BAB I
PENDAHULUAN

1.1 Latar Belakang

Jutaan pengguna Twitter menggunakan *platform* media sosial ini untuk berkomunikasi dan berinteraksi dengan orang-orang di seluruh dunia dengan berbagai tujuan, seperti berbagi informasi dan berkomunikasi. Selain itu, *platform* memungkinkan pengguna menulis, membaca, dan berbagi teks pendek yang disebut *tweet*, yang memiliki panjang maksimal 280 karakter. (Meftah *et al.*, 2018) (Kumar & Gruzd, 2019).

Pengguna Twitter sering menggunakan kata-kata yang tidak baku, yang dapat membuat pesan lebih sulit dipahami dan menyebabkan komunikasi yang kurang efektif. Kata baku didefinisikan sebagai kata yang sesuai dengan norma tata bahasa standar, sementara kata tidak baku didefinisikan sebagai kata yang biasa digunakan dalam percakapan sehari-hari atau tidak sesuai dengan norma tata bahasa (EYD) (Ruhamah *et al.*, 2018).

Normalisasi kata membantu pemrosesan dan analisis *tweet* karena membantu mengatasi masalah seperti singkatan, bahasa gaul, kesalahan ejaan, dan penggunaan bahasa yang tidak pantas. Ini mengubah kata tidak baku menjadi kata baku dalam kalimat atau *tweet* (Ivan & Adikara, 2019).

Untuk mengurangi kata-kata yang tidak relevan dan meningkatkan akurasi klasifikasi, *preprocessing*, yang mencakup *tokenizing*, *cleansing*, *case folding*, *spelling correction*, *filtering*, dan *stemming*, sangat penting. Pengklasifikasi *Naïve Bayes* digunakan untuk klasifikasi, dan metode *Levenshtein Distance* digunakan untuk mengubah kata tidak baku menjadi kata baku. Hasil pengujian menunjukkan tingkat akurasi tertinggi, dengan perbaikan kata tidak baku sebesar 98,33%, dengan *precision*, *recall*, dan *f1-measure* masing-masing sebesar 96.77%, 100%, dan 98.36% (Antinasari *et al.*, 2017).

Beberapa tahap *preprocessing*, seperti *case folding*, *cleansing*, *tokenizing*, *stemming*, dan *stopword*, termasuk dalam penelitian sebelumnya yang menggunakan metode *Levenshtein Distance* untuk mengatasi salah ejaan dan

metode *Naïve Bayes Classifier* untuk mengevaluasi akurasi. Penelitian tersebut menggunakan 450 data pelatihan, dengan 150 kategori positif, 150 negatif, dan 150 netral. Hasil pengujian pada 100 data uji menunjukkan tingkat akurasi menggunakan *Levenshtein Distance* sebesar 67,05% dan tingkat akurasi tanpa *Levenshtein Distance* sebesar 63,83% dalam proses klasifikasi menggunakan *Naive Bayes* (Rozi *et al.*, 2019).

Studi ini berbeda dari penelitian sebelumnya karena berfokus pada *preprocessing* menggunakan kamus opini berbasis *lexicon*, menggunakan KBBI sebagai sumber referensi untuk menemukan kata-kata baku dan tidak baku, dan menggunakan metode *InSet* untuk menghitung nilai sentimen *Confusion Matrix*.

Stopword Removal, yang bertujuan untuk menghilangkan kata-kata yang tidak relevan dari daftar *stopword*, adalah salah satu langkah *preprocessing* teks yang dibahas dalam penelitian ini. Analisis sentimen dilakukan dengan teknik *Naive Bayes*. Dalam penelitian ini, algoritme Random Sampling Berdasarkan Istilah digunakan untuk membuat daftar *stopword* dengan parameter X, Y, dan L. Hasil evaluasi menunjukkan bahwa menggunakan algoritme ini dengan parameter tertentu menghasilkan daftar *stopword* yang paling akurat. Pengujian menunjukkan bahwa menggunakan algoritme Random Sampling Berdasarkan Istilah dalam penghapusan *stopword* atau tanpa proses penghapusan *stopword* mampu mencapai tingkat akurasi yang lebih tinggi (Rinandyaswara *et al.*, 2022).

Dengan tujuan mengembangkan sistem otomatis yang akurat, penelitian ini berfokus pada *preprocessing* dan deteksi kata-kata baku dan tidak baku dalam data Twitter dengan menggunakan Kamus Besar Bahasa Indonesia (KBBI). Deteksi otomatis kata-kata baku dan tidak baku dalam data Twitter berbasis KBBI sangat penting karena memungkinkan peneliti untuk dengan mudah menemukan kata-kata tidak baku atau *slang* yang relevan untuk penelitian mereka.

1.2 Rumusan Masalah

Dari latar belakang yang dipaparkan di atas, dapat di tarik rumusan masalah berupa:

1. Bagaimana melakukan otomatisasi pendeteksi kata baku dan tidak baku pada data Twitter yang diperoleh?
2. Bagaimana performa dari penerapan *InSet Lexicon* dalam melakukan analisis sentimen menggunakan data dari Twitter?

1.3 Batasan Masalah

Dalam latar belakang dan rumusan masalah di atas terdapat Batasan masalah:

1. Penelitian ini menggunakan data dari (Twitter sebelum menjadi X) dan (Twitter sesudah menjadi X) dengan topik pemindahan IKN (Ibu Kota Negara).
2. Penelitian ini menggunakan data *tweet* dari Twitter dengan rentang waktu dari 1 Maret 2022 hingga 17 Maret 2022 dan X dengan rentan waktu 29 Maret 2024 hingga 30 April 2024.
3. Kamus KBBI yang digunakan dalam penelitian ini diperoleh dari Github yang berjumlah 28.526 kata Bahasa Indonesia yang baku.
4. Penghapusan kata yang berimbuhan terdiri dari beberapa kata berikut ini Kata *Prefiks*, Kata *Sufiks,* Kata *infiks,* dan Kata *Konfiks.*
5. Proses normalisasi dari kata tidak baku menjadi baku dilakukan secara manual.
6. Penelitian ini menggunakan metode *lexicon-based* dengan menerapkan InSet (*Indonesian Sentimen*) sebagai penentuan polaritas.
7. Pengujian performa dari hasil penentuan polaritas sentimen menggunakan *Confusion Matrix* dengan menghitung *Accuracy. Precision, Recall,* dan *F1-score*

1.4 Tujuan Penelitian

Tujuan utama dari penelitian ini, seperti yang dijabarkan dalam latar belakang dan rumusan masalah adalah:

1. Membangun sistem otomatisasi untuk mendeteksi kata baku dan tidak baku pada data Twitter berbasis KBBI dengan tujuan meningkatkan keakuratan dan konsistensi dalam analisis teks di lingkungan media sosial.
2. Menilai dampak pendeteksian kata tidak baku terhadap peningkatan akurasi sentimen analisis pada data Twitter, dengan fokus pada interpretasi yang lebih tepat terhadap opini dan perasaan yang diungkapkan oleh pengguna.

3. Menggunakan metode *InSet* sebagai alat otomatis untuk melakukan sentimen analisis, dengan tujuan meningkatkan efisiensi dalam pemrosesan besar data Twitter dan memberikan kontribusi pada perkembangan teknologi kecerdasan buatan.
4. Menganalisis dan mendokumentasikan bagaimana pendeteksian kata tidak baku dapat membantu peneliti dalam mengidentifikasi dan menemukan kata-kata yang tidak baku di *platform* Twitter, sehingga memberikan pemahaman lebih dalam tentang variasi Bahasa.

1.5 Manfaat Penelitian

Terdapat beberapa manfaat yang bisa didapat dari penelitian ini, baik untuk penulis, pembaca, maupun pihak lain yang terlibat yaitu:

1. Manfaat untuk Penulis:
   - Pengembangan Keterampilan: Penulis dapat mengembangkan keterampilan dalam bidang otomatisasi pendeteksian kata baku dan tidak baku, meningkatkan kompetensinya dalam pemrosesan bahasa alami.
   - Kontribusi Ilmiah: Penelitian ini dapat memberikan kontribusi signifikan terhadap literatur ilmiah dengan menyajikan solusi otomatis untuk pendeteksian kata baku dan tidak baku di lingkungan Twitter berbasis KBBI.
2. Manfaat untuk Pembaca:
   - Peningkatan Kualitas Analisis Sentimen: Pembaca dapat memanfaatkan hasil penelitian ini untuk meningkatkan kualitas analisis sentimen di platform Twitter dengan memperhitungkan aspek penggunaan kata tidak baku, memberikan wawasan yang lebih akurat terkait opini dan sentimen pengguna.
3. Manfaat untuk Peneliti:
   - Pemahaman Mendalam tentang Metode *InSet*: Peneliti dapat memperoleh pemahaman mendalam tentang penggunaan metode *InSet* untuk sentimen analisis otomatis, yang dapat menjadi landasan untuk penelitian lebih lanjut dalam bidang analisis bahasa dan kecerdasan buatan.

- Pengembangan Metodologi Baru: Penelitian ini dapat membuka peluang untuk pengembangan metodologi baru dalam analisis sentimen yang memasukkan deteksi kata tidak baku sebagai elemen kritis.

4. Manfaat untuk Masyarakat:
    - Kesadaran Bahasa yang Lebih Baik: Hasil penelitian dapat meningkatkan kesadaran masyarakat tentang pentingnya penggunaan bahasa yang benar dan sesuai dengan KBBI di media sosial, membantu menciptakan lingkungan komunikasi yang lebih baik.
    - Pencegahan Konten Negatif: Deteksi kata tidak baku dapat membantu dalam pencegahan dan penanganan konten negatif atau tidak etis di platform Twitter, menciptakan lingkungan online yang lebih positif dan etis.

1.6 Sistematika Penelitian

Sistematika penulisan tugas akhir ini terdiri dari beberapa tahapan yang dibagi mejadi beberapa BAB yaitu sebagai berikut

BAB I PENDAHULUAN

Pada bagian ini peneliti menjabarkan bagian latar belakang, rumusan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penelitian.

BAB II KAJIAN LITERATUR

Pada bagian ini penulis menjabarkan penelitian terkait judul skripsi dan teori yang digunakan dalam topik penelitian dengan cara yang sama.

BAB III METODOLOGI PENELITIAN

Pada bagian ini penulis menjabarkan alur penelitian yang dilakukan selain itu juga dibahas waktu dan tempat penelitian, alat penelitian, pengummpulan data, penentuan polaritas, dan analisis dan perancangan.

BAB IV HASIL DAN PEMBAHASAN

Pada bagian ini penulis memaparkan hasil pengujian dan pembahasan dari judul penelitian yang dilakukan bagian ini menjelaskan secara lengkap komposisi data, tahap *preprocessing* dan *Confusion Matrix.*

BAB V PENUTUP

Pada bagian ini penulis menyampaikan kesimpulan dan saran dari hasil penelitian yang diperoleh, bagian ini akah berisi tentang hal-hal yang terjadi dalam penelitian ini dan juga menampilkan hasil pengujian.