BAB I

PENDAHULUAN

1.1. Latar Belakang

Bahasa adalah sebuah bagian yang sangat penting dalam hidup manusia karena manusia biasa berkomunikasi satu sama lain dengan mudah melalui bahasa. Hebatnya, makna bahasa bukan hanya terkait pada pemilihan kata, namun bagaimana cara menyampaikannya. Ketika pembicara tidak dapat menyampaikan pesannya dengan menggunakan satu bahasa, pembicara perlu ubah bahasa menjadi lebih dipahami. Saat ini, bahasa gaul telah menjadi fenomena di era generasi milenial, tidak hanya untuk orang-orang yang tinggal di kota begitu juga mereka yang tinggal di desa tahu bahasa gaul, bahasa seolah-olah sudah menjadi tren di kalangan milenial. Sementara di sisi lain, orang tua dan orang dewasa mengeluh tentang tata bahasa dan ucapan bahasa anak-anak mereka yang sulit untuk dipahami. Apalagi media sosial yang semakin berkembang saat ini, yang dimulai dari Short Message Service (SMS) ke Instagram, Facebook, WhatsApp, Twitter, dan lain-lain[1].

Hornby (1974) menyatakan bahwa bahasa gaul adalah kata, frasa, arti kata, yang biasa digunakan dalam percakapan di antara teman-teman atau kolega, tetapi tidak cocok untuk tulisan yang bagus atau acara resmi. Slang adalah bahasa (kata-kata, frase, dan penggunaan) dari register informal bahwa anggota kelompok khusus seperti remaja, musisi, atau penjahat mendukung untuk membangun identitas kelompok. Bahasa Slang biasanya populer di kalangan generasi milenial[2].

Algoritma Naïve Bayes adalah algoritma untuk mengklasifikasikan data berdasarkan Teorema Bayes mengasumsikan independensi parameter satu sama lain. Teorema Bayes menyediakan cara untuk menghitung probabilitas nilai parameter menggunakan nilai yang lain parameter[3]. Probabilitas berdasarkan fitur eksklusif pada data timbul menjadi anggota pada urutan probabilitas dan diperoleh menggunakan menghitung frekuensi setiap nilai fitur kelas berdasarkan pembinaan dataset. Training dataset merupakan subset yang dipakai buat melatih prosedur pemecahan klasifikasi. Proses training memakai nilai yg diketahui buat memprediksi nilai yang belum diketahui[4].

Algoritma Decision Tree adalah metode klasifikasi yang umumnya digunakan di berbagai jenis bidang, seperti *machine learning*, *image processing*, dan identifikasi pola[5]. Algoritma ini menjadi model yang paling banyak digunakan dan populer karena sederhana dan mudah dipahami. Decision Tree dihasilkan dari data training dalam arah top-down. Node akar dari pohon keputusan adalah keadaan awal pohon-simpul keputusan pertama. Setiap node dalam pohon berisi beberapa data. Atas dasar suatu algoritma, beberapa perhitungan diselesaikan dan simpul pohon keputusan dibagi menjadi dua atau lebih cabang. Dalam beberapa kasus, node tidak dapat dipecah, dalam hal ini akan menjadi node keputusan akhir[6].

1.2. Rumusan Masalah

Berdasarkan latar belakang yang sudah dijabarkan diatas, peneliti ingin menganalisa seberapa akurat algoritma Naive Bayes dan Decision Tree dalam mengklasifikasi Bahasa Slang Pada Sosial Media serta melihat apakah algoritma Naive Bayes dan Decision Tree lebih efisien jika dibandingkan dengan penelitian sebelumnya yang menggunakan algoritma berbeda?

1.3. Tujuan dan Manfaat

1.3.1. Tujuan

Tujuan penelitian ini adalah untuk melihat banyaknya penggunaan bahasa Slang di Sosial Media dan bertujuan untuk melihat seberapa akurat algoritma Naive Bayes dan Decision Tree dalam penelitian ini.

1.3.2. Manfaat

1. Agar tidak mengganggu pemahaman bahasa bagi orang lain.
2. Dapat menambahkan kosakata dalam ilmu pengetahuan di bidang linguistik.
3. Penelitian ini diharapkan dapat membagi pemahaman khususnya masalah bahasa slang dan dapat menjadi sebuah rancangan dalam pembuatan kamus bahasa slang.

1.4. Batasan Masalah

Batasan- batasan masalah dalam penelitian ini sebagai berikut :

1. Penelitian ini hanya menggunakan algoritma Naive Bayes dan Decision Tree untuk membandingkan tingkat akurasi mana yang lebih tinggi dalam menentukan hasil bahasa slang yang ada dalam dataset.
2. Dataset yang di ambil berupa komentar dari postingan di Tiktok dan Instagram.
3. Jumlah sampel dataset adalah sebanyak 2.000 sampel dengan 4 fitur.
4. Data didapatkan dengan metode web scraping.
5. Proses pengolahan data dilakukan menggunakan software rapidminer studio.
6. Jumlah data akan digunakan sebagai untuk data training sebanyak 1.200, data validasi sebanyak 400 dan data testing sebanyak 400.

1.5. Keterbaruan

1. Menurut Handayani I, (2020) Hasil pengujian sistem menggunakan K-NN metode klasifikasi mampu mengklasifikasikan 438 data dengan benar. Sebagai perbandingan, Naïve Bayes benar mengklasifikasikan 428 data menggunakan k-fold cross-validation uji dengan $k = 10$. Hal ini menunjukkan bahwa K-NN metode klasifikasi memiliki akurasi lebih tinggi daripada metode klasifikasi Naïve Bayes pada data yang digunakan. Metode K-NN mendapatkan akurasi yang lebih tinggi karena Algoritma Naïve Bayes adalah algoritma parametrik yang mengasumsikan bahwa setiap atribut data adalah independen, yang merupakan properti langka di dunia nyata.
2. Menurut Wibawa AP, (2019) Data diklasifikasikan ke dalam beberapa label yaitu Q1, Q2, Q3, Q4 dan NQ. Dalam penelitian ini variabel yang digunakan adalah H index, SJR, Total Docs. (2017), Total Docs. (3years), Total Refs, Total Cites (3years), Citable Docs. (3years), Cites / Doc. (2years), and Ref. / Doc. Klasifikasi dari kualitas jurnal dapat memudahkan masyarakat untuk memilih jurnal yang berkualitas. Dalam studi ini, peneliti juga menyimpulkan bahwa algoritma Naïve Bayes Classifier mampu mengklasifikasikan kualitas jurnal, meskipun nilai akurasinya tidak terlalu optimal.
3. Menurut Charbuty B, (2021) Decision Tree memiliki beberapa kemungkinan kekhawatiran tentang ketahanan, adaptasi skalabilitas dan optimalisasi ketinggian. Namun, berbeda dengan metode klasifikasi data lainnya, Decision

Tree membuat kumpulan aturan yang efisien yang mudah dipahami. Makalah ini mengulas penelitian terbaru yang dilakukan di banyak bidang, seperti analisis penyakit medis, klasifikasi teks, klasifikasi smartphone pengguna dan gambar, dll. Selanjutnya, detail yang digunakan dalam teknik/algorithm, kumpulan data yang digunakan oleh penulis dan hasil yang dicapai terkait dengan akurasi diringkas untuk decision tree. Akhirnya, akurasi terbaik dicapai untuk algoritma decision tree adalah 99,93% ketika menggunakan mesin belajar repositori sebagai dataset.