BAB I
PENDAHULUAN

1.1 Latar Belakang

Dengan adanya platform daring, tamu dapat dengan mudah meninggalkan ulasan tentang hotel tempat mereka menginap sehingga membantu calon tamu selanjutnya membuat keputusan tentang pilihan tempat penginapan mereka [1]. Ulasan dari tamu mempunyai pengaruh besar terhadap reputasi dan pendapatan hotel, terutama ulasan negatif yang dianggap lebih bisa dipercaya daripada ulasan positif [2]. Sangat penting bagi pihak hotel untuk mengidentifikasi faktor-faktor yang dipentingkan oleh tamu agar bisa meningkatkan kepuasan tamu sehingga mendorong tamu meninggalkan ulasan yang positif [3]. Tanggung jawab tersebut jatuh pada analis data, yang bertanggung jawab mengembangkan keputusan atau rekomendasi aksi terhadap masalah nyata berdasarkan wawasan yang dihasilkan dari riwayat data [4].

Demi mengidentifikasi faktor kunci kepuasan tamu, seorang analis data perlu menelusuri ulasan tamu satu per satu. Namun, jumlah ulasan yang tersedia secara daring sangat banyak dan tidak terstruktur sehingga sulit bagi analis untuk mengekstraksi informasi bermakna secara manual [5]. Untuk mengatasi masalah tersebut, [6] dan [7] menggunakan teknik pemodelan topik untuk menemukan topik yang sering dibahas di dalam ulasan tamu secara otomatis. Pemodelan topik adalah teknik di bidang NLP (*Natural Language Processing*) yang bertujuan untuk menemukan topik-topik di dalam sekumpulan dokumen yang berjumlah banyak sehingga sulit dianalisis secara manual [8]. Selain di industri perhotelan, pemodelan topik juga pernah digunakan untuk analisis dampak berita terhadap ekonomi [9], merangkum perkembangan di bidang AI [10], dan identifikasi penyebab kecelakaan turbin angin [11].

Perbandingan metode pemodelan topik yang sudah ada di [5], [9] dan [12] menunjukkan bahwa dibandingkan dengan metode sebelumnya, BERTopic cenderung menemukan topik yang lebih berkualitas dan spesifik untuk teks pendek seperti pesan pengguna di X dan ulasan pelanggan di *e-commerce*. Selain itu, dibandingkan dengan metode pemodelan topik lainnya, BERTopic dapat menemukan jumlah topik secara otomatis tanpa harus menguji kandidat jumlah topik satu per satu [12], sehingga cocok digunakan dalam analisis data eksploratif yang memerlukan lebih sedikit pengaturan parameter [13].

Untuk mempermudah analisis data eksploratif dengan model topik, diperlukan juga sebuah aplikasi yang bisa mempresentasikan topik-topik yang ditemukan dengan jelas kepada pengguna. Sistem analitik visual (*visual analytics*) muncul sebagai strategi yang cocok untuk membantu proses analisis dengan memanfaatkan visualisasi interaktif. Analitik visual dapat memberikan gambaran umum tingkat tinggi dari data teks dengan memvisualisasikan topik dan memungkinkan pengguna untuk menyaring dan memeriksa sesuai dengan kebutuhan analitis mereka [14]. Tapi, sistem analitik visual yang sudah ada di dalam bidang pemodelan topik seperti VISTopic [15] dan UTOPIAN [16] hanya mempertimbangkan topik-topik yang ditemukan dari algoritma pemodelan topik, tanpa mempertimbangkan hubungan topik dengan data lain.

Penelitian di [1], [17], dan [18] membuktikan bahwa wawasan yang baru dapat diperoleh dengan mencari hubungan antara data teks dengan data lain menggunakan regresi, sehingga membantu penulis dari ketiga penelitian tersebut menarik kesimpulan yang lebih informatif tentang *dataset* yang ditelusuri. Persepsi dan ekspektasi tamu dapat dipengaruhi oleh banyak hal, seperti waktu [1], kewarganegaraan [2], dan lokasi kunjungan [6]. Oleh karena itu, topik-topik yang ditemukan oleh algoritma pemodelan topik perlu dipelajari hubungannya dengan data-data kuantitatif di *dataset* baik jenis ruang, jenis fasilitas, atau kewarganegaraan tamu, demi membantu analis data memperoleh pemahaman yang lebih mendalam tentang pendapat tamu.

Ulasan online dari tamu hotel memiliki peran penting dalam membentuk reputasi dan pendapatan hotel, sehingga sangat penting bagi analis data untuk menelusuri ulasan tamu untuk memahami faktor-faktor kunci kepuasan tamu. Algoritma pemodelan topik yaitu BERTopic dan sistem analitik visual dapat membantu analis memahami informasi yang terkandung di dalam teks ulasan tanpa membutuhkan analis membaca ulasan satu per satu. Selain itu, topik yang ditemukan oleh BERTopic dapat dimodelkan dengan regresi untuk menemukan hubungan antara topik dengan data kuantitatif lain. Oleh karena itu, kami memutuskan untuk mengangkat judul: **IMPLEMENTASI BERTOPIC UNTUK ANALISIS FAKTOR KUNCI KEPUASAN TAMU BERDASARKAN KORELASI ANTAR DATA KUANTITATIF DENGAN DATA TEKS DARI ULASAN HOTEL**.

1.2 Rumusan Masalah

Berdasarkan uraian pada latar belakang di atas, maka rumusan masalah penelitian ini adalah sebagai berikut: Apakah terdapat hubungan antara topik yang ditemukan algoritma BERTopic dengan data kuantitatif lain di dalam *dataset*?

1.3 Tujuan

Tujuan dari penelitian ini adalah untuk menyediakan sebuah aplikasi yang mampu mengidentifikasi topik yang terkandung dalam data ulasan hotel, mencari hubungan topik dengan data-data kuantitatif lain di dalam *dataset*, serta mempresentasikan hasil analisis dengan jelas kepada analis data, sehingga analis data dapat mengidentifikasi faktor yang berkontribusi pada kepuasan tamu dengan mudah.

1.4 Manfaat

Manfaat yang dapat diperoleh dari tugas akhir ini adalah sebagai berikut:

1. Analis data dapat menggunakan aplikasi yang dikembangkan untuk memahami bagaimana tamu memandang layanan dan fasilitas hotel berdasarkan ulasan yang diberikan, sehingga mereka bisa membantu manajer hotel membuat keputusan tentang strategi bisnis selanjutnya.
2. Skripsi ini menjadi referensi bagi penelitian selanjutnya di bidang pemodelan topik.

1.5 Ruang Lingkup

Ruang lingkup dari tugas akhir ini adalah sebagai berikut:

1. Metode pemodelan topik yang digunakan adalah BERTopic.
2. Hubungan antara topik dan data lain dimodelkan menggunakan model regresi linier, logistik, multinomial-logistik, dan *proportional odds* (regresi ordinal).
3. Regresi hanya dilaksanakan menggunakan variabel independen biner.
4. Analisis regresi hanya digunakan untuk membuat kesimpulan tentang apakah topik dari BERTopic bisa menjelaskan variabel dependen atau tidak, dan tidak digunakan untuk memprediksi data kuantitatif atau untuk membuat kesimpulan yang otoritatif tentang faktor apa yang berkontribusi kepada kepuasan tamu.
5. Pengguna aplikasi perlu mengetahui cara menginterpretasi koefisien dari model regresi.

6. Aplikasi dijalankan secara lokal di PC atau *laptop,* dan tampilan dari aplikasi diakses melalui *localhost* menggunakan *browser*.
7. Aplikasi menerima file dalam bentuk CSV atau XLSX sebagai *input*. Aplikasi menghasilkan visualisasi topik yang ditemukan oleh BERTopic, tabel *dataset* yang sudah di-*preprocess*, dan visualisasi korelasi topik dengan data kuantitatif lain di *dataset* sebagai *output*.
8. Aplikasi hanya bertanggung jawab atas *preprocessing* teks sebelum teks diproses oleh model topik. *Preprocessing* data lain di dalam *dataset* harus dilaksanakan oleh pengguna aplikasi sebelum memasukkan *dataset* ke aplikasi.
9. Aplikasi umumnya hanya akan fokus mencari hubungan topik dengan data kuantitatif lain di *dataset*, dan tidak mempertimbangkan hubungan antara data kuantitatif yang tidak meliputi topik.
10. Topik yang ditemukan oleh algoritma pemodelan topik dijabarkan dalam bentuk himpunan kata kunci yang paling bisa menjelaskan informasi di dalam dokumen-dokumen yang ada, bukan teks rangkuman.
11. Aplikasi yang dirancang tidak akan mendukung pemodelan topik berhierarki, pemodelan topik terpandu, atau pemodelan topik *semi-supervised*.
12. Walaupun aplikasi dapat menerima data waktu/temporal, aplikasi hanya mendukung analisis berbentuk regresi ordinal dan tidak mendukung analisis statistika untuk data *time-series*.
13. Penelitian ini mengambil beberapa dataset berupa ulasan hotel berbahasa Inggris dari *Kaggle* untuk mengevaluasi kegunaan dari aplikasi yang akan dibuat:
    a. "Hotel Reviews Booking.com" oleh Michel Hatab. https://www.kaggle.com/datasets/michelhatab/hotel-reviews-bookingcom
    b. "Datafiniti-DB", oleh Datafiniti & Shion https://www.kaggle.com/datasets/datafiniti/hotel-reviews
    c. "*A Dataset of TripAdvisor Guest Reviews for Major Hotels in Salalah, Oman*", oleh Ricardo Biason. https://data.mendeley.com/datasets/dkfwj76kx6/3

14. Sesuai dengan rumusan masalah di atas, metode evaluasi yang akan digunakan adalah sebagai berikut:
    a. Usabilitas aplikasi diuji menggunakan pengujian *blackbox*
    b. Optimisasi *hyperparameter* model topik untuk mendapatkan hasil model topik yang paling koheren.
    c. Bangun model regresi menggunakan topik sebagai variabel independen dan data lain di dalam *dataset* sebagai variabel dependen. Nilai p dan ukuran efek dari model regresi digunakan untuk menjawab rumusan masalah.
15. Aplikasi akan dikembangkan dalam bahasa Inggris agar bisa disebarluaskan melalui GitHub.