JURNAL ILMIAH PERPUSTAKAAN, SISTEM INFORMASI, INFORMATIKA & ELEKTRO
Volume 1, Nomor 1, Juni 2026
hlm. 38-55

Klasifikasi Genre Buku Digital Berdasarkan Ringkasan Teks Menggunakan Metode TF-IDF dan Support Vector Machine

Deswita Ratna Sari Dewi[1*], Egi Abinowi[2]
[1, 2] FISIP, Universitas Widyatama, Jl. Cikutra No. 204A, Sukapada, Kec. Cibeunying Kidul, Kota Bandung, Jawa Barat 40125
E-mail korespondensi: [1*]deswita.ratna@widyatama.ac.id, [2]egi.abinowi@widyatama.ac.id

Keywords: *big data, text classification, machine learning, support vector machine, text mining*

Abstract
The rapid growth of digital book collections poses challenges in effectively and consistently grouping books by genre. This study aims to classify book genres based on text summaries using a text mining approach. The dataset used is the Book Genre Dataset, which consists of 4,657 book documents across ten genre categories. The research process included text preprocessing, feature extraction using Term Frequency-Inverse Document Frequency (TF-IDF), and the development of a classification model using Support Vector Machines (SVM). The data was divided into a training set (80%) and a test set (20%) using stratified sampling. The results show that the classification model achieved an accuracy of 68.78%, a precision of 69.26%, a recall of 68.78%, and an F1-score of 68.52%. The Fantasy genre achieved the best performance with an F1-score of 0.77, while the Romance genre showed the lowest performance due to an imbalance in data distribution and similarities in text characteristics with other genres. The research findings indicate that book summaries contain sufficient information to support the process of automatic genre identification. The results of this study have the potential to be utilized in the development of digital libraries and content-based book recommendation systems.

Kata kunci: *data besar, klasifikasi teks, pembelajaran mesin, mesin vektor pendukung, penambangan teks*

Abstrak
Pertumbuhan koleksi buku digital dalam jumlah besar menimbulkan tantangan dalam proses pengelompokan genre secara efektif dan konsisten. Penelitian ini bertujuan untuk mengklasifikasikan genre buku berdasarkan ringkasan teks menggunakan pendekatan *text mining*. Dataset yang digunakan merupakan *Book Genre Dataset* yang terdiri atas 4.657 dokumen buku dengan sepuluh kategori genre. Proses penelitian meliputi prapengolahan teks, ekstraksi fitur menggunakan *Term Frequency-Inverse Document Frequency*, serta pembangunan model klasifikasi menggunakan *Support Vector Machine*. Data dibagi menjadi data pelatihan sebesar 80% dan data pengujian sebesar 20% dengan metode *stratified sampling*. Hasil penelitian menunjukkan bahwa model klasifikasi menghasilkan tingkat akurasi sebesar 68,78%, *precision* sebesar 69,26%, *recall* sebesar 68,78%, dan *F1-score* sebesar 68,52%. Genre *Fantasy* memperoleh performa terbaik dengan nilai *F1-score* sebesar 0,77, sedangkan genre *Romance* menunjukkan performa terendah akibat ketidakseimbangan distribusi data dan kemiripan karakteristik teks dengan genre lain. Temuan penelitian menunjukkan bahwa ringkasan buku mengandung informasi yang cukup untuk mendukung proses identifikasi genre secara otomatis. Hasil

penelitian berpotensi dimanfaatkan dalam pengembangan perpustakaan digital dan sistem rekomendasi buku berbasis konten.

PENDAHULUAN

Perkembangan teknologi informasi dan komunikasi telah mendorong pertumbuhan data digital dalam jumlah yang sangat besar pada berbagai sektor, termasuk pendidikan, penerbitan, dan perpustakaan digital. Fenomena ini dikenal sebagai *big data*, yaitu kumpulan data yang memiliki volume, variasi, dan kecepatan pertumbuhan yang tinggi sehingga memerlukan teknik analisis khusus untuk menghasilkan informasi yang bernilai. Salah satu jenis data yang terus mengalami peningkatan adalah data teks tidak terstruktur (*unstructured text data*), seperti artikel, berita, ulasan pengguna, dokumen akademik, serta ringkasan buku digital. Berbeda dengan data terstruktur yang tersimpan dalam bentuk tabel, data teks memerlukan proses ekstraksi dan pengolahan khusus agar dapat digunakan sebagai sumber informasi untuk mendukung pengambilan keputusan (Minaee et al., 2021).

Dalam konteks perpustakaan digital dan industri penerbitan, pertumbuhan jumlah koleksi buku elektronik menyebabkan proses pengelolaan informasi menjadi semakin kompleks. Ribuan bahkan jutaan koleksi buku tersedia dalam berbagai kategori sehingga proses identifikasi dan pengelompokan genre secara manual menjadi kurang efisien. Genre merupakan salah satu metadata penting yang digunakan untuk mengorganisasi koleksi, mendukung sistem temu kembali informasi (*information retrieval*), dan meningkatkan kualitas sistem rekomendasi buku. Namun demikian, penentuan genre masih sering dilakukan secara manual oleh penerbit atau pengelola koleksi, sehingga rentan terhadap inkonsistensi dan membutuhkan sumber daya yang besar. Oleh karena itu, diperlukan pendekatan otomatis yang mampu mengidentifikasi genre buku berdasarkan informasi yang terkandung dalam isi dokumen.

Perkembangan bidang *data mining* dan *machine learning* telah membuka peluang untuk melakukan klasifikasi dokumen secara otomatis melalui pendekatan *text mining*. Klasifikasi teks merupakan proses pengelompokan dokumen ke dalam kategori tertentu berdasarkan karakteristik linguistik yang terkandung di dalamnya. Dalam beberapa tahun terakhir, penelitian mengenai klasifikasi teks berkembang pesat seiring meningkatnya kebutuhan pengolahan data tidak terstruktur pada berbagai domain. Menurut Kowsari et al. (2019), klasifikasi teks menjadi salah satu bidang penelitian yang paling aktif dalam *data mining* karena kemampuannya dalam mengubah data teks menjadi informasi yang dapat dimanfaatkan untuk berbagai keperluan analitis dan prediktif. Proses klasifikasi umumnya dilakukan melalui tahapan prapengolahan data,

ekstraksi fitur, dan pembangunan model pembelajaran mesin yang mampu mengenali pola tertentu dalam dokumen.

Salah satu teknik ekstraksi fitur yang masih banyak digunakan hingga saat ini adalah *Term Frequency–Inverse Document Frequency* (TF-IDF). Metode ini memberikan bobot pada kata berdasarkan frekuensi kemunculannya dalam suatu dokumen dan tingkat keunikannya pada keseluruhan koleksi dokumen. TF-IDF terbukti mampu menghasilkan representasi teks yang efektif untuk berbagai tugas klasifikasi karena dapat mengurangi pengaruh kata-kata umum yang tidak memiliki nilai diskriminatif tinggi (Ramos, 2003). Selain itu, algoritma Support Vector Machine (SVM) dikenal sebagai salah satu metode klasifikasi yang memiliki performa baik pada data berdimensi tinggi, termasuk data teks. Penelitian oleh Joachims (1998) menunjukkan bahwa SVM mampu membangun batas klasifikasi yang efektif pada dokumen teks yang memiliki jumlah fitur sangat besar.

Penelitian mengenai klasifikasi teks telah banyak dilakukan dalam satu dekade terakhir. Minaee et al. (2021) menunjukkan bahwa berbagai metode klasifikasi, baik berbasis *machine learning* maupun *deep learning*, mampu menghasilkan performa yang tinggi pada tugas kategorisasi dokumen. Penelitian tersebut menemukan bahwa kualitas representasi fitur memiliki pengaruh yang signifikan terhadap keberhasilan proses klasifikasi. Penelitian lain yang dilakukan oleh Kowsari et al. (2019) menyimpulkan bahwa kombinasi teknik ekstraksi fitur dan algoritma klasifikasi tradisional masih mampu menghasilkan performa yang kompetitif pada berbagai jenis dataset teks. Sementara itu, Sun et al. (2019) menunjukkan bahwa model berbasis *transformer* seperti BERT memberikan peningkatan performa klasifikasi melalui pemahaman konteks bahasa yang lebih baik dibandingkan metode konvensional.

Meskipun penelitian klasifikasi teks telah berkembang pesat, sebagian besar penelitian sebelumnya berfokus pada data berita, ulasan produk, media sosial, atau analisis sentimen. Penelitian yang menggunakan ringkasan buku sebagai objek utama masih relatif terbatas, terutama pada dataset yang memiliki jumlah dokumen besar dan variasi genre yang beragam. Selain itu, sebagian penelitian terdahulu lebih banyak menitikberatkan pada perbandingan algoritma klasifikasi tanpa mengeksplorasi karakteristik data buku digital sebagai bagian dari ekosistem *big data*. Kondisi ini menunjukkan adanya kesenjangan penelitian (*research gap*) terkait penerapan teknik *text mining* untuk klasifikasi genre buku pada koleksi digital berskala besar.

Kesenjangan penelitian tersebut menjadi semakin relevan mengingat meningkatnya penggunaan perpustakaan digital dan platform distribusi buku elektronik. Pengelolaan koleksi dalam jumlah besar membutuhkan sistem klasifikasi yang mampu bekerja secara otomatis dan konsisten. Selain itu, keberagaman genre buku menyebabkan proses identifikasi kategori menjadi

lebih kompleks dibandingkan klasifikasi dokumen pada domain lain. Oleh karena itu, diperlukan penelitian yang tidak hanya menguji kemampuan model klasifikasi, tetapi juga mengevaluasi bagaimana karakteristik ringkasan buku dapat dimanfaatkan untuk membedakan genre secara sistematis.

Penelitian ini berupaya mengisi kesenjangan tersebut dengan memanfaatkan dataset yang terdiri atas 4.657 dokumen buku yang mencakup sepuluh kategori genre, yaitu *thriller*, *fantasy*, *science*, *history*, *horror*, *crime*, *romance*, *psychology*, *sports*, dan *travel*. Berbeda dengan sebagian penelitian sebelumnya yang menggunakan dataset dengan jumlah kategori terbatas, penelitian ini memanfaatkan koleksi buku digital yang lebih beragam sehingga memungkinkan analisis yang lebih representatif terhadap kondisi nyata pengelolaan informasi digital. Selain itu, penelitian ini mengombinasikan metode TF-IDF dan Support Vector Machine untuk mengevaluasi kemampuan klasifikasi genre berdasarkan ringkasan buku yang tersedia.

Secara teoritis, penelitian ini didasarkan pada konsep *Knowledge Discovery in Databases* (KDD), yang menjelaskan bahwa pengetahuan dapat diperoleh melalui serangkaian tahapan yang meliputi seleksi data, pembersihan data, transformasi data, *data mining*, dan interpretasi hasil. Dalam penelitian ini, tahapan tersebut diwujudkan melalui proses prapengolahan teks, ekstraksi fitur menggunakan TF-IDF, pembangunan model klasifikasi menggunakan SVM, dan evaluasi performa model menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Pendekatan tersebut memungkinkan data teks yang awalnya tidak terstruktur diubah menjadi informasi yang dapat digunakan untuk mendukung proses pengambilan keputusan pada sistem pengelolaan buku digital.

Kontribusi penelitian ini terletak pada penerapan teknik *text mining* untuk klasifikasi genre buku menggunakan dataset berskala besar yang merepresentasikan karakteristik data teks dalam lingkungan *big data*. Selain memberikan gambaran mengenai efektivitas kombinasi TF-IDF dan SVM dalam mengidentifikasi genre buku, penelitian ini juga memberikan pemahaman mengenai pengaruh distribusi data dan karakteristik linguistik terhadap performa klasifikasi. Hasil penelitian diharapkan dapat menjadi referensi bagi pengembangan sistem rekomendasi buku, pengelolaan perpustakaan digital, serta penelitian lanjutan dalam bidang *big data analytics* dan klasifikasi teks.

Berdasarkan uraian tersebut, tujuan penelitian ini adalah menganalisis karakteristik ringkasan buku sebagai sumber informasi tekstual, membangun model klasifikasi genre buku menggunakan metode TF-IDF dan Support Vector Machine, serta mengevaluasi performa model dalam mengidentifikasi genre secara otomatis berdasarkan isi ringkasan buku. Melalui penelitian ini diharapkan diperoleh pemahaman yang lebih komprehensif mengenai penerapan *text mining* pada

data buku digital serta kontribusinya terhadap pengembangan ilmu pengetahuan di bidang *big data* dan *data mining*.

METODE

Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode *text mining* dan klasifikasi teks untuk mengidentifikasi genre buku berdasarkan ringkasan (*summary*) yang terdapat pada dataset. Penelitian dilakukan menggunakan pendekatan *supervised learning*, yaitu proses pembelajaran mesin yang memanfaatkan data berlabel sebagai dasar pembentukan model klasifikasi. Genre buku digunakan sebagai label kelas (*class label*), sedangkan ringkasan buku digunakan sebagai sumber fitur yang dianalisis.

Tahapan penelitian dilaksanakan melalui proses akuisisi data, prapengolahan teks (*text preprocessing*), transformasi fitur menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF), pembangunan model klasifikasi menggunakan algoritma Support Vector Machine (SVM), serta evaluasi performa model menggunakan beberapa metrik pengukuran. Pendekatan ini dipilih karena mampu menangani data teks dalam jumlah besar dan menghasilkan model klasifikasi yang efektif pada data berdimensi tinggi (Minaee et al., 2021).

Subjek dan Sumber Data Penelitian

Subjek penelitian berupa dokumen teks yang berasal dari dataset buku digital. Dataset terdiri atas 4.657 data buku yang memuat tiga atribut utama, yaitu judul buku (*title*), genre buku (*genre*), dan ringkasan buku (*summary*).

Data penelitian mencakup sepuluh kategori genre, yaitu *thriller*, *fantasy*, *science*, *history*, *horror*, *crime*, *romance*, *psychology*, *sports*, dan *travel*. Ringkasan buku digunakan sebagai unit analisis karena mengandung informasi yang merepresentasikan isi dan karakteristik setiap genre.

Distribusi data menunjukkan bahwa genre *thriller* merupakan kategori dengan jumlah dokumen terbesar, sedangkan genre *sports*, *travel*, dan *psychology* merupakan kategori dengan jumlah dokumen terkecil. Variasi jumlah dokumen antar-genre memberikan kondisi yang sesuai untuk menguji kemampuan model klasifikasi pada data yang memiliki karakteristik *imbalanced class*.

Variabel Penelitian

Penelitian ini menggunakan dua jenis variabel, yaitu variabel bebas dan variabel terikat.

Variabel Bebas (Independent Variable)

Variabel bebas berupa teks ringkasan buku (*summary*) yang diproses menggunakan teknik *text mining*. Ringkasan buku ditransformasikan menjadi representasi numerik menggunakan metode TF-IDF sehingga dapat diproses oleh algoritma klasifikasi.

Variabel Terikat (Dependent Variable)

Variabel terikat berupa genre buku yang terdiri atas sepuluh kategori:

1. Thriller
2. Fantasy
3. Science
4. History
5. Horror
6. Crime
7. Romance
8. Psychology
9. Sports
10. Travel

Genre digunakan sebagai label target yang diprediksi oleh model klasifikasi.

Instrumen Penelitian

Instrumen penelitian terdiri atas perangkat keras dan perangkat lunak yang digunakan selama proses analisis data.

Perangkat Keras

Penelitian dilakukan menggunakan komputer dengan spesifikasi yang mendukung pengolahan data teks dan proses pelatihan model *machine learning*.

Perangkat Lunak

Perangkat lunak yang digunakan meliputi:

- Python sebagai bahasa pemrograman utama.
- Pandas untuk pengolahan dan manipulasi data.
- NumPy untuk komputasi numerik.
- Scikit-learn untuk ekstraksi fitur TF-IDF, pembangunan model SVM, dan evaluasi performa model.
- Matplotlib untuk visualisasi data.
- NLTK untuk mendukung proses prapengolahan teks.

Selain perangkat lunak tersebut, instrumen utama penelitian adalah dataset buku digital yang berisi 4.657 dokumen teks.

Prosedur Penelitian

Penelitian dilaksanakan melalui beberapa tahapan yang dirancang agar dapat direplikasi oleh peneliti lain.

Seleksi Data

Tahap awal dilakukan dengan memilih atribut yang relevan dengan tujuan penelitian. Atribut yang digunakan meliputi:

- *summary* sebagai sumber fitur teks.
- *genre* sebagai label klasifikasi.

Atribut lain yang tidak berkontribusi terhadap proses klasifikasi tidak digunakan dalam proses pemodelan.

Prapengolahan Teks (*Text Preprocessing*)

Prapengolahan dilakukan untuk meningkatkan kualitas data sebelum proses klasifikasi. Tahapan yang dilakukan meliputi:

1. **Case Folding**

Seluruh huruf diubah menjadi huruf kecil untuk menghindari perbedaan representasi kata akibat penggunaan huruf kapital.

2. **Punctuation Removal**

Karakter khusus, angka, dan tanda baca dihapus dari dokumen.

3. **Tokenization**

Ringkasan buku dipecah menjadi unit kata (*token*) untuk memudahkan proses analisis.

4. **Stopword Removal**

Kata-kata umum yang tidak memiliki kontribusi signifikan terhadap proses klasifikasi dihapus dari dokumen.

Tahapan ini menghasilkan korpus teks yang lebih bersih dan siap digunakan dalam proses ekstraksi fitur.

Transformasi Fitur

Dokumen hasil prapengolahan kemudian diubah menjadi representasi numerik menggunakan metode TF-IDF.

Pembobotan TF-IDF dihitung menggunakan Persamaan (1).

$$TFIDF(t,d) = TF(t,d) \times \log\left(\frac{N}{DF(t)}\right) \quad (1)$$

Keterangan:

- $TF(t,d)$= frekuensi kemunculan kata $t$pada dokumen d
- $DF(t)$= jumlah dokumen yang mengandung kata t
- N= jumlah keseluruhan dokumen

Pada penelitian ini digunakan maksimum 10.000 fitur TF-IDF yang memiliki nilai informasi tertinggi untuk merepresentasikan dokumen.

Pembagian Data

Dataset dibagi menjadi dua kelompok data, yaitu:

- 80% data pelatihan (*training set*)
- 20% data pengujian (*testing set*)

Pembagian dilakukan menggunakan metode *stratified sampling* agar proporsi masing-masing genre tetap terjaga pada data pelatihan maupun data pengujian.

Dari total 4.657 dokumen, sebanyak 3.725 dokumen digunakan sebagai data pelatihan dan 932 dokumen digunakan sebagai data pengujian.

Pembangunan Model Klasifikasi

Model klasifikasi dibangun menggunakan algoritma Linear Support Vector Machine (Linear SVM). Algoritma ini dipilih karena memiliki kemampuan yang baik dalam menangani data berdimensi tinggi dan terbukti efektif pada berbagai penelitian klasifikasi teks (Joachims, 1998). Model dilatih menggunakan matriks TF-IDF yang dihasilkan pada tahap sebelumnya. Setelah proses pelatihan selesai, model digunakan untuk memprediksi genre pada data pengujian.

Teknik Analisis Data

Analisis data dilakukan melalui dua tahap, yaitu analisis deskriptif dan analisis inferensial berbasis *machine learning.*

Analisis Deskriptif

Analisis deskriptif digunakan untuk menggambarkan karakteristik dataset yang meliputi:

- Distribusi jumlah dokumen pada setiap genre.
- Persentase masing-masing genre.
- Statistik panjang ringkasan buku.

Hasil analisis disajikan dalam bentuk tabel dan grafik untuk memberikan gambaran umum mengenai karakteristik data penelitian.

Evaluasi Model Klasifikasi

Kinerja model dievaluasi menggunakan *confusion matrix* dan empat metrik evaluasi, yaitu *accuracy, precision, recall,* dan *F1-score*.

Nilai *accuracy* dihitung menggunakan Persamaan (2).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (2)$$

Nilai *precision* dihitung menggunakan Persamaan (3).

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

Nilai *recall* dihitung menggunakan Persamaan (4).

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

Keterangan:

- TP = *True Positive*
- TN = *True Negative*
- FP = *False Positive*
- FN = *False Negative*

Evaluasi dilakukan untuk mengetahui kemampuan model dalam mengidentifikasi genre buku berdasarkan informasi yang terkandung dalam ringkasan. Hasil evaluasi kemudian dianalisis untuk mengetahui genre yang memiliki performa klasifikasi tertinggi maupun terendah serta faktor-faktor yang memengaruhinya.

HASIL

Statistika Deskriptif

Dataset yang digunakan terdiri atas 4.657 dokumen buku yang tersebar ke dalam sepuluh kategori genre. Analisis deskriptif dilakukan untuk mengetahui distribusi data pada masing-masing kelas sebelum dilakukan proses klasifikasi.

Tabel 1. Distribusi Genre Buku

Genre	Frekuensi	Persentase (%)
Thriller	1.023	21,97
Fantasy	876	18,81
Science	647	13,89
History	600	12,88
Horror	600	12,88
Crime	500	10,74
Romance	111	2,38
Psychology	100	2,15
Sports	100	2,15
Travel	100	2,15

Total	**4.657**	**100,00**

Hasil pada Tabel 1 menunjukkan bahwa genre *thriller* merupakan kategori dominan dengan proporsi sebesar 21,97%, sedangkan genre *psychology*, *sports*, dan *travel* merupakan kategori dengan jumlah data paling sedikit. Ketidakseimbangan distribusi kelas ini menunjukkan adanya karakteristik *imbalanced dataset* yang umum ditemukan pada data teks berskala besar.

Hasil Prapengolahan Data

Tahap *preprocessing* dilakukan melalui proses *case folding*, penghapusan tanda baca, tokenisasi, dan pembobotan menggunakan TF-IDF. Hasil transformasi menghasilkan matriks fitur berdimensi tinggi yang merepresentasikan karakteristik setiap dokumen dalam bentuk numerik.

Pembobotan TF-IDF dihitung menggunakan Persamaan (1).

$$TFIDF(t,d) = TF(t,d) \times \log\left(\frac{N}{DF(t)}\right) \quad (1)$$

dengan:

- $TF(t,d)$= frekuensi kata pada dokumen
- $DF(t)$= jumlah dokumen yang memuat kata
- N= jumlah seluruh dokumen

Uji Asumsi

Pada penelitian berbasis *machine learning*, uji normalitas dan homogenitas tidak diperlukan. Namun dilakukan pemeriksaan terhadap beberapa karakteristik data yang memengaruhi proses klasifikasi.

Distribusi Kelas

Distribusi data menunjukkan adanya ketidakseimbangan jumlah dokumen antar-genre. Rasio antara kelas terbesar (*thriller*) dan kelas terkecil (*sports*, *travel*, dan *psychology*) mencapai lebih dari 10:1.

Kesesuaian Data Teks

Seluruh dokumen memiliki atribut ringkasan yang dapat digunakan sebagai sumber informasi untuk proses klasifikasi. Selain itu, variasi panjang ringkasan yang cukup tinggi

menunjukkan keberagaman informasi yang dapat dimanfaatkan model untuk mempelajari pola masing-masing genre.

Representasi Fitur

Transformasi TF-IDF menghasilkan representasi data berdimensi tinggi yang sesuai dengan karakteristik algoritma Support Vector Machine (Joachims, 1998).

Uji Hipotesis

Hipotesis penelitian dirumuskan sebagai berikut:

H0: Ringkasan buku tidak mampu digunakan untuk mengidentifikasi genre buku secara efektif.

H1: Ringkasan buku mampu digunakan untuk mengidentifikasi genre buku secara efektif.

Pengujian dilakukan menggunakan algoritma Linear Support Vector Machine pada data uji sebanyak 932 dokumen.

Hasil Evaluasi Model

Evaluasi model dilakukan menggunakan metrik *Accuracy*, *Precision*, *Recall*, dan *F1-Score*.

Rumus Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (2)$$

Rumus Precision

$$Precision = \frac{TP}{TP + FP} \qquad (3)$$

Rumus Recall

$$Recall = \frac{TP}{TP + FN} \qquad (4)$$

Rumus F1-Score

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (5)$$

Hasil Evaluasi Keseluruhan

Tabel 2. Hasil Evaluasi Model TF-IDF + SVM

Metrik	Nilai
Accuracy	68,78%
Precision	69,26%
Recall	68,78%
F1-Score	68,52%

Berdasarkan Tabel 2, model berhasil mencapai tingkat akurasi sebesar **68,78%**. Nilai precision, recall, dan F1-score yang relatif seimbang menunjukkan bahwa model memiliki kemampuan yang cukup baik dalam mengenali pola genre berdasarkan ringkasan buku.

Dengan demikian, hipotesis **H1 diterima**, yaitu ringkasan buku mengandung informasi yang cukup untuk digunakan dalam proses identifikasi genre secara otomatis.

Evaluasi Per Genre

Tabel 3. Performa Klasifikasi Setiap Genre

Genre	Precision	Recall	F1-Score
Crime	0,78	0,66	0,71
Fantasy	0,76	0,78	0,77
History	0,70	0,72	0,71
Horror	0,58	0,60	0,59
Psychology	0,93	0,65	0,76
Romance	0,27	0,14	0,18
Science	0,74	0,68	0,71
Sports	0,73	0,80	0,76
Thriller	0,62	0,73	0,67
Travel	1,00	0,45	0,62

Hasil menunjukkan bahwa genre **Fantasy** memperoleh performa yang paling stabil dengan nilai F1-score sebesar 0,77. Genre **Sports** dan **Psychology** juga menunjukkan performa yang baik meskipun jumlah datanya relatif sedikit.

Sebaliknya, genre **Romance** memperoleh nilai F1-score terendah yaitu 0,18. Hasil tersebut menunjukkan bahwa karakteristik teks pada genre romance memiliki kemiripan dengan genre lain sehingga lebih sulit dibedakan oleh model.

Analisis Hasil

Hasil penelitian menunjukkan bahwa kombinasi TF-IDF dan SVM mampu menghasilkan performa klasifikasi yang cukup baik pada dataset buku berskala besar. Akurasi sebesar 68,78% mengindikasikan bahwa model berhasil mengidentifikasi lebih dari dua pertiga dokumen secara benar berdasarkan ringkasan yang tersedia.

Genre dengan karakteristik kosakata yang spesifik seperti *fantasy*, *sports*, dan *psychology* cenderung lebih mudah dikenali oleh model. Sebaliknya, genre yang memiliki kemiripan tema dan kosakata seperti *romance*, *thriller*, dan *crime* menghasilkan tingkat kesalahan klasifikasi yang lebih tinggi.

Temuan ini menunjukkan bahwa kualitas representasi teks memiliki peran penting dalam proses klasifikasi dokumen. Ringkasan buku terbukti mampu memberikan informasi yang cukup untuk membedakan genre, meskipun masih terdapat tantangan akibat ketidakseimbangan kelas dan kemiripan semantik antar-genre.

PEMBAHASAN

Hasil penelitian menunjukkan bahwa pendekatan *text mining* menggunakan representasi TF-IDF dan algoritma Support Vector Machine (SVM) mampu mengklasifikasikan genre buku berdasarkan ringkasan teks dengan tingkat akurasi sebesar 68,78%, *precision* sebesar 69,26%, *recall* sebesar 68,78%, dan *F1-score* sebesar 68,52%. Hasil tersebut menunjukkan bahwa informasi yang terkandung dalam ringkasan buku memiliki kemampuan yang cukup baik dalam merepresentasikan karakteristik genre sehingga dapat dimanfaatkan untuk proses klasifikasi otomatis. Temuan ini mengindikasikan bahwa ringkasan buku tidak hanya berfungsi sebagai deskripsi isi, tetapi juga mengandung pola linguistik yang dapat digunakan untuk membedakan kategori buku secara sistematis.

Keberhasilan model dalam mencapai tingkat akurasi mendekati 70% menunjukkan bahwa metode TF-IDF masih memiliki relevansi yang tinggi dalam pengolahan data teks berskala besar. Meskipun saat ini berkembang berbagai pendekatan berbasis *deep learning*, representasi TF-IDF tetap menjadi metode yang efektif karena mampu menghasilkan fitur yang mudah diinterpretasikan serta memiliki kebutuhan komputasi yang relatif lebih rendah. Menurut Kowsari et al. (2019), kombinasi TF-IDF dengan algoritma klasifikasi tradisional masih menjadi

pendekatan yang kompetitif pada berbagai kasus klasifikasi dokumen, khususnya ketika ukuran dataset tidak mencapai skala yang membutuhkan model jaringan saraf yang kompleks. Hasil penelitian ini memperkuat temuan tersebut dengan menunjukkan bahwa klasifikasi genre buku dapat dilakukan secara efektif menggunakan pendekatan yang relatif sederhana namun memiliki tingkat akurasi yang memadai.

Analisis lebih lanjut menunjukkan bahwa performa model tidak merata pada setiap kategori genre. Genre *Fantasy* memperoleh nilai *F1-score* sebesar 0,77, sedangkan genre *Sports* dan *Psychology* masing-masing memperoleh nilai *F1-score* sebesar 0,76. Tingginya performa pada ketiga genre tersebut menunjukkan bahwa dokumen dalam kategori tersebut memiliki karakteristik kosakata yang relatif spesifik dan berbeda dari genre lainnya. Pada genre *Fantasy*, misalnya, ringkasan buku umumnya memuat istilah yang berkaitan dengan dunia imajinatif, sihir, kerajaan, naga, atau makhluk mitologi. Kosakata yang khas tersebut menghasilkan pola yang lebih mudah dikenali oleh model klasifikasi sehingga meningkatkan ketepatan prediksi.

Temuan ini sejalan dengan penelitian Minaee et al. (2021) yang menjelaskan bahwa performa klasifikasi teks sangat dipengaruhi oleh tingkat keterpisahan semantik antar kategori. Semakin unik kosakata yang dimiliki suatu kategori, semakin mudah model membangun batas klasifikasi yang jelas. Dalam konteks penelitian ini, genre *Fantasy*, *Sports*, dan *Psychology* memiliki identitas linguistik yang relatif kuat sehingga menghasilkan performa klasifikasi yang lebih baik dibandingkan genre lainnya.

Sebaliknya, genre *Romance* menghasilkan nilai *F1-score* terendah yaitu 0,18. Rendahnya performa tersebut menunjukkan bahwa model mengalami kesulitan dalam membedakan genre *Romance* dari genre lain. Salah satu penyebab utama kondisi tersebut adalah jumlah data yang relatif kecil dibandingkan genre lainnya. Dari total 4.657 dokumen, hanya 111 dokumen yang termasuk dalam kategori *Romance*. Ketidakseimbangan jumlah data menyebabkan model memiliki kesempatan yang lebih sedikit untuk mempelajari karakteristik genre tersebut secara optimal.

Selain faktor jumlah data, rendahnya performa genre *Romance* juga dapat dijelaskan melalui kemiripan semantik dengan genre lain seperti *Drama*, *History*, maupun *Thriller* yang sering kali memuat unsur hubungan interpersonal dalam narasi cerita. Akibatnya, terdapat tumpang tindih kosakata yang menyebabkan model kesulitan membangun batas klasifikasi yang jelas. Fenomena ini sesuai dengan temuan Krawczyk (2016) yang menyatakan bahwa ketidakseimbangan kelas (*class imbalance*) merupakan salah satu faktor utama yang menurunkan performa klasifikasi pada kategori minoritas.

Keberadaan *class imbalance* juga terlihat dari distribusi dataset yang didominasi oleh genre *Thriller* (21,97%) dan *Fantasy* (18,81%), sedangkan beberapa genre lainnya memiliki proporsi

kurang dari 3%. Kondisi tersebut merupakan karakteristik umum dalam lingkungan *big data*, di mana distribusi data sering kali tidak merata. Meskipun demikian, hasil penelitian menunjukkan bahwa model tetap mampu menghasilkan performa yang cukup baik secara keseluruhan. Hal ini mengindikasikan bahwa penggunaan TF-IDF berhasil mengurangi sebagian dampak ketidakseimbangan data dengan memberikan bobot yang lebih tinggi pada kata-kata yang memiliki nilai diskriminatif terhadap suatu genre.

Temuan penelitian ini memiliki beberapa kesamaan dengan penelitian terdahulu, namun juga menunjukkan perbedaan yang menjadi kontribusi ilmiah tersendiri. Penelitian Kowsari et al. (2019) menemukan bahwa kombinasi TF-IDF dan algoritma klasifikasi tradisional mampu menghasilkan performa yang kompetitif pada berbagai jenis dokumen teks. Penelitian tersebut lebih banyak berfokus pada dokumen berita, ulasan produk, dan data media sosial. Sementara itu, penelitian ini menerapkan pendekatan yang sama pada ringkasan buku digital yang memiliki karakteristik teks naratif dan kompleksitas semantik yang berbeda. Hasil yang diperoleh menunjukkan bahwa pendekatan TF-IDF dan SVM tetap efektif meskipun diterapkan pada domain yang berbeda.

Perbedaan lain terlihat pada objek penelitian yang digunakan. Sebagian besar penelitian klasifikasi teks dalam satu dekade terakhir menggunakan dataset yang berorientasi pada analisis sentimen, deteksi spam, atau kategorisasi berita. Penelitian oleh Minaee et al. (2021), misalnya, menyoroti perkembangan metode klasifikasi teks berbasis *deep learning* yang banyak diterapkan pada data media sosial dan dokumen umum. Sebaliknya, penelitian ini menggunakan ringkasan buku sebagai sumber data utama. Karakteristik data tersebut memberikan tantangan yang berbeda karena ringkasan buku mengandung narasi yang lebih panjang dan kompleks dibandingkan teks pendek pada media sosial. Dengan demikian, penelitian ini memperluas ruang lingkup penerapan *text mining* pada domain perpustakaan digital dan pengelolaan koleksi buku elektronik.

Kontribusi utama penelitian ini terletak pada penggunaan dataset buku digital yang relatif besar dengan sepuluh kategori genre yang berbeda. Sebagian penelitian sebelumnya menggunakan jumlah kategori yang lebih sedikit atau dataset yang lebih terbatas. Penggunaan 4.657 dokumen memungkinkan identifikasi pola yang lebih beragam dan lebih mendekati kondisi nyata pada sistem perpustakaan digital maupun platform rekomendasi buku. Oleh karena itu, hasil penelitian ini memberikan bukti empiris bahwa pendekatan klasifikasi teks dapat diterapkan secara efektif dalam pengelolaan informasi berbasis buku digital.

Dari perspektif teori *Knowledge Discovery in Databases* (KDD), penelitian ini menunjukkan bahwa data teks yang awalnya tidak terstruktur dapat diubah menjadi pengetahuan yang bernilai melalui tahapan seleksi data, pembersihan data, transformasi fitur, dan analisis pola. Proses tersebut menghasilkan model yang mampu mengidentifikasi genre buku secara otomatis

berdasarkan ringkasan yang tersedia. Temuan ini mendukung pandangan bahwa *data mining* tidak hanya berfungsi sebagai alat analisis data, tetapi juga sebagai sarana untuk mengekstraksi pengetahuan baru dari kumpulan data yang besar dan kompleks.

Implikasi praktis dari penelitian ini cukup signifikan bagi pengelolaan perpustakaan digital dan sistem rekomendasi buku. Model klasifikasi yang dihasilkan dapat digunakan untuk membantu proses kategorisasi koleksi baru secara otomatis sehingga mengurangi ketergantungan pada proses manual. Selain itu, informasi genre yang diperoleh dari hasil klasifikasi dapat digunakan untuk meningkatkan akurasi sistem rekomendasi berbasis konten (*content-based recommendation system*). Dengan demikian, pengguna dapat memperoleh rekomendasi buku yang lebih relevan berdasarkan karakteristik isi buku, bukan hanya berdasarkan metadata yang tersedia.

Meskipun menghasilkan performa yang cukup baik, penelitian ini masih memiliki beberapa keterbatasan. Pertama, representasi fitur yang digunakan masih berbasis TF-IDF sehingga belum sepenuhnya mampu menangkap hubungan kontekstual antar kata. Kedua, distribusi data yang tidak seimbang menyebabkan performa klasifikasi pada beberapa genre minoritas masih relatif rendah. Ketiga, penelitian hanya menggunakan atribut ringkasan buku tanpa memanfaatkan informasi tambahan seperti judul, kata kunci, nama penulis, atau ulasan pembaca. Oleh karena itu, penelitian selanjutnya dapat mengembangkan model berbasis *word embedding* atau *transformer* seperti BERT untuk meningkatkan kemampuan model dalam memahami konteks semantik dokumen. Selain itu, penerapan teknik *oversampling* atau *class weighting* dapat digunakan untuk mengurangi dampak ketidakseimbangan data.

Secara keseluruhan, hasil penelitian menunjukkan bahwa kombinasi TF-IDF dan SVM mampu mengidentifikasi genre buku secara efektif berdasarkan ringkasan teks yang tersedia. Temuan ini tidak hanya mendukung hasil penelitian sebelumnya mengenai efektivitas *text mining* dalam klasifikasi dokumen, tetapi juga memberikan kontribusi baru melalui penerapan pada dataset buku digital berskala besar dengan variasi genre yang beragam. Dengan demikian, penelitian ini memperluas penerapan metode klasifikasi teks dalam bidang perpustakaan digital dan pengelolaan informasi serta membuka peluang pengembangan penelitian lanjutan yang memanfaatkan teknik kecerdasan buatan yang lebih mutakhir.

KESIMPULAN

Penelitian ini bertujuan untuk menganalisis kemampuan ringkasan buku sebagai sumber informasi tekstual dalam proses klasifikasi genre buku digital menggunakan pendekatan *text mining*. Hasil penelitian menunjukkan bahwa ringkasan buku memiliki karakteristik linguistik yang cukup representatif untuk digunakan dalam proses identifikasi genre secara otomatis.

Penerapan metode *Term Frequency-Inverse Document Frequency* dan *Support Vector Machine* mampu membangun model klasifikasi yang efektif dalam mengenali pola antar-genre pada dataset buku digital yang terdiri atas 4.657 dokumen dan sepuluh kategori genre. Temuan penelitian menunjukkan bahwa keberhasilan klasifikasi sangat dipengaruhi oleh karakteristik kosakata setiap genre serta distribusi data pada masing-masing kategori. Genre yang memiliki kosakata lebih spesifik dan konsisten cenderung menghasilkan performa klasifikasi yang lebih baik dibandingkan genre yang memiliki kemiripan semantik dengan kategori lain. Penelitian ini memberikan kontribusi empiris bahwa pendekatan *text mining* berbasis TF-IDF dan SVM masih relevan untuk mengolah data teks berskala besar serta dapat dimanfaatkan sebagai alternatif dalam proses pengelompokan koleksi buku digital secara otomatis. Selain itu, penelitian ini memperluas penerapan klasifikasi teks pada domain perpustakaan digital dan pengelolaan informasi berbasis buku elektronik.

REFERENSI

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. Proceedings of NAACL-HLT. https://doi.org/10.48550/arXiv.1810.04805

Joachims, T. (1998). *Text categorization with support vector machines: Learning with many relevant features*. European Conference on Machine Learning, 137–142. https://doi.org/10.1007/BFb0026683

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L. E., & Brown, D. E. (2019). *Text classification algorithms: A survey*. Information, 10(4), 150. https://doi.org/10.3390/info10040150

Krawczyk, B. (2016). *Learning from imbalanced data: Open challenges and future directions*. Progress in Artificial Intelligence, 5(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). *Deep learning-based text classification: A comprehensive review*. ACM Computing Surveys, 54(3), 1–40. https://doi.org/10.1145/3439726

Ramos, J. (2003). *Using TF-IDF to determine word relevance in document queries*. Proceedings of the First Instructional Conference on Machine Learning. https://doi.org/10.48550/arXiv.cs/0306050

Salton, Gerard, & Buckley, C. (1988). *Term-weighting approaches in automatic text retrieval*. Information Processing & Management, 24(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). *How to fine-tune BERT for text classification?* China National Conference on Chinese Computational Linguistics. https://doi.org/10.1007/978-3-030-32381-3_16

Zhang, X., Zhao, J., & LeCun, Y. (2015). *Character-level convolutional networks for text classification*. Advances in Neural Information Processing Systems, 28. https://doi.org/10.48550/arXiv.1509.01626