Eksplorasi Fitur Seleksi pada SVM dan *Random Forest* dalam Analisis Sentimen Aplikasi GoPay

Rahma Aziz S.[1], Yuni Yamasari[2]

[1,2] Program Studi S1 Teknik Informatika, Universitas Negeri Surabaya
[1]rahma.20025@mhs.unesa.ac.id
[2]yuniyamasari@unesa.ac.id

Abstrak*—Aplikasi GoPay merupakan aplikasi yang cukup populer saat ini. Hal ini ditunjukkan dengan banyaknya ulasan dari penggunanya yaitu sebanyak 44 ribu lebih. Ulasan ini memberikan informasi penting mengenai kepuasan pengguna terhadap layanan yang ditawarkan. Namun, dengan volume data ulasan yang sangat besar, pengolahan secara manual akan menyulitkan, sehingga dibutuhkan metode otomatis untuk mengkategorikan ulasan pengguna berdasarkan sentimennya. Penelitian ini bertujuan untuk memberikan solusi dengan melakukan analisis sentimen terhadap ulasan pengguna aplikasi GoPay. Metode SVM dan *Random Forest* diterapkan sebagai metode klasifikasi dan *Chi-square* dan *Information Gain* sebagai fitur seleksi. Untuk teknik evaluasi, split data dan *K-Fold Cross Validation* diimplementasikan. Data yang digunakan adalah ulasan pengguna pada periode Juli 2023 hingga Februari 2024 sebanyak 1.000 data. Pada periode tersebut, aplikasi GoPay lebih banyak mendapatkan sentimen positif. Hasil pengujian memperlihatkan teknik split data menghasilkan performa yang lebih baik dibandingkan dengan *K-Fold Cross Validation*. Kemudian, fitur seleksi *Chi-square* tidak mampu meningkatkan performa pada model yang dibangun dengan SVM dan *Random Forest*. Sedangkan, fitur seleksi *Information Gain* hanya mampu meningkatkan performa model *Random Forest*. Metode SVM dengan kernel Sigmoid mencapai kinerja terbaik pada pengujian tanpa fitur seleksi. Nilai kinerja yang dicapai adalah akurasi 98%, presisi 98,09%, recall 98%, dan F1-score 98,01%. Sedangkan, metode *Random Forest* memiliki kinerja terbaik ketika dikombinasikan dengan *Information Gain* dengan nilai akurasi yang dihasilkan sebesar 97%, presisi 97,01%, recall 97%, dan F1-score 96,99%. Kinerja terbaik dari kedua metode tersebut dicapai dengan teknik evaluasi split data rasio 90:10. Hasil- hasil yang diperoleh ini menunjukkan bahwa metode SVM memiliki performa yang lebih baik dibandingkan *Random Forest* dengan perbedaan akurasi sebesar 1%, presisi 1,08%, recall 1%, dan F1-score 1,02%.

Kata Kunci*— GoPay, Analisis Sentimen, SVM, *Random Forest*, *Chi-square*, *Information Gain

I. PENDAHULUAN

Kemajuan teknologi membawa beragam kemudahan dalam berbagai aspek kehidupan, termasuk akses informasi, pendidikan, hingga ekonomi atau keuangan. Produk kombinasi antara teknologi dan keuangan merupakan produk *Financial Technology* (*FinTech*). Salah satu bentuk *FinTech* yaitu aplikasi *e-wallet* atau dompet digital yang muncul di kalangan masyarakat [1].

E-wallet memadukan layanan teknologi dan keuangan sehingga mengganti model bisnis yang semula konvensional menjadi bisnis modern. Dengan menggunakan *e-wallet*, seseorang dapat lebih mudah melakukan transaksi tanpa memerlukan dompet fisik [2].

Layanan *e-wallet* telah banyak tersedia di Indonesia, seperti OVO, Dana, GoPay, dan lainnya [3]. GoPay menduduki peringkat pertama sebagai *e-wallet* yang terbanyak digunakan di Indonesia dengan perolehan persentase sebesar 88% [4]. Faktor-faktor yang mempengaruhi kepuasan konsumen untuk menggunakan GoPay didominasi oleh kepercayaan, manfaat, dan kemudahan [5].

Sebelumnya, GoPay merupakan salah satu fitur dalam aplikasi Gojek. Hingga pada tahun 2023, pihak GoTo Financial menciptakan sebuah inovasi baru dengan membuat GoPay menjadi aplikasi tersendiri [6]. Sasaran utama dari aplikasi ini adalah calon pengguna yang belum pernah menggunakan Gojek atau Tokopedia. Meskipun aplikasi GoPay telah diluncurkan, layanan GoPay yang saat ini masih terikat dalam aplikasi Gojek masih terus berjalan dan tidak digantikan.

Aplikasi GoPay secara resmi diluncurkan melalui *platform* Google Play Store. Google Play Store memberikan fitur dimana pengguna dapat memberikan ulasan terhadap aplikasi yang diunduh. Bagi penyedia jasa, kepuasan pelanggan bergantung pada kualitas layanan yang diterima oleh pelanggan [7]. Dalam rangka mengetahui tingkat kepuasan pengguna aplikasi terhadap layanan yang diberikan, penting bagi perusahaan dalam memahami ulasan-ulasan yang ditulis oleh pengguna di Google Play Store.

Hingga bulan November 2023, GoPay telah diunduh oleh 5 juta lebih pengguna, diulas oleh 44 ribu lebih pengguna, dan memperoleh rating 4,3 dari 5,0. Data ulasan tersebut akan memberikan manfaat yang signifikan ketika dianalisis dengan baik. Namun, dengan jumlah data ulasan yang cukup besar, pengolahan secara manual akan cukup sulit dilakukan. Diperlukan metode yang otomatis untuk mengkategorikan ulasan pengguna sesuai dengan kelasnya, salah satunya yaitu metode analisis sentimen.

Analisis sentimen yaitu proses otomatis untuk memahami, mengekstrak, dan mengolah data teks guna mengidentifikasi sentimen dalam sebuah kalimat [8]. Analisis ulasan pengguna aplikasi GoPay di Google Play Store diperlukan untuk menilai *feedback* pengguna terhadap layanan *e-wallet* GoPay. Analisis sentimen dapat dilakukan dengan menggunakan berbagai algoritma klasifikasi, diantaranya SVM, *Naïve Bayes*, *Random Forest*, *Decision Tree*, dan lain-lain.

Proses klasifikasi dalam analisis sentimen sangat dipengaruhi oleh fitur, seperti kata-kata atau frasa. Pada umumnya, karakteristik sentimen teks sangatlah kompleks sehingga penggunaan seluruh atribut akan mengurangi kinerja

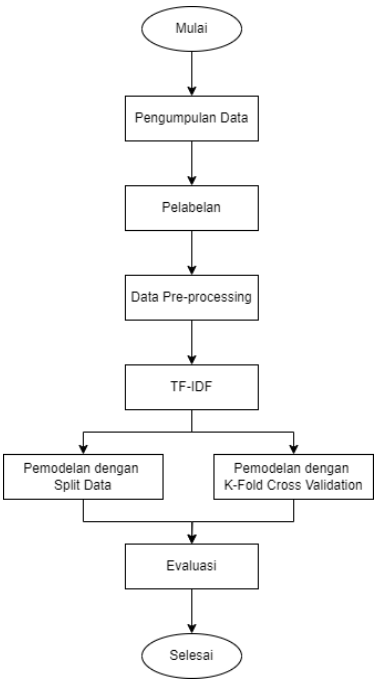
classifier, yang dapat memengaruhi tingkat akurasi [9]. Agar kinerja *classifier* tidak menurun, digunakan metode seleksi fitur untuk mengoptimalkan kinerjanya.

Seleksi fitur berperan untuk mengurangi subset fitur yang tidak memiliki pengaruh yang signifikan pada kelas [10]. Model analisis sentimen dapat menjadi lebih efektif dan akurat dengan memilih fitur seleksi yang relevan. Seleksi fitur juga membantu mengurangi dimensi data yang kompleks, sehingga proses analisis menjadi lebih cepat dan efisien. *Information Gain* dan *Chi-square* merupakan contoh metode yang digunakan untuk seleksi fitur.

Pada penelitian ini dilakukan perbandingan algoritma klasifikasi SVM dan *Random Forest* dalam mengklasifikasikan sentimen ulasan pengguna aplikasi GoPay di Google Play Store. Masing-masing algoritma klasifikasi dikombinasikan dengan seleksi fitur *Information Gain* dan *Chi-square* untuk dilakukan analisis mengenai pengaruh penggunaan seleksi fitur pada tiap model klasifikasi.

II. METODE PENELITIAN

Agar lebih mudah memahami proses pelaksanaan suatu penelitian, diperlukan pembuatan diagram alur penelitian. Diagram alur pada penelitian ini ditunjukkan pada Gbr. 1.



Gbr 1. Alur Penelitian

A. Pengumpulan Data

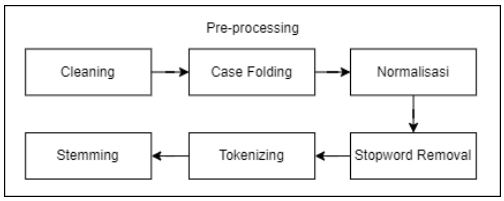
Data yang dikumpulkan adalah data ulasan pada Google Play Store yang dituliskan oleh pengguna aplikasi GoPay. Pengumpulan data dilakukan dengan menggunakan teknik *scraping*.

B. Pelabelan

Data ulasan dikategorikan ke dalam dua label, yaitu positif dan negatif. Pelabelan didasarkan pada *rating* yang diberikan pengguna saat memberikan ulasan. Ulasan dengan *rating* 1-3 diberikan label negatif dan ulasan dengan *rating* 4-5 diberikan label positif. Pelabelan awal dilakukan secara otomatis menggunakan fungsi Python, kemudian dilanjutkan validasi pelabelan secara manual untuk memastikan bahwa hasil pelabelan telah sesuai dengan isi ulasan.

C. Data Pre-processing

Pre-processing mengubah data tidak terstruktur menjadi data terstruktur untuk mempermudah pemrosesan lebih lanjut. Alur *data pre-processing* ditunjukkan pada Gbr 2.



Gbr 2. Alur Data Pre-processing

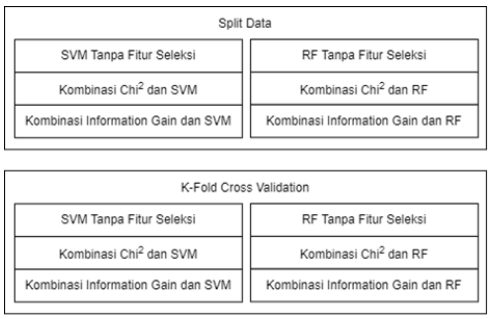
1) *Cleaning*, yaitu membersihkan data dari informasi yang tidak relevan, seperti angka, emotikon, tanda baca, spasi berlebih, dan lain-lain.
2) *Case Folding*, yaitu mengubah seluruh huruf dalam teks menjadi huruf kecil (*lowercase*).
3) Normalisasi, yaitu memperbaiki penulisan kata yang kurang sesuai dan mengubah kata tidak baku menjadi kata baku.
4) *Stopword Removal*, yaitu menghilangkan kata-kata yang tidak memberikan nilai penting.
5) *Tokenizing*, yaitu membagi teks menjadi unit yang lebih kecil yang disebut ‘token’.
6) *Stemming*, yaitu menghilangkan imbuhan pada kata hasil tokenisasi dan diubah menjadi kata dasar.

D. TF-IDF

Pembobotan TF-IDF (*Term Frequency-Inverse Document Frequency*) bertujuan untuk mengevaluasi seberapa penting suatu kata (*term*) dalam dokumen. *Term Frequency* (TF) mengukur seberapa sering *term* tertentu muncul dalam sebuah dokumen. *Inverse Document Frequency* (IDF) mengukur seberapa jarang sebuah *term* muncul di seluruh kumpulan dokumen.

E. Pemodelan

Pemodelan dilakukan menggunakan dua teknik pembagian data yaitu split data dan *K-Fold Cross Validation*. Dalam setiap teknik tersebut dilakukan pemodelan dengan fitur seleksi dan tanpa fitur seleksi.



Gbr 3. Alur Pemodelan

1) Split Data

Split data merupakan tahap membagi data dari hasil *pre-processing* menjadi data *training* dan data *testing* ke dalam perbandingan tertentu [11]. Data *training* adalah data yang digunakan untuk melatih model dalam memahami sentimen (positif dan negatif) dalam teks. Data *testing* adalah data yang digunakan untuk menguji kinerja model dalam memprediksi sentimen ulasan yang belum pernah dilihat sebelumnya [12]. Pada penelitian ini, ditentukan 5 rasio pembagian data yang berbeda yang ditunjukkan pada Tabel I.

TABEL I
PEMBAGIAN DATA LATIH DAN DATA UJI (SPLIT DATA)

Rasio	Data Latih	Data Uji
90 : 10	900 data	100 data
80 : 20	800 data	200 data
70 : 30	700 data	300 data
60 : 40	600 data	400 data
50 : 50	500 data	500 data

2) K-Fold Cross Validation

Metode *K-Fold Cross Validation* adalah teknik pengujian model yang membagi dataset menjadi *k*-bagian (*fold*). Konsep dari metode ini adalah membagi dataset menjadi *k*-bagian yang terdiri dari data latih dan data uji [13]. Di setiap bagian, urutan penggunaan data uji harus berbeda antara satu bagian dengan yang lainnya. Sisa data yang tidak digunakan untuk data uji akan digunakan sebagai data latih di setiap bagian. Pada penelitian ini, ditentukan 5 nilai *k* yang berbeda yang ditunjukkan pada Tabel II.

TABEL II
PEMBAGIAN DATA LATIH DAN DATA UJI (K-FOLD CROSS VALIDATION)

k	Data Latih	Data Uji
2	1 bagian	1 bagian
4	3 bagian	1 bagian
6	5 bagian	1 bagian
8	7 bagian	1 bagian
10	9 bagian	1 bagian

3) Information Gain

Untuk memperoleh nilai *Information Gain*, langkah pertama yang perlu dilakukan adalah dengan menghitung nilai *entropy*. *Entropy* dihitung untuk mengukur tingkat ketidakmurnian sekumpulan objek pada setiap cabang suatu atribut. Nilai *entropy* dihitung menggunakan rumus pada Persamaan (1) [14].

$$Entropy\ (c) = -\sum_{i=1}^{m} p(c_i) \log p(c_i) \quad (1)$$

Keterangan :
m : Jumlah nilai pada atribut target (jumlah kelas)
$p(c_i)$: Probabilitas fitur ke-i

Selanjutnya perhitungan nilai *Information Gain* dilakukan menggunakan rumus pada Persamaan (2).

$$IG\ (c,t) = Entropy(c) + \sum_{j \in value(t)}^{n} \frac{|c_j|}{|c|} Entropy(c_j) \quad (2)$$

Keterangan :
$Entropy(c)$: *Entropy* fitur *c* (sebelum pemisahan)
$Entropy(c_j)$: *Entropy* fitur *c* untuk *class t = j* (setelah pemisahan)
$value(t)$: Himpunan nilai yang mungkin untuk *class t*
n : Jumlah nilai-nilai yang mungkin untuk *class t*
$|c_j|$: Jumlah sampel *class* dengan nilai = *j*
$|c|$: Jumlah sampel untuk seluruh *class*

4) Chi-square

Fungsi *Chi-square* dalam fitur seleksi digunakan untuk menguji estimasi dan independensi dengan tujuan menghitung ketergantungan kelas pada fitur [15]. *Chi-square* terdiri dari 3 pengujian, yakni *observe frequency*, *expected frequency*, dan *test statistic* [16].

$$E_{ij} = \frac{total\ baris\ i \times total\ kolom\ j}{total\ data} \quad (3)$$

$$x^2 = \sum \frac{(O-E)^2}{E} \quad (4)$$

Keterangan :
x^2 : Nilai *Chi-square*
O : Frekuensi yang diamati (*Observed Frequency*)
E : Frekuensi yang diharapkan (*Expected Frequency*)

5) Support Vector Machine

SVM menggunakan *hyperplane* untuk mengklasifikasikan data [17]. Klasifikasi dilakukan dengan mencari *hyperplane* atau kernel maksimal yang memisahkan dua buah kelas. Pada penelitian ini, pemodelan dengan metode SVM diimplementasikan menggunakan 4 kernel, yaitu Linear, RBF (*Radial Basis Function*), Sigmoid, dan Polynomial.

a. Linear
$$K\left(x_i, x_j\right) = x_i^T x_j \quad (5)$$
b. Polynomial
$$K\left(x_i, x_j\right) = \left(\gamma\, x_i^T x_j + r\right)^p,\ \ \gamma > 0 \quad (6)$$
c. Radial Basis Function (RBF)
$$K\left(x_i, x_j\right) = \exp\left(-\gamma \left\|x_i - x_j\right\|^2\right),\ \ \gamma > 0 \quad (7)$$
d. Sigmoid
$$K\left(x_i, x_j\right) = \tanh\left(\gamma\, x_i^T x_j + r\right) \quad (8)$$

Keterangan :
x_i : Data latih
x_j : Data uji
p : Derajat *polynomial*
γ : *Scalling* parameter dari jarak *Euclidean*
r : Nilai konstanta

6) Random Forest

Pada konsep klasifikasinya, *Random Forest* menggabungkan beberapa pohon keputusan (*Decision Tree*), dimana masing-masing *tree* diberikan data secara acak dari sebagian data latih [18]. Pada penelitian ini, pemodelan dengan *Random Forest* diimplementasikan menggunakan 4 *n_estimators* berbeda, yaitu 100, 300, 500, dan 1.000.

Dengan input yang terdiri dari :
D = Dataset yang terdiri dari *d* baris
k = Angka dari jumlah *tree*

Pembentukan *Random Forest* dimulai dengan membuat sampel data D_i dari dataset D sebanyak d baris dengan melakukan pengembalian, lalu menggunakan sampel data D_i untuk membangun *k-tree* secara iteratif. Setiap *tree* dibangun menggunakan metode CART (*Classification and Regression Trees*), yang menentukan *node* berdasarkan *Information Gain* setelah menghitung *entropy*. Selain itu, *gini index* juga dihitung untuk setiap *tree*. Perhitungan *gini index* didefinisikan pada Persamaan (9).

$$Gini\ (A) = 1 - \sum_{i=1}^{n}(pi)^2 \quad (9)$$

Keterangan :
i : Kelas atribut
n : Jumlah kelas variabel Y
pi: Proporsi jumlah kelas dalam atribut i terhadap jumlah kelas n dalam atribut

Setelah semua *tree* dibangun, data uji digunakan untuk memprediksi keluaran klasifikasi berdasarkan aturan dari setiap *tree*. Hasil prediksi kemudian dikumpulkan dan target prediksi akhir ditentukan berdasarkan jumlah suara (*vote*) terbanyak dari seluruh *tree* dalam *Random Forest*.

F. Evaluasi

Tahap evaluasi bertujuan untuk mengetahui seberapa baik performa algoritma klasifikasi yang digunakan dalam penelitian. Algoritma SVM dan *Random Forest* dievaluasi melalui *pengujian Classification Report* dan *Confusion Matrix*. Model *Confusion Matrix* ditunjukkan pada Tabel III [19].

TABEL III
MODEL CONFUSION MATRIX

Klasifikasi	Aktual Positif	Aktual Negatif
Prediksi Positif	TP	FP
Prediksi Negatif	FN	TN

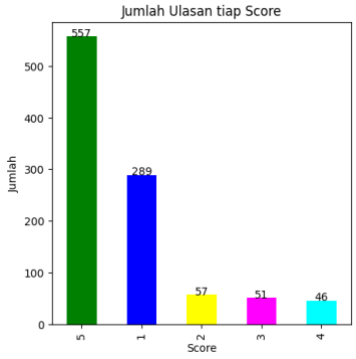
True Positive (TP) yaitu banyaknya data positif yang diprediksi benar oleh model; *True Negative* (TN) yaitu banyaknya data negatif yang diprediksi benar oleh model; *False Positive* (FP) yaitu banyaknya data positif yang diprediksi salah oleh model; *False Negative* (FN) yaitu banyaknya data negatif yang diprediksi salah oleh model. Dengan menggunakan informasi dari *confusion matrix*, selanjutnya dapat dihitung berbagai metrik kinerja (*performance matrix*) model, diantaranya :

1) *Akurasi*, untuk mengukur seberapa tepat model dalam mengklasifikasikan data dengan benar [19].
2) Presisi, untuk mengukur seberapa akurat model dalam memprediksi antara data yang diminta dengan hasil prediksi dari model [19].
3) *Recall*, untuk mengukur seberapa baik bagi model dalam menemukan kembali semua *instance* dari kelas tertentu dalam dataset [19].
4) *F1-Score*, untuk mengukur rata-rata presisi dan *recall* yang dibobotkan, menghasilkan metrik tunggal yang mengukur keseimbangan antara keduanya [19].

III. HASIL DAN PEMBAHASAN

A. Pengumpulan Data

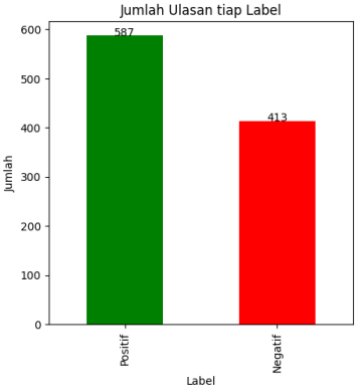
Data yang digunakan yaitu ulasan pengguna aplikasi GoPay pada Google Play Store periode Juli 2023 sampai Februari 2024 sejumlah 1.000 data. Ulasan yang dikumpulkan adalah ulasan dengan bahasa Indonesia, berasal dari negara Indonesia, serta memiliki hubungan yang relevan dengan aplikasi. Hasil proses *scraping* data untuk masing-masing skor ulasan ditunjukkan pada Gbr 4.



Gbr 4. Hasil Proses Scraping Data

B. Pelabelan

Pelabelan otomatis dengan Python dilakukan dengan ketentuan jika skor ulasan lebih besar dari 3, maka data diberi label 'Positif', sedangkan jika tidak, maka diberi label 'Negatif'. Kemudian dilanjutkan validasi pelabelan secara manual untuk memastikan bahwa hasil pelabelan telah sesuai dengan isi ulasan. Hasil proses pelabelan ditunjukkan pada Gbr 5.



Gbr 5. Hasil Pelabelan

C. Data Pre-processing

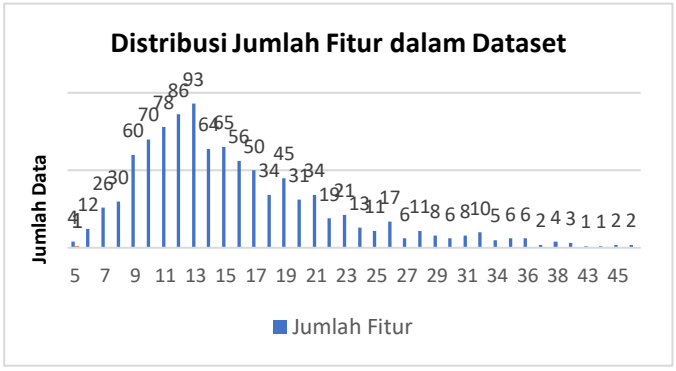
Data yang telah memiliki label selanjutnya dilakukan *pre-processing* untuk menghapus *noise* di dalam data. Hasil dari proses *data pre-processing* ditunjukkan pada Tabel IV.

TABEL IV
HASIL DATA PRE-PROCESSING

Data Awal	
Mudah di gunakan ,tapi Harus di tingkatkan lagi keamanannya, jangsn sampai bisa di bibol orang yg tidak bertanggubg jawab	
Cleaning	Mudah di gunakan tapi Harus di tingkatkan lagi keamanannya jangsn sampai bisa di bibol orang yg tidak bertanggubg jawab
Case Folding	mudah di gunakan tapi harus di tingkatkan lagi keamanannya jangsn sampai bisa di bibol orang yg tidak bertanggubg jawab

Normalisasi	mudah di gunakan tapi harus di tingkatkan lagi keamanannya jangan sampai bisa di bobol orang yang tidak bertanggung jawab
Stopwords Removal	mudah tingkatkan keamanannya bobol orang bertanggung
Tokenizing	['mudah', 'tingkatkan', 'keamanannya', 'bobol', 'orang', 'bertanggung']
Stemming	['mudah', 'tingkat', 'aman', 'bobol', 'orang', 'tanggung']
Hasil Akhir	
mudah tingkat aman bobol orang tanggung	

Setelah melalui seluruh tahap *pre-processing*, didapatkan jumlah total fitur di dalam dataset yaitu sebanyak 15.859 fitur. Jumlah fitur paling banyak yang terkandung dalam satu data adalah sebanyak 46 fitur, sedangkan yang paling sedikit adalah 5 fitur. Data yang mengandung 13 fitur memiliki frekuensi paling banyak di dalam dataset, yaitu sebanyak 93 data. Grafik distribusi jumlah fitur dalam dataset ditunjukkan pada Gbr 6.



Gbr 6. Grafik Distribusi Jumlah Fitur dalam Dataset

D. TF-IDF

Pembobotan TF-IDF dilakukan dengan mengimpor modul *TfidfVectorizer* dalam *scikit-learn*. Hasil proses TF-IDF ditunjukkan pada Gbr 7.

```
(0, 194)      0.2856543557721204  :     :
(0, 857)      0.3832786887263139  (997, 1221)   0.31897584435863413
(0, 437)      0.17382102726087045 (997, 1341)   0.42577419486939283
(0, 1019)     0.22165892782127716 (997, 107)    0.2759471569340936
(0, 230)      0.171161608053574   (997, 110)    0.23987015769162742
(0, 775)      0.1837202539061245  (997, 1333)   0.38052997489475937
(0, 1370)     0.19312215270692168 (997, 69)     0.25705723605340663
(0, 1325)     0.19001193775221756 (997, 230)    0.2608361734162511
(0, 244)      0.21756399002221616 (997, 830)    0.17066429038145858
(0, 605)      0.2890663439333438  (998, 898)    0.46593003081707896
(0, 918)      0.24664363862130897 (998, 1345)   0.43364098072002516
(0, 97)       0.12619949900765115 (998, 1262)   0.3995838981821233
(0, 832)      0.35452913238241623 (998, 1357)   0.33352449641356224
(0, 1359)     0.2890663439333438  (998, 1331)   0.23431591202181792
(0, 1247)     0.2206069239549506  (998, 420)    0.1924044389888953
(0, 830)      0.22398100690118913 (998, 616)    0.407104771504933
(0, 375)      0.19991116439747159 (998, 69)     0.15490282566723307
(1, 1188)     0.161059237744031222 (998, 830)   0.2056847823189747
(1, 327)      0.183842210010973   (999, 64)     0.41524528959623014
(1, 636)      0.17669327115660166 (999, 995)    0.38141032621704046
(1, 1396)     0.19378415408509084 (999, 763)    0.37973712213629773
(1, 1088)     0.14095203262018405 (999, 1331)   0.22567216846340823
(1, 973)      0.20863570947510648 (999, 127)    0.21754085099258993
(1, 166)      0.22522156078100625 (999, 69)     0.14918857310116776
(1, 110)      0.12111670974030715 (999, 437)    0.6149353966774553
:     :                           (999, 830)    0.1980972203096691
```

Gbr 7. Hasil Pembobotan TF-IDF

E. Hasil Pengujian Metode SVM

1) SVM Tanpa Fitur Seleksi

Hasil pengujian metode SVM tanpa fitur seleksi dengan kernel Linear, RBF, Sigmoid, dan Polynomial untuk 5 skenario pengujian split data dan *K-Fold Cross Validation* ditunjukkan pada Tabel V.

TABEL V
HASIL PENGUJIAN METODE SVM TANPA FITUR SELEKSI

Split Data					***K-Fold Cross Validation***				
Rasio	**Akurasi**	**Presisi**	**Recall**	**F1-Score**	**Nilai k**	**Akurasi**	**Presisi**	**Recall**	**F1-Score**
Kernel Linear									
90 : 10	96,00%	96,10%	96,00%	96,01%	2	89,90%	89,91%	89,90%	89,91%
80 : 20	94,50%	94,58%	94,50%	94,52%	4	90,30%	90,38%	90,30%	90,31%
70 : 30	92,33%	92,35%	92,33%	92,34%	6	90,89%	91,05%	90,89%	90,91%
60 : 40	90,75%	90,74%	90,75%	90,73%	8	91,10%	91,26%	91,10%	91,11%
50 : 50	90,20%	90,19%	90,20%	90,19%	10	91,50%	91,75%	91,50%	91,52%
Kernel RBF									
90 : 10	96,00%	96,10%	96,00%	96,01%	2	91,20%	91,28%	91,20%	91,22%
80 : 20	94,50%	94,98%	94,50%	94,54%	4	91,40%	91,52%	91,40%	91,42%
70 : 30	92,67%	92,79%	92,67%	92,69%	6	91,20%	91,35%	91,20%	91,21%
60 : 40	91,50%	91,55%	91,50%	91,51%	8	91,30%	91,50%	91,30%	91,32%
50 : 50	92,00%	92,09%	92,00%	92,02%	**10**	**91,80%**	**92,09%**	**91,80%**	**91,83%**
Kernel Sigmoid									
90 : 10	**98,00%**	**98,09%**	**98,00%**	**98,01%**	2	89,90%	89,92%	89,90%	89,91%
80 : 20	95,00%	95,23%	95,00%	95,03%	4	90,80%	90,89%	90,80%	90,82%
70 : 30	92,33%	92,35%	92,33%	92,34%	6	90,89%	91,10%	90,89%	90,91%
60 : 40	91,50%	91,50%	91,50%	91,50%	8	91,00%	91,21%	91,00%	91,01%
50 : 50	89,80%	89,79%	89,80%	89,80%	10	91,40%	91,64%	91,40%	91,42%
Kernel Polynomial									
90 : 10	88,00%	88,34%	88,00%	87,83%	2	87,10%	87,68%	87,10%	86,85%
80 : 20	91,00%	91,15%	91,00%	91,04%	4	89,50%	89,70%	89,50%	89,41%
70 : 30	90,00%	90,01%	90,00%	89,94%	6	90,20%	90,33%	90,20%	90,14%
60 : 40	87,25%	87,78%	87,25%	87,02%	8	89,10%	89,26%	89,10%	89,02%
50 : 50	86,80%	87,76%	86,80%	86,46%	10	89,50%	89,92%	89,50%	89,43%

Berdasarkan hasil pada Tabel V, ditunjukkan bahwa pengujian SVM tanpa fitur seleksi dengan split data memiliki performa terbaik pada rasio 90:10 dengan kernel Sigmoid. Nilai akurasi yang dihasilkan sebesar 98%, presisi 98,09%, recall 98%, dan F1-score 98,01%. Sedangkan, pengujian dengan *K-Fold Cross Validation* memiliki performa terbaik pada k = 10 dengan kernel RBF. Nilai akurasi yang dihasilkan sebesar 91,80%, presisi 92,09%, recall 91,80%, dan F1-score 91,83%.

2) Kombinasi Chi-square dan SVM

Hasil pengujian kombinasi *Chi-square* dan SVM dengan kernel Linear, RBF, Sigmoid, dan Polynomial untuk 5 skenario pengujian menggunakan teknik split data dan *K-Fold Cross Validation* ditunjukkan pada Tabel VI.

TABEL VI
HASIL PENGUJIAN KOMBINASI CHI-SQUARE DAN SVM

Split Data					*K-Fold Cross Validation*				
Rasio	**Akurasi**	**Presisi**	**Recall**	**F1-Score**	**Nilai k**	**Akurasi**	**Presisi**	**Recall**	**F1-Score**
Kernel Linear									
90 : 10	**96,00%**	**96,10%**	**96,00%**	**96,01%**	2	91,40%	91,45%	91,40%	91,41%
80 : 20	91,50%	92,01%	91,50%	91,57%	4	90,70%	90,81%	90,70%	90,72%
70 : 30	91,33%	91,41%	91,33%	91,35%	6	90,70%	90,89%	90,70%	90,72%
60 : 40	92,00%	92,02%	92,00%	92,01%	8	90,90%	91,08%	90,90%	90,92%
50 : 50	93,20%	93,19%	93,20%	93,20%	10	90,90%	91,16%	90,90%	90,92%
Kernel RBF									
90 : 10	94,00%	94,00%	94,00%	94,00%	2	89,60%	89,66%	89,60%	89,62%
80 : 20	91,00%	91,43%	91,00%	91,07%	4	90,60%	90,75%	90,60%	90,63%
70 : 30	91,33%	91,41%	91,33%	91,35%	6	90,09%	90,37%	90,09%	90,13%
60 : 40	90,25%	90,28%	90,25%	90,26%	8	90,40%	90,58%	90,40%	90,42%
50 : 50	90,60%	90,63%	90,60%	90,61%	10	90,20%	90,58%	90,20%	90,23%
Kernel Sigmoid									
90 : 10	**96,00%**	**96,10%**	**96,00%**	**96,01%**	2	90,90%	90,99%	90,90%	90,92%
80 : 20	92,00%	92,60%	92,00%	92,07%	4	90,40%	90,46%	90,40%	90,42%
70 : 30	92,33%	92,43%	92,33%	92,36%	6	91,10%	91,33%	91,10%	91,12%
60 : 40	90,50%	90,49%	90,50%	90,49%	8	90,80%	91,02%	90,80%	90,82%
50 : 50	92,20%	92,19%	92,20%	92,19%	**10**	**91,40%**	**91,77%**	**91,40%**	**91,43%**
Kernel Polynomial									
90 : 10	84,00%	86,42%	84,00%	83,26%	2	76,30%	80,33%	76,30%	74,18%
80 : 20	84,50%	85,22%	84,50%	83,99%	4	80,00%	82,32%	80,00%	78,89%
70 : 30	82,33%	84,23%	82,33%	81,44%	6	81,40%	83,32%	81,40%	80,47%
60 : 40	78,75%	82,30%	78,75%	77,22%	8	82,30%	84,29%	82,30%	81,44%
50 : 50	76,00%	80,56%	76,00%	73,74%	10	82,90%	84,75%	82,90%	82,15%

Berdasarkan hasil pada Tabel VI, ditunjukkan bahwa pengujian kombinasi *Chi-square* dan SVM dengan split data memiliki performa terbaik pada rasio 90:10 dengan kernel Linear dan Sigmoid. Nilai akurasi yang dihasilkan sebesar 96%, presisi 96,10%, recall 96%, dan F1-score 96,01%. Sedangkan, pengujian dengan *K-Fold Cross Validation* memiliki performa terbaik pada k = 10 dengan kernel Sigmoid. Nilai akurasi yang dihasilkan sebesar 91,40%, presisi 91,77%, recall 91,40%, dan F1-score 91,43%.

3) Kombinasi Information Gain dan SVM

Hasil pengujian kombinasi *Information Gain* dan SVM dengan kernel Linear, RBF, Sigmoid, dan Polynomial untuk 5 skenario pengujian menggunakan teknik split data dan *K-Fold Cross Validation* ditunjukkan pada Tabel VII.

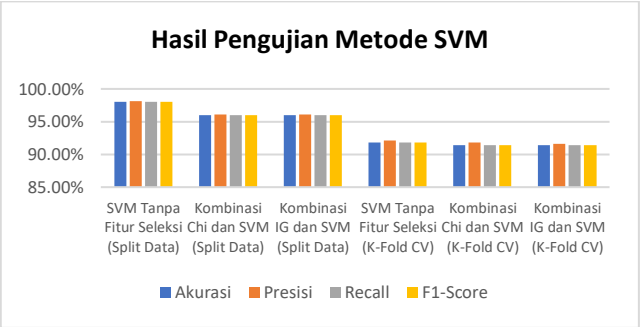
TABEL VII
HASIL PENGUJIAN KOMBINASI INFORMATION GAIN DAN SVM

Split Data					*K-Fold Cross Validation*				
Rasio	**Akurasi**	**Presisi**	**Recall**	**F1-Score**	**Nilai k**	**Akurasi**	**Presisi**	**Recall**	**F1-Score**
Kernel Linear									
90 : 10	**96,00%**	**96,05%**	**96,00%**	**95,99%**	2	90,20%	90,28%	90,20%	90,22%
80 : 20	91,50%	91,52%	91,50%	91,51%	4	89,90%	89,97%	89,90%	89,92%
70 : 30	90,67%	90,70%	90,67%	90,68%	6	90,50%	90,78%	90,50%	90,51%
60 : 40	90,25%	90,32%	90,25%	90,27%	8	90,60%	90,76%	90,60%	90,61%
50 : 50	91,00%	91,11%	91,00%	91,02%	10	91,20%	91,47%	91,20%	91,21%
Kernel RBF									
90 : 10	95,00%	95,03%	95,00%	95,01%	2	90,80%	90,87%	90,80%	90,82%
80 : 20	92,00%	92,26%	92,00%	92,05%	4	91,00%	91,10%	91,00%	91,02%
70 : 30	91,00%	91,05%	91,00%	91,02%	**6**	**91,40%**	**91,63%**	**91,40%**	**91,42%**
60 : 40	90,00%	90,02%	90,00%	90,01%	8	90,80%	90,98%	90,80%	90,82%
50 : 50	91,40%	91,51%	91,40%	91,42%	10	90,80%	90,98%	90,80%	90,82%
Kernel Sigmoid									
90 : 10	95,00%	95,03%	95,00%	95,01%	2	90,40%	90,55%	90,40%	90,43%

80 : 20	91,00%	91,15%	91,00%	91,04%	4	89,50%	89,54%	89,50%	89,51%
70 : 30	90,00%	90,08%	90,00%	90,02%	6	89,90%	90,18%	89,90%	89,90%
60 : 40	90,00%	90,02%	90,00%	90,01%	8	90,00%	90,25%	90,00%	90,01%
50 : 50	90,60%	90,79%	90,60%	90,64%	10	90,50%	90,73%	90,50%	90,52%
Kernel Polynomial									
90 : 10	87,00%	87,17%	87,00%	86,86%	2	86,30%	86,96%	86,30%	86,00%
80 : 20	88,50%	88,61%	88,50%	88,54%	4	87,60%	87,91%	87,60%	87,43%
70 : 30	90,00%	89,99%	90,00%	89,96%	6	87,70%	87,97%	87,70%	87,58%
60 : 40	86,75%	87,16%	86,75%	86,53%	8	87,90%	87,98%	87,90%	87,80%
50 : 50	85,80%	86,96%	85,80%	85,38%	10	88,30%	88,72%	88,30%	88,18%

Berdasarkan hasil pada Tabel VII, ditunjukkan bahwa pengujian kombinasi *Information Gain* dan SVM dengan split data memiliki performa terbaik pada rasio 90:10 dengan kernel Linear. Nilai akurasi yang dihasilkan sebesar 96%, presisi 96,05%, recall 96%, dan F1-score 95,99%. Sedangkan, pengujian dengan *K-Fold Cross Validation* memiliki performa terbaik pada k = 6 dengan kernel RBF. Nilai akurasi yang dihasilkan sebesar 91,40%, presisi 91,63%, recall 91,40%, dan F1-score 91,42%.

Dari seluruh pengujian yang telah dilakukan, model SVM tanpa fitur seleksi memiliki performa yang paling baik di antara pengujian lainnya. Hasil pengujian terbaik dari metode SVM ditunjukkan melalui grafik pada Gbr 8.



Gbr 8. Grafik Hasil Pengujian Terbaik Metode SVM

F. Hasil Pengujian Metode Random Forest

1) Random Forest Tanpa Fitur Seleksi

Hasil pengujian metode *Random Forest* tanpa fitur seleksi dengan 100, 300, 500, dan 1000 *n_estimators* untuk 5 skenario pengujian split data dan *K-Fold Cross Validation* ditunjukkan pada Tabel VIII.

TABEL VIII
HASIL PENGUJIAN METODE RANDOM FOREST TANPA FITUR SELEKSI

Split Data					***K-Fold Cross Validation***				
Rasio	**Akurasi**	**Presisi**	**Recall**	**F1-Score**	**Nilai k**	**Akurasi**	**Presisi**	**Recall**	**F1-Score**
100 *n_estimators*									
90 : 10	**96,00%**	**96,05%**	**96,00%**	**95,99%**	2	89,80%	89,82%	89,80%	89,77%
80 : 20	90,00%	90,06%	90,00%	90,02%	4	90,10%	90,12%	90,10%	90,10%
70 : 30	92,33%	92,33%	92,33%	92,33%	6	91,40%	91,52%	91,40%	91,39%
60 : 40	92,00%	92,02%	92,00%	91,97%	8	90,40%	90,48%	90,40%	90,40%
50 : 50	92,00%	92,01%	92,00%	91,97%	10	91,20%	91,38%	91,20%	91,20%
300 *n_estimators*									
90 : 10	95,00%	95,00%	95,00%	94,99%	2	90,60%	90,64%	90,60%	90,58%
80 : 20	92,00%	92,00%	92,00%	92,00%	4	90,90%	90,91%	90,90%	90,88%
70 : 30	92,33%	92,32%	92,33%	92,32%	**6**	**91,50%**	**91,61%**	**91,50%**	**91,49%**
60 : 40	91,75%	91,80%	91,75%	91,70%	8	90,90%	91,00%	90,90%	90,90%
50 : 50	91,80%	91,87%	91,80%	91,75%	10	91,10%	91,29%	91,10%	91,09%
500 *n_estimators*									
90 : 10	95,00%	95,00%	95,00%	94,99%	2	91,10%	91,13%	91,10%	91,07%
80 : 20	92,50%	92,52%	92,50%	92,51%	4	90,70%	90,71%	90,70%	90,70%
70 : 30	92,67%	92,66%	92,67%	92,65%	6	91,30%	91,41%	91,30%	91,29%
60 : 40	91,75%	91,80%	91,75%	91,70%	8	91,00%	91,10%	91,00%	91,00%
50 : 50	91,60%	91,68%	91,60%	91,54%	10	90,90%	91,11%	90,90%	90,89%
1000 *n_estimators*									
90 : 10	95,00%	95,00%	95,00%	94,99%	2	91,00%	91,03%	91,00%	90,98%
80 : 20	92,00%	92,00%	92,00%	92,00%	4	90,70%	90,72%	90,70%	90,69%
70 : 30	92,67%	92,66%	92,67%	92,65%	6	91,10%	91,20%	91,10%	91,09%
60 : 40	91,50%	91,57%	91,50%	91,45%	8	90,90%	91,01%	90,90%	90,89%
50 : 50	91,20%	91,28%	91,20%	91,14%	10	91,00%	91,22%	91,00%	90,99%

Berdasarkan hasil pada Tabel VIII, ditunjukkan bahwa pengujian *Random Forest* tanpa fitur seleksi dengan split data memiliki performa terbaik pada rasio 90:10 dengan 100 *n_estimators*. Nilai akurasi yang dihasilkan sebesar 96%,

presisi 96,05%, recall 96%, dan F1-score 95,99%. Sedangkan, pengujian dengan *K-Fold Cross Validation* memiliki performa terbaik pada k = 6 dengan 300 *n_estimators*. Nilai akurasi yang dihasilkan sebesar 91,50%, presisi 91,61%, recall 91,50%, dan F1-score 91,49%.

2) Kombinasi Chi-square dan Random Forest

Hasil pengujian kombinasi *Chi-square* dan *Random Forest* dengan 100, 300, 500, dan 1000 *n_estimators* untuk 5 skenario pengujian menggunakan teknik split data dan *K-Fold Cross Validation* ditunjukkan pada Tabel IX.

TABEL IX
HASIL PENGUJIAN KOMBINASI CHI-SQUARE DAN RANDOM FOREST

Split Data					***K-Fold Cross Validation***				
Rasio	**Akurasi**	**Presisi**	**Recall**	**F1-Score**	**Nilai k**	**Akurasi**	**Presisi**	**Recall**	**F1-Score**
100 *n_estimators*									
90 : 10	93,00%	93,09%	93,00%	92,96%	2	89,50%	89,55%	89,50%	89,43%
80 : 20	90,50%	90,48%	90,50%	90,49%	4	89,00%	88,98%	89,00%	88,97%
70 : 30	90,67%	90,66%	90,67%	90,62%	6	89,60%	89,73%	89,60%	89,55%
60 : 40	89,75%	89,81%	89,75%	89,68%	8	89,70%	89,81%	89,70%	89,65%
50 : 50	89,80%	90,04%	89,80%	89,68%	10	89,80%	90,10%	89,80%	89,73%
300 *n_estimators*									
90 : 10	**95,00%**	**95,00%**	**95,00%**	**94,99%**	2	88,70%	88,81%	88,70%	88,61%
80 : 20	91,00%	91,00%	91,00%	91,00%	4	89,60%	89,60%	89,60%	89,56%
70 : 30	90,33%	90,33%	90,33%	90,28%	**6**	89,90%	90,02%	89,90%	89,87%
60 : 40	90,25%	90,28%	90,25%	90,20%	8	89,40%	89,51%	89,40%	89,35%
50 : 50	89,60%	89,81%	89,60%	89,49%	10	89,80%	90,05%	89,80%	89,76%
500 *n_estimators*									
90 : 10	**95,00%**	**95,00%**	**95,00%**	**94,99%**	2	88,80%	88,94%	88,80%	88,70%
80 : 20	91,50%	91,48%	91,50%	91,49%	4	89,90%	89,90%	89,90%	89,87%
70 : 30	91,00%	90,99%	91,00%	90,97%	6	89,60%	89,74%	89,60%	89,55%
60 : 40	89,75%	89,85%	89,75%	89,67%	**8**	**90,00%**	**90,16%**	**90,00%**	**89,95%**
50 : 50	89,60%	89,87%	89,60%	89,47%	10	89,80%	90,07%	89,80%	89,75%
1000 *n_estimators*									
90 : 10	94,00%	94,03%	94,00%	93,98%	2	88,80%	88,93%	88,80%	88,71%
80 : 20	91,00%	91,00%	91,00%	91,00%	4	89,70%	89,70%	89,70%	89,67%
70 : 30	90,00%	90,01%	90,00%	89,94%	6	89,50%	89,63%	89,50%	89,45%
60 : 40	89,75%	89,85%	89,75%	89,67%	8	89,90%	90,00%	89,90%	89,85%
50 : 50	89,80%	90,04%	89,80%	89,68%	10	89,80%	90,03%	89,80%	89,75%

Berdasarkan hasil pada Tabel IX, ditunjukkan bahwa pengujian kombinasi *Chi-square* dan *Random Forest* dengan split data memiliki performa terbaik pada rasio 90:10 dengan 300 dan 500 *n_estimators*. Nilai akurasi yang dihasilkan sebesar 95%, presisi 95%, recall 95%, dan F1-score 94,99%. Sedangkan, pengujian dengan *K-Fold Cross Validation* memiliki performa terbaik pada k = 8 dengan 500 *n_estimators*. Nilai akurasi yang dihasilkan sebesar 90%, presisi 90,16%, recall 90%, dan F1-score 89,95%.

3) Kombinasi Information Gain dan Random Forest

Hasil pengujian kombinasi *Information Gain* dan *Random Forest* dengan 100, 300, 500, dan 1000 *n_estimators* untuk 5 skenario pengujian menggunakan teknik split data dan *K-Fold Cross Validation* ditunjukkan pada Tabel X.

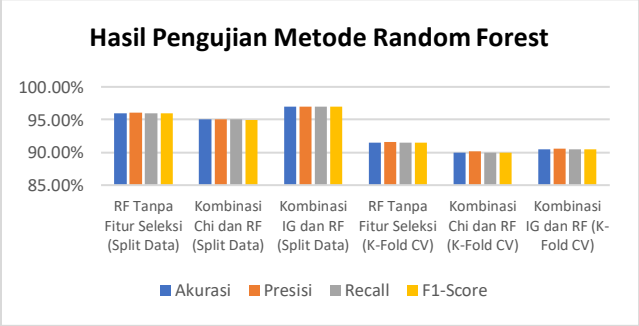
TABEL X
HASIL PENGUJIAN KOMBINASI INFORMATION GAIN DAN RANDOM FOREST

Split Data					***K-Fold Cross Validation***				
Rasio	**Akurasi**	**Presisi**	**Recall**	**F1-Score**	**Nilai k**	**Akurasi**	**Presisi**	**Recall**	**F1-Score**
100 *n_estimators*									
90 : 10	**97,00%**	**97,01%**	**97,00%**	**96,99%**	2	89,50%	89,56%	89,50%	89,44%
80 : 20	92,50%	92,49%	92,50%	92,49%	4	90,20%	90,21%	90,20%	90,18%
70 : 30	91,33%	91,41%	91,33%	91,26%	6	90,00%	90,09%	90,00%	89,97%
60 : 40	89,50%	89,58%	89,50%	89,42%	8	90,20%	90,39%	90,20%	90,15%
50 : 50	90,60%	90,72%	90,60%	90,52%	10	90,40%	90,61%	90,40%	90,38%
300 *n_estimators*									
90 : 10	95,00%	95,00%	95,00%	94,99%	2	89,90%	89,96%	89,90%	89,85%
80 : 20	92,00%	91,98%	92,00%	91,98%	4	90,40%	90,42%	90,40%	90,39%
70 : 30	91,67%	91,72%	91,67%	91,61%	6	89,70%	89,84%	89,70%	89,66%
60 : 40	90,75%	90,76%	90,75%	90,71%	8	89,90%	90,05%	89,90%	89,86%
50 : 50	91,20%	91,31%	91,20%	91,13%	10	90,30%	90,53%	90,30%	90,28%
500 *n_estimators*									

90 : 10	95,00%	95,00%	95,00%	94,99%	2	89,90%	89,93%	89,90%	89,86%
80 : 20	91,50%	91,48%	91,50%	91,49%	4	90,40%	90,42%	90,40%	90,39%
70 : 30	91,00%	91,05%	91,00%	90,94%	6	90,20%	90,30%	90,20%	90,17%
60 : 40	90,25%	90,28%	90,25%	90,20%	8	89,90%	90,02%	89,90%	89,86%
50 : 50	91,00%	91,13%	91,00%	90,93%	10	90,40%	90,60%	90,40%	90,38%
1000 *n_estimators*									
90 : 10	95,00%	95,00%	95,00%	94,99%	2	89,80%	89,84%	89,80%	89,75%
80 : 20	92,50%	92,53%	92,50%	92,45%	**4**	**90,50%**	**90,52%**	**90,50%**	**90,49%**
70 : 30	91,67%	91,72%	91,67%	91,61%	6	90,30%	90,47%	90,30%	90,27%
60 : 40	89,75%	89,81%	89,75%	89,68%	8	89,80%	89,93%	89,80%	89,76%
50 : 50	90,40%	90,50%	90,40%	90,33%	10	90,20%	90,44%	90,20%	90,17%

Berdasarkan hasil pada Tabel X, ditunjukkan bahwa pengujian kombinasi *Information Gain* dan *Random Forest* dengan split data memiliki performa terbaik pada rasio 90:10 dengan 100 *n_estimators*. Nilai akurasi yang dihasilkan sebesar 97%, presisi 97,01%, recall 97%, dan F1-score 96,99%. Sedangkan, pengujian dengan *K-Fold Cross Validation* memiliki performa terbaik pada k = 4 dengan 1000 *n_estimators*. Nilai akurasi yang dihasilkan sebesar 90,50%, presisi 90,52%, recall 90,50%, dan F1-score 90,49%.

Dari seluruh pengujian yang telah dilakukan, model kombinasi *Information Gain* dan *Random Forest* memiliki performa yang paling baik di antara pengujian lainnya. Hasil pengujian terbaik dari metode *Random Forest* ditunjukkan melalui grafik pada Gbr 9.



Gbr 9. Grafik Hasil Pengujian Terbaik Metode SVM

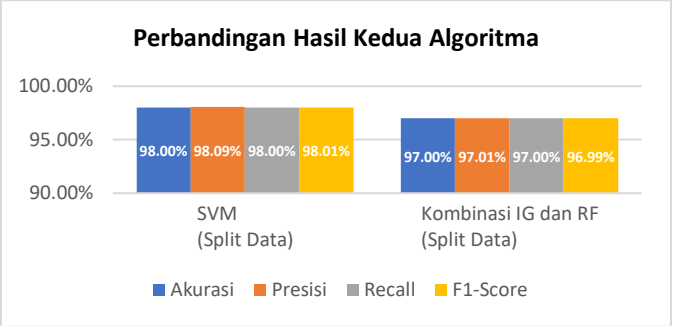
G. Hasil Perbandingan Metode SVM dan Random Forest

Perbandingan hasil pengujian bertujuan untuk mengetahui metode yang memiliki performa lebih baik antara SVM dan *Random Forest*. Tabel XI menampilkan hasil perbandingan dari kedua metode tersebut.

TABEL XI
PERBANDINGAN HASIL PENGUJIAN METODE SVM DAN RANDOM FOREST

Algoritma	Akurasi	Presisi	Recall	F1-Score
SVM (Split Data)	**98,00%**	**98,09%**	**98,00%**	**98,01%**
Kombinasi IG dan RF (Split Data)	97,00%	97,01%	97,00%	96,99%

Berdasarkan perbandingan hasil pada Tabel XI, split data menghasilkan performa yang lebih baik dibandingkan dengan *K-Fold Cross Validation*. Metode SVM memiliki performa terbaik pada pengujian tanpa fitur seleksi, yaitu dengan nilai akurasi sebesar 98%, presisi 98,09%, recall 98%, dan F1-score 98,01%. Sedangkan, metode *Random Forest* memiliki performa terbaik dengan kombinasi fitur seleksi *Information Gain*. Nilai akurasi yang dihasilkan sebesar 97%, presisi 97,01%, recall 97%, dan F1-Score 96,99%. Perbandingan dari kedua performa terbaik tersebut direpresentasikan dalam bentuk grafik seperti pada Gbr 10.



Gbr 10. Grafik Perbandingan Hasil Metode SVM dan Random Forest

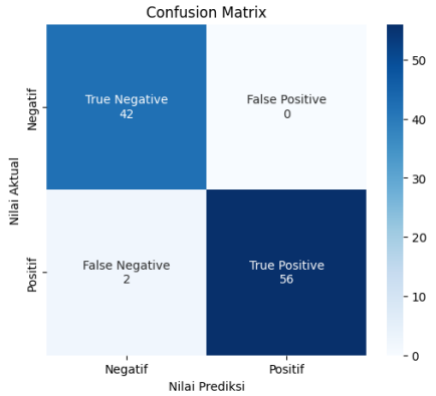
H. Evaluasi

Berikut ini pada Gbr 11 ditampilkan *classification report* untuk model dengan performa terbaik yaitu kernel Sigmoid dari pengujian SVM tanpa fitur seleksi dengan split data rasio 90:10.

	precision	recall	f1-score	support
Negatif	0.95	1.00	0.98	42
Positif	1.00	0.97	0.98	58
accuracy			0.98	100
macro avg	0.98	0.98	0.98	100
weighted avg	0.98	0.98	0.98	100

Gbr 11. Hasil Classification Report

Berikutnya, *confusion matrix* kernel Sigmoid dari pengujian SVM tanpa fitur seleksi dengan split data rasio 90:10 ditunjukkan pada Gbr 12.



Gbr 12. Hasil Confusion Matrix

Berdasarkan hasil visualisasi *confusion matrix*, kernel Sigmoid memprediksi 42 data sebagai negatif dan 0 data sebagai positif dari total 42 data negatif. Kemudian, memprediksi 2 data sebagai negatif dan 56 data sebagai positif dari total 58 data positif.

IV. KESIMPULAN

Berdasarkan hasil pengujian, penelitian ini dapat disimpulkan bahwa teknik evaluasi split data menghasilkan performa yang lebih baik dibandingkan dengan *K-Fold Cross Validation*. Penggunaan fitur seleksi *Chi-square* tidak mampu meningkatkan performa model SVM dan *Random Forest*, baik dengan teknik split data ataupun *K-Fold Cross Validation*. Sedangkan, penggunaan fitur seleksi *Information Gain* hanya mampu meningkatkan performa model *Random Forest* apabila digunakan dengan teknik split data. Metode SVM memiliki performa terbaik pada pengujian tanpa menerapkan fitur seleksi, yaitu dengan split data rasio 90:10 menggunakan kernel Sigmoid. Nilai akurasi yang dihasilkan sebesar 98%, presisi 98,09%, recall 98%, dan F1-score 98,01%. Sedangkan, metode *Random Forest* memiliki performa terbaik dengan kombinasi fitur seleksi *Information Gain*, yaitu dengan split data rasio 90:10 menggunakan 100 *n_estimators*. Nilai akurasi yang dihasilkan sebesar 97%, presisi 97,01%, recall 97%, dan F1-score 96,99%. Hasil yang diperoleh menunjukkan bahwa metode SVM memiliki performa yang lebih baik dibandingkan dengan *Random Forest*.

V. SARAN

Untuk penelitian serupa yang selanjutnya dapat menggunakan data ulasan aplikasi GoPay di Google Play Store yang terbaru, menambah jumlah dataset untuk meningkatkan kemampuan model dalam mengenali pola dalam teks dan membuat prediksi sentimen yang lebih akurat, serta mempertimbangkan penggunaan metode klasifikasi dan fitur seleksi lainnya untuk dibandingkan dengan hasil pada penelitian ini.

UCAPAN TERIMA KASIH

Puji syukur penulis panjatkan kepada Allah SWT atas rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan penelitian ini. Penulis mengucapkan terima kasih yang mendalam kepada orang tua atas doa dan dukungan yang tanpa henti. Rasa terima kasih juga penulis sampaikan kepada dosen pembimbing yang selalu memberikan arahan dan bimbingan terbaiknya. Tak lupa, penulis mengucapkan terima kasih kepada teman-teman yang juga membantu dan mendukung dalam proses penyelesaian penelitian ini.

REFERENSI

[1] S. A. Helmayanti, F. Hamami, and R. Y. Fa'rifah, "Penerapan Algoritma TF-IDF dan Naïve Bayes Untuk Analisis Sentimen Berbasis Aspek Ulasan Aplikasi Flip Pada Google Play Store," *Jurnal Indonesia : Manajemen Informatika dan Komunikasi*, vol. 4, no. 3, pp. 1822–1834, Sep. 2023, doi: 10.35870/jimik.v4i3.415.

[2] W. Karim, M. A. Ulfy, and A. Hossain, "Factors Influencing the Use of E-wallet as a Payment Method among Malaysian Young Adults," *Journal of International Business and Management*, vol. 3, no. 2, 2020, doi: 10.37227/jibm-2020-2-21.

[3] A. Oktian Permana and Sudin Saepudin, "Perbandingan Algoritma K-Nearest Neighbor dan Naïve Bayes Pada Aplikasi Shopee," *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 4, no. 1, pp. 25–32, Apr. 2023, doi: 10.37859/coscitech.v4i1.4474.

[4] Populix, "Consumer Preference Towards Banking and E-Wallet Apps," 2022.

[5] Maghfira, "Faktor-Faktor yang Mempengaruhi Penggunaan Sistem Pembayaran GoPay," 2018.

[6] GoPay, "Tentang GoPay," Gopay.co.id. Accessed: Nov. 03, 2023. [Online]. Available: https://gopay.co.id/tentang-gopay

[7] L. F. Lishobrina, M. P. Arum, C. M. Hidayat, L. I. Widianty, and G. P. Wengkau, "Analisis Faktor Kepuasan Pengguna Gopay dalam Digital Financial Management," 2023.

[8] S. A. Saputra, D. Rosiyadi, W. Gata, and S. M. Husain, "Analisis Sentimen E-Wallet Pada Google Play Menggunakan Algoritma Naive Bayes Berbasis Particle Swarm Optimization," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)* , vol. 3, no. 3, pp. 377–382, 2019.

[9] S. Wang, D. Li, L. Zhao, and J. Zhang, "Sample cutting method for imbalanced text sentiment classification based on BRC," *Knowl Based Syst*, vol. 37, pp. 451–461, 2013, doi: 10.1016/j.knosys.2012.09.003.

[10] D. Abror, "Analisis Sentimen Review Aplikasi PeduliLindungi Menggunakan Seleksi Fitur Information Gain Berbasis SVM," *Indonesian Journal on Software Engineering (IJSE)*, vol. 9, no. 1, pp. 1–8, 2023, [Online]. Available: http://ejournal.bsi.ac.id/ejurnal/index.php/ijse

[11] A. I. Tanggraeni and M. N. N. Sitokdana, "Analisis Sentimen Aplikasi E-Government Pada Google Play Menggunakan Algoritma Naïve Bayes," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 9, no. 2, pp. 785–795, 2022.

[12] D. Cart, "Training Data vs. Validation Data vs. Test Data for ML Algorithms," Applause.com. Accessed: Nov. 06, 2023. [Online]. Available: https://www.applause.com/blog/training-data-validation-data-vs-test-data

[13] F. Oktaviani P and R. Cahya W, "Analisis Sentimen pada Ulasan Pengguna MRT Jakarta Menggunakan Metode Neighbor-Weighted K-Nearest Neighbor dengan Seleksi Fitur Information Gain," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 4, no. 7, pp. 2195–2203, 2020, [Online]. Available: http://j-ptiik.ub.ac.id

[14] K. Schouten, F. Frasincar, and R. Dekker, "An Information Gain-Driven Feature Study for Aspect-Based Sentiment Analysis," 2016.

[15] A. P. P. Wardani, A. Adiwijaya, and M. D. Purbolaksono, "Sentiment Analysis on Beauty Product Review Using Modified Balanced Random Forest Method and Chi-Square," *Journal of Information System Research (JOSH)*, vol. 4, no. 1, pp. 1–7, Oct. 2022, doi: 10.47065/josh.v4i1.2047.

[16] D. Irvantoro, I. Saifudin, and R. Umilasari, "Feature Selection Menggunakan Chi-Square Dan N-Gram Dengan Algoritma Naive Bayes Classifier Untuk Analisis Sentimen Review Produk Elektronik," 2020.

[17] D. N. Fitriana and Y. Sibaroni, "Sentiment Analysis on KAI Twitter Post Using Multiclass Support Vector Machine (SVM)," *RESTI Journal (System Engineering and Information Technology)*, vol. 4, no. 2, pp. 846–853, 2020, [Online]. Available: http://jurnal.iaii.or.id

[18] T. B. Rohman, D. Dwi Purwanto, and J. Santoso, "Sentiment Analysis Terhadap Review Rumah Makan di Surabaya Memanfaatkan Algoritma Random Forest," *Fakultas Sistem Informasi*, 2018.

[19] S. Anggreany, "Confusion Matrix," socs.binus.ac.id. Accessed: Nov. 06, 2023. [Online]. Available: https://socs.binus.ac.id/2020/11/01/confusion-matrix/