INTEGRASI ALGORITMA GENETIK UNTUK SELEKSI FITUR PADA ANALISIS SENTIMEN REVIEW HOTEL MENGGUNAKAN ALGORITMA NAÏVE BAYES

TESIS

14001712 ANDI TAUFIK

PROGRAM PASCASARJANA MAGISTER ILMU KOMPUTER
SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER
NUSA MANDIRI
JAKARTA
2016

INTEGRASI ALGORITMA GENETIK UNTUK SELEKSI FITUR PADA ANALISIS SENTIMEN REVIEW HOTEL MENGGUNAKAN ALGORITMA NAÏVE BAYES

TESIS

Diajukan sebagai salah satu syarat untuk memperoleh gelar
Magister Ilmu Komputer (M.Kom)

14001712 ANDI TAUFIK

PROGRAM PASCASARJANA MAGISTER ILMU KOMPUTER
SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER
NUSA MANDIRI
JAKARTA
2016

SURAT PERNYATAAN ORISINALITAS

Yang bertanda tangan dibawah ini :

Nama	:	Andi Taufik
NIM	:	141001712
Program Studi	:	Magister Ilmu Komputer
Jenjang	:	Stata Dua (S2)
Konsentrasi	:	Management Informasi System (MIS)

Dengan ini menyatakan bahwa tesis yang telah saya buat dangan judul : "Integrasi Algoritma Genetik untuk seleksi fitur pada Analisis Sentimen Review Hotel menggunkan Algoritma Naïve Bayes" adalah hasil karya sendiri , dan semua sumber baik uang dikutip maupun yang dirujuk telah saya nyatakan dengan benar dan tesis belum di terbitkan atau dipublikasikan dimanapun dan dalam bentuk apapun.

Dengan demikian surat pernyataan ini saya buat dengan sebenar-benarnya. Apabila dikemudian hari ternyata saya memberikan keterangan palsu dan ada pihak lain yang mengklaim bahwa tesis yang telah saya buat adalah hasil karya karya milik seseorang atau badan tertentu, saya bersedia diproses baik secara pidana maupun perdata dan kelulusan saya dari Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri dicabut/ dibatalkan.

Jakarta, 12 Agustus 2016

Yang Menyatakan

Andi Taufik

HALAMAN PENGESAHAN

Tesis ini diajukan oleh :

Nama	:	Andi Taufik
NIM	:	141001712
Program Studi	:	Magister Ilmu Komputer
Jenjang	:	Stata Dua (S2)
Konsentrasi	:	Management Informasi System (MIS)
Judul Tesis	:	" Integrasi Algoritma Genetik untuk Seleksi Fitur pada Analisis Sentimen Review Hotel Menggunakan Algoritma Naïve Bayes"

Telah berhasil dipertahankan dihadapan Dewan penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Magister Ilmu Komputer (M.Kom) pada Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri (STMIK Nusa Mandiri).

Jakarta, 27 Agustus 2016
Pascasarjana Magister Ilmu Komputer
STMIK Nusa mandiri
Direktur

Prof. Dr. Ir R. Eko Indrajit,M.Sc, MBA

DEWAN PENGUJI

Penguji I	: Dr. Sfenrianto, M.Kom	.....
Penguji II	: Dr. Windu Gata, M.Kom	.....
Penguji III/ Pembimbing	: Dr. Sularso Budilaksono, M.Kom	.....

STMIK NUSA MANDIRI	**LEMBAR KONSULTASI BIMBINGAN TESIS**
	PASCASARJANA MAGISTER ILMU KOMPUTER STMIK NUSA MANDIRI

- NIM : 14001712
- Nama Lengkap : Andi Taufik
- Dosen Pembimbing : Dr. Sularso Budilaksono
- Judul Tesis : Integrasi Algoritma Genetik untuk Seleksi Fitur pada Analisis Sentimen Review Hotel Menggunakan Algoritma Naïve Bayes

Foto 2x3

NO	Tanggal Bimbingan	Pokok Bahasan	Paraf Dosen Pembimbing
1.	03/06/2016	Pengajuan Judul	
2.	17/06/2016	Pengajuan BAB I	
3.	22/07/2016	Pengajuan BAB II,BAB III dan Revisi BAB I	
4.	05/08/2016	Pengajuan BAB IV, BAB V, Prototype dan Revisi BAB II, BAB III	
5.	09/08/2016	Revisi BAB IV, BAB V dan Prototype	
6.	12/08/2016	Acc Keseluruhan	

Catatan untuk dosen pembimbing

Bimbingan Skripsi
- Dimulai pada tanggal : 03 Juni 2016
- Diakhiri pada tanggal : 12 Agustus 2016
- Jumlah Pertemuan Bimbingan : 6

Disetujui Oleh,
Dosen Pembimbing

[Dr. Sularso Budilaksono, M.Kom]

KATA PENGANTAR

Puji syukur alhamdullillah, penulis panjatkan kehadirat Allah SWT, yang telah melimpahkan rahmat dan karunia-Nya, sehingga pada akhirnya penulis dapat menyelesaikan tesis ini tepat pada waktunya. Di mana tesis ini penulis sajikan dalam bentuk buku yang sederhana. Adapun judul penulisan tesis, yang penulis ambil adalah sebagai berikut: "Integrasi Algoritma Genetik untuk Seleksi Fitur Pada Analisis Sentimen Review Hotel Menggunakan Algoritma Naïve Bayes"

Tujuan penulisan tesis ini dibuat sebagai salah satu syarat utuk mendapatkan gelar Magister Ilmu Komputer (M.Kom) Pada Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri (STMIK Nusa Mandiri).

Penulis melakukan pencarian dan menganalisa berbagai macam sumber referensi, baik dalam bentuk jurnal ilmiah, buku-buku literatur, internet, dan lainlain yang terkait dengan pembahasan pada tesis ini.

Penulis menyadari bahwa tanpa bimbingan dan dukungan dari semua pihak dalam pembuatan tesis ini, maka penulis tidak dapat menyelesaikan tesis ini tepat pada waktunya. Untuk itu ijinkanlah penulis dalam kesempatan ini untuk mengucapkan ucapan terima kasih yang sebesar-besarnya kepada :

1. Bapak Dr. Sularso Budilaksono selaku pembimbing tesis yang telah menyediakan waktu, pikiran dan tenaga dalam membimbing penulis dalam menyelesaikan tesis ini.
2. Kepada kedua Orang Tua dan adik tercinta yang telah memberikan dukungan material dan moral kepada penulis.
3. Retno, Hesti dan teman-teman STMIK Nusa Mandiri
4. Seluruh Staff staff Teknical Support Bsi Group dan ADM BTI
5. Seluruh staff pengajar Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Nusa Mandiri yang telah memberikan pelajaran yang berarti bagi penulis selama menempuh studi.

6. Seluruh staff dan karyawan Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Nusa Mandiri yang telah melayani penulis dengan baik selama kuliah.

Serta semua pihak yang terlalu banyak untuk penulis sebutkan satu persatu sehingga terwujudnya penulisan tesis ini. Penulis menyadari bahwa penulisan tesis ini masih jauh sekali dari sempurna, untuk itu penulis mohon kritik dan saran yang bersifat membangun demi kesempurnaan penulisan karya ilmiah yang penulis hasilkan untuk yang akan datang

Akhir kata semoga tesis ini dapat bermanfaat bagi penulis khususnya dan bagi para pembaca yang berminat pada umumnya.

Jakarta, 12 Agustus 2016

Andi Taufik

Penulis

SURAT PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH

UNTUK KEPENTINGAN AKADEMIS

Yang bertanda tangan dibawah ini, saya:

Nama	:	Andi Taufik
NIM	:	141001712
Program Studi	:	Magister Ilmu Komputer
Jenjang	:	Stata Dua (S2)
Konsentrasi	:	Management Informasi System (MIS)
Jenis Karya	:	Tesis

Demi pengembangan ilmu pengetahuan, dengan ini menyetujui untuk memberikan ijin kepada pihak Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri (STMIK Nusa Mandiri) **Hak Bebas Royalti Non-Eksklusif** ***(Non-exclusive Royalti-Free Right)*** atas karya ilmiah kami yang berjudul : "Integrasi Algoritma Genetik untuk Seleksi Fitur Pada Analisis Sentimen Review Hotel Menggunakan Algoritma Naïve Bayes" beserta perangkat yang diperlukan (apabila ada).

Dengan **Hak Bebas Royalti Non-Eksklusif** ini pihak STMIK Nusa Mandiri berhak menyimpan, mengalih-media atau *bentuk*-kan, mengelolaannya dalam pangkalan data *(database)*, mendistribusikannya dan menampilkan atau mempublikasikannya di *internet* atau media lain untuk kepentingan akademis tanpa perlu meminta ijin dari kami selama tetap mencantumkan nama kami sebagai penulis/pencipta karya ilmiah tersebut.

Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan pihak STMIK Nusa Mandiri, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya ilmiah saya ini.

Demikian pernyataan ini saya buat dengan sebenarnya.

Jakarta, 12 Agustus 2016

Yang menyatakan

Andi Taufik

ABSTRAKSI

Nama	:	Andi Taufik
NIM	:	141001712
Program Studi	:	Magister Ilmu Komputer
Jenjang	:	Stata Dua (S2)
Konsentrasi	:	Management Informasi System (MIS)
Judul Tesis	:	" Integrasi Algoritma Genetik untuk Seleksi Fitur pada Analisis Sentimen Review Hotel Menggunakan Algoritma Naïve Bayes"

Saat ini pengunjung yang menulis pendapat untuk berbagi pengalaman secara online terus meningkat. Setiap pengunjung perlu untuk membuat keputusan saat berlibur sebelum memesan hotel untuk menginap, biasanya menbaca hasil review dari pengunjung sebelumnya, tentunya membutuhkan waktu yang cukup lama apabila membaca review tersebut secara keseluruhan namun, jika hanya sedikit review yang dibaca, informasi yang didapatkan akan bias. Analisa sentimen bertujuan untuk mengatasi masalah ini dengan secara otomatis mengelompokkan review pengguna menjadi opini positif atau negatif . Pengklasifikasi Naïve Bayes adalah teknik *machine learning* yang populer untuk klasifikasi teks, karena sangat sederhana, efisien dan memiliki performa yang baik pada banyak domain. Namun, Naïve Bayes memiliki kekurangan yaitu sangat sensitif pada fitur yang terlalu banyak, yang mengakibatkan akurasi klasifikasi menjadi rendah. Oleh karena itu, dalam penelitian ini digunakan metode pemilihan fitur, yaitu Genetic algorithm agar bisa meningkatkan akurasi pengklasifikasi Naïve Bayes. Penelitian ini menghasilkan klasifikasi teks dalam bentuk review positif atau review negatif dari review hotel yang diambil dari situs www. Tripadvisor.com.Pengukuran berdasarkan akurasi Naive Bayes sebelum dan sesudah penambahan metode pemilihan fitur. Evaluasi dilakukan menggunakan 10 fold cross validation. Sedangkan pengukuran akurasi diukur dengan *confusion matrix* dan kurva ROC. Hasil penelitian menunjukkan peningkatan akurasi Naïve Bayes dari 90.50% menjadi 94.50%.

Kata Kunci :
Analisa sentimen, Review Hotel, Naïve Bayes, Genetic Algorithm, Klasifikasi teks

ABSTRACT

Name	:	Andi Taufik
NIM	:	141001712
Study Of Program	:	Magister Ilmu Komputer
Levels	:	Stata Dua (S2)
Concentration	:	Management Informasi System (MIS)
Titel	:	" Integrasi Algoritma Genetik untuk Seleksi Fitur pada Analisis Sentimen Review Hotel Menggunakan Algoritma Naïve Bayes"

Currently visitors who wrote an opinion to share experiences online continues to increase. Each visitor will need to make a decision while on vacation before ordering a hotel for an overnight stay, usually reading the results of a review of previous visitors, certainly requires quite a long time when reading the review as a whole, however, if just a little review that read, the information obtained will be biased. Sentiment analysis aims to address this problem by automatically classify user review be opinions positive or negative. Pengklasifikasi Naïve Bayes machine learning technique is popular for text classification, because it is very simple, efficient and have good performance in many domains. However, Naïve Bayes has a shortage that is very sensitive on the features too much, resulting in a lower classification accuracy. Therefore, in this study used methods the selection of features, i.e. Genetic algorithm to improve accuracy pengklasifikasi Naïve Bayes. This research resulted in the classification of texts in the form of a positive review or a negative review of a hotel review taken from the website www.Tripadvisor.com. Measurement based on Naive Bayes accuracy both before and after the addition of method selection feature. The evaluation was conducted using a 10 fold cross validation. While the measurement accuracy is measured by the confusion matrix and ROC curves. The results showed an increase in the accuracy of Naïve Bayes from 90.50% to 94,50%

Keywords:
Analysis of sentiment, reviews Hotel, Naïve Bayes, Genetic Algorithm, text Classification

DAFTAR ISI

Halaman

DAFTAR TABEL

Halaman

DAFTAR GAMBAR

Halaman

DAFTAR LAMPIRAN

BAB I
PENDAHULUAN

1.1. Latar Belakang Penulisan

Dengan memanfaatkan perkembangan teknologi informasi dan *website*, Informasi melalui Pengguna jejaringan sosial mengenai *review* hotel menyediakan *review* pengunjung yang digunakan untuk berinteraksi dengan pengunjung lain nya, *platform* digunakan sebagai wadah untuk membuat dan mendengar pendapat pengunjung yang menghasilkan ulasan perjalanan dan jasa perhotelan yang telah dikunjungi pada saat liburan menjadi sumber informasi yang sangat penting bagi pengunjung (Duan at al., 2013). Informasi yang sangat berguna saat ini, karena orang cenderung mencari informasi yang cepat dalam pemesanan. Lebih banyak pengguna yang mencari informasi melalui pendapat orang lain di media sosial, blog dan situs-situs *review*. Pentingnya ulasan hotel sebagai sumber informasi khusus untuk pemesanan hotel (Markopoulos at al., 2015). Memungkinkan para pengelola dunia pariwisata untuk memberikan informasi lebih detail tentang produk pariwisata yang ditawarkan. Banyak orang yang memeriksa pendapat dari pembeli lain sebelum membeli produk untuk membuat pilihan yang tepat. Hotel merupakan salah satu produk pariwisata yang sangat penting untuk dipertimbangkan baik dari segi fasilitas, pelayanan ataupun jarak tempuh perjalanan wisata (Taylor at al., 2013).

Untuk meningkatkan pemasaran dan pemesanan hotel, maka kebutuhan pihak hotel adalah meningkatkan jumlah nilai dan *review* dari media sosial. untuk mendapatkan hasil peringkat maupun total *review* di media sosial memberikan dampak positif pada setiap jumlah rata-rata pemesanan. Selain mendapatkan dampak positif pada transaksi pemesanan, jumlah *review* hotel yang tersedia bisa dijadikan sebagai evaluasi yang dapat meningkatkan kinerja hotel menjadi lebih baik (Suardika, 2016).

Setiap orang perlu untuk membuat keputusan saat berlibur sebelum memesan hotel untuk menginap, biasanya mereka meminta pendapat orang lain, hal ini dapat diperoleh dengan membaca opini atau hasil *review* dari pengalaman pengunjung sebelumnya yang tentunya membutuhkan waktu yang cukup lama.

Terdapat beberapa penelitian yang sudah dilakukan dalam hal pengklasifikasian analisis sentimen terhadap *review yang* tersedia, diantaranya adalah penelitian oleh Duan, Cao, Yu & Levy yang menggunakan algoritma *Naïve Baye*s untuk sentimen analisis kualitas layanan hotel dengan menggunakan dua label klasifikasi yaitu klasifikasi positif dan klasifikasi negatif (Duan at al., 2013) lalu penelitian dari Markopoulos, Mikros, Iliadi & Liontos dimana dalam membuat *classifier sentiment* yang menerapkan *Support Vector Mechines* dengan *fiture Unigram* pada *review* hotel dalam bahasa Yunani modern yang membandingkan dua metodologi yang berbeda. (Markopoulos at al., 2015). Sedangkan penelitian yang dilakukan oleh Suardika, sentimen analisis dilakukan menggunakan metode *Naïve Bayes* yang mencari hubungan peringkat antar hotel pada situs Tripadvisor dengan hasil klasifikasi dalam sentimen positif, sentiment negatif atau sentimen netral. Dengan metode naïve bayes nilai akurasi rata-ratanya adalah 81% dan menghasilkan analisis korelasi membuktikan hipotesis bahwa semakin rendah peringkat hotel, semakin besar persentasi sentimen negatif (Suardika, 2016)

Naïve Bayes merupakan klasifikasi sederhana dan efektif banyak digunakan untuk teknik pengolahan informasi seperti *image recognition*, NPL, pengambilan informasi dan lain-lain (Duan at al., 2013). *Naïve Bayes* sangat sederhana dan efesien, selain kesederhanaan, *Naïve Bayes* adalah *machine learning* yang populer untuk teknik klasifikasi teks, dan memiliki kinerja yang baik di banyak domain dan sebagai algoritma karakteristik. Namun *Naïve Bayes* sebagai klasifikasi yang sangat sederhana dan efesien serta sangat sensitif dalam pemilihan fitur (Chen at al, 2009).

Tingkatan lain yang umumnya ditemukan dalam pendekatan sentimen klasifikasi adalah pemilihan fitur, Pemilihan fitur dapat membuat klasifikasi menjadi lebih baik, efektif dan efisien dengan mengurangi jumlah analisis

identifikasi data atau fitur yang cocok untuk dipertimbangkan dalam proses pembelajaran (Moraes Valiati, & Neto, 2013). Ada dua jenis utama dalam pemilihan fitur dalam *machine learning : wrapper dan filter*. *Wrapper* menggunakan klasifikasi akurasi dari beberapa algoritma sebagai fungsi evaluasi. Salah satu metode *wrapper* yang dapat digunakan di pemilihan fitur adalah *Genetic Algorithm* (GA)

Pada penelitian ini menggunakan pengklasifikasi *Naïves Bayes* dengan algoritma genetika sebagai metode pemilihan fitur pada komentar dari *review* suatu hotel sebagai teknik untuk meningkatkan analisa sentimen.

1.2. Identifikasi Masalah

Permasalahan yang muncul berdasarkan latar belakang dari permasalahan yang diuraikan diatas maka identifikasi masalah dari penelitian ini adalah *Naïve Bayes* yang merupakan algoritma sederhana, efisien dan merupakan teknik *machine learning* yang populer untuk klasifikasi teks, tetapi *Naïve Bayes* masih memiliki kelemahan yang sangat sensitif terhadap jumlah fitur yang terlalu banyak, yang mengakibatkan ketepatan klasifikasi menjadi rendah

1.3. Rumusan Masalah

Rumusan masalah yang diangkat pada penelitian ini adalah untuk melihat apakah terjadi perbedaan atau peningkatan pada akurasi *Naïve Bayes* apabila *Genetic Algorthm* untuk seleksi fitur Pada analisis sentimen *review* hotel diterapkan ?

1.4. Tujuan Penelitian

Tujuan penelitian ini adalah untuk mengetahui seberapa meningkatnya akurasi *Naïve Bayes* dengan menggunakan *Genetic Algorithm* yang diintegrasikan pada analisis sentimen *review* hotel berbahasasa Indonesia

1.5. Manfaat Penetitian

Berdasarkan tujuan penelitian, maka manfaat dari penelitian ini adalah

1. Manfaat dari penelitian ini adalah membantu dalam menggambil keputusan saat ingin melakukan pemesanan hotel yang sesuai dangan

keinginan agar lebih efisien dan efektif dibandingkan jika harus membaca *review* yang memakan waktu cukup lama

2. Memberikan kontribusi keilmuan pada penelitian yang berkaitan dengan analisa sentimen atau *opinion mining* dan menerapkan algoritma *Naïve Bayes* dengan menggunakan pemilihan fitur *Genetic Algoritm* dalam pengklasifikasian *review* atau opini sehingga dapat dijadikan sebagai pemikiran untuk pengembangan teori berikutnya.

1.6. Ruang Lingkup Penelitian

Ruang lingkup penelitian dibatasi pada pengujian algoritma *Naïve Bayes* yang menerapkan Algoritma Genetika untuk seleksi fitur. Ruang lingkup data dilakukan pada *dataset review* hotel. Penelitian ini dilakukan menggunakan perangkat lunak RapidMiner5.3

1.7. Sistematika Penulisan

BAB I Pendahuluan

Membahas mengenai latar belakang, penulisan , identifikasi masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, ruang lingkup penelitian dan sistematika penulisan.

BAB II Landasan Teori/Kerangka Pemikiran

Membahas tentang tinjauan studi tentang teori yang melandasi penelitian yaitu algoritma *Naïve Bayes* dan pemilihan fitur *Genetic Algorithm*. Studi kasus disajikan untuk memeberikan contoh dan langkah mengenai algoritma *Naïve Bayes*.

BAB III Metode Penelitian

Membahas metode pengumpulan data, metode yang diusulkan eksperimen dan pengujian algoritma *Naïve Bayes*, pemilihan fitur

Genetic Algorithm untuk meningkatkan akurasi dalam pengklasifikasian komentar pada review hotel

BAB IV Hasil Penelitian dan Pembahahasan

Membahas hasil penelitian dari pengujian algoritma *Naïve Bayes*, *Genetic Algorithm* sebelum dan sesudah model diterapkan. Membandingkan hasil dari kedua model untuk melihat tingkat akurasi yang paling tinggi

BAB V Penutup

Membahas kesimpulan tentang kekurangan maupun kelebihan dari model yang digunakan dalam penelitian yang dilakukan dan memberikan saran utuk penelitian berikutnya.

BAB II
LANDASAN/KERANGKA PEMIKIRAN

2.1. Tinjauan Pustaka

Tinjauan pustaka yang digunakan dalam penulisan ini yakni menggunakan buku dan jurnal yang berhubungan dengan tema yang dipilih. Menggunakan referensi untuk menjelaskan pengklasifikasi *Naïve Bayes*, dan pembahasan teori lainya.

2.1.1 Data Mining

Menurut (Gorunescu, 2011) Data mining dapat didefenisikan sebagai sebuah proses untuk menemukan pola data.

Menurut Witten (Witten, Frank dan Hall, 2011) *Data mining* merupakan perpaduan dari ilmu statistik, kecerdasan buatan (sistem pakar) dan penelitian dalam bidang *database*, untuk itu diperlukan penyaringan melalui sejumlah besar material data atau melakukan penyelidikan dengan cerdas tentang keberadaan suatu data yang memiliki nilai *Daryl Pregibons*.

2.1.2 Klasifikasi, Validasi dan Evaluasi Algoritma *Data Mining*

Klasifikasi merupakan data baru diklasifikasikan didasarkan pada data *training*. Klasifikasi pada *data mining* untuk memprediksi label *class* dan mengklasifikasi data didasarkan pada data training dan nilai label *class* dalam mengklasifikasikan *atribut* dan menggunakannya saat mengklasifikasikan data baru.

Menurut Han dan Kamber (Han dan Kamber, 2007), langkah dari klasifikasi proses:

1. Data *cleaning*

   Pada tahapan *preprocessing* ini data dihilangkan atau dikurangi kesalahannya dan di *treatment* dari nilai yang kosong
2. *Relevance analysis*

   Menghilangkan atribut-atribut yang tidak *relevan* atau atribut yang redudan
3. Data *transformation* and *reduction*

   Membangun normalisasi data

Menurut Han dan Kamber (Han dan Kamber, 2007), klasifikasi *data mining* memiliki beberapa algoritma, yaitu:

1. *Decision Tree Classification*
2. *Naive Bayes Classification*
3. *Rule-Based Classification*
4. *Neural Network*
5. *Super Vector Machines*
6. *Associative classification*
7. *K-Nearest-Neighbor Classifiers*
8. *Genetic Algorithm*
9. *Rough set Approach*
10. *Fuzzy Set Approaches*

Menurut Gorunescu (Gorunescu, 2011) validasi adalah proses mengevaluasi akurasi prediksi dari suatu model. Ada banyak metode yang digunakan untuk memvalidasi suatu model berdasarkan data yang ada, seperti *holdout, random sub-sampling, cross-validation, stratified sampling, bootstrap*, dan lain sebagainya. *K-fold Cross-validation* merupakan teknik validasi dengan membagi data awal secara acak kedalam *k* bagian yang saling terpisah atau "*fold"* (Han dan Kamber, 2007)

Metode evaluasi klasifikasi menurut Han (Han dan Kamber, 2007);

1. Akurasi, memperkirakan label *class*
2. Kecepatan, waktu untuk membangun model dan waktu dalam menggunakan model
3. Kehandalan, mengatasi *noise* dan *missing values*

Menurut Han (Han & Kamber, 2007) *confusion matrix* adalah alat yang sangat berguna untuk menganalisa seberapa baik pengklasifikasi bisa mengenali *tuple* dari *class* yang berbeda. Dalam *confusion matrix* dikenal beberapa istilah seperti *True positif* yang merujuk pada *tuple positif* yang secara benar dilabeli oleh pengklasifikasi, sementara *True negatif* adalah *tuple negatif* yang secara benar dilabeli oleh pengklasifikasi. Adapula *False positif* yang merupakan *tuple negatif* yang secara tidak benar dilabeli oleh pengklasifikasi, dan *False negatif* yang merupakan *tuple positif* yang secara tidak benar dilabeli oleh pengklasifikasi

Kurva ROC akan digunakan untuk mengukur AUC (*Area Under Curve*). Kurva ROC membagi hasil positif dalam sumbu y dan hasil negatif dalam sumbu x (Witten, Frank, & Hall, 2011). Sehingga semakin besar area yang berada dibawah kurva, semakin baik pula hasil prediksi. Kurva ROC digunakan untuk mengukur nilai *Area Under Curve* (AUC).

Menurut Gorunescu (Gorunescu, 2011) hasil perhitungan dapat divisualisasikan dengan kurve ROC (*Receiver Operating Characteristic*) atau AUC (*Area Under Curve*). Berikut tingkat nilai diagnose dari ROC, yaitu :

a. Nilai AUC 0.90-1.00 = *excellent classification*
b. Nilai AUC 0.80-0.90 = *good classification*
c. Nilai AUC 0.70-0.80 = *fair classification*
d. Nilai AUC 0.60-0.70 = *poor classification*
e. Nilai AUC 0.50-0.60 = *failure*

Kurva ROC memiliki properti yang menarik: mereka tidak sensitif terhadap perubahan distribusi kelas. Jika proporsi positif terhadap kasus negatif berubah dalam satu set tes, kurva ROC tidak akan berubah. Untuk melihat mengapa demikian, dapat dilihat pada *confusion matrix* dalam (Moraes Valiati, & Neto, 2013). Nilai *accuracy* adalah *presentase* dari jumlah *record data* yang diklasifikasikan secara baik dan benar dengan menggunakan sebuah algoritma dan dapat membuat klasifikasi setelah dilakukan pengujian pada hasil klasifikasi tersebut. Nilai *precison* atau yang juga dikenal dengan nama *confidence value* merupakan *proporsi* dari jumlah kasus yang diprediksi mendapatkan hasil positif dimana nilainya juga positif pada data (Han & kamber, 2007) Menurut Powers (Powers, 2011) nilai dari *Recall* atau *sensitivity value* merupakan *proporsi* dari jumlah kasus yang bernilai positif yang sebenarnya dan diprediksi positif secara benar.

Confusion matrix.

	Predicted	
	Positive documents	Negative documents
Actual positive documents	# True Positive samples (TP)	# False Negative samples (FN)
Actual negative documents	# False Positive samples (FP)	# True Negative samples (TN)

Gambar 2.1 *Confusion Matrix*

Berikut ini adalah persamaan model *Confusion Matrix* (Han dan Kamber, 2007)

1. Nilai akurasi (acc) adalah proporsi jumlah predikasi yang benar

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad \text{.....(2.1)}$$

2. *Sensitivity* digunakan untuk membandingkan *proporsi* tp terhadap *tupel* yang positif.

$$Sensitivity = \frac{TP}{TP+FN} \quad \text{.....(2.2)}$$

3. *Specifity* digunakan untuk membandingkan proporsi tn terhdap tupel yang negatif

$$Specifity = \frac{TP}{TN+FP} \quad \text{.....(2.3)}$$

4. PPV (*Positive Predictive Value*) adalah proporsi kasus dengan diagnosa positif

$$PPV = \frac{TP}{TP+FP} \quad \text{.....(2.4)}$$

5. NPV(*negative Predictive Value*) adalah *proporsi* kasus dengan diagnosa negatif

$$NPV = \frac{TN}{TN+FN} \quad \text{.....(2.5)}$$

2.1.3 Text Mining

Menurut Charjan (Charjan dan Pun, 2013). *Text mining* adalah penemuan dari pengetahuan yang menarik pada dokumen teks.Hal ini merupakan tantangan untuk menemukan pengetahuan yang akurat pada teks dokumen untuk menolong pengguna untuk menemukan yang diinginkan. Penemuan pengetahuan dapat menjadi efektif digunakan dan memperbaharui pola penemuan dan menerapkannya ke *text mining*.

Menurut Bramer (Bramer, 2007) teks merupakan sesuatu yang umum dalam melakukan pertukaran informasi. Syarat umum data dan *teks mining* adalah

informasi yang diambil dan dapat menjadi data yang berguna.*Text mining* merupakan proses menganalisa teks untuk menjadi informasi yang berguna untuk tujuan tertentu. Informasi yang diambil harus jelas dan *eksplisit*, karena *text mining* merubah menjadi bentuk yang dapat digunakan oleh *computer* atau orang yang tidak memiliki waktu untuk membaca *full teks*. *Text mining* merupakan variasi dari data *mining* yang berusaha menemukan pola yang menarik dari sekumpulan data tekstual yang berjumlah besar (Feldman dan sanger, 2007).

2.1.4 Sentimen Analisis

Menurut Medhat (Medhat et al., 2014) Analisis sentimen adalah suatu bidang yang sedang berlangsung dalam penelitian berbasiskan teks. Analisis sentimen atau *opinion mining* adalah kajian tetang cara untuk memecahkan masalah dari opini masyarakat, sikap dan emosi suatu entitas, dimana entitas tersebut dapat mewakili individu, peristiwa atau topik.

Menurut Kontopoulus (Kontopoulus et al., 2013) *Opinion mining* atau juga dikenal sebagai analisa sentimen adalah proses yang bertujuan untuk menentukan apakah polaritas kumpulan teks tulisan (dokumen, kalimat, paragrap, dll) cenderung ke arah positif, negatif, atau netral.

Menurut Moraes (Moraes, Valiati dan Neto, 2013) langkah-langkah yang biasa dilakukan dalam analisis sentimen klasifikasi text

1. Definisikan domain set

   Pengumpulan dataset yang melingkupi suatu domain, misalnya *dataset review* restoran, dataset *review* film, dataset *review* produk, dan lain-lain.

2. *Pre-Processing*

   Tahap pemprosesan awal yang umumnya dilakukan dengan proses *tokenization*, *stopwords removal* dan *stemming*.

3. *Transformation.*

   Proses representasi angka yang dihitung dari data tekstual. *Binary representation* yang umumnya digunakan dan hanya menghitung kehadiran atau ketidakhadiran sebuah kata di dalam dokumen. Berapa kali sebuah kata muncul di dalam suatu dokumen juga digunakan sebagai skema pembobotan dari data tekstual. Proses yang umumnya digunakan yaitu TF-IDF, *Binary transformation*, dan *Frequency transformation*.

4. *Feature Selection*

   Pemilihan fitur (*feature selection*) bisa membuat pengklasifikasi lebih efisien/efektif dengan mengurangi jumlah data untuk dianalisa dengan mengidentifikasi fitur yang *relevan* yang selanjutnya akan diproses. Metode pemilihan fitur yang biasanya digunakan adalah *Expert Knowledge*, *Minimum Frequency, Information gain*, *Chi-Square*, dan lain sebagainya.

5. *Classification*

   Proses klasifikasi umumnya menggunakan pengklasifikasi seperti *Naïve Bayes*, *Support Vector Machine*, dan lain sebagainya.

6. *Interpretation/Evaluation*

   Tahap evaluasi biasanya menghitung akurasi, *recall* dan *precision*.

2.1.5 Review

Pendapat dan pengalaman yang ditulis oleh wisatawan pada *platform* web lain pada saat liburan. Hotel merupakan salah satu produk pariwisata yang sangat penting untuk dipertimbangkan baik dari segi fasilitas, pelayanan ataupun jarak tempuh perjalanan wisata. Saat ini sudah banyak website wisata yang menyediakan fasilitas untuk pengguna internet menuliskan opini dan pengalaman pribadinya secara *online.* Banyak orang yang memeriksa pendapat dari pembeli lain sebelum membeli produk untuk membuat pilihan yang tepat. Yang memungkinkan para pengelola dunia pariwisata untuk memberikan informasi lebih detail tentang produk pariwisata yang ditawarkan (Taylor et al, 2013). Tinjauan pengguna *online* telah meledak dalam beberapa tahun terakhir, *merevolusi* industri hotel. Ulasan wisata dari konsumen lain mempengaruhi setengah dari semua keputusan pembelian hotel (Duan et al, 2013).

2.1.6 Pre-Processing

Menurut Haddi (Haddi, Liu dan Shi, 2013) *Preprocessing* data adalah proses pembersihan dan mempersiapkan teks untuk klasifikasi . Seluruh proses melibatkan beberapa langkah: membersihkan teks *online*, penghapusan ruang *spasi*, memperluas singkatan, kata dasar (*stemming)*, penghapusan kata henti *(stopword removal)*, penanganan negasi dan terakhir seleksi fitur.

N-gram didefinisikan sebagai sub-urutan n karakter dari kata diberikan. Misalnya, "*mountain*" dapat diwakili dengan *character* n-gram yang ditunjukkan pada tabel berikut (Gencosman, Ozmutlu dan Ozmutlu, 2014)

2.1.7 TF-IDF (*Term Frequency- Inverse Document Frequency*)

Menurut Robertson (Robertson, 2004) Metode *TF-IDF* merupakan suatu cara untuk memberikan bobot hubungan suatu kata (*term*) terhadap dokumen. Metode ini menggabungkan dua konsep untuk perhitungan bobot yaitu, frekuensi kemunculan sebuah kata didalam sebuah dokumen tertentu dan *inverse* frekuensi dokumen yang mengandung kata tersebut. Frekuensi kemunculan kata didalam dokumen yang diberikan menunjukkan seberapa penting kata tersebut didalam dokumen tersebut. Frekuensi dokumen yang mengandung kata tersebut menunjukkan seberapa umum kata tersebut. Sehingga bobot hubungan antara sebuah kata dan sebuah dokumen akan tinggi apabila frekuensi kata tersebt tinggi didalam dokumen dan frekuensi keseluruhan dokumen yang mengandung kata tersebut yang rendah pada kumpulan dokumen (*database*). Metode TF-IDF merupakan metode untuk menghitung bobobt setiap kata yang paling umum digunakan pada *information retrieval*. Metode ini juga terkenal efisien, mudah dan memiliki hasil yang akurat (Feldman dan Sanger, 2007). Metode ini akan menghitung nilai *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) pada setiap kata di setiap dikomen dalam korpus.

Rumus umum untuk pembobotan TF-IDF (Robertson, 2004) :

$$W = tf * idf \quad \text{.....(2.6)}$$

$$W = tf * log(\frac{N}{df}) \quad \text{.....(2.7)}$$

Keterangan :

W : bobot kata/ *term* terhadap dokumen d

Tf : jumlah kemunculan kata/ *term* t dalam dokumen d

N : Jumlah semua dokumen dalam *database*

Df : Jumlah dokumen yang mengandung kata/ *term* t

d : Dokumen ke- d

Idf : *Inversed Document Frequency* Nilai idf didapat dari :

$$idf = log(\frac{N}{df})$$

Berdasarkan rumus (2.7), berapapun besarnya nilai tf, apabila N = df dimana sebuah kata/*term* muncul di semua dokumen, maka akan didapatkan hasil 0 (nol) untuk perhitungan idf, sehingga perhitungan bobotnya diubah menjadi sebagai berikut:

$$W = tf * (log\left(\frac{N}{df}\right) + 1) \quad \text{.....(2.8)}$$

Rumus (2.8) dapat dinormalisasi dengan rumus (2.9) dengan tujuan menstandarisasi nilai bobot (wtd) ke dalam interval 0 s.d. 1, seperti yang ditulis oleh Intan (Intan dan Defeng, 2006):

$$W = \frac{tf*(log\left(\frac{N}{df}\right)+1)}{\sqrt{\sum_{k=1}^{t}(tf)^2*\left(log\left(\frac{N}{df}\right)+1\right)^2}} \quad \text{.....(2.9)}$$

2.1.8 Pemilihan Fitur

Menurut Maimon (Maimon dan Rokach, 2010) Seleksi fitur adalah untuk mengidentifikasi beberapa fitur dalam kumpulan data yang sama pentingnya, dan membuang semua fitur lain seperti informasi yang tidak *relevan* dan berlebihan. Proses seleksi fitur mengurangi dimensi dari data dan memungkinkan algoritma *learning* untuk beroperasi lebih cepat dan lebih efektif. *Feature selection* digunakan untuk menghilangkan fitur yang tidak *relevan* dan berlebihan, yang mungkin menyebabkan kebingungan dengan menggunakan metode tertentu (Gorunescu, 2011). Menurut Koncz (Koncz dan Paralic, 2011) Metode seleksi fitur memiliki peran penting dalam analisis sentimen, sama seperti dalam tugas *text mining* lainnya. Penggunaan yang tepat dari metode seleksi *fitur* membantu dalam memahami *atribut* yang *relevan* untuk kelas tertentu, serta meningkatkan akurasi klasifikasi. Menurut John, Kohavi, dan Pfleger dalam Chen (Chen at al,

2009) ada dua jenis metode seleksi fitur dalam pembelajaran *machine learning*, yaitu itu *wrappers* dan *filters*. Menurut Koncz (Koncz dan Paralic, 2011) Penggunaan yang tepat dari metode seleksi fitur membantu juga memahami *atribut* yang *relevan* untuk kelas tertentu, serta meningkatkan akurasi klasifikasi. Seleksi *fitur* mempengaruhi beberapa aspek yaitu pola klasifikasi, akurasi klasifikasi, waktu yang diperlukan untuk pembelajaran fungsi klasifikasi, jumlah contoh yang dibutuhkan untuk pembelajaran dan biaya yang terkait dengan *fitur* menurut Yang dan Honavar dalam (Zhao et al., 2011). Seleksi *fitur* merupakan proses optimasi untuk mengurangi satu set besar *fitur* besar sumber asli agar subset *fitur* yang relatif kecil yang signifikan untuk meningkatkan akurasi klasifikasi cepat dan efektif.

Menurut Chen (Chen at al, 2009) *Wrapper* menggunakan klasifikasi akurasi dari beberapa algoritma sebagai fungsi evaluasi. Menurut Gunal, Salah satu metode *wrapper* yang dapat digunakan di pemilihan *fitur* adalah *genetic algorithm* (GA). Menurut Kohavi dan John dalam wibowo, *wrapper* mengusulkan cara sederhana dan ampuh untuk mengatasi masalah seleksi variabel, terlepas dari *mechine learning* yang dipilih. Bahkan, *machine learning* dianggap sebagai *black box* yang sempurna dan metode sesuai untuk penggunaan *off-the-shelf machine learning.*

Menurut Gunal (Gunal,2012) Salah satu metode *wrapper* yang biasa digunakan dalam pemilihan *fitur* adalah *Genetic algorithm* (GA). *Genetic algorithm* mudah disejajarkan dan telah digunakan untuk klasifikasi seperti masalah optimasi lainnya. Dalam *data mining*, *genetic algorithm* dapat digunakan untuk mengevaluasi *fitness* algoritma lainnya

1. Geneti Algoritma

Menurut Goldberg dalam indriyani GA adalah metode pencarian acak, efektif menjelajahi ruang pencarian besar. Menurut Harb (Harb dan Desuky, 2013) *Genetic Algorithm* (GA) adalah algoritma pencarian berdasarkan prinsip-prinsip seleksi alam dan genetika. Dasar-dasar pendekatan algoritma genetik yang dikerahkan untuk memecahkan berbagai masalah. GA beroperasi pada populasi solusi potensial yang menerapkan prinsip hidup untuk menghasilkan ketepatan yang lebih baik dan pendekatan GA berusaha memecahkan solusi masalah yang

lebih baik. Pada setiap generasi, pendekatan set baru dibuat dengan proses pemilihan individu sesuai dengan tingkat nilai *fitness* dalam domain masalah dan pembiakan mereka bersama-sama menggunakan operator dari proses genetik yang dilakukan, yaitu *crossover* dan mutasi. Proses ini mengarah ke *evolusi* populasi individu yang lebih disesuaikan dengan lingkungan mereka daripada individu-individu yang diciptakan seperti yang terjadi secara adaptasi alami.

Menurut Zukhri (Zukhri, 2014) Optimasi adalah proses menyelesaikan suatu masalah tertentu supaya berada pada kondisi yang paling menguntungkan dari suatu sudut pandang. Masalah yang harus diselesaikan berkaitan erat dengan data-data yang dapat dinyatakan dalam satu atau beberapa variabel. Pengertian menguntungkan, biasanya berhubungan dengan pencarian nilai minimum atau pencarian nilai maksimum, bergantung pada sudut pandang yang digunakan. Algoritma Genetika merupakan salah satu algoritma optimasi, yang diciptakan untuk meniru beberapa proses yang diamati dalam evolusi alam. Algoritma Genetika juga merupakan algoritma *stochastic* yang kuat berdasarkan prinsip-prinsip seleksi alam dan natural genetik yang cukup berhasil diterapkan dalam masalah *machine learning* dan optimasi (Guo, Wang & Han, 2010).

Menurut Singh (Singh dan Kakkar, 2015) Algoritma genetika (GAs) adalah algoritma pencarian berdasarkan ide-ide evolusi seleksi alam dan genetik. Dapat dikatakan tepat bahwa algoritma genetik adalah penemuan alam itu sendiri. Dinamakan kelangsungan hidup terkuat oleh teori Charles Darwin, yang digunakan untuk mengembangkan algoritma genetik GA. Pertama kali disajikan oleh J.H. Holland di 1875. *Genetic algorithm* pada dasarnya menunjukkan cara yang cerdas untuk mengeksploitasi dari pencarian acak didefinisikan ruang cari untuk memecahkan masalah.

Menurut Suyanto (Suyanto, 2005) ada beberapa komponen dalam algoritma genetika, yaitu :

1. Inisialisasi poulasi awal merupakan suatu metode untuk menhasilkan kromosom-kromosom awal. Jumlah individu pada populasi awal merupakan masukan dari pengguna. Setelah jumlah individu pada populasi awal ditentukan, dilakukan inisialisasi terhadap kromosom yang terdapat pada

populasi tersebut. Inisialisasi dilakukan secara acark, namun tetap memperlihatkan domain solusi dan kendala permasalahan yang ada.

2. Fungsi evaluasi dalam algoritma genetika merupakan sebuah fungsi yang memberikan penilaian kepada kromosom (*Fitnes value*) untuk dijadikan suatu acuan dalam mencapai nilai optimal pada algoritma genetika Nilai *Fitness* ini kemudian menjadi nilat bobot suatu kromsom. Pada evolusi alam individu yang bernilai *fitness* tinggi yang akan bertahan hidup. Sedangkan individu yg bernilai *fitness* rendah akan mati. Pada masalah *optimasi*, jika solusi yang akan dicari adalah memaksimalkan fungsi h sehingga nilai *fitness* yang digunakan adalah nilai dari fungsi h tersebut, yakni *f*=h (dimana f adalah nilai *fitness*). Tetapi jika maslaahnya adalah meminimalkan fungsi h (masalah minimasi), maka fungsi h tidak bisa digunakan secara langsung. Hal ini disebabkan adanya aturan bahwa individu yang memiliki nilai fitness tinggi lebih mampu bertahan hidup pada generasi berikutnya. Oleh karena itu nilai *fitness* yang bisa digunakan adalah *f=1/h*, yang artinya semakin kecil nilai h, semakin besar nilai *f*. tetapi hal ini akan menjadi masalah jika h bisa bernilai 0, yang mengakibatkan *f* bisa bernilai tak hingga. Untuk mengatasinya, h perlu ditambah sebuah bilangan yang dianggap kecil [0-1] sehingga nilai *fitness*nya menjadi :

$$Fitness = 1/(1+a) \quad \text{….....(2.10)}$$

Dengan a adalah bilangan kecil dan bervariasi [0-1] sesuai dengan masalah yang akan diselesaikan. Oleh karena itu fungsi fitness menjadi maslaah atau penentu utama keberhasilan algoritma genetika.

3. Seleksi adalah proses di mana genom individu yang dipilih dari populasi dan dievaluasi sesuai dengan fungsinya fitness yang didefinisikan. Lebih cocok akan kromosom lebih adalah kesempatan untuk bertahan hidup atau untuk dipilih. Pada tahap ini menentukan nilai total nilai *fitness* menjadi :

$$total\ Fitnes = nilai\ fitnes\ k1 + nilai\ fitnes\ ke..n \quad \text{…….....(2.11)}$$

Keterangan

k : Nilai kromosom

n : Jumlah nilai kromosom yang telah ditentukan

Menentukan nilai probabilitas masing- masing kromosom

$$Probabilitas = \frac{Nilai\ Fitnes\ 1}{Total\ Fitnes} + \frac{Nilai\ Fitnes\ 1}{Total\ Fitnes} + n \quad \text{.....(2.12)}$$

Menentukan kumulatif probabilitasnya dari setiap kromosom :

$$Probabilitas\ kromosom\ 1 + Probabilitas\ kromosom \quad \text{.....(2.13)}$$

4. Pindah silang atau *crossover* adalah sebuah proses yang membentuk kromosom baru dari dua kromosom induk dengan menggabungkan bagian informasi dari masing–masing kromosom. Crossover menghasilkan kromosom baru yang disebut kromosom anak (offspring). Crossover bertujuan untuk menambah keanekaragaman string dalam satu populasi dengan penyilangan antar string yang diperoleh dari reproduksi sebelumnya. Pindah silang juga berakibat buruk jika ukuran populasinya sangat kecil. Dalam suatu populasi yang sangat kecil, suatu kromosom dengan gen-gen yang mengarah ke solusi akan sangat cepat menyebar ke kromosom-kromosom lainya. Untuk mengatasi masalah ini digunakan suatu pindah silang hanya bisa dilakukan dengan probabilitas tertentu (probabilitas crossover). Artinya pindah silang bisa dilakukan hanya jika suatu bilangan random [0- 1] yang dibangkitkan kurang dari probabilitas crossover (Pc) yang ditentukan. Pada umumnya Pc diset mendekati 1, misalnya 0,8. Probabilitas crossover (Pc) bertujuan untuk mengendalikan operator crossover. Jika n adalah banyaknya string pada populasi, maka sebanyak (Pc) x n string akan mengalami crossover. Semakin besar nilai (Pc), semakin cepat pula string baru muncul dalam populasi. Dan juga jika (Pc) terlalu besar, string yang merupakan kandidat solusi terbaik mungkin dapat hilang lebih cepat pada generasi berikutnya.
5. Mutasi merupakan proses mengubah secara acak nilai dari satu atau beberapa gen dalam suatu kromosom. Mutasi adalah operator algoritma genetika yang bertujuan untuk membentuk individu-individu yang baik atau memiliki kualitas diatas rata-rata. Selain itu mutasi dipergunakan untuk mengembalikan kerusakan materi genetik akibat proses *crossover*.

Pada mutasi terdapat satu parameter yang sangat penting, yaitu probabilitas mutasi (Pm) yang bertujuan untuk mengendalikan operator mutasi. Probabilitas mutasi didefinisikan sebagai persentasi dari jumlah total gen dalam populasi yang akan mengalami mutasi. Disetiap generasi diperkirakan terjadi mutasi sebanyak (Pm) x n string. Pada seleksi alam murni, mutasi jarang sekali muncul sehingga probabilitas mutasi yang digunakan umumnya kecil, lebih kecil dari probabilitas *crossover*.

Probabilitas Mutasi (Pm) biasanya diset antara [0-1], misalnya 0.1. Misalkan *offspring* yang terbentuk adalah 100 dengan jumlah gen setiap kromosom adalah 4 dan peluang mutasi adalah 0.10, maka diharapkan terdapat 40 kromosom dari 400 gen yang ada pada populasi tersebut akan mengalami mutasi.

2.1.9 Naïve Bayes

Algoritma Naïve Bayes merupakan suatu bentuk klasifikasi data dengan menggunakan probabilitas dan statistik. Metode ini pertama kali dikenalkan oleh ilmuan inggris Thomas Bayes, yaitu digunakan untuk memprediksi peluang yang terjadi dimasa depan berdasarkan pengalaman dimasa sebelumnya sehingga dikenal sebagaiTeorema bayes (Bramer, 2007)

Menurut Han (Han dan Kamber, 2007) tahapan dalam algoritma Naïves Bayes:

a. Perhatikan D adalah record training dan ketetapan label-label kelasnya dan masing-masing record dinyatakan n atribut (n field) $X = (X_1, X_2 \ldots., X_n)$

b. Misalkan terdapat m kelas $C_1, C_2 \ldots., C_m)$

c. Klasifikasi adalah diperoleh maximum posteriori yaitu maximum P(Ci|X)

d. Ini diperoleh dari teorema Bayes

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad \text{.....(2.14)}$$

Karena P(X) adalah konstan untuk semua kelas, hanya perlu dimaksimalkan.

$$P(C_i|X) = P(X|C_i)P(C_i) \quad \text{.....(2.15)}$$

Sebagai contoh, misalkan seperti tabel 2.1

Tabel 2.1 Contoh Data *Training* Dari *ALL Electronics Costumer Database*

Id	*Age*	*Income*	*Student*	*Create _rating*	*Class* *Buy_computer*
1	<=30	*High*	No	Fair	No
2	<=30	*High*	No	excellent	No
3	31..40	*High*	No	Fair	Yes
4	>40	*Medium*	No	Fair	Yes
5	>40	*Low*	Yes	Fair	Yes
6	>40	*Low*	Yes	excellent	No
7	31..40	*Low*	Yes	excellent	Yes
8	<=30	*Medium*	No	Fair	No
9	<=30	*Low*	Yes	Fair	Yes
10	>40	*Medium*	Yes	Fair	Yes
11	<=30	*Medium*	Yes	excellent	Yes
12	31..40	*Medium*	No	excellent	Yes
13	31..40	*High*	Yes	Fair	Yes
14	>40	*Medium*	No	excellent	No
15	<=30	*Medium*	Yes	fair	?

b. Terdapat dua class dariklasifikasi yang dibentuk:

C1= *buys_computer =yes*

C2= *buys_computer= no*

Misalterdapat data X (belumd iketahui *class*-nya)

X=(*age*="<=30",*income="medium",student="yes",credit_rating="fair*")

Penyelesaian :

DibutuhuntukmemaksimalkanP(X|Ci) P(Ci) untuki=1,2

c. P(Ci) merupakan *prior probability* untuksetiap *class* bedasar data contoh:

- P(*buys_computer=”yes”*) = 9/14=0.643
- P(*buys_computer=”no”*) = 5/14=0.357

HitungP(X|Ci), untuki=1,2

- P(*age* = ”<30” | *buys_computer = “yes”*) = 2/9=0.222
- P(*age* = ”<30” | *buys_computer = “no”*) = 3/5 =0.600
- P*(income = “medium”| buys_computer = “yes”*) = 4/9 = 0.444
- P(*income = “medium”| buys_computer = “no”*) = 2/5 = 0.400
- P(*student= “yes” | buys_computer = “yes*”) = 6/9 = 0.667
- P(*student= “yes” | buys_computer = “no”*) = 1/5= 0.200
- P(*credit_rating = “fair” | buys_computer = “yes”*) = 6/9 = 0.667
- P(*credit_rating = “fair” | buys_computer = “no”*) = 2/5 = 0.400
- P(X | *buys_computer=”yes”*)
  = 0.222 x 0.444 x 0.677 x 0.677 = 0.044
- P(X | *buys_computer=”no*)
  = 0.600 x 0.400 x 0.200 x 0.400 = 0.019

d. *Probabilitas Posterior* :

- P(X|*buys_computer=”yes”)*P(*buys_computer=”yes”*)
  = 0.044 x 0.643 = 0.028
- P(X|*buys_computer=”no*) P(*buys_computer=”no”*)
- = 0.019 x 0.357 = 0.007

Kesimpulan :*buys_computer =”yes”*

2.2. Tinjauan Studi

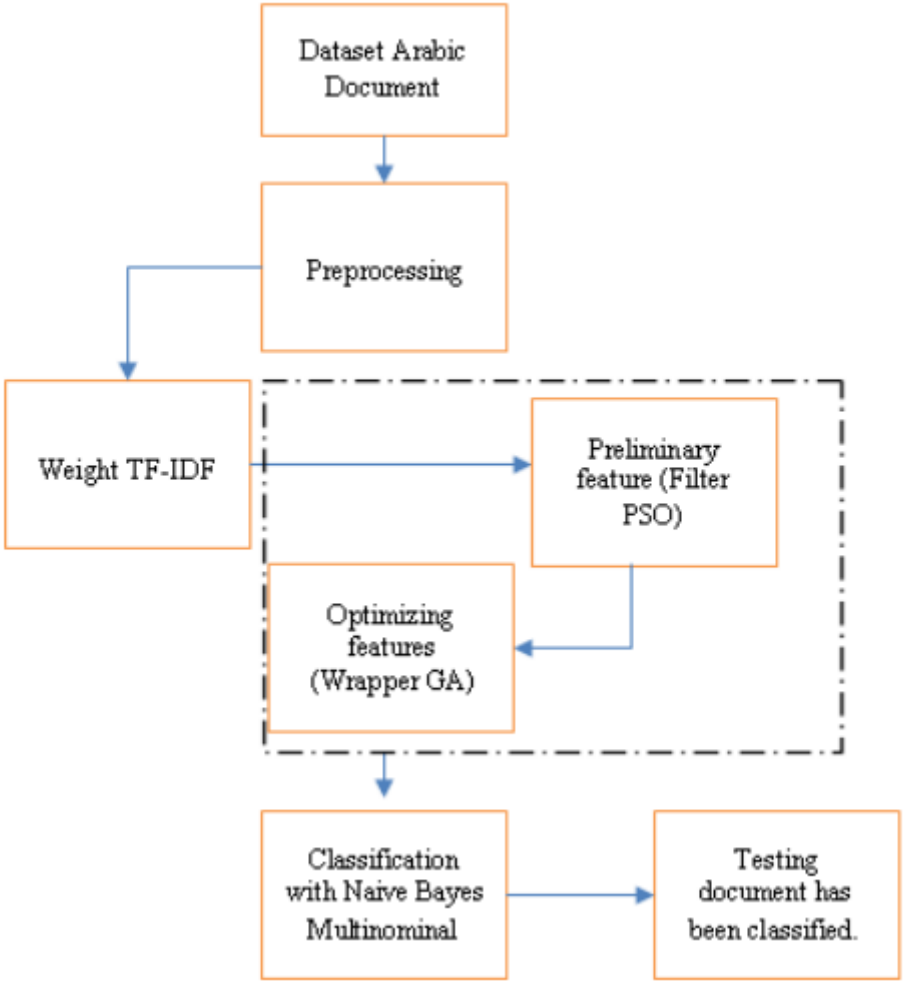
Ada beberapa penelitian yang menggunakan algoritma Naïve Bayes sebagai pengklasifikasi, dan Genetic Algorthm sebagai seleksi fitur dalam klasifikasi teks analisa sentiment pada review, dantaranya :

2.2.1 Model Penelitian Indriyani, Wawan Gunawan dan Ardhon Rakhmadi

Penelitian yang dilakukan oleh Indriyani, Gunawan dan Rakhhmadi (2015). Melakukan penelitian pengklasifikasian korpus dalam Bahasa arab dengan menggunakan Naïve Bayes. Data yang di ambil dari http://www.shamela.ws

menggunakan 478 dokumen terdiri dari tujuh kategori yang menurut paling sering dibahas adalah sholat, Zakat, Shaum, Haji, mutazawwij, Baa'aand Isytaro dan Wakaf. Hasil klasifikasi kedalam tujuh kategori 70% untuk data pelatihan dan 30% data untuk pengujian.

Langkah-langkah yang dilakukan terdiri dari proses *prepocessing*, setiap dokumen di konversi menjadi dokumen berita dengan urutan *Filtering* yaitu penghapusan karakter illegal (angka dan simbol) isi dokumen , penghapusan *stoplist* yaitu penghapusan karakter termasuk dalam kategori *stopword* atau kata-kata dari setiap dokumen diproses dan disusun menjadi vektor istilah yang mewakili dokumen. *Stemming*, yaitu untuk mengembalikan bentuk dasar dari setiap istilah yang ditemukan dalam *vektor* dokumen dan pengelompokan berdasarkan ketentuan yang sama. TF-IDF bobot yaitu melakukan pembobotan TF-IDF pada setiap hal yang ada dalam vektor dokumen.setelah tahap *prepocessing* tahap berikutnya adalah *filtering* dengan menggnakan *Particle Swarm Optimization* (PSO)**.** Setelah tahap penyaringan dilakukan maka tahap berikutnya adalah *wrapper* dengan menggunakan algoritma genetika. Dan langkah yang terakhir yaitu klasifikasi dengan menggunakan *Naïve Bayes*. Gambar 2.2 Menunjukan model yang diusulkan oleh Indriyani.



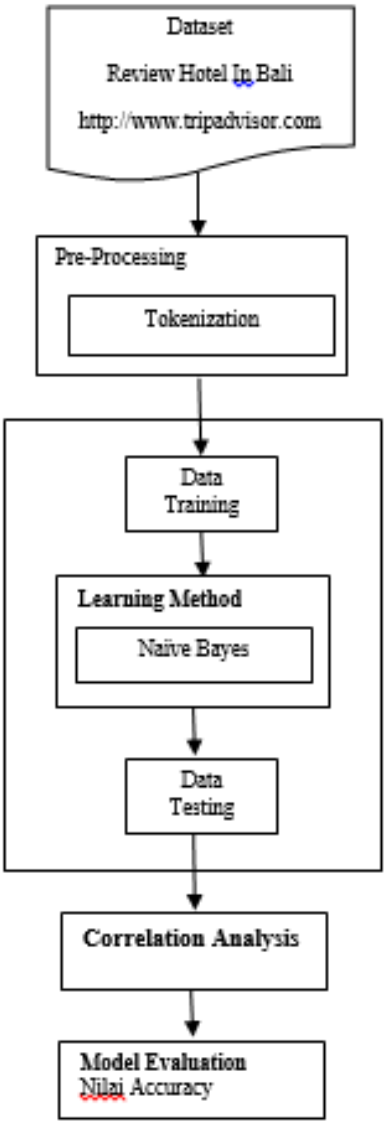
Gambar 2.2 Model Yang Diusulkan oleh Indriyani, Wawan Gunawan dan Ardhon Rakhmadi

2.2.2 Model Penelitian Suardika I Gede

Penelitian yang dilakukan oleh Suardika I Gede (2016). Melakukan penelitian menentukan apakah dahubungan antara peringkat hasil analisis sentiment yang dilakukan dangan analisis korelasi dengan menggunakan Naïve Bayes. Data yang di ambil dari www.tripadvisor.com dataset yang yangdiambil dari ulasan hotel dari beberapa daerah yang dipilih yaitu jimbran, Kuta, Nusa Dua, dan Seminyak. Satu hotel dengan peringkat tinggi dan satu hotel dengan peringkat rendah yang dipilih untuk setiap area tes akan dilakukan untuk menentukan apakah ada hubungan antara penilaian dengan hasil analisis sentimen. Hasil analisis sentimen dengan menggunakan algoritma Naïve Bayes adalah 81% dan

hasil analisis korelasi membuktikan hipotesis semakin besar presentase sentimen negatif maka semakin rendah peringkat hotel di TripAdvisor.

Langkah-langkah yang dilakukan terdiri dari pengolahan data dilakukan berdasarkan data dari situs www.tripadvisor.com mengenai ulasan hotel. Setelah dataset terkumpul maka langkah selanjutnya yaitu sistem analis sentimen yang dibangun menggunakan Naïve Bayes untuk mengukur tingkat akurasi klasifikasi. Data yang sudah diolah kemudian digunakan sebagai dasar pencarian hubungan antara peringkat hotel (tinggi atau rendah) disitus Tripadvisor dengan hasil klasifikasi kata sentiemen positif, sentimen negatif dan sentimen netral. Jadi dalam kesimpulan, ada hubungan yang positif dan koefisien korelasi antara peringkat dengan hasil analisis sentimen 0,836. Data dan koefisien yang diperoleh dalam sampel dapat digeneralisasikan ke populasi yang diambil dari TripAdvisor. Gambar 2.3 menunjukan model yang diusulkan oleh Saurdika



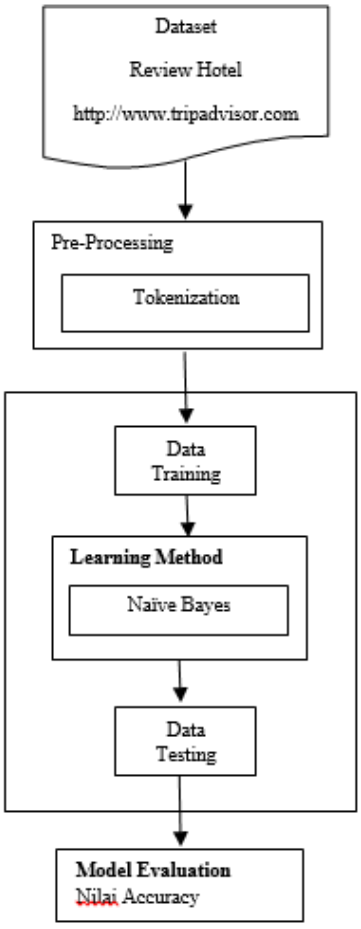
Gambar 2.3 Model Yang Diusulkan oleh Suardika

2.2.3 Model Penelitian Duan, Cao, Yu dan Levy

Penelitian yang dilakukan oleh Duan at al (2013). Melakukan penelitian pada dua domain yaitu pemasaran dan informasi, yang pertama mengidentifikasi dan mengukur kualitas layanan dan kinerja pelayanan. Yang kedua meneliti dampak WOM digital (*Review* pengguna secara *online*). Analisis sentimen dilakukan menggunakan Naïve Bayes berdasrkan data yang diambil dari tripadvisor. Hasil untuk klasifikasi lima dimensi (*Tangibles, Reliability, Responsiveness, Assurance dan Empati*) dari isi *review* yang menggunakan

algoritma Naïve bayes adalah 70 % berkaitan dengan SERVPERF dan jumlah kalimat yang mengandung kata sentimen positif dan kata sentiment negatif.

Langkah-langkah yang dilakukan terdiri dari tiga proses yang terlibat dalam prosedur analisis, SERVPERF klasifikasi dimensi dan klasifikasi sentimen polarisasi, yang pertama membersihkan konten baku dengan menghapus catatan kosong yang digandakan. Setelah itu diakui sebagai potongan *review* dari semua kata dalam setiap kalimat yang dinormalisasi kemudian menyimpanya sebagai korpus halus dalam format komputasi. Dalam proses yang kedua menyempurnakan kalimat dari langkah pertama kedalam satu dimensi kualitas layanan tertentu. Proses yang ketiga adalah melakukan analisis sentimen klasifikasi teks dan menentukan sentimen polaritas untuk setiap unit analisis berdasarkan review pelanggan tentang pelayanan dan kualitas hotel berdasarkan lima dimensi (*Tangibles, Reliability, Responsiveness, Assurance dan Empati*) dari isi *review.*

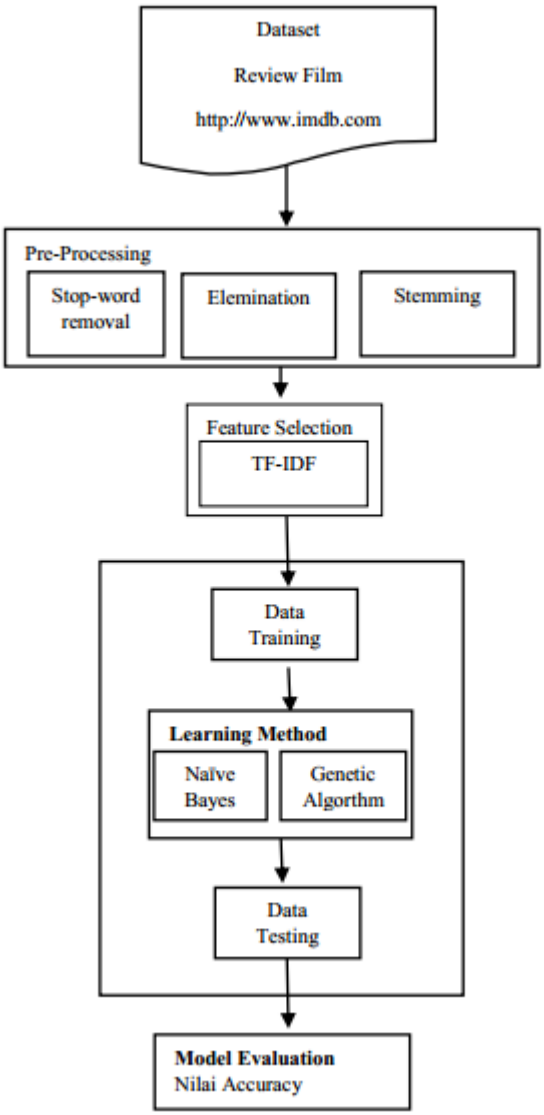


Gambar 2.4 Model Yang Diusulkan oleh Duan, Cao, Yu dan Levy

2.2.4 Model Penelitian M. Govindarajan

Penelitian yang dilakukan oleh Govindarajan (2013) melakukan penelitian analisis sentimen dari ulasan film dengan metode hybrid dari Naïve Bayes dan algoritma genetik. Data set yang diiambil dari http://imdb.com terdiri dari review 2000 film, 1000 label yang mengandung kata positif dan 1000 label yang mengandung kata negatif. Hasil klasifikasi dengan *classifier* Naïve Bayes adalah 91,15 % dan penggabungan antara *classifier* Naïve Bayes dengan Algoritma Genetik adalah 93,80 %

Langkah-langkah yang dilakukan berdasarkan pada lima bagian utama yaitu tahap *pre processing*, pada tahap ini langkah-langkah yang terlibar adalah *pre processing dokumen* dengan menggukan *stop-word, elimination* dan *stemming* setelah *pra processing* dokumen selesai maka dilakukan *fitur extrasi* dengan menggunakan metode TF-IDF (*Term Frequency - Inverse Document Frequency*) untuk mengidentifikasi kaka-kata penting dalam sebuah dokumen teks, selanjutnya setelah pemilihan pemilihan fitur selesai langkah selanjutnya pemilihan model atau algoritma *mechine leraning* yang akan digunakan untuk melatih classifier text. Setelah langkah pre processing selesai langkah selanjutnya adalah fase indexing dokumen. selanjutnya fase pengurangan fitur, setelah pengurangan fitur selesai dilakukan langkah selanjutnya adalah fase kasifikasi text yaitu menggunakan algoritma Naïve Bayes untuk klasifikasinya dan Genetik Algoritma sebagai pemilihan fiturnya. Langkah selanjutnya adalah fase menggabungkan kedua algoritma untuk agregat hasil klasifikasi terbaik.



Gambar 2.5 Model Yang Diusulkan oleh Govindarajan

Tabel 2. Perbandingan Penelitian Terkait

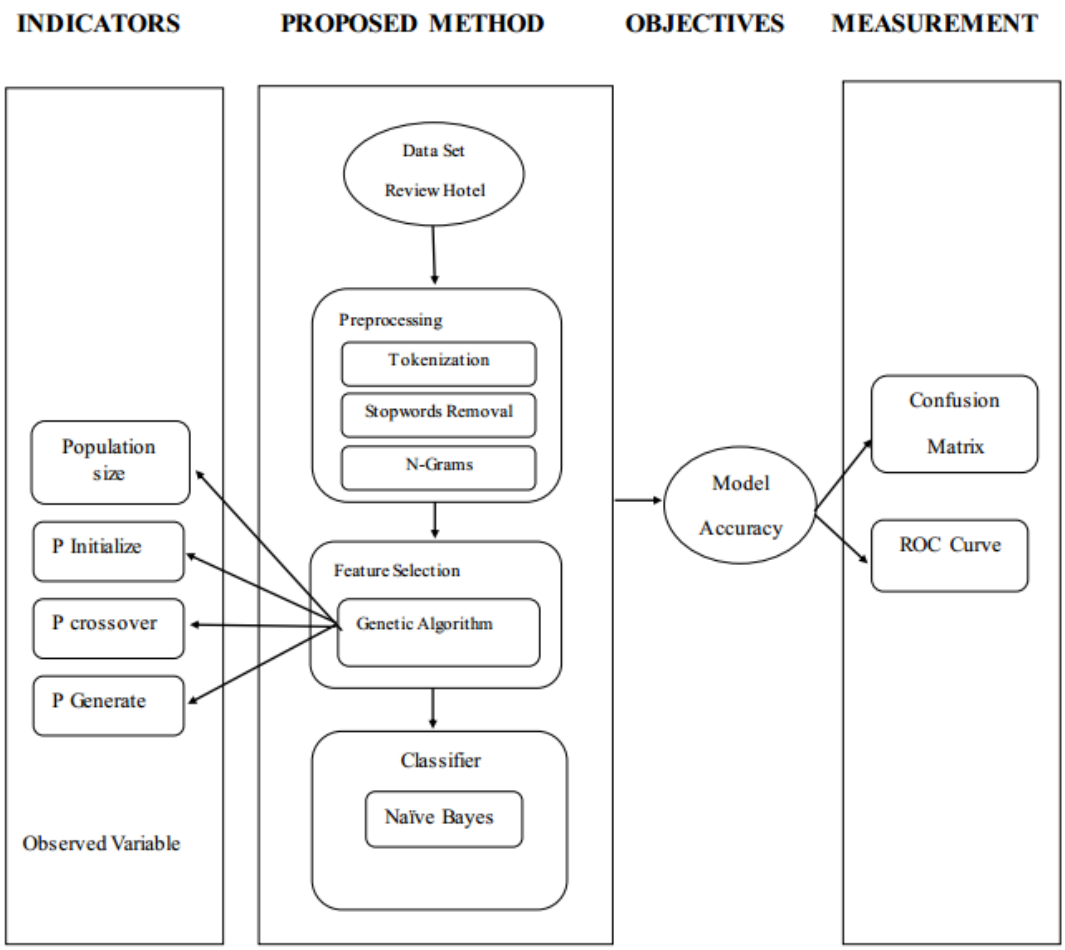
Judul	Text Processing	Feature Selection	Classifier	Accuracy
Filter-wrapper Approach to Feature Using PSO-GA for Arabic Document (Indriyani et al., 2015)	• Filtering • Stoplist Removal • Term Extraction • Stemming	TF-IDF	Naïve Bayes	70%
Sentiment Analysis System And Correlation Analisys On Hospitality In Bali (Suardika, 2016)	• Tokenization • Remove All Token	-	Naïve Bayes	81 %
Sentiment Analysis Tecnique to Study Hotel Service Quality (Duan et al., 2013)	• Tokenization	-	Naïve Bayes	70 %
Sentiment Analysis of	• Stopword Removal	TF-IDF	Naïve Bayes	91,15%

Judul	Text Processing	Feature Selection	Classifier	Accuracy
Movie Reviews (Govindaraja m, 2013)	• Elimination • Stemming			
Model yang diusulkan	• Tokenisasi • Stopword removal • N-gram	Genetic Algorithm	Naïve Bayes	?

Dari tinjauan studi diatas, dapat diketahui bahwa Naïve Bayes merupakan pengklasifikasi terbak untuk memecahkan masalah dalam pengklasifikasifikasian. Pada Tesis ini, penulis menggunakan pengkasifikasi Naïve Bayes yang diintegrasikan dengan metode pemilihan Fitur Genetic Algorithm dan metode AdaBoost yang berfungs untuk meningkatkan tingkat akurasi dan pengklasifikasi

2.3. Tinjauan Organisasi/Objek Penelitian

Penelitian ini adalah mengenai klasifikasi teks review hotel dengan mengggunakan Naïve Bayes. Dataset yang digunakan dari www.tripadvisor.com yang terdiri dari 100 review positif dan 100 review negatif. Untuk preprocessing dilakuakan Tokenisasi, *Stopword Removal*, *N-Grams*, metode pemilihan fitur yang digunakan adalah genetic algorithm, sedangkan pengklasifikasi yang digunakan adalah Naïve Bayes. Penelitian yang ini nantinya menghasilkan akuras dan nilai AUC dan menggunakan RapidMiner Versi 5.3 untuk hasil evaluasi. Kerangka pemikiran ini akan dijelaskan secara singkat seperti Gambar 2.5



Gambar 2.6 Kerangka Pemikiran Penelitian

Pada gambar 2.6 kerangka pemikiran dari penelitian dapat dijelaskan bahwa penelitian dimulai dari penemuan masalah yaitu berapa perbedaan atau dapat meningkatkan nilai akurasi pada klasifikasi alagoritma Naïve Bayes apabila Genetic Algorithm diterapkan. Data uji menggunakan 10 *Fold Cross Validation* untuk pengujian Naïve Bayes sebelum dan sesudah menggunakan metode pemilihan Fitur Gentetic Algorithm. Menurut Gunal (Gunal, 2012) untuk Genetic algorithm, parameter yang dianggap memberikan hasil yang optimal sebagai berikut *population size* diberi angka 50, *number of generation* diberi angka 30, *probability of crossover* diberi angka 0.8, dan *probability of mutation* diberi angka 0.08. Untuk penunjuang penelitian maka digunakan alat bantu untuk mengukur eksperimen dengan menggunakan *software* RapidMiner 5.3. Evaluasi dari pengujian validasi menggunakan Confusion Matrix dalam bentuk Kurva

Menggunakan Kurva ROC. Hasil yang dibandingkan adalah akurasi Naïve Bayes sebelum menggunakan metode pemilihan fitur dengan akurasi setelah menggunakan pemilihan fitur Genetic Algorithm, Naïve Bayes Diuji dalam tahap *wrapper*

BAB III
METODE PENELITIAN

3.1. Perancangan Penelitian

Metode penelitian yang penulis lakukan adalah metode penelitian eksperimen, dengan tahapan sebagai berikut :

1. Pengumpulan Data

   Pengumpulan data ditentukan berdasarkan data yang akan diproses yaitu berupa *review* positif maupun *review* negatif. Data tersebut kemudian diintegrasikan didalam *dataset.*

2. Pengolahan Data Awal

   Dilakukan penyeleksian data. Data dibersihkan dan diintegrasikan kedalam bentuk yang dinginkan sebelum dilakukan pembuatan model.

3. Metode yang diusulkan

   Data yang diteliti dan dianalisa kemudian dikelompokkan ke variabel mana yang berhubungan dengan satu sama lainnya, lalu dibuatkan model yang sesuai dengan jenis data. Pembagian data kedalam data latihan (*training* data) dan data uji (*testing* data) juga diperlukan untuk pembuatan model. Dengan penambahkan metode pemilihan fitur *Genetic Algorithm* untuk meningkatkan akurasi pada pengkasifikasi Naïve Bayes

4. Eksperimen dan Pengujian Metode

   Eksperimen pada model yang akan dilakukan dengan menggunakan RapidMiner 5 untuk mengolah data. Model diuji untuk melihat hasil yang akan dimanfaatkan untuk mengambil keputusan hasil penelitian

5. Evaluasi Dan Validasi Hasil

   Pada sebuah penelitian dilakukan evaluasi terhadap model untuk mengetahui akurasi dari model yang telah digunakan. Validasi hasil digunakan untuk melihat perbandingan dari model yang digunakan dengan hasil yang telah dilakukan sebelumnya.

3.2. Pengumpulan Data

Pada penelitian ini, hanya menggunakan 200 data review hotel yang diambil dari situs http://www.tripadvisor.com. Data yang terdiri dari 100 *review* positif dan 100 data *review* negatif

Berkut ini contoh data *review* Positif

"Lokasinya sangat strategis dan dekat dengan Plasa Senayan. Suasana kamar juga sangat nyaman dan sesuai dengan harga yang dibayar. Menginap di sini untuk urusan pekerjaan, sehingga kenyamanan yang diberikan mampu memberikan istirahat yang baik. Sarapan pagi walau tidak banyak pilihan tetapi rasanya cukup lezat".

Berikut ini conth data *review* Negatif

"Hotel ini sangat kotor. Lantainya basah dan langkah-langkah kami akan menempel di lantai. Banyak nyamuk di dalam kamar. Saluran TV buruk. TV buruk. Saya telah mengambil beberapa gambar bantal-bantalnya. Bantal-bantal terlalu kotor dan bau. Tidak direkomendasikan jika anada mencari hotel murah yang bersih dan nyaman".

3.3. Pengolahan Data Awal

Dataset yang digunakan dalam penelitian ini hanya 100 *review* positif dan 100 *review* negatif yang dijadikan sebagai data *training*. *Dataset* ini dalam tahap *preprocessing* harus melalui 3 proses, yaitu :

1. *Tokenization*

   Yaitu mengumpulkan semua kata yang muncul dan menghilangkan tanda baca maupun simbol yang bukan huruf.

2. *Stopwords Removal*

   Yaitu penghapusan kata-kata yang tidak relevan, seperti kata untuk, hanya, dengan, dan sebagainya.

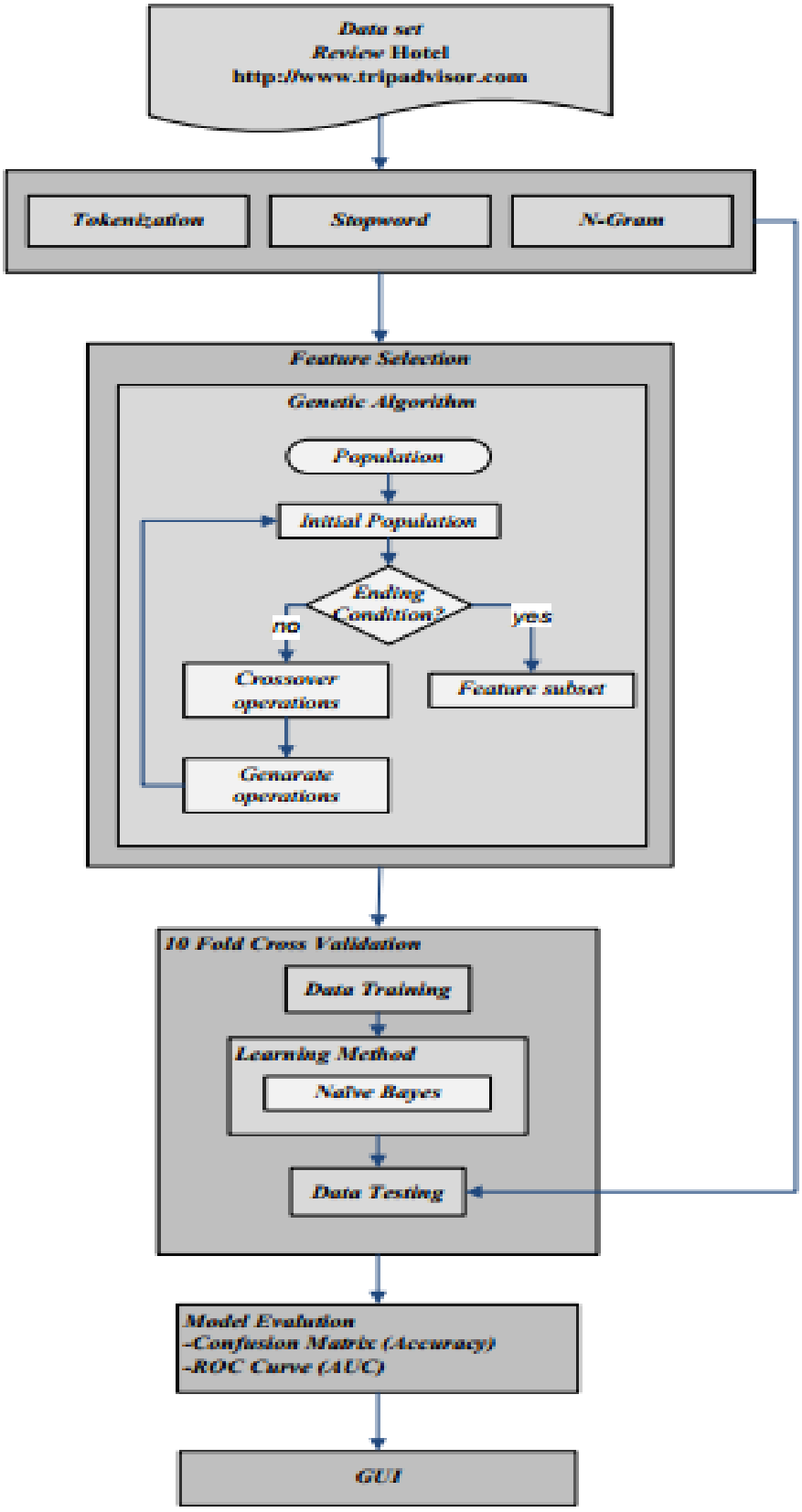
3. *N-Gram*

   Yaitu potongan n karakter dalam suatu string tertentu atau potongan n kata dalam suatu kalimat tertentu.

Sedangkan untuk tahap *transformation* dengan melakukan pembobotan TF-IDF (*Term Frequency - Inverse Document Frequency*) pada masing-masing kata. Dimana prosesnya menghitung kehadiran atau ketidakhadiran sebuah kata didalam sebuah dokumen. Beberapa kali sebuah kata muncul didalam suatu dokumen juga digunakan sebagai skema pembobotan dari kata tekstual

3.4. Metode Yang Diusulkan

Metode yang penulis usulkan adalah menunggunkan metode pemilihan fitur jenis *wrapper*, yaitu *Genetic Algorithm,* yang digunakan untuk meningkatkan akurasi dari pengkasifikasi Naïve Bayes. Penelitian ini mengenai *review* hotel dengan menggunakan pengkasifikasi Naïve Bayes merupakan salah satu algoritma yang memiliki kecepatan yang tinggi saat diaplikasikan kedalam *database* dengan data yang besar. *Dataset* yang digunakan berasal dari www.tripadvisor.com yang terdiri dari 100 *review* positif dan 100 *review* negatif. Untuk *preprocessing* dilakukan *tokenization, stopword removal* dan *N-Grams*. Penelitian ini nantinya menghasilkan akurasi dan nilai *AUC* dengan menggunakan aplikasi *RapidMiner versi* 5.3 untuk hasil evaluasi. Lihat gambar 3.1 untuk model yang diusulkan secara detail.



Gambar 3.1 Model yang diusulkan

Data harus melalui tahap *preprocessing* terlebih dahulu agar didapatkan kata-kata yang sudah dihilangkan simbol-simbolnya dan mendapatkan kata-kata yang relevan untuk diklasifikasi. Proses evaluasi dilakukan untuk pengujian dengan algoritma Naïve Bayes dan evaluasi dilakukan dengan menggunakan menggunakan 10-*fold cross validation* untuk pengujian sebelum dan sesudah menggunakan *Genetic Algorithm*. Pengukuran akurasi diukur dengan *confusion matrix*. Menurut Gunal (Gunal, 2012) untuk *Genetic Algorithm*, parameter yang dianggap memberikan hasil yang optimal sebagai berikut *population size* diberi angka 50*, number of generation* diberi angka 30*, probability of crossover* diberi angka 0.8*,* dan *probability of mutation* diberi angka 0.08.

Hasil yang dibandingkan adalah akurasi *Naïve Bayes* sebelum menggunakan metode pemilihan fitur dengan akurasi setelah menggunakan pemilihan fitur *Genetic Algorithm*, Naïve Bayes Diuji dalam tahap *wrapper*

3.5. Eksperimen dan pengujian Model

Proses Eksperimen yang dilakukan menggunakan aplikasi RapidMiner 5.3 untuk pengujian. Model dilakukan menggunakan *dataset Review* Hotel. Tahap pengujian data untuk mengklasifikasi *review* sebagai berikut :

1. Menyiapkan *dataset* untuk eksperimen
2. *Input review* hotel yang belum pernah dikasifikasikan sebelumnya
3. Semua *text* dari *review* yang telah di *input* hilangkan terlebih dahulu simbol-simbol yang ada didalamnya
4. Mendesain arsitektur algoritma klasifikasi Naïve Bayes
5. Melakukan *training* dan *testing* terhadap algoritma Naïve Bayes dan mencatat hasil *Accuracy* dan *AUC*
6. Melakukan pengujian dengan model 10 *fold cross validation* dan mencari nilai metode pemilihan fitur
7. Mendesain arsitektur algoritma klasifikasi Naïve Bayes dan pemilihan fitur *Genetic Algorthm*.
8. Melakukan *training* dan *testing* terhadap Algoritma Naïve Bayes, *Genetic Algorithm* kemudian mencatat hasil *Accuracy* Dan *AUC*

Sedangkan spesifikasi Komputer yang digunakan untuk eksperimen ini dapat dilihat pada tabel 3.1

Tabel 3.1 Spesifikasi Komputer yang Digunakan

Processor	Intel(R) Core(TM) i3-2350M CPU @ 2.30GHz
Memori	4 GB
Harddisk	320 GB
Sistem Operasi	*Mcrosoft Windows 8.1*
Aplikasi	*RapidMiner 5.1*

3.6. Evaluasi dan Validasi Hasil

Validasi dilakukan menggunakan 10 *fold cross* data eksperimen akan dibagi menjadi 10 bagian. Satu bagian untuk data *testing* sedangkan sembilan bagian lainnya untuk data *training*. Sedangkan pengukuran akurasi diukur dengan *confusion matrix* dan *kurva ROC* (*Receiver Operating Characteristic*) untuk mengukur nilai *AUC*. Dengan *confusion matrix*, akurasi Naïves Bayes sebelum menggunakan pemilihan fitur *Genetic Algorithm* dan setelah menggunakan pemilihan fitur *Genetic Algorithm*. Tabel 3.2 berikut adalah tampilan *confusion matrix* dan rumus perhitungannya menurut Gorunescu (Gorunescu,2011):

Tabel 3.2 *Confusion Matrix*

Classification	***Predicted Class***		
		Class = Yes	*Class = No*
	Class = Yes	a (*True Positif* – TP)	b (*False Negatif* – FN)
	Class = No	c (*False Positif* – FP)	d (*True Negatif* – TN)

$$Accurucy = \frac{a+d}{a+b+c+d} \frac{TP+TN}{TP+FN+FP+TN}$$

BAB IV
HASIL PENELITIAN DAN PEMBAHASAN

4.1. Hasil

Penelitian ini bertujuan untuk mengetahui akurasi dari algoritma Naïve Bayes dan pemilihan fitur *Genetic Algorithm*

4.1.1. Klasifikasi teks menggunakan Algoritma Naïve Bayes

Sebelum diklasifikasikan, dataset harus melalui beberapa tahapan proses agar bisa diklasifikasikan dalam proses selanjutnya, berikut ini adalah tahapan prosesnya :

1. Pengumpulan Data

   Pada penelitian ini menggunakan data *review* hotel yang diambil dari situs http://www.tripadvisor.com . *Review* hotel yang digunakan hanya 200 *review* hotel yang terdiri dari 100 *review* positif dan 100 *review* negatif. Data tersebut masih berupa sekumpulan teks yang terpisah dalam bentuk dokumen. Data *review* positif disatukan dalam satu folder dan diberi nama positif, sedangkan data *review* negatif disatukan dalam satu *folder* dan diberi nama negatif. Tiap dokumen berekstensi .txt yang dapat dibuka dengan menggunakan aplikasi *Notepad*

2. Pengolahan Data Awal

   a. Tokenization

   Dalam proses *tokenization* ini, semua kata yang ada didalam setiap dokumen dikumpulkan dan di hilangkan tanda bacanya, serta dihilangkan juga apabila ada simbol yang bukan huruf. Berikut adalah contoh hasil dari proses *tokenization* dalam *RapidMiner*. Tabel 4.1 menunjukan hasil perbandingan teks sebelum dan sesudah dilakukan proses *tokenization*

Tabel 4.1. Perbandingan teks sebelum dan sesudah dilakukan proses *tokenization*

Sebelum dilakukan proses *tokenization*	Kamar bagus, *modern,* bersih, besar. Lokasi bagus.*Breakfast fantastic*.Pemandangan bagus diluar area hotel.*Service excellence*.satu" nya masalah cm kamar mandi (air tdk bisa mengalir turun krn permukaan lbh tinggi).
Setelah dilakukan proses *tokenization*	Kamar bagus *modern* bersih besar Lokasi bagus *Breakfast fantastic* Pemandangan bagus diluar *area* hotel *Service excellence* satu nya masalah cm kamar mandi air tdk bisa mengalir turun krn permukaan lbh tinggi

b. Stopword Removal

Dalam proses ini, kata-kata yang tidak relevan akan dihapus , seperti kata untuk hanya dengan dan sebagainya yang merupakan kata –kata yang tidak mempunyai makna tersendiri jika dipisahkan dengan kata yang lain dan tidak terkait dengan kata sifat yang berhubungan dangan sentimen. Tabel 4.2 menunjukan hasil perbandingan teks sebelum dan sesudah dilakukan proses *stopword removal*.

Tabel 4.2. Perbandingan teks sebelum dan sesudah dilakukan proses *stopword removal*

Sebelum dilakukan proses *stopword removal*	Lokasi hotelnya masuk gang,kamarnya lembab dan kamar mandinya kotor. seperti tidak ada pembersihan setiap hari. *Front office* nya seperti
Sebelum dilakukan proses	tempat penitipan barang dan

stopword removal	terkesan kotor dan tidak teratur. foto foto yang terpasang di *website* sepertinya foto hotel baru jadi.
Sebelum dilakukan proses *stopword removal*	lokasi hotelnya masuk gang kamarnya lembab kamar mandinya kotor pembersihan hari *Front office* nya tempat penitipan barang terkesan kotor teratur foto foto terpasang *website* foto hotel baru jadi

c. *N-gram (Bi-Gram)*

Dalam proses ini, potongan 2 karakter dalam suatu *string* tertentu atau potongan 2 kata dalam suatu kalimat tertentu. Contoh pemotongan 2 kata Bi-gram dalam kata Hotel cukup nyaman : “hotel”, “hotel_cukup”, “cukup”, “cukup_nyaman”, “nyaman”. Tabel 4.3 menunjukan perbandingan teks sebelum dan sesudah dilakukan proses *N-gram (Bi-gram)*

Tabel 4.3. Perbandingan teks sebelum dan sesudah dilakukan proses *N-gram (Bi-gram)*

Sebelum dilakukan proses *N-gram (Bi-gram)*	Untuk harga dihotel sebanding yg didapat kamar mandi kotor *spot sprei* diganti spot Karpet kamar bau Dibandingkan hotel harga yg hotel kurang baik kondisi kamarnya
Sebelum dilakukan proses *N-gram (Bi-gram)* **Sebelum dilakukan proses *N-***	Untuk Untuk_harga harga harga_dihotel dihotel dihotel_sebanding sebanding sebanding_yg yg yg_didapat didapat didapat_kamar kamar

gram (Bi-gram)	kamar_mandi mandi mandi_kotor kotor kotor_*spot spot spot_sprei sprei sprei*_diganti diganti diganti_*spot spot spot*_Karpet Karpet Karpet_kamar kamar kamar_bau bau bau_Dibandingkan Dibandingkan Dibandingkan_hotel hotel hotel_harga harga harga_yg yg yg_hotel hotel hotel_kurang kurang kurang_baik baik baik_kondisi kondisi kondisi_kamarnya kamarnya

3. Klasifikasi

Proses pengklasifikasian ini adalah menentukan *class* untuk setiap kalimat sebagai anggota *class* positif atau *class* negatif. Penentuan *class* pada setiap kalimat ditentukan melalui perhitungan probabilitas dari rumus Naïve Bayes. *Class* diberikan nilai Positif apabila nilai probabilitas pada dokumen tersebut untuk nilai *class* positifnya lebih besar dibandingkan dengan *class* negatif. Dan suatu kalimat dikatakan *class* negatif apabila nilai probabilitas pada dokumen tersebut untuk nilai *class* negatifnya lebih besar dibandingkan dengan *class* positifnya. Penulis hanya menampilkan 10 dokumen sentimen dari keseluruhan 200 data *training* dan 4 kata yang berhubungan dengan kata sentimen, yaitu bagus, nyaman, kotor dan buruk. Kehadiran kata dalam suatu kalimat akan diwakili oleh angka 1 dan angka 0 jika kata tersebut tidak muncul dalam kalimat pada dokumen.

Tabel 4. 4 Hasil Klasifikasi Text

Dokumen	Bagus	Nyaman	Kotor	Buruk	*Class*
Positif-1.txt	1	1	0	0	Positif
Positif-2.txt	0	1	0	0	Positif
Positif-3.txt	1	0	0	0	Positif
Positif-4.txt	0	1	0	0	Positif
Positif-5.txt	1	0	0	0	Positif
Negatif-1.txt	0	0	1	0	Negatif
Negatif-2.txt	1	0	1	0	Negatif
Negatif-3.txt	0	0	1	0	Negatif
Negatif-4.txt	0	1	0	1	Negatif
Negatif-5.txt	0	0	1	0	?

1. Hitung probabilitas bersyarat *(likehood*) dokumen ke N-5 pada *class* positif dan negatif

   P(Negatif-5.txt|Positif) = P(Bagus=3|Positif) x P(Nyaman=3|Positif) x
   P(Kotor=0|Positif) x P(Buruk=0|Positif)

   P(Negatif-5.txt|Positif) = 3/5 x 3/5 x 0/5 x 0/5

   = 0,6 x 0,6 x 0 x 0

   = 0

   P(Negatif-5.txt|Negatif) = P(Bagus=1|Negatif) x P(Nyaman=1|Negatif)
   x P(Kotor=4| Negatif) x P(Buruk=1|Negatif)

   P(Negatif-5.txt|Negatif) = 1/4 x 1/4 x 4/4 x /1/4

   = 0.25 x 0.25 x 1 x 0.25

   = 0.015

2. Probabilitas prior darai *class* positif dan negatif dihitung dengan proporsi dokumen pasa setiap class

   P(Positif) = 5/9 = 0,55

   P(Negatif) = 4/9 =0,44

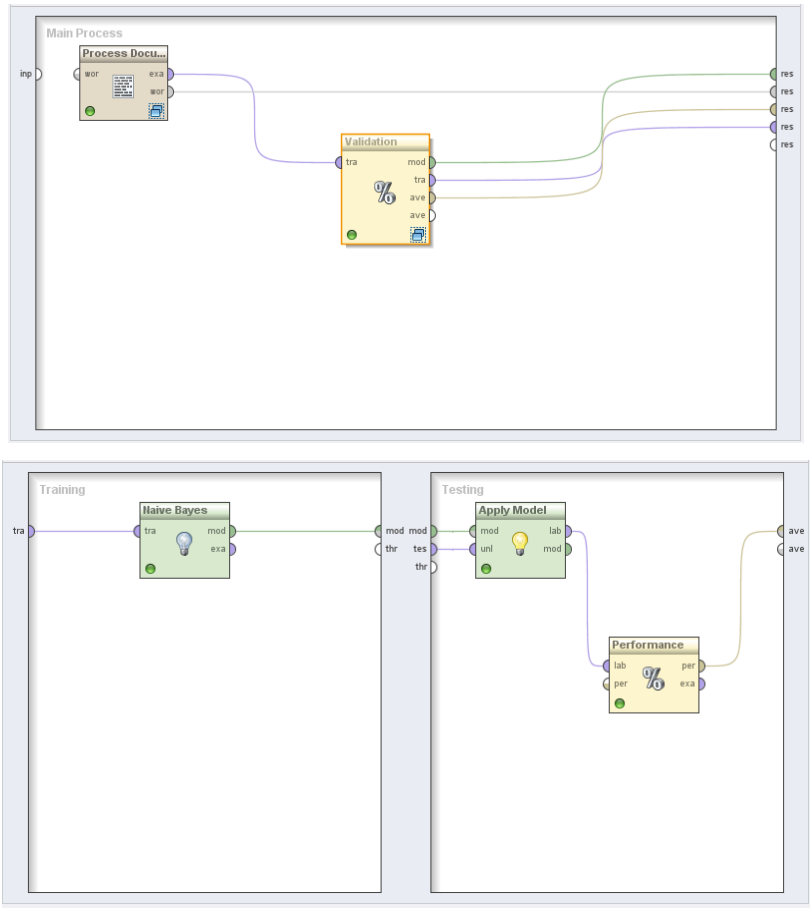
3. Hitung probabilitas posterior dengan memasikan nilai Bayes dan menghilangkan penyebut (N-5)

P(Positif|N-5) $= \frac{(0)(0.55)}{(negatif-5)} = 0$

P(Negatif|N-5) $= \frac{(0.015)(0.44)}{(negatif-5)} = 0{,}0066$

Berdasarkan probobilitas diatas, maka dapat disimpulkan bahwa dokumen Negatif-5.txt termasuk kedalam *class* negatif, karena P(Positif|Negatif-5.txt) lebih kecil dari pada P(Negatif|Negatif-5.txt)

Perhitungan diatas dapat dibuat suatu model dengan *RapidMiner* 5.3 Desain model arsitektur klasifikasi Naïve Bayes dapat dilihat pada gambar 4.1



Gambar 4.1 Desain Model Arsitektur Klasifikasi Naïve Bayes

4.1.2. Hasil Eksperimen Menggunakan Algoritma Naïve Bayes

Dari data sebanyak 200 data *review* hotel yang terdiri dari 100 data *review* positif dan 100 data *review* negatif. Sebanyak 96 data di prediksi *class* negatif sesuai yaitu termasuk kedalam prediksi *class* negatif dan sebanyak 11 data di prediksi *class* negatif ternyata termasuk kedalam *class* positif, 85 data di prediksi

class positif sesuai yaitu termasuk kedalam prediksi *class* positif dan sebanyak 15 data di prediksi *class* positif ternyata termasuk kedalam *class* negatif. Hasil yang diperoleh dengan menggunakan algoritma Naïve Bayes menggunakan *RapidMiner* 5.3 mendapatkan nilai *Accuracy* = 90.50% seperti pada tabel 4.4 dan mendapatkan nilai *AUC* : 0.500, Seperti gambar 4.2

Tabel 4.5 *Confusion Matrix* Algoritma Naïve Bayes

***Accuracy* : 90.50% +/- 9.07% (mikro :90.50%)**			
	True Class Negatif	***True Class Positif***	***Class Precision***
Pred. Class Negatif	**96**	**15**	**86.49%**
Pred. Class Positif	**4**	**85**	**95.51%**
Class Recall	**96.00%**	**85.00%**	

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

$$Accuracy = \frac{96 + 85}{95 + 15 + 4 + 85}$$

$$Accuracy = \frac{181}{200} = 0.905 = 90.50\%$$

Gambar 4.2 *Grafik Area Under Curve (AUC)* Naïve Bayes

4.1.3. Pengujian Model dengan 10 *Fold Cross Validation*

Pada penelitian ini, penulis melakukan pengujian model dengan menggunakan teknik 10 *cross validation,* di mana proses ini membagi data secara acak ke dalam 10 bagian. Proses pengujian dimulai dengan pembentukan model dengan data pada bagian pertama. Model yang terbentuk akan diujikan pada 9 bagian data sisanya. Setelah itu proses akurasi dihitung dengan melihat seberapa banyak data yang sudah terklasifikasi dengan benar.

Teknik 10 *fold Cross Validation* ditentukan berdasarkan dari hasil uji coba penulis untuk mendapatkan hasil akurasi yang tinggi, dalam hal ini yang akan di uji coba untuk meningkatkan akurasi adalah nilai *validation*. Tabel indikator dan hasil pengujian dapat dilihat pada tabel 4.6 pengujian model 10 *fold cross Validation*

Tabel 4.6 Pengujian Model 10 *fold cross Validation*

Validation	*Accuracy (%)*
1	85.00%
2	85.00%
3	87.52%
4	85.00%
5	89.00%
6	88.47%
7	89.06%
8	87.50%
9	90.47%
10	**90.50%**

Pada tabel 4.6 menunjukan nilai akurasi teringgi yaitu dengan menggunakan 10 *fold cross validation*. Dimana nilai akurasi menggunakan 10 *fold cross validation* menddapatkan nilai 90.50 %

4.1.4. Model dengan Metode Pemilihan Fitur *Genetic Algorithm*

Sebelum melakukan perhitungan *Genetic Algorithm* langkah pertama mencari nilai bobot (w) dari *TF-IDF*. Melakukan pembobotan *TF-IDF* (*Term Frequency - Inverse Document Frequency*) pada masing-masing kata. Dimana prosesnya menghitung kehadiran atau ketidak hadiran sebuah kata didalam sebuah dokumen. Beberapa kali sebuah kata muncul didalam suatu dokumen juga digunakan sebagai skema pembobotan dari kata tekstual.

Tabel 4. 7 Dokumen Pembangunan Index Kamus Kata untuk *TF-IDF*

Dokumen Ke	Nama Dokumen	*Review*
D1	Positif-1.txt	Saya bepergian anak pikir *ritz carlton* pp **hotel** bisnis **hotel keluarga** ternyata dugaan salah turun mobil sapa ke**ramah**an masuk pikir **hotel** lobbynya banget gak petunjuk dimana resepsionis nya arahkan lantai anak sapa ke**ramah**an **kamar** resepsionisnya jumpai **hotel kamar** x lift hehe lantai hubnya ***check in*** cepat **ramah** anak **kamar** nya **luas kasur** ukuran king size nyaman internet free cepat **kamar** mandi *bathtub shower tv mini* dalamnya sayangnya amenities gak **bagus** keluhan cepat tanggapi tambahan air mineral layani mengambil paket *breakfast* anak prediksi bangun telat kolam renang **bagus** dewasa sauna spa terpisah laki perempuan ***fasilitas*** *gym* terhubung langsung *pacific place mall* lantai dasar disukai anak *kidz overall* **lokasi** pusat kota **pelayanan** *staff* **ramah** cocok bisnis membawa **keluarga** perhatikan *traffic* jam sudirman gak kepengen ngabisin jalan
D2	Positif-2.txt	Ruangan **luas** dan perlengkapan sangat lengkap terutama jika ingin memasak***,room*** **nyaman**,**lokasi** strategis ditengah kota,dan **pelayanan** sangat baik,menu *breakfast* tidak begitu banyak tapi semuanya sangat cukup *worth it*.
D3	Positif-3.txt	**Kamar** yang **luas** dengan pemandangan menghadap gelora bung karno. sangat dekat dengan pusat perbelanjaan. **kamar** yang **luas** dan ***fasilitas*** yang sangat **bagus**. makanan yang disediakan sangat

Dokumen Ke	Nama Dokumen	*Review*
		bervariasi. proses *check in* dan *check out* yang cepat.

Pada dokumen D1, dokumen D2 dan dokumen D3 yang akan dilakukan pembangunan *indeks* untuk *lexicon* (kamus kata) dan pembobotan dengan *TF-IDF* sehingga dapat nilai bobot w sesuai dengan rumus *TF-IDF* standar yang belum dinormalisasi. Sebelum menghitung nilai w terlebih dahulu mencari nilai tf pada setiap dokumen misalkan kata chek in terdapat di dokumen D1 sebanyak 1 dan dokumen D3 sebanyak 1. Mencari nilai df didapat dari jumlah kata yang mucul di dalam dokumen. Dan mencari nilai idf = *log* (jumlah dokumen/ df)

Menentukan nilai tf, df dan idf

Tabel 4.8 Perhitungan tf, df dan idf

	Tf			df	idf
	D1	D2	D3		
Ramah	4	0	0	1	0,477
Kamar	4	0	2	2	0,176
Check in	1	0	1	2	0,176
Luas	1	1	2	3	0
Kasur	1	0	0	1	0,477
Bagus	2	0	1	2	0,176
Fasilitas	1	0	1	2	0,176
Keluarga	2	0	0	1	0,477
Pelayanan	1	1	0	2	0,176
Nyaman	0	1	0	1	0,477
Hotel	4	0	0	1	0,477
Room	0	1	0	1	0,477
Lokasi	1	1	0	2	0,176

Menentukan nilai idf

$$idf\,(ramah) = \log(\frac{N}{df})$$

$$idf\,(ramah) = \mathrm{Log}(\frac{3}{1}) = 0{,}477$$

Setelah diketahui nilai tf, df dan idf untuk masing-masing dokumen dan masing-masing kata , kemudian hitung nilai w untuk setiap kata dalam setiap dokumen

Tabel 4.9 Nilai Bobot (w) Sebelum Normalisasi

	tf			Df	idf	w		
	D1	D2	D3			D1	D2	D3
Ramah	4	0	0	1	0,477	1,908	0	1,908
Kamar	4	0	2	2	0,176	0,704	0	0,704
Check in	1	0	1	2	0,176	0,176	0	0,176
Luas	1	1	2	3	0	0	0	0
Kasur	1	0	0	1	0,477	0,477	0	0,477
Bagus	2	0	1	2	0,176	0,352	0	0,352
Fasilitas	1	0	1	2	0,176	0,176	0	0,176
Keluarga	2	0	0	1	0,477	0,954	0	0,954
Pelayanan	1	1	0	2	0,176	0,176	0,176	0,176
Nyaman	0	1	0	1	0,477	0	0,477	0
Hotel	4	0	0	1	0,477	1,908	0	1,908
Room	0	1	0	1	0,477	0	0,477	0
Lokasi	1	1	0	2	0,176	0,176	0,176	0,176

Nilai w sebelum Normalisasi

$$w = tf\,(ramah) * idf(ramah)$$

$$w = 4 * 0{,}447$$

$$w = 1{,}908$$

Tabel 4.10 Nilai Bobot (w) Setelah Normalisasi

	tf			df	idf	w		
	D1	D2	D3			D1	D2	D3
Ramah	4	0	0	1	0,477	0,551	0	0
Kamar	4	0	2	2	0,176	0,439	0	0,633
Check in	1	0	1	2	0,176	0,110	0	0,317
Luas	1	1	2	3	0	0,093	0,351	0,538
Kasur	1	0	0	1	0,477	0,138	0	0
Bagus	2	0	1	2	0,176	0,219	0	0,317
Fasilitas	1	0	1	2	0,176	0,110	0	0,317
Keluarga	2	0	0	1	0,477	0,276	0	0
Pelayanan	1	1	0	2	0,176	0,110	0,412	0
Nyaman	0	1	0	1	0,477	0	0,518	0
Hotel	4	0	0	1	0,477	0,551	0	0

	tf			df	idf	w		
	D1	D2	D3			D1	D2	D3
Room	0	1	0	1	0,477	0	0,518	0
Lokasi	1	1	0	2	0,176	0,110	0,412	0

$$w(ramah) = \frac{4 * (0{,}477 + 1)}{\sqrt{(4^2 * [0{,}477]^2) + (4^2 + [0{,}176]^2) + (1^2 * [0{,}176]^2) \ldots . n}}$$

$$w(ramah) = \frac{5{,}980}{\sqrt{114{,}926}}$$

$$w(ramah) = \frac{5{,}908}{10{,}720}$$

$$w(ramah) = 0{,}551$$

Setelah pembobotan nilai (w) pada *TF-IDF* sudah diketahui, maka bisa menghitung *genetic algorithm*.

1. Langkah pertama adalah menentukan populasi awal dimana nilai populasi awal di dapat dari pembobotan (w) pada *tf-idf* yang sudah di normalisasi dijadikan sebagai nilai populasi awal.

   D1 = [0,551_0,439_0,110_0,093_0,138_0,219_0,110_0,276_0,110_0_0,551_ 0_0,110]

   D2 = [0_0_0_0,351_0_0_0_0_0,412_0,518_0_0,518_0,412]

   D3 = [0,_0,633_0,317_0,538_0_0,317_0,317_0_0_0_0_0_0]

2. Langkah ke 2 adalah evaluasi nilai *fitnes*, maka akan menghasilkan nilai *fitness* pada setiap *cromosom*

   Fitnes D1 = 1/(1+(0,551 + 0,439 + 0,110 + 0,093 + 0,138 + 0,219 + 0,11 0 + 0,276 + 0,110 + 0+ 0,551 + 0 + 0,110))

   = 0,270

   Fitnes D2 =1/(1+(0 + 0 + 0 + 0,351 + 0 + 0 + 0 + 0 + 0,412 + 0,518 + 0 + 0,518 + 0,412))

   = 0,311

Fitnes D3 =1(1+(0 + 0,633 + 0,317 + 0,538 + 0 + 0,317 + 0,317 + 0 + 0 + 0

+ 0 + 0 +0))

=0,320

Total nilai fitness adalah 0,270+0,311+0,320 = 0,902

Probabilitas masing-masing *cromosom* menjadi :

P(D1) = 0,270/0,902= 0,299

P(D2) =0,311/0,902= 0,345

P(D3) =0,320/0,902 = 0,355

Dari hasil probabilitas tertinggi, dihasilkan bahwa *cromosom* 3 mempunyai nilai *fitness* paling tinggi. Maka *cromosom* 3 juga mempunyai kesempatan paling besar dalam proses seleksi selanjutnya dengan *Roullete Wheel*.

3. Penentuan *Cromosom* Induk

Untuk proses seleksi digunakan *Roullete Wheel*, untuk itu diperlukan nilai kumulatif propabilitasnya dari setiap *cromosom*, yakni sebagai berikut :

D1 = 0,299

= 0,299

D2 = 0,299 + 0,345

= 0,645

D3 = 0,299+0,345+0,355

= 1

Langkah selanjutnya adalah dengan menggunkan bilangan acak R antara 0 sampai dengan 1, bilangan acak dipilih sesuai dengan jumlah *cromosom*:

R(D1) = 0, 375

R(D2) = 0, 725

R(D3) = 0, 255

Memilih *cromosom* ke x sebagai *Parent* dengan syarat D[x-1] < R < D [x]. Angka acak R[D1] < nilai kumulatif dari D2 sehingga D2 nanti akan

dilakukan *crossover* dengan D1 dan nilai R[D2] < nilai *komulatif* dari D3, Sehingga D3 akan dilakukan *crossover* dengan D2 .

Hasil seleksi *Roullete Whell* pada populasi ini untuk *crossover* menjadi :

D1 dengan D2

0,299 < 0,375 < 0,645

D1 menjadi D2

0_0_0_0,351_0_0_0_0_0,412_0,518_0_0,518_0,412

D2 dengan D3

0,645 < 0,725 < 1

D2 menjadi D3

0,_0,633_0,317_0,538_0_0,317_0,317_0_0_0_0_0_0

D3 dengan D1

1 < 0,255 < 0,299

D3 tetap D3

0_0,633_0,317_0,538_0_0,317_0,317_0_0_0_0_0_0

4. Perkawinan Silang atau *Crossover*

Dalam *crossover* menentukan probability (pr) yaitu 0,5 atau 50%. Hanya *cromosom* yang nilai R lebih kecil dari 0,5 yang akan bermutasi. Maka *cromosom* ke y akan dipilih menjadi induk jika R(y) < pr, dari bilangan acak R diatas maka yang akan menjadi parent adalah D1 dan D3. Sedangkan D2 > 0,5 sehingga tidak dilakukan seleksi. Selanjutnya setelah melakukan pemilihan *parent*, dilanjutkan menentukan *kromosom* yang akan dilakukan perkawinan silang dengan sejumlah atribut 1-13. Dalam hal ini posisi *cut-point* (cp) dipilih menggunakan bilangan acak 1-13 sesuai banyaknya *crosover* yang terjadi. misalnya didapatkan posisi *crossover* adalah 9, maka *cromosom parent* dipotong mulai gen ke 8 kemudian potongan gen tersebut saling ditukarkan antar *parent.*

D1><D3

Cp (D1) = 9

Cp (D3) = 13

Nilai *Offspring*[1] = D1><D3 dengan cp (D1)

0_0_0_0,351_0_0_0_0_**0**_0,518_0_0,518_0,412

><

0_0,633_0,317_0,538_0_0,317_0,317_0_**0,412**_0_0_0_0

=

0_0_0_0,351_0_0_0_0_**0** _0,518_0_0,518_0,412

Nilai *Offspring*[2] =D3><D1

0_0,633_0,317_0,538_0_0,317_0,317_0_0_0_0_0_**0,412**

><

0_0_0_0,351_0_0_0_0_0,412_0,518_0_0,518_**0**

=

0_0,633_0,317_0,538_0_0,317_0,317_0_0_0_0_0_**0,412**

Sehingga populasi baru yang dihasilkan dari *crossover* adalah

D1= 0_0_0_0,351_0_0_0_0_**0** _0,518_0_0,518_0,412

D2= 0_0,633_0,317_0,538_0_0,317_0,317_0_0_0_0_0_0

D3= 0_0,633_0,317_0,538_0_0,317_0,317_0_0_0_0_0_**0,412**

5. Mutasi

cromosom Jumlah *cromosom* yang mengalami mutasi dalam satu populasi ditentukan oleh persentase p *mutation*. Proses mutasi dilakukan dengan cara mengganti satu gen yang terpilih secara acak dengan suatu nilai baru yang didapat secara acak. Total gen = (gen dalam kromosom) * jumlah kromosom

= 13 * 3

= 39

Tentukan posisi gen yang akan mengalami mutasi dengan menggunakan bilangan acak antara 1 sampai dengan total gen, yaitu antara 1 sampai 39. Misalkan ρm kita tentukan 10% maka jumlah gen yang mengalami mutasI adalah 10% dari 39 yaitu 3,9 atau 3 gen. Kemudian gunakan bilangan acak dari total gen misalkan yang terpilih adalah posisi gen 21, 22 dan 23 yang akan mengalami mutasi. Dengan demikian yang akan mengalami mutasi adalah kromosome ke- 2 gen nomor 8,9 dan 10. Maka nilai gen pada

posisi tersebut akan diganti dengan bilangan acak 0,00-1,00. Misalkan bilangan acak yang digunakan adalah 0,110, 0,276 0,551 maka kromosom ke-2 berubah menjadi 0_0,633_0,317_0,538_0_0,317_0,317_**0,110_0,276_0,551**_0_0_0

Populasi pada generasi pertama menjadi:

D1 = 0_0_0_0,351_0_0_0_0_**0** _0,518_0_0,518_0,412

D2 = 0_0,633_0,317_0,538_0_0,317_0,317_**0,110_0,276_0,551**_0_0_0

D3 = 0_0,633_0,317_0,538_0_0,317_0,317_0_0_0_0_0_**0,412**

Kromosom-kromosom pada populasi ini akan mengalami proses yang sama seperti generasi sebelumnya yaitu proses evaluasi, seleksi, *crossover* dan mutasi yang kemudian akan menghasilkan kromosom-kromosom baru untuk generasi yang selanjutnya. Proses ini akan berulang sampai sejumlah generasi yang telah ditetapkan sebelumnya.

4.1.5. Hasil Eksperimen Menggunakan Algoritma Naïve Bayes dan *Genetic Algorithm*

Untuk mendapatkan model yang baik penulis mencoba menyesuaikan beberapa nilai agar mendapatkan hasil akurasi yang tinggi. Untuk genetic algorithm indikator yang disesuaikan adalah *population size*= 10, *p initialize*, *p crossover* dan *generate* = 1.0. sedangkan yang diuji coba untuk meningkatkan nialai akurasi dan AUC adalah merubah nilai *P initialize dan P crossover* . Tabel indikator dan hasil pengujian accuracy dan AUC dapat dilihat pada tabel 4.11

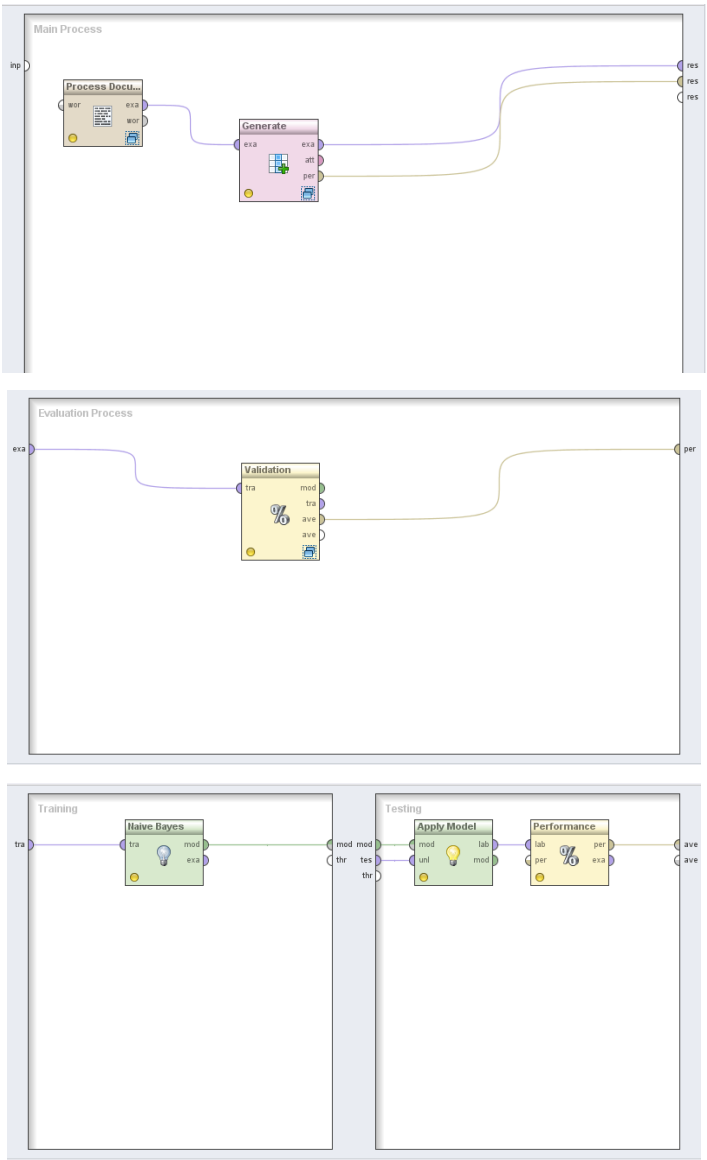
Tabel 4.11 Tabel Indikator Dan hasil pengujian

P initialize	***P Crossover***	***Accuracy***	***AUC***
0.5	0.5	93.00%	0.550
0.6	0.6	93.00%	0.550
0.7	**0.7**	**94.50%**	**0.635**
0.8	0.8	94.00%	0.591
0.9	0.9	92.50%	0.500
1.0	1.0	94.50%	0.500

Dalam menyesuaikan nilai indikator pada *genetic Algorithm*, akurasi dan *AUC* yang paling tinggi diperoleh dengan kombinasi *population size* = 10, *p initialize* = 0,7 *p crossover* = 0.7 dan *generate* 1.0. hasil akurasi mencapai 94.50

% dan nilai *AUC* 0.635. didalam proses pengujian *Genetic Algorithm* hanya merubah nilai *p initialize* dan *p crossover* karena jika indikator lain ikut dirubah nilainya, dapat menyebabkan proses pengolahan data menjadi semakin lama.

Desain model Naïve Bayes dan pemilihan fitur *Genetic Algorithm* ini dapat dilihat pada gambar 4.3



Gambar 4.3 Desain model Naïve Bayes dan pemilihan fitur *Genetic Algorithm* penggunakan *RapidMiner*

4.2. Pembahasan

Dengan memiliki model klasifikasi teks pada *review*, pembaca dapat mengidentifikasi dengan mudah mana yang termasuk *review* yang negatif dan *review* yang positif tanpa harus membaca satu persatu disetiap kalimatnya. Data *review* yang sudah ada, ditentukan terlebih dahulu kata yang mengandung kata sentimen positif dan kata sentimen negatif pada setiap *review* yang sering muncul, kemudian dipisahkan menjadi kata-kata, lalu diberikan bobot pada masing-masing kata tersebut. Dapat dilihat kata mana saja yang berhubungan dengan sentimen yang sering muncul dan memiliki bobot yang paling tinggi. Dengan demikian dapat diketahui *review* tersebut masuk kedalam *review* positif ataun *review* negatif.

Dalam penelitian ini hasil pengujian model akan dibahas melalui *confusion matrix* untuk menunjukan seberapa baik model dibentuk. Tanpa menggunakan metode pemilihan fitur, algoritma Naïve Bayes sendiri sudah menghasilkan akurasi sebesar 90.50% dan nilai *AUC* 0.500. Ditambahkan metode pemilihan fitur menggunakan *Genetic Algorithm* untuk meningkatkan nilai akurasi dan nilai *AUC* nya. Setelah menggunakan metode pemilihan fitur jenis *wrapper* dengan *Genetic Algorithm* , akurasi algoritma Naïve Bayes meningkat menjadi 94.50% dan nilai AUC 0.586 seperti bisa dilihat pada tabel 4.12

Tabel 4.12 Model algoritma Naïve Bayes sebelum dan sesudah menggunakan metode pemilihan fitur

	Algoritma Naïve Bayes	**Algoritma Naïve Bayes dan *Genetic Algorithm***
Prediksi klasifikasi *review* positif	**85**	**94**
Prediksi klasifikasi *review* negatif	**96**	**95**
Akurasi Model	**90.50%**	**94.50%**
AUC	**0.500**	**0.635**

4.2.1. Pengukuran dengan *Confusion* Matrix

Pengukuran *Confusion Matrix* yang akan ditampilkan dalam penelitian ini adalah perbandingan dari hasil akurasi model Naïve Bayes sebelum menggunakan metode pemilihan fitur yang bisa dilihat pada gambar 4.4 Dan setelah menggunakan metode pemiliha fitur jenis wrapper yaitu *Genetic Algoritm* yang bisa dilihat pada gambar 4.5

accuracy: 90.50% +/- 9.07% (mikro: 90.50%)

	true Class Negatif	true Class Positif	class precision
pred. Class Negatif	96	15	86.49%
pred. Class Positif	4	85	95.51%
class recall	96.00%	85.00%	

Gambar 4.4 *Confusion matrix* model Naïve Bayes Sebelum menggunakan *Genetic Algorithm*

$$Accuracy = \frac{181}{96 + 15 + 4 + 85} = 0.905 \text{ x } 100\% = 90.50\%$$

Pada gambar 4.4 Dapat dilihat hasil pengklasifikasian *review* hotel menggunakan algoritma Naïve Bayes dari 100 review positif dan 100 *review* negatif yang diambil dari situs www.tripadvisor.com. Menghasilkan nilai akurasi sebesar 90.50 %

accuracy: 94.50% +/- 6.10% (mikro: 94.50%)

	true Class Negatif	true Class Positif	class precision
pred. Class Negatif	95	6	94.06%
pred. Class Positif	5	94	94.95%
class recall	95.00%	94.00%	

Gambar 4.5 *Confusion Matrix* model Naïve Bayes Sesudah menggunakan *Genetic Algorithm*

$$Accuracy = \frac{189}{95+6+5+94} = 0.945 \text{ x } 100\% = 94.50\%$$

Pada gambar 4.5 Dapat dilihat hasil pengklasifikasian *review* hotel dengan menggunakan algoritma Naïve Bayes dan penambahan metode pemilihan fitur

Geneteic Algorithm dari 100 *review* positif dan 100 *review* negatif yang diambil dari situs www.tripadvisor.com menghasilkan nilai akurasi sebesar 94.50% .

Nilai akurasi ini mengalami peningkatan sebesar 4% dari penggunaan algoritma Naïve Bayes sebelum menambahkan metote pemilihan fitur *genetic Algorithm*

4.2.2. *Curva ROC (Receiver Operating Characteristic*)

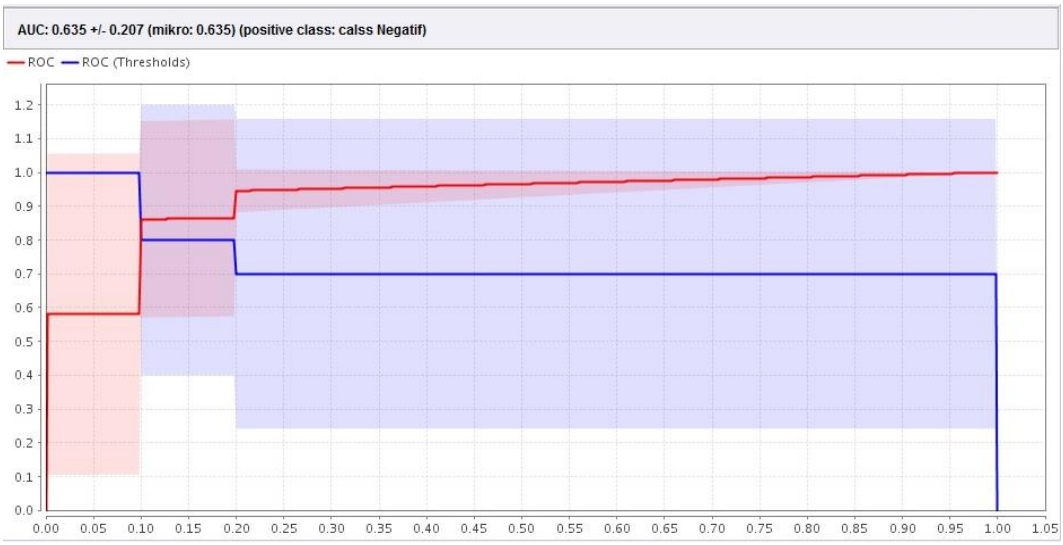
ROC merupakan sebuah grafik untuk menilai hasil prediksi. Dalam model klasifikasi *ROC* merupakan teknik visualisasi, pengaturan dan pemilihan klasifikasi berdasarkan hasil performance. *Kurva ROC* merupakan *tools* untuk membanfingkan model klasifikasi (Gorunescu, 2011).

Dari grafik *ROC* didapatkan pula nilai *AUC* (*Area Under Curve)* untuk menganalisa hasil prediksi klasifikasi. Penentuan hasil prediksi klasifikasi dilihat dari batasan nilai *AUC* sebagai berikut (Gorunescu, 2011):

a. Nilai *AUC* 0.90-1.00 = *excellent classification*
b. Nilai *AUC* 0.80-0.90 = *good classification*
c. Nilai *AUC* 0.70-0.80 = *fair classification*
d. Nilai *AUC* 0.60-0.70 = *poor classification*
e. Nilai *AUC* 0.50-0.60 = *failure*

Berikut adalah tampilan kurva *ROC* yang akan dihitung nilai *AUC*–nya dari 100 *review* positif dan 100 *review* negatif yang diambil dari situs www.tripadvisor.com. Gambar 4.6 adalah *kurva ROC* untuk model Naïve Bayes sebelum menggunakan metode pemilihan fitur *Genetic Algorithm* dan gambar 4.7 adalah kurva *ROC* setelah menggunakan metode pemilihan fitur *Genetic Algorithm.*

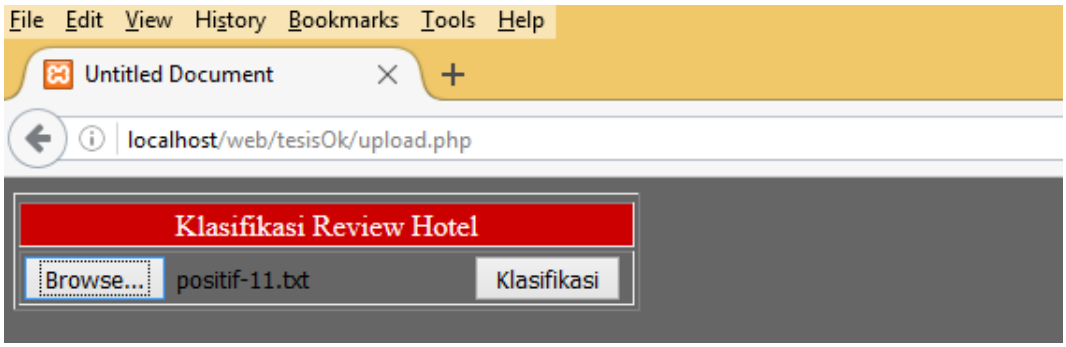
Gambar 4.6 *Kurva ROC* Model Naïve Bayes



Gambar 4.7 *Kurva ROC* Model Naïve Bayes dan *Genetic Algorithm*

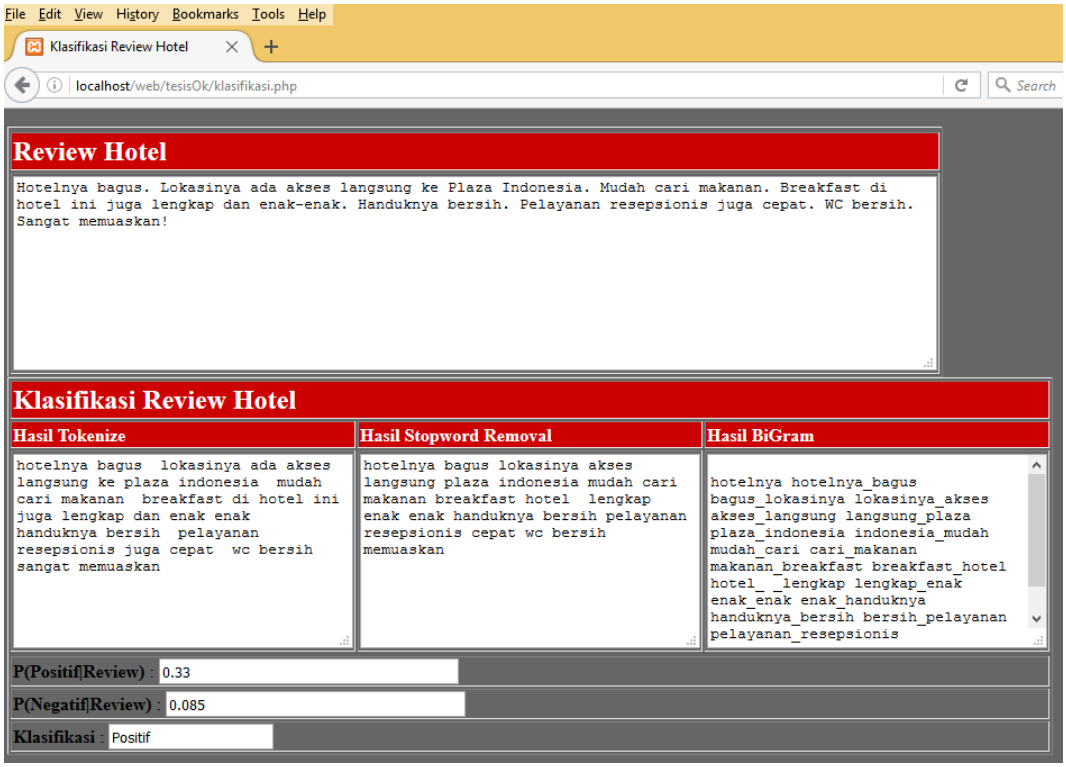
4.3. Pengembangan *Prototype* Pengklasifikasian Review

Penulis membuat aplikasi untuk untuk menguji model yang sudah ada menggunakan *dataset* yang ada untuk diketahui *class* positif dan negatifnya dalam *review* hotel. Hasil klasifikasi dari penelitian akan diterapakan kedalam pembuatan aplikasi untuk klasifikasi *review* hotel menggunakan perangkat lunak *dreamweaver* CS 3 menggunakan bahasa pemrograman php, sehingga dapat mengetahui hasil klasifikasi termasuk kedalam *class* positif atau termasuk kedalam klasifikasi *class* negatif. Seperti gambar 4.8 dibawah ini



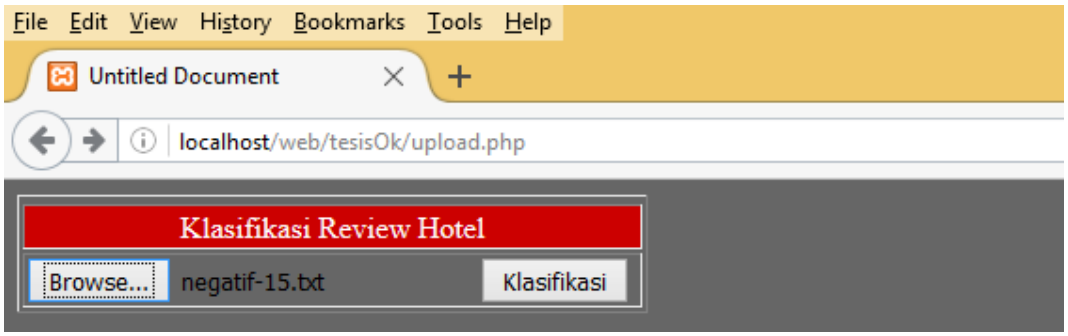
Gambar 4.8 Tampilan aplikasi mengklasifikasi *review* positif

Gambar 4.8 merupakan tampilan aplikasi untuk mengklasifikasi *review* hotel, dimana dalam aplikasi ini terdapat komentar yang diindikasi sebagai *review* positif. Pada aplikasi tersebut terlebih dahulu *upload* file .txt yang akan di klasifikasi, misalkan positif-11.txt setelah itu tekan tombol klasifikasi, maka akan tampil hasil dari klasifikasi.



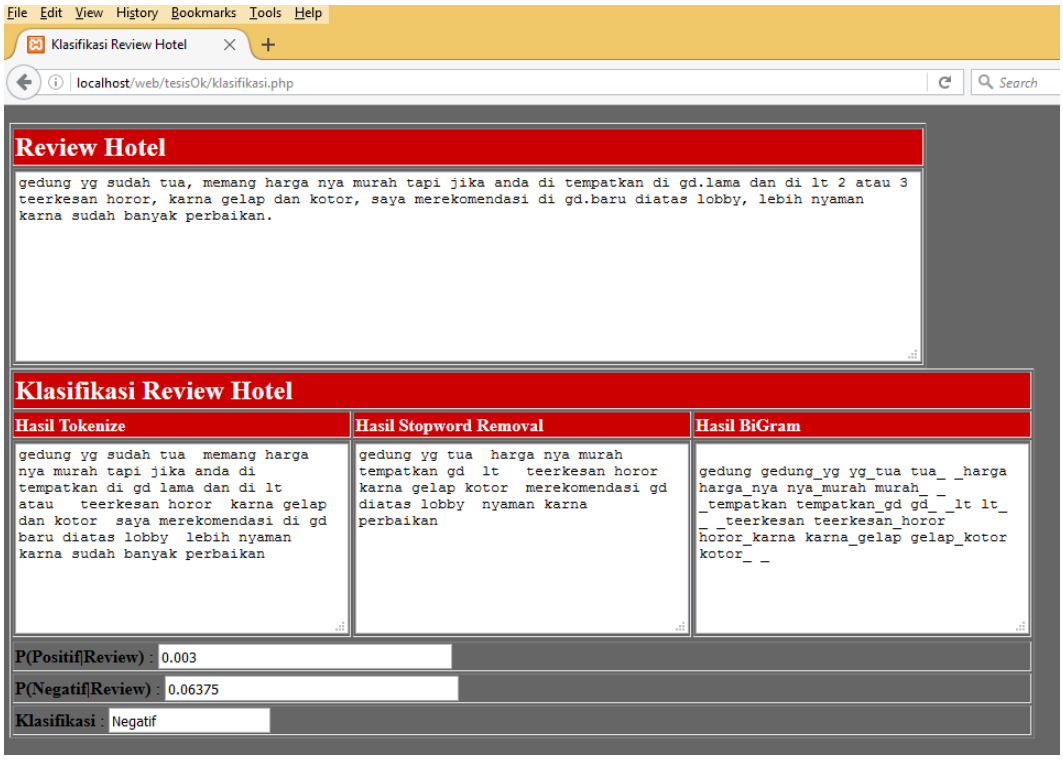
Gambar 4.9 Tampilan aplikasi Hasil mengklasifikasi *Review* Positif

Gambar 4.9 merupakan hasil dari klasifikasi gambar 4.8. Dimana pada gambar ini *review* hotel yang di *input* dapat diketahui hasil klasifikasinya dan menampilkan *review* hotel, hasil *tokenize*, hasil *Stopword* dan hasil *Bi-Gram*. Hasil *inputan* pada gambar 4.8 merupakan *review* positif , dikarena nilai p(positif|*Review*) lebih besar dari P(Negatif|*Review*)



Gambar 4.10 Tampilan aplikasi mengklasifikasi *Review* Negatif

Gambar 4.10 merupakan tampilan aplikasi untuk mengklasifikasi *review* hotel, dimana dalam aplikasi ini terdapat komentar yang diindikasi sebagai *review* negatif. Pada aplikasi tersebut terlebih dahulu *upload file* .txt yang akan di klasifikasi, misalkan negatif-15.txt setelah itu tekan tombol klasifikasi, maka akan tampil hasil dari klasifikasi seperti gambar 4.11 dibawah ini



Gambar 4.11 Tampilan aplikasi mengklasifikasi Hasil *Review* Negatif

Gambar 4.11 merupakan hasil dari klasifikasi gambar 4.8. Dimana pada gambar ini *review* hotel yang di *input* dapat diketahui hasil klasifikasinya dan menampilkan *review* hotel, hasil *tokenize*, hasil *Stopword* dan hasil *Bi-Gram*. Hasil *inputan* pada gambar 4.10 merupakan *review* negatif , dikarena nilai P(Negatif|*Review*) lebih besar dari p(Positif|*Review*)

Pada tabel 4.13 adalah hasil dari klasifikasi *review* hotel dengan menggunakan *prototype* yang sudah di buat.

Tabel 4.13 Hasil uji data klasifikasi *review* hotel menggunakan aplikasi

Nama dokumen	***Review***	**Hasil *Tokenization***	***Hasil Stopword***	***Hasil Bi-Gram***	***Class***
Positf-11.txt	Hotelnya bagus. Lokasinya ada akses langsung ke Plaza Indonesia. Mudah cari makanan. Breakfast di hotel ini juga lengkap dan enak-enak. Handuknya bersih. Pelayanan resepsionis juga cepat. WC bersih. Sangat memuaskan!	Hotelnya bagus Lokasinya ada akses langsung ke Plaza Indonesia Mudah cari makanan Breakfast di hotel ini juga lengkap dan enak enak Handuknya bersih Pelayanan resepsionis juga cepat WC bersih Sangat memuaskan	Hotelnya bagus Lokasinya akses langsung Plaza Indonesia Mudah cari makanan Breakfast hotel lengkap enak enak Handuknya bersih Pelayanan resepsionis cepat WC bersih memuaskan	Hotelnya Hotelnya_bagus bagus bagus_Lokasinya Lokasinya Lokasinya_akses akses akses_langsung langsung langsung_Plaza Plaza Plaza_Indonesia Indonesia Indonesia_Mudah Mudah Mudah_cari cari cari_makanan makanan makanan_Breakfast Breakfast Breakfast_hotel hotel hotel_lengkap lengkap lengkap_enak enak enak_enak enak enak_Handuknya Handuknya Handuknya_bersih bersih bersih_Pelayanan Pelayanan Pelayanan_resepsionis resepsionis resepsionis_cepat cepat cepat_WC WC WC_bersih bersih bersih_memuaskan memuaskan	Positif
Positif-12.txt	Grand Hyatt Jakarta (GHJ) mempunyai lokasi terbaik di Jakarta. - Kamar luas dan nyaman - Kamar mandi bagus & lega - Disediakan buah di kamar pada kedatangan - Staff sangat ramah dan sigap membantu - Makan pagi sangat lezat - Lokasi Terbaik di Jakarta - Terletak di atas Plaza	Grand Hyatt Jakarta GHJ mempunyai lokasi terbaik di Jakarta Kamar luas dan nyaman Kamar mandi bagus lega Disediakan buah di kamar pada kedatangan Staff sangat ramah dan sigap membantu Makan pagi sangat lezat Lokasi Terbaik di Jakarta Terletak di atas Plaza Indonesia sehingga mudah untuk berbelanja kebutuhan	Grand Hyatt Jakarta GHJ mempunyai lokasi terbaik Jakarta Kamar luas nyaman Kamar mandi bagus lega Disediakan buah kamar kedatangan Staff ramah sigap membantu Makan pagi lezat Lokasi Terbaik Jakarta Terletak atas Plaza Indonesia mudah untuk berbelanja kebutuhan untuk makan luar Tips Minta kamar persis menghadap air mancur HI untuk	Grand Grand_Hyatt Hyatt Hyatt_Jakarta Jakarta Jakarta_GHJ GHJ GHJ_mempunyai mempunyai mempunyai_lokasi lokasi lokasi_terbaik terbaik terbaik_Jakarta Jakarta Jakarta_Kamar Kamar Kamar_luas luas luas_nyaman nyaman nyaman_Kamar Kamar Kamar_mandi mandi mandi_bagus bagus bagus_lega lega	positif

Nama dokumen	*Review*	Hasil *Tokenization*	*Hasil Stopword*	*Hasil Bi-Gram*	*Class*
	Indonesia, sehingga mudah untuk berbelanja kebutuhan atau untuk makan di luar. Tips: - Minta kamar yang persis menghadap air mancur HI untuk pemandangan terbaik	atau untuk makan di luar Tips Minta kamar yang persis menghadap air mancur HI untuk pemandangan terbaik	pemandangan terbaik	lega_Disediakan Disediakan Disediakan_buah buah buah_kamar kamar kamar_kedatangan kedatangan kedatangan_Staff Staff Staff_ramah ramah ramah_sigap sigap sigap_membantu membantu membantu_Makan Makan Makan_pagi pagi pagi_lezat lezat lezat_Lokasi Lokasi Lokasi_Terbaik Terbaik Terbaik_Jakarta Jakarta Jakarta_Terletak Terletak Terletak_atas atas atas_Plaza Plaza Plaza_Indonesia Indonesia Indonesia_mudah mudah mudah_untuk untuk untuk_berbelanja berbelanja berbelanja_kebutuhan kebutuhan kebutuhan_untuk untuk untuk_makan makan makan_luar luar luar_Tips Tips Tips_Minta Minta Minta_kamar kamar kamar_persis persis persis_menghadap menghadap menghadap_air air air_mancur mancur mancur_HI HI HI_untuk untuk untuk_pemandangan pemandangan pemandangan_terbaik terbaik	
Positif-13.txt	Berharap dapat datang kembali dengan keluarga. Namun untuk dapat berbelanja sebaiknya menggunakan kendaraan karena aga jauh dari pusat pertokoan. Suasana hotel	berharap dapat datang kembali dengan keluarga Namun untuk dapat berbelanja sebaiknya menggunakan kendaraan karena aga jauh dari pusat	berharap datang kembali keluarga untuk berbelanja sebaiknya menggunakan kendaraan aga jauh pusat pertokoan Suasana hotel nyaman bersih pelayanan ramah	berharap berharap_datang datang datang_kembali kembali kembali_keluarga keluarga keluarga_untuk untuk untuk_berbelanja	Positif

Nama dokumen	*Review*	Hasil *Tokenization*	*Hasil Stopword*	*Hasil Bi-Gram*	*Class*
	sangat nyaman, bersih, dan pelayanan yang ramah	pertokoan Suasana hotel sangat nyaman bersih dan pelayanan yang ramah		berbelanja berbelanja_sebaiknya sebaiknya sebaiknya_menggunakan menggunakan menggunakan_kendaraan kendaraan kendaraan_aga aga aga_jauh jauh jauh_pusat pusat pusat_pertokoan pertokoan pertokoan_Suasana Suasana Suasana_hotel hotel hotel_nyaman nyaman nyaman_bersih bersih bersih_pelayanan pelayanan pelayanan_ramah ramah	
Positif-14.txt	Sabtu, 30 April 2016 adalah pengalaman pertama saya menginap di hotel Shangri-La Jakarta. Dari pengalaman pertama tersebut, saya mengalami pengalaman luar biasa. Yang bahkan pertama kali nya saya menemukan hotel yang sangat memperhatikan detail, terutama dalam hal keperluan bisnis. Untuk hal umum seperti kebersihan, kenyamanan, saya memberikan nilai 9,5/10. Dan untuk keramahan saya berikan 10/10. Karena (maaf), biasanya hotel bintang 5 di Jakarta sedikit "biasa" saja menganggap customer lokal, terutama yang masih cukup muda. Tetapi, Shangri-la mematahkan anggapan saya. Mereka sangat ramah. Thumbs Up. Dan sesuai judul, detail nya menjadi hal yang	Sabtu April adalah pengalaman pertama saya menginap di hotel Shangri La Jakarta Dari pengalaman pertama tersebut saya mengalami pengalaman luar biasa Yang bahkan pertama kali nya saya menemukan hotel yang sangat memperhatikan detail terutama dalam hal keperluan bisnis Untuk hal umum seperti kebersihan kenyamanan saya memberikan nilai Dan untuk keramahan saya berikan Karena maaf biasanya hotel bintang di Jakarta sedikit biasa saja menganggap customer lokal terutama yang masih cukup muda Tetapi Shangri la mematahkan anggapan saya Mereka sangat ramah Thumbs Up Dan sesuai judul detail nya menjadi hal yang sangat saya perhatikan Di dalam kamar terdapat detail	Sabtu April pengalaman pertama menginap hotel Shangri La Jakarta pengalaman pertama mengalami pengalaman luar pertama kali nya menemukan hotel memperhatikan detail terutama keperluan bisnis Untuk umum kebersihan kenyamanan memberikan nilai untuk keramahan berikan maaf hotel bintang Jakarta menganggap customer lokal terutama cukup muda Shangri la mematahkan anggapan ramah Thumbs Up sesuai judul detail nya menjadi perhatikan kamar terdapat detail jarang dapatkan hotel semir sepatu setrika dock speaker iphone PERALATAN KERJA Peralatan kerja lengkap strappler gunting penjepit kertas selotape senter pebisnis membantu detail sebenernya murah jarang hotel menyediakan	Sabtu Sabtu_April April April_pengalaman pengalaman pengalaman_pertama pertama pertama_menginap menginap menginap_hotel hotel hotel_Shangri Shangri Shangri_La La La_Jakarta Jakarta Jakarta_pengalaman pengalaman pengalaman_pertama pertama pertama_mengalami mengalami mengalami_pengalaman pengalaman pengalaman_luar luar luar_pertama pertama pertama_kali kali kali_nya nya nya_menemukan menemukan menemukan_hotel hotel hotel_memperhatikan memperhatikan memperhatikan_detail detail detail_terutama terutama	Positif

Nama dokumen	*Review*	Hasil *Tokenization*	*Hasil Stopword*	*Hasil Bi-Gram*	*Class*
	sangat saya perhatikan. Di dalam kamar, terdapat detail yang jarang saya dapatkan di hotel lain. Seperti semir sepatu, setrika, dock speaker iphone, dan PERALATAN KERJA. Peralatan kerja ini lengkap seperti strappler, gunting, penjepit kertas, selotape, bahkan senter. Buat para pebisnis, hal ini sangat membantu tentunya. Dan detail ini, sebenernya murah, tetapi jarang hotel yang menyediakan sebagai standar fasilitas kamar. Menurut saya, dengan pelayanan sedetail itu, secara penuh saya berikan nilai 9,8/10	yang jarang saya dapatkan di hotel lain Seperti semir sepatu setrika dock speaker iphone dan PERALATAN KERJA Peralatan kerja ini lengkap seperti strappler gunting penjepit kertas selotape bahkan senter Buat para pebisnis hal ini sangat membantu tentunya Dan detail ini sebenernya murah tetapi jarang hotel yang menyediakan sebagai standar fasilitas kamar Menurut saya dengan pelayanan sedetail itu secara penuh saya berikan nilai	standar fasilitas kamar Menurut pelayanan sedetail secara penuh berikan nilai	terutama_keperluan keperluan keperluan_bisnis bisnis bisnis_Untuk Untuk Untuk_umum umum umum_kebersihan kebersihan kebersihan_kenyamanan kenyamanan kenyamanan_memberikan memberikan memberikan_nilai nilai nilai_untuk untuk untuk_keramahan keramahan keramahan_berikan berikan berikan_maaf maaf maaf_hotel hotel hotel_bintang bintang bintang_Jakarta Jakarta Jakarta_menganggap menganggap menganggap_customer customer customer_lokal lokal lokal_terutama terutama terutama_cukup cukup cukup_muda muda muda_Shangri Shangri Shangri_la la la_mematahkan mematahkan mematahkan_anggapan anggapan anggapan_ramah ramah ramah_Thumbs Thumbs Thumbs_Up Up Up_sesuai sesuai sesuai_judul judul judul_detail detail detail_nya nya nya_menjadi menjadi menjadi_perhatikan perhatikan perhatikan_kamar kamar kamar_terdapat terdapat terdapat_detail detail detail_jarang jarang jarang_dapatkan dapatkan dapatkan_hotel hotel hotel_semir semir semir_sepatu sepatu	

Nama dokumen	*Review*	Hasil *Tokenization*	*Hasil Stopword*	*Hasil Bi-Gram*	*Class*
				sepatu_setrika setrika setrika_dock dock dock_speaker speaker speaker_iphone iphone iphone_PERALATAN PERALATAN PERALATAN_KERJA KERJA KERJA_Peralatan Peralatan Peralatan_kerja kerja kerja_lengkap lengkap lengkap_strappler strappler strappler_gunting gunting gunting_penjepit penjepit penjepit_kertas kertas kertas_selotape selotape selotape_senter senter senter_pebisnis pebisnis pebisnis_membantu membantu membantu_detail detail detail_sebenernya sebenernya sebenernya_murah murah murah_jarang jarang jarang_hotel hotel hotel_menyediakan menyediakan menyediakan_standar standar standar_fasilitas fasilitas fasilitas_kamar kamar kamar_Menurut Menurut Menurut_pelayanan pelayanan pelayanan_sedetail sedetail sedetail_secara secara secara_penuh penuh penuh_berikan berikan berikan_nilai nilai	
Positif-15.txt	Hotelnya nyaman untuk keluarga dan pemandangan dari kamar ke taman / kolam renang sangat bagus, terlihat	Hotelnya nyaman untuk keluarga dan pemandangan dari kamar ke taman kolam renang sangat bagus	Hotelnya nyaman untuk keluarga pemandangan kamar taman kolam renang bagus terlihat jelas kolam berbentuk	Hotelnya Hotelnya_nyaman nyaman nyaman_untuk untuk untuk_keluarga keluarga	Positif

Nama dokumen	*Review*	Hasil *Tokenization*	*Hasil Stopword*	*Hasil Bi-Gram*	*Class*
	jelas kolam berbentuk kupu-kupu.	terlihat jelas kolam berbentuk kupu kupu	kupu kupu	keluarga_pemandangan pemandangan pemandangan_kamar kamar kamar_taman taman taman_kolam kolam kolam_renang renang renang_bagus bagus bagus_terlihat terlihat terlihat_jelas jelas jelas_kolam kolam kolam_berbentuk berbentuk berbentuk_kupu kupu kupu_kupu kupu	
Negatif-11.txt	Kamar dan pelayanannya sangat standard, ac dingin, tempat tidur bersih dan nyaman, dinding kamar tidak soundproof, tidak ada keset kamar mandi, parkiran kecil hanya muat beberapa mobil dan motor. secara keseluruhan, sesuai dengan harga	Kamar dan pelayanannya sangat standard ac dingin tempat tidur bersih dan nyaman dinding kamar tidak soundproof tidak ada keset kamar mandi parkiran kecil hanya muat beberapa mobil dan motor secara keseluruhan sesuai dengan harga	Kamar pelayanannya standard ac dingin tempat tidur bersih nyaman dinding kamar soundproof keset kamar mandi parkiran muat mobil motor secara keseluruhan sesuai harga	Kamar Kamar_pelayanannya pelayanannya pelayanannya_standard standard standard_ac ac ac_dingin dingin dingin_tempat tempat tempat_tidur tidur tidur_bersih bersih bersih_nyaman nyaman nyaman_dinding dinding dinding_kamar kamar kamar_soundproof soundproof soundproof_keset keset keset_kamar kamar kamar_mandi mandi mandi_parkiran parkiran parkiran_muat muat muat_mobil mobil mobil_motor motor motor_secara secara secara_keseluruhan keseluruhan keseluruhan_sesuai sesuai sesuai_harga harga	positif
Negatif-12.txt	Hotel yang cukup buruk dari kejauhan, dan lebih buruk saat didekati. Lobinya sangat butuh pembaruan agar menjadikannya lebih mengundang. Saya menginap satu malam karena menanti penerbangan.	Hotel yang cukup buruk dari kejauhan dan lebih buruk saat didekati Lobinya sangat butuh pembaruan agar menjadikannya lebih mengundang Saya menginap satu malam karena menanti	Hotel cukup buruk kejauhan buruk didekati Lobinya butuh pembaruan menjadikannya mengundang menginap satu malam menanti penerbangan Memakan waktu jauh dugaan Bandara Jakarta	Hotel Hotel_cukup cukup cukup_buruk buruk buruk_kejauhan kejauhan kejauhan_buruk buruk buruk_didekati didekati didekati_Lobinya Lobinya Lobinya_butuh butuh butuh_pembaruan	Negatif

Nama dokumen	*Review*	Hasil *Tokenization*	*Hasil Stopword*	*Hasil Bi-Gram*	*Class*
	Memakan waktu jauh lebih lama daripada dugaan saya dari Bandara Jakarta. Ini kesalahan yang mahal sebab saya tak bisa mendapatkan taksi dari bandara menuju hotel lebih murah daripada Rp 250.000,00 (perjalanan balik Rp 90.000,00). Ada banyak hotel yang jauh lebih dekat dengan bandara, banyak yang menyertakan angkutan gratis dari dan menuju bandara. Saat memperhitungkan ongkos taksi, khususnya bagi wisatawan tunggal, ini mungkin pilihan lebih baik. Saya membayar sekitar $30 untuk hotel ini semalam, tidak termasuk sarapan. Kamar-kamarnya butuh diperbarui, namun ranjangnya nyaman. Pancurannya menyedihkan. WiFi tak berfungsi di kamar, namun berfungsi di lobi. Mengesalkan. Saya bisa mendengar jelas kebisingan jalan di bawah kamar saya. Posisi hotel ini buruk. Tak ada apa-apa lagi di sekitarnya. Saya tak bisa menemukan restoran. Ada pusat kebugaran dan sauna di hotel, meskipun tak saya manfaatkan. Stafnya dengan sigap menyiapkan taksi untuk mengantar saya ke bandara pagi-pagi. Secara keseluruhan, hindari sebisa mungkin	penerbangan Memakan waktu jauh lebih lama daripada dugaan saya dari Bandara Jakarta Ini kesalahan yang mahal sebab saya tak bisa mendapatkan taksi dari bandara menuju hotel lebih murah daripada Rp perjalanan balik Rp Ada banyak hotel yang jauh lebih dekat dengan bandara banyak yang menyertakan angkutan gratis dari dan menuju bandara Saat memperhitungkan ongkos taksi khususnya bagi wisatawan tunggal ini mungkin pilihan lebih baik Saya membayar sekitar untuk hotel ini semalam tidak termasuk sarapan Kamar kamarnya butuh diperbarui namun ranjangnya nyaman Pancurannya menyedihkan WiFi tak berfungsi di kamar namun berfungsi di lobi Mengesalkan Saya bisa mendengar jelas kebisingan jalan di bawah kamar saya Posisi hotel ini buruk Tak ada apa apa lagi di sekitarnya Saya tak bisa menemukan restoran Ada pusat kebugaran dan sauna di hotel meskipun tak saya manfaatkan Stafnya dengan sigap menyiapkan taksi untuk mengantar saya ke bandara pagi pagi Secara keseluruhan hindari sebisa mungkin karena ada banyak pilihan lebih baik	kesalahan mahal mendapatkan taksi bandara menuju hotel murah Rp perjalanan balik Rp hotel jauh bandara menyertakan angkutan gratis menuju bandara memperhitungkan ongkos taksi wisatawan tunggal pilihan baik membayar untuk hotel semalam termasuk sarapan Kamar kamarnya butuh diperbarui ranjangnya nyaman Pancurannya menyedihkan WiFi berfungsi kamar berfungsi lobi Mengesalkan mendengar jelas kebisingan jalan bawah kamar Posisi hotel buruk menemukan restoran pusat kebugaran sauna hotel manfaatkan Stafnya sigap menyiapkan taksi untuk mengantar bandara pagi pagi Secara keseluruhan hindari sebisa pilihan baik	pembaruan pembaruan_menjadikannya menjadikannya menjadikannya_mengundang mengundang mengundang_menginap menginap menginap_satu satu satu_malam malam malam_menanti menanti menanti_penerbangan penerbangan penerbangan_Memakan Memakan Memakan_waktu waktu waktu_jauh jauh jauh_dugaan dugaan dugaan_Bandara Bandara Bandara_Jakarta Jakarta Jakarta_kesalahan kesalahan kesalahan_mahal mahal mahal_mendapatkan mendapatkan mendapatkan_taksi taksi taksi_bandara bandara bandara_menuju menuju menuju_hotel hotel hotel_murah murah murah_Rp Rp Rp_perjalanan perjalanan perjalanan_balik balik balik_Rp Rp Rp_hotel hotel hotel_jauh jauh jauh_bandara bandara bandara_menyertakan menyertakan menyertakan_angkutan angkutan angkutan_gratis gratis gratis_menuju menuju menuju_bandara bandara bandara_memperhitungkan memperhitungkan memperhitungkan_ongkos ongkos ongkos_taksi taksi	

Nama dokumen	*Review*	Hasil *Tokenization*	*Hasil Stopword*	*Hasil Bi-Gram*	*Class*
	karena ada banyak pilihan lebih baik.			taksi_wisatawan wisatawan wisatawan_tunggal tunggal tunggal_pilihan pilihan pilihan_baik baik baik_membayar membayar membayar_untuk untuk untuk_hotel hotel hotel_semalam semalam semalam_termasuk termasuk termasuk_sarapan sarapan sarapan_Kamar Kamar Kamar_kamarnya kamarnya kamarnya_butuh butuh butuh_diperbarui diperbarui diperbarui_ranjangnya ranjangnya ranjangnya_nyaman nyaman nyaman_Pancurannya Pancurannya Pancurannya_menyedihkan menyedihkan menyedihkan_WiFi WiFi WiFi_berfungsi berfungsi berfungsi_kamar kamar kamar_berfungsi berfungsi berfungsi_lobi lobi lobi_Mengesalkan Mengesalkan Mengesalkan_mendengar mendengar mendengar_jelas jelas jelas_kebisingan kebisingan kebisingan_jalan jalan jalan_bawah bawah bawah_kamar kamar kamar_Posisi Posisi Posisi_hotel hotel hotel_buruk buruk buruk_menemukan menemukan menemukan_restoran restoran restoran_pusat pusat pusat_kebugaran	

Nama dokumen	*Review*	Hasil *Tokenization*	*Hasil Stopword*	*Hasil Bi-Gram*	*Class*
				kebugaran kebugaran_sauna sauna sauna_hotel hotel hotel_manfaatkan manfaatkan manfaatkan_Stafnya Stafnya Stafnya_sigap sigap sigap_menyiapkan menyiapkan menyiapkan_taksi taksi taksi_untuk untuk untuk_mengantar mengantar mengantar_bandara bandara bandara_pagi pagi pagi_pagi pagi pagi_Secara Secara Secara_keseluruhan keseluruhan keseluruhan_hindari hindari hindari_sebisa sebisa sebisa_pilihan pilihan pilihan_baik baik	
Negatif-13.txt	Jangan pernah menginap di hotel ini! Hotel ini sangat buruk. Semua stafnya sangat menyebalkan! Mulai dari resepsionis, tata graha, dan layanan kebersihan menyebalkan! Staf yang bekerja di hotel ini tak berpendidikan. Mereka bahkan tak tahu cara tersenyum untuk menyapa pelanggan. Jangan tertipu harganya. Hanya karena murah maka Anda menginap. Anda akan menyesal dalam hidup Anda MENGAPA bisa memilih hotel ini. Bahkan mereka minta Anda membayar jika Anda ingin memakai handuk. Lebih baik Anda tambahkan sedikit lagi untuk mendapatkan hotel yang lebih baik. Kisah paling MENAKUTKAN yang baru saya ketahui	Jangan pernah menginap di hotel ini Hotel ini sangat buruk Semua stafnya sangat menyebalkan Mulai dari resepsionis tata graha dan layanan kebersihan menyebalkan Staf yang bekerja di hotel ini tak berpendidikan Mereka bahkan tak tahu cara tersenyum untuk menyapa pelanggan Jangan tertipu harganya Hanya karena murah maka Anda menginap Anda akan menyesal dalam hidup Anda MENGAPA bisa memilih hotel ini Bahkan mereka minta Anda membayar jika Anda ingin memakai handuk Lebih baik Anda tambahkan sedikit lagi untuk mendapatkan hotel yang lebih baik Kisah	menginap hotel Hotel buruk stafnya menyebalkan Mulai resepsionis tata graha layanan kebersihan menyebalkan Staf bekerja hotel berpendidikan tahu cara tersenyum untuk menyapa pelanggan tertipu harganya murah menginap menyesal hidup memilih hotel minta membayar memakai handuk baik tambahkan untuk mendapatkan hotel baik Kisah MENAKUTKAN baru ketahui menginap orang restoran toko toko berkata hotal terbakar bangunan hangus kerusuhan Jakarta memperbaruinya beroperasi kembali puas menginap	menginap menginap_hotel hotel hotel_Hotel Hotel Hotel_buruk buruk buruk_stafnya stafnya stafnya_menyebalkan menyebalkan menyebalkan_Mulai Mulai Mulai_resepsionis resepsionis resepsionis_tata tata tata_graha graha graha_layanan layanan layanan_kebersihan kebersihan kebersihan_menyebalkan menyebalkan menyebalkan_Staf Staf Staf_bekerja bekerja bekerja_hotel hotel hotel_berpendidikan berpendidikan berpendidikan_tahu tahu tahu_cara cara cara_tersenyum tersenyum tersenyum_untuk untuk untuk_menyapa	Negatif

Nama dokumen	*Review*	Hasil *Tokenization*	*Hasil Stopword*	*Hasil Bi-Gram*	*Class*
	saat menginap di sana, beberapa orang di restoran dan toko-toko di sekitarnya berkata, hotal ini sebelumnya terbakar. Seluruh bangunan hangus saat kerusuhan 1998 di Jakarta. Mereka memperbaruinya dan beroperasi kembali. Saya sangat tak puas saat menginap di sana.	paling MENAKUTKAN yang baru saya ketahui saat menginap di sana beberapa orang di restoran dan toko toko di sekitarnya berkata hotal ini sebelumnya terbakar Seluruh bangunan hangus saat kerusuhan di Jakarta Mereka memperbaruinya dan beroperasi kembali Saya sangat tak puas saat menginap di sana		menyapa menyapa_pelanggan pelanggan pelanggan_tertipu tertipu tertipu_harganya harganya harganya_murah murah murah_menginap menginap menginap_menyesal menyesal menyesal_hidup hidup hidup_memilih memilih memilih_hotel hotel hotel_minta minta minta_membayar membayar membayar_memakai memakai memakai_handuk handuk handuk_baik baik baik_tambahkan tambahkan tambahkan_untuk untuk untuk_mendapatkan mendapatkan mendapatkan_hotel hotel hotel_baik baik baik_Kisah Kisah Kisah_MENAKUTKAN MENAKUTKAN MENAKUTKAN_baru baru baru_ketahui ketahui ketahui_menginap menginap menginap_orang orang orang_restoran restoran restoran_toko toko toko_toko toko toko_berkata berkata berkata_hotal hotal hotal_terbakar terbakar terbakar_bangunan bangunan bangunan_hangus hangus hangus_kerusuhan kerusuhan kerusuhan_Jakarta Jakarta Jakarta_memperbaruin	

Nama dokumen	*Review*	Hasil *Tokenization*	*Hasil Stopword*	*Hasil Bi-Gram*	*Class*
				ya memperbaruinya memperbaruinya_beroperasi beroperasi beroperasi_kembali kembali kembali_puas puas puas_menginap menginap	
Negatif-14.txt	Apa yang dapat saya katakan tentang hotel ini. Foto-foto websitenya sangat mengundang; namun, setelah tiba, sama sekali tidak mengesankan saya. Setelah tiba di pintu masuk hotel, saya dikecewakan dengan pemandangan betapa "sederhana"nya hotel ini. Area lobi dan resepsionis tampak sangat "kuno" yang tidak memberi saya perasaan modernitas sebagaimana yang ditunjukkan di website. Setelah mendapatkan kunci kamar, bellboy membantu saya membawakan koper yang hanya berupa satu koper troli kecil. Mungkin, itu merupakan prosedur mereka bagi resepsionis untuk memberikan kunci pada bellboy dan membantu saya membawa koper ke kamar, yang saya rasa tidak perlu karena hanya berupa koper kecil yang dapat saya jinjing. Namun, itu tidak mengganggu saya begitu banyak karena saya juga bekerja di perhotelan, rasanya sangat hebat mendapatkan pelayanan seperti itu jika mempunyai bawaan yang banyak dan/atau jika bepergian bersama anak kecil dengan banyak barang bawaan. Begitu saya memasuki kamar,	Apa yang dapat saya katakan tentang hotel ini Foto foto websitenya sangat mengundang namun setelah tiba sama sekali tidak mengesankan saya Setelah tiba di pintu masuk hotel saya dikecewakan dengan pemandangan betapa sederhana nya hotel ini Area lobi dan resepsionis tampak sangat kuno yang tidak memberi saya perasaan modernitas sebagaimana yang ditunjukkan di website Setelah mendapatkan kunci kamar bellboy membantu saya membawakan koper yang hanya berupa satu koper troli kecil Mungkin itu merupakan prosedur mereka bagi resepsionis untuk memberikan kunci pada bellboy dan membantu saya membawa koper ke kamar yang saya rasa tidak perlu karena hanya berupa koper kecil yang dapat saya jinjing Namun itu tidak mengganggu saya begitu banyak karena saya juga bekerja di perhotelan rasanya sangat hebat mendapatkan pelayanan seperti itu jika mempunyai bawaan yang banyak	katakan hotel Foto foto websitenya mengundang tiba mengesankan tiba pintu masuk hotel dikecewakan pemandangan betapa sederhana nya hotel Area lobi resepsionis tampak kuno memberi perasaan modernitas ditunjukkan website mendapatkan kunci kamar bellboy membantu membawakan koper berupa satu koper troli prosedur resepsionis untuk memberikan kunci bellboy membantu membawa koper kamar rasa perlu berupa koper jinjing mengganggu bekerja perhotelan rasanya hebat mendapatkan pelayanan mempunyai bawaan bepergian anak barang bawaan memasuki kamar merasakan berda jaman kuno Dinding dindingnya dikotori jejak kaki lemari kayu meja punya bekas goresan kusam toiletnya pengap tempat pancurannya setengah dibatasi kaca dimana mandi airnya meluap bagian toilet memiliki celah untuk memisahkan kamar mandi area WC Pancurannya membuat frustasi waktu airnya dingin merasakan hangat panas memutar kerannya penuh Tempat tidurnya seindah harapkan mendapatkan tidur nyenyak seharian bekerja	katakan katakan_hotel hotel hotel_Foto Foto Foto_foto foto foto_websitenya websitenya websitenya_mengundang mengundang mengundang_tiba tiba tiba_mengesankan mengesankan mengesankan_tiba tiba tiba_pintu pintu pintu_masuk masuk masuk_hotel hotel hotel_dikecewakan dikecewakan dikecewakan_pemandangan pemandangan pemandangan_betapa betapa betapa_sederhana sederhana sederhana_nya nya nya_hotel hotel hotel_Area Area Area_lobi lobi lobi_resepsionis resepsionis resepsionis_tampak tampak tampak_kuno kuno kuno_memberi memberi memberi_perasaan perasaan perasaan_modernitas modernitas modernitas_ditunjukkan ditunjukkan ditunjukkan_website website website_mendapatkan mendapatkan mendapatkan_kunci kunci kunci_kamar kamar kamar_bellboy bellboy bellboy_membantu	Negatif

Nama dokumen	*Review*	Hasil *Tokenization*	*Hasil Stopword*	*Hasil Bi-Gram*	*Class*
	saya dapat merasakan bahwa saya berda di jaman "kuno". Dinding-dindingnya dikotori jejak kaki, lemari kayu dan meja punya bekas goresan dan kusam, toiletnya pengap dan tempat pancurannya hanya setengah dibatasi kaca 6" dimana setiap saat anda mandi, airnya akan meluap ke seluruh bagian toilet karena tidak memiliki celah untuk memisahkan kamar mandi dan area WC. Pancurannya juga membuat saya frustasi sepanjang waktu, airnya dingin dan anda dapat merasakan sedikit hangat (sama sekali tidak panas) meskipun anda memutar kerannya sampai penuh. Tempat tidurnya tidak seindah yang saya harapkan, selama saya mendapatkan tidur yang nyenyak setelah seharian bekerja keras, sudah cukup buat saya. Rekan saya yang menginap di lantai non smoking, juga memiliki masalah yang lebih buruk dari saya. Saya merasa hotel ini sangat membutuhkan perbaikan besar-besaran. Saya rasa, untuk perjalanan bisnis kali ini, saya tidak keberatan menginap paling tidak di hotel bintang 3, hotel, kamar, suasananya pasti lebih BERSIH dengan pancuran yang berfungsi baik.	dan atau jika bepergian bersama anak kecil dengan banyak barang bawaan Begitu saya memasuki kamar saya dapat merasakan bahwa saya berda di jaman kuno Dinding dindingnya dikotori jejak kaki lemari kayu dan meja punya bekas goresan dan kusam toiletnya pengap dan tempat pancurannya hanya setengah dibatasi kaca dimana setiap saat anda mandi airnya akan meluap ke seluruh bagian toilet karena tidak memiliki celah untuk memisahkan kamar mandi dan area WC Pancurannya juga membuat saya frustasi sepanjang waktu airnya dingin dan anda dapat merasakan sedikit hangat sama sekali tidak panas meskipun anda memutar kerannya sampai penuh Tempat tidurnya tidak seindah yang saya harapkan selama saya mendapatkan tidur yang nyenyak setelah seharian bekerja keras sudah cukup buat saya Rekan saya yang menginap di lantai non smoking juga memiliki masalah yang lebih buruk dari saya Saya merasa hotel ini sangat membutuhkan perbaikan besar besaran Saya rasa untuk perjalanan bisnis kali ini saya tidak keberatan menginap paling tidak di hotel bintang hotel kamar suasananya pasti lebih BERSIH dengan	keras cukup Rekan menginap lantai non smoking memiliki masalah buruk merasa hotel membutuhkan perbaikan besar besaran rasa untuk perjalanan bisnis kali keberatan menginap hotel bintang hotel kamar suasananya BERSIH pancuran berfungsi baik	membantu membantu_membawakan membawakan membawakan_koper koper koper_berupa berupa berupa_satu satu satu_koper koper koper_troli troli troli_prosedur prosedur prosedur_resepsionis resepsionis resepsionis_untuk untuk untuk_memberikan memberikan memberikan_kunci kunci kunci_bellboy bellboy bellboy_membantu membantu membantu_membawa membawa membawa_koper koper koper_kamar kamar kamar_rasa rasa rasa_perlu perlu perlu_berupa berupa berupa_koper koper koper_jinjing jinjing jinjing_mengganggu mengganggu mengganggu_bekerja bekerja bekerja_perhotelan perhotelan perhotelan_rasanya rasanya rasanya_hebat hebat hebat_mendapatkan mendapatkan mendapatkan_pelayanan pelayanan pelayanan_mempunyai mempunyai mempunyai_bawaan bawaan bawaan_bepergian bepergian bepergian_anak anak anak_barang barang barang_bawaan bawaan bawaan_memasuki memasuki	

Nama dokumen	*Review*	Hasil *Tokenization*	*Hasil Stopword*	*Hasil Bi-Gram*	*Class*
		pancuran yang berfungsi baik		memasuki_kamar kamar kamar_merasakan merasakan merasakan_berda berda berda_jaman jaman jaman_kuno kuno kuno_Dinding Dinding Dinding_dindingnya dindingnya dindingnya_dikotori dikotori dikotori_jejak jejak jejak_kaki kaki kaki_lemari lemari lemari_kayu kayu kayu_meja meja meja_punya punya punya_bekas bekas bekas_goresan goresan goresan_kusam kusam kusam_toiletnya toiletnya toiletnya_pengap pengap pengap_tempat tempat tempat_pancurannya pancurannya pancurannya_setengah setengah setengah_dibatasi dibatasi dibatasi_kaca kaca kaca_dimana dimana dimana_mandi mandi mandi_airnya airnya airnya_meluap meluap meluap_bagian bagian bagian_toilet toilet toilet_memiliki memiliki memiliki_celah celah celah_untuk untuk untuk_memisahkan memisahkan memisahkan_kamar kamar kamar_mandi mandi mandi_area area area_WC WC WC_Pancurannya Pancurannya Pancurannya_membuat membuat membuat_frustasi frustasi frustasi_waktu waktu waktu_airnya	

Nama dokumen	*Review*	Hasil *Tokenization*	*Hasil Stopword*	*Hasil Bi-Gram*	*Class*
				airnya airnya_dingin dingin dingin_merasakan merasakan merasakan_hangat hangat hangat_panas panas panas_memutar memutar memutar_kerannya kerannya kerannya_penuh penuh penuh_Tempat Tempat Tempat_tidurnya tidurnya tidurnya_seindah seindah seindah_harapkan harapkan harapkan_mendapatkan mendapatkan mendapatkan_tidur tidur tidur_nyenyak nyenyak nyenyak_seharian seharian seharian_bekerja bekerja bekerja_keras keras keras_cukup cukup cukup_Rekan Rekan Rekan_menginap menginap menginap_lantai lantai lantai_non non non_smoking smoking smoking_memiliki memiliki memiliki_masalah masalah masalah_buruk buruk buruk_merasa merasa merasa_hotel hotel hotel_membutuhkan membutuhkan membutuhkan_perbaikan perbaikan perbaikan_besar besar besar_besaran besaran besaran_rasa rasa rasa_untuk untuk untuk_perjalanan perjalanan perjalanan_bisnis bisnis bisnis_kali kali	

Nama dokumen	*Review*	Hasil *Tokenization*	*Hasil Stopword*	*Hasil Bi-Gram*	*Class*
				kali_keberatan keberatan keberatan_menginap menginap menginap_hotel hotel hotel_bintang bintang bintang_hotel hotel hotel_kamar kamar kamar_suasananya suasananya suasananya_BERSIH BERSIH BERSIH_pancuran pancuran pancuran_berfungsi berfungsi berfungsi_baik baik	
Negatif -15.txt	Gedung yg sudah tua, memang harga nya murah tapi jika anda di tempatkan di gd.lama dan di lt 2 atau 3 teerkesan horor, karna gelap dan kotor, saya merekomendasi di gd.baru diatas lobby, lebih nyaman karna sudah banyak perbaikan.	gedung yg sudah tua memang harga nya murah tapi jika anda di tempatkan di gd lama dan di lt atau teerkesan horor karna gelap dan kotor saya merekomendasi di gd baru diatas lobby lebih nyaman karna sudah banyak perbaikan	gedung yg tua harga nya murah tempatkan gd lt teerkesan horor karna gelap kotor merekomendasi gd baru diatas lobby nyaman karna perbaikan	gedung gedung_yg yg yg_tua tua tua_harga harga harga_nya nya nya_murah murah murah_tempatkan tempatkan tempatkan_gd gd gd_lt lt lt_teerkesan teerkesan teerkesan_horor horor horor_karna karna karna_gelap gelap gelap_kotor kotor kotor_merekomendasi merekomendasi merekomendasi_gd gd gd_baru baru baru_diatas diatas diatas_lobby lobby lobby_nyaman nyaman nyaman_karna karna karna_perbaikan perbaikan	Negatif

Perhitungan Manual *Prototype* Naïve Bayes adalah sebagai berikut

Berdasarkan penguumpulan data *review* hotel dalam penelitian ini 100 *review* positif dan 100 *review* negatif yang diambil dari situs www.tripadvisor.com dengan menggunakan kata sentimen Bagus, Nyaman, Kotor dan Buruk dapat dihitung secara manual berdasarkan jumlah kata yang sering muncul disetiap *review* dengan jumlah kata bagus pada *review* positif sebanyak 66 *review*, jumlah kata nyaman pada *review* positif sebanyak 60 *review*, jumlah kata

buruk pada *review* positif sebanyak 1 *review* dan jumlah kata buruk pada *review* positif sebanyak 1 *review*. Sedangkan jumlah kata bagus pada *review* negatif sebanyak 17 *review*, jumlah kata nyaman pada *review* negatif sebanyak 17 *review*, jumlah kata buruk pada *review* negatif 75 *review* dan jumlah kata buruk sebanyak 44 *review*.

Setelah sudah mendapatkan jumlah dari semua kata sentimen, bagus, nyaman, kotor dan buruk maka di bagi dengan jumlah keseluruhan *review* positif sebanyak 100 *review* dan jumlah review negatif sebanyak 100. Seperti tabel dibawah ini

Tabel 4.14 Perhitungan kata sentimen bagus, nyaman, kotor dan buruk

Review	Bagus	Kotor	Kotor	Buruk
Positif	66/100 =0.66	60/100= 0.60	1/100 = 0.01	1/100 = 0.01
Negatif	17/100 = 0.17	17/100= 0.17	75/100 = 0.75	44/100= 0.44

Berikut ini adalah tabel perhitungan manual *prototype* Naïve Bayes yang dimana yang dimana penulis hanya menampilkan 10 dokumen sentimen dari 200 data *training* dan 4 kata yang berhubungan yaitu bagus, nyaman, kotor dan buruk. Kehadiran kata dalam suatu kalimat akan diawali dengan angka hasil perhitungan kata sentimen dan dan angka 1 jika kata tersebut tidak muncul dalam kalimat pada dokumen dan dihasil akhirnya dikalikan dengan jumlah nilai *review* dibagikan jumlah keseluruhan *review*

Tabel 4.15 Perhitungan Manual *Prototype* Akurasi Naïve Bayes

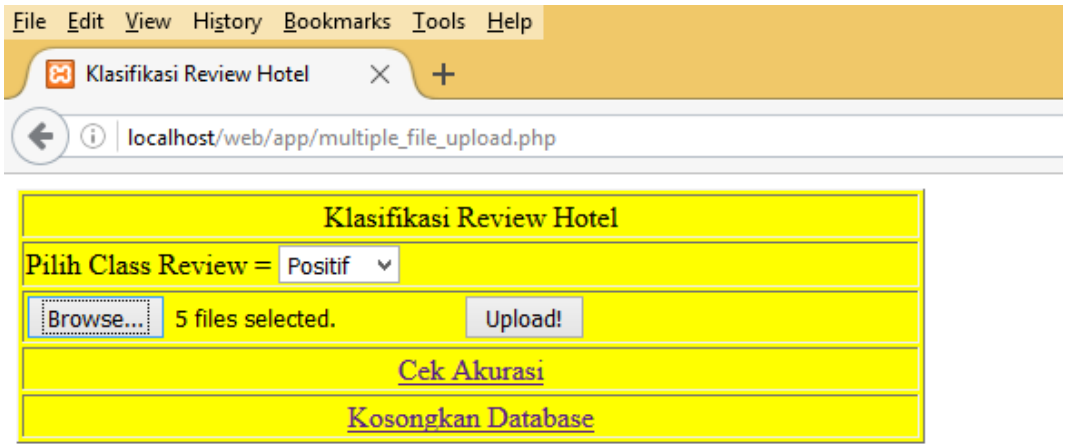
Nama Dokumen	Prediksi	Bagus	Nyaman	Kotor	Buruk	pos/neg	Hasil	Class
positif-11.txt	Positif	0,66	1	1	1	0,5	0,33	Positif
	Negatif	0,17	1	1	1	0,5	0,085	
positif-12.txt	Positif	0,66	0,6	1	1	0,5	0,198	Positif
	Negatif	0,17	0,17	1	1	0,5	0,01445	
positif-13.txt	Positif	1	0,6	1	1	0,5	0,3	Positif
	Negatif	1	0,17	1	1	0,5	0,085	
positif-14.txt	Positif	1	0,6	1	1	0,5	0,3	Positif
	Negatif	1	0,17	1	1	0,5	0,085	
positif-	Positif	0,66	0,6	1	1	0,5	0,198	Positif

15.txt	Negatif	0,17	0,17	1	1	0,5	0,01445	
negatif-11.txt	Positif	1	0,6	1	1	0,5	0,3	Positif
	Negatif	1	0,17	1	1	0,5	0,085	
negatif-12.txt	Positif	1	0,6	1	0,01	0,5	0,003	Negatif
	Negatif	1	0,17	1	0,44	0,5	0,0374	
negatif-13.txt	Positif	1	1	1	0,01	0,5	0,005	Negatif
	Negatif	1	1	1	0,44	0,5	0,22	
negatif-14.txt	Positif	1	1	0,01	0,01	0,5	0,00005	Negatif
	Negatif	1	1	0,75	0,44	0,5	0,165	
negatif-15.txt	Positif	1	0,6	0,01	1	0,5	0,003	Negatif
	Negatif	1	0,17	0,75	1	0,5	0,06375	

Perhitungan prediksi *review* positif dan *review* negatif pada setiap dokumen diatas adalah mengkalikan nilai kata bagus x nilai kata nyaman x nilai kata kotor x nilai kata buruk dikali 0,5 sama dengan hasil dari masing- masing nilai prediksi, apabila nilai prediksi positif lebih besar prediksi negatif maka termasuk kelas positif begitupun sebaliknya.

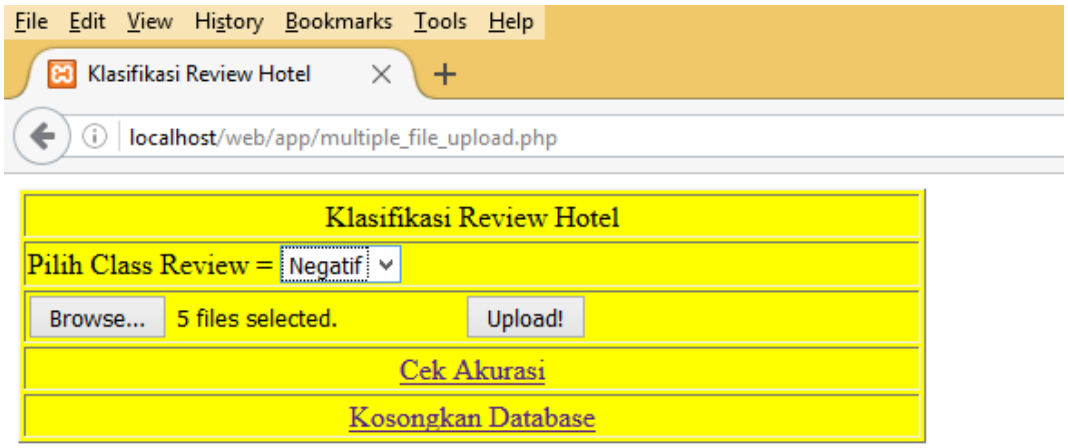
4.4. Pengembangan *Prototype* Hitung Nilai Akurasi Naïve Bayes

Penulis membuat aplikasi untuk untuk menghitung nilai akurasi untuk menguji model yang sudah ada menggunakan *dataset* dalam *review* hotel. Hasil akurasi dari penelitian akan diterapakan kedalam pembuatan aplikasi untuk klasifikasi *review* hotel menggunakan perangkat lunak *dreamweaver* CS 3 menggunakan bahasa pemrograman php, sehingga dapat mengetahui nilai akurasi dari jumlah review menggunakan Algoritma Naïve Bayes. Seperti gambar 4.12 dibawah ini



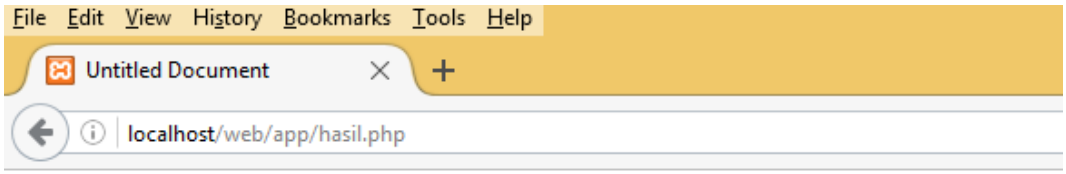
Gambar 4.12 Tampilan Aplikasi *Input Review* positif

Gambar 4.12 merupakan tampilan aplikasi untuk *input review* positif, dimana dalam aplikasi ini diminta untuk memasukan komentar yang diprediksi positif kedalam aplikasi. Pada aplikasi tersebut terlebih dahulu pilih *class review* positif untuk *upload* file .txt *review* positif yang akan di upload, klik tombol *Browse* untuk mencari data *review* positif-11.txt –positif -15.txt setelah itu tekan tombol upload, maka dokumen akan tersimpan kedalam *database* langkah selanjutknya adalah memasukan review negatif seperti gambar 4.13 dibawah ini.



Gambar 4.13 Tampilan Aplikasi *Input Review* Negatif

Gambar 4.13 merupakan tampilan aplikasi untuk *input review* negatif, dimana dalam aplikasi ini diminta untuk memasukan komentar yang diprediksi negatif kedalam aplikasi. Pada aplikasi tersebut terlebih dahulu pilih *class review* negatif untuk *upload* file .txt *review* negatif yang akan di *upload*, klik tombol *Browse* untuk mencari data *review* negatif misalkan negatif-11.txt –negatif -15.txt setelah itu tekan tombol *upload*, maka dokumen akan tersimpan kedalam *database* langkah selanjutknya adalah klik tombol cek akurasi untuk mengetahui nilai akurasi dari dokumen positif-11.txt sampai dokumen positif-15.txt dan negatif-11.txt sampai negatif-15.txt hasil akurasi seperti gambar 4.14 dibawah ini.



Nilai Akurasi Naive Bayes

	True Negatif	True Positif
Pred Negatif	4	1
Pred Positif	0	5
Akurasi	90	

Gambar 4.14 Tampilan Aplikasi Nilai Akurasi *Prototype* Naïve Bayes

Gambar 4.14 merupakan tampilan aplikasi untuk hasil nilai akurasi Naïve Bayes, dari data sebayak 10 data *review* hotel yang terdiri dari 5 data *review* positif dan 5 data *review* negatif. Sebanyak 5 data di prediksi kedalam *class* positif yaitu termasuk kedalam prediksi *class* positif, 4 data diprediksi kedalam *class* negatif sesuai termasuk kedalam prediksi data negatif dan 1 diprediksi kedalam *class* negatif ternyata masuk kedalam *class* positif.

Dari perhitungan manual *prototype* diatas dapat dihutung nilai akurasinya dengan rumus :

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \; x \; 100$$

$$Accuracy = \frac{5 + 4}{5 + 1 + 0 + 4} \; x \; 100$$

$$Accuracy = \frac{9}{10} = 0{,}9 \; x \; 100 = 90$$

Berikut ini adalah hasil perhitungan nilai akurasi menggunakan *RapidMiner* Naïve Bayes dengan data yang sama :

accuracy: 90.00% +/- 30.00% (mikro: 90.00%)

	true Class Negatif	true Class Positif	class precision
pred. Class Negatif	4	0	100.00%
pred. Class Positif	1	5	83.33%
class recall	80.00%	100.00%	

Gambar 4.15 Hasil Nilai Akurasi Naïve Bayes Menggunakan *RapidMiner* 5.3

4.5. Implikasi Penelitian

Implikasi penelitian ini mencakup beberapa aspek, diantaranya :

1. Implikasi terhadap *system*

   Hasil evaluasi menunjukan bahwa penerapan *Genetic Algorithm* untuk seleksi fitur jenis *wrapper* dapat meningkatkan nilai Akurasi dan *AUC* Naïve Bayes dan merupakan metode yang cukup baik dalam mengklasifikasi teks *review* hotel berbahasa Indonesia. Dengan menerapkan metode tersebut dapat membantu para pengunjung yang akan memesan kamar hotel dalam mengambil keputusan dengan cepat dan mengurangi waktu dalam membaca *review* atau komentar hotel pada pengunjung sebelumnya dan dapat memberikan informasi dalam menentukan kamar yang disediakan sesuai dengan keinginan pengunjung hotel, untuk meningkatkan kenyamanan dan pelayanan hotel kedepanya.

2. Implikasi Terhadap Manajerial

   Dapat membantu para pengembang sistem yang berkaitan dengan *review* hotel baik dari sumber tripadvisor maupun dari situs yang menawarkan jasa penginapan dimedia sosial lainya agar menggunakan aplikasi RapidMiner dalam mengklasifikasi *review* komentar pada saat membangun suatu sistem

3. Implikasi terhadap aspek penelitian selanjutnya

   Penelitian selanjutnya bisa menggunakan pemilihan fitur atau metode lainnya yang bisa meningkatkan akurasi Naïve Bayes pada *dataset* berbahasa Indonesia yang berbeda atau komparasi *dataset* berbahasa Indonesia dengan Bahasa Inggris dari domain yang berbeda seperti *review* film, *review* produk dan *review* restoran dan sebagainya.

BAB V

PENUTUP DAN SARAN

5.1 Kesimpulan

Untuk klasifikasi teks dengan data *review* hotel, yang terdiri dari 100 *review* positif dan 100 *review* negatif yang diambil dari situs www.tripadvisor.com, salah satu metode klasifikasi yang dapat digunakan adalah pengklasifikasi Naïve Bayes. Dalam hal ini Algoritma Naïve Bayes merupakan metode klasifikasi yang sangat sederhana dan efisien. Selain itu Naïve Bayes merupakan pengklasifikasi teks yang sangat popular yang memiliki performa yang sangat baik pada banyak domain baik dalam klasifikasi teks.

Dari pengolahan data yang sudah dilakukan, menggunakan metode pemilihan fitur jenis *wrapper* yaitu *Genetic Algorithm* terbukti dapat meningkatkan akurasi pada pengklasifikasi Naïve Bayes. Data *review* hotel berbahasa Indonesia diklasifikasi dengan baik kedalam *review* positif maupun *review* negatif. Akurasi model Naïve Bayes sebelum menggunakan metode pemilihan fitur *Genetic Algorithm* mencapai 90.50%, sedangkan setelah menggunakan metode pemilihan fitur *Genetic Algorithm*, akurasinya meningkat menjadi 94.50%, dapat meningkatkan akurasi sebesar 4%, Dalam mendukung klasifikasi teks berbahasa Indonesia, penulis mengembangkan aplikasi *review* hotel untuk mengklasifikasi *review* positif dan *review* negatif menggunakan bahasa pemrograman PHP.

Model yang dibentuk dapat diterapkan pada seluruh *review* hotel, sehingga dapat langsung hasilnya dalam mengklasifikasi teks pada *review* termasuk kedalam *review* positif atau *review* negatif. Sehinga dapat membantu pengunjung atau pemesan hotel dalam mengambil keputusan dangan cepat dan efisien saat memesan penginapan tanpa harus khawatir adanya pemberian rating yang tidak sesuai dengan *review*nya dan dapat memberikan informasi dalam menentukan kamar yang disediakan sesuai dengan keinginan pengunjung hotel, untuk meningkatkan kenyamanan dan pelayanan hotel kedepanya.

\

5.2 Saran

Walaupun pengklasifikasi Naïve Bayes sering digunakan dan mempunyai performa yang baik dalam pengklasifikasi teks, namun ada beberapa hal yang dapat ditambahkan untuk penelitian selanjutnya

1. Mengklasifikasi teks dengan *dataset* Bahasa Indonesia dan Bahasa Inggris dengan menggunakan Algoritma Naïve Bayes, sehingga bisa mengetahui keunggulan dari hasil klasifikasi Bahasa Indonesia dan Bahasa Inggris.
2. Menggunakan metode pemilihan fitur yang lain, seperti *Information Gain*, *Mutual Information*, *Chi Square, Gini Index* dan lain-lain agar hasilnya bisa dibandingkan dengan metode yang umum digunakan. Baik penggunaan metode-metode terpisah atau digabung.
3. Menggunakan data *review* berbahasa Indonesia dari domain yang berbeda, misalnya *review* film, *review* restoran, *review* produk dan sebagainya

DAFTAR PUSTAKA

Bramer, Max.(2007). Principles of Data Mining. London: Springer.

Charjan, Miss Dipti S dan Pun, Mukesh A. (2013). Pattern Discovery For Text Mining Using Pattern Taxonomy. International Journal of Engineering Trends and Technology. Volume 4 Issue 10. 4550-4555

Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications, 36,* no 3 pp. 5432– 5435.

Duan, W., Cao, Q., Yu, Y., dan Levy, S., (2013). *Mining Online User-Generated Content : Using Sentimen Analisis Technique to Study Hotel Quality*. 2013 46th Hawaii International Conference System Sciences.

Elden, S. A., Moustafa, A. M., Harb, M. H., dan Emara, H. A., (2013). *AdaBoost Ensemble With Simple Genetic Algoritm For Srudent Prediction Model*.

Feldman, R & Sanger, J. 2007. The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press : New York.

Gencosman, B. C., Ozmutlu, H. C., dan Ozmutlu, S. (2014). Character n-gram application for automatic new topic identification. *Information Processing and Management, 50,* 821-856. doi:10.1016/j.ipm.2014.06.005

Guo, P., Wang, X. & Han, Y., 2010. The Enhanced Genetic Algorithms for the Optimization Design. , (Bmei), pp.2990–2994.

Gunal, S. (2012). Hybrid feature selection for text classification ¨, vol. *20*.

Gorunescu, F. (2011). *Data Mining: Concepts, Models and Techniques*. Berlin: Springer.

Govindarajam, M., 2013. Sentiment Analysis of Movie Using Hybrid Method of Naïve Bayes and Genetic Algorithm.

Haddi, E., Liu, X., dan Shi, Y. (2013). The Role of Text Pre-processing in Sentiment Analysis. *Procedia Computer Science, 17,* 26-32. doi:10.1016/j.procs.2013.05.005

Han, J., & Kamber, M. (2007). *Data Mining Concepts and Techniques*.

Harb, M. H., dan Desuky, S. A., (2011). *Adaboost Ensemble with Genetic Algoritm Post Optimization for Intrucsion Detection*

Intan, R. dan Defeng, A. (2006). *Subject Based Search Engine Menggunakan TF-IDF dan Jaccard's Coefficient*, Universitas Kristen Petra, Surabaya

Koncz, P. & Paralic, J., 2011. *An approach to feature selection for sentiment analysis*. 2011 15th IEEE International Conference on Intelligent Engineering Systems, pp.357–362.

Kontopoulos, E., Berberidis, C., Dergiades, T., dan Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter post. *Expert Systems with Applications, 40,* 4065-4074. doi:10.1016/j.eswa.2013.01.001

Markopoulos, G., Mikros, G., Iliadi, A., Dan Liontos, M., (2015). *Sentiment Analysis of Hotel Reviews in Greek: A Comparison of Unigram Features*. Springer Proceeding in Business and Economics. DOI 10.1007/978-3-319-15859-4_3.

Medhat, W., Hassan, A. & Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), pp.1093–1113.

Moraes, R., Valiati, F., & Neto, W. P. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications*, *40*, 621–633.

O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*, Second. Boston, MA: Springer US, 2010, pp. 1, 86, 97.

Powers, D.M.W. (2011). *Evaluation: From Precision, Recall and F-Measure to ROC*, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies, ISSN: 2229- 3981 & ISSN: 2229-399X, Volume 2, Issue 1, 2011, pp-37-63.

Robertson, S. (2004). *Understanding Inverse Document Frequency: On Theoritical*
Arguments for IDF, Journal of Documentation; 2004; 60, 5; ABI/IN*FORM* Global.

Singh, A., dan Kakkar, V., (2015). A Study of Text Mining Techniques and Optimization Using Genetic Algorithm.

Suardika, I. G., (2016). *Sentiment Analysis System And Correlation Analysis On Hospitality In Bali*. Journal Of Theoretical and Applied Information Technology (Vol.84. No.1).

Suyanto. *Algoritma Genetika dalam Matlab*. Andi, Yogyakarta, 2005.

Taylor, E. M., Velasquez, J. D., Marquez, F. B., dan Matsuo, Y., (2013). Indentifying Customer Preferences about Tourism Products using an Aspect-Based Opinion Mining Approach. *Procedia Computer Science, 22,* 182-191. doi:10.1016/j.procs.2013.09.094

Wang, Ruihu. (2012). AdaBoost for Feature Selection, Classification and Its Relation with SVM, A Review. 2012 International Conference on Solid State Device and Material Science. 800-807.

Wibowo, Setyoningsih. (2014). *Neural Network* Dengan Algoritma Genetika Sebagai Pemilihan Fitur Pada Prediksi Loyalitas Pelanggan. Majalah Ilmiah Pawiyatan Vol :XXI, No : 2.

Witten, H. I., Frank, E., & Hall, M. A. (2011). Data Mining Practical Machine Learning Tools And Technique. Burlington: Elsevier Inc.

Zhao, M., Fu, C., Ji, L., Tang, K., & Zhou, M. (2011). Feature selection and parameter optimization for support vector machines: A new approach based on genetic algorithm with feature chromosomes. Expert Systems with Applications, 38(5), 5197–5204. doi:10.1016/j.eswa.2010.10.041

Zukhri, Zainudin.(2014). Algoritma Genetika Metode Komputasi untuk Menyelesaikan Maslah Optimasi. Yogyakarta: Andi Offset

DAFTAR RIWAYAT HIDUP

I. Biodata Mahasiswa

NIM	: 14001712
Nama lengkap	: Andi Taufik
Tempat, tanggal lahir	: Bogor, 30 November 1991
Alamat Lengkap	: Jl. Raya Jembatan Besi, RT 010/001 Jembatan Besi Tambora, Jakarta Barat

Riwayat Pendidikan Formal

1. SD Negeri Cibunian 01, Bogor lulus tahun 2004
2. SMP Negeri'1 Pamijahan , Bogor lulus tahun 2007
3. SMK PANDU, Bogor lulus tahun 2010
4. Akademik AMIK Bina Sarana Informatika, Jakarta Lulus tahun 2013,
5. STMIK Nusa Mandiri, Jakarta Lulus tahun 2014

II. Riwayat Pengalaman Pekerjaan

1. Asisten Teknical Support BSI Group Juni 2012 sampai dengan maret 2014
2. Technical Support BSI Group April 2014 Sampai dengan sekarang

Jakarta, 12 Agustus 2016

Andi Taufik