1. Pendahuluan

Latar Belakang

Ketika wisatawan ingin memilih hotel yang nyaman untuk perjalanannya, mereka akan mencari ulasan dari wisatawan lainnya[1]. 93% manajemen hotel menyatakan bahwa ulasan wisatawan *online* sangat penting bagi masa depan properti mereka[2]. Ulasan yang ada di internet lebih relevan, aktual dan rinci dibandingkan ulasan yang ditemukan pada brosur hotel[3]. Tetapi dengan jumlah ulasan yang sangat banyak, mereka tidak dapat memahami dan menyimpulkan semua ulasan hotel apakah mengandung opini positif atau opini negatif.

Beberapa website ulasan hotel hanya memberikan *ratings* yang dianggap tidak bersifat objektif sehingga tidak dapat digunakan sebagai perbandingan antar hotel[3]. Misalnya pada ulasan Hotel Ibis, terdapat ulasan: "Bad service and disappointing facilities"[15]. Tetapi tamu hotel tersebut memberikan skor 4 bintang. Sedangkan pada Hotel Aston terdapat ulasan: "I like to stay here a lot. The wall and lobby decoration full of Indonesia reliefs and carving"[16]. Tamu hotel tersebut memberikan skor 3 bintang. Oleh karena itu dari hasil ulasan tersebut Hotel Ibis belum tentu lebih baik dibandingkan Hotel Aston dan begitu juga sebaliknya. Contoh diatas dapat diambil kesimpulan bahwa penilaian dengan sistem *ratings* berbeda dengan ulasan yang ditulis sehingga tidak dapat digunakan untuk perbandingan hotel karena informasi yang diberikan tidak jelas.

Pada riset sebelumnya dilakukan klasifikasi sentimen menggunakan NLP dan *Bayesian Classification*[4]. Metode algoritma *Naïve Bayes* terbukti efektif di dalam klasifikasi sentimen dan memiliki performa yang lebih baik dibandingkan metode lainnya seperti *k-neighbor classifier* (KNN) [4]. Pada tugas akhir ini penulis melakukan penelitian terhadap klasifikasi sentimen otomatis untuk ulasan hotel yang diberikan oleh tamu hotel. Dataset diambil dari *Datafiniti's Business Database* yang merupakan situs database ulasan produk, hotel dan properti. Penelitian ini memilih metode *Multinomial Naïve Bayes* pada ulasan hotel dengan mengharapkan akurasi yang lebih bisa dicapai. *Multinomial Naïve Bayes* ini juga digunakan karena kecepatan dan kesederhanaanya pada klasifikasi teks[5]. Selain itu metode *Multinomial Naïve Bayes* mengikuti prinsip dari distribusi multinomial yang digunakan untuk pengolahan teks. Pada penelitian sebelumnya digunakan *Bag-of-Words* untuk ekstraksi fiturnya yang memiliki kelemahan seperti *running time* yang lebih lama dan hasil yang tidak optimal karena menampung banyak fitur[1]. Pada klasifikasi teks, seleksi fitur dapat meningkatkan efisiensi dan akurasi sehingga digunakan untuk penelitian ini [18]. Penulis melakukan *preprocessing* terhadap data yaitu *case folding, remove punctuation, stopword removal, lemmatization* dan *tokenization.* Oleh karena itu penelitian ini memberikan solusi membuat model seleksi fitur pada klasifikasi teks dengan *Multinomial Naïve Bayes* untuk membandingkan performa model pada *Bag-of-Words.* Seleksi fitur yang diusulkan ada dua yaitu seleksi fitur *frequency-based* dan seleksi fitur dengan menghapus fitur yang memiliki nilai probabilitas positif dan negatif minimal. Selain itu penulis menambahkan *preprocessing* untuk mengetahui pengaruh performa pada setiap model.

Topik dan Batasannya

Berdasarkan latar belakang yang sudah dijelaskan, penelitian ini membuat klasifikasi sentimen terhadap ulasan hotel menjadi orientasi positif dan negatif. Metode klasifikasi menggunakan *Multinomial Naive Bayes* sebagai *classifier* dan melakukan *preprocessing* terhadap dataset ulasan hotel. Selain itu melakukan perbandingan terhadap pengaruh *Bag-of-Words* dan seleksi fitur terhadap performa model yang dibuat.

Rumusan masalah pada penelitian ini adalah bagaimana pengaruh *preprocessing* terhadap performa klasifikasi dan bagaimana perbandingan performa *Bag-of-Words* dan seleksi fitur untuk mencari performa yang paling optimal. Penelitian ini membuat beberapa skenario sebagai perbandingan model yang telah dibuat.

Pada penelitian ini terdapat beberapa batasan masalah, yaitu dataset yang digunakan adalah dataset ulasan hotel bahasa inggris dan jumlah *record* pada dataset adalah 5000 *record*. Pada dataset ini dilakukan pengurangan dimensi sesuai kebutuhan penelitian sehingga hanya satu dimensi yang digunakan yaitu dimensi teks. Penelitian ini menggunakan teknik pelabelan secara manual dengan pelabelan positif dan negatif. *Preprocessing* yang digunakan adalah *case folding, remove punctuation, stopword removal, lemmatization* dan *tokenization.* Untuk validasi penelitian ini menggunakan 10 cross fold validation karena bisa meningkatkan akurasi yang lebih optimal[14].

Tujuan

Penelitian ini memiliki beberapa tujuan untuk mengatasi latar belakang yang sudah dijelaskan. Tujuan dari penelitian ini adalah melakukan pengujian untuk mengetahui pengaruh performa *preprocessing* terhadap klasifikasi dan melakukan perbandingan performa antara *Bag-of-Words* dan seleksi fitur yang ditambahkan pada proses klasifikasi untuk mengetahui metode yang paling optimal pada klasifikasi sentimen.

Organisasi Tulisan

Bagian selanjutnya pada penelitian ini adalah bagian 2 yang membahas studi terkait pada penelitian yang dilakukan, bagian 3 yang membahas teori dan perancangan sistem penelitian, bagian 4 yang membahas evaluasi model penelitian, dan bagian 5 yang membahas kesimpulan.