Jurnal Teknik Mesin, Elektro dan Ilmu Komputer(TEKNIK)
Vol.5, No.2 JULI 2025
P-ISSN : 2809-9095; E-ISSN : 2809-9125, Hal 84-96
DOI: https://doi.org/10.55606/teknik.v5i2.7433

PENERAPAN INFORMATION GAIN UNTUK SELEKSI FITUR PADA ALGORITMA NAÏVE BAYES UNTUK ANALISIS SENTIMEN IDENTITAS KEPENDUDUKAN DIGITAL

Yuliana Nogo Welan
Institut Keguruan dan Teknologi Larantuka

Alfian Nara Weking
Institut Keguruan dan Teknologi Larantuka

Dominikus Boli Watomakin
Institut Keguruan dan Teknologi Larantuka

Alamat: Jl. Ki Hajar Dewantara Kec. Larantuka - Kab. Flores Timur - Prov. Nusa Tenggara Timur

Korespondensi penulis: yulianawelan6@gmail.com

Abstract. *The rise of digital technology has driven the Indonesian government to implement Digital Population Identity (IKD) as a solution to enhance public services. However, user reviews on Google Play Store show diverse responses, requiring sentiment analysis to understand public perception. This study aims to improve sentiment classification accuracy on IKD app reviews using the Naïve Bayes algorithm optimized with Information Gain feature selection. The dataset consists of 1,000 Indonesian-language reviews manually labeled and preprocessed using text cleaning and TF-IDF feature representation. To address class imbalance, the SMOTE technique was applied. Experiments were conducted by comparing models without feature selection and balancing against those using Information Gain and SMOTE. Results indicate that the combination of Information Gain and SMOTE significantly enhances model performance, achieving 68,5% accuracy and 53,0% positive F1-Score. These findings confirm that Information Gain is effective in improving sentiment classification efficiency and accuracy. This study provides valuable insights for developing strategies to improve digital service quality.*
Keywords *: Feature Selection, Information Gain, Naïve Bayes, Sentiment Analysis, SMOTE.*

Abstrak. Pemerintah Indonesia telah menerapkan Identitas Kependudukan Digital (IKD) sebagai upaya meningkatkan pelayanan publik di era digital. Namun, beragam respons dari pengguna di Google Play Store menunjukkan perlunya analisis sentimen

Received May 13, 2025; Revised May 17, 2025; Accepted June 28, 2025
**Corresponding author, e-mail address*

untuk memahami persepsi masyarakat. Tujuan penelitian ini adalah meningkatkan akurasi klasifikasi sentimen pada ulasan aplikasi IKD dengan menggunakan algoritma Naïve Bayes yang dioptimalkan melalui seleksi fitur Information Gain. Dataset yang digunakan terdiri dari 1.000 ulasan berbahasa Indonesia yang telah dilabeli secara manual dan diproses melalui tahapan pre-processing teks serta representasi fitur TF-IDF. Untuk mengatasi masalah ketidakseimbangan data, teknik SMOTE digunakan dalam penelitian ini. Eksperimen dilakukan dengan membandingkan kinerja model tanpa seleksi fitur dan balancing dengan model yang menggunakan seleksi fitur Information Gain dan SMOTE. Hasil penelitian menunjukkan bahwa kombinasi Information Gain dan SMOTE secara signifikan meningkatkan performa model, dengan akurasi mencapai 68,5% dan F1-Score positif sebesar 53,0%. Temuan ini mengindikasikan bahwa seleksi fitur Information Gain efektif dalam meningkatkan efisiensi dan akurasi klasifikasi sentimen, sehingga dapat menjadi referensi dalam pengembangan strategi peningkatan kualitas layanan digital.
Kata kunci: **Analisis Sentimen, Information Gain, Naïve Bayes, Seleksi Fitur, SMOTE.**

LATAR BELAKANG

Transformasi digital di Indonesia telah membawa perubahan signifikan dalam layanan publik, termasuk melalui implementasi Identitas Kependudukan Digital (IKD). Aplikasi seluler ini, yang dikembangkan oleh Direktorat Jenderal Dukcapil, bertujuan untuk mempermudah akses data kependudukan dan meningkatkan kualitas layanan publik. Dengan mengimplementasikan Identitas Kependudukan Digital (IKD), masyarakat dapat lebih mudah mengakses layanan administrasi tanpa perlu membawa dokumen fisik, sehingga meningkatkan efisiensi dan menghemat biaya (Lestari et al., 2023) . Di daerah seperti Kabupaten Flores Timur, teknologi ini sangat krusial untuk mengatasi tantangan geografis dan memperluas akses layanan publik.

Meskipun aplikasi IKD menawarkan banyak manfaat, implementasinya juga dihadapkan pada beberapa tantangan. Masyarakat sering kali melaporkan masalah seperti kesulitan verifikasi wajah, masalah login, dan kualitas antarmuka aplikasi melalui ulasan di Google Play Store. Ulasan pengguna di platform digital seperti Google Play Store mencerminkan sentimen positif dan negatif terhadap layanan IKD. Untuk memahami persepsi masyarakat secara lebih mendalam, diperlukan pendekatan sistematis seperti analisis sentimen berbasis teks.

Analisis sentimen adalah metode yang efektif untuk menilai pandangan masyarakat terhadap suatu produk atau layanan berdasarkan teks yang mereka tulis. Dalam kasus aplikasi IKD, analisis ini dapat membantu mengidentifikasi masalah

teknis, kebutuhan pengguna, dan area pengembangan yang diperlukan. Analisis sentimen pada data teks, seperti ulasan pengguna, memiliki beberapa tantangan yang perlu diatasi. Di antaranya adalah tingginya dimensi fitur yang dihasilkan dari teks dan ketidakseimbangan dalam distribusi label sentimen, yang dapat mempengaruhi akurasi hasil analisis (Utami et al., 2015).

Penelitian sebelumnya telah menguji berbagai algoritma klasifikasi, termasuk Naïve Bayes, Support Vector Machine (SVM), dan model deep learning seperti BERT. Meskipun ada banyak pilihan, Naïve Bayes tetap populer karena kesederhanaannya dan efisiensi komputasi dalam analisis teks. Namun, performa algoritma ini sangat bergantung pada kualitas dan relevansi fitur yang dipilih. Oleh karena itu, proses seleksi fitur menjadi krusial untuk menghilangkan noise dan meningkatkan akurasi model. Information Gain adalah salah satu teknik yang populer digunakan untuk seleksi fitur, yang mengukur seberapa besar kontribusi setiap fitur terhadap hasil klasifikasi (Harahap, 2021). Selain itu, distribusi sentimen pada data ulasan aplikasi IKD yang tidak seimbang, dengan ulasan negatif yang lebih banyak daripada ulasan positif, menjadi tantangan tersendiri dalam proses pelatihan model klasifikasi. Untuk mengatasi masalah distribusi kelas yang tidak seimbang, teknik Synthetic Minority Over-sampling Technique (SMOTE) digunakan untuk meningkatkan jumlah data pada kelas minoritas, sehingga model dapat lebih baik dalam mengenali pola pada kelas tersebut (Farhan et al., 2025).

Tujuan dari penelitian ini adalah untuk menilai bagaimana pengaruh seleksi fitur dengan Information Gain terhadap kinerja algoritma Naïve Bayes dalam klasifikasi sentimen ulasan aplikasi IKD. Selain itu, penelitian ini juga menganalisis dampak penggunaan teknik SMOTE dalam mengatasi masalah distribusi data yang tidak seimbang. Kontribusi utama penelitian ini adalah pengembangan pendekatan analisis sentimen yang lebih akurat dan efisien melalui kombinasi teknik seleksi fitur dan balancing data. Hasil penelitian ini diharapkan dapat membantu pengembang dan pemangku kebijakan dalam meningkatkan kualitas layanan aplikasi IKD, serta memberikan kontribusi pada bidang text mining dan kecerdasan buatan dalam layanan publik digital.

KAJIAN TEORITIS

1. **Analisis sentimen**

   Analisis sentimen adalah metode yang digunakan untuk menganalisis pendapat, sentimen, evaluasi, sikap, penilaian, dan emosi terkait dengan produk, layanan, individu, atau kegiatan tertentu. Tujuannya adalah untuk mengidentifikasi opini seseorang berdasarkan topik tertentu. Proses analisis sentimen melibatkan beberapa tahapan, seperti pengidentifikasian domain dataset, preprocessing, pemilihan fitur, pelabelan, klasifikasi, dan evaluasi. Analisis sentimen memanfaatkan teknik Natural Language Processing (NLP), analisis teks, dan Linguistik Komputasi untuk mengekstraksi dan menganalisis informasi subjektif dari berbagai sumber, termasuk media sosial (Farhan et al., 2025).

2. **Media Sosial**

   Media sosial adalah platform daring yang memungkinkan pengguna untuk berpartisipasi, berinteraksi, berbagi, dan melakukan berbagai aktivitas dalam dunia virtual tanpa batasan ruang dan waktu. Contoh media sosial termasuk platform seperti Facebook, Twitter, Instagram, dan lainnya (Yunarfi et al., 2021).

3. **Text Mining**

   Text mining adalah teknik algoritmik berbasis komputer yang digunakan untuk mendapatkan pengetahuan baru dari sekumpulan teks. Text mining merupakan bagian dari sekumpulan informasi yang bekerja pada data teks tidak terstruktur. Meskipun memiliki kemiripan dengan data mining, perbedaan utama antara keduanya terletak pada tipe data yang digunakan, di mana text mining fokus pada data tidak terstruktur sedangkan data mining pada data terstruktur. Kombinasi keduanya dapat digunakan untuk menyelesaikan masalah klasifikasi, klastering, dan prediksi pada informasi tekstual (Priyanto & Ma'arif, 2018).

4. Kependudukan Digital

Identitas Kependudukan Digital (IKD) adalah bentuk digital dari data kependudukan yang dikemas dalam aplikasi perangkat seluler, menggunakan foto atau QR Code. IKD dirancang untuk mengurangi kebutuhan dokumen fisik seperti Kartu Tanda Penduduk (KTP) dalam administrasi dan menghemat anggaran. Aplikasi ini terintegrasi dengan berbagai layanan seperti kesehatan, pendidikan, perbankan, dan pajak melalui aktivasi tunggal, sehingga memudahkan penggunaannya di era digital (Wahyuningsih & Hendry, 2023).

5. Algoritma Naïve Bayes

Naïve Bayes Classifier merupakan teknik prediksi berbasis probabilistik sederhana dengan asumsi independensi yang kuat pada fitur, dalam artian sebuah fitur pada sebuah data tidak berkaitan dengan ada atau tidaknya fitur lain dalam data yang sama. Ciri utama dari metode Naïve Bayes ini adalah asumsi yang kuat akan independensi dari masing-masing kondisi (Harahap, 2021).

6. Ekstraksi Fitur TF-IDF (Term Frequency-Inverse Document Frequency).

Metode TF-IDF adalah suatu cara untuk untuk memberikan bobot hubungan suatu kata (term) terhadap dokumen. Metode ini menggabungkan dua konsep untuk perhitungan bobot yaitu frekuensi kemunculan sebuah kata didalam sebuah dokumen tertentu dan inverse frekuensi dokumen yang mengandung kata tersebut. Pembobotan ini menunjukan bawah dapat menghasilkan akurasi yang cukup tinggi (Afifi, 2022).

7. Seleksi Fitur dalam Analisis Sentimen

Seleksi fitur adalah proses penting dalamanalisis sentimen yang bertujuan memilih subset fitur yang paling relevan dari seluruh fitur yang tersedia dalam teks. Fitur-fitur ini bisa berupa kata-kata, frasa, atau karakteristik linguistik tertentu yang berkontribusi pada prediksi sentimen. Dengan menerapkan seleksi fitur, model analisis sentimen menjadi lebih efisien dan akurat karena hanya menggunakan fitur-fitur yang paling informatif. Proses ini meningkatkan performa dan efisiensi model secara keseluruhan (Edwar et al., 2023).

8. Information Gain

Information Gain adalah teknik seleksi fitur yang menggunakan metode *scoring* dalam menentukan nominal ataupun pembobotan atribut kontinu yang didiskretkan menggunakan nilai entropy. (Harahap, 2021). Python

9. Python

Python adalah bahasa pemrograman interpretatif yang dikenal mudah dipelajari dan memiliki keterbacaan kode yang baik. Bahasa ini mendukung berbagai paradigma pemrograman seperti berorientasi objek, imperatif, dan fungsional. Beberapa kelebihan Python meliputi koleksi kepustakaan yang luas dengan modul-modul siap pakai, struktur bahasa yang sederhana dan jelas, pengelolaan memori otomatis, serta sifat modular yang memudahkan pengembangan dan pengelolaan kode (Romzi1 & Budi Kurniawan2, 2018)

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan desain eksperimental untuk mengkaji pengaruh seleksi fitur Information Gain terhadap kinerja algoritma Naïve Bayes dalam klasifikasi sentimen ulasan aplikasi IKD. Sumber data penelitian ini adalah 1.000 ulasan pengguna aplikasi IKD yang dikumpulkan melalui web scraping dari Google Play Store pada April hingga Mei 2024. Ulasan yang dipilih merupakan komentar berbahasa Indonesia yang dinilai populer dan relevan dengan layanan IKD. Setelah pengumpulan data, ulasan pengguna dikategorikan menjadi sentimen positif dan negatif melalui proses pelabelan manual oleh tiga orang pakar. Langkah ini bertujuan untuk memastikan validitas dan konsistensi data yang digunakan dalam penelitian. Selanjutnya, dilakukan pre-processing data yang mencakup cleansing, case folding, tokenisasi, penghapusan stopword, dan filtering token berdasarkan panjang kata. Semua proses ini diimplementasikan menggunakan Python di Google Colab. Fitur teks kemudian diekstraksi menggunakan metode TF-IDF, yang menghasilkan representasi numerik dari dokumen dalam bentuk vektor dengan total 3568 fitur. Selanjutnya, seleksi fitur dilakukan menggunakan metode Information Gain untuk meningkatkan efisiensi dan akurasi model klasifikasi. Penelitian ini membandingkan dua seleksi fitur yaitu tanpa seleksi fitur (seluruh fitur TF-IDF digunakan) dan dengan seleksi fitur (1.500 fitur teratas berdasarkan skor Information Gain). Untuk mengatasi masalah ketidakseimbangan

distribusi kelas sentimen, teknik SMOTE (Synthetic Minority Over-sampling Technique) diterapkan pada data latih.

Teknik SMOTE digunakan untuk menghasilkan data sintetis pada kelas minoritas (positif) untuk meningkatkan proporsinya dan mengurangi bias model terhadap kelas mayoritas. Pembagian data dilakukan dengan rasio 80:20 untuk pelatihan dan pengujian model. Algoritma Multinomial Naïve Bayes dipilih sebagai model klasifikasi karena kemampuannya yang efektif dalam menangani data teks yang berbasis frekuensi kata. Model dilatih dengan berbagai konfigurasi fitur dan dievaluasi menggunakan metrik akurasi, precision, recall, dan F1-score, dengan fokus pada sentimen positif sebagai kelas minoritas. Perbandingan hasil model dilakukan untuk menilai dampak seleksi fitur dan balancing data terhadap performa klasifikasi.

HASIL DAN PEMBAHASAN

1. Pelabelan

Penelitian ini mengumpulkan 1.000 ulasan pengguna aplikasi IKD dari Google Play Store, yang kemudian dilabeli secara manual oleh tiga pakar menjadi dua kategori sentimen, yaitu positif dan negatif. Hasil dari pelabelan sentimen negatif 717 dan positif 283. Dalam ulasan menunjukkan adanya ketidakseimbangan data yang signifikan karena lebih dominan ke data ulsan negatif.

Tabel 1. Hasil Pelabelan Manual

No	Conten	Label
1	Sebenarnya aplikasi ini membantu, hanya kenapa suka error, saya mengajukan permohonan untuk mencetak KK baru karena ada penambahan anggota keluarga tidak di verifikasi oleh admin, 1 bulan tidak ada kemajuan dari permohonan saya	Positif
2	habis update. minta login . suruh scan wajah . tombol kamera gak berfungsi . sama dengan user lainnya . masalah masih di identifikasi wajah pakai kamera.	Negatif
3	Saya harap Aplikasinya segera di Update atau diperbaharui kembali, Lalu kualitas Gambar KTP dengan KK atau Dokumen lain Resolusinya di perbagus kualitas Jernih HD, soalnya bentuk Tanda Tangan nya jadi ngeblur Hitam Putih, kualitasnya jelek gambarnya nah. Terus halaman bentuk aplikasi & Logo dipercantik lagi, dari segi manfaat tolong ditambah karena belum bisa pakai KTP Digital untuk Verifikasi Data Pribadi atau pengurusan Layanan Online maupun Offline... walaupun Uji Coba tapi tetap wajib baik.	Positif
4	Ribet. Ktp buram, buat daftar keperluan mendesak gak bisa karena blanko hbs. Trs denger berita bisa pakai digital. Pas masuk malah gak terdaftar. Ini gimana sih. Katanya melek teknologi, tp ssshh mbuhlaaahh. diperbaharui malah masuk lagi wkwk	Negatif

…	aplikasi dan di dunia nyata pelayanan nya sama saja sangat buruk , untuk cetak ektp saja ,birokrasi lempar sana sini ,kirain adanya aplikasi dan teknologi membuat masyarakat cepat,mudah dalam mengurus administrasi kependudukan tetapi dugaan saya salah semua masih sama saja	Negatif
1000	Saya harap Aplikasinya segera di Update atau diperbaharui kembali, Lalu kualitas Gambar KTP dengan KK atau Dokumen lain Resolusinya di perbagus kualitas Jernih HD, soalnya bentuk Tanda Tangan nya jadi ngeblur Hitam Putih, kualitasnya jelek gambarnya nah. Terus halaman bentuk aplikasi & Logo dipercantik lagi, dari segi manfaat tolong ditambah karena belum bisa pakai KTP Digital untuk Verifikasi Data Pribadi atau pengurusan Layanan Online maupun Offline... walaupun Uji Coba tapi tetap wajib baik.	positif

2. Pre-processing Data

Dalam pre-processing terdapat beberapa tahapan penting, yaitu cleansing, case folding, tokenisasi, penghapusan stopword, dan filtering token berdasarkan panjang karakter.

a. Proses cleansing bertujuan untuk menghilangkan tanda baca, simbol, dan karakter yang tidak relevan dari teks, yang hasilnya dapat dilihat pada Gambar 1.

	content_lower	content_clean
0	sebenarnya aplikasi ini membantu, hanya kenapa...	sebenarnya aplikasi ini membantu hanya kenapa ...
1	habis update. minta login . suruh scan wajah	habis update minta login suruh scan wajah tomb...
2	aplikasi kocak, ngapain online kalau masih per...	aplikasi kocak ngapain online kalau masih perl...
3	saya kasih bintang 2 dulu ya. karena belum mem...	saya kasih bintang dulu ya karena belum memuda...
4	ribet. ktp buram, buat daftar keperluan mendes...	ribet ktp buram buat daftar keperluan mendesak...

Gambar 1. Hasil Cleansing

b. Transform Case digunakan untuk mengubah semua teks menjadi huruf kecil, sehingga menyeragamkan format teks, seperti ditunjukkan pada Gambar 2 di bawa ini:

```
Sebelum transform case:
0    sebenarnya aplikasi ini membantu, hanya kenapa...
1    habis update. minta login . suruh scan wajah ....
2    Aplikasi kocak, ngapain online kalau masih per...
3    saya kasih bintang 2 dulu ya. karena belum mem...
4    Ribet. Ktp buram, buat daftar keperluan mendes...
Name: content, dtype: object

Setelah transform case:
0    sebenarnya aplikasi ini membantu, hanya kenapa...
1    habis update. minta login . suruh scan wajah ....
2    aplikasi kocak, ngapain online kalau masih per...
3    saya kasih bintang 2 dulu ya. karena belum mem...
4    ribet. ktp buram, buat daftar keperluan mendes...
Name: content_lower, dtype: object
```

Gambar 2. Hasil Transform Case

c. Tokenisasi digunakan untuk memecah teks menjadi unit-unit yang lebih kecil, seperti kata atau token, yang hasilnya dapat dilihat pada gambar 3 di bawa ini:

	content_clean	content_token
0	sebenarnya aplikasi ini membantu hanya kenapa ...	[sebenarnya, aplikasi, ini, membantu, hanya, k...
1	habis update minta login suruh scan wajah tomb...	[habis, update, minta, login, suruh, scan, waj...
2	aplikasi kocak ngapain online kalau masih perl...	[aplikasi, kocak, ngapain, online, kalau, masi...
3	saya kasih bintang dulu ya karena belum memuda...	[saya, kasih, bintang, dulu, ya, karena, belum...
4	ribet ktp buram buat daftar keperluan mendesak...	[ribet, ktp, buram, buat, daftar, keperluan, m...

Gambar 3. Hasil Tokenisasi

d. Stopword berfungsi untuk menghilangkan kata-kata yang tidak bermakna, seperti yang terlihat pada gambar 4 di bawah ini:

	content_token_filtered	content_final
0	[sebenarnya, aplikasi, ini, membantu, hanya, k...	[sebenarnya, aplikasi, membantu, suka, error, ...
1	[habis, update, minta, login, suruh, scan, waj...	[habis, update, minta, login, suruh, scan, waj...
2	[aplikasi, kocak, ngapain, online, kalau, masi...	[aplikasi, kocak, ngapain, online, kalau, perl...
3	[saya, kasih, bintang, dulu, karena, belum, me...	[kasih, bintang, dulu, memudahkan, masyarakat,...
4	[ribet, ktp, buram, buat, daftar, keperluan, m...	[ribet, ktp, buram, buat, daftar, keperluan, m...

Gambar 4. Hasil Stopword

e. Tahap Filtering Token by Length digunakan untuk menghapus kata-kata yang terlalu pendek atau terlalu panjang berdasarkan panjang karakter yang dapat di lihat di bawa ini:

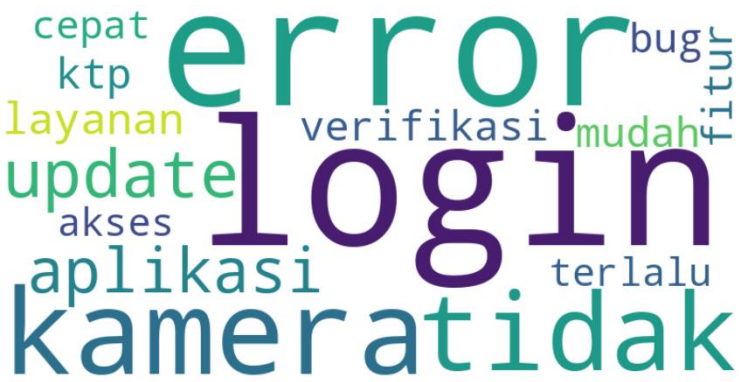
	content_token	content_token_filtered
0	[sebenarnya, aplikasi, ini, membantu, hanya, k...	[sebenarnya, aplikasi, ini, membantu, hanya, k...
1	[habis, update, minta, login, suruh, scan, waj...	[habis, update, minta, login, suruh, scan, waj...
2	[aplikasi, kocak, ngapain, online, kalau, masi...	[aplikasi, kocak, ngapain, online, kalau, masi...
3	[saya, kasih, bintang, dulu, ya, karena, belum...	[saya, kasih, bintang, dulu, karena, belum, me...
4	[ribet, ktp, buram, buat, daftar, keperluan, m...	[ribet, ktp, buram, buat, daftar, keperluan, m...

Gambar 5. Hasil Filtering Token by Length

3. Ekstraksi dan Seleksi Fitur

Setelah proses pembersihan data, dilakukan ekstraksi fitur menggunakan metode TF-IDF (Term Frequency-Inverse Document Frequency), yang menghasilkan 3.568 fitur kata unik dari seluruh dokumen. Jumlah fitur yang besar dapat menyebabkan overfitting dan meningkatkan kompleksitas komputasi. Untuk mengatasi hal ini, dilakukan seleksi fitur menggunakan metode Information Gain (IG) untuk memilih fitur-fitur yang paling informatif dalam membedakan sentimen. Sebanyak 1.500 fitur dengan skor IG tertinggi dipilih untuk pelatihan model. Fitur-fitur penting seperti error, verifikasi, login, mudah, dan kamera

memiliki nilai IG tinggi, yang mencerminkan fokus pengguna pada fungsi utama aplikasi IKD, terutama dalam proses pendaftaran, otentikasi, dan penggunaan kamera.



Gambar 6. Wordcloud fitur dengan nilai IG tertinggi.

4. Penyeimbangan Data dengan SMOTE

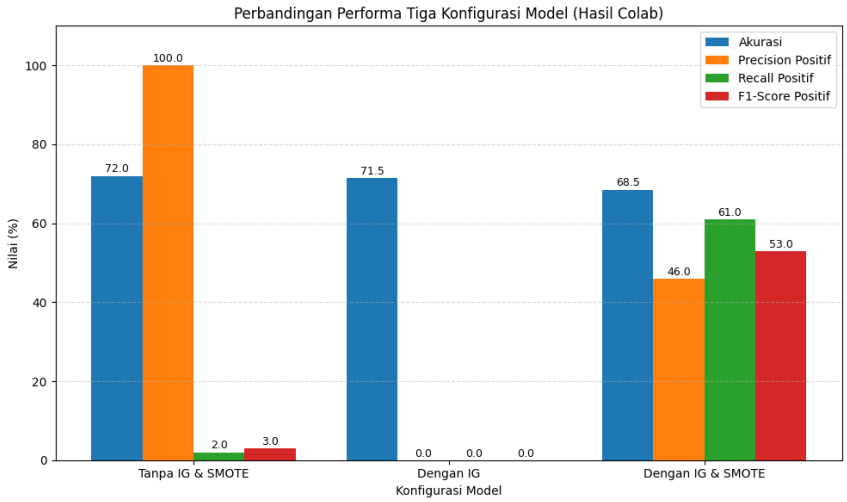
Karena distribusi data tidak seimbang, dengan kelas negatif yang lebih dominan, menyebabkan model menjadi bias. Untuk mengatasinya, teknik SMOTE (Synthetic Minority Over-sampling Technique) diterapkan pada data latih untuk menciptakan data sintetis dari kelas minoritas (positif) dan mencapai keseimbangan antar kelas. Dengan menerapan SMOTE, distribusi data menjadi seimbang, yaitu 574 data negatif dan 574 data positif. Dengan data yang seimbang, model dapat belajar secara lebih efektif dan adil untuk mengenali kedua kelas, tanpa bias terhadap salah satu kelas. Yang dapat di lihat pada tabel di bawa ini:

Tabel 1. Distribusi Data Latih Sebelum dan Sesudah SMOTE

No	Kondisi	Negatif	Positif	Keterangan
1	Data latih sebelum SMOTE	574	226	Distribusi data latih tidak seimbang. Model cenderung bias ke kelas negatif.
2	Data latih setelah SMOTE	574	574	Data latih seimbang. SMOTE menambahkan data sintetis di kelas positif hingga jumlahnya setara kelas negatif. Model dapat belajar lebih adil dalam mengenali kedua kelas.

5. Evaluasi dan Perbandingan Model

Evaluasi model Naïve Bayes dilakukan menggunakan metrik akurasi, precision, recall, dan F1-score pada kelas positif, dengan hasil evaluasi yang dapat di lihat pada gambar di bawa ini :

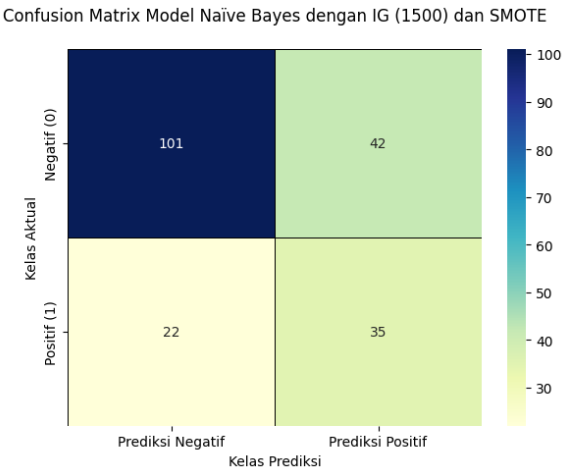


Gambar 7. Perbandingan Performa Naïve Bayes

Dari gambar diatas menunjukan bahwa Secara keseluruhan, penerapan Information Gain terbukti dapat meningkatkan performa model dengan menyaring fitur yang paling relevan, sedangkan penerapan SMOTE efektif dalam mengatasi ketidakseimbangan data dan meningkatkan performa model dalam mengklasifikasikan ulasan positif.

6. Confusion Matrix Model Terbaik

Penerapan Information Gain dan SMOTE pada model klasifikasi sentimen ulasan aplikasi Identitas Kependudukan Digital terbukti efektif dalam meningkatkan performa model, terutama pada kelas minoritas (positif). Hasil confusion matrix menunjukkan bahwa model dapat mengenali ulasan positif dengan lebih baik, meskipun masih ada kesalahan klasifikasi. Yang dapat di lihat pada gambar dibawa ini :



Gambar 8 Grafik Confusion Matrix

Penerapan Information Gain dan SMOTE terbukti saling melengkapi dalam meningkatkan performa model. Information Gain mengurangi fitur yang tidak penting, sementara SMOTE menyeimbangkan data dan meningkatkan sensitivitas model terhadap kelas minoritas. Hasilnya adalah model klasifikasi yang lebih akurat dan adil.

KESIMPULAN DAN SARAN

Penelitian ini menunjukkan bahwa kombinasi Information Gain untuk seleksi fitur dan SMOTE untuk balancing data dapat meningkatkan akurasi klasifikasi sentimen pada ulasan aplikasi Identitas Kependudukan Digital (IKD) dengan menggunakan algoritma Naïve Bayes. Hasil pengujian tiga konfigurasi model menunjukkan bahwa kombinasi Information Gain dan SMOTE memberikan performa terbaik dengan akurasi 68,5% dan F1-score positif sebesar 53,0%. Hal ini menunjukkan bahwa pemilihan fitur yang relevan dan penyeimbangan data yang tepat sangat penting dalam meningkatkan akurasi dan keadilan model klasifikasi terhadap kelas minoritas. Penelitian selanjutnya dapat difokuskan pada eksplorasi algoritma klasifikasi lain dan perbandingan metode seleksi fitur serta teknik balancing yang berbeda untuk meningkatkan performa model yang lebih optimal.

DAFTAR REFERENSI

Afifi, W. (2022). Analisis sentimen pengguna twitter terhadap layanan internet pt indosat tbk dengan metode k-nearest neighbor (k-nn) dan naive bayes classifier (nbc). In *Repository.Uinjkt.Ac.Id*.

Edwar, E., Semadi, I. G. A. N. R., Samsudin, M., & Dharmendra, I. K. (2023). Perbandingan Metode Seleksi Fitur Pada Analisis Sentimen (Studi Kasus Opini PILKADA DKI 2017). *INFORMATICS FOR EDUCATORS AND PROFESSIONAL : Journal of Informatics*, *8*(1), 11. https://doi.org/10.51211/itbi.v8i1.2408

Farhan, A., Rahman, A. Y., Informatika, T., Malang, U. W., & Store, G. P. (2025). *DIGITAL DI GOOGLE PLAY STORE DENGAN BERT*. *9*(3), 3776–3783.

Harahap, R. habibie. (2021). *analisis sentimen pembelajaran daring pada mahasiswa menggunakan metode maximum entropy dengan seleksi fitur information gain*.

Lestari, R. A., Erfina, A., & Jatmiko, W. (2023). Penerapan Algoritma Support Vector Machine pada Analisis Sentimen Terhadap Identitas Kependudukan Digital. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *10*(5), 1063–1070. https://doi.org/10.25126/jtiik.20231057264

Priyanto, A., & Ma'arif, M. R. (2018). Implementasi Web Scrapping dan Text Mining untuk Akuisisi dan Kategorisasi Informasi dari Internet (Studi Kasus: Tutorial Hidroponik). *Indonesian Journal of Information Systems*, *1*(1), 25–33. https://doi.org/10.24002/ijis.v1i1.1664

Romzi1, M., & Budi Kurniawan2. (2018). Implementasi Pemrograman Python Menggunakan Visual Studio Code. *Unnes*, 1–76. https://journal.unmaha.ac.id/index.php/jik/article/view/198/176

Utami, L. D., Tinggi, S., Informatika, M., Komputer, D., Mandiri, N., & Wahono, R. S. (2015). Integrasi Metode Information Gain Untuk Seleksi Fitur dan Adaboost Untuk Mengurangi Bias Pada Analisis Sentimen Review Restoran Menggunakan Algoritma Naïve Bayes. *Journal of Intelligent Systems*, *1*(2). http://journal.ilmukomputer.org

Wahyuningsih, N., & Hendry, H. (2023). Perbandingan Metode Klasifikasi Dalam Analisis Sentimen Masyarakat Terhadap Identitas Kependudukan Digital (Ikd). *JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, *8*(4), 1218–1227. https://doi.org/10.29100/jipi.v8i4.4155

Yunarfi, G. R., Ricky Simdy, & Jackson. (2021). Implementasi Text Mining untuk Mengetahui Kata Abreviasi dalam Percakapan Media Sosial. *Journal of Digital Ecosystem for Natural Sustainability (JoDENS)*, *1*(2), 78–83.