BAB III LANDASAN TEORI

Pada subbab ini dijabarkan teori yang mendukung bagi penulis dalam menyusun thesis ini.

A. *Review* (Ulasan) Restoran Cepat Saji di Media Sosial

Pelanggan sering menuliskan ulasan atau pendapat mereka tentang pengalaman mengenai suatu instansi atau tempat yang pernah mereka kunjungi [20]. Pendapat dari pelanggan menjadi sumber informasi yang penting agar dapat memahami bagaimana pandangan pelanggan tersebut tentang pelayanan dari sebuah instansi [1]. Pendapat pelanggan yang terdapat pada media sosial termasuk data yang *real-time* dibandingkan dengan data yang didapatkan melalui pengisian kuisioner. Ulasan konsumen dalam bentuk teks memberikan gambaran tentang perasaan konsumen terhadap restoran tersebut. Hal ini juga dapat mempengaruhi perilakuk dari konsumen lain yang berinteraksi dengan ulasan ini. Ulasan dari konsumen termasuk berpengaruh kuat dikarenakan berasal langsung dari perspektif konsumen [21].

B. Penambangan Teks atau *Text Mining*

Penelitian ini menerapkan konsep penambangan teks pada media sosial Twitter. Penambangan teks adalah variasi pada bidang yang disebut penambangan data, yaitu mencoba menemukan pola yang menarik dari basis data besar. Penambangan teks adalah teknologi baru yang berupaya melakukan ekstraksi informasi bermakna dari sekumpulan data tekstual. Penambangan teks adalah perpanjangan dari penambangan data ke data tekstual. Untuk mendapatkan

informasi yang berguna dari sejumlah besar dokumen teknis dengan cepat, telah menjadi keharusan untuk menggunakan teknik komputer otomatis [22]. Adopsi luas alat media sosial telah menghasilkan banyak data tekstual, yang berisi pengetahuan tersembunyi bagi bisnis [23].

C. *Topic Modelling*

Teks yang didapat dari proses pengambilan data selanjutnya akan diolah dengan metode *topic modelling*. Peneliti pembelajaran mesin telah mengembangkan pemodelan topik (*topic modelling*) probabilistik, sebagai serangkaian algoritma yang berfungsi dalam menemukan dan memberi catatan pada arsip besar dokumen dengan menyajikan informasi yang bersifat tematik. Pemodelan topik adalah metode statistik yang memproses sekumpulan kata dalam sebuah untuk mendapatkan tema pada teks dan bagaimana tema-tema tersebut saling terhubung. Algoritma pemodelan topik tidak membutuhkan pelabelan dokumen sebelumnya (topik muncul berdasarkan hasil analisis teks asli). Pemodelan topik membantu dalam mengatur arsip elektronik, terutama dalam skala yang sulit dilakukan oleh manusia. Tujuan pemodelan topik adalah untuk secara otomatis menemukan topik dari kumpulan dokumen [24].

D. *Latent Dirichlet Allocation*

Latent Dirichlet Allocation (LDA) adalah model probabilistik generatif dari *corpus*. LDA adalah algoritma *Machine Learning* tanpa pengawasan (*unsupervised*) yang menganalisis dan menemukan topik di antara koleksi dokumen yang besar. Teknik ini bergantung pada pendekatan "kantong kata", yang menganggap setiap dokumen yang ada sebagai vektor jumlah kata. Setiap dokumen

direpresentasikan menjadi berupa distribusi probabilitas pada sejumlah topik, dimana setiap topik direpresentasikan menjadi berupa distribusi probabilitas pada sejumlah kata [25]. Untuk setiap dokumen dalam koleksi, dihasilkan kumpulan kata dalam proses yang meliputi dua tahap [24], yaitu :

a. Memilih sebuah distribusi topik secara acak.

b. Untuk setiap kata yang terdapat dalam dokumen :

   1) Memilih secara acak dari distribusi topik pada langkah pertama.

   2) Memilih sebuah kata secara acak dari distribusi kata yang sesuai.

Algoritma *sampling* yang paling umum digunakan untuk pemodelan topik adalah *Gibbs Sampling* [24], seperti yang dapat dilihat pada formula (1).

$$P(z_i = j|t_i, d_i, z_{-i}) \propto \frac{C_{t,j}^{(TZ)} + \beta}{\sum_t C_{t,j}^{TZ} + T\beta} \frac{C_{t,j}^{(DZ)} + \alpha}{\sum_z C_{d,z}^{DZ} + Z\alpha} \quad (1)$$

Dimana C^{TZ} mengelola perhitungan dari semua *topic-term assignments*, C^{DZ} melakukan perhitungan dari *document-topic assignments*, z_{-i} melambangkan semua *topic-term* dan *document-topic assignments* kecuali untuk penetapan saat ini dari z_i untuk *term* t_i, lalu α dan β adalah parameter *Dirichlet* [25]. Untuk menghitung probabilitas topik dari setiap kata digunakan rumus formula (2) dimana φ_{tz} adalah probabilitas dari kata kata t untuk topik z.

$$\varphi_{tz} \propto \frac{C_{t,j}^{(TZ)} + \beta}{\sum C_{t,j}^{TZ} + T\beta} \quad (2)$$

Sedangkan untuk menghitung θ_{dz} yang merupakan proporsi topik z dari setiap dokumen d dapat digunakan rumus seperti pada formula (3) [25]:

$$\theta_{dz} \propto \frac{C_{d,z}^{(DZ)} + \alpha}{\sum C_{d,z}^{DZ} + Z\alpha} \quad (3)$$

Setelah dilakukan analisis pemodelan topik, dilanjutkan dengan melakukan visualisasi dari hasil tersebut agar lebih dapat dipahami bagaimana penyebaran dari topik-topik tersebut. Visualisasi yang dilakukan menggunakan metode *Intertopic Distance Map, Intertopic Distance Map* adalah visualisasi topik dalam ruang dua dimensi. Teknik ini memperlakukan kata sebagai simpul (*node*) dalam sebuah jaringan dan dapat memperlihatkan keterkaitan antara kata-kata tertentu [18]. Hasil dari analisis ini memberikan hasil yang tidak sepenuhnya dikategorikan, namun lebih menekankan pada hubungan antar objek (kata). Lingkaran pada grafik dikalkulasikan menggunakan algoritma penskalaan multidimensi yang tidak dapat atau sulit sekali dibayangkan dengan otak manusia, lalu kemudian dibentuk menjadi sejumlah dimensi yang wajar berdasarkan kata-kata yang dianalisis, sehingga topik yang lebih dekat memiliki lebih banyak kata yang sama [11]. Pemahaman atas suatu fenomena yang dianalisis menggunakan metode ini menjadi dapat dipahami dengan lebih jelas [19].