BAB 2
TINJAUAN PUSTAKA

2.1 Text Mining

Text mining adalah suatu bidang baru yang sedang berkembang yang mencoba untuk mengumpulkan informasi yang memiliki arti dari teks bahasa alami [**5**]. Bidang ini mungkin lebih dikenali sebagai proses dalam menganalisis teks untuk mengekstrak informasi yang berguna untuk suatu tujuan tertentu. Dibandingkan dengan jenis data yang tersimpan dalam *database, text mining* menggunakan data teks yang tidak terstruktur, tidak memiliki bentuk yang jelas dan sulit untuk diuraikan dengan pendekatan algoritma. Namun, dalam budaya modern, teks adalah perantara yang paling umum untuk pertukaran secara formal dari informasi. Bidang *text mining* biasanya berkaitan dengan teks yang memiliki fungsi untuk komunikasi dari informasi yang faktual atau opini, dan keinginan untuk mencoba mengekstrak informasi dari sebuah teks secara otomatis merupakan hal yang menarik meskipun tingkat keberhasilan yang diperoleh hanyalah sebagian [**6**].

Text mining umumnya digunakan untuk menunjukkan sistem yang menganalisis sejumlah besar teks bahasa alami dan mendeteksi pola penggunaan leksikal atau bahasa dalam upaya untuk mengekstrak informasi yang mungkin berguna [**6**].

2.2 Analisis Sentimen

Sentiment analysis atau *opinion mining* mengacu pada penggunaan Pengolahan Bahasa Alami, Analisis Teks dan Komputasi Linguistik untuk mengidentifikasi dan mengekstraksi informasi dari bahan sumber yang berupa teks. Secara umum, analisis sentimen bertujuan untuk menentukan sikap pembicara atau penulis terhadap suatu topik atau keseluruhan polaritas kontekstual pada suatu dokumen. Sikap dapat berupa penilaian atau evaluasi, sisi

emosional penulis pada saat menulis atau efek komunikasi emosional yang penulis inginkan terhadap pembacanya [**2**].

Tugas dasar dalam analisis sentimen adalah mengklasifikasikan polaritas dari teks yang ada dalam dokumen, kalimat, atau fitur/tingkat aspek – apakah pendapat yang dikemukakan dalam dokumen, kalimat atau fitur entitas/aspek bersifat positif, negatif atau netral. Lebih lanjut manfaat lain dari analisis sentimen adalah dapat mengklasifikasikan ungkapan emosional seperti sedih, gembira, atau marah.

Ekspresi atau sentimen mengacu pada fokus topik tertentu, pernyataan pada satu topik mungkin akan berbeda makna dengan pernyataan yang sama pada subjek yang berbeda. Sebagai contoh, adalah hal yang baik untuk mengatakan alur film tidak terprediksi, tapi adalah hal yang tidak baik jika 'tidak terprediksi' dinyatakan pada kemudi dari kendaraan. Bahkan pada produk tertentu, kata – kata yang sama dapat menggambarkan makna kebalikan, contoh adalah hal yang buruk untuk waktu *start-up* pada kamera digital jika dinyatakan "lama", namun jika "lama" dinyatakan pada usia baterai maka akan menjadi hal positif. Oleh karena itu pada beberapa penelitian, terutama pada *review* produk, pekerjaan didahului dengan menentukan elemen dari sebuah produk yang sedang dibicarakan sebelum memulai proses *opinion mining* [**4**].

2.3 Twitter

Twitter adalah jejaring sosial *online* dan layanan *microblogging* yang memungkinkan penggunanya untuk mengirim dan membaca pesan berbasis teks yang dibatasi sebanyak 140 karakter dikenal dengan sebutan *tweet*. Pengguna yang terdaftar dapat membaca dan mencatat *tweet*, sedangkan pengguna yang tidak terdaftar hanya dapat membacanya. Pengguna mengakses Twitter melalui tampilan situs, pesan singkat, atau melalui aplikasi untuk perangkat selular.

Twitter didirikan pada bulan Maret 2006 oleh Jack Dorsey, Evan Williams, Biz Stone dan Noah Glass, lalu diluncurkan pada bulan Juli 2006. Layanan Twitter berkembang secara pesat dan mendunia dalam waktu singkat, dengan 500

juta pengguna terdaftar pada tahun 2012 yang mencatat 340 juta *tweet* per harinya. Layanan Twitter juga menangani 1,6 milyar kata kunci pencarian per harinya. Twitter menjadi satu dari sepuluh situs yang paling sering dikunjungi dan dikenal dengan sebutan "Pesan singkat melalui Internet".

Tingginya popularitas Twitter menyebabkan layanan ini telah dimanfaatkan untuk berbagai keperluan dalam berbagai aspek, misalnya sebagai sarana protes, kampanye politik, sarana pembelajaran, dan sebagai media komunikasi darurat. Twitter juga dihadapkan pada berbagai masalah dan kontroversi seperti masalah keamanan dan privasi pengguna, gugatan hukum, dan penyensoran.

2.4 Naive Bayes

Bayes merupakan teknik prediksi berbasis probabilistik sederhana yang berdasar pada penerapan teorema Bayes dengan asumsi independensi (ketidaktergantungan) yang kuat (naif) [**3**].

Dalam Bayes (terutama Naive Bayes), maksud independensi yang kuat pada fitur adalah bahwa sebuah fitur pada sebuah data tidak berkaitan dengan ada atau tidaknya fitur lain dalam data yang sama. Contohnya pada kasus klasifikasi hewan dengan fitur penutup kulit, melahirkan, berat dan menyusui. Dalam dunia nyata, hewan yang berkembang biak dengan cara melahirkan dipastikan juga menyusui. Disini ada ketergantungan pada fitur menyusui karena hewan yang menyusui biasanya melahirkan, atau hewan yang bertelur biasanya tidak menyusui. Dalam Bayes, hal tersebut tidak dipandang sehingga masing – masing fitur seolah tidak memiliki hubungan apa pun [**3**].

Prediksi Bayes didasarkan pada teorema Bayes dengan persamaan umum seperti terlihat pada Persamaan 2.1

$$\boldsymbol{P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}}$$

(2. 1)

Keterangan :

P(H|E) = Probabilitas akhir bersyarat (*conditional probability*) suatu hipotesis H terjadi jika diberikan bukti (*evidence*) E terjadi.

P(E|H) = Probabilitas sebuah bukti E akan mempengaruhi hipotesis H

P(H) = Probabilitas awal (priori) hipotesis H terjadi tanpa memandang bukti apapun

P(E) = Probabilitas awal (priori) bukti E terjadi tanpa memandang hipotesis/bukti yang lain

Ide dasar dari aturan Bayes adalah bahwa hasil dari hipotesis atau peristiwa (H) dapat diperkirakan berdasarkan pada beberapa bukti (E) yang diamati. Ada beberapa hal penting dari aturan Bayes tersebut, yaitu

1. Sebuah probabilitas awal/priori H atau P(H) adalah probabilitas dari suatu hipotesis sebelum bukti diamati.
2. Sebuah probabilitas akhir H atau P(H|E) adalah probabilitas dari suatu hipotesis setelah bukti diamati.

Tergantung pada situasi yang tepat dari model probabilitas, Naive Bayes dapat dilatih sangat efisien dalam *supervised learning*. Dalam aplikasi praktis, parameter estimasi untuk model Naive Bayes menggunakan metode *likelihood* maksimum, dengan kata lain, seseorang dapat bekerja dengan model Naïve Bayes tanpa mempercayai probabilitas Bayesian atau menggunakan metode Bayesian lainnya [**4**].

Sebuah keuntungan dari Naive Bayes adalah bahwa ia memerlukan sejumlah kecil data pelatihan untuk mengestimasi parameter (rata – rata dan varian dari variabel) yang diperlukan untuk klasifikasi. Karena variabel diasumsikan independen, hanya varian dari variabel – variabel untuk setiap kelas yang perlu ditentukan dan bukan keseluruhan *covariance matrix* [**4**].

2.4.1 Naive Bayes untuk Klasifikasi

Kaitan antara Naive Bayes dengan klasifikasi, korelasi hipotesis, dan bukti dengan klasifikasi adalah bahwa hipotesis dalam teorema Bayes merupakan label kelas yang menjadi target pemetaan dalam klasifikasi, sedangkan bukti merupakan fitur – fitur yang menjadi masukan dalam model klasifikasi. Jika X adalah vektor masukan yang berisi fitur dan Y adalah label kelas, Naive Bayes dituliskan dengan P(Y|X). Notasi tersebut berarti probabilitas 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 [**3**].

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 nilai Y' dengan memaksimalkan nilai P(Y'|X') yang didapat [**3**].

Formulasi Naive Bayes untuk klasifikasi seperti dapat dilihat pada Persamaan 2.2

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{q} P(X_i|Y)}{P(X)} \quad \textbf{(2. 2)}$$

Keterangan

P(Y|X) = Probabilitas data dengan vektor X pada kelas Y.

P(Y) = Probabilitas awal kelas Y.

$\prod_{i=1}^{q} P(X_i|Y)$ = Probabilitas independen kelas Y dari semua fitur dalam vektor X.

Nilai P(X) selalu tetap sehingga dalam perhitungan prediksi nantinya kita tinggal menghitung bagian $P(Y)\prod_{i=1}^{q} P(X_i|Y)$ dengan memilih yang terbesar

sebagai kelas yang dipilih sebagai hasil prediksi. Sementara probabilitas independen $\prod_{i=1}^{q} P(X_i|Y)$ tersebut merupakan pengaruh semua fitur dari data terhadap setiap kelas Y, yang dinotasikan dengan Persamaan 2.3

$$P(Y|X = x) = P(Y) \prod \prod_{i=1}^{q} P(X_i|Y = y) \quad \textbf{(2. 3)}$$

Setiap set fitur X= $\{X_1, X_2, X_3, \dots, X_q\}$ terdiri atas q atribut (q dimensi).

Umumnya, Bayes mudah dihitung untuk fitur bertipe kategoris seperti pada kasus klasifikasi hewan dengan fitur "penutup kulit" dengan nilai {bulu, rambut, cangkang}, atau kasus fitur "jenis kelamin" dengan nilai {pria, wanita}. Namun untuk fitur dengan tipe numerik (kontinu) ada perlakuan khusus sebelum dimasukkan dalam Naive Bayes [**3**]. Caranya adalah :

1. Melakukan diskretisasi pada setiap fitur kontinu dan mengganti nilai fitur kontinu tersebut dengan nilai interval diskret. Pendekatan ini dilakukan dengan mentransformasi fitur kontinu ke dalam fitur ordinal.
2. Mengasumsikan bentuk tertentu dari distribusi probabilitas untuk fitur kontinu dan memperkirakan parameter distribusi dengan data pelatihan. Distribusi *Gaussian* biasanya dipilih untuk merepresentasikan probabilitas bersyarat dari fitur kontinu pada sebuah kelas $P(X_i|Y)$, sedangkan distribusi *Gaussian* dikarakteristikkan dengan dua parameter : mean, μ, dan varian, σ^2. Untuk setiap kelas y_j, probabilitas bersyarat kelas y_j untuk fitur X_i seperti terlihat pada Persamaan 2.4

$$P(X_i = x_i \mid Y = y_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} exp^{-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}} \quad \textbf{(2. 4)}$$

Parameter μ_{ij} bisa didapat dari mean sampel X_i $(\bar{x})$ dari semua data latih yang menjadi milik kelas y_j, sedangkan σ_{ij}^2 dapat diperkirakan dari varian sampel (s^2) dari data latih [**3**].

2.4.2 Karakteristik Naive Bayes

Klasifikasi dengan Naive Bayes bekerja berdasarkan teori probabilitas yang memandang semua fitur dari data sebagai bukti dalam probabilitas. Hal ini memberikan karakteristik Naive Bayes sebagai berikut :

1. Metode Naive Bayes teguh (*robust*) terhadap data – data yang terisolasi yang biasanya merupakan data dengan karakteristik berbeda (*outlier*). Naive Bayes juga dapat menangani nilai atribut yang salah dengan mengabaikan data latih selama proses pembangunan model dan prediksi.
2. Tangguh menghadapi atribut yang tidak relevan.
3. Atribut yang mempunyai korelasi bisa mendegradasi kinerja klasifikasi Naive Bayes karena asumsi independensi atribut tersebut sudah tidak ada [**3**].

2.5 UML

UML (*Unified Modeling Language*) merupakan standarisasi bahasa pemodelan untuk pembangunan perangkat lunak yang dibangun dengan menggunakan teknik pemrograman berorientasi objek. UML muncul karena adanya kebutuhan pemodelan visual untuk menspesifikasikan, menggambarkan, membangun, dan dokumentasi dari sistem perangkat lunak. UML merupakan bahasa visual untuk pemodelan dan komunikasi mengenai sebuah sistem dengan menggunakan diagram dan teks – teks pendukung [**7**].

2.5.1 Diagram UML

Pada UML terdapat 13 macam yang dikelompokkan ke dalam 3 kategori berikut :

a. *Structure Diagram*

   Kumpulan diagram yang digunakan untuk menggambarkan suatu struktur statis dari sistem yang dimodelkan. Terdiri atas *Class Diagram*, *Object Diagram*, *Component Diagram*, *Composite Structure Diagram*, *Package Diagram* dan *Deployment Diagram*.

b. *Behavior Diagram*

Kumpulan diagram yang digunakan untuk menggambarkan kelakuan sistem atau rangkaian perubahan yang terjadi pada sebuah sistem. Terdiri atas *Use Case Diagram*, *Activity Diagram* dan *State Machine Diagram*.

c. *Interaction Diagram*

Kumpulan diagram yang digunakan untuk menggambarkan interaksi sistem dengan sistem lain maupun interaksi antar subsistem pada suatu sistem. Terdiri atas *Sequence Diagram, Communication Diagram, Timing Diagram* dan *Interaction Overview Diagram*.

Adapun dari 13 diagram tersebut yang akan digunakan pada sistem yang dibangun adalah sebagai berikut :

2.5.1.1 Use Case Diagram

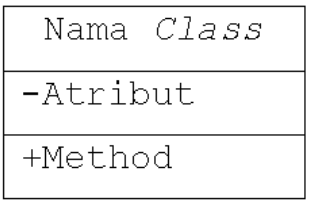
Merupakan pemodelan untuk kelakuan (*behavior*) sistem yang akan dibuat. *Use Case* mendeskripsikan sebuah interaksi antara satu atau lebih aktor dengan sistem yang akan dibuat. Secara kasar, *Use Case* digunakan untuk mengetahui fungsi apa saja yang ada di dalam sebuah sistem dan siapa saja yang berhak menggunakan fungsi – fungsi itu.

2.5.1.2 Activity Diagram

Menggambarkan *workflow* (aliran kerja) atau aktivitas dari sebuah sistem atau proses bisnis atau menu yang ada pada perangkat lunak. Menggambarkan aktivitas dari sistem bukan apa yang dilakukan aktor.

2.5.1.3 Class Diagram

Menggambarkan struktur sistem dari segi pendefinisian kelas – kelas yang akan dibuat untuk membangun sistem. Terdiri dari atribut yang merupakan variabel – variabel yang dimiliki oleh suatu kelas, serta metode atau operasi yang merupakan fungsi – fungsi yang dimiliki oleh suatu kelas.



Gambar 2. 1 Contoh *Class* Diagram

2.5.1.4 Sequence Diagram

Menggambarkan kelakuan objek pada *Use Case* dengan mendeskripsikan waktu hidup objek dan *message* yang dikirimkan dan diterima antar objek.

2.6 Cross Validation

Untuk tujuan replikasi, yang dibutuhkan adalah mendapatkan jumlah dari data latih dan pasangan data uji dari *dataset* X (setelah lebih dahulu memisahkan sebagian data sebagai data uji). Untuk mendapatkannya, jika sampel X cukup besar, kita dapat membaginya secara acak kedalam K bagian, lalu secara acak memisahkan masing-masing bagian ke dalam 2 kelompok dan menggunakan setengah data pertama untuk data latih dan setengah data lainnya untuk data uji. Pengujian ini dilakukan berulang dengan menggunakan data *split* yang berbeda, dinamakan *cross validation*.

K-fold cross-validation adalah sebuah pengujian dimana *dataset* X dibagi secara acak kedalam K bagian dengan ukuran yang sama, Xi, i=1,...,K. Untuk memproses tiap pasangnya, diambil terlebih dahulu salah satu bagian dari K bagian sebagai data tes, dan mengkombinasikan sisanya sebanyak K-1 bagian sebagai data latih. Proses pengujian dilakukan sebanyak K kali [**8**]. Nilai K yang umum digunakan adalah 5 atau 10 [**9**] [**10**].

Dari pengujian yang dilakukan, dievaluasi dengan menggunakan *confusion matrix* untuk mengetahui tingkat kebenaran dari penelitian.

Tabel 2. 1 *Confusion* Matrix

	Tested Class		
Predicted Class	Positive	Negative	Total
Positive	True Positive	False Negative	*p*
Negative	False Positive	True Negative	*n*
Total	*p'*	*n'*	N

Keterangan

True Positive : Hasil diprediksi bernilai *True*, dan pengujian menghasilkan *True*

False Positive : Hasil diprediksi bernilai *False*, dan pengujian menghasilkan *True*

False Negative : Hasil diprediksi bernilai *True*, dan pengujian menghasilkan *False*

True Negative : Hasil diprediksi bernilai *False*, dan pengujian menghasilkan *False.*

Setelah hasil pengujian menggunakan *k-fold cross validation* dievaluasi menggunakan *confusion matrix*, selanjutnya dapat dihitung nilai *error* dari keseluruhan iterasi dengan Persamaan 2.5

$$CV(\alpha) = \frac{1}{K}\sum_{i=1}^{K} Error_i(f(\times, \alpha))$$

(2. 5)

Keterangan

$CV(\alpha)$ = Nilai *error* pada pengujian

K = Jumlah *fold*

$f(\times, \alpha)$ = Jumlah error pada masing – masing iterasi