1. Pendahuluan

1.1 Latar Belakang

Analisis sentimen adalah bidang studi yang menganalisis opini, penilaian dan sikap terhadap entitas seperti produk, layanan, organisasi, individu, topik, dan atributnya [1]. Analisis sentimen merupakan klasifikasi yang mengelompokkan teks berdasarkan opini atau sentimen yang terkandung di dalamnya [2]. Proses klasifikasi dapat dilakukan dengan menggunakan metode naive bayes. Konsep naive bayes dapat dikatakan sederhana dan intuitif, metode yang efisien dan memberikan akurasi yang tinggi [2, 3, 4].

Di samping penggunaan metode naive bayes yang terbukti memberikan hasil klasifikasi yang baik pada banyak penelitian, data opini yang digunakan adalah data tidak terstruktur, dan banyaknya jumlah fitur dapat menimbulkan masalah tingginya dimensionalitas data. Dimensionalitas data yang tinggi dapat menyebabkan data *overload*, menyimpang dan dapat mempengaruhi proses klasifikasi [5]. Maka dari itu diperlukan proses untuk memilih fitur yang relevan, dengan kriteria untuk mendapatkan fitur yang optimal [6]. Seleksi fitur dapat dilakukan dengan melihat nilai berdasarkan hasil evaluasi dengan kriteria tertentu [5]. Seleksi fitur yang pada beberapa penelitian [7, 8, 9] memberikan performansi yang baik adalah metode chi-square dengan konsep untuk melihat ketergantungan antar dua variabel. Oleh karena itu pada penelitian ini seleksi fitur akan dilakukan dengan *chi square*, dan hasilnya akan dipakai untuk klasifikasi sentimen menggunakan naive bayes.

Sehubungan dengan analisis sentimen, berdasarkan data hingga bulan Januari 2019, pengguna ponsel terdaftar di Indonesia mencapai 355,5 juta atau 133% dari populasi Indonesia yang berjumlah 268,2 juta jiwa [10]. Berdasarkan data tersebut, sentimen masyarakat tentang penyedia layanan telekomunikasi dapat dijadikan data untuk dianalisis menggunakan metode klasifikasi naive bayes dan seleksi fitur *chi square*.

1.2 Topik dan Batasan

Berdasarkan latar belakang, masalah yang dapat dirumuskan pada penelitian ini adalah bagaimana pengaruh seleksi fitur *chi square* terhadap metode naive bayes apabila diterapkan pada klasifikasi opini berbahasa Indonesia terhadap penyedia layanan telekomunikasi. Studi kasus yang dilakukan terbatas pada penyedia layanan telekomunikasi terkemuka di Indonesia yang dilansir dari situs *Finder* [11] yaitu Telkomsel, IM3 Ooredoo, XL Axiata dan Tri. Data yang digunakan berupa tweet terhadap akun resmi penyedia layanan telekomunikasi terpilih dengan username Twitter @Telkomsel, @IM3Ooredoo, @triindonesia, dan @myXLCare. Data teks yang digunakan sebagai dataset klasifikasi berjumlah 1200 tweet, dimana berdasarkan penelitian sebelumnya, klasifikasi dengan naive bayes memberikan akurasi tinggi dengan 70% data latih dan 30% data uji [12]. Metode naive bayes, berdasarkan persebaran datanya dapat dibedakan menjadi tiga jenis [13], yaitu bernoulli naive bayes, gaussian naive bayes, dan satu lagi yang digunakan pada penelitian ini yaitu multinomial naive bayes yang cocok digunakan dalam pengklasifikasian data teks atau dokumen [14]. Selanjutnya, pengukuran yang digunakan untuk melihat performansi metode yang diterapkan adalah akurasi, presisi, *recall* dan *f1-score*.

1.3 Tujuan

Pendapat seseorang terhadap penyedia layanan telekomunikasi sekarang makin mudah untuk disampaikan secara langsung ke akun resmi Twitter penyedia telekomunikasi terkait. Pendapat tersebut dapat dijadikan data untuk selanjutnya diklasifikasikan berdasarkan polaritas pendapat yang terkandung di dalamnya. Naive Bayes adalah metode klasifikasi dengan performansi yang baik [2, 3, 4], dengan asumsi ketidaktergantungan atau independensi yang kuat. *Chi square* merupakan metode statistik yang mengukur seberapa dependen atau tingkat ketergantungan antara dua variabel, dimana pada kasus ini berarti ketergantungan antara *term* dengan kategori [15] yang pada beberapa penelitian sebelumnya [7, 8, 9] dapat meningkatkan performansi untuk proses klasifikasi.

Tujuan dari penelitian ini adalah untuk mengetahui bagaimana pengaruh seleksi fitur *chi square* terhadap metode klasifikasi naive bayes untuk data opini berbahasa Indonesia tentang penyedia layanan telekomunikasi.

1.4 Organisasi Tulisan

Organisasi penulisan ini dimulai dengan latar belakang, topik, batasan dan tujuan penelitian yang termasuk dalam pendahuluan. Selanjutnya ada studi terkait yang berisi tentang penelitian-penelitian yang berhubungan dengan analisis sentimen dan metode yang digunakan. Ketiga, sistem yang dibangun, gambaran umum sistem beserta penjelasan setiap tahapnya. Keempat, dituliskan hasil implementasi dan analisis berdasarkan penelitian yang dilakukan. Kelima adalah poin kesimpulan serta saran.