BAB III

LANDASAN TEORI

3.1. Data mining

Kamagi & Hansun (2014) mendefinisikan *data mining* sebagai suatu proses untuk menemukan hubungan, pola, dan *trend* baru yang bermakna dengan menyaring data yang sangat besar, yang tersimpan dalam penyimpanan, menggunakan teknik pengelolaan pola, seperti teknik statistik dan matematika. *Data mining* bertujuan untuk mengatasi jumlah data yang besar, dimensi data yang tinggi, data yang heterogen dan berbeda sifat. Mardi (2014) juga memberikan definisi tentang *data mining* yaitu proses mencari pola atau informasi menarik dalam data yang terpilih dengan menggunakan teknik atau metode tertentu.

Data mining juga merupakan penerapan suatu algoritma untuk melakukan ekstraksi pola dari suatu data sehingga dapat diketahui pengetahuan yang tersembunyi dari sebuah data (Saputra, Adji, & Permanasari, 2015). *Data mining* telah banyak digunakan untuk berbagai bidang seperti bisnis, ilmuwan dan pemerintahan untuk menyaring *volume* data yang begitu besar. Contoh penyaringan informasi seperti perjalanan penumpang pesawat, data populasi dan data pemasaran untuk menghasilkan laporan riset pasar (Li,2010). Algoritma atau metode yang dipilih berdasarkan proses *Knowledge Discovery in Database (KDD).*

Menurut Mabrur & Lubis (2012) proses *Knowledge Discovery in Database (KDD)* secara garis besar yaitu :

1. ***Data Selection***

   Data selection digunakan untuk memilih data dari kumpulan data operasional yang dilakukan sebelum menggali informasi.

2. ***Pre-processing* atau *cleaning***

   Pre-pocessing atau *cleaning* dilakukan untuk membuang data yang berulang-ulang, memeriksa data yang tidak konsistensi dan memperbaiki data yang salah apabila terjadi salah pengetikan.

   Tahap-tahap dalam *preprocessing* adalah :

   a. ***Tokenization***

      Tokenization merupakan bagian pemotongan urutan karakter dan sebuah set dokumen dimana pemotongan tersebut dibuat menjadi kata atau karakter yang sesuai dengan kebutuhan sistem (Indranandita et al., 2008). Proses ini juga merupakan proses pembersihan karakter tertentu seperti tanda baca dengan melakukan penghilangan tanda baca serta merubah huruf menjadi huruf kecil (Rizqiyani, Mulwinda, & Putri, 2017) .

      Contoh proses *tokenization* :

      Input : aku suka dengan sinetron Indonesia.

Output :

aku	suka	dengan	sinetron	indonesia

b. Steeming

Steeming merupakan suatu proses mengubah token yang berimbuhan menjadi kata dasar dengan menghilangkan semua imbuhan yang ada pada token tersebut (Ling, Kencana, & Oka, 2014).

Sebagai contoh, kata berhasil, keberhasilan akan diubah secara otomatis menjadi hasil.

c. Stopword removal

Stopword removal merupakan proses dimana melakukan pembuangan terhadap kata-kata yang tidak berpengaruh pada proses klasifikasi (Gusriani et al., 2016). Kata-kata yang tidak berpengaruh di dalam proses klasifikasi tersebut misalnya yang, dari, di, dan kata penghubung lainnya.

3. Transformation

Transformation dilakukan untuk pengkodean pada data yang telah dipilih sehingga data sesuai dengan *data mining*.

4. Data mining

Data mining adalah proses mencari informasi yang menarik pada data-data yang terpilih dengan menggunakan salah satu metode yang dianggap cocok.

5. Evaluation

Evaluation adalah tahap dimana dilakukan proses pemeriksaan informasi yang telah ditemukan apakah bertentangan dengan fakta yang telah ada sebelumnya.

3.2. Analisis Sentimen

Gusriani, Wardhani, & Zul (2016) mengatakan bahwa analisis sentimen adalah bidang ilmu yang menganalisa pendapat, sentimen, evaluasi, penilaian, sikap dan emosi publik terhadap entitas seperti produk, jasa, organisasi, individu, masalah, peristiwa, topik dan atribut mereka. Selain itu Zuhri & Alamsyah (2017) mengatakan bahwa analisis sentimen merupakan bidang studi dari *data mining* yang digunakan untuk menganalisis, memahami, mengolah dan mengekstrak data tekstual yang berupa opini, sentimen dan lain sebagainya. Analisis sentimen bertujuan untuk mengelompokkan atau mengklasifikasikan data ke dalam sentimen positif, negatif dan netral tentang polaritas teks.

Berdasarkan beberapa pendapat diatas dapat diketahui bahwa analisis sentimen pada umumnya adalah sebuah bidang studi dari *data mining* yang dapat digunakan untuk menganalisis pendapat, opini atau sentimen seseorang terhadap sesuatu hal tertentu dimana berupa opini positif, negatif maupun netral.

3.3. Metode Klasifikasi Naïve Bayes

Menurut Indranandita, Susanto, & C (2008) metode Naïve Bayes adalah salah satu metode klasifikasi dengan menggunakan teori probabilitas sebagai dasar teori. Metode Naïve Bayes adalah sebuah metode klasifikasi dengan probabilitas sederhana yang menggunakan teorema Bayes dengan asumsi ketergantungan yang tinggi (Nurhuda et al., 2014).

Selain itu metode Naïve Bayes juga didefinisikan sebagai sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi dari dataset yang diberikan dan mengasumsikan semua atribut independen dan tidak saling bergantung dengan nilai yang diberikan oleh variabel kelas lain. Naïve Bayes dikatakan menguntungkan karena hanya membutuhkan jumlah data pelatihan yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian dan sering bekerja jauh lebih baik dari yang diharapkan (Saleh, 2015).

Theorema bayes :

$$P(Ci|X) = \frac{P\ X\ Ci\ *\ P(Ci)}{P(X)}$$

Persamaan diatas yaitu P(Ci|X) merupakan probabilitas Ci terjadi jika X sudah terjadi, P(Ci)adalah probabilitas Ci dalam data dengan sifat *independent* terhadap X, X adalah kumpulan atribut, P(X|Ci) adalah probabilitas X terjadi jika Ci sudah terjadi berdasarkan data training. Untuk dapat mengklasifikasikan sebuah tweet, dalam penelitian ini penulis menggunakan metode klasifikasi Naive Bayes untuk klasifikasi teks, seperti berikut ini :

$$P(v1|C=c) = \frac{CountTerms(v1,docsv(c))}{AllTerms(docs(c))}$$

Dimana v1 merupakan suatu kata tertentu dalam *tweet*, sedangkan *CountTerms(v1,docsv(c))* menunjuk pada jumlah kemunculan suatu kata berlabel c (positif, negatif atau netral). *AllTerms(docs(c))* menunjuk pada jumlah semua kata berlabel c yang ada pada dataset. Untuk menghindari adanya nilai 0 pada probabilitas, maka diberlakukan Laplace(*add-one*) *smoothing*. Tujuannya adalah untuk mengurangi probabilitas dari hasil keluaran yang terobservasi dan juga sekaligus meningkatkan atau menambah probabilitas hasil/keluaran yang belum

terobservasi, sehingga persamaan menjadi sebagai berikut :

$$P(v1|C=c) = \frac{CountTerms(v1,docsv(c))+1}{AllTerms(docs(c))+|V|}$$

Dimana |V| menunjuk pada jumlah semua kata dalam *tweet* yang ada pada dataset.

3.4. Twitter

Twitter adalah sebuah media sosial yang diciptakan oleh Jack Dorsey pada bulan Juli 2006 di bawah perusahaan Odeo Corp dimana media sosial ini dikatakan tampak lebih mudah penggunaannya daripada media sosial lainnya. Twitter adalah sebuah situs *web* yang dioperasikan oleh Twitter inc, yang memungkinkan pengguna untuk mengirim dan membaca pesan yang biasa disebut kicauan atau *tweets* (Marpaung, 2017). *Tweet*s bisa dilihat oleh orang-orang tertentu sesuai dengan keinginan dari pemilik *tweet*. Pengguna juga dapat melihat *tweet* atau kicauan dari pengguna lain ketika sudah mengikuti atau biasa disebut dengan *follow*.

Habibi, Setyohadi, & Ernawati (2016) mengatakan bahwa twitter merupakan sebuah layanan *microblogging* yang memposting sesuatu yang pendek *(tweet)* melalui *website* atau *mobile* dengan panjang maksimal 140 karakter. Pesan dari twitter memiliki atribut yang unik, yang membedakan dengan sosial media lainnya yaitu panjang karakter maksimal 140

karakter, memiliki Twitter API yang dapat mengakses data twitter secara gratis sehingga mempermudah dalam proses pengumpulan data *tweets* dalam jumlah yang banyak, memiliki banyak model bahasa yang berbeda, dan pengguna twitter dapat mengirim pesan singkat dengan berbagai topik secara global.

3.5. Sinetron

Takariani (2013) mengatakan sinetron merupakan salah satu acara di televisi yang pertama kali diperkenalkan oleh seorang pendiri Institut Kesenian Jakarta atau dikenal dengan IKJ yaitu Soemardjono. Sinetron disebut sebagai soap opera dalam bahasa Inggris atau dalam bahasa Indonesia opera sabun. Istilah opera sabun muncul ketika drama serial terkenal di dunia pertelevisian Amerika Serikat. Drama-drama serial tersebut membuat banyak perusahaan sabun memasang iklan pada televisi.

Sinetron yang muncul pertama kali di Indonesia adalah sinetron dengan judul Losmen yang bernaung di bawah stasiun TVRI pada tahun 1980-an. Pada saat ini sinetron terus mengalami perkembangan dengan berbagai cerita yang menarik untuk ditonton. Dengan adanya sinetron juga menimbulkan banyak stasiun televisi swasta yang bermunculan dan membuat banyak sinetron sesuai dengan cerita masing-masing (Takariani, 2013).