

BAB I

PENDAHULUAN

1.1 Latar Belakang

Twitter, sebagai platform *mikroblogging* dengan batasan 280 karakter per tweet, telah menjadi wadah utama bagi publik untuk menyalurkan pemikiran, opini, dan kegelisahan sehari-hari. Pada Januari 2023, terdapat 167 juta pengguna aktif media sosial di Indonesia, yang setara dengan 60,4% dari total populasi negara tersebut, menurut data We Are Social [1]. Sejak diluncurkan pada tahun 2006, Twitter telah tumbuh pesat di Indonesia, angkanya naik dari sekitar 20 juta pengguna pada 2015 menjadi lebih dari 170 juta pada 2024, menjadikannya salah satu negara dengan pertumbuhan tercepat di Asia Tenggara. Pertumbuhan pengguna ini didorong oleh semakin mudahnya akses internet dan *smartphone*, serta pergeseran budaya komunikasi ke arah *digital-first*, terutama di kalangan milenial dan Gen Z. Di samping itu, tren penggunaan Twitter untuk diskusi politik, kampanye sosial, dan *customer service* korporasi menunjukkan bagaimana platform ini bertransformasi dari sekadar media sosial menjadi alat pengambilan keputusan strategis.

Ekspresi emosi, baik verbal maupun nonverbal, merupakan aspek *fundamental* dari interaksi manusia. Dalam komunikasi tatap muka, nada suara dan bahasa tubuh memudahkan pendengar menangkap emosi pembicara [2]. Namun, komunikasi via teks di media sosial bersifat tidak langsung (non-vokalis), sehingga mengaburkan sinyal emosional yang seharusnya tampak jelas. Hal ini menimbulkan tantangan tersendiri: bagaimana menafsirkan emosi pengguna hanya dari rangkaian kata, emoji, atau simbol dalam sebuah tweet [3]. Analisis emosi teks memiliki potensi besar dalam berbagai bidang, mulai dari pemasaran digital (untuk memahami resonansi kampanye iklan) hingga kesehatan mental (untuk deteksi dini kecemasan atau depresi berbasis teks pengguna). Beberapa studi terkini juga menunjukkan penggunaan *emotion analytics* dalam mendukung keputusan bisnis *real-time*, seperti memantau opini publik saat peluncuran produk atau merespons krisis media dengan cepat.

Sejumlah penelitian telah mencoba menjawab tantangan ini melalui teknik *Natural Language Processing* (NLP) dan algoritma *machine learning*. Misalnya, Agustia & Suryono (2025) membandingkan *Naïve Bayes*, *Random Forest*, dan *Logistic Regression* dalam analisis sentimen tweet tentang perjudian online; hasilnya *Random Forest* unggul dengan akurasi 78% setelah penerapan SMOTE dan TF-IDF [4]. Penelitian lain oleh Fadillah dkk. (2025) memanfaatkan komentar Instagram untuk menganalisis sentimen pembatasan BBM Peralite, di mana *Random Forest* juga menunjukkan performa terbaik (akurasi 73%) dibandingkan *K-Nearest Neighbor* [5].

Meskipun demikian, sebagian besar studi masih fokus pada sentimen polaritas (positif–negatif–netral) dan belum mengkaji detail kategori emosi seperti *anger*, *fear*, *happy*, *love*, dan *sadness* yang lebih halus namun penting untuk pemahaman psikologis pengguna. Di sisi lain, variasi bahasa informal, seperti penggunaan *slang*, singkatan, *typo*, serta kehadiran *emoji* dan tanda baca yang tidak baku dapat menimbulkan kompleksitas tambahan dalam preprocessing teks. Model-model konvensional sering kali mengalami penurunan akurasi karena *noise* linguistik ini dan distribusi data yang timpang antar-kelas emosi. Data tidak seimbang dapat membuat model cenderung bias ke kelas mayoritas, sehingga performa pada kelas minoritas menurun drastis.

Beberapa pendekatan *deep learning* mulai diterapkan untuk mengatasi isu kontekstual dan nuansa bahasa ini, tetapi memerlukan sumber daya komputasi besar dan data berlabel dalam jumlah masif, yang tidak selalu tersedia di lingkungan penelitian Indonesia.

Berdasarkan latar tersebut, penelitian ini bertujuan mengembangkan *pipeline* NLP terintegrasi yang mencakup pembersihan teks lanjutan (normalisasi *slang*, koreksi *typo*, konversi *emoji*), ekstraksi fitur TF-IDF, dan penyeimbangan data menggunakan SMOTE, lalu mengaplikasikannya pada algoritma *Random Forest*. Dengan harapan, model yang dihasilkan mampu mengklasifikasikan lima kategori emosi utama dalam tweet berbahasa Indonesia secara lebih akurat dan stabil.

1.2 Rumusan Masalah

1.2.1 Permasalahan

Berdasarkan tinjauan pustaka, terdapat sejumlah permasalahan utama yang mendorong dilakukannya penelitian ini:

1. Variasi Bahasa Informal: Tweet berbahasa Indonesia sering mengandung *slang*, singkatan, *typo*, *emoji*, dan emotikon yang tidak baku, sehingga menyulitkan model NLP konvensional dalam menangkap makna dan nuansa emosi.
2. Distribusi Data Tidak Seimbang: Pada banyak dataset emosi, sebagian kelas (misalnya *happy* atau *sadness*) memiliki jumlah sampel jauh lebih sedikit dibanding kelas lainnya, yang mengakibatkan model cenderung bias terhadap kelas mayoritas.
3. Keterbatasan Kategori Emosi: Sebagian besar studi sebelumnya hanya membedakan sentimen positif, negatif, dan netral; sedikit yang mengklasifikasikan emosi secara *granular* ke dalam kategori *anger*, *fear*, *happy*, *love*, dan *sadness*.

1.2.2 Metode Usulan

Untuk mengatasi permasalahan di atas, penelitian ini mengusulkan *pipeline* klasifikasi emosi sebagai berikut:

[1] *Preprocessing* Lanjutan

- Normalisasi *slang* dan singkatan Bahasa Indonesia
- Koreksi otomatis *typo* berbasis kamus baku
- Konversi *emoji* dan emotikon menjadi *token* kata
- Penghapusan URL, *mention*, *hashtag*, dan tanda baca tidak relevan

[2] Ekstraksi Fitur TF-IDF

- Pembobotan term *frequency-inverse document frequency*.

[3] Penyeimbangan Data dengan SMOTE

- Sintesis sampel minoritas untuk menyeimbangkan distribusi kelas emosi

[4] Klasifikasi dengan *Random Forest*

- Pelatihan model menggunakan hutan dengan banyak pohon keputusan
- Voting mayoritas untuk menentukan prediksi akhir

1.2.3 Pertanyaan Penelitian

Berdasarkan rumusan masalah dan metode usulan di atas, penelitian ini akan menjawab pertanyaan-pertanyaan berikut:

1. Bagaimana pengaruh *preprocessing* (normalisasi *slang*, koreksi *typo*, konversi *emoji*) terhadap akurasi model *Random Forest* dalam mengklasifikasikan emosi tweet Bahasa Indonesia?
2. Sejauh mana penggunaan TF-IDF dapat meningkatkan performa model?
3. Seberapa efektif SMOTE dalam menyeimbangkan distribusi data emosi dan memperbaiki metrik *precision*, *recall*, dan *F1-score* untuk kelas minoritas?
4. Konfigurasi *hyperparameter Random Forest* (jumlah pohon, kedalaman maksimum) manakah yang memberikan kombinasi optimal antara akurasi keseluruhan dan kestabilan performa antar-kelas?

Dengan menjawab pertanyaan-pertanyaan tersebut, penelitian diharapkan mampu menghasilkan model klasifikasi emosi yang tidak hanya akurat dan stabil, tetapi juga aplikatif untuk keperluan *monitoring* opini publik secara langsung.

1.3 Tujuan dan Manfaat

1.3.1 Tujuan Penelitian

Penelitian ini memiliki tujuan sebagai berikut:

1. Mengembangkan Model Klasifikasi Emosi: Membangun *pipeline end-to-end* yang mengintegrasikan teknik NLP (normalisasi *slang*, koreksi *typo*, konversi *emoji*, *stemming*) dengan ekstraksi fitur TF-IDF dan balancing data menggunakan SMOTE.
2. Mengevaluasi Performa Model: Menilai akurasi, *precision*, *recall*, *F1-score*, dan kestabilan performa antar-kelas (*anger*, *fear*, *happy*, *love*, *sadness*) pada data uji.
3. Mengidentifikasi Dampak Setiap Tahap *Preprocessing*: Menganalisis kontribusi masing-masing langkah preprocessing (normalisasi *slang*, koreksi *typo*, konversi *emoji*) terhadap peningkatan metrik evaluasi.
4. Menentukan Konfigurasi *Hyperparameter* Optimal: Mencari nilai terbaik untuk parameter *Random Forest* (*n_estimators*, *max_depth*, *max_features*) yang

memberikan *trade-off* optimal antara akurasi keseluruhan dan konsistensi performa kelas minoritas.

1.3.2 Manfaat Penelitian

1. Memperkaya Literatur NLP Bahasa Indonesia: Menambah referensi penelitian tentang pengolahan bahasa informal (*slang*, *typo*, *emoji*) dalam klasifikasi emosi.
2. Panduan Metodologis: Menyajikan framework *best-practice* bagi peneliti selanjutnya dalam memilih dan mengombinasikan teknik *preprocessing*, ekstraksi fitur, dan penyeimbangan data.
3. *Customer Experience Management*: Menunjang tim *customer service* dalam merespon keluhan atau pujian dengan cepat berdasarkan emosi yang terdeteksi.
4. Deteksi Dini Isu Sosial & Kesehatan Mental: Potensi penerapan untuk memantau gejala kecemasan atau depresi berbasis pola bahasa, sebagai alat pendukung intervensi kesehatan masyarakat.

Dengan tujuan dan manfaat tersebut, diharapkan penelitian ini tidak hanya memberikan kontribusi teoretis, tetapi juga solusi aplikatif yang dapat langsung diadopsi dalam berbagai konteks pemantauan opini dan emosi di media sosial.

1.4 Batasan Masalah

Untuk menjaga fokus dan kesesuaian lingkup penelitian, maka ditetapkan batasan-batasan sebagai berikut:

1. *Preprocessing* Teks: Proses pembersihan meliputi normalisasi *slang*, koreksi *typo* berbasis kamus baku, konversi *emoji* ke *token*, penghapusan URL/*mention*/*hashtag*, pengurangan karakter berulang, *stopword removal*, dan *stemming*.
2. Ekstraksi Fitur: Ekstraksi fitur terbatas pada representasi TF-IDF. Teknik embedding (Word2Vec, BERT) atau feature lain (sentiment lexicon) tidak digunakan.
3. Penyeimbangan Data: Penyeimbangan kelas dilakukan hanya dengan *manual oversampling* dan SMOTE. Metode *undersampling* lain tidak diuji.
4. Model dan Evaluasi: Klasifikasi menggunakan algoritma *Random Forest* dengan eksperimen *hyperparameter* (*n_estimators*, *max_depth*, *max_features*).

Model lain (SVM, *Neural Network*, CNN) tidak dibandingkan secara langsung. Evaluasi dibatasi pada metrik akurasi, *precision*, *recall*, dan *F1-score* melalui pembagian data *train-test* (80:20).

Dengan batasan-batasan tersebut, penelitian ini difokuskan pada pengembangan dan evaluasi model *Random Forest* untuk klasifikasi emosi tweet dalam skala eksperimen.

1.5 Sistematika Penulisan

Sistematika penulisan merupakan sebuah metode atau urutan dalam menyelesaikan sebuah penelitian. Penulisan skripsi ini dikelompokkan menjadi beberapa sub bab dengan sistematika penulisan sebagai berikut :

BAB I PENDAHULUAN

Bab ini membahas latar belakang penelitian tentang klasifikasi emosi dalam tweet berbahasa Indonesia, mulai dari peran media sosial dan tantangan NLP pada teks informal, hingga urgensi membangun model otomatis untuk deteksi emosi publik. Di dalamnya juga dijelaskan rumusan masalah (permasalahan, metode usulan, dan pertanyaan penelitian), tujuan dan manfaat penelitian, serta batasan-batasan ruang lingkup yang ditetapkan.

BAB II KAJIAN PUSTAKA

Bab ini menguraikan teori dan konsep dasar yang mendasari penelitian, meliputi: pengertian emosi dan analisis sentimen di media sosial; teknik-teknik preprocessing teks Bahasa Indonesia (normalisasi *slang*, koreksi *typo*, konversi *emoji*, *stemming*, *stopword removal*); representasi fitur menggunakan TF-IDF; metode penyeimbangan data SMOTE; serta algoritma *Random Forest*.

BAB III METODE PENELITIAN

Pada Bab III ini dijelaskan rangkaian metodologi penelitian yang dimulai dengan pengumpulan *dataset*. Data kemudian menjalani *preprocessing* teks, termasuk normalisasi *slang*, koreksi *typo*, konversi *emoji*, penghapusan *noise* (URL, *mention*, *hashtag*), tokenisasi, *stopword removal*, dan *stemming*, sebelum diekstraksi fiturnya menggunakan TF-IDF. Untuk mengatasi ketidakseimbangan kelas, SMOTE diterapkan pada data pelatihan, lalu model *Random Forest* dilatih dengan eksperimen *hyperparameter* (*n_estimators*, *max_depth*, *max_features*) dan

dievaluasi menggunakan metrik akurasi, *precision*, *recall*, *F1-score*, serta *confusion matrix* pada data uji.

BAB IV HASIL DAN PEMBAHASAN

Pada Bab IV ini disajikan hasil eksperimen dan analisis performa model klasifikasi emosi, dimulai dengan ringkasan statistik dataset setelah *preprocessing* (dampak normalisasi *slang* dan konversi *emoji*), kemudian perbandingan metrik evaluasi (akurasi, *precision*, *recall*, *F1-score*) untuk tiap kombinasi TF-IDF, serta sebelum dan sesudah penerapan SMOTE. Selanjutnya dipaparkan efek variasi *hyperparameter Random Forest* (*n_estimators*, *max_depth*, *max_features*) terhadap performa keseluruhan dan kestabilan antar-kelas, dilengkapi visualisasi *confusion matrix* untuk setiap konfigurasi utama.

BAB V PENUTUP

Bab ini memuat kesimpulan dari penelitian yang disusun berdasarkan hasil analisis dan pertanyaan penelitian yang telah dirumuskan. Bab ini juga menyajikan saran untuk pengembangan atau perluasan penelitian selanjutnya.

DAFTAR PUSTAKA

Berisi seluruh referensi ilmiah, jurnal, artikel, dan dokumen yang dijadikan acuan dalam menyusun penelitian ini