# BAB II
# TINJAUAN PUSTAKA



## 2.1 *Data Mining*

*Data mining* adalah proses yang memperkerjakan satu atau lebih teknik pembelajaran computer (*machine learning*) untuk menganalisis dan mengekstraksi pengetahuan (*knowledge*) secara otomatis. *Data mining* adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual (Amalia, 2018).

*Data mining* adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam *database*. *Data mining* adalah proses yang menggunakan teknik statistic, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar (Amalia, 2018).

Berdasarkan definisi-definisi di atas tentang *data mining* dapat disimpulkan bahwa *data mining* adalah sebuah proses pencarian secara otomatis untuk menemukan pola atau model dari suatu *database* yang besar.

## 2.2 *Text Mining*

*Text mining* adalah proses penemuan akan informasi atau *trend* baru yang sebelumnya tidak terungkap dengan memproses dan menganalisa data dalam jumlah besar. Dalam menganalisa sebagian atau keseluruhan *unstructured text*, *text mining* mencoba untuk mengasosiasikan satu bagian *text* dengan yang lainnya berdasarkan aturan-aturan tertentu. Hasil yang diharapkan adalah informasi baru atau "*insight*" yang tidak terungkap jelas sebelumnya (Kurniasari, 2018).

*Text mining* adalah penggalian data untuk menyelesaikan masalah kebutuhan informasi dengan menerapkan teknik *data mining*, *machine learning*, *natural language processing*, pencarian informasi, dan manajeman pengetahuan. *Text mining* melibatkan praproses dokumen seperti kategorisasi teks, ekstraksi

Informasi, dan ekstraksi kata. Metode ini digunakan untuk mengekstraksi informasi dari sumber data melalui identifikasi dan eksplorasi pola yang menarik (Kurniawan, 2017).

*Text mining* merupakan teknologi yang digunakan untuk menganalisis data tak terstruktur data berbentuk teks. Dalam analisis *text mining* terdapat dua fase utama yaitu (1) *Preprocessing* dan integrasi dari data tak terstruktur, (2) analisis statistik data yang telah dilakukan *preprocessing* untuk mengekstraksi konten dari yang terdapat dalam teks menurut Fransis dan Flynn. Sedangkan menurut *Shollow Wiess* dalam bukunya menyatakan bahwa *text mining* merupakan transformasi dari data teks menjadi data numerik, dengan kata lain *text mining* mengubah data tak terstruktur menjadi data terstruktur (Fakhri, 2019).

*Text mining* merupakan proses menambang data yang berupa teks dimana sumber data biasanya didapatkan dari dokumen dan tujuannya adalah mencari kata-kata yang dapta mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen. Tujuan dari *text mining* adalah mengekstrak informasi yang berguna dari sumber data. Jadi, sumber data yang digunakan pada *text mining* adalah sekumpulan dokumen yang memiliki format yang tidak terstruktur melalui identifikasi dan eksplorasi pola yang menarik. Adapun tugas khusus *text mining* antara lain, pengkategorisasian teks (*text categotization*) dan pengelompokan teks (*text clustering*) (Putri, 2014).

*Text mining* adalah salah satu bidang khusus dalam *data mining* yang memiliki definisi menambang data berupa teks dimana sumber data biasanya didapatkan dari dokumen dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen (Dewi, 2018).

*Text mining* dapat menganalisa dokumen, mengelompokkan dokumen berdasarkan kata-kata yang terkandung di dalamnya, serta menentukan kesamaan di antara dokumen untuk mengetahui bagaimana mereka berhubungan dengan variabel lainnya. Penerapan yang paling umum pada penggunaan *text mining* misalnya penyaringan spam, analisa sentiment, mengukur preferensi pelanggan, meringkas dokumen, pengelompokan topic penelitian, dan banyak lainnya (Dewi, 2018).

Berdasarkan definisi-definisi di atas tentang *text mining* dapat disimpulkan bahwa *text mining* adalah suatu proses pencarian informasi berupa teks data untuk memperoleh informasi yang berguna pada suatu data.



## 2.3 Klasifikasi

Klasifikasi teks merupakan kegiatan untuk mengelompokkan sebuah dokumen ke dalam sebuah kelas dokumen. Seiring berkembangnya jumlah dokumen, diperlukan sebuah tools untuk melakukan klasifikasi secara otomatis. Proses klasifikasi secara otomatis dapat dilakukan oleh komputer.

Berdasarkan jumlah kelas terdapat 2 tipe klasifikasi, yaitu *binary classification* dan *multi-class classification*. *Binary classification* merupakan klasifikasi sebuah obyek ke salah satu kelas dari dua kelas yang ditentukan. Sedangkan *multi-class classification* adalah klasifikasi sebuah obyek ke satu atau lebih kelas.

Dalam klasifikasi teks, obyek data dibagi menjadi 2, yaitu data latih dan data uji. Data latih adalah data dokumen yang sudah diklasifikasikan secara manual, sedangkan data uji adalah data dokumen yang belum diklasifikasikan dan akan digunakan sebagai data pengujian. Dalam penelitian ini menggunakan klasifikasi sederhana atau tipe klasifikasi *binary classification* karena kategori kelas hanya ada 2 kelas yaitu kelas positif dan kelas negatif.

## 2.4 Pembobotan Kata

TF-IDF (*Term Frequency & Inverse Document Frequency*) merupakan metode pembobotan secara statistik. Metode TF-IDF menunjukkan seberapa penting sebuah kata pada sebuah dokumen yang terletak pada sebuah kelompok. Metode pembobotan TF-IDF biasanya digunakan dalam *text mining*. *Term frequency* (TF) adalah jumlah sebuah kata pada sebuah dokumen sedangkan *inverse document frequency* atau IDF adalah nilai yang digunakan untuk mengukur seberapa penting sebuah kata pada koleksi dokumen (Sebastian,2019). Menurut (Dinata, Pande Made R. C., dan Rakhmawati Nur Aini, 2020) persamaan IDF dapat dilihat pada persamaan (2.1).

$$idf_{t,d} = ln\left(\frac{1+N}{1+df}\right) + 1 \quad (2.1)$$



Dimana:

$N$ = jumlah koleksi dokumen

$df$ = jumlah dokumen dimana terdapat *term*$(t)$

Jadi pada metode pembobotan TF-IDF, perhitungan bobot *term* dalam sebuah dokumen dilakukan dengan mengalikan nilai TF dengan nilai IDF. Persamaan (2.2) merupakan perhitungan bobot *term* ($W$):

$$W_{t,d} = tf_{t,d} \times idf_{t,d} \quad (2.2)$$

dimana:

$W_{t,d}$ = bobot *term* $t$ terhadap dokumen $d$

$tf_{t,d}$ = frekuensi kemunculan *term* $t$ pada dokumen $d$

$idf_{t,d}$ = nilai $idf$ dari *term* $t$

## 2.5 Multinomial Naïve Bayes (MNB)

MNB merupakan suatu metode yang mengambil jumlah kata yang muncul dalam setiap dokumen disuatu kelas $C$, dengan mengasumsikan dokumen memiliki kejadian dalam kata tidak bergantung dari kelasnya dokumen. Probabilitas sebuah dokumen $d$ berada di kelas $C$, kondisi dapat dinyatakan sebagai berikut (Harjitol dkk,2018).

$$P(C|term\ dokumen\ d) = P(t_1|C) \times P(t_2|C) \times \ldots \times P(t_k|C) \times P(C) \quad (2.3)$$

$P(C)$ : probabilitas prior dari kelas $C$

$P(C|term\ dokumen\ d)$ : probabilitas sebuah dokumen $d$ berada di kelas $C$

$P(t_k|C)$ : probabilitas kata ke-$n$ dengan diketahui kelas $C$

Probabilitas prior kelas $C$ ditentukan dengan rumus:

$$P(C) = \frac{N_C}{N} \quad (2.4)$$

$N_C$ : jumlah kelas $C$ pada seluruh dokumen,

$N$ : jumlah seluruh dokumen.

Probabilitas kata ke-$n$ ditentukan dengan menggunakan persamaan sebagai berikut:

$$P(t_k|C) = \frac{T_{ct}}{\sum_{t \in V} T_{ct}} \quad , k = (1, \dots, k) \tag{2.5}$$



$T_{ct}$ : jumlah kemunculan *term* $t$ pada dokumen dengan kelas $C$

$\sum_{t \in V} T_{ct}$ : jumlah *term* di seluruh data *training* dengan kelas $C$

Untuk menghindari nilai probabilitas masing-masing kata bernilai nol, digunakan *laplace smoothing* atau *add-one* yaitu proses penambahan nilai 1 pada setiap nilai $T_{ct}$ pada perhitungan *conditional* probabilitas sebagai berikut:

$$P(t_k|C) = \frac{T_{ct}+1}{(\sum_{t \in V} T_{ct})+B} \tag{2.6}$$

dimana:

$B$ : jumlah seluruh kata pada data training

Sehingga untuk rumus Multinomial Naïve Bayes yang digunakan dengan pembobotan kata TF-IDF adalah sebagai berikut:

$$P(t_k|C) = \frac{W_{ct}+1}{(\sum_{t \in V} W_{ct})+B} \tag{2.7}$$

dimana:

$W_{ct}$ :bobot kata TF-IDF pada dokumen dengan kelas $C$

$\sum_{t \in V} W_{ct}$ : jumlah bobot kata TF-IDF seluruh kata pada dokumen dengan kelas $C$

Untuk menentukan kelas terbaik suatu dokumen dalam klasifikasi multinomial naïve bayes ditentukan dengan mencari *maximum a posterior* (MAP) kelas $C$ sebagai berikut (sabrani dkk,2020).

$$C_{map} = argmax\ P(C) \prod_{k=1}^{k} P(t_k|C) \tag{2.8}$$

## 2.6 *E-marketplace*

Dunia maya yang tercipta karena berkembangnya teknologi internet, secara tidak langsung membentuk sebuah pasar atau arena perdagangan tersendiri yang kerap dinamakan sebagai *e-marketplace*. Sebagaimana pasar dalam pengertian konvensional yaitu tempat bertemunya penjual dan pembeli, didalam *e-marketplace* berinteraksi pula berbagai perusahaan-perusahaan di dunia tanpa dibatasi oleh teritori ruang (geografis) maupun waktu. Beragam produk dan jasa dalam berbagai bentuknya dicoba ditawarkan oleh perusahaan-perusahaan yang telah "go internet" ini dalam berbagai domain industri, sehingga menghasilkan

suatu nilai dan volume perdagangan yang tidak kalah besar dari pasar Konvensional (Kodong, 2012).



*E-marketplace* adalah sebuah sistem infomasi antar organisasi dimana pembeli dan penjual di pasar mengkomunikasikan informasi tentang harga, produk dan mampu menyelesaikan transaksi melalui saluran komunikasi elektronik. Suatu *e-marketplace* merepresentasikan suatu struktur sosial, konsep ekonomi pasar, dan penggunaan teknologi. *E-marketplace* dapat memberikan peluang untuk melakukan bisnis dan melaksanakan transaksi melalui saluran elektronik, biasanya pada platform yang berbasiskan internet (Marco, 2017).

## 2.7 *Text Preprocessing*

Struktur data yang baik dapat memudahkan proses komputerisasi secara otomatis. Pada *text mining*, informasi yang akan digali berisi informasi-informasi yang strukturnya sembarang. Oleh karena itu, diperlukan proses pengubahan bentuk menjadi data yang terstruktur sesuai kebutuhannya untuk proses dalam data mining, yang biasanya akan menjadi nilai-nilai numerik. Proses ini sering disebut *Text Preprocessing*. Setelah data menjadi data terstruktur dan berupa nilai numerik maka data dapat dijadikan sebagai sumber data yang dapat diolah lebih lanjut.

## 2.8 Penelitian Terdahulu

Berikut ini adalah rangkuman hasil penelitian terdahulu yang berkaitan dengan penelitian yang dilakukan.

**Tabel 2.1** Penelitian Terdahulu

| No. | Nama dan Tahun Publikasi | Hasil |
|---|---|---|
| 1 | Purwanto, Devi Dwi dan Santoso, Joan (2015) | Pada penelitian ini menentukan review positif satau negative pelanggan website penjualan menggunakan multinomial naïve bayes. Hasil uji coba memiliki keakurasian sebesar 85,6%. |

| | | |
|---|---|---|
| 2 | Harjitol, Bambang, Aini, Kuni Nur, dan Murtiyasa, Budi (2018) | Penelitian ini mengkategorikan dokumen bahasa Inggris yang terkait dengan serangan jaringan menggunakan Multinomial Naïve Bayes dan Term frequency-Inverse Document Frequency (TF-IDF). Hasil percobaan menunjukkan bahwa MNB dengan TF-IDF mendapatkan akurasi 76,00%. |
| 3 | Yulianto, Alfian, Herdiani, Anisa, dan Sardi Indra Lukmana (2019) | Penelitian ini mengklasifikasikan keberpihakan *tweet* pada pemilihan presiden 2019 dengan menggunakan Multinomial Naïve Bayes. Hasil pengujian menunjukkan bahwa MNB mendapatkan akurasi 72%. |
| 4 | Rahman, Umar S. A., Yudi, Wibisono dan Nugroho, Eddy Prasetyo (2020) | Penelitian dilakukan menggunakan Multinomial Naïve Bayes untuk melakukan klasifikasi ujaran kebencian pada dataset kicauan (twitter) Bahasa Indonesia. Hasil yang diperoleh dengan rasio data testing 10% memiliki hasil akurasi sebesar 76,6%. |
| 5 | Sabrani, Alif., dkk (2020) | Dalam penelitian ini pengujian dilakukan dengan multinomial naïve bayes dalam mengelompokkan artikel online tentang gempa bumi di Indonesia. Pembobotan kata juga dilakukan menggunakan TF-IDF dengan hasil uji coba diperoleh keakurasian sebesar 95.20%. |