BAB III

LANDASAN TEORI

3.1. Data Mining

Data mining merupakan serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basis data atau bisa disebut dengan KDD (*Knowledge Discovery in Database*). Informasi yang dihasilkan diperoleh dengan cara mengekstraksi dan mengenali pola yang penting atau menarik dari data yang terdapat dalam basis data (Han, et al., 2011).

Knowledge Discovery in Database adalah keseluruhan proses non-trivial untuk mencari dan mengidentifikasi pola atau *pattern* dalam data dimana pola yang ditemukan bersifat sah, baru, dapat bermanfaat dan dapat dimengerti.

3.2. Opinion Mining/Analisis Sentimen

Opinion Mining/Analisis Sentimen adalah cabang penelitian dari *text mining* yang menganalisa opini, sentimen, evaluasi, penilaian, sikap, dan emosi publik mengenai suatu entitas seperti produk, layanan, organisasi, individu, isu, kejadian, topik, dan berbagai atributnya (Liu, 2012). Terdapat tiga buah subproses dari *opinion mining* yaitu *document subjectivity, opinion orientation,* dan *target detection.* Dalam dunia bisnis *opinion mining* banyak digunakan untuk menganalisis secara otomatis opini pelanggan tentang produk dan pelayanannya.

3.3. Algoritma Naïve Bayes

Naïve Bayes merupakan salah satu metode *classifier* yang memanfaatkan metode probabilitas dan statistik yang dikemukakan oleh ilmuan Inggris Thomas Bayes, yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya (Raschka, 2014). Dasar dari teorema ini adalah:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \quad \text{.....(3.1)}$$

P(A|B) = peluang A menghasilkan kondisi B
P(B|A) = peluang B yang muncul saat kondisi A
P(A) = peluang keseluruhan kemunculan A
P(B) = peluang keseluruhan kemunculan B

Peluang A terjadi saat kondisi B ditentukan dari peluang kondisi B saat menghasilkan hasil A, dikali peluang A, dibagi peluang B. Pada pengaplikasiannya nanti rumus ini berubah menjadi :

$$P(C_i|D) = \frac{P(D|C_i) \times P(C_i)}{P(D)} \quad \text{.....(3.2)}$$

$P(C_i|D)$ = peluang kelas C_i menghasilkan kondisi D
$P(D|C_i)$ = peluang kondisi D yang muncul di kelas C_i
$P(C_i)$ = peluang keseluruhan kemunculan kelas C_i
P(D) = peluang keseluruhan kemunculan kondisi D

Pada rumus diatas, hasil dari peluang kelas C_i menghasilkan kondisi D dapat dihitung dengan membagi hasil perkalian peluang D yang muncul di kelas C_i dan

peluang keseluruhan kemunculan kelas C_i dengan peluang keseluruhan kemunculan kondisi D.

3.4. Presisi & Recall

Presisi adalah sebuah ukuran seberapa sering hasil analisis sentimen benar. Presisi dihitung dari berapa kali hasil dinilai benar dibagi dengan berapa kali analisis dilakukan, *Recall* mengukur kelengkapan atau sensitivitas sebuah *classifier.* Semakin tinggi *recall* semakin rendah jumlah *false negative* atau data positif yang terdeteksi sebagai negatif. (Martin, 2014).

Perhitungan presisi positif dengan rumus:

$$P_p = \frac{TP}{TP+FP} \text{.....(3.3)}$$

Perhitungan presisi negatif dengan rumus:

$$P_n = \frac{TN}{TN+FN} \text{.....(3.4)}$$

Perhitungan recall positif dengan rumus:

$$R_p = \frac{TP}{TP+FN} \text{.....(3.5)}$$

Perhitungan recall negatif dengan rumus:

$$R_n = \frac{TN}{TN+FP} \text{.....(3.6)}$$

Dengan keterangan sebagai berikut:

TP= data true positive atau data teranalisis bersentimen positif dan benar

TN= data true negative atau data teranalisis bersentimen negatif dan benar

FP= data false positive atau data teranalisis bersentimen positif namun salah

FN= data false negative atau data teranalisis bersentimen negatif namun salah