Prediksi Sentimen untuk Meningkatkan Layanan Publik dengan Analisis Ulasan Google Play dan CatBoost

1st Muhammad Raffy Ibnu Mustofa
Fakultas Informatika
S1 Data Sains Universitas Telkom
Bandung, Indonesia
raffyibnu45@gmail.com

2nd Muhammad Rizki Wiratama
Fakultas Informatika
S1 Data Sains Universitas Telkom
Bandung, Indonesia
rizkiwirat@gmail.com

3rd Muhammad Sya'bani Falif
Fakultas Informatika
S1 Data Sains Universitas Telkom
Bandung, Indonesia
msfalif404@gmail.com

***Abstract*—Pemerintah Indonesia telah meluncurkan aplikasi Identitas Kependudukan Digital untuk mendigitalkan KTP, sebagai langkah menuju negara maju. Aplikasi ini telah diunduh lebih dari 10 juta kali sejak 2023, memicu berbagai tanggapan pengguna yang positif, negatif, dan netral. Proyek ini menganalisis sentimen masyarakat terhadap aplikasi tersebut menggunakan metode machine learning, khususnya CatBoost. Data diperoleh dari komentar pengguna di PlayStore melalui proses crawling. Analisis ini bertujuan untuk memahami harapan, kekhawatiran, dan tanggapan pengguna, serta memberikan wawasan mendalam untuk perbaikan aplikasi dan memastikan penerimaan luas di masyarakat Indonesia.**

***Index Terms*—identitas kependudukan digital, analisis sentimen, opinion mining, CatBoost**

I. Introduction

Indonesia saat ini mencoba melakukan digitalisasi KTP yang merupakan salah satu langkah yang signifikan untuk menjadi negara maju. Dilansir dari CNBC, pada tahun sebelumnya, Kementerian Dalam Negeri (Kemendagri) sedang menyiapkan aplikasi yang dikenal sebagai Identitas Kependudukan Digital [1].

Pemerintah Indonesia mengeluarkan aplikasi ini dengan tujuan menggantikan KTP fisik dengan KTP digital atau e-KTP, yang dapat diakses melalui perangkat gawai dan menampilkan data pribadi sebagai identitas penduduk. Langkah digitalisasi KTP ini sejalan dengan upaya pemerintah dalam meningkatkan efisiensi dan keterjangkauan pelayanan kependudukan, mencerminkan komitmen negara ini untuk terus berinovasi dalam penerapan teknologi informasi.

Pada era dimana konektivitas digital semakin sering muncul dalam kehidupan sehari-hari, Aplikasi Identitas Kependudukan Digital menciptakan potensi untuk memudahkan proses identifikasi dan pelayanan publik secara lebih canggih. Melalui digitalisasi ini, diharapkan masyarakat Indonesia dapat menikmati manfaat dan kemudahan akses dalam berbagai layanan publik, menciptakan dasar yang kokoh untuk pemerintahan elektronik yang lebih terintegrasi.

Meskipun demikian, keberhasilan implementasi dan penggunaan aplikasi ini masih dipengaruhi oleh berbagai faktor, termasuk kesiapan teknologi, literasi digital masyarakat Indonesia, serta berbagai perubahan dalam pola pikir terkait keamanan dan privasi data

Analisis sentimen merupakan proses ekstraksi dan klasifikasi opini, emosi, dan sentimen dari teks. Teknik pemrosesan bahasa alami digunakan untuk menganalisis data teks dan mengidentifikasi sentimen positif, negatif, atau netral.

Menurut IBM [2], analisis sentimen mampu mengolah bahasa dalam unggahan di media sosial, tanggapan, ulasan, dan sejenisnya. Proses ini bertujuan untuk mengekstrak sikap dan emosi yang muncul dalam respons terhadap produk, promosi, dan acara. Informasi yang dihasilkan dapat menjadi bahan bagi perusahaan dalam merancang produk, kampanye iklan, dan berbagai keperluan lainnya.

Analisis sentimen masyarakat Indonesia terhadap Aplikasi Identitas Kependudukan Digital menjadi hal penting untuk memahami bagaimana aplikasi ini diterima dan digunakan oleh masyarakat. Dengan menggali sentimen ini, akan didapatkan beberapa sentimen yang mengidentifikasikan harapan, kekhawatiran, serta kendala yang mungkin dihadapi oleh pengguna aplikasi [3].

Untuk mendapatkan pemahaman yang lebih mendalam mengenai sentimen masyarakat terhadap aplikasi Identitas Kependudukan Digital, kami akan mengimplementasikan metode analisis sentimen menggunakan teknik machine learning, dengan fokus utama pada algoritma CatBoost. Metode ini akan membantu mengklasifikasikan opini, emosi, dan sentimen dari teks-teks yang diperoleh dari berbagai sumber, termasuk media sosial, ulasan, dan respons publik lainnya.

Setelah menerapkan metode analisis sentimen menggunakan CatBoost, hasil klasifikasi sentimen masyarakat terhadap Aplikasi Identitas Kependudukan Digital akan dievaluasi

menggunakan metrik-metrik tertentu. Melalui evaluasi kinerja metode seperti akurasi, presisi, recall, dan F1-score, akan diperoleh pemahaman yang lebih dalam tentang efektivitas dan ketepatan dalam mengidentifikasi tanggapan pengguna. Evaluasi berbasis metrik ini diharapkan dapat memberikan gambaran yang lebih komprehensif, memandu pemilihan metode terbaik, dan mengarahkan upaya perbaikan serta peningkatan analisis sentimen identitas kependudukan digital di Indonesia.

Proyek ini juga akan melibatkan aspek prediksi data baru, di mana akan dilakukan analisis sentimen secara berkala untuk merespons ulasan terbaru pada Aplikasi Identitas Kependudukan Digital. Pendekatan ini diimplementasikan untuk memberikan pandangan dinamis terhadap perubahan sentimen masyarakat, sehingga dapat membantu peningkatan aplikasi secara proaktif.

II. Pengumpulan Data

A. Deskripsi Data

Dataset yang digunakan dalam analisis sentimen terhadap aplikasi Identitas Kependudukan Digital (IKD) terdiri dari teks komentar pengguna, rating, tanggal, dan metadata terkait lainnya. Data ini mencakup berbagai tanggapan pengguna, baik positif, negatif, maupun netral, yang mencerminkan pengalaman dan opini mereka mengenai aplikasi IKD.

B. Sumber Data

Dataset untuk analisis sentimen terhadap aplikasi Identitas Kependudukan Digital (IKD) diperoleh melalui proses scrapping komentar pengguna di PlayStore. Proses ini melibatkan pengumpulan data dari ulasan pengguna yang telah mengunduh dan menggunakan aplikasi IKD. Scrapping dilakukan secara otomatis untuk mengekstrak teks komentar, rating, tanggal, dan metadata terkait lainnya. Data yang dikumpulkan ini mencakup berbagai tanggapan pengguna yang mencerminkan pengalaman dan opini mereka mengenai aplikasi IKD, yang kemudian digunakan dalam analisis sentimen.

III. Metode

A. Definisi Model

CatBoost (Categorical Boosting) adalah algoritma machine learning berbasis gradient boosting yang dikembangkan oleh Yandex. Algoritma ini dirancang untuk menangani data kategorikal secara efektif, yang merupakan salah satu keunggulan utamanya.

B. Cara Kerja

CatBoost bekerja dengan memanfaatkan kekuatan gradient boosting untuk membangun model prediktif secara bertahap. Proses dimulai dengan inisialisasi model menggunakan parameter seperti jumlah iterasi, tingkat pembelajaran, dan kedalaman pohon. Data dikemas dalam objek Pool untuk memungkinkan penanganan fitur numerik dan kategorikal secara efisien. Dalam setiap iterasi, model baru ditambahkan untuk memperbaiki kesalahan model sebelumnya. Salah satu keunggulan CatBoost adalah cara uniknya dalam menangani fitur kategorikal menggunakan teknik target-based encoding, menggantikan nilai kategorikal dengan statistik target yang dihitung dari data pelatihan. Selain itu, CatBoost menghitung berbagai statistik internal untuk mengoptimalkan pemrosesan data kategorikal dan mengurangi risiko overfitting. Proses ini melibatkan pemisahan data pelatihan dalam skema tertentu untuk memastikan statistik yang andal, sehingga setiap iterasi model dapat lebih akurat dalam memprediksi hasil.

C. Kelebihan

CatBoost menonjol dengan penanganan data kategorikal yang efisien dan efektif tanpa perlu encoding kategorikal sebelumnya, mengurangi kompleksitas pra-pemrosesan data dan mempercepat waktu pelatihan model. Keunggulan ini dikombinasikan dengan kemampuannya untuk mengurangi risiko overfitting melalui mekanisme internal yang cermat, serta performa tinggi dalam berbagai jenis tugas prediktif seperti klasifikasi dan regresi. Dengan dukungan untuk hyperparameter tuning yang fleksibel, skalabilitas, dan dukungan untuk berbagai platform, CatBoost menjadi pilihan yang menarik dalam pengembangan model prediktif. Ditambah dengan dokumentasi yang kuat dan dukungan komunitas yang solid, CatBoost menjadi alat yang andal untuk penyelesaian masalah prediktif yang kompleks.

D. Pembersihan Teks

Data yang digunakan pada projek ini akan melewati serangkaian proses pembersihan teks yang dirangkum pada Figure 1. Proses pembersihan teks tersebut meliputi penghapusan simbol, angka, tanda baca, dan kata henti. Misalnya, kalimat ”Saya mempunyai 2 buah apel, dan saya sangat menyukainya!” akan diubah menjadi ”saya mempunyai buah apel saya sangat menyukainya”.

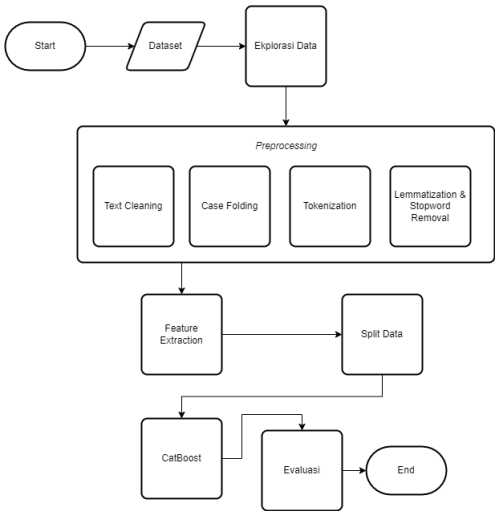

Fig. 1. Rangkaian Proses Pembersihan Teks

E. Stemming

Setelah teks melewati serangkaian proses pembersihan dan teks sudah dalam kondisi bersih, maka selanjutnya teknik stemming akan diterapkan pada teks. Stemming adalah proses mengubah kata menjadi bentuk dasarnya. Contohnya, kata ”berlari”, ”berlarian”, dan ”berlarilah” akan diubah menjadi ”lari”.

IV. Hasil dan Analisis

A. Eksplorasi Data

Berdasarkan Figure 2 dapat diketahui bahwa akhir-akhir ini aplikasi IKD didominasi oleh sentiment negatif yang mengartikan bahwa aplikasi ini tidak berjalan dengan baik dan ada sesuatu yang malfungsi.

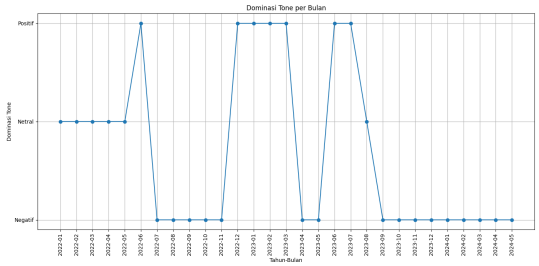

Fig. 2. Perkembangan Dominasi Tone Berdasarkan Waktu

Data pada penelitian ini akan dibagi menjadi dua kelas berdasarkan skor rating playstore dengan skor diatas 3 akan diklasifikasikan sebagai positif, skor 3 akan diklasifikasikan sebagai netral, dan skor dibawah 3 akan diklasifikasikan sebagai negatif. Lalu, Berdasarkan Figure 2 dapat diketahui bahwa review dengan tone negatif lebih banyak daripada tone positif.

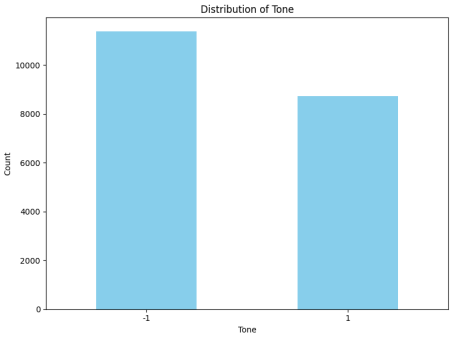

Fig. 3. Distribusi Tone Positif dan Negatif

Visualisasi kata terbanyak dengan n-gram sama dengan 2 untuk tone negatif diperlukan untuk mengetahui dua kata apa saja yang paling banyak disebutkan. Didapatkan hasilnya ialah kata-kata seperti ”kesalahan koneksi”, ”scan barcode”, ”kantor dukcapil”, ”tidak bisa”, merupakan salah satu dari sekian kata yang banyak disebut. Hal ini mengartikan bahwa bagian jaringan, fitur scan barcode perlu diperbaiki. Selain itu hal administratif lain juga perlu dibenahi agar pengguna tidak perlu data ke kantor dukcapil untuk mengurusnya.

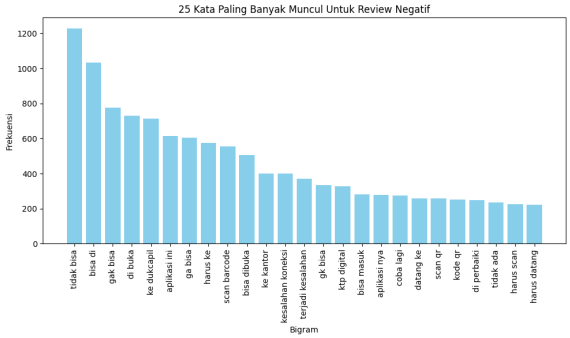

Fig. 4. Rangkaian Proses Pembersihan Teks

B. Performa Model

Setelah data yang dikumpulkan selesai dibersihkan dan dilakukan proses stemming, maka model machine learning menggunakan CatBoost Classifier akan dilatih terhadap data tersebut. Tabel I merangkum performa dari model yang telah dilatih dengan menggunakan ketiga metrik evaluasi, yaitu akurasi, F1-Score, dan juga ROC-AUC.

TABLE I
Model Performances

F1-Score	Accuracy	ROC-AUC
0.87	0.90	0.89

Akurasi model mencapai 90%, yang berarti model CatBoost Anda mampu mengklasifikasikan 90% dari semua data uji dengan benar. Ini menunjukkan bahwa model bekerja dengan baik secara umum dalam mengidentifikasi sentimen positif dan negatif. F1-Score sebesar 87% menunjukkan bahwa model memiliki keseimbangan yang baik antara presisi dan recall. Artinya, model ini tidak hanya akurat dalam prediksinya tetapi juga konsisten dalam mendeteksi sentimen positif dan negatif secara proporsional. Nilai ROC-AUC sebesar 89% mengindikasikan bahwa model memiliki kemampuan yang sangat baik dalam membedakan antara sentimen positif dan negatif. Semakin dekat nilai ini ke 1, semakin baik performa model dalam klasifikasi.

Hal ini menunjukan bahwa model CatBoost menunjukkan performa yang sangat baik dalam analisis sentimen untuk aplikasi IKD dengan nilai akurasi, F1-Score, dan ROC-AUC yang tinggi. Model ini dapat diandalkan untuk mengidentifikasi sentimen positif dan negatif dari data dengan tingkat keakuratan dan keseimbangan yang baik.

V. Kesimpulan

Indonesia sedang melakukan langkah signifikan dalam digitalisasi KTP dengan meluncurkan aplikasi Identitas Kependudukan Digital (IKD). Aplikasi ini bertujuan menggantikan KTP fisik dengan e-KTP yang dapat diakses melalui perangkat digital, sejalan dengan upaya pemerintah dalam meningkatkan efisiensi dan keterjangkauan pelayanan kependudukan. Analisis sentimen terhadap penerimaan masyarakat terhadap aplikasi ini penting untuk memahami bagaimana aplikasi ini

diterima dan digunakan oleh masyarakat. Dengan menggunakan teknik machine learning, khususnya algoritma CatBoost, proyek ini bertujuan mengklasifikasikan opini, emosi, dan sentimen dari data teks yang dikumpulkan dari berbagai sumber.

Hasil evaluasi menunjukkan bahwa model CatBoost memiliki performa yang sangat baik dengan akurasi 90%, F1-Score 87%, dan ROC-AUC 89%. Ini menunjukkan bahwa model mampu mengklasifikasikan data dengan tingkat keakuratan dan keseimbangan yang baik antara presisi dan recall, serta memiliki kemampuan yang sangat baik dalam membedakan sentimen positif dan negatif. Kesimpulannya, model CatBoost terbukti andal untuk analisis sentimen aplikasi IKD, memberikan pemahaman yang lebih dalam tentang tanggapan pengguna. Dengan pemantauan berkala dan evaluasi kinerja model yang terus-menerus, hasil analisis sentimen ini dapat membantu pemerintah dalam meningkatkan aplikasi secara proaktif dan menyesuaikan dengan kebutuhan serta harapan masyarakat, mendukung proses digitalisasi yang lebih efektif dan efisien di Indonesia.

REFERENCES

[1] CNBC Indonesia. (Dec. 14, 2023). ”Ada Identitas Kependudukan Digital, Nasib e-KTP Gimana?.” [Online]. Available: https://www.cnbcindonesia.com. [Accessed: March 9, 2024].

[2] IBM. ”Apa itu Analisis Sentimen.” [Online]. Available: https://www.ibm.com/id-id/topics/sentiment-analysis. [Accessed: March 9, 2024].

[3] Srinivasan, S. M., Shah, P., & Surendra, S. S. (2021). An approach to enhance business intelligence and operations by sentimental analysis. Journal of System and Management Sciences, 11(3), 27-40. [Accessed: March 9, 2024]