# ANALISIS TESTIMONIAL WISATAWAN MENGGUNAKAN TEXT MINING DENGAN METODE *NAIVE BAYES* DAN *DECISION TREE*, STUDI KASUS PADA HOTEL – HOTEL DI JAKARTA

**Yonathan Sunoto[1)] dan Budi Wasito[2)]**
1)Alumni Program Studi Sistem Informasi
2) Sstaf Pengajar Program Studi Sistem Informasi
Institut Bisnis dan Informatika Kwik Kian Gie
Jl. Yos Sudarso Kav.87 Sunter Jakarta Utara 14350
http://www.kwikkiangie.ac.id

## ABSTRACT

The ability to express the opinion of lines of text can be extremely useful, and this is a good area to be studied, no doubt because of the possibility of commercial value because most information is now stored as text. In this age of the internet today many reviews, opinions, comments or opinions are so abundant and scattered in internet media in the form of text, thus giving rise to the term or overflow of text that can be used as the object of the new knowledge that is what is called Text Mining. Currently, Text mining is believed to have a high potential commercial value. Text Mining is a process that aims to find the information or the latest trends previously revealed, to process and analyze large amounts of data. In analyzing part or all unstructured text, text mining to try to associate one with the other parts of the text based on certain rules.

Besides text mining is also interpreted as a data mining activities from the data in the form of text or a document, with the aim of searching for words that can represent what is in the document so it can be analyzed in text mining connectedness, In the processing Text Mining conducted prior Tokenizing process, Filtering, Stemming, Tagging and Analyzing. Stages of the process is carried out with the help of tools Semantria. Results semantria process tool is a classification based sentiment analysis. After appearing classification sentiment analysis, the next step was measured by the method of Naive Bayes and Decision Tree. Baselines to generate corresponding processed products is to ensure the characteristics of the data related to the objectives to be achieved from the study.

In the context in the field of Text Mining There are a variety of processing one of which is with Process Mining with a focus on the classification.The processed text mining based on sentiment classification, the region with the sequence that has the highest positive sentiment Central Jakarta (80.7%) and North Jakarta (71.2%), East Jakarta (65.1%), West Jakarta (65% ) and South Jakarta (63.8%).

**Key Words:** *Tags, Text Processing, Tokenizing, Filtering, RapidMiner*

## 1. PENDAHULUAN

Penggunaan Internet saat ini sudah tidak lepas dari aktifitas keseharian.. Apapun yang kita lakukan, pasti menggunakan internet mulai dari keseharian kita, pekerjaan kita bahkan media komunikasi kita sekarang ini tergantung dari internet. Dalam internet mengandung banyak hal dan untuk visualisasinya sendiri disebut sebagai *website*. Website adalah suatu halaman web yang saling berhubungan yang umumnya berada pada suatu tempat yang sama berisikan kumpulan informasi yang disediakan secara perorangan, kelompok, atau organisasi. Sebuah situs web biasanya ditempatkan setidaknya pada sebuah server web yang dapat diakses melalui jaringan seperti Internet, ataupun jaringan wilayah lokal (LAN) melalui alamat Internet yang dikenali sebagai URL. URL adalah singkatan dari Uniform Resource Locator, yaitu rangkaian karakter menurut suatu format standar tertentu, yang digunakan untuk menunjukkan alamat suatu sumber seperti dokumen dan gambar di Internet. Gabungan atas semua situs yang dapat diakses

publik di Internet disebut pula sebagai World Wide Web atau lebih dikenal dengan singkatan WWW. Setiap kita membuka sebuah website, itu tidak terlepas dari hal-hal tersebut.

Halaman web adalah tempat yang "hijau" bagi orang - orang untuk mengekspresikan pendapat mereka dengan topik yang beragam .Bahkan pemberi opini secara profesional, seperti ulasan perjalanan, memiliki *blog* dimana publik dapat mengomentari dan merespon apa yang mereka pikirkan karena seperti yang dijelaskan di awal bahwa internet terhubung dengan dunia. Artinya kita yang telah menggunakan layanan internet, telah terhubung dengan banyak orang dan bisa mencari informasi apa saja dan berinteraksi dengan apa saja. Kemampuan untuk menyatakan pendapat tersebut dari baris-baris teks dapat menjadi sangat berguna, dan ini adalah area yang baik untuk dikaji, tidak diragukan karena kemungkinan nilai komersialnya dikarenakan kebanyakan informasi saat ini disimpan sebagai teks. Di era kemajuan internet saat ini banyaknya ulasan, opini, komentar atau pendapat yang demikian melimpah dan bertebaran di media internet dalam bentuk teks, sehingga memunculkan istilah atau limpahan teks yang bisa dijadikan sebagai objek pengetahuan baru yakni apa yang disebut dengan *Text Mining*. Saat ini, *Text mining* diyakini memiliki potensi nilai komersial tinggi.

Namun, dalam membagi komentar ke dalam kategori-kategori tersebut untuk saat ini masih dilakukan secara manual, artinya dalam mengunggah komentar, kita harus terlebih dahulu mengetahui isi dari komentar yang akan diunggah secara kesuluruhan untuk selanjutnya dimasukkan ke dalam kategori yang tepat. Hal ini sangat merepotkan bagi para calon tamu apabila jumlah hotel yang ingin dicari cukup banyak. Oleh karena itu, perlu adanya sistem dimana sistem tersebut dapat mengklasifikasikan komentar secara otomatis sesuai dengan kategori-kategori sentimen yang ada sehingga bisa membantu para calon tamu dalam mencari hotel terbaiknya. Text mining adalah salah satu teknik yang dapat digunakan untuk melakukan klasifikasi dimana, text mining merupakan variasi dari data mining yang berusaha menemukan pola yang menarik dari sekumpulan data tekstual yang berjumlah besar. Selain klasifikasi, *text mining* juga digunakan untuk menangani masalah *clustering, information extraction*,dan *information retrival* .

*Text mining* adalah salah satu teknik yang dapat digunakan untuk melakukan klasifikasi dokumen dimana t*ext mining* merupakan variasi dari data mining yang berusaha menemukan pola yang menarik dari sekumpulan data tekstual yang berjumlah besar (Feldman & Sanger, 2007).

*Text mining*, mengacu pada proses mengambil informasi berkualitas tinggi dari teks. Informasi berkualitas tinggi biasanya diperoleh melalui peramalan pola dan kecenderungan melalui sarana seperti pembelajaran pola statistik. Proses *text mining* yang khas meliputi :

1. Kategorisasi teks,
2. *Text clustering*, ekstraksi konsep/ entitas,
3. Produksi taksonomi granular,
4. *Sentiment analysis*, penyimpulan dokumen, dan
5. Pemodelan relasi entitas (yaitu, pembelajaran hubungan antara entitas bernama)

*Sentiment analysis* atau *opinion mining* adalah studi komputasional dari opini-opini orang, sentimen dan emosi melalui entitas dan atribut yang dimiliki yang diekspresikan dalam bentuk teks (Liu, 2012). Analisis sentimen akan mengelompokkan polaritas dari teks yang ada dalam kalimat atau dokumen untuk2 mengetahui pendapat yang dikemukakan dalam kalimat atau dokumen tersebut

## 2. TINJAUAN PUSTAKA

### 2.1. Sistem Informasi

Sistem informasi adalah kumpulan dari elemen-elemen yang saling berhubungan ataupun sekelompok komponen yang mengambil (*input*), memanipulasi (*proses*), menyimpan, dan mengeluarkan (*output*) data dan informasi serta menyediakan sebuah reaksi umpan balik (*feedback*) untuk mencapai tujuan sistem[8]. Sistem informasi secara teknis didefinisikan sebagai kumpulan elemen-elemen yang saling berhubungan yang mengumpulkan (atau menerima), mengolah, menyimpan, dan menyebarkan informasi untuk selanjutnya digunakan dalam mendukuk pengambilan keputusan serta mengatur sebuah organisasi[4]. Sistem informasi dapat berupa kombinasi yang terorganisir dari manusia, perangkat keras,

perangkat lunak, jaringan komunikasi, *data resources*, dan kebijakan maupun prosedur untuk menyimpan, mengambil, mengubah, maupun memilah informasi di dalam sebuah organisasi[6].

### 2.2. Data

Data adalah kumpulan fakta mentah ataupun hasil pengamatan, biasanya berupa fenomena fisik ataupun transaksi bisnis[6]. Data dapat diartikan sebagai sebuah aliran fakta-fakta mentah yang mewakili kejadian tertentu yang terjadi di dalam sebuah organisasi ataupun lingkungan fisik sebelum disusun dan diurutkan kedalam sebuah bentuk yang dapat dimengerti dan digunakan oleh orang lain[4].

### 2.3. *Database*

*Database* adalah sebuah kumpulan elemen data yang secara logika saling berhubungan[6].

### 2.4. *Data Warehouse*

*Data warehouse* adalah sebuah *database* besar yang mengumpulkan informasi-informasi bisnis dari berbagai sumber, mencakup segala aspek proses, produk, dan pelanggan yang ada di perusahaan, untuk mendukung manajemen pengambilan keputusan di perusahaan itu[8].

### 2.5. *Data Mining*

*Data mining* adalah sebuah bidang penelitian yang berfokus pada pencarian atau pendefinisian pola-pola pada data. *Data mining* adalah sebuah istilah yang berkaitan dengan penggunaan algoritma-algoritma dan komputer untuk menemukan pola-pola menarik dalam data[9].

### 2.6. Jenis *Data Mining*

Menurut [4] *data mining* dapat dibagi menjadi:

1. Asosiasi (*Association*)

Asosiasi adalah hubungan kejadian-kejadian dengan satu peristiwa tertentu. Misalnya, penelitian dari pola pembelian yang terjadi di suatu supermarket mengungkapkan bahwa ketika keripik jagung terjual, maka enam puluh lima persen dari penjualannya, kola ikut terjual. Tetapi ketika diadakan promosi, maka penjualan kola menjadi delapan puluh lima persen besamaan dengan penjualan keripik jagung. Informasi ini membantu manajer untuk mengambil keputusan yang lebih baik karena mereka telah belajar dari keuntungan berpromosi.

2. Pengurutan (*Sequences*)

Di dalam pengurutan, kejadian-kejadian saling berhubungan sejalan dengan waktu. Sebagai contoh kita sering menemukan ketika seseorang membeli rumah, enam puluh lima persen orang akan membeli kulkas dalam waktu dua minggu, dan empat puluh lima persen orang akan membeli oven dalam kurun waktu satu bulan sejak pembelian rumah.

3. Klasifikasi (*Classification*)

Klasifikasi digunakan untuk mengenal pola-pola dimana suatu data harus dikelompokkan dengan cara memeriksa data-data terdahulu yang sebelumnya telah dikelompokkan berdasarkan syarat-syarat tertentu. Misalnya perusahaan kartu kredit ataupun telekomunikasi khawatir kehilangan pelanggannya. Klasifikasi membantu untuk mengenali mana pelanggan yang berpotensi untuk berhenti sehingga dapat memberikan gambaran untuk membantu manajer memprediksi pelanggan-pelanggan seperti itu. Maka manajer dapat memberikan penawaran khusus untuk mempertahankan pelanggan tersebut.

4. Segmentasi (*Clustering*)

Segmentasi bekerja mirip dengan klasifikasi, hanya saja kelompok data belum ditentukan. Alat *data mining* dapat membantu menentukan perbedaan pengelompokkan dalam data-data seperti menemukan kelompok afinitas untuk kartu bank maupun memisahkan data menjadi kelompok-kelompok pelanggan berdasarkan demografis dan jenis investasi pribadi.

5. Prediksi (*Forecasting*)

Prediksi menggunakan serangkaian nilai-nilai yang sudah ada untuk meramalkan nilai lain. Contohnya prediksi dapat menemukan pola-pola di dalam data untuk dapat membantu para manajer meramalkan nilai masa depan seperti penjualan.

Menurut [3] dalam *data mining* ada pola-pola yang dapat dipahami, yaitu:

1. *Class/Concept Description: Characterization and Discrimination*

Entri data dapat dikelompokkan atau dikonsepkan. Deskripsi ini dapat diterangkan menggunakan:

a. *Data Characterization*

Karakterisasi data adalah hasil kumpulan dari karakteristik umum atau fitur dari target kelompok data.

b. *Data Discrimination*

Diskriminasi data adalah perbandingan fitur umum suatu kelompok data dengan fitur umum objek dari satu atau lebih kelompok lain yang berlawanan.

2. *Mining Frequent Patterns, Associations, and Correlations*

a. *Frequent Patterns*

Pola-pola yang sering muncul di dalam data.

b. *Associations*

Pola dimana sebuah variabel memiliki tingkat keyakinan dengan variabel lainnya dan tingkat pendukung dimana variabel lain memiliki kesamaan.

c. *Correlations*

Tingkat hubungan satu variabel dengan variabel lainnya.

3. *Classification and Regression for Predictive Analysis*

a. *Classification*

Klasifikasi adalah sebuah proses untuk menemukan permodelan (atau fungsi) yang menjabarkan dan membedakan konsep atau kelompok data.

b. *Regression*

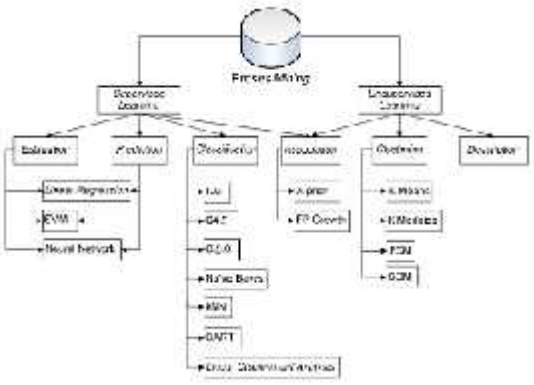
Analisis regresi adalah suatu metodologi statistik untuk memperkirakan hubungan antar variabel yang sering digunakan untuk prediksi angka.

4. *Cluster Analysis*

Analisis klaster adalah pengelompokan analisis objek data tanpa label kelas.

### 2.7. *Process Mining*

Merupakan suatu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data. Beberapa metode yang dapat digunakan berdasarkan pengelompokan data mining dapat dilihat pada Gambar .



Gambar 1.1. Process Mining

### 2.8. Text Mining

*Text Mining* (penambangan teks) adalah suatu proses yang bertujuan untuk menemukan informasi atau tren terbaru yang sebelumnya tidak terungkap, dengan memproses dan menganalisa data dalam jumlah besar. Dalam menganalisa sebagian atau keseluruhan *unstructured text*, *text mining* mencoba untuk mengasosiasikan satu bagian teks dengan yang lainnya berdasarkan aturan-aturan tertentu. Selain itu *text mining* juga diartikan sebagai kegiatan menambang data dari data yang berupa teks atau dokumen, dengan tujuan mencari kata-kata yang dapat mewakili apa yang ada dalam dokumen sehingga dapat dilakukan analisa keterhubungan dalam *text mining* adalah sebagai berikut :

*1. Tokenizing*

Proses ini memotong setiap kata dalam teks, dan mengubah huruf dalam dokumen menjadi huruf kecil. Hanya huruf “a” sampai “z” yang diterima, sedangkan karakter selain huruf dihilangkan. Jadi hasil proses tokenizing adalah kata yang merupakan penyusun kalimat /*string* yang dimasukan.

*2. Filtering*

Pada tahap ini dilakukan proses filter atau penyaringan kata hasil dari proses *tokenizing*, dimana kata yang tidak relevan dibuang. Proses ini menggunakan pendekatan *stoplist*. Yang termasuk *stoplist* adalah "yang", "di","dari" dan lain-lain.

3. *Stemming*

*Stemming* adalah proses untuk menggabungkan atau memecahkan setiap varian-varian suatu kata menjadi kata dasar. *Stem* (akar kata) adalah bagian dari akar yang tersisa setelah dihilangkan imbuhannya (awalan dan akhiran).

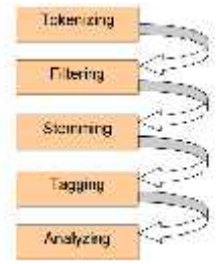
*4. Tagging*

*Tagging* adalah suatu proses mencari bentuk asal dari kata bentuk lampau. Tahap ini tidak digunakan pada teks bahasa karena kata dalam bahasa indonesia tidak mempuyai bentuk lampau.

5. *Analizing*

Pada tahap ini dilakukan proses perhitungan bobot dokumen agar diketahui seberapa jauh tingkat similaritas antara *keyword* yang dimasukan dengan dokumen.

Klasifikasi/ kategorisasi dokumen adalah masalah dalam ilmu informasi yaitu untuk menetapkan dokumen elektronik masuk dalam satu atau lebih kategori, berdasarkan isinya. Tugas klasifikasi dokumen dapat dibagi menjadi dua macam yaitu klasifikasi dokumen terawasi di mana beberapa mekanisme eksternal (seperti *feedback* manusia) memberikan informasi mengenai klasifikasi yang tepat untuk dokumen, dan klasifikasi dokumen tak terawasi, dimana klasifikasi harus dilakukan sepenuhnya tanpa merujuk ke informasi eksternal. Ada juga klasifikasi dokumen semi-diawasi, dimana bagian dari dokumen diberi label oleh mekanisme eksternal.




Gambar 2.1. *Text Processing*

Pendekatan manual *Text Mining* secara intensif dalam laboratorium pertama muncul pada pertengahan 1980-an, namun kemajuan teknologi telah memungkinkan ranah tersebut untuk berkembang selama dekade terakhir. *Text Mining* adalah bidang interdisipliner yang mengacu pada pencarian informasi, pertambangan data, pembelajaran mesin, statistik, dan komputasi linguistik. Pada dasarnya proses kerja dari *Text Mining* banyak mengapdopsi 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). Tahap-tahap *Text Mining* secara umum adalah *text preprocessing* dan *feature selection* (Feldman & Sanger 2007, Berry & Kogan 2010) .

## 2.9. Sentiment Analysis

*Sentiment analysis* atau *opinion mining* mengacu pada bidang yang luas dari pengolahan bahasa alami, komputasi linguistik dan *text mining* yang bertujuan menganlisa pendapat, sentimen, evaluasi, sikap, penilaian dan emosi seseorang apakah pembicara atau penulis berkenaan dengan suatu topik, produk, layanan, organisasi, individu, ataupun kegiatan tertentu (Liu, 2011). Tugas dasar dalam analisis sentimen adalah mengelompokkan teks yang ada dalam sebuah kalimat atau dokumen kemudia menentukan pendapat yang dikemukakan dalam kaliamat atau dokumen tersebut apakah bersifat positif, negatif atau netral (Dehaff, M., 2010). *Sentiment analysis* juga dapat menyatakan perasaan emosional sedih, gembira, atau marah. Kita dapat mencari pendapat tentang produk-produk, merek atau orang-orang dan menentukan apakah mereka dilihat positif atau negatif di web (Saraswati, 2011). Hal ini memungkinkan kita untuk mencari informasi tentang:

(1) Deteksi *Flame* (rants buruk)
(2) Persepsi produk baru.
(3) Persepsi Merek.
(4) Manajemen reputasi.

d. *Sentences Sentiment Classification*

Jika kalimat diklasifikasikan sebagai subjektif, kita menentukan apakah itu mengungkapkan pendapat positif atau negatif.

*Supervised Learning* dapat diterapkan begitu saja untuk beberapa dokumen-tingkat klasifikasi sentimen, dan *Lexicon-based Method*. Sebelum membahas algoritma yang ada (beberapa algoritma tidak menggunakan subjektivitas klasifikasi langkah), mari kita menunjukkan asumsi implisit yang dibuat dalam banyak penelitian pada subjek. Asumsi kalimat-tingkat klasifikasi sentimen: Sebuah kalimat mengungkapkan sentimen tunggal dari pemegang pendapat tunggal.

Menurut Yu dan Hatzivassiloglou (2003) "*For sentiment classification of subjective sentences, used a method similar to that in (Turney, 2002). Instead of using one seed word for positive and one for negative as in (Turney, 2002), this work used a large set of seed adjectives. Furthermore, instead of using PMI, this work used a modified log-likelihood ratio to determine the positive or negative orientation for each adjective, adverb, noun and verb. To assign an orientation to each sentence, it used the average log-likelihood scores of its words. Two thresholds were chosen using the training data and applied to determine whether the sentence has a positive, negative, or neutral orientation. The same problem was also studied in (Hatzivassiloglou and Wiebe, 2000) considering gradable adjectives*."

Untuk klasifikasi sentimen kalimat subjektif, Yu dan Hatzivassiloglou (2003) menggunakan metode yang sama dengan publikasi oleh *Turney* di tahun 2002, karya ini menggunakan set besar kata sifat utama. Selain itu, alih-alih menggunakan *Pointwise Mutual Information* (definisi atau titik informasi timbal balik, adalah ukuran dari asosiasi yang digunakan dalam teori informasi dan statistik. Berbeda dengan *informasi mutual* (MI) yang dibangun berdasarkan PMI, mengacu pada kejadian tunggal, sedangkan MI mengacu pada rata-rata dari semua peristiwa yang mungkin.), karya ini menggunakan modifikasi *log-likehood ratio* untuk menentukan orientasi positif atau negatif untuk setiap kata sifat, kata keterangan, kata benda dan kata kerja. Untuk menetapkan orientasi untuk setiap kalimat, digunakan skor log-kemungkinan rata-rata yang terdapat di kata-kata. Dua ambang dipilih menggunakan data pelatihan dan diterapkan untuk menentukan apakah kalimat memiliki orientasi positif, negatif, atau netral. Masalah yang sama juga belajar untuk mempertimbangkan kata sifat *gradable*.

Orientasi sentimen kalimat ditentukan dengan menjumlahkan nilai orientasi semua kata sentimen dalam kalimat. Sebuah kata positif diberi nilai sentimen dari +1 dan kata negatif diberi nilai sentimen -1. Kata negasi dan kata-kata yang bertentangan (misalnya, tapi dan namun) juga dipertimbangkan. Dalam (Kim dan Hovy, 2004), pendekatan yang sama juga digunakan. Metode kompilasi leksikon sentimen juga serupa. Namun, mereka menentukan orientasi sentimen kalimat dengan mengalikan nilai dari kata sentimen dalam kalimat. Sekali lagi, kata positif diberi nilai sentimen dari +1 dan kata negatif diberi nilai sentimen -1. Para penulis juga bereksperimen dengan dua metode lain menggabungkan nilai sentimen tapi mereka lebih rendah. Dalam (Kim dan Hovy, 2004), digunakan untuk mengidentifikasi beberapa jenis tertentu dari pendapat. Dalam (Nigam dan Hurst 2004), Nigam dan Hurst menerapkan leksikon tertentu dan pendekatan NLP dangkal untuk menilai orientasi sentimen kalimat.

### 2.10. Naive Bayes Classifer (NBC)

*NBC* merupakan salah satu algoritma dalam teknik data mining yang menerapkan teori *Bayes* dalam klasifikasi . Teorema keputusan *Bayes* adalah adalah pendekatan statistik yang fundamental dalam pengenalan pola (pattern recoginition). *Naive bayes* didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara konditional saling bebas jika diberikan nilai *output*. Dengan kata lain, diberikan nilai *output*, probabilitas mengamati secara bersama adalah produk dari probabilitas individu. Dengan memasukkan Persamaan 1 ke Persamaan 2 akan diperoleh pendekatan yang digunakan dalam *NBC*.

### 2.11. Decision Tree

*Decision tree* merupakan salah satu metode klasifikasi yang menggunakan representasi struktur pohon (*tree*) di mana setiap node merepresentasikan atribut, cabangnya merepresentasikan nilai dari atribut, dan daun merepresentasikan kelas. *Node* yang paling atas dari decision tree disebut sebagai *root*. *Decision tree* merupakan metode klasifikasi yang paling populer digunakan. Selain karena pembangunannya relatif cepat,

hasil dari model yang dibangun mudah untuk dipahami.

Pada *decision tree* terdapat 3 jenis node, yaitu:

a. ***Root Node***, merupakan node paling atas, pada node ini tidak ada input dan bisa tidak mempunyai output atau mempunyai output lebih dari satu.

b. ***Internal Node*** , merupakan node percabangan, pada node ini hanya terdapat satu input dan mempunyai output minimal dua.

c. ***Leaf node*** atau ***terminal node*** , merupakan node akhir, pada node ini hanya terdapat satu input dan tidak mempunyai output.

## 3. METODE PENELITIAN

### 3.1. Teknik Pengumpulan Data

Dalam penelitian ini data yang digunakan adalah data sekunder dari komentar – komentar yang diambil dari http://*TripAdvisor*.co.id/.

### 3.2. Metode dan Teknik Pengumpulan Data

*Naive Bayes*

$$P(a_1, a_2, \ldots, a_n \mid V_j) = \prod_i P(a_i \mid v_j) \quad \ldots\ldots (1)$$

$$V_{NB} = \arg\max_{v_j \in V} P(V_j) \prod_i P(a_i \mid v_j) \quad \ldots\ldots (2)$$

dengan

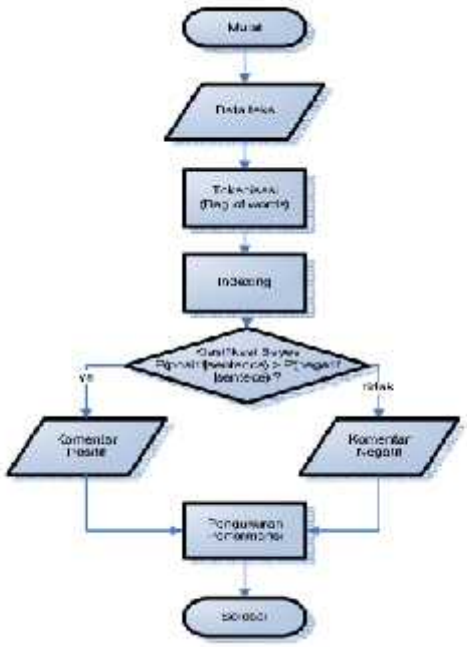
$V_{NB}$ : nilai output hasil klasifikasi Naive Bayes

$P(a_i \mid v_j)$ : rasio antara $n_i/n$, di mana $n_i$ adalah jumlah data *training* untuk $v=v_j$ dan $a=a_i$; dan

n adalah total kemungkinan output

Gambar 3.1 Rumus dasar dari *Naive Bayes*

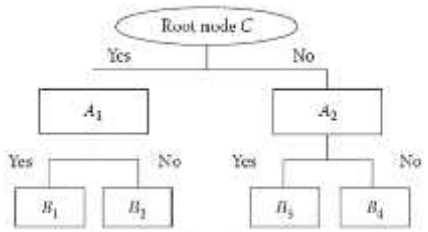
Tahapan proses klasifikasi opini dengan *NBC* ditunjukan dalam diagram seperti berikut :



Gambar 3.2 *Tahapan Text Mining* dengan *Naive Bayes*

*a) Decision Tree*

Struktur sederhana dan dapat ditafsirkan memungkinkan decision tree untuk memecahkan masalah atribut multi-type. Decision tree juga dapat mengelola nilai-nilai yang hilang atau data noise (Dua & Xian, 2011).




Gambar 3.3 Contoh Struktur *Decision Tree*

Gambar 3.3 Contoh Struktur Decision Tree Sumber: Dua & Xian, 2011 Membangun klasifikasi dengan Decision Tree melalui beberapa tahapan sebagai berikut (Larose, 2005):

a. Pertama siapkan *data training* yang biasanya diambil dari data histori atau data masa lampau yang kemudian dibuat ke dalam kelas-kelas tertentu.
b. Menghitung nilai *entropy* yang akan digunakan untuk menghitung nilai *gain* dari

masing-masing atribut sehingga diperoleh atribut dengan nilai *gain* yang tertinggi yang selanjutnya akan digunakan menjadi akar pohon. Entropy adalah suatu parameter untuk mengukur tingkat keberagaman (heterogenitas) dari kumpulan data. Rumus menghitung *entropy* dan *gain* seperti yang ditunjukkan dalam persamaan (1) dan (2).

$$E \quad (S) = \sum_{i=1}^{n} -p \; . \; l \; _{2} . P \qquad (1)$$

Gambar 3=.4 Rumus Menghitung *Entropy*

Keterangan:
S= Himpunan kasus
n = jumlah partisi S
Pi = proporsi Si terhadap S

$$G \quad (S, A) = E \quad (S) - \sum_{i=1}^{n} E \quad (S)$$

Gambar 3.5 Rumus Menghitung *Gain*

Keterangan:
S = Himpunan Kasus
A = Fitur
n = jumlah partisi atribut A
|Si| = Proporsi Si terhadap S
|S| = jumlah kasus dalam S

c. Ulangi terus langkah sebelumnya yaitu menghitung nilai tiap atribut berdasarkan nilai gain yang tertinggi hingga semua record terpartisi.
d. Proses dari Decision Tree ini akan berhenti jika semua record dalam simpul N mendapat kelas yang sama, tidak ada atribut di dalam record yang dipartisi lagi, dan tidak ada record di dalam cabang yang kosong.

## 4. ANALISIS DAN PEMBAHASAN

Pembahasan ini dibuat berdasarkan tahapan dan proses yang terdiri dari :

1. Input Data dari Tripadvisor.co.id ke dalam tabel
2. Penerapan *Text Processing*
3. Pengujian Data dengan metode *Naive Bayes* dan *Decision Tree*
4. Hasil Uji dari metode *Naive Bayes* dan *Decision Tree*
5. Rancangan Prototype *Graphical User Interface* berbasis Web

### 4.1. Jakarta Utara

*Decision Tree* ini menghasilkan suatu angka yang bisa dijadikan batasan untuk standar deviasinya, khususnya untuk *Detected Sentiment* pada hasil perhitungan *Naive Bayes*. Bisa disimpulkan bahwa tamu yang menginap di hotel-hotel daerah Jakarta Utara berada di taraf puas karena memiliki nilai 70,2% untuk sentimen positifnya dari hasil perhitungan menggunakan algoritma *Naive Bayes*. Sedangkan untuk nilai batas bawah dan batas atas pada *Document Sentiment* kali ini mencapai -0,057 ≤ x ≤ 0,221 dengan menggunakan metode *Decision Tree*.

| Sentiment | Naive Bayes | Decision Tree | Memenuhi Syarat? |
|---|---|---|---|
| Positive | 0,476 | >0 ,221 | Yes |
| Negative | - 0,233 | - ≤ 0,057 | Yes |
| Neutral | 0,008 | ≥ 0,221 | No |

### 4.2. Jakarta Pusat

Tamu yang menginap di hotel-hotel daerah Jakarta Pusat berada di taraf sangat puas karena memiliki nilai 80,7% untuk sentimen positifnya dari hasil perhitungan menggunakan algoritma *Naive Bayes*. Sedangkan untuk nilai batas bawah dan batas atas pada *Document Sentiment* kali ini mencapai -0,035 ≤ x ≤ 0,220 dengan menggunakan metode *Decision Tree*. Berikut tabel perbandingannya :

| Sentiment | Naive Bayes | Decision Tree | Memenuhi Syarat? |
|---|---|---|---|
| Positive | 0,469 | >0 ,220 | Yes |
| Negative | - 0,162 | < - 0,035 | Yes |
| Neutral | 0,125 | ≤ 0,220 | Yes |

### 4.3. Jakarta Timur

Tamu yang menginap di hotel-hotel daerah Jakarta Timur berada di taraf yang biasa saja karena memiliki nilai 65,1% untuk sentimen positifnya dari hasil perhitungan menggunakan algoritma *Naive Bayes*. Sedangkan untuk nilai batas bawah dan batas atas pada *Document Sentiment* kali ini

mencapai -0,008 ≤ x ≤ 0,215 dengan menggunakan metode *Decision Tree*

| Sentiment | Naive Bayes | Decision Tree | Memenuhi Syarat? |
|---|---|---|---|
| Positive | 0,451 | > 0 ,215 | Yes |
| Negative | - 0,241 | < - 0,008 | Yes |
| Neutral | 0,121 | < 0,215 | Yes |

### 4.4. Jakarta Selatan

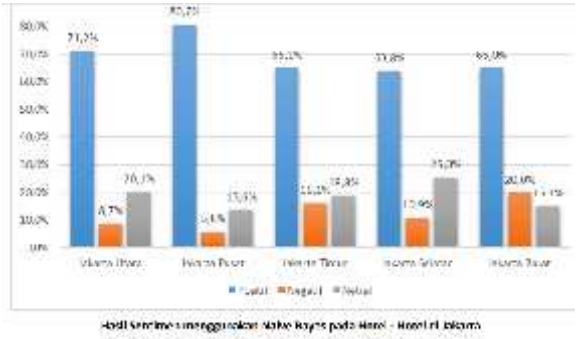
Nilai batas bawah dan batas atas pada *Document Sentiment* kali ini mencapai -0,068 ≤ x ≤ 0,222 dengan menggunakan metode *Decision Tree.* Bisa dikatakan bahwa Hotel – hotel di daerah Jakarta Selatan ini memiliki tingkat kepuasan yang

cukup karena sentimen postifnya berada di angka 63,8%, lebih rendah dari Jakarta Timur yang meraih sentimen positif sebesar 65,1%.

| Sentiment | Naive Bayes | Decision Tree | Memenuhi Syarat? |
|---|---|---|---|
| Positive | 0,472 | > 0 ,222 | Yes |
| Negative | - 0,287 | < - 0,068 | Yes |
| Neutral | 0,091 | < 0,222 | Yes |

### 4.5. Jakarta Barat

**Gambar 4.1**
**Hasil Sentimen dengan *Bootstrap***



Hotel – hotel di daerah Jakarta Barat ini memiliki tingkat kepuasan yang cukup karena sentimen postifnya berada di angka 65%, lebih rendah sedikit dari Jakarta Timur yang meraih sentimen positif sebesar 65,1%.

| Sentiment | Naive Bayes | Decision Tree | Memenuhi Syarat? |
|---|---|---|---|
| Positive | 0,459 | > 0,209 | Yes |
| Negative | - 0,190 | < - 0,050 | Yes |
| Neutral | 0,084 | < 0,209 | No |

### 4.6. *Graphical User Interface*

Penulis mengimplentasikan hasil tersebut kedalam GUI dengan menggunakan *bootstrap*. Bootstrap merupakan *framework* untuk membangun desain web secara responsif. Artinya, tampilan web yang dibuat oleh *bootstrap* akan menyesuaikan ukuran layar dari browser yang kita gunakan baik di *desktop*, *tablet* ataupun *mobile device*. Bootstrap merupakan *framework* untuk membangun desain web secara responsif. Artinya, tampilan web yang dibuat oleh *bootstrap* akan menyesuaikan ukuran layar dari browser yang kita gunakan baik di *desktop*, *tablet* ataupun *mobile device*.

## 5. SIMPULAN

Berdasarkan analisis dan pembahasan sebagaimana yang tertulis pada bab IV ,maka dapat disimpulkan sebagai berikut :

1. Hasil olahan *text mining* berdasarkan klasifikasi sentimen, maka wilayah dengan urutan yang memiliki sentimen positif tertinggi adalah Jakarta Pusat (80,7%) lalu Jakarta Utara (71,2%), Jakarta Timur (65,1%), Jakarta Barat (65%) dan Jakarta Selatan (63,8%).

2. Dengan hasil olahan *Text Mining*, komentar tamu – tamu yang pernah menginap, pengelola hotel dapat memastikan secara cermat bagaimana kondisi penilaian sebenarnya dari para tamu yang pernah menginap.

## 6. REKOMENDASI

Tentunya dalam pembuatan penelitian ini peneliti ingin memberkan rekomendasi bagi peneliti berikutnya, yaitu:

1. Peneliti berikutnya disarankan melakukan olah *Text Mining* dengan

metode selain *Naive Bayes* yaitu dengan metode Support Vector Machine dan kNN.

2. Untuk rancangan GUI bagi peneliti berikutnya, dapat menggunakan php murni yang dapat menampilkan hasilnya tanpa bantuan *tools RapidMiner*, sehingga *user* dapat menegtahui hasilnya.

3. Bagi peneliti berikutnya yang ingin melakukan penelitian di bidang *text mining*, masih banyak tema – tema yang belum dikaji khususnya terkait dengan kuliner atau objek wisata berikut fasilitas yang berada di wilayah Jakarta Utara.

## 7. DAFTAR PUSTAKA


[1] Aston, N., Liddle, J., & Hu, W. (2014). *"Twitter Sentiment in Data Streams with Perceptron"*, Journal of Computer and Communications, 2014.

[2] Asur, S., & Huberman, B. A. (2010, August). *Predicting the future with social media. In Web Intelligence and Intelligent Agent Technology* (WI-IAT), 2010 IEEE/WIC/ACM International Conference on (Vol. 1, pp. 492-499). IEEE.

[3] Berry, Michael Wand & Jacob Kogan (2010), *Text mining: applications and theory*

[4] John Wiley and Sons, Ltd

[5] Bing Liu (2012). *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, May 2012.

[6] Bollen, J., Mao, H., Zeng, X., (2011). *Twitter mood predicts the stock market, Journal of Computational Science 2*, pp.1-8

[7] Bowo Prasetyo. *"Mengenal RapidMiner : Tool Open Source untuk Data Mining"*, http://www.slideshare.net/bowoprasetyo/rapidminer.

[8] Connolly, Thomas dan Carolyn Begg (2014), *Database Systems : Practical Approach to Design, Implementation, and Management*, Edisi ke-6, England:Addison Wesley.

[9] Dua, S. & Xian Du. 2011. *Data Mining and Machine Learning in Cybersecurity*. USA:

[10] Taylor & Francis Group. ISBN-13: 978-1-4398-3943-0

[11] Davidov, D., Tsur, O., & Rappoport, A. (2010, August*). Enhanced sentiment learning using twitter hashtags and smileys. In Proceedings of the 23rd International Conference on Computational Linguistics:* Posters (pp. 241-249). Association for Computational Linguistics.

[12] Dehaff, M. 2010. *Sentiment Analysis, Hard But Worth It!*. [Online]. Tersedia di:

[13] http://www.customerthink.com/blog/sentiment_analysis_hard_but_worth_it

[14] Han, Jiawei and Micheline Kamber (2006), *Data Mining: Concepts and Techniques, 2nd ed.*, The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor

[15] Hurst, M., and Nigam, K. (2004). Retrieving topical sentiments from online document collections. In *Proceedings of the 11th Conference on Document Recognition and Retrieval.*

[16] Morgan Kaufmann Publishers, March 2006. ISBN 1-55860-901-6

[17] Jason Jong (2011). *Predicting Rating with Sentiment Analysis .*, http://cs229.stanford.edu/proj2011/Jong%20PredictingRatingwithSentimentAnalysis.pdf

[18] Kim, Soo-Min & Eduard Hovy (2004), *Determining the Sentiment of Opinions*, Proceedings of the COLING conference, Geneva.

[19] Laudon, Kenneth C., Jane P. Laudon dan Ahmed Elragal (2012), *Sistem Informasi Manajemen : Mengelola Perusahaan Digital*, Edisi ke-12, Jakarta:Salemba Empat.

[20] Larose, D. T. 2005. *Discovering Knowledge in Data*. New Jersey: John Willey & Sons, Inc. ISBN0-471-66657-2.

[21] Linus Philip Lawrence, *Reliability of Sentiment Mining Tools: A comparison of Semantria and Social Mention.*, http://essay.utwente.nl/65302/

[22] Markus Hofmann, Ralf Klinkenberg, "*RapidMiner: Data Mining Use Cases and Business Analytics Applications (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series)*," *CRC Press*, October 25, 2013.

[23] O'Brien, James A. (2010), *Pengantar Sistem Informasi : Perspektif Bisnis dan Manajerial*, Edisi ke-12, Jakarta : Salemba Empat.

[24] Sani Susanto dan Dedy Suryadi (2010), *Pengantar Data Mining : Menggali Pengetahuan Dari Bongkahan Data*, Yogyakarta:Penerbit Andi Ofset Yogyakarta.

[25] Sanger, James and Ronen Feldman, (2006), *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press, Dec 11, 2006

[26] Saraswati, Ni Wayan Sumartini. (2011), "*Text Mining* dengan Metode *Naive Bayes Classifier* dan *Support Vector Machines* untuk *Sentiment Analysis*", Tesis Program Pascasarjana Universitas Udayana, Denpasar.

[27] Sharda, Ramesh., Dursun Delen dan Efraim Turban. (2014), *Business Intelegence : "A Managerial Perspective on Analytics", Third Edition* : Pearson

[28] Tala, F. Z. (2003). *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*. M.S. thesis. M.Sc. Thesis. Master of Logic Project. Institute for Logic, Language and Computation. Universiteti van Amsterdam The Netherlands

[29] Turney, P. (2002) Thumps up or thumbs down? Semantic orientation applied to unsupervisedclassification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*

[30] Yu, H., and Hatzivassiloglou, V. (2003) *Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-03).