BAB III
LANDASAN TEORI

Pada bab ini, penulis akan membahas bagian-bagian yang menjadi landasan dalam melakukan analisis sentiment terhadap opini masyarakat Indonesia tentang Debat Pilpres 2019 pada *Twitter*. Penelitian ini menggunakan metode *Naïve Bayes classifier* mengenai penggunaan metode, algoritma dan *tools.*

3.1 *Twitter*

Dalam sebuah tulisan yang menjelaskan bahwa *Twitter* merupakan media *online* dioperasikan oleh *Twitter Inc* yang buat pada tahun 2006 oleh Jack Dorsey. *Twitter* dibuat pertama kali di San Brunomor, California, San Francisco (Puspita, 2016). *Twitter* digunakan oleh banyak kalangan dengan tujuan yang beraneka ragam, salah satunya karena mudah dalam mengoperasikannya. Pengguna menggunakan *Twitter* biasanya bertujuan untuk berkomunikasi, berbagi pendapat, informasi yang sedang hangat-hangatnya (*trending*), bisa juga sebagai ajang bisnis promosi, juga sebagai sarana untuk mendebatkan sesuatu.

Para pengguna tidak perlu bersusah payah dalam menggunakan *Twitter* karena kemudahannya. Dengan kemudahan dalam menggunakan merupakan beberapa alasan mengapa *Twitter* sampai saat ini masih memiliki jumlah pengguna yang banyak di Indonesia. Para pengguna *Twitter* hanya diberi maksimal 140 karakter disetiap *tweet*nya (Nurhuda et al., 2014). Dengan *Twitter* kita tidak hanya dapat berkomunikasi dengan teman saja, tetapi kita dapat berkomunikasi dengan orang lain. *Twitter* juga sering sekali diramaikan dengan fenomena *hashtag* di mana suatu kejadian atau preistiwa menjadi informasi yang disebar luaskan dengan *hashtag*.

3.1.1 *Twitter* API

a. REST API

Fungsi *REST (Representational State Transfer) API* adalah menyediakan pengaksesan program *Twitter* untuk membaca dan menulis opini di *Twitter*.

REST API bisa mengidentifikasi media sosial *Twitter* menggunakan tanggapan yang tersedia di *JSON (Java Script Object Notation)*.

b. Streaming API

Streaming API yaitu sebuah implementasi dari pengguna *streaming* akan meneruskan pesan yang menunjukan opini-opini lain sudah berjalan dengan pengambilan *endpoint REST*.

c. Ads API

Ads API yaitu sebuah *platform* iklan yang ada di *Twitter* bisa mengintegrasikan sebuah iklan yang akan muncul di *recent Twitter* agar tidak mengganggu pengguna dengan pemakaian media sosial *Twitter* perlu dikelola oleh mitra *Twitter*.

3.2 Data *Mining*

Data *mining* merupakan sebuah tahapan tertentu dalam menggali nilai tambah berupa informasi dari suatu data dengan melakukan penggalian pola-pola untuk menemukan pola menarik dari sejumlah data yang besar dengan tujuan untuk memanipulasi data menjadi informasi yang lebih berisi atau berharga. Data dapat tersimpan di dalam *database, warehouse* dan repositori infromasi lainnya. Data *mining* ini adalah *young interdisciplinary field*, yang diambil dari berbagai bidang seperti *system database*, data *warehousing*, statistik, visualisasi data, dan komputasi berkinerja tinggi. Bidang lain yang berkontribusi dengan data *mining* adalah saraf jaringan, pengenalan pola, dan yang lainnya (Gorunescu, 2011).

Pada proses data *mining* layaknya menggunakan beberapa metode yang bisa digunakan. Metode tersebut digunakan untuk menemukan pengetahuan yang nantinya digunakan. Metode yang ada seperti *classification, clustering, regression, dependency modeling, deviation change detection,* dan *summarization.*

3.3 Analisis Sentimen

Analisis sentimen dapat disebut juga sebagai penambangan opini, seperti dalam bidang studi yang menganalisis pendapat, sentimen, evaluasi, penilaian, sikap dam emosi seseorang terhadap entitas seperti produk, layanan, organisasi, individu, masalah, peristiwa, topi dan atributnya (Zhang & Liu, 2016).

Selanjutnya, analisis sentimen dapat dikatakan sebagai *opinion mining* yang dapat dikatakan sebagai proses dalam memahami sampai kepada mengolah data secara sistem demi mendapatkan informasi yang berkaitan dengan bebuah opini dan pendapat. Analisis sentimen biasanya dilakukan dengan tujuan untuk mengetahui sifat sebuah opini yang sedang dibahas, yang biasanya bersifat pada opini yang negatif atau positif (Buntoro, Adji, & Purnamasari, 2014).

Berdasarkan beberapa pendapat tentang analisis sentimen dapat diketahui bahwa pada dasarnya analisis sentimen merupakan cara bagaimana menganalisis pendapat, tanggapan, komentar, opini dan sentimen dari masyarakat yang biasanya dicurahkan kedalam sosial media ataupun media *online*. Dengan opini yang bermacam-macam, dapat dilihat opini yang bersifat *positif*, *netral* ataupun Negatif.

Seperti yang dikatakan oleh L.Lee dan B.Pang analisis sentimen dibagi menjadi 2 kategori, yaitu *Coarse-grained sentiment analysis* dan *Fined-grained sentiment analysis. Coarse-grained sentiment analysis* merupakan sebuah proses klasifikasi dilakukan berdasarkan orientasi sebuah dokumen secara keseluruhan. Dalam orientasi tersebut dibagi menjadi tiga jenis yaitu negatif, netral, dan positif. Sedangkan *Fined-grained sentiment analysis* menggunakan objek yang berupa sebuah kalimat melainkan bukan sebuah dokumen secara keseluruhan. Seperti contoh kalimat yang mengandung sifat positif "semoga debat pilpres 2019 dapat memudahkan masyarakat memilih Presidennya" dan kalimat negatif "debat pilpres tahun ini banyak menuai perpecahan masyarakat Indonesia". Pada dasarnya tidak semua *tweet* menggunakan kata baku dan memiliki banyak kata yang artinya sama seperti "saya dan aku" dengan kata lain "*gua, gue, aq, ak, sa*" dan sebagainya.

3.4 Metode Naïve Bayes Classifier

Metode *Naïve Bayes classifier* merupakan salah satu metode paling popular dan sering digunakan dalam menyelesaikan masalah klasifikasi dengan pertimbangannya yaitu mudah diimplementasikan. Metode baik dalam melakukan klasifikasi data karena kesederhanaannya metode dan implementasi (Ting, Ip, & Tsang, 2011). Pada dasarnya setiap tulisan merujuk pada proses dan cara

implentasinya yang cukup sederhana. Dengan melakukan analisa suatu topik yang memiliki sifat *positif*, Negatif dan *netral*. Metode *Naïve Bayes classifier* sendiri mempunyai kemampuan untuk klarifikasi data serupa dengan *decision tree* dan *neural network*. Metode *Naïve Bayes classifier* telah terbukti mempunyai akurasi kecepatan tinggi yang dapat diaplikasikan kedalam *database* dengan data yang besar. Metode *Naïve Bayes classifier* merupakan suatu teknik untuk memprediksi sebuah hal probabilistik yang berdasarkan pada penerapan *theorema bayes* oleh fitur-fitur yang di asumsikan oleh *theorema bayes*. Dengan arti dari fitur-fitur yang diasumsikan oleh *theorema bayes* dari data tersebut tidak berkaitan dengan ada atau tidak dalam fitur-fitur lain dengan menggunakan data yang sama. Rumus dasar yang digunakan dalam metode *Naïve Bayes classifier* adalah sebagai berikut:

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

Gambar 3.1. Rumus Dasar Dalam Metode *Naïve Bayes classifier*

Keterangan:

P (X|Ci) = Merupakan probabilitas X terjadi jika Ci sudah terjadi berdasarkan data *training.*

P (Ci) = Merupakan probabilitas Ci dalam data dengan sifat independen terhadap X.

X = X merupakan kumpulan atribut.

P (Ci|X) = Merupakan probabilitas Ci terjadi jika X sudah terjadi.

P (X) = Probalitas dari X.

3.5 Pengumpulan data

Pada tahap ini melakukan pencarian data dan mengumpulkan data yang berkaitan dengan Debat Pemilihan Presiden dan Wakil Presiden 2019. Pencarian dan pengumpulan data diambil dari media sosial *Twitter* di mana data yang digunakan berdasarkan hasil dari *tweet* dengan metode *crawling* yang berhubungan dengan Debat Pemilihan Presiden dan Wakil Presiden 2019. Untuk

memperbanyak data peneliti juga menggunakan *hashtag* untuk mendukung dalam pencarian dan lanjut ketahap *preprocessing* atau *cleaning*.

3.6 *Preprocessing* atau *cleaning*

Prepocessing 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 di mana pemotongan tersebut dibuat menjadi kata atau karakter yang sesuai dengan kebutuhan *system*. Proses ini juga merupakan proses pembersihan karakter tertentu seperti tanda baca dengan melakukan penghilangan tanda baca serta merubah huruf menjadi huruf kecil. Seperti contoh "saya bangga dengan Indonesia" menjadi "saya bangga dengan indonesia".

b. Stemming

Stemming merupakan suatu proses mengubah token yang berimbuhan menjadi kata dasar dengan menghilangkan semua imbuhan yang ada pada token tersebut. Sebagai contoh, kata menyempurnakan, kesempurnaan akan diubah secara otomatis menjadi sempurna.

c. Case folding

Case folding berfungsi menghilangkan angka dan bentuk tanda baca sehingga data yang diambil hanya mengandung karakter huruf a sampai z.

d. Stopword removal

Stopword removal merupakan proses di mana melakukan pembuangan terhadap kata-kata yang tidak berpengaruh pada proses klasifikasi. Kata-kata yang tidak berpengaruh di dalam proses klasifikasi tersebut misalnya yang, dari, di, dan kata penghubung lainnya.

3.7 WEKA

Lunak WEKA (*Waikato Enviroment for Knowledge Analysis*) data *mining* yang memiliki sekumpulan algoritma data *mining* untuk menjalankan proses klasifikasi, *clustering* (pengelompokkan), regresi, asosiasi dan visualisasi. WEKA merupakan perangkat lunak yang dibangun menggunakan bahasa pemrograman

Java yang kemudian didistribusikan menjadi perangkat lunak *open source* (Frank, Hall, & Witten, 2016).

3.8 K-fold Cross Validation

Cross Validation adalah algortima pembelajaran dengan membagi data menjadi dua bagian, satunya digunakan sebagai data *training* dan yang lainnya digunakan sebagai data *testing*. Salah satu metode dari *Cross validation* adalah *k-fold Cross Validation* yang merupakan teknik pembelajaran yang memecah *dataset* sebanyak k-buah secara acak, kemudian dilakukan sejumlah eksperimen sebanyak k-kali yang di mana eksperimen ini menggunakan *dataset* ke-k sebagai data *testing* dan menggunakan sisa *dataset* lainnya sebagai data *training* (Mudry & Tjellström, 2011).

3.9 Confusion Matrix

Suatu metode yang biasanya digunakan untuk melakukan perhitungan akurasi pada sebuah metode klasifikasi. Tabel *confusion matrix* yang digunakan untuk membantu dalam proses perhitungan evaluasi *system* (Tiara, Sabariah, & Effendy, 2015).

	Predicted Positives	Predicted Negatives
Actual Positives instances	Number of True Positives instances (TP)	Number of False Negatives instances (FN)
Actual Negatives instances	Number of False Positives instances (FP)	Number of True Negatives instances (TN)

Gambar 3.2 Tabel *Confusion Matrix*

Rumus tersebut digunakan sebagai sistem dalam menilai sebuah objek yang digunakan dalam analisis. Rumus tersebut dapat dijelaskan seperti berikut:

a. Akurasi

Akurasi merupakan patokan seberapa baik metode yang digunakan dalam mengklasifikasikan data, adapun rumus untuk menghitung nilai akurasi adalah:

Akurasi = (TP + TN) / (TP + TN +FP + FN)

b. Presisi

Presisi merupakan nilai dari ketepatan dari metode yang digunakan dalam klasifikasi. Nilai tersebut menunjukkan banyaknya data yang dapat

terklasifikasi di kelas yang benar dalam beberapa pengujian, adapun rumus untuk menghitung nilai presisi adalah:

Presisi = TP / (TP + FN)

c. Recall

Recall adalah nilai yang dapat mengukur hasil berapa persen data yang terklasifikasikan dengan benar, adapun rumus yang digunakan untuk menghitung nilai *recall* adalah:

Recall = TP / (TP + FN)

Setelah mengetahui nilai akurasi, *error rate*, presisi dan *recall*, penulis menggunakan kurva ROC untuk memvisualisasikan akurasi metode yang digunakan. Kurva ini menggambarkan TPR pada sumbu y dan FPR pada sumbu x. Dengan menggunakan kurva ROC kita dapat mengetahui besar daerah *area under curve (AUC)*. Lalu nilai AUC akan digunakan untuk menguji performa model klasifikasi. AUC yang memiliki nilai 0.5 atau mendekati 0.5 menandakan metode yang digunakan mengklasifikasikan dengan tidak benar. AUC yang memiliki nilai 1.0 menandakan metode yang digunakan dapat mengklasifikasikan dengan benar.

3.10 Zero-R

Metode *Zero-R* merupakan sebuah metode klasifikasi yang paling sederhana yang bergantung pada target dan mengabaikan semua prediktor. Klasifikasi *Zero-R* hanya memprediksi kategori mayoritas (kelas). Meskipun metode *Zero-R* tidak ada kekuatan dalam prediktabilitas, namun hal ini berguna didalam menentukan kinerja dasar sebagai patokan untuk metode klasifikasi lainnya. Pada tahap ini dilakukan proses validasi data menggunakan metode *Zero-R* terhadap *dataset* yang telah terbentuk. Tujuan dari tahap ini untuk menguji apakah metode *Zero-R* dapat melakukan proses klasifikasi dengan baik terhadap *dataset* yang telah dibentuk.

3.11 One-R

One-R merupakan singkatan dari *One Rule*. Algoritma dari *One-R* akan membangkitkan sebuah *rule* untuk setiap atribut kemudian memilih *rule* dengan

error paling dan digunakan sebagai *One Rule*nya. Pada tahap ini dilakukan proses validasi data menggunakan metode *One-R* terhadap *dataset* yang telah terbentuk. Tujuan dari tahap ini untuk menguji apakah metode *One-R* dapat melakukan proses klasifikasi dengan baik terhadap *dataset* yang telah dibentuk.

3.12 N-gram

N-gram merupakan sekumpulan kata yang diambil dari sebuah paragraf dan kalimat. Metode *N-gram* digunakan untuk mengambil kata perkata dari sebuah kontinuitas dibaca dari teks sumber hingga akhir dari dokumen. *N-gram* dibedakan berdasarkan jumlah potongan karakter sebesar n. Untuk membantu dalam mengambil kata perkata yang berupa karakter huruf tersebut, maka dilakukan *padding* dengan *blank* diawal dan diakhir suatu kata. Seperti pada contoh berikut : kata ”*TEXT*” dapat diuraikan ke dalam beberapa *N-gram* berikut (”_” merepresentasikan *blank*):

uni-grams: T, E, X, T

bi-grams: _T, TE, EX, XT, T_

tri-grams: _TE, TEX, EXT, XT_

quad-grams: _TEX, TEXT, EXT_

quint-grams: _TEXT, TEXT_