BAB I
PENDAHULUAN

1.1 Latar Belakang

Analisis sentimen merupakan salah satu fokus dalam pengolahan bahasa alami (NLP) yang bertujuan untuk mengenali perasaan atau sentimen berdasarkan opini yang disampaikan melalui bentuk teks [1]. Opini berbentuk teks dapat dianalisis sehingga menghasilkan pengetahuan. Opini tersebut dapat dikategorikan menjadi dua kelas kategori yaitu positif dan negatif dalam analisis sentimen. Selain dua kelas klasifikasi tersebut, beberapa penelitian lain telah menggunakan tiga kelas klasifikasi yaitu klasifikasi positif, netral, dan negatif [2], [3].

Data opini tersedia pada berbagai sumber seperti media sosial, platform berita, platform penyedia data, dan lain – lain. Data yang tersedia pada media online memiliki sifat yang tidak terstruktur baik dari segi tata bahasa maupun segi penulisan. Umumnya opini pada media online mengandung kata *slang,* yaitu bahasa gaul atau prokem. *Slangword* adalah kata atau frasa yang berbentuk informal dan sering dipakai dalam komunikasi, terutama oleh kelompok sosial atau usia tertentu, dan mengalami perubahan yang terus-menerus. I Gede Budiasa, dkk mendefinisikan kata *slang* sebagai variasi bahasa dari suatu kelompok yang digunakan saat pergaulan dengan penuh kreatifitas menghasilkan ujaran baru dan mengkombinasikan berbagai variasi bahasa dalam komunikasi sehari – hari [4]. Kata *slang* sering dijumpai oleh kalangan remaja dan merupakan salah faktor yang menyebarluaskan pada media sosial.

Kata *slang* yang ada dalam teks akan menjadi tantangan dalam klasifikasi sentimen. Kata-kata *slang* memiliki makna yang berbeda dari arti kata baku atau formal. Hal ini dapat menyebabkan kesalahpahaman jika model tidak mengenali makna *slang* yang sesungguhnya. Sehingga kata *slang* harus dikonversi menjadi kata baku sesuai dengan KBBI. Tahap pengubahan kata *slang* menjadi kata baku disebut dengan tahap normalisasi.

Selain kata *slang*, teks dalam dataset seringkali mengandung *noise* berupa kesalahan ejaan, penggunaan tanda baca yang tidak konsisten, serta berbagai anomali lain yang bisa berdampak pada akurasi model analisis sentimen. *Noise* dalam teks adalah elemen yang tidak relevan atau tidak diinginkan yang bisa mengganggu pemahaman dan interpretasi data.

Dalam analisis sentimen, tahapan *preprocessing* teks yang umum digunakan meliputi *case folding, tokenizing, penghapusan stop words*, dan *stemming* [5]. Namun, proses standar ini sering kali tidak mencakup koreksi ejaan, yang bisa menyebabkan kesalahan pengejaan mempengaruhi pembobotan kata dalam analisis. Selain itu, keberadaan singkatan dan penggunaan kata *slang* juga berpotensi menurunkan akurasi dari hasil analisis sentimen.

Dalam satu kalimat opini setidaknya terdapat minimal satu kata *slang* yang ada di dalamnya. Dengan mengabaikan atau menghapus kata-kata *slang* dapat mengakibatkan interpretasi yang salah terhadap sentimen sebenarnya dari kalimat tersebut. Salah satu contoh kalimat *review* dari aplikasi Lazada yaitu "Kapasitasnya cetek banget, aturannya kalo sepi lumayan, nggak ada tuh virus. Tapi kalo duh gile, ampun, makin penuh aja, bikin keki.". Dalam satu kalimat tersebut terdapat lebih dari satu kata *slang* yaitu kata "cetek", "duh gile", "keki". Jika kata- kata *slang* dihapus, kemungkinan besar akan mempengaruhi sentimen yang ingin diungkapkan dalam kalimat tersebut. Alasan mengapa kata *slang* tidak boleh dihapus sepenuhnya adalah karena kata *slang* memberikan nuansa dan emosi yang penting dalam bahasa sehari-hari. Apabila kata *slang* tetap ada dalam teks, pada tahap fitur atau vektorisasi, kata-kata tersebut berpotensi menjadi elemen baruatau fitur baru. Sehingga, pentingnya normalisasi kata-kata *slang* menjadi bentuk standar perlu dilakukan untuk memastikan akurasi analisis teks. Dengan melakukan normalisasi text, representasi numerik dalam vektorisasi tetap konsisten dan memberikan pemahaman yang lebih akurat terhadap emosi dan sentimen dalam teks.

Menurut penelitian [5] penerapan kamus *slang word* dalam analisis sentimen ini perlu dilakukan. Data yang digunakan dalam penelitian tersebut 88 % kalimat mengandung kata *slang.* Dalam penelitian tersebut penerapan kamus

slang word dapat meningkatkan akurasi model sentimen dibandingkan dengan tanpa penerapan kamus *slang word*. Penelitian terdahulu yang membahas normalisasi kata *slang* juga dilakukan oleh Riri & Ade [3] dengan mengangkat studi kasus *tweet* produk *gadget*. Penelitian tersebut memiliki tujuan untuk membandingkan hasil akurasi setelah dilakukan tahapan normalisasi dan tanpa normalisasi. Dengan menerapkan model *word2vec*, TF-IDF untuk ekstraksi fitur dan algoritma *Naive Bayes* untuk klasifikasi, hasil yang diperoleh menunjukkan peningkatan akurasi tertinggi hingga 91% setelah normalisasi, berbanding dengan 88% tanpa normalisasi.

Berdasarkan penjelasan sebelumnya, penerapan normalisasi teks memiliki pengaruh yang baik untuk akurasi model klasifikasi sehingga penelitian ini akan membuat sebuah korpus yang berisi kumpulan kata *slang* dengan mengkombinasikan kata *slang* yang terdapat pada data dan kata *slang* yang sudah ada di *internet*. Perlunya pembuatan kamus *slang* pada penelitian ini karena bahasa *slang* selalu berubah dan berkembang seiring waktu. Dengan membuat korpus *slang word* memungkinkan untuk tetap mengikuti perkembangan kata *slang* terbaru. Kospus *slang* yang sudah dibuat akan digunakan untuk tahapan normalisasi dan membandingkan hasil akurasi beberapa model klasifikasi dengan skenario penerapan normalisasi dan tanpa normalisasi. dengan pelabelan data yang akan digunakan yaitu pelabelab *IndoBERT*. Pelabelan menggunakan model *IndoBERT* dipilih berdasarkan penelitian [6] yang menunjukkan bahwa *IndoBERT* efektif dalam menangani variasi bahasa dan nuansa kontekstual bahasa Indonesia. Selain itu, *IndoBERT* mampu menangani subjektivitas dan konteks spesifik dari teks berbahasa Indonesia, menjadikannya pilihan untuk pelabelan menggunakan *IndoBERT*.

Algoritma klasifikasi yang dipakai merupakan algoritma machine learning yaitu algoritma *Naïve Bayes Classifier* dan Algoritma *Decision Tree Classifier*. Algoritma *Naive Bayes* dipilih karena kesederhanaannya, kecepatan komputasi yang tinggi, dan efektivitasnya dalam menangani data teks yang besar dan bervariasi. Penelitian menunjukkan bahwa *Naive Bayes* memberikan akurasi yang tinggi dalam berbagai tugas klasifikasi teks di Indonesia [7]. *Decision Tree* juga dipertimbangkan karena kemampuannya untuk menangani data yang *non-*

linear dan memberikan interpretabilitas yang baik, yang membuatnya cocok untuk analisis data yang kompleks. Studi terbaru mengonfirmasi bahwa *Decision Tree* efektif dalam klasifikasi teks dengan hasil akurasi yang kompetitif dibandingkan dengan algoritma lain [8].

1.2 Rumusan Masalah

Kata *slang* dalam teks opini di media online merupakan sebuah *noise*, yang apabila kata *slang* dihapus akan menghilangkan sentimen, dan apabila dipertahankan dapat mengurasi akurasi. Sehingga diperlukannya normalisasi kata *slang* menjadi kata baku sesuai standar KBBI.

1.3 Pertanyaan Penelitian

Pertanyaan penelitian dalam penelitian ini yaitu bagaimana pengaruh normalisasi *slang word* terhadap akurasi model klasifikasi teks bahasa Indonesia?

1.4 Batasan Masalah

Terdapat batasan masalah yang ada dalam penelitian ini yaitu:

a. Data yang akan digunakan hanya terdiri dari dua data yaitu data *review* aplikasi Lazada dan data *review* Tokopedia bersumber dari platform Kaggle.
b. Data yang digunakan merupakan data teks.
c. Algoritma yang digunakan terdiri dari Algoritma *Naïve Bayes Classifier* dan *Decision Tree.*
d. Pelabelan otomatis menggunakan *IndoBERT*
e. Klasifikasi dalam penelitian ini dibagi menjadi dua kategori yaitu kelas positif dan kelas negatif.

1.5 Tujuan Penelitian

Tujuan penelitian dalam penelitian ini sebagai berikut:

a. Menganalisis pengaruh penerapan kamus *slang word* pada tahap preprocessing terhadap akurasi model sentimen analisis Bahasa Indonesia.
b. Membuat korpus yang berisi kata *slang* berbahasa Indonesia.

1.6 Manfaat Penelitian

Manfaat penelitian dalam penelitian ini sebagai berikut:

a. Sebagai bahan referensi untuk penelitian analisis sentimen selanjutnya dalam menerapkan kamus *slang word* dalam tahapan *preprocessing*.
b. Menyediakan korpus yang berisi *slang word* untuk penelitian analisis sentimen selanjutnya.