BAB II
LANDASAN TEORI

2.1 Tinjauan Pustaka

2.1.1 Penelitian Sebelumnya

Adapun perbandingan antara penelitian analisis data yang telah dibuat dengan penelitian analisis sebelumnya terkait dengan Skripsi ini dapat dilihat pada tabel 2.1.

Tabel 2.1 Perbandingan dengan Penelitian Sebelumnya

Perbandingan	Penelitian 1	Penelitian 2	Penelitian Sekarang
Objek Yang Diteliti	Penelitian Kepuasan Penyedia Layanan Telekomunikasi	Analisis sentimen terhadap review sebuah restoran	Analisis Sentimen Menjelang Pilplres 2019
Tujuan	Untuk Mengetahui tingkat kepuasan customer	Untuk menganalisa sentimen pada review pengunjung terhadap restoran	Untuk Mengetahui Sentimen Masyarakat Terhadap Pilpres 2019
Studi Kasus	Twitter	Zomato, blog, twitter	Twitter
Algoritma	*SVM dan Lexicon Based*	*NaÏve Bayes*	*NaÏve Bayes*
Software Pengembang	-	*Web aplication*	*Dekstop*

A. Analisis sentimen tingkat kepuasan pengguna penyedia layanan telekomunikasi seluler indonesia pada twitter dengan metode *Support Vector Machine* (SVM) dan *Lexicon Based Features* (Umi Rofiqoh, 2017). Pada penelitian ini dilakukan analisa data untuk mengetahui tingkat kepuasan customer dan sentimen terhadap jasa penyedia layanan tersebut. Penelitian ini data diperoleh dengan mengambil data dari media sosial *twitter*, data yang diambil dari twitter berjumlah 300 dengan rasio data latih 70% dan data uji sebanyak 30%. Selanjutnya data data ini dilakukan pembobotan kata dengan proses klasifikasi *term frequency(tf), document frequency(df)* dan *inverse document frequency(idf)*. Data yang sudah dilakukan proses klasifikasi selanjutnya dihitung dengan algoritma *Support*

Vector Machine (SVM). Untuk mendapatkan hasil dari sentimen negatif dan positif penelitian ini menggunakan fitur *Lexicon Based*. Hasil yang didapatkan dari analisis sentimen pada tingkat kepuasan penggunaan penyedia layanan telekomunikasi seluler menggunakan *Support Vector Machine* dan *Lexicon Based* adalah 79% nilai *accuracy*, *precision* 65%, *recall* sebesar 97%, dan f-measure sebesar 78%. Apabila tidak menggunakan *Lexicon Based* nilai *recall* akan menurun menjadi 78%. Hal ini menjadi bukti jika *Lexicon Based* pada analisis sentimen mempunyai pengaruh yang besar.

B. Analisis Sentimen Pada Review Restoran dengan Teks Bahasa Indonesia Menggunakan Algoritma *NaÏve Bayes* (Mutia, 2017). Penelitian ini hampir sama dengan penelitian pertama, namun yang membedakan adalah algoritma yang dipakai dan fitur yang digunakan. Pada penelitian ini menggunakan *NaÏve Bayes* sebagai algoritma dan untuk fiturnya menggunakan *Genetics Algorithm.* Untuk hasilnya diperoleh 86.50% dari hasil yang sudah dihitung menggunakan *NaÏve Bayes*, tetapi tingkat akurasi dari perhitungan tersebut akan naik bila didalam proses perhitungan sudah menggunakan fitur *Genetics Algorithm.* Penggunaan fitur ini mempengaruhi hasil dari perhitungan yang menyebabkan kenaikan akurasi sebanyak 4%, sehingga jumlah akurasi jika *NaÏve Bayes* dan *Genetics Algorithm* digabungkan adalah sebesar 90.50%

2.2 Dasar Teori

2.2.1 Pengertian Analisis Sentimen

Opinion Mining atau analisis sentimen merupakan salah satu bidang dari ilmu komputer yang mempelajari komputasi linguistik, pengolahan bahasa alami, dan text mining yang bertujuan untuk menganalisa emosi, penilaian, sikap, pendapat, sentimen, evaluasi seorang terhadap seorang pembicara atau penulis berkenaan dengan suatu produk, layanan, organisasi, individu, tokoh publik, topik, acara, ataupun kegiatan tertentu (Liu, 2012).

Proses utama dalam analisis sentimen yaitu mengelompokkan teks yang terdapat dalam suatu kalimat atau suatu dokumen yang kemudian diproses untuk mengetahui apakah teks tersebut bersifat negatif, positif ataupun netral. Analisis sentimen dapat digunakan untuk mencari pendapat tentang produk, merk, atau tokoh publik sekalipun dan dapat menentukan apakah mereka dilihat sebagai positif, negatif atau netral (Saraswati, 2011). Hal ini memungkinkan pengguna untuk mencari informasi tentang: 1) Deteksi Flame (rants buruk), 2) Persepsi produk baru, 3) Persepsi merk, 4) Manajemen reputasi.

Analisis sentimen difokuskan untuk *review* klasifikasi berdasarkan polaritas. Berdasarkan polaritas tersebut, analisis sentimen dibagi menjadi dua bagian yaitu dokumen klasifikasi ke pendapat atau fakta dan dokumen klasifikasi ke dalam positif dan negatif atau yang lebih sering 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.

2.2.2 Pengertian Text Mining

Text mining merupakan bagian dari data mining dimana proses yang dilakukan utamanya adalah melakukan esktraksi pengetahuan dari informasi dari pola-pola yang terdapat dalam sekumpulan dokumen teks menggunakan alat analisis tertentu (R. Feldman, 2016). Text mining diolah untuk keperluan mencari dokumen teks pada analisis sentimen.

Text mining bertujuan untuk mencari kata-kata yang dapat mewakili apa yang ada didalam dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen. Text mining mempunyai 5 tahapan yaitu *Tokenizing*, *Filtering*, *Stemming*, *Tagging*, dan *Analyzing*.

Tahapan *Tokenizing* adalah proses pemotongan string masukan berdasarkan tiap kata yang menyusunnya. Pada prinsipnya proses ini adalah memisahkan setiap kata yang menyusun suatu dokumen tersebut. Tahap *Filtering* adalah suatu proses dimana diambil sebagian data tertentu, dan membuang data frekuensi yang lain. Tahapan *Stemming* adalah proses pemetaan dan penguraian berbagai bentuk atau varian dari suatu kata menjadi bentuk kata dasarnya. Tahapan *Tagging* adalah kata yang belum lama dilahirkan (singkatan singkatan). Tahapan *Analyzing* adalah

untuk mencari seberapa jauh keterhubungan antar kata-kata setiap pada setiap dokumen.

2.2.3 Pengertian Twitter

Twitter adalah sebuah situs *web* yang dimiliki dan dioperasikan oleh Twitter Inc., yang menawarkan jaringan sosial berupa *microblog* sehingga memungkinkan penggunanya untuk mengirim dan membaca pesan *tweet* (Twitter, 2013). Microblog adalah suatu jenis alat komunikasi *online* dimana penggunanya dapat memperbaharui status tentang mereka yang sedang memikirkan dan melakukan sesuatu, apa pendapat mereka tentang suatu objek atau fenomena tertentu yang sedang terjadi. *Tweet* adalah teks tulisan yang hanya dibatasi dengan jumlah 280 karakter yang ditampilkan pada profil pengguna tersebut. *Tweet* bisa dilihat secara publik, namun juga bisa dilihat hanya oleh para pengikut (*follower*) saja karena terdapat fitur private akun yang membatasi ruang dari *tweet* itu sendiri.

Tidak seperti *Facebook, LinkedIn,* dan *Myspace*. *Twitter* merupakan sebuah jejaring sosial yang dapat digunakan sebagai sebuah graph berarah (Wang, 2010). Yang berarti penggunanya dapat mengikuti pengguna lainnya, namun pengguna kedua tidak diperlukan untuk mengikutinya kembali. Kebanyakan akun bersifat publik, dimana orang lain yang belum mengikuti akun tersebut dapat melihat *tweet* dari akun tersebut. Sedangkan akun private hanya orang yang mengikuti dari akun tersebut yang dapat melihat isi dari *tweet* dan profil dari akun tersebut.

Pengguna dari *Twitter* dapat menulis pesan berdasarkan topik yang sedang *booming* maupun tidak dengan menyisipkan tanda (#) yang biasa disebut dengan *hashtag*. Sedangkan untuk menyebutkan orang lain atau dalam *twitter* terkenal dengan istilah *mention,* pengguna dapat menyisipkan tanda (@).

Pada awal kemunculannya *Twitter* mempunyai batasan sampai 140 karakter saja, hal ini disesuaikan dengan kompatibilitas dengan pesan sms. Batasan karakter yang berjumlah 140 ini juga menjadi peningkatan dalam penggunaan layanan untuk memperpendek URL seperti *bit.ly, goo.gl, dan tr.im*. Namun seiring berjalannya waktu Twitter resmi memperpanjang jumlah karakter

menjadi 280 pada tanggal 08 November 2017 (Tirto.ID, 2017). Fitur-fitur yang terdapat pada twitter antara lain :

1. Halaman Utama (*Home*)

   Pada halaman utama ini kita dapat melihat *tweet* yang dikirimkan dari akun yang sudah diikuti (*following*).
2. Notifikasi (*Notifications*)

   Notifikasi pada *Twitter* berisi tentang sekumpulan pemberitahuan tentang adanya *mention* yang masuk maupun orang yang sudah mengikuti akun ataupun notifikasi tentang aktivitas dari *followers*.
3. Pesan Langsung (*Direct Message*)

   Pada halaman pesan ini berisi tentang sebuah pesan yang masuk daripada orang lain ke akun *twitter*.
4. Pencarian (*Search*)

   Kolom *search* berguna untuk mencari sesuatu yang berhubungan dengan *tweet*, topik maupun akun akun yang ingin dicari.
5. Profil (*Profile*)

   Pada halaman ini menampilkan tentang seluruh profil dari akun. Jumlah pengikut, jumlah yang diikuti serta jumlah *tweet* akan dapat dilihat.
6. *Followers*

   Pengikut adalah sebuah akun akun yang sudah mengikutin akun yang kita buat. Jika kita menuliskan sebuah *tweet* maka otomatis akan selalu terlihat oleh pengikut dari akun kita dan masuk ke halaman utama.
7. *Following*

   Sebaliknya *following* adalah akun akun yang kita ikuti. *Tweet* dari akun yang kita ikuti dapat dilihat pada halaman utama tanpa harus mengunjungi profil akun tersebut.
8. *Mentions*

   Fitur ini mempunyai fungsi untuk memulai percakapan antar dua akun atau lebih dan untuk menandai akun pada suatu *tweet*.

9. *Re-tweet*

   Sebuah fitur untuk mengutip *tweet* dari orang lain dan akan muncul pada halaman profil.

10. *Like* (*Favorit*)

    Tweet ditandai sebagai favorit akan *tweet* tersebut tidak hilang oleh halaman sebelumnya.

11. *Hashtag* (#)

    Hashtag "#" yang ditulis didepan topik tertentu agar pengguna lain bisa mencari topik yang sejenis yang ditulis oleh orang lain juga.

12. *List*

    Pengguna *twitter* dapat mengelompokkan ikutan mereka dalam satu group sehingga memudahkan untuk dapat melihat secara keseluruhan para nama pengguna (*username*) yang mereka ikuti (*follow*).

13. Topik Terkini (*Trending Topic*)

    Topik yang sedang banyak dibicarakan oleh banyak pengguna dalam satu waktu yang bersamaan.

2.2.4 Pengertian Klasifikasi

Klasifikasi merupakan proses menemukan sebuah model atau fungsi yang mendeskripsikan dan membedakan data ke dalam kelas-kelas. Klasifikasi melibatkan proses pemeriksaan karakteristik dari objek dan memasukkan objek ke dalam salah satu kelas yang sudah didefinisikan sebelumnya. Pemilahan banyak data yang sesuai dengan persamaan atau perbedaan yang dikandungnya dinamakan klasifikasi. Berikut adalah jenis-jenis dari klasifikasi data :

a. Klasifikasi menurut jenis data.
   - Data hitung (*enumeration/counting data*).

     Data hitung adalah hasil perhitungan atau jumlah tertentu, yang termasuk data hitung adalah persentase dari suatu jumlah tertentu. Contoh mencatat jumlah tweet dalam suatu kumpulan atau

persentase data tweet menjadi kelas positif dan negatif yang kemudian menghasilkan suatu hitungan.

- Data Ukur (*Measurement Data*).
  Data ukur adalah data yang menunjukkan ukuran mengenai nilai sesuatu.

b. Klasifikasi data menurut sifat data.
   - Data Kuantitatif (*Quantitative Data*).
     Data kuantitatif adalah data mengenai penggolongan dalam hubungannya dengan penjumlahan.
   - Data Kualitatif (*Qualitative Data*).
     Data kualitatif adalah data mengenai penggolongan dalam hubungannya dengan kualitas atau sifat tertentu.

Sedangkan model dari klasifikasi tersebut dibagi menjadi 2 yaitu permodelan deskriptif dan permodelan prediktif. Permodelan deskriptif adalah model yang dapat bertindak sebagai suatu alat yang bersifat menjelaskan untuk membedakan antara objek dengan kelas yang berbeda. Pemodelan prediktif adalah model klasifikasi juga dapat menggunakan prediksi label kelas yang belum diketahui recordnya.

Tujuan dari klasifikasi itu sendiri adalah sebagai berikut :

1. Menemukan model dari training set yang membedakan record kedalam kategori atau kelas yang sesuai, model tersebut kemudian digunakan untuk mengklasifikasikan record yang kelasnya belum diketahui sebelumnya pada test set.
2. Mengambil keputusan dengan memprediksikan suatu kasus, berdasarkan hasil klasifikasi yang diperoleh.

2.2.5 Pengertian Algoritma NaÏve Bayes

NaÏve Bayes Classifier adalah salah satu metode klasifikasi yang berakar pada teorema *Bayes*. Ciri utama dari *NaÏve Bayes Classifier* ini adalah asumsi yang sangat kuat (naif) terhadap tingkat independensi dari masing-masing kondisi

atau kejadian. Terdapat dua tahap klasifikasi dokumen *tweet* pada penelitian ini. Tahap pertama adalah proses training terhadap dokumen yang sudah diketahui kategorinya. Sedangkan tahap kedua adalah menghitung data yang belum diketahui kelasnya sehingga melalui perhitungan tersebut dapat diketahui dokumen atau data tersebut masuk dalam kategori positif atau negatif.

Dalam algoritma *NaÏve Bayes Classifier* setiap dokumen direpresentasikan dengan pasangan atribut "$x_{1,}\, x_{2}$, $x_{3\,\ldots}\, x_{n}$" dimana x_1 adalah kata pertama, x_2 adalah kata kedua dan seterusnya. Sedangkan V adalah himpunan kategori tweet sebagai berikut :

$$P(V\,|x_1,\ldots,\, x_n) = \frac{\mathrm{P(V)P(V\,|x1,\ldots,xn|V)}}{\mathrm{P(x1,\ldots,xn\,)}} \qquad (1)$$

Dimana variabel V mepresentasikan kelas, sementara variabel $x_{1,\ \ldots,}\ x_{n}$ mempresentasikan karakteristik-karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi. Maka rumus tersebut menjelaskan bahwa peluang masuknya sampel dengan karakteristik tertentu dalam kelas V (*posterior*) adalah peluang munculnya kelas V (sebelum masuknya sampel tersebut, disebut *prior*), dikali dengan peluang kemunculan karakteristik-karakteristik sampel pada kelas V (*likelihood*), dibagi dengan peluang kemunculan karakteristik-karakteristik sampel secara global (*evidence*).

Nilai evidence selalu tetap untuk setiap kelas pada satu sampel. Nilai dari posterior tersebut yang nantinya akan dibandingkan dengan nilai-nilai *posterior* kelas lainnya untuk menentukan ke kelas apa suatu sampel akan diklasifikasikan. Penjabaran lebih lanjut rumus *NaÏve Bayes Classifier* dapat dilakukan dengan menjabarkan P(x_1, ... , x_n| V) menggunakan aturan perkalian, menjadi sebagai berikut:

$$\begin{aligned} P(x1,\ldots,xn|V) &= P(x1|V)\ P(x2,\ldots,xn|V,x1) \\ &= P(x1|V)P(x2|V,x1)P(x3,\ldots,xn|V,xn|V,x1,x2) \\ &= P(x1|V)P(x2|V,x1)\ldots P(xn|V,x1,x2,\ldots\, xn-1) \end{aligned} \qquad (2)$$

Hasil penjabaran persamaan (2) memperlihatkan semakin banyak dan semakin kompleksnya faktor-faktor syarat yang mempengaruhi nilai probabilitas,

sehingga menjadi rumit untuk dianalisa satu-persatu. Akibatnya, perhitungan tersebut menjadi sulit untuk dilakukan.

Disinilah digunakan asumsi independensi yang sangat tinggi (naif), bahwa masing-masing fitur (x_1, x_2, x_3 ... x_n) saling bebas (*independent*) satu sama yang lain. Dengan asumsi tersebut, maka berlaku suatu kesamaan sebagai berikut:

$$P(x_i|x_j) = \frac{P(xi \cap xj)}{P(xj} = P(x_i) \quad (3)$$

Untuk i ≠ j, sehingga persamaan (3) menjadi

$$P(xi|V,xj) = P(xi|V) \quad (4)$$

Dari persamaan (4) dapat disimpulkan bahwa asumsi independensi naif tersebut membuat syarat peluang kejadian sederhana, sehingga perhitungan menjadi mungkin untuk dilakukan. Selanjutnya, penjabaran $P(x1, \ldots, xn|V)$ dapat disederhakan menjadi seperti berikut :

$$P(x1,\ldots,xn|V) = P(x1|V)P(x2|V)\ldots P(xn|V)P(x1,\ldots,xn|V) \quad (5)$$

$$P(x1,\ldots,xn|V) = \Pi P(xi|V)ni=1 \quad (6)$$

Dari persama (6), persamaan (1) *NaÏve Bayes Classifier* dapat dituliskan sebagai berikut :

$$P(V\ |x_1, \ldots, x_n) = \frac{P(V)\prod_{i=1}^{n} P(xi|V)}{P(x1,x2\ \ldots,xn)} \quad (7)$$

Persaman (7) merupakan model dari teorema *NaÏve Bayes Classifier* yang selanjutnya akan digunakan pada klasifikasi data *tweet* (Dharmawan, 2014). Pada saat klasifikasi algoritma mencari *probabilitas* tertinggi dari semua dokumen yang diujikan, dimana persamaan (7) menjadi sebagai berikut :

$$P(V_j|x_1, \ldots, x_n) = \underset{Vj \in V}{argmax} \ \frac{P(V)\prod_{i=1}^{n} P(xi|Vj)}{P(x1,x2\ \ldots,xn)} \quad (8)$$

Adapun V_j adalah kategori *tweet* dimana penelitian ini j1 = kategori *tweet* sentimen negatif, j2 = kategori *tweet* sentimen positif, dan j3 = kategori *tweet* sentimen netral. Sedangkan $P(x1, x2 \ldots, xn)$ mempresentasikan *evidence* yang nilainya konstan untuk semua kelas pada satu sampel. Penjabaran *evidence* tersebut yaitu :

$$P(x_1,x_2,x_3,\dots x_n) = P(x_1 \cup x_2 \cup x_3 \cup \dots \cup x_n)$$
$$= P(x_1 + x_2 + x_3 + \dots + x_n)$$
$$= P(x_1) + P(x_2) + P(x_3) + \dots + P(x_n)$$
$$= \sum_{i=1}^{n} P(x_i)$$
$$= 1$$

Sehingga persamaan (8) dapat disederhanakan menjadi sebagai berikut:

$$P(V_j|x_1, \dots, x_n) = \underset{Vj \in v}{argmax}\ P(V) \prod_{i=1}^{n} P(xi|Vj) \qquad (9)$$

Keterangan :

V_j = Kategori *tweet* j = 1, 2, 3, ..., n. Dimana dalam penelitian ini j1 = kategori *tweet* sentimen negatif, j2 = kategori *tweet* sentimen positif, j3 = kategori *tweet* sentimen netral.

P(xi|Vj) = *Probabilitas* x_i pada kategori Vj

$P(V_j)$ = *Probabilitas* dari Vj

Untuk P(Vj) dan P(xi|Vj) persamaannya adalah sebagai berikut:

$$P(Vj) = \frac{|docs\ j|}{|all\ docs|} \qquad (10)$$

$$P(xi|Vj) = \frac{nk+1}{n+|kosakata|} \qquad (11)$$

Keterangan :

|docs j| = jumlah dokumen setiap kategori j

|all docs| = jumlah dokumen semua kategori

n_k = jumlah frekuensi kemunculan setiap n-gram

n = jumlah frekuensi kemunculan n-gram kata dari setiap kategori

|kosakata| = jumlah semua n-gram kata dari semua kategori

2.3 Teori Umum

2.3.1 Aplikasi Weka

Weka adalah aplikasi data mining *Open Source* berbasis *Java*. Aplikasi ini dikembangan pertama kali oleh Universitas Waikato di Selandia Baru sebelum menjadi bagian dari Pentaho. *Weka* terdiri dari koleksi algoritma *machine learning* yang dapat digunakan untuk melakukan *generalisasi / formulasi* dari sekumpulan data sampling. Walaupun kekuatan *Weka* terletak pada algoritman yang makin lengkap dan canggih, kesuksesan data mining tetap terletak pada

faktor pengetahuan manusia implementornya. Tugas pengumpulan data yang berkualitas tinggi dan pengetahuan pemodelan dan pengguna algoritma yang tepat diperlukan untuk menjamin keakuratan formulasi yang diharapkan.

UNIVERSITAS SAHID SURAKARTA