

BAB II

LANDASAN TEORI

2.1 Twitter

Twitter adalah layanan jaringan sosial *microblog*. Dikatakan *microblog* karena pengguna *twitter* mengirim dan membaca pesan blog seperti pada umumnya. Pesan yang disiarkan oleh pengguna *twitter* biasa disebut dengan *tweet*. *Tweet* dapat dilihat secara publik, namun pengirim pesan dapat membatasinya hanya kepada teman-teman tertentu saja. Beberapa istilah yang ada pada *twitter* adalah sebagai berikut (Novantirani, Sabariah, dan Effendy, 2015):

1. *Tweet*, adalah fitur umum dimana pengguna *twitter* dapat mengirimkan pesan berupa teks, foto, dan video.
2. *Retweet*, adalah *tweet* yang dibagikan ulang oleh pengikut dari pengguna *twitter*.
3. *Hashtag*, merupakan simbol “#” yang berguna untuk menandai topik atau kata kunci agar mempermudah pencarian.
4. *Mention*, adalah cara untuk menandai atau memberi balasan kepada pengguna lainnya dengan memberi awalan “@” kemudian dilanjutkan dengan nama pengguna *twitter*.

2.2 Data Mining

Data mining merupakan suatu istilah yang digunakan untuk menggali atau menemukan pengetahuan di dalam suatu *database*. *Data mining* adalah suatu proses yang menggunakan teknik statistika, matematika, kecerdasan tiruan, dan *machine learning* untuk memperoleh serta mengidentifikasi informasi dari *database* yang besar. *Data mining* bisa juga diartikan sebagai pengekstrakan informasi baru yang diperoleh dari *database* yang besar sehingga dapat berguna dalam pengambilan keputusan (Tan, Steinbach, dan Kumar, 2006).

Istilah *data mining* juga disebut *knowledge discovery* (KDD). Han dan Kamber (2006) menyatakan bahwa langkah-langkah *knowledge discovery* (KDD) sebagai berikut:

1. *Data cleaning*, merupakan proses penghapusan data yang tidak konsisten serta menghilangkan *noise*.
2. *Data integration*, merupakan proses penggabungan atau mengombinasikan data dari berbagai sumber.
3. *Data selection*, merupakan proses yang digunakan untuk menyeleksi data yang relevan yang akan digunakan pada proses *data mining*.
4. *Data transformation*, merupakan proses yang digunakan untuk mentransformasikan data menjadi bentuk yang sesuai dalam proses *data mining*.
5. *Data mining*, merupakan proses metode-metode atau algoritma untuk menghasilkan suatu pola pada data.
6. *Pattern evaluation*, merupakan proses yang digunakan untuk menguji kebenaran dari pola data yang diperoleh.
7. *Knowledge representation*, merupakan proses visualisasi yang bertujuan untuk menyajikan hasil *mining* kepada pengguna.

2.3 Text Mining

Text mining adalah sebuah proses pengetahuan intensif dimana pengguna bekerja dengan sekumpulan dokumen dengan menggunakan beberapa alat analisis. Tujuan dari *text mining* adalah untuk memperoleh informasi dari data berupa teks atau bahasa. Teks merupakan data yang tidak terstruktur dan sulit untuk menemukan algoritma untuk menyelesaikannya. Namun untuk saat ini teks merupakan alat yang paling sering digunakan dalam pertukaran informasi.

Sesuatu yang membedakan antara *text mining* dan *data mining* terletak pada sumber data yang digunakan. Pola-pola data pada *text mining* diperoleh dari data tekstual yang tidak terstruktur yang bukan berasal dari suatu *database*. Kesamaan antara *text mining* dan *data mining* adalah jumlah data yang digunakan merupakan data besar dan berdimensi tinggi serta struktur data yang terus berubah. *Data mining* mudah diproses oleh mesin atau komputer karena data tersebut terstruktur dari proses *warehousing*. Sedangkan *text mining* memiliki struktur yang kompleks,

tidak lengkap, arti yang tidak standar, dan bahasa yang berbeda (Feldman dan Sanger, 2007).

2.4 Machine Learning

Fokus utama dari *machine learning* adalah program komputer dapat belajar secara otomatis untuk mengenali pola yang kompleks dengan akurasi model yang tinggi. *Machine learning* juga dapat membuat keputusan yang tepat berdasarkan data yang digunakan. Beberapa jenis *machine learning* yang memiliki hubungan dengan *data mining* adalah (Sa'adah, Rochayani, Lestari, dan Lusia, 2021):

1. *Supervised learning* atau klasifikasi yang berdasarkan pada contoh yang telah ada pada data *training*.
2. *Unsupervised learning* atau *clustering* yang memiliki contoh input tanpa kelas label.
3. *Semi-supervised learning*, merupakan teknik dari *machine learning* yang melibatkan kelas yang memiliki label dan tidak memiliki label Ketika mempelajari model.
4. *Active learning*, merupakan pendekatan *machine learning* yang dapat membiarkan pengguna berperan dalam proses pembelajaran.

2.5 Natural Language Processing (NLP)

Natural language processing adalah bidang ilmu komputer dan linguistik yang berkaitan antara komputer dan bahasa manusia. NLP juga berarti bahwa sistem bahasa alami yang mencakup percakapan informasi dari basis data komputer ke dalam bahasa yang dapat dibaca manusia. NLP harus memperhatikan pengetahuan bahasa itu sendiri, seperti kata yang digunakan, bagaimana kata-kata tersebut digabung sehingga menghasilkan suatu kalimat, arti dari sebuah kata, dan fungsi sebuah kata pada kalimat. Tingkatan pengelolaan bahasa alami adalah sebagai berikut (Putranti dan Winarko, 2013):

1. Fonetik dan fonologi, merupakan sesuatu yang berhubungan dengan suara yang menghasilkan kata yang dapat dikenali.

2. Morfologi, merupakan pengetahuan tentang kata dan bentuknya dapat dimanfaatkan untuk membedakan setiap kata.
3. Sintaksis, merupakan pengetahuan tentang urutan kata dalam suatu kalimat dan menjadi bentuk yang sistematis.
4. Semantik, mempelajari arti dari suatu dan mengetahui bagaimana arti dari setiap kata membentuk kalimat yang utuh.
5. Pragmatik, merupakan pengetahuan yang berkaitan dengan masing-masing konteks, tergantung situasi dan tujuan pembuatan sistem.
6. *Discourage knowledge*, merupakan pengenalan mengenai suatu kalimat yang sudah dikenali sebelumnya mempengaruhi arti dari kalimat selanjutnya.
7. *Word knowledge*, mencakup arti kata secara umum dan memiliki arti khusus bagi suatu kata dalam suatu percakapan pada konteks tertentu.

2.6 Analisis Sentimen

Analisis sentimen atau biasa juga disebut sebagai *opinion mining* merupakan cara yang digunakan untuk menganalisis sentimen pendapat orang, penilaian, evaluasi, sikap, dan emosi terhadap entitas seperti produk, jasa, topik, dan masalah lainnya. Faktor pendukung yang menyebabkan analisis sentimen sangat meningkat adalah sebagai berikut (Pang dan Lee, 2008):

1. Munculnya *machine learning* dalam pengolahan bahasa dan pengambilan informasi.
2. Tersedianya dataset untuk algoritma *machine learning* khususnya pada *world wide web*.
3. Realisasi tantangan intelektual dengan aplikasi *intelligence* dan komersial yang ditawarkan

2.7 Analisis Korelasi

Analisis korelasi merupakan uji statistik yang digunakan untuk mengukur keeratan hubungan antara dua variabel. Ukuran yang digunakan untuk mengukur kuatnya hubungan antara dua variabel disebut dengan koefisien korelasi. Nilai

koefisien korelasi dapat dinyatakan sebagai berikut (Nugroho, Akbar, dan Vusvitasari, 2008):

$$-1 \leq r \leq 1 \quad (2.1)$$

dengan:

$r = 1$, hubungan X dan Y sangat kuat dan positif

$r = -1$, hubungan X dan Y sangat kuat dan negatif

$r = 0$, tidak ada hubungan antara X dan Y

Koefisien korelasi antara dua variabel pada statistik parametrik yang sering digunakan adalah koefisien korelasi Pearson. Skala pengamatan yang digunakan paling tidak berupa skala interval atau rasio dan berdistribusi normal. Persamaan yang digunakan untuk menentukan koefisien korelasi Pearson adalah sebagai berikut (Nugroho, Akbar, dan Vusvitasari, 2008):

$$r = \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left(\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right) \left(\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2\right)}} \quad (2.2)$$

Keeratan hubungan antara dua variabel tidak hanya dilihat dari besarnya koefisien korelasi, namun juga perlu dilakukan pengujian hipotesis sebagai berikut:

H_0 : tidak terdapat korelasi diantara kedua variabel

H_1 : terdapat korelasi diantara kedua variabel

Statistik uji koefisien korelasi dapat dihitung dengan menggunakan persamaan berikut (Dewanti, Suparti, Prahutama, 2020):

$$t_{hitung} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad (2.3)$$

Kriteria uji yang digunakan tolak H_0 jika nilai $|t_{hitung}| > t_{\frac{\alpha}{2}; n-2}$. Jika H_0 ditolak maka kesimpulan yang dapat diambil adalah bahwa terdapat korelasi diantara kedua variabel.