Klasifikasi Data Saran Pemustaka di Perpustakaan STIKOM Bali Menggunakan TF-IDF dan Multinomial Naive Bayes

I Made Bhaskara Gautama
Institut Teknologi dan Bisnis STIKOM Bali
e-mail: bhaskara@stikom-bali.ac.id

Diajukan: 19 November 2022; Direvisi: 22 Desember 2022; Diterima: 23 Desember 2022

Abstrak

Perpustakaan STIKOM Bali berperan penting dalam mendukung pembelajaran dan riset oleh dosen maupun mahasiswa. Strategi peningkatan layanan perpustakaan dilakukan dengan mengumpulkan data saran dari anggota perpustakaan melalui kuesioner. Saran tersebut dikelompokkan menjadi beberapa kategori. Prioritas perbaikan diambil berdasarkan kategori dengan jumlah saran terbanyak. Namun, pustakawan dan staf memiliki kesulitan untuk mengelompokkan saran sehingga sulit untuk menentukan prioritas perbaikan. Permasalahan ini diatasi dengan mengklasifikasi data saran pemustaka menggunakan machine learning. Data awal berisi 1015 saran dengan 5264 kata. 608 saran tidak valid karena bukan merupakan kata atau diisi dengan huruf acak. Data tersebut diproses menjadi beberapa versi menggunakan pra pemrosesan yang berbeda. Transformasi teks menjadi bentuk numerik dilakukan menggunakan metode TF-IDF. Sebelum diklasifikasi, data diseimbangkan menggunakan beberapa metode class balancing yaitu SMOTE, ROS, RUS, dan TL. Proses klasifikasi dilakukan dengan menggunakan metode Multinomial Naive Bayes. Hasil klasifikasi dievaluasi menggunakan metrik performa yaitu accuracy, precission, recall, dan F1-score. Nilai tertinggi diperoleh oleh data yang tidak mengalami pra pemrosesan dan data dengan pra pemrosesan lowecase conversion dengan class balancing ROS dengan nilai accuracy 0.798, precission 0.869, recall 0.798, dan F1-score 0.815. Penelitian ini menemukan bahwa teknik pra pemrosesan tidak selalu meningkatkan performa klasifikasi. Pra pemrosesan sebaiknya dilakukan dengan mempertimbangkan karakteristik data yang digunakan.

Kata kunci: *Klasifikasi, TF-IDF, Multinomial Naive Bayes, Machine learning.*

Abstract

The STIKOM Bali Library plays a crucial role in supporting learning and research for both faculty and students. The strategy for enhancing library services involves collecting suggestion data from library members through questionnaires. These suggestions are categorized into several groups, with improvement priorities based on the categories with the most suggestions. However, librarians and staff face challenges in categorizing suggestions, making it difficult to determine improvement priorities. This issue is addressed by classifying suggestion data using machine learning. The initial data contains 1015 suggestions with 5264 words. 608 suggestions are invalid as they are either not words or filled with random letters. This data is processed into several versions using different preprocessing techniques. Text transformation into numerical form is done using the TF-IDF method. Before classification, the data is balanced using several class balancing methods, including SMOTE, ROS, RUS, and TL. The classification process is performed using the Multinomial Naive Bayes method. The classification results are evaluated using performance metrics, namely accuracy, precision, recall, and F1-score. The highest values were obtained by data without preprocessing and data with lowercase conversion preprocessing with ROS class balancing, with accuracy 0.798, precision 0.869, recall 0.798, and F1-score 0.815. This study found that preprocessing techniques do not always improve classification performance. Preprocessing should be done considering the characteristics of the data used.

Keywords*: Klasifikasi, TF-IDF, Multinomial Naive Bayes, Machine learning.*

1. Pendahuluan

Perpustakaan adalah sebuah institusi atau departemen yang tidak hanya menyediakan akses terhadap beragam literatur dan sumber daya informasi, tetapi juga menjadi pusat pengetahuan bagi komunitas akademik [1]. Institut Teknologi dan Bisnis STIKOM Bali memiliki perpustakaan yang

memegang peran vital dalam mendukung proses belajar mengajar dan penelitian. Perpustakaan ini memiliki koleksi yang luas dan *up-to-date*, sehingga tidak hanya menjadi tempat bagi dosen, staf, dan mahasiswa untuk mendapatkan bahan bacaan yang relevan, tetapi juga sebagai tempat untuk mendalami pengetahuan dan memperluas wawasan mereka. Perpustakaan memegang peran penting dalam mendukung pembelajaran, penelitian, dan pengembangan intelektual di berbagai institusi pendidikan. Sebagai tempat yang menyediakan akses ke berbagai sumber daya informasi, mulai dari buku cetak hingga jurnal elektronik, perpustakaan memberikan kesempatan bagi pengguna untuk mengeksplorasi berbagai topik, memperdalam pemahaman pemustaka, dan mengembangkan keterampilan literasi. Selain itu, perpustakaan juga menciptakan lingkungan yang mendukung kolaborasi dan pertukaran ide, di mana pemustaka dapat berinteraksi, berdiskusi, dan belajar bersama. Dengan demikian, perpustakaan tidak hanya menjadi gudang pengetahuan, tetapi juga menjadi pusat kegiatan intelektual dan sosial yang vital bagi pertumbuhan dan perkembangan masyarakat akademis.

Perpustakaan STIKOM Bali berupaya untuk terus meningkatkan pelayanan dan fasilitas dengan mengadopsi pendekatan proaktif yaitu menyebarkan kuesioner kepada pemustaka melalui *website* perpustakaan. Melalui kuesioner ini, pemustaka memiliki kesempatan untuk menyampaikan pendapat, kebutuhan, dan saran mereka secara langsung kepada pihak perpustakaan. Perpustakaan dapat mengumpulkan umpan balik yang berharga untuk memahami secara lebih baik kebutuhan pengguna dan merancang strategi perbaikan yang sesuai guna meningkatkan pengalaman pengguna dan memenuhi harapan mereka. Kuesioner yang disebarkan melalui *website* perpustakaan menjadi salah satu cara efektif untuk menggali persepsi dan preferensi pemustaka secara luas. Platform *online* digunakan agar perpustakaan dapat mencapai lebih banyak pemustaka, termasuk yang mungkin tidak memiliki kesempatan untuk memberikan masukan secara langsung. Hasil dari kuesioner ini memberikan wawasan yang berharga bagi perpustakaan dalam mengidentifikasi area-area yang perlu ditingkatkan, baik itu dari segi koleksi, fasilitas, atau layanan yang ditawarkan. Dengan menganalisis data yang terkumpul, perpustakaan dapat membuat keputusan yang lebih terinformasi dan bertindak responsif terhadap kebutuhan pemustaka, sehingga meningkatkan kualitas layanan secara keseluruhan.

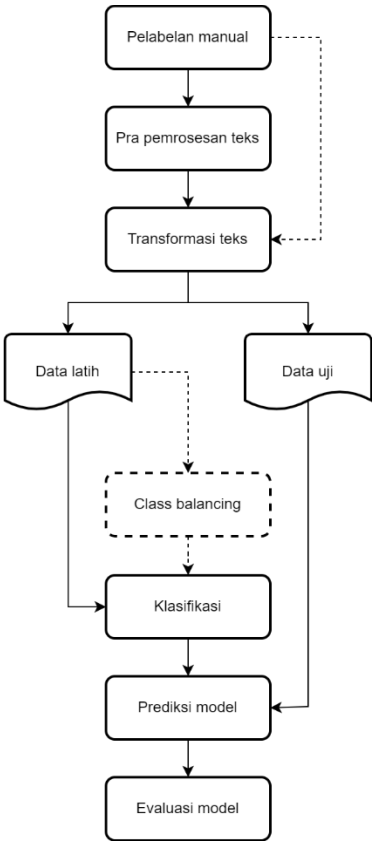
Meskipun kuesioner adalah alat yang berguna untuk mengumpulkan umpan balik, pihak perpustakaan sering menghadapi tantangan dalam memilah dan mengelola data yang masuk. Sebagian besar kuesioner mungkin diisi dengan informasi yang berharga, namun ada juga yang diisi teks acak, bukan merupakan sebuah/sekumpulan kata, atau tanpa relevansi yang jelas. Hal ini dapat menyebabkan kesulitan bagi pihak perpustakaan dalam mengidentifikasi pola atau tren yang signifikan. Selain hal tersebut, dengan jumlah data yang besar, pengelompokan saran-saran menjadi tugas yang memakan waktu dan membingungkan. Untuk mengatasi tantangan ini, perpustakaan dapat menggunakan alat atau teknik analisis data, seperti pemrosesan bahasa alami atau algoritma pembelajaran mesin, untuk membantu dalam mengidentifikasi dan mengelompokkan saran-saran berdasarkan tema atau topik tertentu. Dengan demikian, pihak perpustakaan dapat menghemat waktu dan sumber daya, sambil tetap fokus pada saran-saran yang paling relevan dan bermanfaat untuk diterapkan dalam meningkatkan pelayanan dan fasilitas perpustakaan.

Penelitian ini dilakukan untuk mengatasi permasalahan yang dihadapi oleh Perpustakaan STIKOM Bali dalam mengelompokkan jawaban dari kuesioner yang masuk. Metode Term Frequency-Inverse Document Frequency (TF-IDF) digunakan untuk mengekstraksi fitur-fitur yang relevan dari teks jawaban kuesioner. TF-IDF mengukur seberapa penting sebuah kata dalam suatu dokumen dalam kumpulan dokumen dengan akurasi yang tinggi [2]. Kata-kata yang muncul lebih sering dalam dokumen tertentu namun jarang muncul di dokumen lain akan memiliki bobot yang lebih tinggi. Setelah fitur-fitur diekstraksi, algoritma klasifikasi Multinomial Naive Bayes digunakan untuk mengklasifikasikan jawaban-jawaban tersebut ke dalam kategori atau label yang sesuai. Algoritma ini merupakan pendekatan yang umum digunakan dalam klasifikasi teks karena kecepatan dan keandalannya [3]. Algoritma ini belajar untuk mengidentifikasi pola-pola yang ada dalam teks dan mengklasifikasikan teks-teks baru ke dalam kategori yang tepat berdasarkan kemungkinan terbesar dengan menggunakan data latih yang telah diberi label. Penerapan metode TF-IDF dan algoritma Multinomial Naive Bayes diharapkan dapat membantu Perpustakaan STIKOM Bali dalam mengelompokkan jawaban-jawaban dari kuesioner secara efisien dan efektif, sehingga memungkinkan mereka untuk dengan cepat mengeksplorasi dan memahami umpan balik dari pemustaka dan mengambil tindakan yang sesuai untuk meningkatkan pelayanan dan fasilitas perpustakaan.

2. Metode Penelitian

Penelitian ini menggunakan metode umum yang digunakan untuk melakukan klasifikasi seperti yang ditunjukkan pada Gambar 1 [4]. Data yang telah tersedia yaitu data saran pemustaka atau anggota

perpustakaan yang dikumpulkan melalui *website* Perpustakaan STIKOM Bali untuk anggota yang sudah melakukan *login*. Data tersebut dibuatkan label atau class secara manual. Pada tahap pra pemrosesan teks, terdapat beberapa skenario yang dilakukan yaitu dengan menggunakan kombinasi antara stop-word and punctuation removal, lowercase conversion, dan spelling correction. Hal ini dilakukan karena pra pemrosesan teks dapat mempengaruhi hasil evaluasi dari klasifikasi. Penelitian sebelumnya mengatakan bahwa tidak ada kombinasi yang pasti dari pra pemrosesan teks yang menjamin dapat meningkatkan hasil evaluasi dari model klasifikasi karena sangat bergantung dari dataset dan bahasa yang digunakan [5]. Oleh karena itu, eksperimen dalam penggunaan pra pemrosesan perlu untuk dilakukan.



Gambar 1. Proses klasifikasi [4]

Proses transformasi teks untuk data yang sudah siap dilakukan dengan mengubah teks mentah menjadi representasi numerik atau vektor yang dapat digunakan untuk analisis lebih lanjut. Tujuan dari transformasi teks adalah untuk menghasilkan representasi yang sesuai untuk digunakan dalam algoritma pembelajaran mesin atau analisis statistik. Metode yang digunakan pada tahap ini adalah TF-IDF. TF-IDF adalah teknik yang umum digunakan dalam pemrosesan teks untuk mengukur pentingnya sebuah kata dalam suatu dokumen dalam sebuah koleksi dokumen. Metode ini berguna untuk mengekstraksi dan memberi bobot pada kata-kata dalam dokumen berdasarkan seberapa sering kata tersebut muncul dalam dokumen tertentu dan seberapa jarang kata tersebut muncul di seluruh koleksi dokumen. Proses TF-IDF terdiri dari dua langkah utama yaitu Term Frequency (TF) dan Inverse Document Frequency (IDF).

Term Frequency (TF): Mengukur seberapa sering sebuah kata muncul dalam suatu dokumen tertentu. Perhitungan dilakukan dengan membagi jumlah kemunculan kata tersebut dengan total jumlah kata dalam dokumen. Jadi, untuk sebuah kata t dalam sebuah dokumen d, TF ditentukan sebagai:

$$TF(t,d) = \frac{jumlah\ kemunculan\ kata\ t\ dalam\ dokumen\ d}{total\ jumlah\ kata\ dalam\ dokumen\ d} \quad (1)$$

Inverse Document Frequency (IDF): Mengukur seberapa jarang sebuah kata muncul di seluruh koleksi dokumen. Perhitungan dilakukan dengan membagi jumlah total dokumen dengan jumlah dokumen yang mengandung kata tersebut, kemudian mengambil logaritma dari hasilnya. Perhitungan ini membantu

dalam menurunkan bobot kata-kata yang umum dan meningkatkan bobot kata-kata yang jarang muncul dalam koleksi dokumen. Jadi, untuk sebuah kata t dalam sebuah koleksi dokumen, IDF ditentukan sebagai:

$$IDF(t) = log \frac{total\ jumlah\ dokumen}{jumlah\ dokumen\ yang\ mengandung\ kata\ t} \quad (2)$$

Kemudian, bobot TF-IDF untuk sebuah kata t dalam sebuah dokumen d adalah hasil perkalian dari TF dan IDF:

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t) \quad (3)$$

Bobot TF-IDF yang dihitung menggunakan persamaan tersebut memberikan representasi vektor numerik untuk setiap dokumen dalam koleksi dokumen, di mana kata-kata yang penting dalam suatu dokumen diberi bobot yang lebih tinggi, sementara kata-kata yang umum atau sering muncul diberi bobot yang lebih rendah.

Hasil dari transformasi teks tersebut dibagi secara acak menjadi data latih (80%) dan data uji (20%). Sebelum proses klasifikasi dilakukan, jika terdapat ukuran kelas yang tidak seimbang, dibutuhkan metode untuk menyeimbangkan ukuran kelas atau class balancing dari data latih. Tujuan penyeimbangan ukuran kelas ini adalah untuk memastikan bahwa model pembelajaran mesin tidak memihak pada kelas mayoritas dan mengabaikan kelas minoritas dalam dataset yang tidak seimbang. Penelitian ini menggunakan teknik class balancing yang umum digunakan yaitu Synthetic Minority Oversampling Technique (SMOTE), Random Over Sampling (ROS), Random Under Sampling (RUS), dan Tomek Links karena terbukti mampu meningkatkan performa dari klasifikasi untuk data yang tidak seimbang [6], [7], [8].

Klasifikasi dilakukan menggunakan metode Binomial Naive Bayes [9]. Naive Bayes adalah salah satu algoritma klasifikasi yang populer dalam pembelajaran mesin, yang didasarkan pada teorema Bayes dengan asumsi bahwa fitur-fitur yang digunakan dalam klasifikasi adalah independen satu sama lain. Binomial Naive Bayes adalah variasi dari algoritma Naive Bayes yang digunakan khusus untuk klasifikasi biner, di mana terdapat dua kelas yang mungkin untuk setiap sampel. Proses ini menghasilkan model klasifikasi yang kemudian diuji atau digunakan untuk memprediksi kelas dari data uji. Evaluasi dilakukan dengan mengukur metrik akurasi, presisi, recall, dan F1-score.

3. Hasil dan Pembahasan

Proses pengolahan data dan klasifikasi dilakukan menggunakan bahasa pemrograman Python. Data yang digunakan pada penelitian ini terdiri dari satu kolom untuk setiap baris datanya. Kolom tersebut berisi saran, tanggapan, atau komentar terhadap Perpustakaan STIKOM Bali yang diisi melalui *website* oleh pemustaka atau anggota perpustakaan. Total data yang terkumpul yaitu sebanyak 1015 baris data. Tampilan potongan data mentah ditampilkan pada Gambar 2.

response ▾ 1
ya
ya
ya
xasdasd asd asd
wsedrfds
WIFI KAMPUS YANG FREE BELUM SAMPAI DI PERPUS
Websitenya mengarahkan ke halaman survey
vewbewbww
Upload koleksi skripsi lebih lengkap di library st...
Upgrade meja dan kursi perpus
Update TA lebih banyak lagi

Gambar 2. Data respons kuesioner yang tersimpan pada *database*.

3.1. Pelabelan Manual

Setiap data pada Gambar 2 diberikan label secara manual. Sebelum label diberikan, kriteria klasifikasi ditentukan terlebih dahulu yaitu: (1) tidak valid; (2) ketersediaan informasi; (3) tata kelola bahan pustaka; (4) layanan perpustakaan; (5) bahan pustaka; (6) infrastruktur; dan (7) fitur *website*. Pelabelan manual dilakukan dengan memperhatikan konteks dan makna dari setiap komentar, serta mengidentifikasi isu-isu yang paling sering muncul atau paling penting bagi pengguna perpustakaan. Proses pelabelan ini merupakan langkah awal yang krusial dalam penelitian, karena label yang tepat dan akurat akan menjadi dasar untuk analisis selanjutnya, termasuk pembangunan model klasifikasi untuk mengidentifikasi pola-pola dan tren-tren dalam data komentar. Gambar 3 menunjukkan penggalan data yang telah diberi label.

```
import pandas as pd

# Load the CSV file
file_path = "/content/komentar.csv"
data = pd.read_csv(file_path)

# Display the head of the dataframe
print(data.head())
```

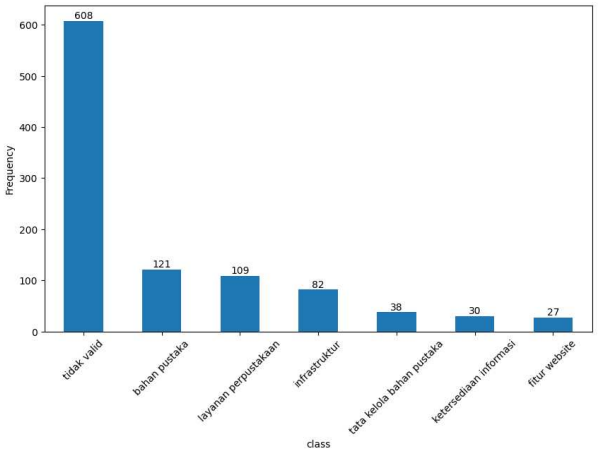
```
                                          suggestion                  class
0  Kelengkapan informasi katalog bahan pustaka ya...  ketersediaan informasi
1                                 semoga semaki baik            tidak valid
2  semoga lampiran pada tugas akhir di isikan kar...  ketersediaan informasi
3                                          tidak ada            tidak valid
4                              semoga lebih baik lagi            tidak valid
```

Gambar 3. Data yang telah diberi label (class).

3.2. Pra Pemrosesan Teks

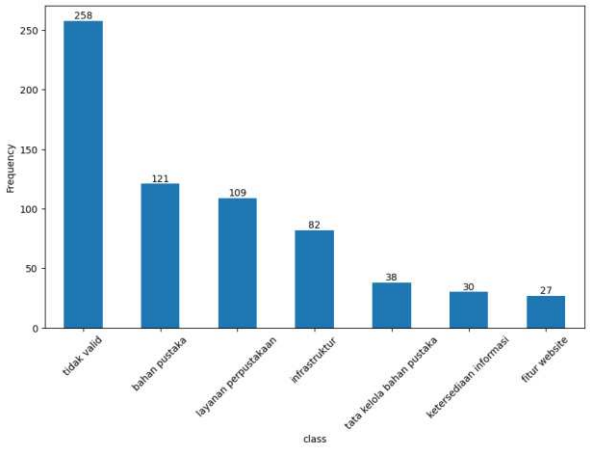
Pra pemrosesan teks dilakukan untuk membersihkan dan menyiapkan data teks sehingga dapat diolah lebih lanjut oleh algoritma pembelajaran mesin atau analisis teks. Penelitian ini melakukan eksperimen dengan menggunakan tiga teknik pra pemrosesan yang banyak digunakan yaitu stop-word and punctuation removal, lowercase conversion, dan spelling correction [5], [10]. Berikut adalah kombinasi pra pemrosesan yang digunakan.

a. Tidak menggunakan pra pemrosesan (BASE)
   Data ini tidak mengalami pra pemrosesan yang selanjutnya disebut sebagai BASE. Jumlah baris data tidak mengalami perubahan dari data awal yaitu 1015 data. Gambar 4 menunjukkan distribusi data ini di masing-masing kelasnya.



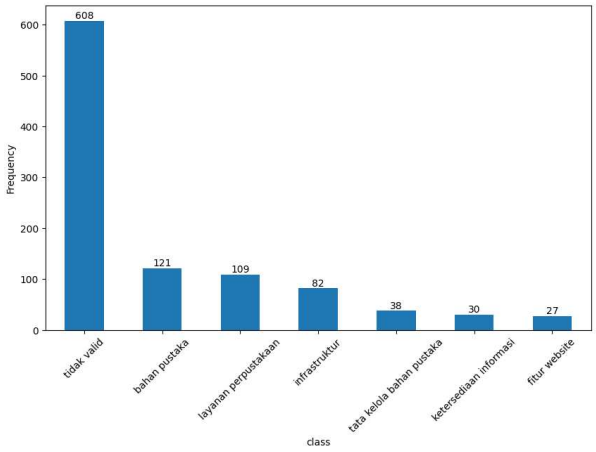
Gambar 4. Jumlah respons berdasarkan kelas (BASE).

b. Stop-word and punctuation removal (DS-1)
   Data ini mengalami pra pemrosesan dengan menghapus stop-word dan tanda baca. Terdapat kasus di mana satu kalimat saran merupakan stop-word semua karena kuesioner tidak diisi dengan baik oleh responden, oleh karena itu terdapat data komentar yang kosong. Baris data ini kemudian dihapus karena dapat mempengaruhi hasil klasifikasi. Proses tersebut menyisakan 665 dari 1015 data. Gambar 5 menunjukkan distribusi data ini di masing-masing kelasnya.



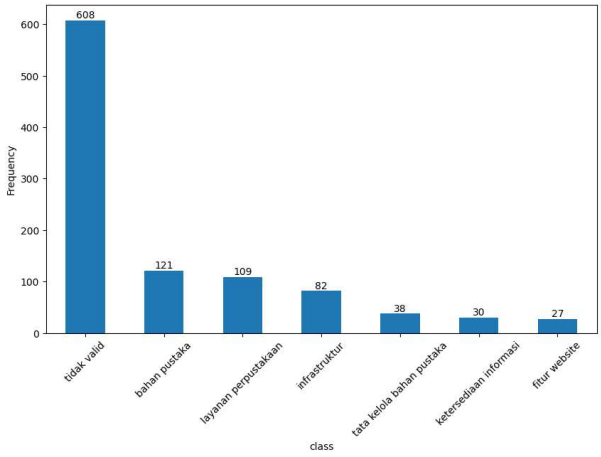
Gambar 5. Jumlah respons berdasarkan kelas (DS-1).

c. Lowercase conversion (DS-2)
Data ini mengalami pra pemrosesan dengan mengubah seluruh huruf menjadi huruf kecil. Jumlah baris data tidak mengalami perubahan dari data awal yaitu 1015 data. Gambar 6 menunjukkan distribusi data ini di masing-masing kelasnya.



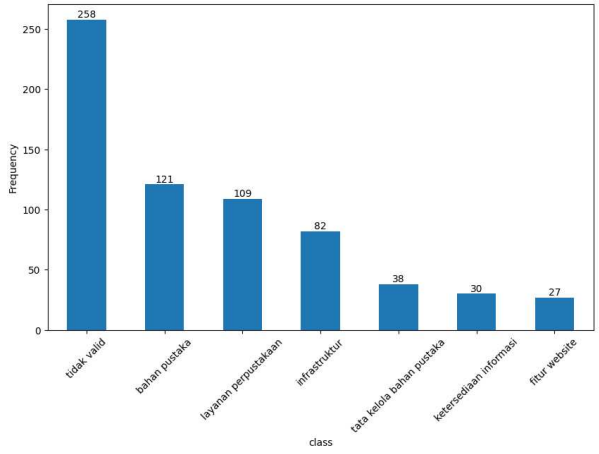
Gambar 6. Jumlah respons berdasarkan kelas (DS-2).

d. Spelling correction (DS-3)
Data ini mengalami pra pemrosesan dengan memperbaiki kesalahan penulisan dan mengubah sejumlah kata tidak baku menjadi kata baku dalam bahasa Indonesia. Jumlah baris data tidak mengalami perubahan dari data awal yaitu 1015 data. Gambar 7 menunjukkan distribusi data ini di masing-masing kelasnya.



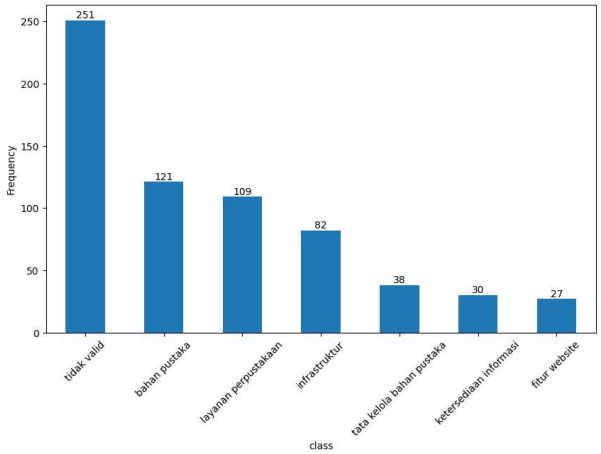
Gambar 7. Jumlah respons berdasarkan kelas (DS-3).

e. Stop-word and punctuation removal dan Lowercase conversion (DS-4)
Pra pemrosesan yang digunakan pada data ini merupakan kombinasi dari penghapusan stop-word, tanda baca, dan konversi semua huruf menjadi huruf kecil. Karena melibatkan penghapusan stop-word, jumlah data ini mengalami pengurangan menjadi 665 data. Gambar 8 menunjukkan distribusi data ini di masing-masing kelasnya.



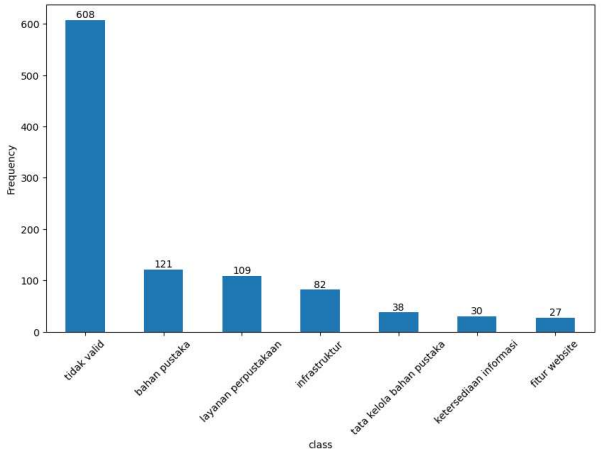
Gambar 8. Jumlah respons berdasarkan kelas (DS-4).

f. Stop-word and punctuation removal dan Spelling correction (DS-5)
Pra pemrosesan yang digunakan pada data ini merupakan kombinasi dari penghapusan stop-word, tanda baca, dan perbaikan kesalahan penulisan dan mengubah sejumlah kata tidak baku menjadi kata baku dalam bahasa Indonesia. Jumlah data ini mengalami pengurangan menjadi 658 data. Data yang berkurang pada pra pemrosesan ini jumlahnya lebih banyak daripada penghapusan stop-word saja karena pengaruh dari perbaikan kesalahan pengetikan kata. Kata yang merupakan stop-word, yang sebelumnya salah ketik, menjadi terhapus pada pra pemrosesan ini. Gambar 9 menunjukkan distribusi data ini di masing-masing kelasnya.



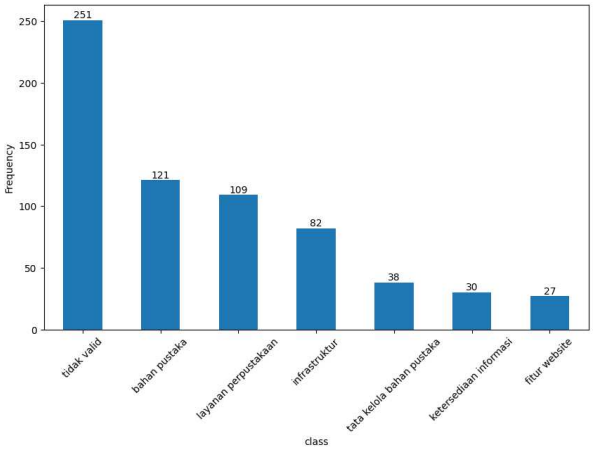
Gambar 9. Jumlah respons berdasarkan kelas (DS-5).

g. Lowercase conversion dan Spelling correction (DS-6)
Data ini mengalami pra pemrosesan dengan mengubah semua huruf menjadi huruf kecil dan memperbaiki kesalahan penulisan serta mengubah sejumlah kata tidak baku menjadi kata baku dalam bahasa Indonesia. Jumlah baris data tidak mengalami perubahan dari data awal yaitu 1015 data. Gambar 10 menunjukkan distribusi data ini di masing-masing kelasnya.



Gambar 10. Jumlah respons berdasarkan kelas (DS-6).

h. Stop-word and punctuation removal, Lowercase conversion, dan Spelling correction (DS-7)
Data ini diproses menggunakan ketiga teknik pra pemrosesan. Jumlah data ini mengalami pengurangan menjadi 658 data. Gambar 11 menunjukkan distribusi data ini di masing-masing kelasnya.



Gambar 11. Jumlah respons berdasarkan kelas (DS-7).

Dataset yang telah mengalami pra pemrosesan selanjutnya dipecah menjadi data training dan data testing. Rasio pemecahan yang dilakukan adalah 20% dari keseluruhan data digunakan sebagai data testing dan 80% dari keseluruhan data digunakan untuk melatih model klasifikasi.

3.3. Transformasi Teks

Transformasi teks dilakukan dengan menggunakan metode TF-IDF. Berikut adalah hasil dari transformasi teks berupa bentuk matriks TF-IDF.

Tabel 1. Bentuk matriks TF-IDF untuk masing-masing dataset.

Dataset	n_docs	n_features
BASE	1015	871
DS-1	665	685
DS-2	1015	871
DS-3	1015	779
DS-4	665	685
DS-5	658	589
DS-6	1015	779
DS-7	658	589

Kolom dataset merepresentasikan data yang telah mengalami proses pra pemrosesan menggunakan kombinasi yang telah dipaparkan sebelumnya. Kolom n_docs merupakan jumlah dokumen atau teks di dalam korpus yang digunakan untuk analisis. Kolom n_feature yaitu jumlah istilah atau teks unik yang diekstrak dari keseluruhan dokumen.

3.4. Class Balancing

Berdasarkan distribusi respons berupa komentar atau saran pada masing-masing kelas yang ditunjukkan pada Gambar 4 sampai dengan Gambar 11, terlihat bahwa kelas tidak valid, atau komentar yang diisi tanpa konteks, memiliki jumlah paling banyak dan menyebabkan distribusi data yang tidak seimbang. Penyeimbangan kelas dilakukan menggunakan dua teknik oversampling (SMOTE dan ROS) dan dua teknik undersampling (RUS dan TL). Masing-masing dataset yang telah mengalami pra pemrosesan mengalami proses class balancing yang sama.

3.5. Klasifikasi

Proses klasifikasi dilakukan dengan menggunakan metode Binomial Naive Bayes. Pada proses ini, model klasifikasi dilatih menggunakan data latih kemudian diuji menggunakan data uji.

3.6. Evaluasi

Proses evaluasi dilakukan dengan mengukur metrik performa atau kinerja berdasarkan pengujian yang dilakukan pada fase sebelumnya. Metrik yang diukur yaitu accuracy, precision, recall, dan F1-score. Tabel 2 menunjukkan hasil pengukuran metrik kinerja dari semua kombinasi eksperimen yang dilakukan.

Tabel 2. Pengukuran metrik kinerja.

		BASE	SMOTE	ROS	RUS	TL
BASE	**accuracy**	0.759	0.783	0.798	0.724	0.754
	precision	0.732	0.862	0.869	0.835	0.723
	recall	0.759	0.783	0.798	0.724	0.754
	f1 score	0.688	0.805	0.815	0.761	0.679
DS-1	**accuracy**	0.579	0.684	0.714	0.647	0.571
	precision	0.501	0.698	0.710	0.676	0.501
	recall	0.579	0.684	0.714	0.647	0.571
	f1 score	0.489	0.684	0.708	0.652	0.484
DS-2	**accuracy**	0.759	0.783	0.798	0.724	0.754
	precision	0.732	0.862	0.869	0.835	0.723
	recall	0.759	0.783	0.798	0.724	0.754
	f1 score	0.688	0.805	0.815	0.761	0.679
DS-3	**accuracy**	0.783	0.788	0.798	0.739	0.773
	precision	0.732	0.862	0.862	0.832	0.743
	recall	0.783	0.788	0.798	0.739	0.773
	f1 score	0.729	0.807	0.813	0.769	0.714
DS-4	**accuracy**	0.579	0.684	0.714	0.647	0.571
	precision	0.501	0.698	0.710	0.676	0.501
	recall	0.579	0.684	0.714	0.647	0.571
	f1 score	0.489	0.684	0.708	0.652	0.484
DS-5	**accuracy**	0.697	0.758	0.765	0.705	0.689
	precision	0.663	0.790	0.791	0.731	0.657
	recall	0.697	0.758	0.765	0.705	0.689
	f1 score	0.638	0.765	0.771	0.705	0.631
DS-6	**accuracy**	0.783	0.788	0.798	0.739	0.773
	precision	0.732	0.862	0.862	0.832	0.743
	recall	0.783	0.788	0.798	0.739	0.773
	f1 score	0.729	0.807	0.813	0.769	0.714
DS-7	**accuracy**	0.697	0.758	0.765	0.705	0.689
	precision	0.663	0.790	0.791	0.731	0.657
	recall	0.697	0.758	0.765	0.705	0.689
	f1 score	0.638	0.765	0.771	0.705	0.631

Nilai accuracy merepresentasikan rasio prediksi yang benar (positif dan negatif) dengan keseluruhan prediksi. Akurasi mengukur seberapa sering model klasifikasi memberikan prediksi yang benar dinyatakan dalam persentase. Precision mengukur seberapa tepat atau relevan prediksi positif model dengan membandingkan rasio prediksi positif yang benar (True Positive) dengan total prediksi positif yang dilakukan oleh model. Recall mengukur seberapa banyak dari keseluruhan kelas positif yang berhasil diidentifikasi oleh model dengan membandingkan rasio prediksi positif yang benar (True Positive) dengan total jumlah kelas positif yang sebenarnya. F1-score adalah rata-rata harmonik dari precision dan recall. Nilai ini memberikan keseimbangan antara kedua metrik tersebut dan berguna ketika ada ketidakseimbangan antara kelas-kelas atau ketertarikan yang sama terhadap presisi dan recall.

Secara keseluruhan, metrik kinerja tertinggi didapatkan oleh dua kombinasi dataset dan teknik class balancing yaitu BASE dan ROS, DS-2 dan ROS. Pada kasus ini, teknik class balancing menggunakan ROS meningkatkan kinerja dari model klasifikasi. Peningkatan yang terjadi menggunakan teknik class balancing ROS yaitu accuracy rata-rata meningkat sebesar 10.06%, precision 24.32%, recall 10.06%, dan F1-score 23.82%. Tanpa menggunakan teknik class balancing, dataset yang mengalami perbaikan kesalahan pengetikan dan kata baku memiliki nilai metrik kinerja tertinggi. Penghapusan stop word dan tanda baca secara keseluruhan memiliki nilai metrik kinerja terendah karena sebanyak 2779 kata dari 5264 kata atau 53% kata di dalam dataset merupakan stop word. Penghapusan stop word pada data yang digunakan dalam penelitian ini berpengaruh signifikan terhadap performa klasifikasi karena menyebabkan data kehilangan konteksnya.

4. Kesimpulan

Penelitian ini dilakukan untuk mengklasifikasi data saran pemustaka di Perpustakaan STIKOM Bali. TF-IDF digunakan untuk mentransformasi teks menjadi representasi numerik. Model klasifikasi yang digunakan adalah Multinomial Naive Bayes. Sebelum proses klasifikasi dilakukan, data diberi label atau kelas secara manual. Data awal berisi 1015 saran dengan 5264 kata. 608 saran tidak valid karena bukan merupakan kata atau diisi dengan huruf acak. 2779 kata dari keseluruhan kata, atau 53% di antaranya merupakan stop word. Eksperimen dilakukan dengan menggunakan teknik pra pemrosesan dan class balancing yang berbeda. Teknik pra pemrosesan yang digunakan merupakan kombinasi antara stop-word and punctuation removal, lowercase conversion, dan spelling correction. Kombinasi tersebut menghasilkan

8 versi data yang berbeda, termasuk data yang tidak mengalami pra pemrosesan. Teknik class balancing yang digunakan yaitu SMOTE, ROS, RUS, dan TL.

Penelitian ini menemukan bahwa dengan menggunakan data yang memiliki karakteristik tersebut di atas, teknik pra pemrosesan lowercase conversion tidak berpengaruh terhadap hasil klasifikasi, sedangkan penghapusan stop word berpengaruh signifikan dalam menurunkan performa klasifikasi. Spelling correction dapat meningkatkan hasil klasifikasi dibandingkan dengan data yang tidak mengalami pra pemrosesan, namun dengan menggunakan teknik class balancing ROS, menurunkan nilai precision dan F1-score namun tidak signifikan. Secara keseluruhan teknik class balancing ROS dapat meningkatkan nilai metrik kinerja atau performa klasifikasi pada kasus ini. Nilai tertinggi diperoleh oleh data yang tidak mengalami pra pemrosesan dan data dengan pra pemrosesan lowecase conversion dengan class balancing ROS dengan nilai accuracy 0.798, precission 0.869, recall 0.798, dan F1-score 0.815.

Berdasarkan hasil evaluasi, model klasifikasi yang terbentuk belum memiliki performa yang baik sehingga belum dapat digunakan oleh Perpustakaan STIKOM Bali secara optimal. Oleh karena itu, diperlukan penelitian lebih lanjut dengan melakukan eksperimen terhadap penggunaan metode transformasi teks atau metode klasifikasi yang berbeda. Selain itu, berdasarkan karakteristik data yang dikumpulkan, perlu dilakukan validasi pengisian kuesioner untuk mencegah saran yang tidak valid.

Daftar Pustaka

[1] M. Djaenudin and C. Trianggoro, "Inovasi Layanan Perpustakaan Khusus Dalam Ekosistem E-Research Dalam Mendukung Open Science: Studi Kasus Perpustakaan PDDI LIPI," *Al Maktabah*, vol. 19, no. 1, 2020.

[2] A. Alessa and M. Faezipour, "Tweet classification using sentiment analysis features and TF-IDF weighting for improved flu trend detection," in *Machine Learning and Data Mining in Pattern Recognition: 14th International Conference, MLDM 2018, New York, NY, USA, July 15-19, 2018, Proceedings, Part I 14*, 2018, pp. 174–186.

[3] G. Singh, B. Kumar, L. Gaur, and A. Tyagi, "Comparison between multinomial and Bernoulli naive Bayes for text classification," in *2019 International conference on automation, computational and technology management (ICACTM)*, 2019, pp. 593–596.

[4] N. L. P. M. Putu, A. Z. Amrullah, and others, "Analisis Sentimen dan Pemodelan Topik Pariwisata Lombok Menggunakan Algoritma Naive Bayes dan Latent Dirichlet Allocation," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 1, pp. 123–131, 2021.

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

[6] D. S. Sisodia, N. K. Reddy, and S. Bhandari, "Performance evaluation of class balancing techniques for credit card fraud detection," in *2017 IEEE International Conference on power, control, signals and instrumentation engineering (ICPCSI)*, 2017, pp. 2747–2752.

[7] B. Drury and A. de Andrade Lopes, "A comparison of the effect of feature selection and balancing strategies upon the sentiment classification of Portuguese news stories," *Proceedings of ENIAC*, 2014.

[8] M. A. Al-Asadi and S. Tasdemir, "Empirical comparisons for combining balancing and feature selection strategies for characterizing football players using FIFA video game system," *IEEE Access*, vol. 9, pp. 149266–149286, 2021.

[9] A. Kelly and M. A. Johnson, "Investigating the statistical assumptions of Naive Bayes classifiers," in *2021 55th annual conference on information sciences and systems (CISS)*, 2021, pp. 1–6.

[10] A. K. Uysal and S. Gunal, "The impact of preprocessing on text classification," *Inf Process Manag*, vol. 50, no. 1, pp. 104–112, 2014.