DATA DAN KLASIFIKASI UNTUK PREDIKSI KINERJA DATA PREPROCESSING AND CLASSIFICATION FOR PREDICTING STUDENT," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 1, pp. 215–224, 2024, doi: 10.25126/jtiik.20241118074.

[10] D. Lestari and S. Lestari, "Penerapan Data Mining Klasifikasi Tingkat Pemahaman Siswa Pada Kegiatan Belajar Mengajar dengan Metode Decision Tree (Studi Kasus SDN Malaka Jaya 11 Duren Sawit)," *Junrnal Indones. Manag. Inform. dan Komun.*, vol. 5, no. 2, pp. 1260–1268, 2024.

[11] M. Liu *et al.*, "Handling missing values in healthcare data : A systematic review of deep learning-based imputation techniques," *Artif. Intell. Med.*, pp. 1–34, 2022, [Online]. Available: 10.1016/j.artmed.2023.102587

[12] A. Ramadhani and N. S. Harahap, "Performance Comparison of Multilabel Text Classification Methods on Translated Hadiths of Bukhari Using Support Vector Machine and Long Short Term Memory Perbandingan Performa Metode Klasifikasi Teks Multilabel Hadis Terjemahan Bukhari Menggunakan Support," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 5, no. July, pp. 896–907, 2025.

[13] C. Kaope and Y. Pristyanto, "The Effect of Class Imbalance Handling on Datasets Toward Classification Algorithm Performance," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 22, no. 2, pp. 227–238, 2023, doi: 10.30812/matrik.v22i2.2515.

[14] Z. Abidin, T. Suratno, and M. F. Putri, "PENERAPAN RANDOM OVERSAMPLING DAN PRINCIPAL COMPONENT ANALYSIS UNTUK MENINGKATKAN AKURASI PREDIKSI KEBANGKRUTAN APPLICATION OF RANDOM OVERSAMPLING AND PRINCIPAL COMPONENT ANALYSIS TO ENHANCE THE ACCURACY OF BANKRUPTCY PREDICTION FOR," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 12, no. 5, 2025.

[15] H. Rifaldi, Dianda, Abdul Fadlil, "TEKNIK PREPROCESSING PADA TEXT MINING MENGGUNAKAN DATA TWEET 'MENTAL HEALTH,'" *Decod. J. Pendidik. Teknol. Inf.*, vol. 3, no. 2, pp. 161–171, 2023.

[16] R. R. Salam, M. F. Jamil, and Y. Ibrahim, "Sentiment Analysis of Cash Direct Assistance Distribution for Fuel Oil Using Support Vector Machine Analisis Sentimen Terhadap Bantuan Langsung Tunai (BLT) Bahan Bakar Minyak (BBM) Menggunakan Support Vector Machine," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. April, pp. 27–35, 2023.

[17] M. Amien, "SEJARAH DAN PERKEMBANGAN TEKNIK NATURAL LANGUAGE PROCESSING (NLP) BAHASA INDONESIA: TINJAUAN TENTANG SEJARAH, PERKEMBANGAN TEKNOLOGI, DAN APLIKASI NLP DALAM BAHASA INDONESIA," *ELANG J. Interdiscip. Res.*, no. 2007, pp. 1–7, 2022, [Online]. Available: https://doi.org/10.32664/elang.v1i01

[18] A. Nur, A. Saputra, R. E. Saputro, D. Intan, and S. Saputra, "Enhancing Sentiment Analysis Accuracy Using SVM and Slang Word Normalization on YouTube Comments," *Sink. J. dan Penelit. Tek. Inform.*, vol. 9, no. 2, pp. 687–699, 2025.

[19] S. J. Angelina, A. Bijaksana, P. Negara, and H. Muhardi, "Analisis Pengaruh Penerapan Stopword Removal Pada Performa Klasifikasi Sentimen Tweet Bahasa Indonesia Analyzing The Impact Of Applying Stopword Removal On Indonesian Tweet Sentiment Classification," *Juara J. Apl. dan Ris. Inform. Curr. Arch. About*, vol. 02, no. 1, pp. 165–173, 2023, doi: 10.26418/juara.v2i1.69680.

[20] R. Hafeez *et al.*, "Contextual Urdu Lemmatization using Recurrent Neural Network Models," *Multidiscip. Digit. Publ. Inst.*, pp. 1–20, 2023.

[21] D. Septiani, I. Isabela, F. Teknik, U. N. Jakarta, F. Teknik, and U. N. Jakarta, "Analisis term frequency inverse document frequency (tf-idf) dalam temu kembali informasi pada dokumen teks," *SINTESIA J. Sist. dan Teknol. Inf. Indones.*, vol. 25, pp. 81–88, 2023.

[22] K. Hadi and E. Utami, "Analysis of K-NN with the Integration of Bag of Words , TF-IDF , and N-Grams for Hate Speech Classification on Twitter," *JUITA J. Inform.*, vol. 12, no. 2, pp. 289–298, 2024.

[23] N. H. Ovirianti, M. Zarlis, and H. Mawengkang, "Support Vector Machine Using A Classification Algorithm," *Sink. J. dan Penelit. Tek. Inform.*, vol. 6, no. 3, pp. 2103–2107, 2022.

[24] S. Dewi, "Arus Jurnal Sains dan Teknologi (AJST) Analisis Sentimen Aplikasi WhatsApp berdasarkan Ulasan di," *Arus J. Sains dan Teknol.*, vol. 2, no. 2, 2024.

[25] A. Indrajaya, "PERBANDINGAN KLASIFIKASI KACANG KERING DENGAN ALGORITMA KNN DAN SVM LINEAR," *J. Komput. dan Inform.*, vol. 20, no. 1, pp. 62–70, 2025.

[26] N. Arifin, U. Enri, and N. Sulistiyowati, "PENERAPAN ALGORITMA SUPPORT VECTOR MACHINE (SVM) DENGAN TF-IDF N-GRAM UNTUK TEXT CLASSIFICATION," vol. 6, no. 2, 2021.