BAB II
LANDASAN TEORI

2.1 Sentimen Analysis

Sentiment analysis (Analisis Sentimen) mengacu pada bidang yang luas dari pengolahan Bahasa alami , komputasi *linguistic* dan *text mining*. Analisis sentiment bertujuan menganalisis pendapat, sentimen, evaluasi, sikap, penilaian dan emosi seseorang apakah pembicara atau penulis berkenan dengan suatu topik , produk , layanan , organisasi, individu ataupun kegiatan tertentu (Liu, 2012). Ditambahkan bahwa analisis sentimen adalah alat untuk memproses koleksi hasil pencarian yang bertujuan dengan mencari atribut suatu produk dan proses memperolah hasil pendapatnya. Tugas dasar dalam analisis sentimen adalah mengelompokkan polaritas dari teks yang ada dalam dokumen, apakah pendapat yang dikemukakan dalam dokumen bersifat positif , negatif atau netral. Analisis sentimen difokuskan untuk me-review klasifikasi berdasarkan polaritas. Berdasarkan klasifikasi, analisis sentimen dibagi menjadi dua kelompok utama. Yaitu dokumen klasifikasi ke pendapat atau fakta, atau dikenal sebagai klasifikasi subjektivitas *(subjectivity classification)* dan dokumen klasifikasi ke dalam positif atau negatif, atau dikenal sebagai analisis sentimen. Hal ini adalah proses yang penting untuk menentukan dokumen yang memiliki opini dan dokumen yang menyimpulkan opini bernilai positif, negatif maupun netral [5].

2.2 Text Mining

Text Mining adalah salah satu bidang khusus dari *data mining*. Text Mining didefinisikan sebagai proses menggali informasi dimana seorang user berinteraksi dengan sekumpulan dokumen menggunakan tools analisis yang merupakan komponen - komponen dalam *data mining* yang salah satunya adalah kategorasi (Fedman dan Sanger, 2007). Tujuan dari text mining adalah untuk mendapatkan informasi implisit dari sekumpulan dokumen sehingga bisa digunakan oleh user untuk mengambil keputusan. Sumber data yang digunakan pada text mining adalah kumpulan text yang memiliki format yang tidak terstruktur maupun semi terstruktur. Pada dasarnya proses kerja dari *text mining* banyak mengadopsi dari penelitian Data Mining namun yang menjadi perbedaan adalah pola yang digunakan oleh text mining diambil dari sekumpulan bahasa alami yang tidak

terstruktur sedangkan dalam *Data Mining* pola yang diambil dari database yang terstruktur (Han & Kamber, 2006)[5].

2.3 Twitter

Twitter adalah sebuah situs web yang dimiliki dan dioperasikan oleh Twitter Inc., yang menawarkan jaringan sosial berupa *mikroblog* sehingga memungkinkan penggunanya untuk mengirim dan membaca pesan *Tweets*. *Microblog* adalah salah satu jenis alat komunikasi online dimana pengguna dapat memperbarui status tentang mereka yang sedang memikirkan dan melakukan sesuatu, apa pendapat mereka tentang suatu objek atau fenomena tertentu [23].

Micro blogging merupakan sumber informasi baru yang menarik untuk pengolahan data, *twitter* merupakan sebuah jaringan sosial yang memberikan layanan *micro blogging* per pesan yang biasa disebut *Tweet*, dikirim oleh pengguna kepada pembacanya yang disebut *follower*. *Tweets* adalah teks tulisan hingga 140 karakter yang ditampilkan pada halaman profil pengguna dan dikirim ke pembacanya yang disebut *follower*. *Twitter* dapat diakses melalui web, pesan singkat (SMS), atau *third party applications* seperti *Ubersocial*.

Twitter menjadi alat yang menarik bagi banyak kalangan untuk mengikuti keinginan para pengguna terhadap setiap kondisi secara langsung. Hal ini menjadi sumber data yang potensial untuk dipakai oleh jutaan orang. *Twitter* membuat segalanya tersedia dalam sebuah data yang bisa dimanfaatkan dengan memakai teknik *Crawling, streaming mining*. Secara prinsip hal ini dapat mengetahui opini public secara umum.

Fitur yang terdapat dalam twitter, antara lain:

1. Laman Utama (*Home*)

   Pada halaman utama kita bisa melihat *Tweets* yang dikirimkan oleh orang-orang yang menjadi teman atau yang di ikuti (*following*).

2. Profil (*Profile*)

   Halaman yang berisikan data diri dari pengguna *Twitter*.

3. *Followers*

   Followers adalah pengguna lain yang ingin menjadikan sebagai teman. Bila pengguna lain menjadi pengikut akun seseorang, maka *Tweets* seseorang yang di ikuti tersebut akan masuk ke dalam halaman utama.

4. *Following*

   Following adalah akun seseorang yang mengikuti pengguna lain agar *Tweets* yang dikirim oleh orang yang diikuti tersebut masuk ke dalam halaman utama.

5. *Mentions*

   Konten ini merupakan balasan dari percakapan, supaya sesama pengguna bisa langsung menandai orang yang diajak bicara.

Twitter menyediakan *Application Programming Interface* (*API*) untuk memfasilitasi data *Crawling*. *API* memudahkan pengguna untuk mengambil data *Tweet* secara *real time* [24].

2.4 *Twitter Api*

Twitter API (*Application Programming Interface*) merupakan sejumlah fungsi yang dapat digunakan pengembang perangkat lunak untuk mengolah data saat membangun perangkat lunak. *Twitter API* menyediakan beberapa fungsi untuk melakukan suatu tugas tertentu, sehingga pengembang perangkat lunak hanya memanggil fungsi tersebut di dalam perangkat lunak yang dibangun. Twitter *API* menggunakan arsitektur *REST* (*Representational State Transfer*) sehingga *Twitter API* dapat digunakan pada format data yang beragam seperti *XML* atau *JSON*. *Twitter API* terdiri atas *Twitter Search API* dan *Twitter Streaming API*. Perbedaan keduanya yaitu, *Twitter Search API* menitik beratkan fungsi pencarian ke masa lampau sedangkan *Twitter Streaming API* menitik beratkan fungsi pencarian ke masa yang akan datang [10].

Tujuan awal dibentuknya *Twitter API* ini adalah untuk mengetahui relasi dan interaksi antara pengguna, namun sebaliknya *Twitter API* banyak digunakan untuk menggali informasi komunitas tertentu atas pandangannya terhadap topik yang sedang *trending* [24]. Jenis-jenis Twitter *API* terbagi menjadi :

1. *Twitter Streaming API*

   Merupakan *API* yang digunakan untuk mendapatkan data *Tweets* secara *real-time* sesuai dengan *keyword* yang digunakan.

2. *Twitter REST API*

   Merupakan *API* yang ditujukan untuk proses *read and write* data *Twitter*, melakukan pengiriman *Tweets*, membaca profil pengguna dan data pengikut. REST *API* menggunakan OAuth dalam pengenalan terhadap aplikasi dan tersedia dalam format JSON.

Dalam terminologi *twitter*, setiap pesan mengGambarkan status seorang pengguna. Berdasarkan *API streaming* para pengguna dapat mengakses status publik hampir secara *realtime*, termasuk *reply* dan *mention* yang dibuat oleh akun-akun publik, status yang dibuat oleh akun-akun terproteksi juga pesan yang tidak dapat di akses.

API membutuhkan akun *Twitter* yang valid agar dapat digunakan. Penelitian ini juga melakukan cara yang sama dengan memanfaatkan *Twitter API*, kemudian dijalankan pada program untuk mengambil informasi mengenai pemilihan presiden 2019 (Pilpres 2019) [25].

2.5 *R Studio*

R adalah suatu bahasa komputer dan merupakan lingkungan pemograman interaktif untuk analisis data dan grafik. Bahasa R memungkinkan kita untuk menghitung, melihat data dan program secara interaktif dengan umpan balik yang cepat sehingga memungkinkan untuk belajar dan memahami tentang data. *R* dapat digunakan pada berbagai bidang seperti analisis keuangan, penelitian statistika, manajemen, akademis, matematika, grafik dan analisis data. *R* merupakan bahasa fungsi. Setiap perintah diinterpretasikan sebagai evaluasi fungsi. Komputasi numerik, grafik atau lainnya pada *R* dilakukan oleh fungsi dengan cara memanggil fungsi tersebut dengan memberikan nama fungsi diikuti atau tanpa diikuti oleh argumen-argumen di dalam tanda kurung ( ). *R* menyediakan banyak fungsi-fungsi yang bisa digunakan. *R* juga dimungkinkan untuk mendefinisikan fungsi-fungsi baru yang tidak disediakan *R* sesuai dengan fungsi.

2.6 Data Crawling

Web crawler atau yang dikenal juga dengan istilah *web spider* atau *web robot* adalah program yang bekerja dengan metode tertentu dan secara otomatis mengumpulkan semua informasi yang ada dalam suatu situs *web*. *Web crawler* akan mengunjungi setiap alamat situs *web* yang diberikan kepadanya, kemudian menyerap dan menyimpan semua informasi yang terkandung di dalam situs tersebut. Setiap kali *web crawler* mengunjungi sebuah situs web, maka dia juga akan mendata semua situs yang ada di halaman yang dikunjunginya secara bergiliran. Ketika *crawlers* menemukan halaman *web*, tugas selanjutnya adalah mengambil data-data dari halaman *web* dan menyimpannya ke dalam suatu media penyimpanan (*hard disk*). Data-data yang disimpan ini, nantinya dapat diakses pada saat dilakukan *query* yang berhubungan dengan data tersebut.

Proses *web crawler* dalam mengunjungi setiap dokumen *web* disebut dengan *web crawling* atau *spidering*. Proses *crawling* dalam suatu dimulai dari mendata seluruh situs dari halaman *web*, menelusurinya satu-persatu, kemudian memasukkannya dalam daftar halaman pada indeks.

Web crawler biasa digunakan untuk membuat salinan secara sebagian atau keseluruhan halaman *web* yang telah dikunjunginya agar dapat diproses lebih lanjut oleh sistem penyusun indeks. *Crawler* dapat juga digunakan untuk proses pemeliharaan sebuah situs *web*, seperti memvalidasi kode html sebuah *web*, dan *crawler* juga digunakan untuk memperoleh data yang khusus seperti mengumpulkan alamat surel[6].

2.7 Data Preprocessing

Tahap *pre-processing* atau pra proses data merupakan proses untuk mempersiapkan data mentah sebelum dilakukan proses lain. Pada umumnya, pra proses data dilakukan dengan cara mengeliminasi data yang tidak sesuai atau mengubah data menjadi bentuk yang lebih mudah diproses oleh sistem. Pra proses sangat penting dalam melakukan analisis sentimen, terutama untuk media sosial yang sebagian besar berisi kata-kata atau kalimat yang tidak formal dan tidak terstruktur serta memiliki *noise* yang besar. Untuk mempelajari suatu data teks, terlebih dahulu ditentukan fitur – fitur yang mewakili setiap kata untuk setiap fitur

yang ada pada dokumen. Sebelum menentukan fitur tersebut diperlukan tahap preprocessing yang dilakukan antara lain:

1. *Cleansing*

   Tahapan ini bertujuan untuk mengeliminasi kata atau simbol yang tidak diperlukan untuk mengurangi *noise* saat proses klasifikasi.

2. *Case Folding*

   Tahapan ini bertujuan untuk mengubah mengubah seluruh huruf kapital menjadi huruf kecil. Tidak semua dokumen konsisten dengan penggunaan huruf kapital, sehingga *case folding* diperlukan untuk mengonversi keseluruhan teks dalam dokumen.

3. *Tokenization*

   Tahapan *tokenization* merupakan tahapan untuk memecah kalimat menjadi kata atau yang biasa disebut *token*. Tahap ini bertujuan untuk menghilangkan karakter yang dianggap sebagai tanda baca.

4. *Remove Duplicate*

   Tahapan ini bertujuan untuk menghilangkan kalimat yang *redundant* atau berulang. Proses ini menghilangkan kalimat yang sama persis dengan kalimat sebelumnya.

5. Eliminasi Stopword

   Kata-kata *stopwords* dapat menambah dimensidata pada proses Klasifikasi. Kamus data *stopwords* yang secara umum digunakan (terdiri dari yang,di,ke,dari,dll) akan ditambahkan dengan stopwords khusus *twitter*, seperti “wkkwk”,”hihihi”,”xoxoxox”, dsb. Kamus data *stopwords twitter* dan *stopwords* Bahasa Indonesia selengkapnya terdapat di dalam lampiran. Khusus untuk kamus data *stopwords twitter* Bahasa Indonesia dikumpulkan secara manual dari twitter.

2.8 Naive Bayes Classifier

Penelitian ini menggunakan Naïve bayes karena dalam proses klasifikisi dalam perhitungan probabilistik, naïve bayes memiliki lebih banyak keuntungan. Salah satu keuntungannya ialah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class. Naive Bayes didasarkan pada teorema Bayes yang memiliki kemampuan klasifikasi serupa dengan decision

tree dan neural network. Naive Bayes terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam database dengan data yang besar. Selain itu berikut kelebihan yang terdapat pada naïve bayes secara keseluruhan, yaitu:

a. Menangani kuantitatif dan data diskrit
b. Kokoh untuk titik noise yang diisolasi, misalkan titik yang dirata – ratakan ketika mengestimasi peluang bersyarat data.
c. Hanya memerlukan sejumlah kecil data pelatihan untuk mengestimasi parameter (rata – rata dan variasi dari variabel) yang dibutuhkan untuk klasifikasi
d. Menangani nilai yang hilang dengan mengabaikan instansi selama perhitungan estimasi peluang.
e. Cepat dan efisiensi ruang.
f. Kokoh terhadap atribut yang tidak relevan.

Kaitan antara Naive Bayes dengan klasifikasi, korelasi hipotesis dan bukti klasifikasi adalah bahwa hipotesis dalam teorema Bayes merupakan label kelas yang menjadi terget pemetaan dalam klasifikasi, sedangkan bukti merupakan fitur-fitur yang memjadi masukkan dalam model klasifikas. Jika X adalah vektor masukkan yang berisi fitur dan Y adalah label kelas, *Naive Bayes* dituliskan dengan P(X|Y). Notasi tersebut berarti probabolitas label kelas Y didapatkan setelah fitur-fitur X diamati. Notasi ini disebut juga probabilitas akhir (*posterior probability*) untuk Y, sedangkan P(Y) disebut probabilitas awal (*prior probability*) Y.

Selama proses pelatihan harus dilakukan pembelajaran probabilitas akhir P(Y|X) pada model untuk setiap kombinasi X dan Y berdasarkan informasi yang didapat dari data latih. Dengan membangun model tersebut, suatu data uji X' dapat diklasifikasikan dengan mencari Y' dengan memaksimalkan nilai P(X'|Y') yang didapat.

Formulasi Naive Bayes untuk klasifikasi adalah :

$$P(Y, X) = \frac{P(y)\Pi^{q}_{i=1} P(Xi|Y)}{P(x)}$$

P(Y|X) adalah probabilitas data dengan vektor X pada kelas Y. P(Y) adalah probabilitas awal kelas Y. $P(y)\Pi^{q} P(Xi|Y)$ adalah probabilitas

independen kelas Y dari semua fitur dalam vektor X. Nilai P(X) selalu tetap sehingga dalam perhitungan prediksi nantinya kita tinggal menghitung bagian $P(y)\Pi^{q}\ P(Xi|Y)$ dengan memilih yang terbesar sebagai kelas yang dipilih sebagai hasil prediksi. Sementara probabilitas independent

$\Pi^{q}\ P(Xi|Y)$ tersebut merupakan pengaruh semua fitur dari data terhadap setiap kelas Y, yang dinotasikan dengan

$$P(Y|X) = \frac{\Pi^{q}\ P(Xi|Y = y)}{i=1}$$

Setiap set fitur X= {X1, X2, X3, ..., Xq} terdiri atas q atribut (q dimensi).

2.9 Hidden Markov Model

Sebuah HMM menggabungkan dua atau lebih rantai Markov dengan hanya satu rantai yang terdiri dari state yang dapat diobservasi dan rantai lainnya membentuk state yang tidak dapat diobservasi *(hidden)*, yang mempengaruhi hasil dari state yang dapat diobservasi. Probabilitas dari satu state ke state lainnya dinamakan *transition probability*. Setiap state mungkin dibentuk oleh sejumlah elemen atau simbol. Untuk sequence, terdapat dua puluh buah simbol. Nilai probabilitas yang berasosiasi dengan setiap simbol dalam setiap state disebut emission probability. Untuk menghitung probabilitas total dari suatu jalur dalam model, baik transition probability maupun *emission probability* yang menghubungkan semua hidden state dan state yang dapat diobservasi harus dimasukkan dalam perhitungan.

Sebuah Hidden Markov Model dikarakteristikkan dengan parameter berikut:

a. N, jumlah state dalam model.
b. M, jumlah simbol pengamatan yang dimiliki setiap state.
c. A = {aij}, aij = P(qt+1 = Sj | qt = Si), himpunan distribusi kemungkinan perpindahan state (transition probability) .
d. B = {bj(k)}, bj(k) = P(vk pada t | qt = Sj), himpunan distribusi kemungkinan simbol pengamatan pada state j (emission probability).
e. π = {πi}, πi = P(q1 = Si), himpunan distribusi kemungkinan state awal. Bentuk ringkas dari HMM adalah λ = (A, B, π) [9].

Analisis rantai Markov adalah suatu metode yang mempelajari sifat-sifat suatu peubah pada masa sekarang yang didasarkan pada sifat-sifatnya di masa lalu dalam usaha menaksir sifat- sifat peubah tersebut dimasa yang akan datang. Model rantai Markov dikembangkan oleh seorang ahli Rusia A.A. Markov pada tahun 1896. Pada analisis Markov yang dihasilkan adalah suatu informasi probabilistik yang dapat digunakan untuk membantu pembuatan keputusan. Konsep dasar analisis markov adalah state dari sistem atau state transisi. Sifat dari proses ini adalah jika diketahui proses berada dalam suatu keadaan tertentu, maka peluang berkembangnya proses di masa mendatang hanya tergantung pada keadaan saat ini dan tidak tergantung pada keadaan sebelumnya. Dengan kata lain rantai Markov adalah rangkaian proses kejadian dimana peluang bersyarat kejadian yang akan datang tergantung pada kejadian sekarang. Informasi yang dihasilkan tidak mutlak menjadi suatu keputusan karena sifatnya yang hanya memberikan bantuan dalam proses pengambilan keputusan.

2.9.1 Syarat dalam Analisis Markov

Beberapa syarat yang harus dipenuhi pada analisis Markov adalah sebagai berikut:

1. Jumlah probabilitas transisi untuk suatu keadaan awal dari sistem sama dengan 1.
2. Probabilitas tersebut berlaku untuk semua partisipan dalam sistem.
3. Probabilitas transisi konstan sepanjang waktu.
4. Kondisi merupakan kondisi yang independen sepanjang waktu.

Penerapan analisis Markov cukup terbatas karena sulit menemukan masalah yang memenuhi semua syarat yang diperlukan untuk analisis Markov, terutama persyaratan bahwa probabilitas transisi harus konstan sepanjang waktu. Keadaan transisi adalah perubahan dari suatu keadaan (*state*) ke keadaan (*state*) lainnya pada periode berikutnya. Keadaan transisi ini merupakan suatu proses acak dan dinyatakan dalam bentuk probabilitas. Probabilitas ini dikenal sebagai probabilitas transisi dan dapat digunakan untuk menentukan probabilitas keadaan atau periode berikutnya