Volume 6 Nomor 2 Juli 2024 JJEEE e-ISSN : 2715-0887
p-ISSN : 2654-7813

Jambura Journal of Electrical and Electronics Engineering

Pendekatan Ensemble pada Analisis Sentimen Ulasan Aplikasi Google Play Store

Ensemble Approach to Sentiment Analysis of Google Play Store App Reviews

Yasin Aril Mustofa*
Program Studi Informatika
Universitas Ichsan Sidenreng Rappang
Sidrap, Indonesia
arieldcc@gmail.com

Irma Surya Kumala Idris
Program Studi Teknik Informatika
Universitas Ichsan Gorontalo
Gorontalo, Indonesia
mhaladp@gmail.com

Diterima : Mei 2024
Disetujui : Juni 2024
Dipublikasi : Juli 2024

Abstrak*—Dalam era digital saat ini, analisis sentimen pada ulasan aplikasi *Google Play Store* telah menjadi kunci penting untuk memahami opini publik terhadap produk teknologi. Penelitian ini bertujuan untuk mengevaluasi efektivitas pendekatan *ensemble* dalam analisis sentimen, dibandingkan dengan algoritma klasifikasi individu. Metode yang digunakan meliputi teknik *ensemble* seperti *Random Forest* dan *Boosting*, serta algoritma individu seperti *Naive Bayes* dan *Support Vector Machine (SVM)*. Penelitian ini mengintegrasikan langkah *preprocessing* yang ekstensif, termasuk *cleaning, case folding, tokenization, stopword removal*, dan normalisasi, untuk mempersiapkan data sebelum proses klasifikasi. Hasil penelitian menunjukkan bahwa model ensemble, khususnya *Random Forest,* menghasilkan kinerja yang superior dalam klasifikasi sentimen ulasan aplikasi, dengan akurasi mencapai 94.15% untuk ulasan aplikasi *Zoom* dan 80.69% untuk ulasan aplikasi *Shopee*. Kinerja ini menegaskan bahwa pendekatan *ensemble* lebih efektif dalam mengatasi kompleksitas dan variasi data ulasan dibandingkan dengan algoritma yang beroperasi secara individu. Penelitian ini memberikan wawasan berharga bagi pengembang aplikasi untuk meningkatkan produk berdasarkan *feedback* pengguna. Namun, masih ada ruang untuk perbaikan dalam hal pengoptimalan algoritma terhadap data yang sangat tidak seimbang dan pengembangan metode yang dapat menangani nuansa bahasa yang lebih kompleks. Saran untuk penelitian mendatang meliputi penggunaan teknik *Deep Learning* dan pengujian lintas domain untuk menilai efektivitas model ini dalam berbagai setting analisis sentimen.

Kata Kunci—Analisis Sentimen; Ensemble Learning; Random Forest; Support Vector Machine; Preprocessing Data.

Abstract—In the current digital era, sentiment analysis of Google Play Store application reviews has become a critical key to understanding public opinion on technology products. This study aims to evaluate the effectiveness of ensemble approaches in sentiment analysis compared to individual classification algorithms. The methods employed include ensemble techniques such as Random Forest and Boosting, along with individual algorithms like Naive Bayes and Support Vector Machine (SVM). This research incorporates extensive preprocessing steps, including cleaning, case folding, tokenization, stopword removal, and normalization, to prepare the data before classification. The results demonstrate that ensemble models, particularly Random Forest, achieve superior performance in sentiment classification of app reviews, with accuracy reaching 94.15% for Zoom app reviews and 80.69% for Shopee app reviews. This performance confirms that ensemble approaches are more effective in handling the complexity and variability of review data compared to individually operated algorithms. The study provides valuable insights for application developers to enhance their products based on user feedback. However, there is still room for improvement in terms of optimizing algorithms for highly unbalanced data and developing methods that can handle more complex language nuances. Recommendations for future research include the use of Deep Learning techniques and cross-domain testing to assess the effectiveness of these models in various sentiment analysis settings.

Keywords—Sentiment Analysis; Ensemble Learning; Random Forest; Support Vector Machine; Data Preprocessing.

I. PENDAHULUAN

Analisis sentimen ulasan aplikasi telah menjadi area penelitian yang penting dengan pertumbuhan eksponensial dalam penggunaan aplikasi seluler. Dengan meningkatnya jumlah pengguna yang berbagi pengalaman mereka melalui platform ulasan, data ulasan ini dapat dimanfaatkan untuk memahami opini publik secara mendalam. Dalam konteks *Google Play Store*, ulasan ini mencakup berbagai aplikasi dari kategori seperti *e-commerce*, *streaming*, dan video konferensi. Studi mengenai aplikasi seperti *Shopee* menunjukkan bahwa analisis sentimen memberikan wawasan berharga bagi pengembang untuk meningkatkan kualitas produk mereka [1]. Selain itu, analisis opini masyarakat terhadap topik tertentu juga dapat dipahami melalui media sosial seperti *Twitter* [2].

Pendekatan ensemble telah muncul sebagai solusi yang efektif dalam meningkatkan kinerja klasifikasi analisis sentimen dibandingkan algoritma individu. *Ensemble learning*, seperti *Random Forest* dan *Boosting*, menggabungkan beberapa model untuk memaksimalkan akurasi dalam mengatasi data yang kompleks. Studi oleh [3] menyoroti efektivitas pendekatan ini, di mana penggabungan model memberikan keunggulan signifikan. Dalam analisis sentimen, *Soft Voting Classifier* yang menggabungkan berbagai algoritma seperti *Logistic Regression* dan *Random Forest* menghasilkan kinerja yang superior [4]. Studi lain menunjukkan metode *ensemble* seperti *Random Forest* dan *Boosting* dapat secara optimal memprediksi sentimen dengan kinerja yang lebih baik dibandingkan algoritma individu [5].

Sementara itu, algoritma individu seperti *Naive Bayes*[6] dan *Support Vector Machine* (SVM) juga telah digunakan secara luas dalam analisis sentimen, meskipun memiliki keterbatasan dalam menghadapi data yang lebih beragam. *Naive Bayes*, sebagai salah satu metode populer, dapat menghasilkan akurasi hingga 89,9% dalam klasifikasi opini publik terhadap UU Cipta Kerja di media sosial [2] . Dalam analisis sentimen aplikasi *Shopee*, SVM mampu mencapai akurasi hingga 98% dalam mengklasifikasi komentar pengguna ke dalam ulasan positif atau negatif [1].

Selain pendekatan algoritmik, langkah *preprocessing* seperti *case folding, stemming, stopword removal*, dan *tokenization* memainkan peran penting dalam meningkatkan akurasi model[7]. Teknik negation handling dan normalisasi data juga memiliki dampak positif terhadap kinerja klasifikasi sentimen [8]. Hal ini memungkinkan algoritma ensemble untuk lebih optimal dalam menganalisis sentimen yang berbeda-beda dari ulasan aplikasi.

Dalam penelitian ini, pendekatan *ensemble* seperti *Random Forest* dan *Boosting* akan digunakan untuk menganalisis ulasan aplikasi di *Google Play Store* dari berbagai kategori. *Random Forest* menggunakan pohon keputusan untuk meningkatkan akurasi dan mengurangi *overfitting*, sedangkan *Boosting* mengurangi bias dengan menyesuaikan bobot *instance* yang salah klasifikasi, meskipun meningkatkan risiko *overfitting* dan kompleksitas komputasi. Pemilihan *Random Forest* dan *Boosting* didasarkan pada efektivitasnya dalam analisis sentimen yang terbukti dalam penelitian. Misalnya, penelitian menunjukkan bahwa *Random Forest* lebih unggul daripada *AdaBoost* dan *Gradient Boosting* dalam analisis sentimen pada komentar *YouTube* berbahasa Indonesia [9]. Efektivitas *Random Forest* dan *Gradient Boosting* dalam tugas klasifikasi kompleks juga disoroti dalam penelitian [10]. Pendekatan seperti *Stacking* dan *Voting* dapat menambah kompleksitas tanpa peningkatan kinerja signifikan, sementara *AdaBoost* dan *Gradient Boosting* mungkin tidak selalu mengungguli *Random Forest* [11]. Pada penelitian yang dilakukan Evaluasi kinerja *ensemble learning* akan dilakukan dengan *K-Fold Cross Validation* dan diukur dengan akurasi, *presisi*, *recall*, serta *F1-score* untuk menemukan pendekatan yang lebih baik daripada algoritma individu.

II. METODE

A. Populasi dan Sample Penelitian

Populasi yang digunakan dalam penelitian ini adalah ulasan aplikasi *Google Play Store* dari berbagai kategori, seperti *e-commerce, streaming,* dan video konferensi [10]. Sampel akan diambil dengan teknik sampling acak sederhana untuk memastikan keterwakilan berbagai kategori aplikasi. Data ulasan akan dibersihkan dan difilter untuk memperoleh ulasan yang relevan dan bervariasi dari sudut pandang sentimen positif dan negatif [1][11].

B. Pengumpulan Data

Data ulasan aplikasi akan diambil langsung dari *Google Play Store* menggunakan teknik *web scraping*, dengan fokus pada aplikasi populer di setiap kategori [12]. Proses pengumpulan data ini merujuk pada beberapa penelitian yang relevan. Misalnya, pada penelitian yang membahas pengambilan data konsumsi makanan yang didokumentasikan pengguna dari aplikasi yang tersedia secara publik, menggunakan perangkat lunak *web data crawling* [13]. Penelitian lainnya juga memberikan wawasan melalui analisis konten kualitatif ulasan pengguna untuk aplikasi pelacakan kesehatan, yang menunjukkan bahwa ulasan pengguna dari *Google Play Store* dapat memberikan wawasan berharga terkait penggunaan aplikasi kesehatan mental [14]. Serta penelitian yang membahas metode dinamis untuk memprioritaskan ulasan pengguna dengan menggunakan database besar dari *Google Play Store* [15].

Untuk menjaga kualitas data dan menghindari bias, ulasan yang dipilih harus memiliki jumlah kata minimum dan telah diposting dalam jangka waktu tertentu [16]. Setiap ulasan akan diberi label sebagai positif atau negatif berdasarkan isi dan konteks ulasan tersebut. Dataset yang diciptakan telah dipublikasikan di Kaggle [17]. Publikasi dataset ini memungkinkan para peneliti lain untuk memanfaatkan data yang sama dalam studi mereka.

C. Teknik Preprocessing Data

Tahapan preprocessing penting dalam meningkatkan akurasi analisis sentimen[4][18]. Langkah-langkah yang akan diterapkan meliputi:

- *Cleaning:* Menghapus karakter yang dianggap kurang penting seperti *url*, *mention*, *hastag*, tanda baca dan angka
- *Case Folding:* Mengubah semua teks menjadi huruf kecil untuk memastikan konsistensi.
- *Tokenization:* Memisahkan teks ulasan menjadi unit-unit kata.
- *Stopword Removal:* Menghilangkan kata-kata umum yang tidak memberikan banyak informasi kontekstual.
- *Steaming:* Mengubah kata menjadi bentuk dasarnya.
- *Normalization:* Menstandarisasi variasi Bahasa dan kesalahan ejaan.

D. Term Weigting

Term weighting mengacu pada pemberian bobot numerik untuk setiap kata dalam dokumen, yang mengindikasikan pentingnya kata tersebut dalam konteks dokumen atau keseluruhan corpus. Bobot ini seringkali bergantung pada frekuensi kemunculan kata dalam dokumen *(Term Frequency, TF)* dan seberapa unik atau jarang kata tersebut muncul di seluruh dokumen dalam corpus *(Inverse Document Frequency, IDF)* [1][2].

Persamaan TF-IDF:

$$TF(t,d) = \frac{f_{t,d}}{max\{f_{t',d} : t' \in d\}} \quad (1)$$

$$IDF(t,d) = log\left(\frac{N}{|\{d \in D : t \in d\}|}\right) \quad (2)$$

$$TF - IDF(t,d,D) = TF(t,d)\ x\ IDF(t,d) \quad (3)$$

E. Pemodelan

Data yang telah dipreproses akan diubah menjadi representasi fitur menggunakan metode TF-IDF dan Word Embedding[19][20], sebelum kemudian digunakan untuk melatih model-model ini. Algoritma *ensemble* seperti *Random Forest* akan menggabungkan hasil prediksi dari beberapa pohon keputusan untuk memperoleh hasil akhir, sementara *Gradient Boosting* secara iteratif akan memperbaiki kesalahan prediksi model sebelumnya[21][22].

Model yang digunakan dalam penelitian ini:

- *Random Forest:*

Random Forest adalah metode *ensemble* yang membangun sekumpulan pohon keputusan untuk meningkatkan kinerja klasifikasi. Setiap pohon bekerja secara independen dengan subset data yang berbeda, lalu suara dari semua pohon dikombinasikan untuk mendapatkan prediksi akhir. *Random Forest* mampu menangani dataset yang besar dan kompleks, dengan memberikan hasil yang akurat dan mengurangi risiko *overfitting* [23]. Algoritma ini juga efektif untuk analisis sentimen, karena mampu menangkap hubungan kompleks antara fitur ulasan dan klasifikasi.

Rumus:

$$f(x) = \frac{1}{N}\sum_{n=1}^{N} f_n\,(x) \quad (4)$$

Dimana: $f_n(x)$ adalah prediksi dari pohon ke-n, dan N adalah jumlah pohon dalam hutan

- *Boosting:*

Boosting adalah teknik yang meningkatkan kinerja model dengan memberi bobot lebih besar pada sampel yang salah diklasifikasikan pada iterasi sebelumnya. Salah satu metode *boosting* yang populer adalah *AdaBoost*, yang menggabungkan beberapa model lemah *(weak learners)* menjadi model kuat *(strong learner)* [5][4][21]. *Boosting* efektif dalam meningkatkan kinerja analisis sentimen dengan cara berfokus pada ulasan yang sulit diklasifikasikan.

Rumus:

Skor akhir dihitung sebagai jumlah tertimbang dari model-model lemah:

$$f(x) = \sum_{m=1}^{M} \sigma_m f_m(x) \quad (5)$$

Dimana $f_m(x)$ adalah model ke-m, dan σ_m adalah bobotnya.

- *Support Vector Machine (SVM)*

SVM adalah algoritma klasifikasi yang memisahkan kelas menggunakan *hyperplane* terbaik. Dalam konteks analisis sentimen, SVM memaksimalkan *margin* antara kelas positif dan negatif pada ulasan aplikasi [1][24]. Teknik ini terkenal dalam menangani data dengan jumlah fitur yang banyak.

Rumus:

Hyperplane terbaik ditemukan dengan meminimalkan:

$$L(w,b) = \frac{1}{2}w^T w \quad (6)$$

Dengan kendala $y_i(w^T + b) \geq 1$ untuk semua i.

- *Naive Bayes:*

Naive Bayes adalah algoritma probabilistik yang mengasumsikan independensi antar fitur. Meskipun sederhana, NB sering kali memberikan hasil yang baik pada analisis sentimen ulasan aplikasi *Google Play Store*, seperti dalam klasifikasi opini publik tentang undang-undang [2].

Rumus:

Probabilitas sentiment dari ulasan dihitung menggunakan *teorema Bayes*:

$$P(C_k|X) = \frac{P(C_k)\prod_{i=1}^{n} P(x_i|C_k)}{P(X)} \quad (7)$$

Dimana C_k adalah kelas sentiment (misalnya, positif atau negatif), dan x_i adalah fitur.

F. Evaluasi dan Validasi Model

Model akan dievaluasi menggunakan *K-Fold Cross Validation* untuk memaksimalkan validitas hasil[25]. Setiap model akan diukur kinerjanya berdasarkan:

- *Akurasi:* Presentase klasifikasi yang benar dari keseluruhan data.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FX} \quad (8)$$

- *Presisi:* Proporsi prediksi yang tepat terhadap sentiment tertentu dari semua yang diprediksi memiliki sentiment tersebut.

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

- *Recall:* Proporsi ulasan dengan sentimen tertentu yang berhasil diklasifikasikan dengan benar.

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

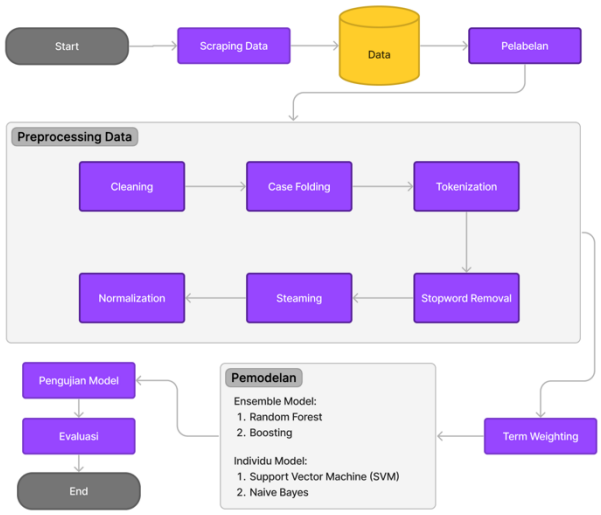
- *F1-Score:* Rata-rata harmonis antara *presisi* dan *recall*.

$$F1 = 2.\frac{precision\ .\ recall}{precision+recall} \quad (11)$$

Dimana :
TP: *True Positive*, data actual yang bernilai positif diprediksi benar
TN: *True Negative*, data actual yang bernilai negatif diprediksi benar
FP: *False Positive*, data actual yang bernilai negatif diprediksi positif
FN: *False Negative*, data actual yang bernilai positif diprediksi negative

Selain itu, metrik ROC-AUC akan digunakan untuk mengukur kemampuan model dalam membedakan antara ulasan positif dan negatif.

Alur Penelitian digambarkan pada Gambar 1:



Gambar 1. Alur Penelitian

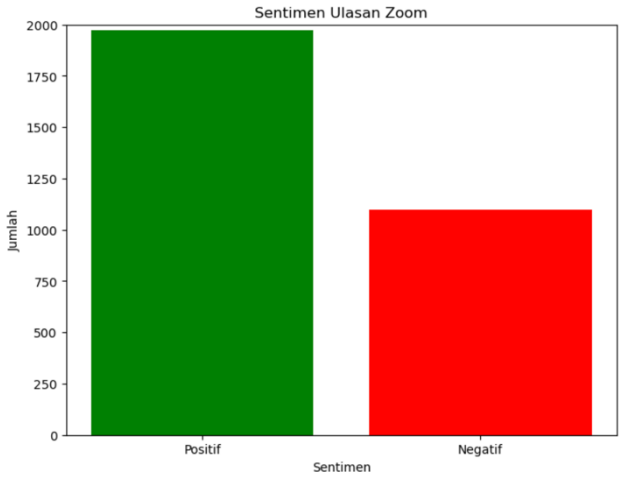
III. HASIL DAN PEMBAHASAN

A. *Pengumpulan Data*

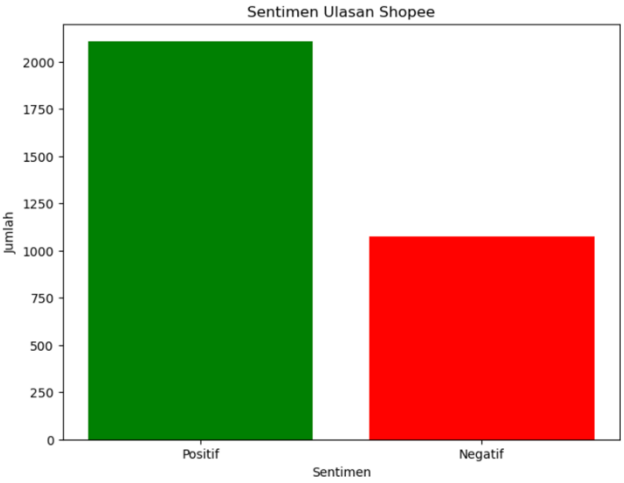
Dalam penelitian ini, teknik *scraping Google Play Store* digunakan untuk mengumpulkan data. Proses ini terfokus pada dua kategori aplikasi: *e-commerce* dan video konferensi. Untuk kategori *e-commerce*, aplikasi *Shopee* dipilih sebagai subjek pengumpulan data. Melalui teknik *scraping*, berhasil diperoleh sebanyak 3.184 ulasan pengguna. Sementara itu, dalam kategori video konferensi, aplikasi *Zoom* menjadi fokus utama. Dari aplikasi ini, sebanyak 3.073 ulasan berhasil dikumpulkan. Metode scraping yang efisien memungkinkan pengambilan data secara langsung dari *Google Play Store*, memastikan bahwa data yang dihasilkan adalah aktual dan relevan dengan kondisi terkini dari masing-masing aplikasi.

B. *Pelabelan data*

Dalam rangka memahami sentimen pengguna terhadap aplikasi *Shopee* dan *Zoom*, dilakukan proses pelabelan data berdasarkan rating yang diberikan oleh pengguna dalam ulasan mereka di *Google Play Store*. Proses pelabelan ini dibagi menjadi dua kategori: positif dan negatif. Label positif diberikan untuk ulasan dengan skor antara 4 dan 5, menunjukkan kepuasan pengguna, sedangkan label negatif diberikan untuk ulasan dengan skor dari 1 sampai 3, menandakan ketidakpuasan pengguna. Hasil pelabelan untuk aplikasi *Shopee* menghasilkan 2,108 ulasan positif dan 1,076 ulasan negatif. Sementara itu, aplikasi *Zoom* menghasilkan 1,973 ulasan positif dan 1,100 ulasan negatif. Proses pelabelan ini penting untuk membedakan nuansa sentimen pengguna yang berkontribusi pada analisis sentimen lebih lanjut.



Gambar 2. Sentimen Ulasan Aplikasi *Zoom*



Gambar 3. Sentimen Ulasan Aplikasi *Shopee*

	userName	score	at	content	sentimen
0	Irwan Cervo	1	2024-05-05 21:12:28	Pelayanannya belum bisa di selesaikan saya bel...	Negatif
1	Karim Mul	1	2024-05-05 21:01:45	Sering error setiap mau beli, setiap mau mengg...	Negatif
2	Irvvan Syah	5	2024-05-05 20:37:17	Pelayanan Shopee sgt baik. Tapi tolong untuk p...	Positif
3	Embun	1	2024-05-05 20:19:11	Ukuran Huruf nya kurang kecil, kalo bisa kecil...	Negatif
4	surati suripno	1	2024-05-05 20:10:19	Kepada pihak shopee yang terhormat kenapa apk ...	Negatif
...	...	...	...	...	...
3179	Ara	2	2024-02-17 21:01:46	Udah hampir 2 tahun stay sama si oren ini tapi...	Negatif
3180	umi malikah	1	2024-02-17 16:08:48	Agak mengganggu soalnya setelah di update mala...	Negatif
3181	Farah Lath	4	2024-02-17 15:56:12	Sudah 2 minggu ini lambat aplikasinya, padahal...	Positif
3182	Agus Prayitno	4	2024-02-16 23:40:02	sip, nambah peningkatan, tinggal supportnya di...	Positif
3183	Dwi Maya Wijayanti	3	2024-01-23 22:10:37	Masih terkendala di bagian keranjang belanja y...	Negatif

Gambar 4. Data diberi label

C. *Preprocessing Data*

Setelah penyelesaian proses pelabelan data, tahap selanjutnya yang dilaksanakan adalah *preprocessing* data. Tahapan ini mencakup serangkaian proses pembersihan dan normalisasi data untuk memastikan kualitas analisis yang lebih baik. Proses *cleaning* melibatkan penghapusan karakter *non-esensial* seperti *URL*, *mention*, *hashtag*, spasi berlebih, serta angka. Selanjutnya, dilakukan *case folding* yang bertujuan untuk mengubah semua teks menjadi huruf kecil guna menghindari duplikasi yang tidak perlu akibat perbedaan kapitalisasi.

Proses *tokenization* memecah teks menjadi satuan terkecil, atau token, yang memudahkan analisis lanjutan. Pemisahan kata-kata ini diikuti dengan *stopword removal*, di mana kata-kata umum yang tidak memberikan informasi

signifikan, seperti "dn", "dgn", "yg", dihilangkan berdasarkan daftar *stopword* yang telah ditentukan sebelumnya. Proses *stemming* menggunakan *library Sastrawi* untuk Bahasa Indonesia juga diterapkan untuk mengurangi kata-kata ke bentuk dasar mereka.

Tahap akhir *preprocessing* adalah normalisasi, yang bertujuan untuk menyamakan variasi bahasa dan memperbaiki kesalahan ejaan yang terjadi. Proses-proses ini membantu dalam menghasilkan dataset yang seragam dan siap analisis. Hasil akhir dari tahapan *preprocessing* ini ditampilkan dalam Tabel 1, memberikan gambaran yang jelas tentang kondisi data sebelum masuk ke proses analisis lebih lanjut.

Tabel 1. Hasil Prepocessing

No	Ulasan
1	voucer bayar oke voucer diskon gratis ongkos kirim tinggal kalimantan diskon tidak main main beli sisi masuk aplikasi shopee buka situs lag delay lumayan pancing emosi saran tingkat halus jalan guna
2	versi bagus buka langsung boro boro belanja lihat produk saja susah
3	transaksi proses transfer tolak dengan kode nomor akun valid copot ganti email ganti nomor telepon coba aplikasi shopee tolak transaksi transfer tolong solusi shopee maaf kurang bintang bintang
4	terima kasih barang nya terima sesuai pesan paking aman rapi rekomendasi pokok jual respon kirim cepat kurir nya moga amanah ramah moga amanah terima kasih shopee
5	shopee penuh butuh jadi sumber potong belanja tambah expedisi khusus shopee maju tapi kalau cepat toko hubungi

D. TF-IDF

Dalam rangka menentukan bobot setiap kata pada data ulasan dari aplikasi *Shopee* dan *Zoom*, digunakan algoritma TF-IDF *(Term Frequency-Inverse Document Frequency)*. Proses ini menghasilkan matriks yang menggambarkan bobot setiap kata terhadap dokumen dalam dataset. Untuk data ulasan Shopee, matriks yang dihasilkan memiliki 5,081 atribut yang merepresentasikan berbagai kata dari total 3,184 data ulasan yang dianalisis. Sementara itu, untuk data ulasan Zoom, matriks yang terbentuk mencakup 2,401 atribut dari 3,073 data ulasan. Matriks-matriks ini merefleksikan pentingnya setiap kata dalam menentukan nuansa sentiment pada ulasan, dan memberikan wawasan tentang distribusi frekuensi kata-kata kunci yang berperan dalam pembentukan opini pengguna terhadap kedua aplikasi tersebut.

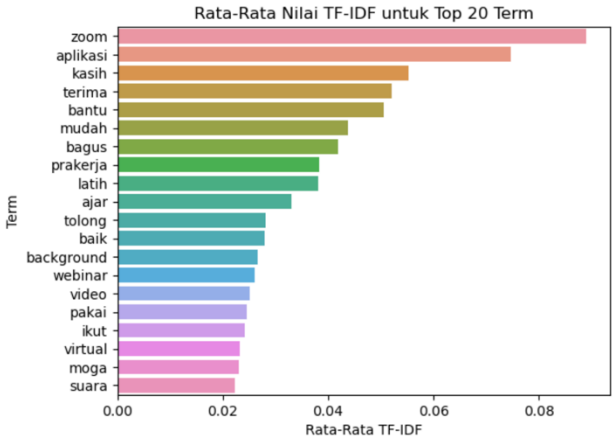
	Term	Average_TFIDF
2399	zoom	0.089010
108	aplikasi	0.074675
942	kasih	0.055252
2135	terima	0.052131
185	bantu	0.050614
...	...	...
1361	mnginstall	0.000048
1359	mngecewakan	0.000048
1357	mnampung	0.000048
1355	mmang	0.000048
1832	sbgian	0.000048

2401 rows × 2 columns

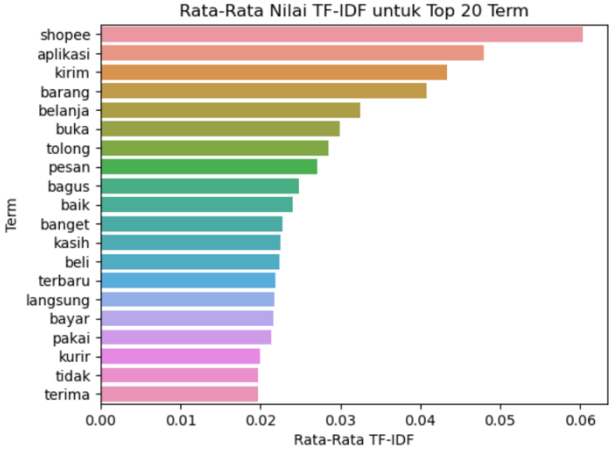
	Term	Average_TFIDF
4149	shopee	0.060383
221	aplikasi	0.047980
2165	kirim	0.043395
361	barang	0.040819
430	belanja	0.032532
...	...	...
284	ayam	0.000049
2958	nasi	0.000049
1494	geprek	0.000049
2786	mie	0.000049
2239	kontra	0.000049

5081 rows × 2 columns

Gambar 5. Pembobotan kata pada aplikasi Zoom dan Shopee



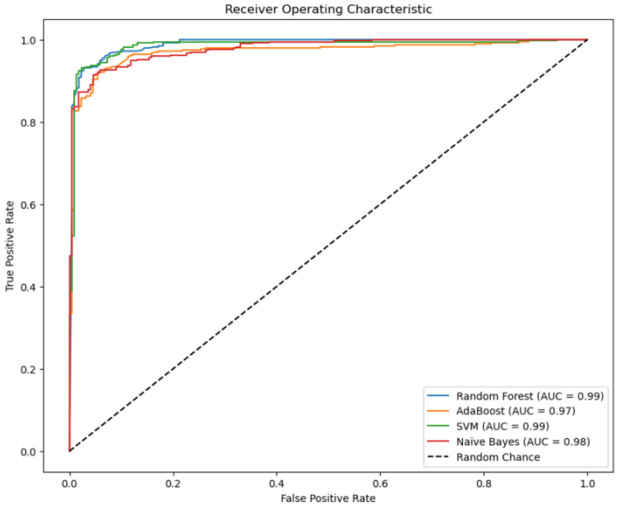
Gambar 6. Rata-rata Nilai TF-IDF pada aplikasi Zoom



Gambar 7. Rata-rata Nilai TF-IDF pada aplikasi Shopee

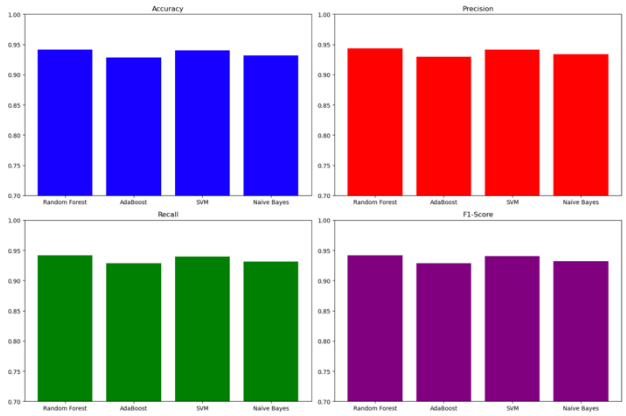
E. Evaluasi

Dalam penelitian ini, kami telah melakukan evaluasi komprehensif terhadap beberapa model pembelajaran mesin untuk menganalisis sentimen dari ulasan aplikasi di *Google Play Store*, khususnya aplikasi *Shopee* dan *Zoom*. Evaluasi dilakukan berdasarkan metrik akurasi, presisi, *recall*, dan *F1-score*, yang memberikan wawasan mendalam mengenai efektivitas setiap model.



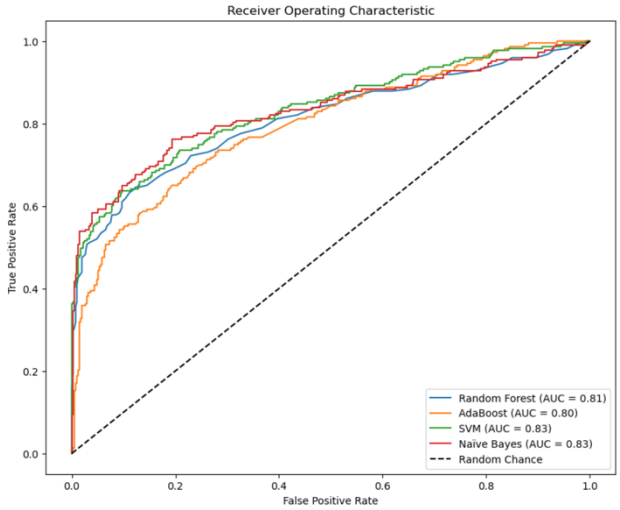
Gambar 8. Grafik ROC-AUC analisis sentiment Data Ulasan Zoom

Gambar 8. menunjukkan *Receiver Operating Characteristic* (ROC) dengan *Area Under Curve* (AUC) untuk berbagai model klasifikasi yang digunakan dalam analisis sentimen ulasan aplikasi *Zoom*. *Random Forest* dan SVM menunjukkan kinerja yang luar biasa dengan AUC mendekati 0.99, menandakan kemampuan yang sangat baik dalam membedakan antara kelas positif dan negatif. *AdaBoost* dan *Naive Bayes* juga menunjukkan kinerja yang sangat kompetitif dengan AUC di atas 0.97. Performa ini menegaskan efektivitas metode *ensemble* dan teknik klasifikasi individu dalam konteks analisis sentimen yang kompleks.



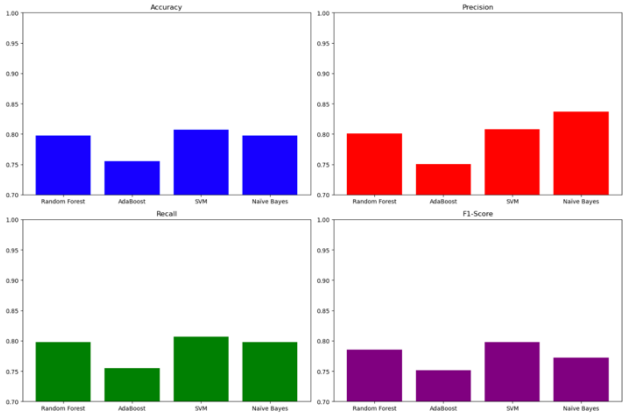
Gambar 9. Grafik Hasil Analisis Data Ulasan Zoom

Gambar 9. memvisualisasikan hasil evaluasi empat model yang digunakan pada dataset ulasan Zoom, mencakup metrik akurasi, presisi, *recall*, dan *F1-score*. *Random Forest* mengungguli model lain dengan mencatat akurasi tertinggi sebesar 94.15% dan *F1-Score* 94.19%, mengindikasikan bahwa model ini sangat cocok untuk mengatasi varian data yang kompleks dalam ulasan aplikasi *streaming* dan konferensi. Model SVM dan *Naïve Bayes* juga menunjukkan hasil yang impresif dengan akurasi di atas 93%, memperkuat argumen bahwa teknik yang lebih canggih diperlukan untuk mendekati keakuratan yang sempurna dalam klasifikasi teks.



Gambar 10. Grafik ROC-AUC analisis sentiment Data Ulasan Shopee

Pada Gambar 10. kita dapat melihat kinerja model klasifikasi yang digunakan untuk analisis sentimen ulasan *Shopee*. Berbeda dengan aplikasi *Zoom*, model di sini memiliki AUC yang lebih rendah, dengan *Random Forest* dan SVM masing-masing mencatat 0.81 dan 0.83. Ini menunjukkan adanya tantangan dalam dataset yang mungkin memerlukan optimasi model atau pendekatan *preprocessing* yang lebih mendalam untuk meningkatkan keakuratan prediksi.



Gambar 11. Grafik Hasil Analisis Data Ulasan Shopee

Pada Gambar 11. memperlihatkan evaluasi metrik akurasi, presisi, *recall*, dan *F1-score* untuk model yang digunakan dalam dataset ulasan *Shopee*. Meskipun performanya tidak seoptimal pada dataset *Zoom*, model SVM menunjukkan performa yang paling baik dengan akurasi sebesar 80.69%, yang menandakan kemampuan tinggi dalam mengklasifikasikan sentimen ulasan secara akurat. Meskipun begitu, model *Random Forest* dan *Naïve Bayes* juga memberikan hasil yang hampir serupa, dengan akurasi 79.75%. Namun, *Naïve Bayes* unggul dalam hal presisi, yaitu sebesar 83.70%, menunjukkan keefektifan model ini dalam mengidentifikasi label kelas dengan tepat. Model *AdaBoost*, meskipun memiliki akurasi lebih rendah, yaitu 75.51%, tetap memberikan kontribusi penting dalam analisis ini dengan menggarisbawahi pentingnya adaptabilitas dalam kondisi data yang beragam.

Hasil evaluasi ini mendukung hipotesis penelitian bahwa pendekatan *ensemble* memberikan akurasi klasifikasi yang lebih baik daripada algoritma individu. Lebih lanjut, hasil ini juga menekankan pentingnya langkah *preprocessing* dalam meningkatkan kinerja model. Langkah-langkah seperti *case folding*, *stemming*, dan normalisasi bukan hanya membantu dalam mengurangi kerumitan data tetapi juga meningkatkan konsistensi dan akurasi hasil analisis sentimen.

IV. KESIMPULAN

Dalam penelitian ini, kami telah menguji keefektifan pendekatan *ensemble* dalam analisis sentimen ulasan aplikasi di *Google Play Store*, dengan fokus pada aplikasi *e-commerce Shopee* dan aplikasi video konferensi *Zoom*. Hasil penelitian menunjukkan bahwa model *ensemble*, khususnya *Random Forest* dan SVM, menawarkan peningkatan signifikan dalam akurasi klasifikasi sentimen dibandingkan dengan algoritma individu. Model *Random Forest* pada ulasan aplikasi *Zoom* mencapai akurasi tertinggi sebesar 94.15%, presisi 94.36%, dan *F1-Score* 94.19%, sementara SVM pada aplikasi *Shopee* mencatat akurasi 80.69%, presisi 80.76%, dan *F1-Score* 79.79%. Penggunaan teknik

preprocessing yang meliputi *cleaning, case folding, tokenization, stopword removal, steaming*, dan normalisasi telah terbukti fundamental dalam mempersiapkan data untuk analisis lebih lanjut, memungkinkan model untuk lebih akurat dalam mengklasifikasikan sentimen. *Preprocessing* ini membantu dalam mengurangi variasi dan kesalahan data yang bisa mengganggu proses klasifikasi. Tujuan penelitian untuk meningkatkan efisiensi dan akurasi analisis sentimen melalui pendekatan *ensemble* telah tercapai, dengan pengamatan bahwa pendekatan ini lebih superior dalam mengatasi kompleksitas data ulasan dibandingkan algoritma yang beroperasi secara individu. Pendekatan ini tidak hanya memperluas pemahaman tentang dinamika opini pengguna tetapi juga memberikan wawasan bagi pengembang aplikasi untuk meningkatkan produk mereka berdasarkan umpan balik pengguna. Namun, terdapat ruang untuk peningkatan, terutama dalam pengoptimalan model untuk data yang sangat tidak seimbang dan dalam pengembangan algoritma yang lebih adaptif yang bisa secara efektif menangani nuansa bahasa yang lebih beragam dan subtansial. Penelitian masa depan dapat mengeksplorasi penggunaan model *Deep Learning* yang lebih canggih, seperti *neural networks*, untuk meningkatkan pemahaman tentang konteks lebih jauh dan aplikasi teknik *Natural Language Processing* (NLP) yang lebih baru. Saran untuk penelitian mendatang termasuk integrasi model *ensemble* dengan teknologi pemrosesan bahasa alami canggih, dan pengujian lintas domain untuk menilai keberlakuan model ini dalam berbagai jenis data ulasan. Selain itu, studi mendatang dapat memfokuskan pada pengembangan metode *preprocessing* yang lebih otomatis yang dapat menyesuaikan diri dengan karakteristik unik dari data ulasan yang berbeda.

REFFERENSI

[1] I. S. K. Idris, Y. A. Mustofa, and I. A. Salihi, "Analisis Sentimen Terhadap Penggunaan Aplikasi Shopee Mengunakan Algoritma Support Vector Machine (SVM)," *Jambura J. Electr. Electron. Eng.*, vol. 5, no. 1, pp. 32–35, 2023, doi: 10.37905/jjeee.v5i1.16830.

[2] T. N. Wijaya, R. Indriati, and M. N. Muzaki, "Analisis Sentimen Opini Publik Tentang Undang-Undang Cipta Kerja Pada Twitter," *Jambura J. Electr. Electron. Eng.*, vol. 3, no. 2, pp. 78–83, 2021, doi: 10.37905/jjeee.v3i2.10885.

[3] P. Yang, Y. H. Yang, B. B. Zhou, and A. Y. Zomaya, "A Review of Ensemble Methods in Bioinformatics," *Curr. Bioinform.*, 2010, doi: 10.2174/157489310794072508.

[4] Y. B. Lasotte, E. J. Garba, Y. M. Malgwi, and M. A. Buhari, "An Ensemble Machine Learning Approach for Fake News Detection and Classification Using a Soft Voting Classifier," *Eur. J. Electr. Eng. Comput. Sci.*, 2022, doi: 10.24018/ejece.2022.6.2.409.

[5] M. J. Sai, P. Chettri, R. Panigrahi, A. Garg, A. K. Bhoi, and P. Barsocchi, "An Ensemble of Light Gradient Boosting Machine and Adaptive Boosting for Prediction of Type-2 Diabetes," *Int. J. Comput. Intell. Syst.*, 2023, doi: 10.1007/s44196-023-00184-y.

[6] S. R. Puspita Sari Jan, Y. A. Mustofa, and I. S. K. Idris, "Analisis Sentimen Terhadap Data Kuisioner Evaluasi Dosen Menggunakan Algoritma Naïve Bayes," *J. Inform. Upgris*, vol. 9, no. 2, pp. 67–72, 2023, doi: 10.26877/jiu.v9i2.17001.

[7] W. Gata and A. Bayhaqy, "Analysis Sentiment About Islamophobia When Christchurch Attack on Social Media," *Telkomnika (Telecommunication Comput. Electron. Control.*, 2020, doi: 10.12928/telkomnika.v18i4.14179.

[8] A. W. Pradana and M. Hayaty, "The Effect of Stemming and Removal of Stopwords on the Accuracy of Sentiment Analysis on Indonesian-Language Texts," *Kinet. Game Technol. Inf. Syst. Comput. Netw. Comput. Electron. Control*, 2019, doi: 10.22219/kinetik.v4i4.912.

[9] S. Khomsah, A. F. Hidayatullah, and A. S. Aribowo, "Comparison of the Effects of Feature Selection and Tree-Based Ensemble Machine Learning for Sentiment Analysis on Indonesian YouTube Comments," pp. 161–172, 2021, doi: 10.1007/978-981-33-6926-9_15.

[10] E. Fersini, E. Messina, and F. Pozzi, "Sentiment Analysis: Bayesian Ensemble Learning," *Decis. Support Syst.*, 2014, doi: 10.1016/j.dss.2014.10.004.

[11] J. Kavanagh, K. A. Greenhow, and A. Jordanous, "Assessing the Effects of Lemmatisation and Spell Checking on Sentiment Analysis of Online Reviews," 2023, doi: 10.1109/icsc56153.2023.00046.

[12] A. A. A. Shamsi and S. Abdallah, "Sentiment Analysis of Emirati Dialect," *Big Data Cogn. Comput.*, 2022, doi: 10.3390/bdcc6020057.

[13] M. Maringer *et al.*, "User-Documented Food Consumption Data From Publicly Available Apps: An Analysis of Opportunities and Challenges for Nutrition Research," *Nutr. J.*, vol. 17, no. 1, 2018, doi: 10.1186/s12937-018-0366-6.

[14] A. Polhemus *et al.*, "Health Tracking via Mobile Apps for Depression Self-Management: Qualitative Content Analysis of User Reviews," *Jmir Hum. Factors*, vol. 9, no. 4, p. e40133, 2022, doi: 10.2196/40133.

[15] M. R. Dehkordi, "Dynamic PScore: A Dynamic Method to Prioritize User Reviews," 2023, doi: 10.21203/rs.3.rs-3790587/v1.

[16] S. N. Apsariny, S. Sediono, N. Chamidah, E. Ana, and A. Kurniawan, "Sentiment Analysis of User Reviews Based on Naïve Bayes Classifier Algorithm With Hyperparameter Optimization: A Case Study on Application 'Kredit Pintar,'" *Syntax Lit. J. Ilm. Indones.*, 2022, doi: 10.36418/syntax-literate.v7i1.6012.

[17] Y. A. Mustofa, "Data Ulasan Shopee dan Zoom app," 2024. https://www.kaggle.com/datasets/yasinarilmustofa/data-ulasan-shopee-and-zoom-app.

[18] H.-T. Duong and T.-A. Nguyen-Thi, "A Review: Preprocessing Techniques and Data Augmentation for Sentiment Analysis," *Comput. Soc. Networks*, 2021, doi: 10.1186/s40649-020-00080-x.

[19] A. Muhaddisi, "Sentiment Analysis With Sarcasm Detection on Politician's Instagram," *Ijccs (Indonesian J. Comput. Cybern. Syst.*, 2021, doi:

10.22146/ijccs.66375.

[20] J. Andoh, L. Asiedu, A. Lotsi, and C. Chapman-Wardy, "Statistical Analysis of Public Sentiment on the Ghanaian Government: A Machine Learning Approach," *Adv. Human-Computer Interact.*, 2021, doi: 10.1155/2021/5561204.

[21] N. Al-Twairesh and H. Al-Negheimish, "Surface and Deep Features Ensemble for Sentiment Analysis of Arabic Tweets," *Ieee Access*, 2019, doi: 10.1109/access.2019.2924314.

[22] S. Poria, H. Peng, A. Hussain, N. Howard, and Z. Wang, "Ensemble Application of Convolutional Neural Networks and Multiple Kernel Learning for Multimodal Sentiment Analysis," *Neurocomputing*, 2017, doi: 10.1016/j.neucom.2016.09.117.

[23] A. Alsayat, "Improving Sentiment Analysis for Social Media Applications Using an Ensemble Deep Learning Language Model," *Arab. J. Sci. Eng.*, 2021, doi: 10.1007/s13369-021-06227-w.

[24] W. Sharif *et al.*, "An Empirical Approach for Extreme Behavior Identification Through Tweets Using Machine Learning," *Appl. Sci.*, 2019, doi: 10.3390/app9183723.

[25] J. M. Deriu, M. Gonzenbach, F. Uzdilli, A. Lucchi, V. D. Luca, and M. Jaggi, "SwissCheese at SemEval-2016 Task 4: Sentiment Classification Using an Ensemble of Convolutional Neural Networks With Distant Supervision," 2016, doi: 10.18653/v1/s16-1173.