 ISSN:2302-8149
e-ISSN:2540-9719

Optimalisasi Seleksi Fitur dalam Analisis Sentimen Bank Saqu: Studi Perbandingan SVM dan Random Forest Menggunakan Information Gain dan Chi-Square

Optimizing Feature Selection in Sentiment Analysis of Bank Saqu: A Comparative Study of SVM and Random Forest using Information Gain and Chi-Square

[1]Anelta Tirta Putri Subandono*, [2]Dhani Ariatmanto
[1,2]Magister Teknik Informatika, Fakultas Ilmu Komputer, Universitas AMIKOM
[1,2]Jl. Ring Road Utara, Ngringin, Condongcatur, Kec. Depok, Kabupaten Sleman, Daerah Istimewa Yogyakarta 55281
*e-mail: anelta.putri2000@gmail.com

(***received*:** 27 February 2025, ***revised*:** 5 March 2025, ***accepted*:** 12 March 2025)

Abstrak

Pemilihan metode seleksi fitur yang optimal menjadi faktor krusial dalam meningkatkan akurasi dan efisiensi model klasifikasi teks. Fitur yang tidak relevan dapat menyebabkan penurunan performa model, meningkatkan kompleksitas komputasi serta menyebabkan overfitting. Berbagai teknik seleksi fitur telah digunakan dalam analisis sentimen, namun kajian secara sistematis membandingkan efektivitas Information Gain dan Chi-Square dalam meningkatkan kinerja model klasifikasi masih terbatas.Bagaimanapun, tujuan penelitian ini untuk mengevaluasi dan mengoptimalkan perbandingan metode seleksi fitur terhadap performa Support Vector Machine (SVM) dan Random Forest (RF) dalam analisis sentimen. Eksperimen dilakukan dalam delapan skema pengujian yang mencakup model tanpa seleksi fitur, model dengan Information Gain, Chi-Square, serta kombinasi keduanya. Hasil pengujian menunjukkan bahwa SVM dengan Chi-Square mencapai akurasi tertinggi sebesar 93% , sedangkan Random Forest dengan Chi-Square memperoleh akurasi terbaik sebesar 91%. Temuan ini mengindikasikan bahwa Chi-Square lebih efektif dibandingkan Information Gain dalam meningkatkan akurasi, serta SVM memiliki performa lebih unggul dibandingkan Random Forest dalam klasifikasi teks. Kesimpulannya, pemilihan metode seleksi fitur yang tepat berkontribusi signifikan dalam meningkatkan akurasi model klasifikasi teks. Hasil penelitian ini dapat menjadi referensi dalam optimalisasi teknik seleksi fitur untuk pengembangan sistem berbasis machine learning yang lebih akurat dan efisien

Kata kunci: analisis sentimen, support vector machine (SVM), random forest (RF), chi-square, information gain

Abstract

The selection of an optimal feature selection method is a crucial factor in improving the accuracy and efficiency of text classification models. Irrelevant features can degrade model performance, increase computational complexity, and lead to overfitting. Although various feature selection techniques have been employed in sentiment analysis, systematic studies comparing the effectiveness of Information Gain and Chi-Square in enhancing classification performance remain limited. This study aims to evaluate and optimize the impact of different feature selection methods on the performance of Support Vector Machine (SVM) and Random Forest (RF) in sentiment analysis. Experiments were conducted using eight testing schemes, including models without feature selection, with Information Gain, Chi-Square, and a combination of both. The results showed that SVM with Chi-Square achieved the highest accuracy at 93%, while Random Forest with Chi-Square achieved the best performance at 91%. These findings indicate that Chi-Square is more effective than Information Gain in improving accuracy, and that SVM outperforms Random Forest in text classification tasks. In conclusion, selecting the appropriate feature selection method significantly contributes to enhancing the accuracy

 ISSN:2302-8149
e-ISSN:2540-9719

of text classification models. This research can serve as a reference for optimizing feature selection techniques in the development of more accurate and efficient machine learning-based systems.

Keywords: *sentiment analysis, support vector machine (SVM), random forest (RF), chi-square, information gain*

1 Pendahuluan

Fintech telah mengubah interaksi masyarakat dengan layanan keuangan, menghadirkan efisiensi dan kenyamanan melalui inovasi seperti sistem pembayaran digital dan mobile banking. Di Indonesia, pertumbuhan pengguna internet dan smartphone semakin mendorong adopsi layanan perbankan digital yang memungkinkan transaksi keuangan dilakukan kapan saja dan di mana saja. Seiring dengan tuntutan akan kepraktisan, teknologi keuangan terus berkembang pesat. Kepuasan pengguna terhadap layanan perbankan digital menjadi salah satu faktor utama yang mendorong ekspansi industri ini[1] [2]

Dengan adanya ulasan atau review secara umum akan menunjukkan tingkat penerimaan yang positif terhadap inovasi ini [3], Ulasan ini akan mencakup beberapa aspek , baik kelebihan maupun kekurangan dan ini akan sangat mempengaruhi penilaian akan kualitas bank digital tersebut [2] Ulasan atau review juga akan mengungkap masalah teknis pada aplikasi bank digital seperti gangguan aplikasi dan adanya kendala dalam autentifikasi, hal ini juga akan mengurangi kepuasan pengguna [4]. Metode analisis sentimen ini juga digunakan untuk menerjemahkan data opini, pemahaman mengolah data tekstual dengan otomatis untuk melihat sentimen public baik itu nilai positif, negatif maupun netral [5], Penting untuk dilakukannya analisis sentimen agar nilai reputasi dapat terukur [6]. Dengan semakin banyaknya ulasan pengguna, diperlukan metode yang efisien untuk mengklasifikasikan sentimen secara otomatis menggunanan teknik machine learning. Namun, pemilihan fitur yang tepat menjadi tantangan utama karena fitur yang tidak relevan dapat menyebabkan penurunan akurasi model dan meningkatkan kompleksitas perhitungan. Information Gain dan Chi-Square adalah dua metode seleksi fitur yang umum digunakan, namun kurangnya kajian sistematis dalam membandingkan efektivitas keduanya dalam meningkatkan akurasi model.

Teknik yang umum digunakan yakni Support Vector Machine (SVM) dan Radom Forest (RF) SVM menjadi teknik pembelajaran yang memiliki dasar teoritis yang kuat dibanding kebanyakan algoritma lain [7], Random Forest akan bekerja dengan menggabungkan beberapa pohon keputusan (Decision Tree) yang dianggap juga optimal dalam analisis sentimen dengan volume data yang besar dan kompleks [8]. Kelebihan yang dimiliki oleh SVM adalah penentuan jarak dengan Support Vector Machine sehingga proses komputasi akan lebih cepat [9], sementara Random Forest dengan menggunakan pendekatan ensambel mampu menangani dara kompleks dan memberikan interpretasi fitur yang baik [10]

Penggunaan Seleksi fitur juga merupakan salah satu bagian penting dalam analisis data sentimen dengan SVM untuk menentukan faktor-faktor yang paling mempengaruhi kepuasan pengguna [4], TF-IDF dalam pembobotan fitur digunakan untuk meniliai seberapa sering suatu kata muncul dalam sebuah dokumen [2] Pada studi ini akan membahas mengenai optimalisasi analisis sentimen dengan membandingkan nilai akurasi dengan dan tanpa penggunaan feature selection . Feature selection yang dipilih yakni Information Gain dan Chi Square, salah satu metode dalam pemilihan fitur yang bekerja dengan cara mengukur relevansi fitur serta memilih fitur informasi yang signifikan, serta mampu menghilangkan overfitting yang akan meningkatkan akurasi pengklasifikasian [11].

Penelitian ini bertujuan untuk menganalisis sentimen terhadap Bank Saqu menggunakan algoritma Support Vector Machine (SVM) dan Random Forest, serta membandingkan kinerjanya dengan dan tanpa penerapan feature selection Information Gain dan Chi-Square. Fokus utama penelitian ini adalah mengevaluasi pengaruh feature selection terhadap akurasi klasifikasi sentimen dan menentukan algoritma yang lebih optimal. Hasil penelitian ini diharapkan dapat memberikan kontribusi ilmiah serta wawasan praktis dalam optimalisasi seleksi fitur untuk meningkatkan efisiensi dan akurasi model klasifikasi, khususnya dalam pengembangan layanan financial technology.

2 Tinjauan Literatur

Berbagai penelitian sebelumnya telah membahasa analisis sentimen dalam berbagai domain, termasuk layanan digital dan perbankan, serta dalam pengujian efektivitas algoritma machine learning

 ISSN:2302-8149
e-ISSN:2540-9719

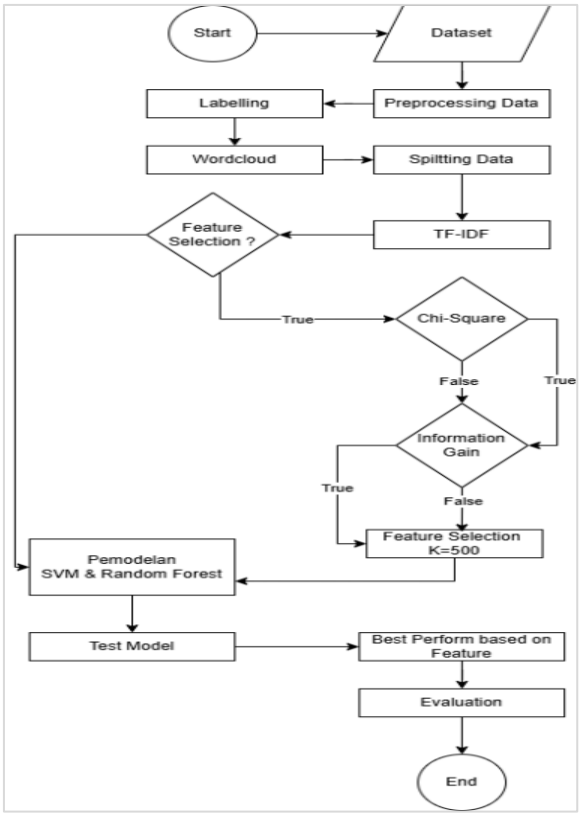
seperti SVM dan Random Forest. Dari segi metode studi rujukan utma yang digunakan oleh [7] menghasilkan nilai akurasi yang tinggi yakni 98% dalam klasifikasi sentimen Shoppe, yang memperkuat alasan pemilihannya dalam penelitian ini. Studi [12] menunjukkan bahwa Information Gain dapat meningkatkan akurasi model, namun belum ada studi yang membandingkan dengan metode lain dalam konteks analisis sentimen perbankan. Studi [13] menghasilkan tingkat akurasi mencapai 87% , namun memiliki keterbatasan dalam menanganai teks konteks kompleks karena tidak menggunakan feature selection. Selanjutnya pada studi [14] menunjukkan bahwa seleksi fitur akan berperan penting dalam peningkatan akurasi klasifikasi sentimen dengan hasil akurasi sebelum digunakan feature selection 60,81% , kemudian setelah menggunakan feature selection didapati peningkatan akurasi sebesar 63,10%.

Berdasarkan penelitian yang sebelumnya, metode seleksi fitur seperti Information Gain dan Chi-Square telah digunakan untuk meningkatkan akurasi model machine learning dalam berbagai domain. Namun masih terdapat kesenjangan penelitian dalam perbandingan sistematis kedua metode ini dalam konteks analisis sentimen. Selain itu, penelitian sebelumnya lebih banyak berfokus pada penggunaan salah satu metode seleksi fitur, tanpa mengeksplorasi kemungkinanan kombinasi keduanya untuk meningkatkan kinerja model klasifikasi. Oleh karena itu, penelitian ini berkontribusi dengan mengevaluasi efektivitas Infromation Gain dan Chi-Square secara individual maupun kombinasi dalam meningkatkan akurasi algoritrma Support Vector Machine dan Random Forest dalam analisis sentimen.

3 Metode Penelitian

Untuk mengevaluasi pengaruh metode seleksi fitur terhadap performa klasifikasi, penelitian ini mengusulkan penggunaan Infromation Gain dan Chi-Square dalam analisis sentimen terhadap data ulasan Bank Saqu. Akan ada delapan skema pengujian yang akan diterapkan, mencakup model tanpa seleksi fitur, model dengan Infromation Gain, model dengan Chi-Square, serta kombinasi keduanya. Pengujian akan dilakukan dengan algoritma Support Vector Machine (SVM) dan Random Forest (RF) untuk menentukan metode seleksi fitur yang paling efektif dalam meningkatkan akurasi model

Tahapan penelitian dilakukan dengan beberapa proses yang saling terkait. Berikut adalah bagan tahapan penelitian, pada Gambar 1



Gambar 1. Diagram alur penelitian

 ISSN:2302-8149
e-ISSN:2540-9719

Langkah awal ialah pengumpulan data berasal dari ulasan aplikasi Bank Saqu yang diperoleh dengan metode web – scraping dari halaman Google Play Store. Teknik ini digunakan karena bisa memungkinkan pengambilan data secara otomatis dengan volume yang cukup besar. Data yang diperoleh meliputi teks ulasan yang memiliki isi opini dan pengalaman pengguna mengenai aplikasi tersebut. setelah data didapatkan , Langkah selanjutnya adalah tahap pra-pemrosesan data yang mencakup pembersihan data yakni penghapusan stopword,tokenisasi, labeling, dan stemming untuk memastikan data yang digunakan dalam analisis dapat diolah dengan baik. Data yang sudah selesai pada tahap pembersihan akan diproses dalam penerapan algoritma klasifikasi dengan Support Vector Machine (SVM) dan Random Forest dengan dan tanpa feature selection Information Gain dan Chi-Square. Kemudian tahap akhir akan dilakukan evaluasi model untuk menguji performa terbaik dalam pengklasifikasian sentimen.

3.1 Pengumpulan Data

Pengumpulan dataset yang relevan dilakukan dalam penelitian ini. Dataset yang digunakan dari ulasan pengguna aplikasi Bank Saqu di Google Play Store dalam rentang waktu November 2023 – November 2024. Data dikumpulkan dengan menggunakan teknik web scarping dan disimpan dalam format CSV dengan berbagai varia-ble yang relevan untuk dianalisis.

3.2 Data Pre-Processing

Pembersihan data ini termasuk penghapusan data yang tidak relevan, data duplikat, data kosong, dsb. Berikut merupakan berbagai proses tahapan dalam preprocessing data. Berikut adalah tahapan dalam pre-processing data.

a. Case Folding
Case folding ini mengonversi semua huruf dengan format huruf kecil yang bertujuan untuk meminimalisir ragam dalam representasi teks [15]. Berikut pada Tabel 1 merupakan contoh dari Case Folding

Tabel 1. Contoh case folding

Sebelum Case Folding	Setelah Case Folding
Saya ini Mahasiswi di Universitas Amikom	saya ini mahasiswa di universitas amikom

b. Normalisasi
Normalisasi akan mengurangi variasi kata yang tidak diperlukan, konversi angka ke bentuk kata, menyamakan singkatan, dan memperbaiki ejaan [15]

c. Stopwords Removal
Proses yang berfungsi untuk memperbaiki kualitas teks dengan menghilangkan kata-kata yang tidak bermakna signifikan atau yang tidak menyampaikan informasi yang relevan [16]. Berikut pada Tabel 2 merupakan contoh dari Stopwords Removal

Tabel 2. Contoh stopwords removal

Sebelum Case Folding	Setelah Case Folding
Saya ini Mahasiswi di Universitas Amikom	[“Saya”, “Mahasiswi”,“Universitas”, “Amikom’]

d. Stemming
Proses pengubahan kata berimbuhan menjadi kata dasar, dengan menghapus semua imbuhan awalan atau akhiran yang ada pada kata dalam data ulasan [11] Berikut adalah contoh dari stemming pada Tabel 3.

Tabel 3. Contoh stemming

Sebelum Case Folding	Setelah Case Folding
Buku bacaan di perpustakaan Universitas Amikom menarik untuk dibaca	Buku **bacaan** di perpustakaan Universitas Amikom menarik untuk **dibaca** Buku baca di perpustakaan Universitas Amikom menarik untuk baca

 ISSN:2302-8149
e-ISSN:2540-9719

e. Spliting Data
Membagi data menjadi beberapa persen bagian yang akan digunakan untuk Training Set, dan Test Set. Dataset yang telah dilakukan preprocessing akan dilakukan pembagian, berikut adalah skema pembagian data yang akan dilakukan, Berikut jumlah presentasi data bagi yang akan dilakukan pada Tabel 4.

Tabel 4 .Presentase pembagian data

Training Set	Test Set
80%	20%

f. Labelling
Proses di mana data diberikan tag atau label tertentu sehingga bisa dikenali oleh sistem atau model pembela-jaran mesin. Label-label ini biasanya mewakili kategori atau klasifikasi tertentu yang membantu model da-lam memahami dan memproses data secara lebih efektif. mana model dilatih menggunakan data yang telah diberi label untuk melakukan prediksi atau mengambil keputusan [17]. Labeling akan dilakukan dengan pendekatan lexicon karena memiliki keunggulan tidak perlu adanya data pelatihan dan dapat langsung digunakan pada teks.

3.3 Wordcloud

Merupakan salah satu teknik dalam visualisasi yang akan menampilkan Kumpulan kata-kata dengan menggunakan ukuran font yang menjadi cerminan frekuensi atau kepentingannya. akan lebih mengeksplorasi dengan menggunakan analisis matematis dari kata kata yang memiliki frekuensi kemunculan tertinggi [18]

3.4 Feature Selection

Teknik dalam machine learning dimaksudkan untuk menentukan fitur paling signifikan dalam sebuah dataset guna meningkatkan performa model klasifikasi. Dengan mengurangi fitur yang redundan, ini akan meningkatkan efisiensi komputasi serta akurasi model. Dalam penelitian ini Information Gain dan Chi-Square digunakan sebagai sebuah metode feature selection dalam pengukuran relevansi fitur dan mengurangi kemungkinan overfitting yang diharapkan model klasifikasi mampu bekerja lebih optimal.

a. Information Gain
Information Gain akan diberi penilaian dan di urutkan, urutan fitur yang paling tinggu merupakan fitur yang paling berkaitan dan memiliki hubungan yang erat dengan dataset yang digunakan. menecerminkan seberapa banyak informasi yang diperoleh dari fitur tersebut kemudian akan dimasukan ke sistem klasifikasi, semakin banyak informasi yang dibawa maka fitur tersebut akan semakin penting, fitur dengan perolehan informasi yang tinggi akan menunjukkan kemampuan klasifikasi yang lebih besar [19]. Berikut persamaan hitungan dalam Information Gain di tunjukan pada (1) (2)

$$IG\ (S, A) = Entropy(S) - \sum_{v \in Values\ (A)} \frac{|S_v|}{|S|}\ x\ Entropy\ (S_v) \quad (1)$$

$$IG = Entropy\ Sebelum - Entropy\ Setelah \quad (2)$$

b. Chi-Square
Metode ini akan menghitung seberapa besar frekuensi fitur yang di teliti dalam suatu klasfikasi. Nilai Chi-Square untuk setiap term diurutkan berdasarkan urutan kata yang digunakan sebagai fitur [20]. Dan akan membantu mengidentifikasi istilah-istilah penting yang berkontribusi pada keputusan klasifikasi . Chi-Square dapat dirumuskan sebagai berikut [21]. Berikut persamaan hitungan dalam Chi-Square ditunjukan pada (3)

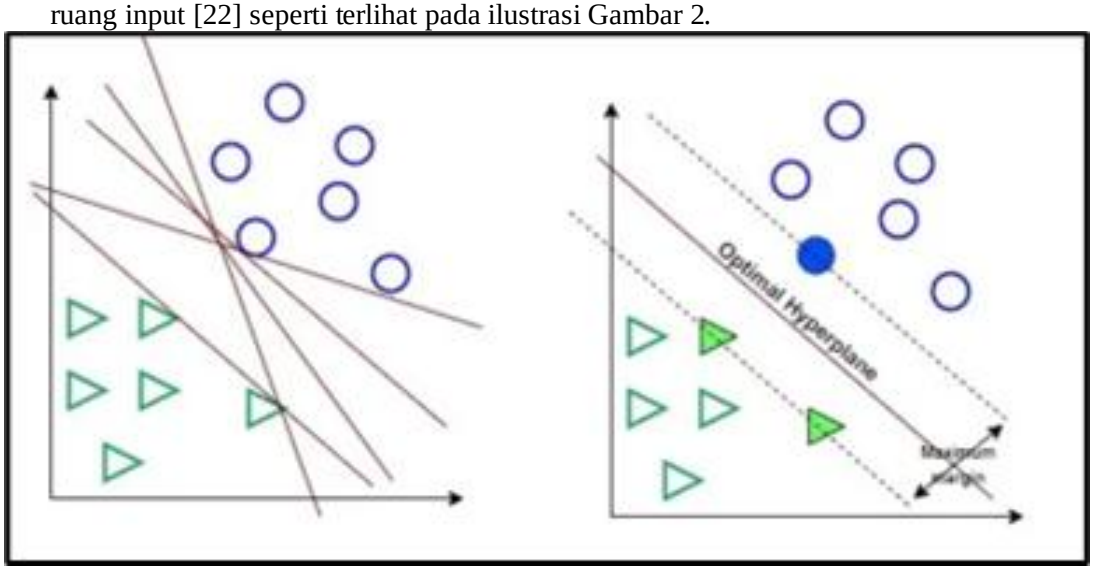
$$X^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(0_{ij} - E_{ij})^2}{E_{ij}} \quad (3)$$

 ISSN:2302-8149
e-ISSN:2540-9719

3.5 Modelling

Dalam analisis sentimen pemilihan algoritma kalsifikasi akan menajdi faktor yang krusial dalam upaya mendapatkan hasil yang akurat dan dapat diandalkan. Support Vector Machine (SVM) dan Random Forest (RF) merupakan dua algoritma yang sering kali digunakan dalam klasifikasi teks karena memiliki keunggulan dalam menangani dara dengan karakteristik kompleks. SVM sendiri meiliki kemampuan yang kuat dalam mencari hyperlane optimal untuk memisahkan data dalam ruang dimensi tinggi, sehingga akan sangat efektif dalam klasifikasi teks dengan jumlah fitur yang besar. Sementara dengan Random Forest mampu bekerja dengan mengombinasikan sejumlah pohon keputusan guna meningkatkan ketepatan akurasi serta meminimalkan resiko overfitting. Pada penelitian ini evaluasi dilakukan untuk menentukan algoritma yang lebih optimal dalam pengklasifikasian senti-ment positif dan negative, terutama setelah penerapan feature selection dengan Information Gain dan Chi-Square.

a. Support Vector Machine (SVM)
   SVM termasuk dalam supervised learning, mampu mengenali data dari label tertentu dari dataset yang sudah dilabeli sebelumnya. sering digunakan dalam klasifikasi teks, memiliki kelebihan yakni mampu menyelesaikan permasalahan dalam klasifikasi teks, SVM memiliki hasil yang lebih baik dalam opinion mining [9]. Konsep sederhana dari SVM ialah menemukan hyperlane optimal yang berfungsi untuk membedakan dua buah kelas pada ruang input [22] seperti terlihat pada ilustrasi Gambar 2.



Gambar 2. Ilustrasi konsep SVM

SVM akan mencari hyperlane optimal yang berfungsi untuk membedakan dua buah entitas pada ruang input [22], SVM mampu mengeneralisasi tinggi tanpa adanya persyaratan pengetahuan tambahan dan dengan tingkat dimensi yang tinggi. pelatihan model SVM linear akan dilakukan pemisahan data menjadi dua kelas dengan menggunakan fungsi berikut pada persamaan (3) (4) (5)

$$Minimize\ \frac{1}{2}||w||^2 + C\ \sum_{i=1}^{N} \max\ (0,1 - y_i(w^T x_i + b)) \quad (3)$$

Objective Function, memiliki tujuan dalam menemukan hyperlane optimal yang akan memisahkan kelas-kelas dalam data.

$$w^T x_i + b = 0 \quad (4)$$

Apabila $w^T x_i + b > 0$ sampel x diprediksi masuk kelas positif, jika $w^T x_i + b < 0$ sampel x diprediksi masuk kelas negatif

 ISSN:2302-8149
e-ISSN:2540-9719

$$Margin = \frac{2}{||w||} \quad (5)$$

Jarak antara hyperlane dengan sampel yang paling dekat dari masing-masing kelas. SVM memiliki tujuan yakni memaksimalkan margin.

b. Random Forest

Metode dalam machine learning yang menggabungkan sejumlah pohon keputusan (decision trees) dalam peningkatan akurasi prediksi [23]. Dengan begitu random forest mampu menghasilkan model yang lebih stabil dan akurat. Proses bagging dalam random forest untuk mempertimbangkan hasil prediksi akhir dengan jumlah mayoritas [24]. Random forest membangun set pelatihan yang berbeda untuk meningkatkan perbedaan model klasifikasi, yang akan mampu meningkatkan pula prediksi ekstraplorasi dari klasifikasi gabungan mode. Berikut persamaan yang digunakan pada (6)

$$H(x) = \arg\max \sum_{i=1}^{k} I(h_i(x) = Y) \quad (6)$$

Dimana H adalam model klasifikasi gabungan, h_i adalah model klasifikasi dari pohon keputusan tunggal , Y adalah output kelas, dan I adalah fungsi indikatif, persamaan i=1 mengilustrasikan penggunaan data terbanyak dalam menentukan klasifikasi akhir [25].

3.6 Evaluasi Model

Merupakan tahapan yang dilaksanakan untuk menilai seberapa baik performa dari model algoritma yang digunakan dalam penelitian [26]. Confusion matriks merupakan alat evaluasi yang umum digunakan dalam klasifikasi data mining, yang mana akan memberikan Gambaran secara menyeluruh mengenai prediksi model yang telah dibuat [27]. Didalam confusion matriks memiliki persamaan lanjutan yang digunakan dalam perhitungan evaluasi, Akurasi yakni Penilaian akurasi ini membantu untuk mempertimbangkan Langkah-langkah yang dilakukan dan Langkah apa saja yang diabaikan [28]. Presisi Merupakan rasio prediksi perbandingan hasil jawaban benar positif dengan seluruh hasil prediksi positif [28], Dengan mengetahui nilai presisi, ini akan dapat terlihat apakah model yang digunakan sudah tepat dan berhasil melakukan klasifikasi yang lebih terfokus pada kelas positif dan mengurangi kesalahan false negatif. Kemudian Recall rasio perbandingan antara true positif dengan data keseluruhan yang true positif. Recall akan memberitau berapa banyak kasus positif telah terprediksi dengan cocok pada model, ini berguna jika false negatif mengalahkan false positive [29], Recall yang memiliki nilai tinggi bermakna kasus positif (TP+FN) akan positif dan diberi label positif (TP). F1 Score evaluasi yang mengkolaborasikan keseimbangan antara presisi dan recall, dengan menghasilkan informasi mengenai seberapa sukses dan baik nya model yang digunakan dalam kombinasi presisi dan recall.

4 Hasil dan Pembahasan

Pendekatan klasifikasi yang diterapkan dalam penelitian ini adalah Support Vector Machine (SVM) dan Random Forest (RF). Dengan jumlah dataset yang digunakan yakni 5.472 data ulasan pengguna terhadap aplikasi Bank Saqu, dengan 2 fitur utama yakni content sebagai teks ulasan dan score sebagai label sentimen. Feature Selection dengan menggunakan metode Information Gain dan Chi-Square diterapkan dalam pemilihan kata-kata yang paling berperngaruh dalam klasifikasi sentimen. kerangka kerja penelitian yang telah dijelaskan pada bagian sebelumnya menghasilkan hasil analisis sebagai berikut.

4.1 Pengambilan Dataset

Penggunaan data dalam penelitian akan dikumpulkan dari ulasan Playstore terkait aplikasi Bank Saqu. Data dikumpulkan dengan menerapkan metode ‘web scraping’. Dalam pengumpulan data ini menggunakan teknik pemrograman untuk mengekstraksi informasi yang didapatkan dari halaman web play store. Proses yang dilakukan adalah mengirim permintaan HTTP ke halaman web Play Store Aplikasi Bank Saqu menggunakan Pustaka request, kemudian setelah berhasil dilakukan data HTML yang diterima akan diproses menggunakan Pustaka, pada penelitian ini menggunakan Pustaka ‘google-play-scraper’. Hasil crawling data di tunjukan pada Gambar 3.

 ISSN:2302-8149
e-ISSN:2540-9719

	reviewId	userName	userImage	content	score	thumbsUpCount	reviewCreatedVersion	at	replyContent	repliedAt	appVersion
0	d71c6824-aadd-4e33-bb1a-55bed3858cbd	Pengguna Google	https://play-lh.googleusercontent.com/EGemol2N...	Saya tarik saldo saya, kaga jelas promo nya Ud...	1	0	24.10.3	2024-11-18 06:36:06	Hai Warga Bank Saqu! Terima kasih atas ulasann...	2024-11-19 01:05:14	24.10.3
1	a4ee4a70-b118-43d4-b1fc-00aef1ca5cda	Pengguna Google	https://play-lh.googleusercontent.com/EGemol2N...	aplikasi simpel mudah d pahami	5	0	24.10.3	2024-11-18 05:25:06	Hai Warga Bank Saqu! Terima kasih atas rating ...	2024-11-19 00:55:33	24.10.3
2	166be4e9-8846-459e-a499-cdde082cfeb8	Pengguna Google	https://play-lh.googleusercontent.com/EGemol2N...	bagus	4	0	24.10.3	2024-11-18 05:00:27	Hai Warga Bank Saqu! Terima kasih atas dukunga...	2024-11-19 00:55:58	24.10.3
3	f70ac660-8a9b-4b78-9ab1-138fe42f392b	Pengguna Google	https://play-lh.googleusercontent.com/EGemol2N...	amanah	5	0	None	2024-11-18 04:43:01	Hai Warga Bank Saqu! Terima kasih atas ulasan ...	2024-11-18 08:55:38	None
4	f3d6ee94-2774-45f9-9857-e075e489356f	Pengguna Google	https://play-lh.googleusercontent.com/EGemol2N...	saya suka rekening ini...	5	0	24.10.3	2024-11-18 04:13:35	Hai Warga Bank Saqu! Terima kasih banyak atas ...	2024-11-18 08:14:03	24.10.3

Gambar 3. Hasil crwaling data

4.2 Pre-Processing Data

Dalam penelitian ini pre-processing data menjadi aspek yang esensial untuk mengurangi noise pada data teks ulasan Bank Saqu. Pada Gambar 4 merupakan hasil dari cleaning data.

	content	casefolding	textnormalize	stopwordremoval	stemming
0	Saya tarik saldo saya, kaga jelas promo nya Ud...	saya tarik saldo saya kaga jelas promo nya ud...	saya tarik saldo saya tidak jelas promo nya su...	tarik saldo promo nya beli bayar normal	tarik saldo promo nya beli bayar normal
1	aplikasi simpel mudah d pahami	aplikasi simpel mudah d pahami	aplikasi simpel mudah d pahami	aplikasi simpel mudah d pahami	aplikasi simpel mudah d paham
2	bagus	bagus	bagus	bagus	bagus
3	amanah	amanah	amanah	amanah	amanah
4	saya suka rekening ini...	saya suka rekening ini	saya suka rekening ini	suka rekening	suka rekening

Gambar 4. Hasil pre-processing data

Dengan memastikan bahwa model dapat menangkap pola sentimen dengan lebih akurat. Pemilihan strategi pre-processing yang tepat berperan penting pada nilai akurasi [30]. Teknik stopword removal dan stemming akan membantu dalam menghilangkan kata-kata yang tidak relevan serta menyederhanakan bentuk kata agar lebih seragam [31]

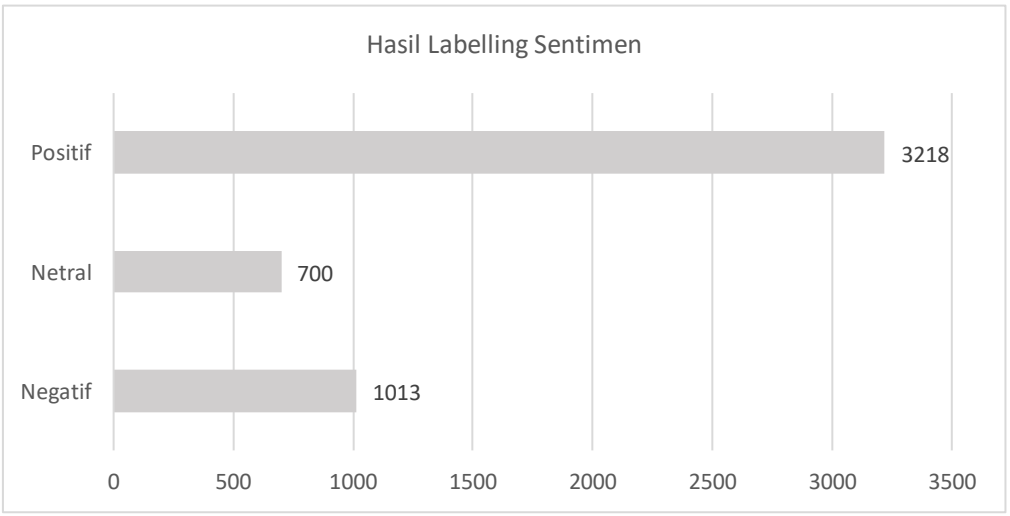
4.3 Labelling

Pada proses labeling menggunakan pendekatan lexicon. Pendekatan ini menjadi salah satu peran penting dalam mengidentifikasi opini berdasarkan daftar kaya yang telah diberikan nilai sentimen. metode ini akan sangat efektif dalam menganalisis sentimen dengan menggunakan kamus kata [32]. Hasi pelabelan yakni seperti pada Tabel 5.

Tabel 5. Label sentimen

Tipe	Nilai Tipe
Positif	>0
Netral	0
Negatif	>0

Perhitungan skor sentimen dengan membuat fungsi yang akan mengkonersi teks menjadi huruf kecil, memecah menjad kata-kata dan menghitung total bobot berdasarkan kemunculan kata dalam kamus positif negatif. Bobot kemudian dikategorikan menjadi 3 label sentimen dari hasil pemrosesan data labeling dengan menggunakan lexicon ini , didapati hasil pada Gambar 5.



Gambar 5. Grafik hasil labelling

 ISSN:2302-8149
e-ISSN:2540-9719

Hasil labeling sentimen, sesuai pada Gambar 5, mendapati jumlah dengan sentimen positif sebanyak 3218, sentimen netral sebanyak 700, dan untuk sentimen negatif sebanyak 1013.

4.4 Wordcloud

Visualisasi dilakukan berdasarkan analisis sentimen teks dengan membagi menjadi dua jenis teks yakni sentimen positif dan sentimen negatif yang ditampilakan. Visualisasi ini akan membantu dalam pemahaman kata-kata yang memiliki frekuensi tertinggi untuk muncul pada masing-masing kategori. Berikut hasil dari visualisasi pada Gambar 6.

Gambar 6. Visualisasi wordcloud

Hasil dari visualisasi wordcloud menunjukkan sentimen positif didominasi oleh kata “bagus”, “ mudah” , “mantap”, “bantu”, “cepat”, yang menandakan kepuasan pengguna terhadap kemudahaan penggunaan dan kecepatan transaksi pada aplikasi. Sebaliknya, sentimen negatif mengandung kata “masuk” , “ daftar” , “data”, “ gagal” , “salah”, yang menunjukkan kendala utama terkait login, pendaftaran, serta proses verifikasi akun.

4.5 Feature Selection

Pada penelitian ini menggunakan tiga pendekatan utama yakni feature selection dengan Information Gain, Chi-Square dan kombinasi anatara keduanya. Langkah awal yang dilakukan yakni mengkonversi teks mentah menjadi representasi numerik dengan TF-IDF yang akan menghitung bobot tiap kata berdasarkan frekuensi nya dalam dokumen dibandingan dengan keseluruhan korpus. Berikut adalah hasil pembobotan dengan TF-IDF pada Gambar 7.

	aaakkk	aaao	aaaqtg	aajuan	aamiien	aamiin	abal	abalan	abang	abank	...	ymynet	youth	youtube	youtubku	yukk	zaman	zelas	zero	zonk	zr
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Gambar 7. Hasil TF-IDF

Proses seleksi fitur dilakukan dengan memilih kata-kata yang memiliki kontribusi paling signifikan dalam analisis sentimen. Metode pertama yang digunakan adalah Chi-Square, yang mengidentikasi fitur dengan hubungan paling kuat terhadap kelas sentimen. Hanya fitur dengan nilai tertinggi yang dipertahankan, sementara fitur dengan kontribusi rendah dieliminasi untuk mengurangi dimensi data tanpa mengorbankan informasi penting. Selanjutnya, dilakukan seleksi fitur dengan Information Gain, yang mengukur seberapa besar suatu fitur memberikan informasi dalam menentukan kelas sentimen. Fitur dengan nilai Information Gain tertinggi dipertahankan, sementara yang kurang informatif dihilangkan agar model dapat bekerja lebih efisien.

Selain menggunakan kedua metode secara terpisah, dilakukan pendekatan kombinasi yang mengintegrasikan Chi-Square dan Information Gain. Fitur yang dipilih dari masing-masing metode kemudian dievaluasi ulang dengan perhitungan skor gabungan, sehingga hanya fitur dengan relevansi tertinggi yang digunakan dalam pemodelan. Pendekatan ini bertujuan untuk meningkatkan kualitas data yang digunakan, mengoptimalkan performa model serta mengurangi overfitting.

 ISSN:2302-8149
e-ISSN:2540-9719

4.6 Modelling

Setelah melalui tahap preprocessing data yang mencakup pembersihan dan ekstraksi fitur,tahapan selanjutnya ialah pemodelan dengan Support Vector Machine (SVM) dan Random Forest (RF) . Dalam tahapan modelling ini akan dilakukan dengan delapan skema pengujian. Skema pengujian ini dirancang dengan menggabungkan berbagai teknik seleksi fitur yakni Chi-Sqaure, Information Gain, dan keduanya dengan tujuan untuk mempertahankan fitur yang dianggap paling relevan dalam klasifikasi sentimen. Setiap skema akan diuji dengan menggunakan kedua algoritma tersebut untuk mengurkur seberapa baik mereka dapat menghasilkan klasifikasi sentimen dari dataset yang digunakan. Kemudian hasil dari tiap skema akan dibandingkan berdasarkan metrik evaluasi dengan akurasi, presisi, recall, dan F-1 Score untuk menentukan kombinasi seleksi fitur dan algoritma dengan performa terbaik.

1. **Support Vector Machine Tanpa Penggunan Feature Selection**
   Pada skema pengujian pertama yakni menguji dataset dengan algoritma Support Vector Machine (SVM) dan tidak menggunakan penambahan feature selection dalam datesetnya. Hasil dari pengujian yang telah dilakukan ini menunjukkan bahwa model memiliki tingkat akurasi sebesar 93%, seperti yang di tunjukkan pada Gambar 8

	precision	recall	f1-score	support
negatif	0.89	0.79	0.84	183
positif	0.94	0.97	0.96	664
accuracy			0.93	847
macro avg	0.92	0.88	0.90	847
weighted avg	0.93	0.93	0.93	847

Gambar 8. Hasil pengujian skema 1

Untuk sentimen negatif memiliki presisi 89% , recall 79%, dan f1-score 84%, yang menunjukkan bahwa model cukup baik dalam mengenali teks dengan sentimen negatif meskipun pada bagian recall memiliki nilai yang lebih rendah yakni 79% yang mana masih menunjukkan beberapa sampel negatif masih salah diklasifikasikan sebagai positif. Kemudian untuk sentimen positif performa model yang ditunjukkan dengan nilai presisi 94%, recall 97% dan f1-score 96% yang berarti model sangat akurat dalam mengidentifikasi teks sentimen positif.

2. **Support Vector Machine menggunakan Feature Selection Chi-Square**
   Fitur yang digunakan telah diseleksi dengan Chi-Square untuk memilih fitur yang paling relevan sehingga model bisa bekerja dengan optimal namun tetap informatif. Berdasarkan pengujian yang dilakukan menghasilkan akurasi sebesar 93%. Sebagaimana yang diperlihatkan pada Gambar 9.

	precision	recall	f1-score	support
negatif	0.81	0.86	0.84	183
positif	0.96	0.95	0.95	664
accuracy			0.93	847
macro avg	0.89	0.90	0.90	847
weighted avg	0.93	0.93	0.93	847

Gambar 9. Hasil pengujian skema 2

Sentimen negatif diperoleh nilai presisi 81%, recall 86% dan f1-score 84% yang mana model cukup baik dalam mengenali sentimen negatif. Sementara itu, dalam kelas sentimen positif menjunukkan performa model yang lebih tinggi yakni dengan nilai presisi 96%, recall 95%, dan f1-score 95%.

 ISSN:2302-8149
e-ISSN:2540-9719

3. **Support Vector Machine menggunakan Feature Selection Information Gain**
Dilakukan dengan algoritma Support Vector Machine dan menggunakan feature selection Infromation Gain. Hasil dari pemodelan menunjukkan performa cukup baik tetapi menhasilkan nilai yang lebih rendah dibandingkan dengan metode seleksi fitur Chi-Square. Berdasarkan hasil akurasi menunjukkan nilai 90% dari total uji . Sebagaimana yang diperlihatkan pada Gambar 10.

	precision	recall	f1-score	support
negatif	0.77	0.75	0.76	183
positif	0.93	0.94	0.93	664
accuracy			0.90	847
macro avg	0.85	0.84	0.85	847
weighted avg	0.90	0.90	0.90	847

Gambar 10. Hasil pengujian skema 3

Untuk kelas sentimen negatif memiliki nilai presisi 77% , recall 75%, dan f1-score 76% yang menunjukkan bahwa model memiliki sedikit kesulitan dalam pengidentifikasian sentimen negatif dengan baik. Sementara untuk kelas sentimen positif memiliki performa yang lebih tinggi dengan presisi 93% , recall 94%, dan f1-score 93%

4. **Support Vector Machine menggunakan Feature Selection Information Gain dan Chi-Square**
Pengujian dengan skema penggabungan antara dua feature selection memiliki performa yang baik dengan nilai akurasi sebesar 93%, yang mana model mampu mengklasifikasikan 93% dari total uji. Berikut detail pada Gambar 11.

	precision	recall	f1-score	support
negatif	0.81	0.86	0.84	183
positif	0.96	0.95	0.95	664
accuracy			0.93	847
macro avg	0.89	0.90	0.90	847
weighted avg	0.93	0.93	0.93	847

Gambar 11. Hasil pengujian skema 4

Secara spesifik untuk kelas sentimen negatif memiliki nilai presisi 81%, recall 86% , dan f1-score 84%, yang mana nilai ini lebih baik dari penggunaan Information Gain secara terpisah. Sedangkan untuk kelas positif memiliki nilai presisi 96%, recall 95%, dan f1-score 95%, yang juga nilainya lebih baik dari nilai dengan Information Gain saja.

5. **Random Forest Tanpa Penggunan Feature Selection**
Pengujian dilakukan dengan model algoritma yang berbeda, yakni menggunakan Random Forest (RF). Pengujian ini dilakukan tanpa menggunakan feature selection, dan hanya menguji data dengan model Random Forest saja Berikut pada Gambar 12.

	precision	recall	f1-score	support
negatif	0.76	0.82	0.79	183
positif	0.95	0.93	0.94	664
accuracy			0.90	847
macro avg	0.85	0.87	0.86	847
weighted avg	0.91	0.90	0.91	847

Gambar 12. Hasil pengujian skema 5

 ISSN:2302-8149
e-ISSN:2540-9719

Dalam pengujian dengan RF tanpa seleksi fitur, model menghasilkan nilai akurasi 90%, dengan spesifik untuk kelas positif memiliki nilai presisi 95% , recall 93%, dan f1-score 94%. Sedangkan kelas negatif memiliki nilai presisi 76% , recall 82%, dan f1-score 79%.

6. **Random Forest menggunakan Feature Selection Chi-Square**

Pengujiam Random Forest dengan Chi-Square memiliki hasil akurasi 91% yang menunjukkan kinerja yang cukup baik dalam pengklasifikasian.. Berikut adalah detail pada Gambar 13.

	precision	recall	f1-score	support
negatif	0.79	0.81	0.80	183
positif	0.95	0.94	0.94	664
accuracy			0.91	847
macro avg	0.87	0.87	0.87	847
weighted avg	0.91	0.91	0.91	847

Gambar 13. Hasil pengujian skema 6

Kelas positif memiliki nilai presisi 95%, recal 94% dan f1-score 94% yang mana nilai ini lebih tinggi dari kelas negatif dengan nilai presisi 79% , recall 81% , dan f1-score 80%. Nilai recall lebih tinggi pada kelas positif berarti model mampu mengidentifikasi lebih banyak instance positif dengan benar

7. **Random Forest menggunakan Feature Selection Information Gain**

Hasil pengujian Random Forest dengan Infromation Gain memiliki nilai akurasi 88% yang mana nilai ini lebih rendah dibandingkan dengan dengan Ch-Square. Berikut detail pada Gambar 14.

	precision	recall	f1-score	support
negatif	0.69	0.77	0.73	183
positif	0.93	0.91	0.92	664
accuracy			0.88	847
macro avg	0.81	0.84	0.82	847
weighted avg	0.88	0.88	0.88	847

Gambar 14. Hasil pengujian skema 7

Secara spesifik, kelas positif memiliki nilai presisi 93%, recall 91%, dan f1-score 92% yang lebih tinggi dari kelas negatif dengan nilai presisi 69%, recall 77%, dan f1-score 73%.

8. **Random Forest menggunakan Feature Selection Information Gain dan Chi-Square**

Pengujian skema ini yang mana menggabungkan antara Random Forest dengan kombinasi Chi-Square dan Infromation Gain menghasilkan nilai akurasi 90% yang mana nilai ini lebih baik daripada hanya menggunakan Information Gain(88%), dan lebih rendah sedikit dibandingkan dengan RF menggunakan Chi-Square (91%). Berikut detail pada Gambar 15.

	precision	recall	f1-score	support
negatif	0.76	0.81	0.78	183
positif	0.95	0.93	0.94	664
accuracy			0.90	847
macro avg	0.85	0.87	0.86	847
weighted avg	0.90	0.90	0.90	847

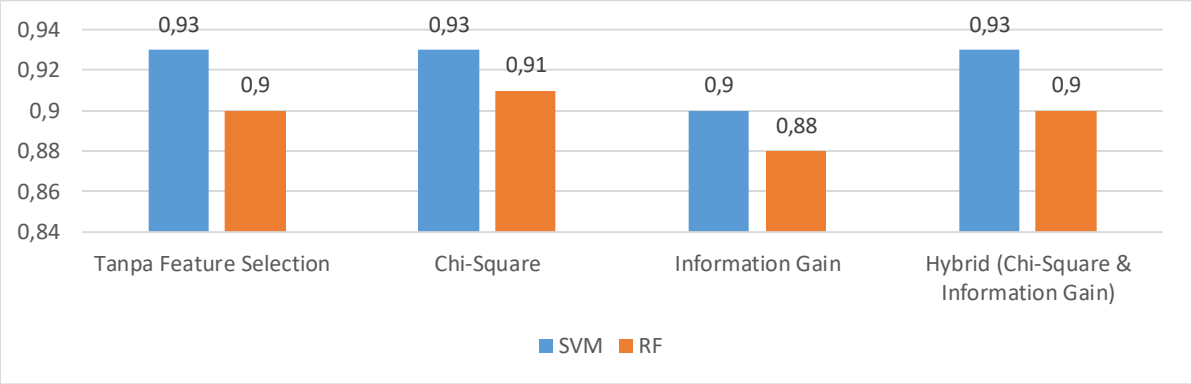
Gambar 15. Hasil pengujian skema 8

 ISSN:2302-8149
e-ISSN:2540-9719

Secara spesifik, kelas positif memiliki nilai presisi 95% , recall 93% , dan f1-score 94% yang lebih tinggi dibandingkan kelas negatif yang memiliki niali presisi 76% , recall 81%, dan f1-score 78%.

4.7 Hasil Perbandingan

Berdasarkan hasil pengujian delapan skema yang melibatkan algoritma Support Vector Machine (SVM) dan Ran-dom Forest (RF) dengan serta tanpa feature selection (Chi-Square, Infromation Gain, dan Hybrid). Yang kemudian akan membandinkan skema mana yang akan menunjukkan performa terbaik berdasarkan metrik evaluasi yang su-dah dihitung dengan berbagai komponen didalamnya seperti presisi, recall, dan f1-score. Gambar 16 berikut akan menunjukkan grafik perbandingan anatara ke-delapan skema tersebut.



Gambar 16. Grafik perbandingan hasil

Dari hasil keseluruhan ini, didapati beberapa temuan penting. SVM secara umum menunjukkan hasil performa yang lebih tinggi dibandingkan Random Forest, terutama dengan skema feature selection dengan metode Chi-Square serta Hybrid. Model SVM dengan Chi-Squre dan hybrid memili nilai akurasi tertinggi sebesar 93%, sementara Random Forest dengan skema terbaiknya , yakni menggunakan Chi-Square hanya mencapai akurasi 91%. Dari segi feature selection, penggunaan Chi-Square cenderung meningkatkan performa baik pada SVM maupun Random Forest dibandingakan dengan Information Gain. Ini telihat dari hasil pengujian SVM maupun RF dengan Chi-Square memiliki hasil akurasi yang lebih tinggi dari Information Gain. Dan secara keseluruhan, SVM lebih unggul dari Random Forest dilihat dari hasil pengujiannya.

5 Kesimpulan

Penelitian ini berfokus pada optimalisasi analisis sentimen terhadap data Bank Saqu dengan membandingkan performa model algoritma Support Vector Machine dan Random Forest, baik tanpa atau menggunakan feature selection yakni Chi-Square, Infromation Gain, dan Hybrid. Dari delapan skema pengujian yang telah dilakukan, analisis hasil menunjukkan SVM memiliki hasil yang lebih baik dari Random Forest, terutama jika model ini dikombinasikan dengan feature selection Chi-Square, yang mana model menghasilkan nilai akurasi tertinggi dengan 93%. Ini mengindikasikan bahwa SVM lebih efisien dalam penanganan karakteristik data teks, sementara feature selection terbukti mampu meningkatkan performa model dengan penyaringan fitur yang relevan. Disisi lain Random Forest menunjukkan nilai performa terbaiknya dengan kolaborasi Chi-Sqaure dan memiliki nilai akurasi 91%. Baik dari SVM ataupun RF, dari hasil dapat disimpulkan bahwa penggunaan feature selection Chi-Square lebih efektif dibandingkan dengan Information Gain dalam hal peningkatan akurasi model karena fitur ini mampu lebih optimal dalam pemilihan fitur yang memiliki signifikansi tinggi dalam penentuan sentimen. Implikasi penting pada penelitian ini, yakni penggunaan seleksi fitur yang tepat, terutama Chi-Square yang memiliki nilai akurasi dengan SVM mencapai 93%, menjadikannnya referensi bagi riset serupa. Seleksi fitur juga meningkatkan efisiensi komputasi dengan mengurangi fitur yang tidak relevan, sehingga akan mempercepat proses dalam pelatihan model. Kemudian perbandingan algoritma menunjukkan bahwa SVM lebih unggul dibandingkan Random Forest dalam klasifikasi teks, sehingga memberikan panduan bagi pemilihan model yang lebih optimal. Penelitian ini membuka peluang eksplorasi lebih lanjut untuk

 ISSN:2302-8149
e-ISSN:2540-9719

meningkatkan akurasi serta pemahaman konteks pada analisis sentimen. Penelitian ini memberikan wawasan bahwa kombinasi machine learning dengan feature selection yang tepat dapat meningkatkan hasil akurasi analisis sentimen yang bisa diimplementasikan lebih luas dalam berbagai sektor, serta untuk pemahaman opini pelanggan secara lebih akurat. Selain itu pengembangan lebih lanjut dapat dilakukan dengan penggunaan dataset yang lebih besar serta pengeksplorasian teknik feature selection lainnya.

Referensi

[1] A. C. Situru, *Pengaruh Sikap terhadap Pemilihan melalui Minat Penggunaan Fintech pada Generasi Milenial Kota Makassar*. 2021.

[2] Alun Sujjadaa, Somantri, Juwita Nurfazri Novianti, dan Indra Griha Tofik Isa, "Analisis Sentimen terhadap Review Bank Digital pada Google Play Store menggunakan Metode Support Vector Machine (SVM)," *J. Rekayasa Teknol. Nusa Putra*, vol. 9, no. 2, pp. 122–135, 2023. https://doi.org/10.52005/rekayasa.v9i2.345

[3] Y. H. Hoang, V. M. Ngo, and N. Bich Vu, "Central Bank Digital Currency: A Systematic Literature Review using Text Mining Approach," *Res. Int. Bus. Financ.*, vol. 64, no. May 2022, p. 101889, 2023.

[4] A. Kumar, S. Chakraborty, and P. K. Bala, "Text Mining Approach to Explore Determinants of Grocery Mobile App Satisfaction using Online Customer Reviews," *J. Retail. Consum. Serv.*, vol. 73, no. June 2022, p. 103363, 2023.

[5] S. Lavianto and I. W. D. P. Adnyana, "Analisa Sentimen terhadap Review Layanan Fintech dengan Metode Naive Bayes Classifier," *J. Teknol. Inf. dan Komput.*, vol. 8, no. 1, pp. 43–51, 2022.

[6] M. I. Fikri, T. S. Sabrila, and Y. Azhar, "Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter," *Smatika J.*, vol. 10, no. 02, pp. 71–76, 2020.

[7] I. S. K. Idris, Y. A. Mustofa, and I. A. Salihi, "Analisis Sentimen terhadap Penggunaan Aplikasi Shopee menggunakan Algoritma Support Vector Machine (SVM)," *Jambura J. Electr. Electron. Eng.*, vol. 5, no. 1, pp. 32–35, 2023.

[8] R. A. S and Y. Yamasari, "Eksplorasi Fitur Seleksi pada SVM dan Random Forest dalam Analisis Sentimen Aplikasi GoPay," vol. 06, pp. 55–65, 2024.

[9] A. Salsabila, J. J. Sihombing, and R. I. Sitorus, "Implementasi Algoritma Support Vector Machine Untuk Analisis Sentimen Aplikasi OLX di Playstore," *J. Informatics Data Sci.*, vol. 1, no. 2, 2022.

[10] T. Wahyuningsih, D. Manongga, I. Sembiring, and S. Wijono, "Comparison of Effectiveness of Logistic Regression, Naive Bayes, and Random Forest Algorithms in Predicting Student Arguments," *Procedia Comput. Sci.*, vol. 234, pp. 349–356, 2024.

[11] S. Alfarizi and E. Fitriani, "Analisis Sentimen Kendaraan Listrik menggunakan Algoritma Naive Bayes dengan Seleksi Fitur Information Gain dan Particle Swarm Optimization," *Indones. J. Softw. Eng.*, vol. 9, no. 1, pp. 19–27, 2023.

[12] B. Zhang, Z. Wang, H. Li, Z. Lei, J. Cheng, and S. Gao, "Information Gain-Based Multi-Objective Evolutionary Algorithm for Feature Selection," *Inf. Sci. (Ny).*, vol. 677, no. May, p. 120901, 2024.

[13] N. Sari, M. Jazman, T. K. Ahsyar, Syaifullah, and A. Marsal, "Penerapan Algoritma Klasifikasi Naive Bayes dan Support Vector Machine untuk Analisis Sentimen Cyberbullying Bilingual di Aplikasi X Implementation of Naive Bayes and Support Vector Machine Classification Algorithms for Sentiment Analysis of Bilingual Cyb," vol. 14, pp. 211–224, 2025.

[14] W. Utomo, M. I. Komputer, F. T. Informasi, U. B. Luhur, P. Utara, and J. Selatan, "Optimalisasi Metode Support Vector Machine (SVM) berbasis Optimized Weight Evolutionary dalam Penentuan Sentimen Komentar Optimized Weight Evolutionary - based Support Vector Machine (SVM) Optimization for Comment Sentiment," vol. 14, pp. 147–171, 2025.

[15] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations," *Organ. Res. Methods*,

 ISSN:2302-8149
e-ISSN:2540-9719

vol. 25, no. 1, pp. 114–146, 2022.
[16] Rachmawati Oktaria Mardiyanto, K. Kusrini, dan Ferry Wahyu Wibowo, "Analisis Sentimen Pengguna Aplikasi Bank Syariah Indonesia dengan menggunakan Algoritma Support Vector Machine (SVM)," *artikel jurnal Tek. Teknol. Inf. dan Multimed.*, vol. 4, no. 1, pp. 9–15, 2023. DOI tidak tersedia.
[17] J. Andrade-Hoz, J. M. Alcaraz-Calero, dan Q. Wang, "NetLabeller: Architecture with Data Extraction and Labelling Framework for Beyond 5G Networks," *artikel jurnal J. Commun. Networks*, vol. 26, no. 1, pp. 80–98, 2024. https://doi.org/10.23919/JCN.2024.000006
[18] B. Valarmathi, N. S. Gupta, V. Karthick, T. Chellatamilan, K. Santhi, and D. Chalicheemala, "Sentiment Analysis of Covid-19 Twitter Data using Deep Learning Algorithm," *Procedia Comput. Sci.*, vol. 235, no. 2023, pp. 3397–3407, 2024.
[19] K. Liu, Z. Deng, and M. Zhang, "Research on Capability Maturity Evaluation Model of Power Grid Data Management," *Procedia Comput. Sci.*, vol. 228, pp. 1030–1037, 2023.
[20] E. Hokijuliandy, H. Napitupulu, and Firdaniza, "Application of SVM and Chi-Square Feature Selection for Sentiment Analysis of Indonesia's National Health Insurance Mobile Application," *Mathematics*, vol. 11, no. 17, 2023.
[21] T. Ernayanti, M. Mustafid, A. Rusgiyono, and A. R. Hakim, "Penggunaan Seleksi Fitur Chi-Square dan Algoritma Multinomial Naïve Bayes untuk Analisis Sentimen Pelangggan Tokopedia," *J. Gaussian*, vol. 11, no. 4, pp. 562–571, 2023.
[22] Z. Alhaq, A. Mustopa, S. Mulyatun, and J. D. Santoso, "Optimasi Algoritma Support Vector Machine untuk Analisis Sentimen pada Ulasan Produk Tokopedia menggunakan PSO," *Media Inform.*, vol. 20, no. 2, pp. 97–108, 2021.
[23] D. K. Anuradha, D. B. Mallik, and D. M. V. Krishna, "Cucconi Feature Extracted Random Decision Forest Classification for Efficient Sentiment Analysis," *Migr. Lett.*, vol. 20, no. S13, pp. 520–533, 2023.
[24] S. Wahyuni Kalumbang, "Perbandingan Regresi Logistik, Klasifikasi Naive Bayes, dan Random Forest (Comparison the Logistic Regression, Naive Bayes Classification, and Random Forest)," *J. Mat. Thales*, vol. 03, no. 02, pp. 1–13, 2021.
[25] D. Yuan, J. Huang, X. Yang, and J. Cui, "Improved Random Forest Classification Approach based on Hybrid Clustering Selection," *Proc. - 2020 Chinese Autom. Congr. CAC 2020*, pp. 1559–1563, 2020.
[26] A. I. Tanggraeni and M. N. N. Sitokdana, "Analisis Sentimen Aplikasi E-Government pada Google Play menggunakan Algoritma Naïve Bayes," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 9, no. 2, pp. 785–795, 2022.
[27] H. Azis, F. Tangguh Admojo, and E. Susanti, "Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah," *Techno.Com*, vol. 19, no. 3, pp. 286–294, 2020.
[28] K. Riehl, M. Neunteufel, and M. Hemberg, "Hierarchical Confusion Matrix for Classification Performance Evaluation," no. August, 2023.
[29] D. Chicco and G. Jurman, "The Advantages of the Matthews Correlation Coefficient (MCC) Over F1 Score and Accuracy in Binary Classification Evaluation," pp. 1–13, 2020.
[30] M. Siino, I. Tinnirello, and M. La Cascia, "Is Text Preprocessing Still Worth the Time ? A Comparative Survey on the Influence of Popular Preprocessing Methods on Transformers and Traditional Classifiers," *Inf. Syst.*, vol. 121, no. July 2023, p. 102342, 2024.
[31] S. Khairunnisa and S. Al Faraby, "Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi," vol. 5, no. April, pp. 406–414, 2021.
[32] O. Alsemaree, A. S. Alam, S. S. Gill, and S. Uhlig, "Heliyon an Analysis of Customer Perception using Lexicon-based Sentiment Analysis of Arabic Texts Framework," *Heliyon*, vol. 10, no. 11, p. e30320, 2024.