IMPLEMENTASI MACHINE LEARNING UNTUK KLASIFIKASI BUKU OTOMATIS PADA PERPUSTAKAAN DIGITAL

Berta Erwin SLAM[1], Feri Irawan[2], Nolan Efranda[3], Rifaldi Herikson[4]
[1,2,3,4] Teknik Informatika, Fakultas Teknik dan Teknologi Kemaritiman, Universitas Maritim Raja Ali Haji, Indonesia
[1]bertaerwinslam@umrah.ac.id, [2]feriirawan@umrah.ac.id, [3]nolanefranda@umrah.ac.id, [4]rifaldiherikson@umrah.ac.id

Abstrak

Permasalahan klasifikasi buku dalam sistem perpustakaan digital, khususnya di tingkat sekolah menengah atas (SMA), masih menjadi tantangan karena banyak institusi belum mengadopsi sistem klasifikasi otomatis. Proses manual dinilai tidak efisien dan rawan inkonsistensi. Penelitian ini bertujuan mengembangkan sistem klasifikasi otomatis berbasis machine learning menggunakan algoritma *Naïve Bayes*, yang dikenal efektif dalam pengolahan teks. Data yang digunakan terdiri dari 10.000 entri buku digital, yang masing-masing mencakup metadata berupa judul, sinopsis, dan kata kunci. Proses preprocessing dilakukan melalui normalisasi teks, penghapusan stopword bahasa Indonesia, serta transformasi ke dalam representasi vektor menggunakan metode TF-IDF. Model dilatih untuk mengenali sepuluh kategori utama dengan berbagai rasio pembagian data latih dan uji, mulai dari 90:10 hingga 50:50. Hasil evaluasi menunjukkan bahwa model mampu menghasilkan akurasi tinggi di berbagai skenario, dengan rentang akurasi antara 89,2% hingga 90,3%. Menariknya, performa model justru meningkat secara konsisten seiring meningkatnya proporsi data uji. Precision dan recall makro juga menunjukkan tren serupa, yang menandakan bahwa model Naïve Bayes cukup robust bahkan saat data latih terbatas. Secara keseluruhan, sistem ini terbukti efektif dalam meningkatkan efisiensi dan konsistensi klasifikasi koleksi perpustakaan digital. Temuan ini merekomendasikan integrasi sistem klasifikasi otomatis ke dalam platform perpustakaan SMA, serta membuka peluang eksplorasi algoritma lanjutan dan pengembangan fitur rekomendasi cerdas di masa depan.

Kata kunci : Perpustakaan Digital, Klasifikasi Buku, Machine Learning, Naive Bayes

1. Pendahuluan

Klasifikasi buku secara otomatis dalam sistem perpustakaan digital, khususnya di tingkat sekolah menengah atas (SMA), masih menghadapi berbagai tantangan. Meskipun digitalisasi koleksi semakin meningkat, mayoritas sekolah belum mengadopsi sistem klasifikasi cerdas yang mampu mengelompokkan buku secara akurat dan efisien. Proses klasifikasi manual tidak hanya memakan waktu tetapi juga rentan terhadap ketidakkonsistenan, sehingga menghambat optimalisasi layanan informasi perpustakaan (Hadiwibowo & Rahani, 2022; Wilian & Sriyanto, 2025). Untuk menjawab tantangan tersebut, pengembangan sistem berbasis machine learning telah menjadi perhatian utama dalam penelitian terkini (Palanivinayagam et al., 2023; Wang et al., 2023).

Salah satu metode populer dalam pengolahan data teks adalah algoritma *Naive Bayes*, yang telah terbukti efektif dan efisien dalam berbagai kasus klasifikasi dokumen (Nugroho et al., 2020; Ogundeji et al., 2022). Implementasi *Naive Bayes* pada sistem perpustakaan menunjukkan hasil yang menjanjikan dalam klasifikasi koleksi berbasis metadata seperti judul, sinopsis, dan kata kunci (Lestari et al., 2023; Murlena & Syahindra, 2024). Optimalisasi *Naive Bayes* dengan teknik seleksi fitur seperti *information gain* dapat meningkatkan akurasi klasifikasi koleksi perpustakaan (Mulyani et al., 2025).

Di sisi lain, beberapa pendekatan alternatif seperti *Support Vector Machine* (SVM), TF-IDF, dan Word2Vec juga telah diterapkan dalam konteks yang serupa (Cahyani & Saraswati, 2023), meskipun memiliki kompleksitas komputasi yang lebih tinggi dibandingkan *Naive Bayes*. Selain itu, pendekatan klasifikasi lain seperti algoritma C4.5 dan *K-Nearest Neighbor* (K-NN) juga telah dibandingkan untuk menilai efektivitasnya dalam lingkungan perpustakaan (Blanquero et al., 2021; Dasuki et al., 2024),namun *Naive Bayes* tetap unggul dalam hal efisiensi dan skalabilitas untuk data teks sederhana.

Dalam kerangka peningkatan efisiensi perpustakaan sekolah, penerapan klasifikasi otomatis berbasis web juga telah dikembangkan (Wulandari et al., 2021). Seiring berkembangnya teknologi pengenalan citra dan pembelajaran mendalam, pendekatan berbasis CNN (*Convolutional Neural Network*) pun telah dicoba untuk klasifikasi koleksi buku berbasis visual (Hu et al., 2024), namun belum banyak diterapkan dalam konteks metadata teks. Di sisi lain, riset dalam bidang digital humanities menyoroti pentingnya kompetensi profesional tenaga pustakawan dalam memanfaatkan teknologi digital untuk mendukung proses klasifikasi dan layanan informasi (Madhav & Muthumari, 2021).

Lebih jauh, pengembangan *framework* adaptif berbasis multi-label dan multi-modal *classification* telah menjadi fokus penelitian terkini, meskipun belum banyak diterapkan pada lingkungan

pendidikan menengah (Nareti et al., 2025). Pendekatan-pendekatan ini memberikan peluang besar dalam memperluas cakupan klasifikasi genre buku, namun membutuhkan dukungan data dan infrastruktur yang lebih kompleks.

Urgensi dari penelitian ini terletak pada kebutuhan mendesak akan sistem klasifikasi otomatis yang dapat diimplementasikan secara praktis dan efisien di tingkat SMA, tanpa memerlukan sumber daya teknologi yang kompleks. Mengingat keterbatasan sumber daya pustakawan dan sistem di banyak sekolah, penerapan solusi berbasis machine learning seperti Naive Bayes dapat menjadi langkah strategis untuk meningkatkan kualitas pengelolaan koleksi dan layanan informasi. Dengan memanfaatkan metadata yang sudah tersedia, institusi pendidikan dapat mempercepat proses klasifikasi dan meningkatkan akurasi pengelompokan buku sesuai kebutuhan kurikulum dan minat baca siswa.

Penelitian ini bertujuan untuk mengembangkan sistem klasifikasi otomatis buku perpustakaan SMA berbasis algoritma *Naive Bayes*. Sistem akan dilatih menggunakan data metadata buku dan diuji untuk mengklasifikasikan koleksi ke dalam kategori subjek secara otomatis. Diharapkan pendekatan ini mampu meningkatkan efisiensi, akurasi, dan kualitas layanan informasi dalam sistem perpustakaan digital sekolah.

2. Metode

Metode penelitian ini dirancang secara sistematis untuk menggambarkan tahapan-tahapan dalam mengembangkan sistem klasifikasi otomatis koleksi buku berbasis algoritma *machine learning*, khususnya *Naive Bayes*, dalam konteks perpustakaan digital. Alur metodologi secara umum digambarkan dalam Gambar 1.

Tahapan awal dimulai dengan proses identifikasi dan pengumpulan data. Data dikumpulkan melalui observasi langsung terhadap sistem pengelolaan koleksi di perpustakaan sekolah serta wawancara dengan pustakawan. Tujuan dari tahapan ini adalah untuk memahami struktur metadata buku yang tersedia serta sistem klasifikasi subjek yang digunakan secara lokal.

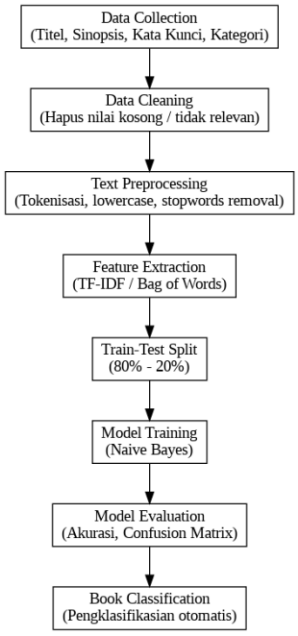
Sumber data (*data source*) berupa koleksi digital metadata buku diperoleh dari sistem manajemen perpustakaan internal. Dataset terdiri atas 10.000 entri buku, yang mencakup informasi judul, sinopsis, kata kunci (*keyword*), dan label kategori subjek. Jumlah kategori klasifikasi (*class*) ditetapkan sebanyak sepuluh (10) kelas utama, Bahasa dan Sastra, Ilmu Pengetahuan Alam, Ilmu Sosial, Teknologi Informasi, Pendidikan Agama, Matematika, Kesehatan dan Olahraga, Seni dan Budaya, Bahasa Asing, dan Geografi.

Fitur (*features*) utama yang digunakan untuk pelatihan model meliputi tiga komponen teks, yaitu judul, sinopsis, dan kata kunci. Ketiga elemen ini diproses melalui teknik pemrosesan bahasa alami (*Natural Language Processing*), termasuk pembersihan teks, tokenisasi, penghapusan *stopwords*, dan normalisasi. Selanjutnya, representasi numerik dari teks dilakukan melalui metode TF-IDF (*Term Frequency–Inverse Document Frequency*) untuk menghasilkan vektor fitur yang dapat dibaca oleh model klasifikasi.

Proses data cleaning dilakukan untuk menghilangkan atribut redundan seperti ISBN, lokasi rak, atau nomor katalog, yang tidak berkontribusi terhadap klasifikasi isi. Setelah data dibersihkan, dilakukan pembagian data (data split) menjadi data latih (training data) dan data uji (testing data) dengan variasi rasio berbeda, mulai dari 90:10 hingga 50:50, guna mengukur stabilitas performa model terhadap jumlah data pelatihan.

Model *Naive Bayes* kemudian dilatih menggunakan data latih tersebut. Model ini menghitung probabilitas *posterior* dari setiap fitur terhadap seluruh kelas yang ada, berdasarkan prinsip probabilistik yang dimiliki oleh *Naive Bayes*. Estimasi dilakukan dengan mengasumsikan independensi antar fitur, sehingga setiap komponen (judul, sinopsis, kata kunci) dihitung kontribusinya secara terpisah dalam menentukan kemungkinan sebuah buku termasuk ke dalam suatu kategori.

Tahapan akhir adalah pengujian model menggunakan data uji. Pengujian dilakukan untuk mengevaluasi akurasi, presisi, recall, dan f1-score dari hasil klasifikasi. Evaluasi performa ini disajikan dalam bentuk *confusion matrix* dan metrik evaluasi makro. Hasil dari pengujian ini menjadi dasar untuk menilai efektivitas model serta menentukan potensi integrasi sistem ke dalam platform perpustakaan digital sekolah.



Gambar. 1. Flowchart Metode Penelitian.

2.1 *Naïve Bayes*

Naïve Bayes adalah salah satu metode klasifikasi yang didasarkan pada teori probabilitas

Bayes. Metode ini digunakan untuk mengklasifikasikan data dengan cara menghitung probabilitas posterior suatu kelas berdasarkan atribut-atribut yang dimiliki oleh data tersebut (Han et al., 2011). Naïve Bayes termasuk dalam kelompok classifier probabilistik, yang memberikan probabilitas untuk setiap kelas sehingga dapat memilih kelas dengan probabilitas tertinggi sebagai hasil klasifikasi.

Dalam konteks *Naïve Bayes*, misalkan terdapat sebuah dataset pelatihan D yang berisi sejumlah tuple, di mana setiap *tuple* memiliki label kelas. Setiap *tuple* direpresentasikan sebagai sebuah vektor atribut berdimensi *n*, yaitu X= ($x1$, $x2$,..., xn) , yang menggambarkan nilai atribut-atribut $A1$, $A2$,..., An. Dataset memiliki m kelas berbeda yang dapat menjadi label dari *tuple* tersebut, yaitu $C1$, $C2$,..., Cm. *Naïve Bayes* bekerja dengan memilih kelas Ci yang memaksimalkan probabilitas posterior $P(C_i|X)$, yakni probabilitas kelas C_i ketika data X diberikan seperti pada persamaan 1.

$P(C_i|X) > P(C_j|X)$ untuk semua j≠i, 1≤j≤m (1)

Probabilitas posterior ini dihitung menggunakan *Teorema Bayes*, yang dapat dilihat pada persamaan 2.

$$P(C_i|X) = \frac{P(X|C_i) \times P(C_i)}{P(X)} \quad (2)$$

Di mana:

- $P(C_i|X)$ adalah probabilitas posterior kelas C_i berdasarkan data X,
- $P(X|C_i)$ adalah likelihood data X apabila berada pada kelas C_i ,
- $P(C_i)$ adalah probabilitas prior kelas C_i, yaitu probabilitas kelas sebelum melihat data,
- $P(X)$ adalah probabilitas total data X.

Karena $P(X)$ adalah konstan untuk semua kelas, maka proses klasifikasi hanya perlu memaksimalkan nilai, seperti ditunjukkan pada Persamaan (3).

$$P(X|C_i) \times P(C_i) \quad (3)$$

Prior $P(C_i)$ biasanya dapat dihitung dari proporsi kelas dalam data pelatihan, seperti ditunjukkan pada Persamaan (4)

$$P(C_i) = \frac{|C_i,D|}{|D|} \quad (4)$$

di mana $|C_i, D|$ adalah jumlah *tuple* dalam kelas C_i, dan $|D|$ adalah total jumlah tuple dalam dataset.

Perhitungan langsung $P(X|C_i)$ akan sangat kompleks jika X memiliki banyak atribut, karena harus memperhitungkan kemungkinan seluruh kombinasi nilai atribut. Oleh karena itu, *Naïve Bayes* menggunakan asumsi independensi kondisional antar atribut, yaitu setiap atribut x_k dianggap independen satu sama lain apabila sudah diketahui kelasnya. Dengan asumsi ini, probabilitas $P(X|C_i)$ dapat dihitung sebagai hasil perkalian probabilitas atribut secara terpisah, seperti ditunjukkan pada Persamaan (5).

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \quad (5)$$

Probabilitas atribut $P(x_k|C_i)$ dihitung berdasarkan tipe atribut tersebut:

- **Atribut kategorikal**: Probabilitas dihitung sebagai frekuensi nilai atribut x_k pada kelas C_i, yaitu jumlah *tuple* dalam kelas C_i yang memiliki nilai x_k pada atribut ke-k, dibagi dengan total jumlah *tuple* dalam kelas C_i .
- **Atribut kontinu (numerik)**: Biasanya diasumsikan mengikuti distribusi *Gaussian* (normal), sehingga probabilitas dihitung menggunakan fungsi densitas probabilitas *Gaussian*, sebagaimana ditunjukkan pada Persamaan (6).

$$P(x_k|C_i) = \frac{1}{\sqrt{2\pi u_i}} \exp(-\frac{(x_k - \mu C_i)^2}{2u_i} \quad (6)$$

Di mana

- x_k = nilai dari atribut ke-k
- C_i = kelas ke-i
- μC_i = rata-rata (mean) dari atribut ke-k untuk kelas C_i
- u_i = varians dari atribut ke-k untuk kelas C_i
- exp = fungsi eksponensial (e^x)
- π = konstanta pi (≈ 3.1416)

Meskipun asumsi independensi antar atribut sering kali tidak realistis dalam banyak kasus nyata, Naïve Bayes tetap menunjukkan performa yang baik dan tahan terhadap data dengan banyak atribut. Hal ini dikarenakan, meskipun nilai probabilitas yang dihitung bisa jadi kurang akurat, ranking relatif antar kelas tetap cukup tepat untuk menghasilkan klasifikasi yang benar dalam banyak kasus (Dietterich et al., 2022)

Selain itu, *Naïve Bayes* memiliki keunggulan lain, yaitu kemampuannya untuk menangani data dengan jumlah atribut yang besar dan proses pelatihan yang cepat. Ini menjadikannya sangat populer di berbagai bidang, seperti pengolahan bahasa alami, deteksi spam, dan sistem rekomendasi (Russell et al., 2021)

Secara teoritis, *Naïve Bayes* merupakan classifier dengan tingkat kesalahan minimum jika asumsi independensi benar-benar terpenuhi. Namun, dalam praktik, akurasi dapat menurun apabila asumsi ini tidak sesuai. Meski begitu, berbagai penelitian empiris menunjukkan bahwa *Naïve Bayes* tetap memberikan hasil yang kompetitif dibandingkan metode klasifikasi lain seperti pohon keputusan dan jaringan syaraf tiruan (Han et al., 2011).

3. Hasil dan Pembahasan

Penelitian ini bertujuan untuk mengembangkan sistem klasifikasi otomatis buku perpustakaan digital tingkat sekolah menengah atas (SMA) menggunakan pendekatan *machine learning* berbasis algoritma *Naïve Bayes*. Peroses dimulai dengan proses

identifikasi dan pengumpulan data. Data dikumpulkan melalui observasi langsung terhadap sistem pengelolaan koleksi di perpustakaan sekolah serta wawancara dengan pustakawan sehingga menadapatkan *data collection*, yang mencakup 10.000 entri buku dengan atribut teks berupa judul, sinopsis, dan kata kunci, serta satu label kategori subjek buku seperti "Bahasa dan Sastra", "Teknologi Informasi", "Seni dan Budaya", dan sebagainya.

Pada tahap kedua yaitu *data cleaning*, peneliti menghapus data duplikat dan mengeliminasi entri yang tidak lengkap atau kosong. Setelah itu, dilakukan tahap text *preprocessing* meliputi tokenisasi, lowercasing, penghapusan stopwords Bahasa Indonesia.

Tahap berikutnya *feature extraction* yaitu teks diubah menjadi representasi vektor numerik menggunakan metode TF-IDF (*Term Frequency-Inverse Document Frequency*). Alternatif seperti *Bag of Words* dapat digunakan, namun TF-IDF dipilih untuk mempertimbangkan bobot pentingnya kata dalam konteks keseluruhan dokumen.

Setelah dilakukan *feature extraction*, data kemudian dibagi ke dalam dua bagian, yaitu data training sebanyak 80% (8.000 record) dan data testing sebanyak 20% (2.000 record). Proses pembagian dilakukan secara acak namun merata terhadap distribusi kelas untuk memastikan keseimbangan data. Setelah itu data dimasukan ke dalam *model training*, model ini dibangun menggunakan algoritma *Naïve Bayes* yang diterapkan pada data hasil transformasi TF-IDF. Model kemudian diuji dengan data testing dan dievaluasi menggunakan metrik akurasi, *precision*, *recall*, *f1-score*, serta confusion matrix.

Hasil evaluasi menunjukkan bahwa model menghasilkan akurasi sebesar 0.892 atau 89,2% yang dapat dilihat pada Gambar 2. Nilai ini menunjukkan performa klasifikasi yang baik dalam mengelompokkan buku ke dalam kategori yang sesuai. Gambar 2 juga menyajikan nilai metrik evaluasi untuk masing-masing kelas target.

```
Akurasi: 0.892

Confusion Matrix:
 [[169   2   3   1   2   6   5   1   2   2]
 [  1 161   4   3   3   3   3   0   2   3]
 [  2   2 179   0   1   4   1   2   2   3]
 [  1   4   3 181   0   2   2   2   4   2]
 [  4   2   3   5 176   0   0   6   5   2]
 [  0   2   2   3   2 194   2   1   4   4]
 [  1   2   3   4   1   3 193   1   2   1]
 [  2   3   1   4   2   1   1 156   1   1]
 [  3   5   1   3   2   5   1   5 201   3]
 [  1   2   3   1   3   3   4   5   2 174]]
```

Gambar. 2. Hasil Akurasi dan Confusion Matrix.

```
Classification Report:
                     precision    recall  f1-score   support

       Bahasa Asing       0.92      0.88      0.90       193
Bahasa dan Sastra         0.87      0.88      0.88       183
           Ekonomi        0.89      0.91      0.90       196
          Geografi        0.88      0.90      0.89       201
       Ilmu Sosial        0.92      0.87      0.89       203
        Matematika        0.88      0.91      0.89       214
Pendidikan Agama          0.91      0.91      0.91       211
            Sains         0.87      0.91      0.89       172
   Seni dan Budaya        0.89      0.88      0.89       229
Teknologi Informasi       0.89      0.88      0.89       198

          accuracy                            0.89      2000
         macro avg        0.89      0.89      0.89      2000
      weighted avg        0.89      0.89      0.89      2000
```

Gambar. 3. Hasil Classification Report.

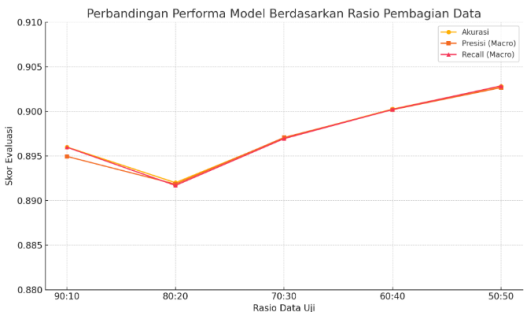
Hasil klasifikasi yang diperoleh dari penerapan algoritma *Naïve Bayes*, yang dapat dilihat pada Gambar 3 menunjukkan bahwa model mampu mencapai akurasi sebesar 89%. Nilai ini menunjukkan bahwa sistem klasifikasi otomatis berbasis pembelajaran mesin cukup efektif dalam mengelompokkan buku ke dalam kategori subjek yang sesuai berdasarkan informasi metadata seperti judul, sinopsis, dan kata kunci. Tingkat akurasi yang tinggi ini mengindikasikan bahwa mayoritas buku berhasil diklasifikasikan secara benar ke dalam kategori yang telah ditentukan, dan ini menjadi indikator awal bahwa sistem dapat diandalkan untuk mendukung operasional perpustakaan digital di tingkat SMA.

Berdasarkan *confusion matrix* pada gambar 2 dan *classification report* pada gambar 3 yang dihasilkan, performa klasifikasi cenderung stabil dan merata pada sebagian besar kelas. Nilai *precision*, *recall*, dan *f1-score* untuk seluruh kelas berkisar antara 0.87 hingga 0.92. Kelas dengan performa terbaik adalah "Pendidikan Agama" dengan nilai *f1-score* 0.91, sedangkan kelas "Bahasa dan Sastra" menunjukkan nilai yang sedikit lebih rendah namun masih dalam kategori baik, yaitu 0.88. Hal ini menunjukkan bahwa algoritma *Naïve Bayes* mampu melakukan klasifikasi dengan baik meskipun atribut yang digunakan berupa teks bebas yang memiliki keragaman tinggi secara semantik.

Untuk mengetahui pengaruh rasio pembagian data terhadap performa model, dilakukan eksperimen tambahan dengan mencoba berbagai rasio split data antara data latih dan data uji, yaitu: 90:10, 80:20, 70:30, 60:40, dan 50:50. Model tetap menggunakan algoritma *Naïve Bayes* dan metode vektorisasi TF-IDF yang sama, serta dilakukan pengukuran terhadap metrik akurasi, *precision* (*macro*), dan *recall* (*macro*).

Tabel 1. Hasil Evaluasi Berbagai Rasio Split Data

Rasio Data Uji	Akurasi	Presisi (Macro)	Recall (Macro)
(90:10)	0.89600	0.894946	0.895973
(80:20)	0.89200	0.891839	0.891691
(70:30)	0.89700	0.897061	0.896935
(60:40)	0.90025	0.900180	0.900205
(50:50)	0.90280	0.902657	0.902848



Gambar. 4. Grafik Perbandingan Performa Model Berdasarkan Rasio Pembagian Data

Model menunjukkan performa yang stabil di berbagai rasio pembagian data yang dapat dilihat

pada table 1, dengan akurasi berkisar antara 89,2% hingga 90,3%. Menariknya, akurasi, *precision*, dan *recall* justru cenderung meningkat, yang dapat dilihat pada gambar 4 seiring bertambahnya proporsi data uji hingga 50%. Temuan ini menunjukkan bahwa algoritma *Naïve Bayes* cukup robust meskipun jumlah data latih berkurang, serta tetap efektif dalam menangani representasi teks yang spars seperti TF-IDF. Hasil penelitian ini sejalan dengan penelitian terdahulu seperti penelitian seperti yang dilakukan oleh Hadiwibowo & Rahani, 2022juga menghasilkan performa baik dari *Naïve Bayes* untuk klasifikasi buku perpustakaan, dengan akurasi di atas 85%. Demikian pula, Murlena & Syahindra, 2024menunjukkan efektivitas algoritma ini dalam mengklasifikasikan minat baca pengunjung perpustakaan. Namun, perbandingan dengan studi lain seperti Cahyani & Saraswati, 2023yang menggunakan kombinasi SVM dengan TF-IDF dan Word2Vec menghasilkan performa yang lebih tinggi (di atas 92%), menunjukkan bahwa metode lain mungkin lebih unggul dalam menangani kompleksitas semantik pada teks panjang atau sinopsis. Hal ini juga didukung oleh Saputra et al., 2025, yang menunjukkan bahwa SVM mampu mengungguli Naïve Bayes secara signifikan dalam klasifikasi opini publik terhadap kendaraan listrik, dengan selisih akurasi yang cukup besar, meskipun kedua algoritma sama-sama bekerja pada data teks. Selain itu, Mulyani et al., 2025menggabungkan Naïve Bayes dengan teknik seleksi fitur seperti *Information Gain* untuk meningkatkan presisi klasifikasi buku, mengindikasikan bahwa penggabungan metode dapat memberi hasil lebih optimal. Dengan demikian, secara umum hasil penelitian yang didapat konsisten dengan literatur yang ada, namun dapat ditingkatkan melalui teknik lanjutan seperti kombinasi metode atau optimasi fitur.

Kelebihan utama dari metode *Naïve Bayes* dalam konteks ini adalah kemampuannya menangani data dengan jumlah atribut (fitur) yang besar dan sparsity yang tinggi, yang umum terjadi dalam representasi teks seperti TF-IDF. Meskipun metode ini didasarkan pada asumsi independensi antar fitur yang dalam kenyataan sering kali tidak terpenuhi dalam data teks hasil klasifikasi tetap menunjukkan kinerja yang baik. Ini disebabkan oleh kemampuan *Naïve Bayes* dalam mengevaluasi distribusi probabilistik dari fitur-fitur individu secara efisien, sehingga tetap mampu memisahkan kelas-kelas dengan cukup jelas.

Kesalahan klasifikasi (misclassification) yang terjadi dalam model umumnya muncul pada kelas-kelas yang memiliki keterkaitan tematis atau tumpang tindih konteks, seperti antara "Seni dan Budaya" dengan "Bahasa Asing", atau "Ilmu Sosial" dengan "Geografi". Hal ini wajar terjadi mengingat sinopsis atau kata kunci yang digunakan oleh buku-buku dalam kategori tersebut sering kali memiliki kemiripan leksikal ataupun semantik, yang menyulitkan model untuk melakukan pemisahan kelas secara eksplisit. Selain itu, ada kemungkinan bahwa beberapa buku memiliki konten yang relevan dengan lebih dari satu kategori, tetapi sistem klasifikasi berbasis supervised learning hanya mengizinkan satu label untuk setiap instance.

Model juga menunjukkan performa yang cukup stabil terhadap kelas-kelas yang memiliki jumlah data yang seimbang. Hal ini penting, karena model pembelajaran mesin umumnya lebih sensitif terhadap data yang tidak seimbang (class imbalance). Dalam penelitian ini, distribusi data pada tiap kelas dijaga agar seimbang saat pembagian data training dan testing, sehingga hasil evaluasi dapat mencerminkan kemampuan generalisasi model secara adil untuk semua kategori.

4. Kesimpulan

Penelitian ini menunjukkan bahwa algoritma *Naive Bayes* dapat digunakan secara efektif untuk klasifikasi otomatis buku dalam sistem perpustakaan digital, dengan tingkat akurasi berkisar antara 89,2% hingga 90,3% berdasarkan 10.000 data yang mencakup judul, sinopsis, dan kata kunci. Evaluasi menggunakan *confusion matrix* dan *classification report* mengindikasikan performa yang stabil pada berbagai kategori subjek, meskipun masih terdapat beberapa kesalahan klasifikasi.

Keterbatasan utama dari penelitian ini adalah penggunaan representasi teks yang sederhana dan asumsi independensi antar fitur, yang tidak selalu sesuai dalam konteks data teks. Selain itu, pengujian hanya dilakukan pada satu domain data, sehingga generalisasi model belum sepenuhnya terverifikasi.

Sebagai arah penelitian selanjutnya, disarankan untuk mengintegrasikan teknik ekstraksi fitur berbasis *word embeddings* atau metode *deep learning* untuk meningkatkan representasi semantik teks. Penggunaan algoritma klasifikasi lain seperti *Support Vector Machine, Random Forest*, atau *ensemble learning* juga direkomendasikan untuk membandingkan dan mengoptimalkan performa sistem klasifikasi.

Daftar Pustaka:

Blanquero, R., Carrizosa, E., Ramírez-Cobo, P., & Sillero-Denamiel, M. R. (2021). Variable selection for Naïve Bayes classification. *Computers and Operations Research*, *135*. https://doi.org/10.1016/j.cor.2021.105456

Cahyani, S. N., & Saraswati, G. W. (2023). Implementation of Support Vector Machine Method in Classifying School Library Books With Combination of TF-IDF and Word2vec. *Jurnal Teknik Informatika (Jutif)*, *4*(6), 1555–1566. https://doi.org/10.52436/1.jutif.2023.4.6.1536

Dasuki, A. U., Kom, S., & Kom, M. (2024). *Perbandingan Algoritma Naive Bayes Classifier dengan K-Nearest Neighbor (K-NN)*

Pada Ulasan Aplikasi Youtube Di PlayStore. *9*, 2024.

Dietterich, T., Bishop, C., Heckerman, D., Jordan, M., & Kearns, M. (2022). *Adaptive Computation and Machine Learning*. https://lccn.loc.gov/2021027430

Hadiwibowo, M. I., & Rahani, F. F. (2022). Data Mining Dalam Penentuan Pemesanan Buku Perpustakaan UAD dengan Menggunakan Metode Naïve Bayes. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *6*(4), 2165. https://doi.org/10.30865/mib.v6i4.4381

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)*.

Hu, J., Yan, Y., & Xie, Z. (2024). Automatic Recognition Technology of Library Books Based on Convolutional Neural Network Model. *HighTech and Innovation Journal*, *5*(1), 200–212. https://doi.org/10.28991/HIJ-2024-05-01-015

Lestari, N., Riza, O. S., & Ardinal, R. (2023). *Implementation Of Text Mining And Pattern Discovery With Naive Bayes Algorithm For Classification Of Text Documents*. https://doi.org/https://doi.org/10.31849/digitalzone.v14i1.13596

Madhav, A., & Muthumari, P. (2021). *Digital Humanities in India: A Developing Country Perspective Use Of Social Networking Sites (SNS) by Lis Professionals To Build Professional Competency-A Study*.

Mulyani, E., Sari, M., Ishlakhuddin, F., Teknik, J., Politeknik, I., & Indramayu, N. (2025). Multinomial Naïve Bayes Optimization with Information Gain for Library Book Classification. In *TeknoIS: Jurnal Ilmiah Teknologi Informasi dan Sains* (Vol. 15).

Murlena, M., & Syahindra, W. (2024). Application of the Naïve Bayes Algorithm in Classifying the Reading Interests of Regional Library Visitors. *Knowbase : International Journal of Knowledge in Database*, *4*(1), 94–105. https://doi.org/10.30983/knowbase.v4i1.8680

Nareti, U. K., Chattopadhyay, S., Mallick, P., Kumar, S., Daga, A. V., Adak, C., Wase, A., & Roy, A. (2025). *An Adaptive Data-Resilient Multi-Modal Framework for Hierarchical Multi-Label Book Genre Identification*. http://arxiv.org/abs/2505.03839

Nugroho, K. S., Istiadi, I., & Marisa, F. (2020). Naive Bayes classifier optimization for text classification on e-government using particle swarm optimization. *Jurnal Teknologi Dan Sistem Komputer*, *8*(1), 21–26. https://doi.org/10.14710/jtsiskom.8.1.2020.21-26

Ogundeji, R., Akinyemi, J., & Tijani, M. (2022). Naïve Bayes Algorithm for Document Classification. In *Annals of Mathematics and Computer Science* (Vol.7). https://www.researchgate.net/publication/372891724

Palanivinayagam, A., El-Bayeh, C. Z., & Damaševičius, R. (2023). Twenty Years of Machine-Learning-Based Text Classification: A Systematic Review. *Algorithms*, *16*(5). https://doi.org/10.3390/a16050236

Russell, S. J. ., Norvig, Peter., Chang, M.-Wei., Devlin, Jacob., Dragan, Anca., Forsyth, David., Goodfellow, Ian., Malik, Jitendra., Mansinghka, Vikash., Pearl, Judea., & Wooldridge, M. J. . (2021). *Artificial intelligence : a modern approach*. Pearson.

Saputra, D., Wantoro, A., Damayanti, & Rusliyawati. (2025). *Perbandingan Metode Naïve Bayes dan Support Vector Machine (SVM) pada Analisis Sentimen Kendaraan Listrik Pada Media Sosial "X."*

Wang, D., Tan, B., Wei, M., Cui, X., & Huang, X. (2023). Using natural language processing and machine learning algorithm for book categorization. *Applied and Computational Engineering*, *2*(1), 856–867. https://doi.org/10.54254/2755-2721/2/20220551

Wilian, D., & Sriyanto, S. (2025). Comparison of the Performance of the C.45 Algorithm with Naive Bayes in Analyzing Book Borrowing at the Library Pringsewu Muhammadiyah University. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, *14*(1), 101–106. https://doi.org/10.32736/sisfokom.v14i1.2300

Wulandari, M. S., Noveandini, R., & Nugroho, D. F. (2021). Optimalisasi Proses Pencarian Buku Dongeng Berbasis Web dengan Menerapkan Metode Klasifikasi Naïve Bayes. *Seminar Nasional Teknologi Informasi Dan Komunikasi STI&K (SeNTIK)*, *5*(1).