

BAB 1

PENDAHULUAN

Pada bab ini akan menjelaskan gambaran umum mengenai penelitian yang dilakukan. Dimulai dari latar belakang, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan penelitian, serta manfaat penelitian.

1.1. Latar Belakang

Surabaya sebagai ibu kota Provinsi Jawa Timur merupakan kota metropolitan yang mengalami pertumbuhan pesat dalam bidang ekonomi dan industri, khususnya sektor perdagangan dan jasa (Gunawan & Farid Ma, 2020). Peningkatan tersebut secara tidak langsung mendorong lonjakan jumlah penduduk, kendaraan bermotor, serta intensitas aktivitas masyarakat, yang berdampak pada dinamika lalu lintas kota. Kondisi ini turut memicu berbagai persoalan seperti kemacetan, kepadatan kendaraan pada ruas-ruas utama, serta meningkatnya risiko kecelakaan lalu lintas, terutama di waktu-waktu sibuk (Gunawan & Farid Ma, 2020). Pemerintah telah mengantisipasi permasalahan tersebut dengan menerapkan kebijakan seperti Undang-Undang Nomor 22 Tahun 2009 tentang Lalu Lintas dan Angkutan Jalan, yang mendorong peran aktif masyarakat dalam menciptakan lalu lintas yang aman dan tertib (Fajarudin et al., 2020). Meski begitu, pelanggaran lalu lintas dan ketidaktertiban pengguna jalan masih menjadi persoalan krusial yang memengaruhi keselamatan berkendara (MAULANA & ASTUTI, 2023).

Selain tingginya tingkat lalu lintas, jumlah kecelakaan di Surabaya juga menunjukkan pola fluktuatif yang cukup signifikan dalam lima tahun terakhir. Menurut data dari BPBD Kota Surabaya, tercatat sebanyak 1.741 kecelakaan lalu lintas pada tahun 2020, turun menjadi 1.262 kasus pada tahun 2021, namun kembali melonjak tajam menjadi 2.132 kasus pada tahun 2022 (Ahmad Riski Maulana et al., 2023). Pada tahun 2023, kepolisian mencatat adanya peningkatan jumlah kecelakaan dibanding tahun sebelumnya, meskipun tingkat keparahan cenderung menurun. Dari 1.356 kecelakaan yang terjadi hingga 13 Desember 2023, terdapat 159 korban meninggal, 20 luka berat, dan 1.177 luka ringan. Sedangkan pada tahun 2024, tercatat 1.051 kecelakaan, dan hingga pertengahan tahun 2025 sudah terjadi

396 kasus, yang mengindikasikan adanya perbaikan dalam upaya penanganan dan pencegahan kecelakaan di Surabaya (Ahmad Riski Maulana et al., 2023).

Dalam konteks tersebut, keluhan masyarakat menjadi indikator penting dalam mengungkap permasalahan publik, terutama di sektor transportasi. Aduan terkait infrastruktur, kemacetan, hingga kecelakaan lalu lintas sering kali disalurkan melalui berbagai media, termasuk media sosial. Tingginya volume aduan menunjukkan adanya kesenjangan antara pelayanan yang diberikan dengan ekspektasi public (Rahimi et al., 2019), serta menjadi parameter penting dalam mengevaluasi kualitas layanan dan efektivitas sistem transportasi (Hasan et al., 2019). Oleh karena itu, pengelolaan aduan secara cepat dan akurat menjadi kebutuhan yang semakin mendesak.

Namun, mengingat besarnya jumlah data yang masuk setiap hari, pengelolaan aduan secara manual cenderung lambat dan tidak efisien. Proses identifikasi, klasifikasi, dan analisis aduan satu per satu membutuhkan sumber daya besar dan rentan terhadap keterlambatan dalam penanganan (misalnya sistem manual memakan waktu signifikan) yang juga pernah dikonfirmasi dalam konteks pengaduan publik di sektor layanan publik (Tian et al., 2022). Hal ini berdampak pada menurunnya kepercayaan masyarakat terhadap institusi terkait. Maka dari itu, diperlukan pendekatan berbasis teknologi yang dapat mengotomatisasi proses ini secara sistematis dan efisien. Dengan strategi ini, lembaga pengelola dapat merespons aduan secara lebih cepat, tepat, dan transparan.

Di era digital saat ini, masyarakat lebih banyak menyampaikan keluhan melalui media sosial. Di Surabaya, Radio Suara Surabaya dan akun X @e100ss menjadi media interaktif yang memfasilitasi pelaporan kondisi lalu lintas dan permasalahan kota secara langsung. Menurut laporan Bagian Riset dan Survei Radio Suara Surabaya, setiap bulan tercatat sekitar 13.859 panggilan telepon yang diterima, di mana 80% dari panggilan tersebut berisi informasi terkait kondisi lalu lintas di wilayah Surabaya dan sekitarnya (Ma'aruf et al., 2021). Melalui kanal ini, partisipasi publik dalam pengawasan sosial meningkat dan menciptakan ekosistem pelaporan yang lebih responsif.

Namun, untuk memahami isi aduan yang beragam dan sering kali tidak terstruktur, diperlukan pendekatan berbasis *Natural Language Processing* (NLP). Identifikasi kata-kata seperti “mohon”, “tolong”, atau “belum” menjadi indikator kuat dalam mengenali laporan aduan (Rahmy Lidiawaty et al., 2025). Data tersebut selanjutnya digunakan sebagai dasar kebenaran (*ground truth*) untuk membangun sistem klasifikasi. Mengingat jumlah data yang besar, pendekatan otomatis melalui machine learning menjadi pilihan ideal untuk menyederhanakan proses klasifikasi.

Penelitian terdahulu menunjukkan bahwa kombinasi metode ekstraksi fitur seperti TF-IDF, *CountVectorizer*, N-Gram, atau LSA dengan algoritma klasifikasi seperti SVM, Random Forest, dan XGBoost mampu menghasilkan akurasi tinggi (Intani et al., 2022a). Di antara kombinasi tersebut, metode TF-IDF dan *Random Forest* menonjol dengan tingkat akurasi hingga 90%. Hal ini menunjukkan potensi besar dalam meningkatkan efisiensi dan akurasi pengolahan data teks dalam konteks aduan masyarakat.

Oleh karena itu, penelitian ini membangun model klasifikasi otomatis berbasis algoritma *Extreme Gradient Boosting* (XGBoost) yang dipadukan dengan tiga metode fitur ekstraksi—TF-IDF, *CountVectorizer*, dan *Word2Vec*. XGBoost dipilih karena kemampuannya dalam menangani data berdimensi tinggi dan kompleksitas relasi antar fitur (M. J. Setiawan & Nastiti, 2024). Data diperoleh melalui *scraping* akun X Suara Surabaya untuk periode Februari hingga Desember 2023 menggunakan *tools* TweetHarvester. Model dikembangkan untuk mengukur efektivitas tiap metode representasi teks dalam proses klasifikasi aduan dan bukan aduan.

Penelitian ini memanfaatkan data hasil web *scraping* dari media sosial X, khususnya dari akun Suara Surabaya (@e100ss), dengan rentang waktu pengambilan data dari Februari hingga Desember 2023. Data tersebut dianalisis menggunakan algoritma *Extreme Gradient Boosting* (XGBoost) untuk membedakan antara teks aduan dan bukan aduan. Proses klasifikasi dilakukan dengan memanfaatkan berbagai metode ekstraksi fitur, yaitu TF-IDF, *CountVectorizer*, dan *Word2Vec*. Model dengan performa terbaik berdasarkan

representasi fitur tersebut kemudian diimplementasikan ke dalam aplikasi berbasis Streamlit agar dapat digunakan secara interaktif.

Kontribusi penelitian ini tidak hanya dalam aspek teknis, tetapi juga strategis, yakni meningkatkan kecepatan dan akurasi penanganan aduan publik. Model ini dapat menjadi alat bantu bagi institusi pemerintahan dalam memprioritaskan masalah, meningkatkan efisiensi pelayanan, dan memperkuat hubungan dengan masyarakat. Pendekatan ini juga dapat dikembangkan lebih lanjut untuk kota lain atau sektor layanan publik lain yang membutuhkan respons cepat dan berbasis data.

1.2. Rumusan Masalah

Dari latar belakang diatas didapatkan permasalahan berikut:

1. Bagaimana mengidentifikasi teks aduan dan bukan aduan pada data yang bersumber dari media sosial secara otomatis menggunakan metode klasifikasi dengan model XGBoost?
2. Bagaimana mengevaluasi performa model klasifikasi XGBoost dalam membedakan antara teks aduan dan bukan aduan secara akurat?
3. Bagaimana membangun *dashboard* interaktif untuk memvisualisasikan hasil klasifikasi teks aduan dan bukan aduan secara *real-time* kepada pengguna?

1.3. Tujuan Penelitian

Berdasarkan rumusan masalah sebelumnya, tujuan penelitian ini adalah:

1. Mengembangkan model klasifikasi untuk membedakan antara teks aduan dan bukan aduan pada komentar yang berasal dari akun media sosial X Suara Surabaya (@e100ss), dengan menggunakan model *Extreme Gradient Boosting* (XGBoost) serta menerapkan metode fitur ekstraksi berupa *Term Frequency-Inverse Document Frecuency* (TF-IDF), *Count Vectorizer*, dan *Word2Vect* untuk menghasilkan representasi teks yang optimal.
2. Mengevaluasi kinerja model klasifikasi dengan menggunakan metrik evaluasi seperti *accuracy*, *recall*, *precision*, dan *f1-score*, yang divisualisasikan dalam bentuk *confusion matrix* dan *ROC Curve* untuk

menilai efektivitas model dalam mengklasifikasikan teks aduan dan bukan aduan.

3. Membangun aplikasi *dashboard* interaktif berbasis streamlit untuk memudahkan pengguna dalam melakukan klasifikasi teks serta menampilkan hasil analisis model secara visual dan *real-time*.

1.4. Batasan Penelitian

Batasan dan asumsi dalam penelitian ini adalah sebagai berikut:

1. Penelitian ini hanya menggunakan model *Extreme Gradient Boosting* (XGBoost) sebagai metode klasifikasi dalam mengidentifikasi teks aduan dan bukan aduan.
2. Metode fitur ekstraksi yang digunakan dalam penelitian ini meliputi *Term Frequency-Inverse Document Frequency* (TF-IDF), *CountVectorizer*, dan *Word2Vect*.
3. Sumber data terbatas pada unggahan dan komentar yang berasal dari akun media sosial X Suara Surabaya (@e100ss), dengan periode pengambilan data dimulai dari bulan Februari 2023 hingga Desember 2023.
4. Bahasa yang dianalisis dalam penelitian ini dibatasi hanya pada Bahasa Indonesia.
5. Jenis aduan yang dianalisis dibatasi pada aduan masyarakat yang berkaitan dengan permasalahan transportasi dan lalu lintas.

1.5. Manfaat Penelitian

Manfaat dari penelitian ini adalah:

1. Bagi Universitas Telkom, memberikan sebuah kontribusi terhadap pengembangan ilmu pengetahuan di bidang analisis sentimen dan *machine learning*. Hasil penelitian ini dapat digunakan sebagai referensi atau rujukan bagi akademis dan penelitian lain dalam melakukan penelitian serupa. Serta membuka peluang kerjasama strategis dengan lembaga pemerintah maupun organisasi non-pemerintah dalam pengelolaan akademik

2. Bagi peneliti, peneliti dapat meningkatkan keterampilan dan pengetahuan dalam bidang analisis data dan *machine learning*. Memberikan solusi nyata terhadap permasalahan sosial yang dihadapi oleh masyarakat, khususnya dalam konteks transportasi dan lalu lintas. Penelitian ini memberikan kesempatan kepada peneliti untuk menerapkan metode ilmiah dalam praktik, termasuk dalam pengumpulan, analisis, dan interpretasi data.
3. Bagi petugas operasional penanganan aduan, mendukung petugas dalam menangani aduan publik dengan lebih cepat dan efektif berkat sistem *machine learning* yang dikembangkan. Mengurangi kerja manual dengan penerapan model otomatis yang langsung mengkategorikan aduan secara tepat dan akurat. Meningkatkan kecepatan dan kualitas respons terhadap keluhan warga, sehingga layanan publik menjadi lebih optimal.

1.6. Sistematika Penulisan

Berikut merupakan sistematika penulisan pada penelitian ini yang terdiri dari tujuh bagian sebagai berikut:

1. Halaman Awal & Daftar
Memuat lembar pengesahan, pernyataan orisinalitas, abstrak (bahasa Indonesia & Inggris), kata pengantar, daftar isi, tabel, gambar, lampiran, dan istilah.
2. BAB 1 – Pendahuluan
Menguraikan latar belakang, rumusan masalah, tujuan penelitian, batasan dan asumsi, manfaat studi, serta gambaran singkat sistematika penulisan.
3. BAB 2 – Landasan Teori
Menyajikan tinjauan pustaka dan teori penunjang—mulai dari Media Sosial X, teknik *web scraping*, kerangka CRISP-DM, text mining, Streamlit, hingga *tools & library Python*—serta alasan pemilihan model XGBoost.

4. BAB 3 – Metodologi

Menjelaskan langkah-langkah CRISP-DM: *business understanding*, *data understanding*, *data preparation*, *modelling*, *evaluation*, dan *deployment*, dan rencana jadwal kegiatan.

5. BAB 4 – Hasil & Pembahasan

Membahas eksplorasi dan pra-pemrosesan data, penerapan model dengan berbagai teknik ekstraksi fitur (TF-IDF, *CountVectorizer*, *Word2Vec*, dan kombinasi), evaluasi kinerja menggunakan metrik dan *confusion matrix*, serta implementasi *deployment*.

6. BAB V – Kesimpulan & Saran

7. Daftar Pustaka & Lampiran