1. PENDAHULUAN

Kereta api telah menjadi pilihan transportasi yang cukup diminati di Indonesia. PT Kereta Api Indonesia (Persero) didirikan pada tahun 1945, merupakan perusahaan milik negara yang mengawasi, memasok, dan mengendalikan sistem transportasi kereta api di negara ini [1]. Data Badan Pusat Statistik (BPS) mencatat total penumpang mencapai 37,9 juta pada bulan Juli 2024, mencakup Pulau Jawa dan Sumatra, termasuk layanan kereta bandara.

Gambar 1. Data Penumpang Kereta Api

Kebutuhan akan transportasi serta kemajuan teknologi mendorong perkembangan transportasi secara online. Kemajuan teknologi informasi juga dimanfaatkan oleh PT KAI untuk memberikan kemudahan kepada masyarakat, salah satunya dengan meluncurkan aplikasi *"Access by KAI"* untuk mempermudah masyarakat melakukan pemesanan tiket secara *online* [2],[3]. PT KAI meluncurkan aplikasi *"Access by KAI"* pada 4 september 2014 yang saat itu memiliki nama *"KAI Access"*, aplikasi ini menawarkan fitur pemesanan tiket, pengecekan jadwal dan menyediakan informasi terkait perjalanan [4].

Namun, terlepas dari berbagai kemudahan layanan aplikasi *"Access by KAI"*, aplikasi tersebut mendapatkan respon yang beragam . Hal ini tercermin dari penilaian yang cukup rendah di *Google Play Store* [5]. Penilaian aplikasi yang rendah mencerminkan adanya masalah atau ketidakpuasan dari sisi pengguna terhadap aplikasi tersebut, yang menunjukkan bahwa taraf kegunaan yang relatif rendah diindikasikan oleh nilai rating dan respons yang rendah [6]. Penelitian mengenai analisis sentimen semakin sering dilakukan seiring dengan

berkembangnya media sosial yang diakses secara daring oleh masyarakat. Analisis sentimen adalah teknik pada pengolahan data teks yang memanfaatkan *Natural Language Processing* (NLP) serta *Machine Learning* (ML) guna secara otomatis mengekstrak data teks dan mengkategorikan emosi atau sikap penulis dengan tujuan guna memperoleh pengetahuan sentimen yang nilainya positif, negatif, atau netral [7].

Analisa sentimen atau *opinion mining* bertujuan untuk memahami tanggapan publik terhadap suatu produk atau layanan. Hasil dari analisis sentimen ini dapat dimanfaatkan oleh manajemen perusahaan dalam proses pengambilan keputusan dan merumuskan kebijakan baru demi menjaga kelangsungan bisnis [8], dalam konteks aplikasi *'Access by KAI'* analisis sentimen pada penelitian ini diharapkan dapat berguna memberikan pemahaman yang lebih mendalam terkait dengan ulasan pengguna sehingga dapat menjadi rekomendasi untuk peningkatan layanan PT KAI.

Pada penelitian ini pelabelan sentimen dilakukan dengan memanfaatkan library *VADER Lexicon*. *VADER* adalah alat analisis sentimen yang dirancang untuk pemrosesan bahasa alami, digunakan untuk menilai emosi dalam teks. Dikembangkan oleh tim di SocialCog, *VADER* merupakan bagian dari pustaka Python NLTK dan disesuaikan untuk menganalisis sentimen di media sosial [9]. *VADER* menghasilkan skor compound yang merupakan gabungan dari skor positif, negatif, dan netral. Nilai di atas nol menunjukkan sentimen positif, di bawah nol menunjukkan sentimen negatif, dan mendekati nol mengindikasikan sentimen netral. Setiap kata dihitung untuk menentukan nilainya [9]. Kinerja algoritma *machine learning* dapat memberikan *accuracy* yang baik dalam melakukan analisis sentimen. Analisis sentimen menggunakan machine learning yang memanfaatkan *lexicon* pada proses pelabelan data terbukti dapat meningkatkan *accuracy* pada model [8], [10].

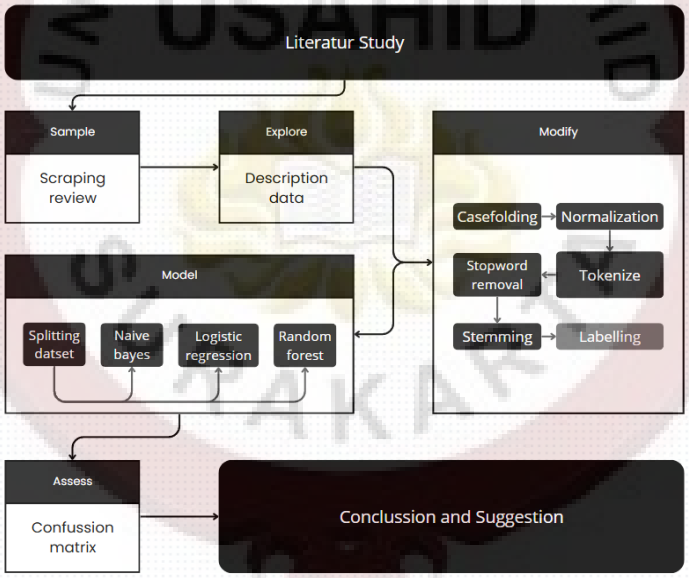
Hasil pelabelan dengan akan disosialisasikan ke dalam *word cloud. Word cloud* merupakan visualisasi istilah tekstual, di mana istilah yang paling banyak muncul ditunjukkan lebih besar. *Word cloud* membantu secara visual menggambarkan frekuensi dan signifikansi relatif dari kata-kata dalam teks, memudahkan pemahaman tema atau fokus utama dengan cepat [11]. Tujuan dilakukannya visualisasi adalah untuk mengetahui kata apa yang paling sering muncul di masing masing kategori sentimen untuk mengetahui konteks yang paling sering dibahas dalam ulasan.

Performa algoritma *machine learning* akan dievaluasi menggunakan tabel berupa *confusion matrix* yang mengemukakan klasifikasi banyaknya data uji yang benar serta data uji yang salah, juga dapat digunakan guna mengevaluasi akurasi model estimasi objek [12].

2. METODE PENELITIAN

Penelitian ini menjadikan aplikasi *"Access by KAI"* sebagai objek penelitian, akan dilaksanakan analisis sentimen pada ulasan pengguna aplikasi *"Access by KAI"* dengan tujuan memahami opini pengguna terhadap layanan yang diberikan. Metode ekstraksi data pada bentuk teks dengan analisis sentimen berguna untuk memperoleh informasi sentimen yang memiliki kategori positif, negatif, atau netral [13].

Penelitian ini akan menerapkan metode SEMMA sebagai kerangka kerja. SEMMA adalah singkatan dari *Sample, Explore, Modify, Model* dan *Assess*. Di dalam proses *data mining,* SEMMA adalah pendekatan yang diterapkan untuk mengelola dan menganalisis data secara sistematis [14]. Alur penelitian dengan metode SEMMA bisa diamati di Gambar 2.



Gambar 2. Alur Penelitian

Tahap pengumpulan sampel data pada penelitian ini dilakukan dengan metode *web scraping,* dengan parameter yang digunakan adalah *count. Count* melambangkan jumlah maksimal ulasan yang ingin didapatkan. Semakin besar nilai *count* yang dimasukkan maka akan semakin banyak ulasan yang diperoleh. Pada penelitian ini digunakan parameter nilai *count* sebesar 100.000. Ulasan hasil scraping disimpan dalam bentuk CSV *(Comma Separated Value)* untuk selanjutnya dilakukan eksplorasi dan pengolahan.

Tahap selanjutnya adalah eksplorasi data yaitu langkah yang dilakukan untuk mendeskripsikan kondisi data hasil *scraping.* Kemudian melakukan visualisasi data pada awal penelitian berupa diagram batang dari distribusi jumlah ulasan berdasarkan bulan dan distribusi

jumlah ulasan berdasarkan skor.

Tahap *modify* melibatkan pembersihan dan pemodifikasian data untuk menyeleksi data yang tidak lengkap atau nilai yang tidak valid untuk diidentifikasi dan diperbaiki. Tahap ini bertujuan untuk menghasilkan dataset yang sudah ternormalisasi, bertujuan untuk pemrosesan lebih lanjut pada tahap model. Beberapa proses yang dilakukan pada tahap modify yaitu *text-preprocessing* diawali dengan *case folding* yang merupakan tahapan mengubah semua huruf teks menjadi standar yang seragam dengan mengubah semua huruf menjadi huruf kecil dengan cara *lower casing*. Tahap berikutnya adalah *normalization* merupakan tahapan mengubah kata atau frasa yang tidak formal menjadi bentuk formal sesuai KBBI. Tahap berikutnya *tokenization* merupakan tahapan memecah teks menjadi bagian kecil perkata yang disebut token. Dilanjutkan dengan *Stopword Removal* yang merupakan tahapan menghapus istilah yang umum digunakan tetapi tidak menyampaikan informasi penting. Tahapan terakhir dalam *text-preprocessing* adalah *Stemming* tahapan mengembalikan kata ke bentuk paling dasar. Setelah dataset melalui tahapan *text-processing,* tahapan modify dilanjutkan dengan proses pelabelan data menggunakan *lexicon-based* untuk menentukan label sentimen pada setiap data sebelum memasukan tahapan model

Tahap selanjutnya adalah model, langkah pertama dalam proses pemodelan adalah memisahkan kumpulan data menjadi data pelatihan serta data uji. Data pelatihan dipakai guna membuat dan melatih model pengetahuan, sedangkan data uji dipakai untuk mengevaluasi seberapa baik model tersebut mengklasifikasikan sentimen ulasan. Pada penelitian ini digunakan 3 algoritma *machine learning* yaitu *Naive Bayes, Logistic Regression* dan *Random Forest. Naive Bayes* adalah algoritma yang dapat memproses data dengan cepat serta menerima masukan dalam format apapun, *Naive Bayes* adalah teknik probabilitas yang menggunakan teorema *Bayes* dengan asumsi ketidaktergantungan antar atribut [15],[16]. *Logistic Regression* adalah metode *supervised machine learning* yang bisa dipakai guna menganalisa data serta menggambarkan antara satu ataupun lebih variabel prediksi melalui satu variabel respon, *Logistic Regression* bekerja dengan menangani variabel respon *biner* dengan prediktor kontinu atau kategorik [17],[18], sedangkan *Random Forest* adalah metode pada analisa yang tersusun atas sejumlah pohon keputusan selaku classifier *Random Forest* bekerja melalui pemecahan data dengan acak ke dalam beberapa pohon keputusan untuk meningkatkan akurasi klasifikasi [19],[20].

Tahap *Assess* menjadi langkah terakhir, pada tahap *assess* hasil dari pengujian model

atau *testing* disebut sebagai prediksi sentimen yang dipresentasikan dalam bentuk *confusion matrix*. Performa suatu model dapat diketahui menggunakan *confusion matrix* yang menampilkan sejumlah prediksi benar serta salah, dengan *confusion matrix* dapat diketahui beberapa parameter turunan yaitu *accuracy, precision, Recall* dan *F1 Score.*