# Integrasi Metode Information Gain Untuk Seleksi Fitur dan Adaboost Untuk Mengurangi Bias Pada Analisis Sentimen Review Restoran Menggunakan Algoritma Naïve Bayes

Lila Dini Utami
*Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri*
*lila.ldu@gmail.com*

Romi Satria Wahono
*Fakultas Ilmu Komputer, Universitas Dian Nuswantoro*
*romi@brainmatics.com*

*Abstrak*: Internet merupakan bagian penting dari kehidupan sehari-hari. Saat ini, tidak hanya dari anggota keluarga dan teman-teman, tetapi juga dari orang asing yang berlokasi diseluruh dunia yang mungkin telah mengunjungi restoran tertentu. Konsumen dapat memberikan pendapat mereka yang sudah tersedia secara online. Ulasan yang terlalu banyak akan memakan banyak waktu dan pada akhirnya akan menjadi bias. Klasifikasi sentimen bertujuan untuk mengatasi masalah ini dengan cara mengklasifikasikan ulasan pengguna ke pendapat positif atau negatif. Algoritma Naïve Bayes (NB) adalah tekhnik *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, sehingga membuat akurasi menjadi rendah. Oleh karena itu, dalam penelitian ini menggunakan Information Gain (IG) untuk seleksi fitur dan metode adaboost untuk mengurangi bias agar dapat meningkatkan akurasi algoritma naïve bayes. Penelitian ini menghasilkan klasifikasi teks dalam bentuk positif dan negatif dari *review* restoran. Pengukuran naïve baye berdasarkan akurasi sebelum dan sesudah penambahan metode seleksi fitur. Validasi dilakukan dengan menggunakan 10 *fold cross validation*. Sedangkan pengukuran akurasi diukur dengan *confusion matrix* dan kurva ROC. Hasil penelitian menunjukkan peningkatan akurasi naïve bayes dari 73.00% jadi 81.50% dan nilai AUC dari 0.500 jadi 0.887. Sehingga dapat disimpulkan bahwa integrasi metode information gain dan adaboost pada analisis sentimen *review* restoran ini mampu meningkatkan akurasi algoritma naïve bayes.

*Kata Kunci*: analis sentimen, *review*restoran, klasifikasi teks, adaboost, information gain, naïve bayes.

## 1 PENDAHULUAN

Pertumbuhan jaringan sosial yang ada saat ini, membuat konsumen menggunakan konten dalam media untuk membuat keputusan yang lebih baik. Lebih banyak konsumen yang melihat pendapat dari konsumen lain sebelum memilih sebuah restoran. Di sisi lain, untuk restoran, sejumlah besar informasi publik yang tersedia bisa dijadikan sebagai bahan intropeksi untuk menjadikan restoran yang lebih baik (Reyes & Rosso, 2012). Beberapa konsumen menuangkan opini atau pengalaman mereka melalui media sosial seperti Facebook, Twitter, atau situs media yang lainnya. *Review* restoran yang dibuat secara online adalah saluran yang menghubungkan pengunjung satu dengan pengunjung lainnya. Hal ini merupakan layanan penyaringan yang dirancang untuk membantu konsumen. Hasil pencarian biasanya disajikan sebagai daftar restoran yang cocok, ditampilkan dengan singkat melalui sebuah gambar yang disertakan nama restoran, alamat serta *review* keseluruhan makanan dan layanan dan sebuah *hyperlink* ke halaman web yang berdedikasi restoran (Zhang, Ye, Law, & Li, 2010). Jika membaca *review* tersebut secara keseluruhan bisa memakan waktu dan sebaliknya jika hanya sedikit *review* yang dibaca, evaluasi akan bias. Klasifikasi sentimen bertujuan untuk mengatasi masalah ini dengan secara otomatis mengelompokkan *review* pengguna menjadi opini positif atau negatif.

Ada beberapa penelitian yang sudah dilakukan dalam hal pengklasifikasian sentimen terhadap *review* yang tersedia, diantaranya adalah penelitian oleh Kang, Yoo & Han, yang menggunakan algoritma naïve bayes dan mengkombinasikan kata sifat dengan N-grams (Kang, Yoo, & Han, 2012b). Lalu ada pula penelitian dari Zhang, Ye, Zhang & Li, dimana pengklasifikasian sentimen pada *review* restoran di internet yang ditulis dalam bahasa Canton menggunakan algoritma klasifikasi naïve bayes dan Support Vector Machine (SVM) (Zhang, Ye, Zhang, & Li, 2011). Sedangkan penelitian oleh Yulan He, menggunakan information gain dan naïve bayes untuk mempelajari ulasan pelanggan tentang sebuah film dan produk lainnya (He & Zhou, 2011).

Naïve bayes banyak digunakan untuk klasifikasi teks dalam *machine learning* yang didasarkan pada fitur probabilitas (Zhang & Gao, 2011). Naïve bayes sangat sederhana dan efisien. Sebagai teknologi *preprocessing* yang penting dalam klasiseleksi fitur dapat meningkatkan skalabilitas, efisiensi dan akurasi dari klasifikasi teks. Secara umum, metode seleksi fitur yang baik harus mempertimbangkan domain dan algoritma karakteristik. Sebagai *classifier*, naïve bayes sangat sederhana dan efisien serta sangat sensitif terhadap seleksi fitur (Chen, Huang, Tian, & Qu, 2009). Klasifikasi positif yang muncul 10% lebih tinggi dari akurasi klasifikasi negatif dan tampak beberapa kasus seperti star atau bintang dengan *review* yang tidak cocok. Algoritma naïve bayes diusulkan dan diukur melalui eksperimen komparatif dengan Unigrams dan Bigrams sebagai fiturnya. Dalam hal ini, naïve bayes membuktikan tingkat akurasi yang bagus saat klasifikasi dianggap seimbang (Kang, Yoo, & Han, 2012a). Akan tetapi, akurasi menjadi tidak akurat saat menghadapi sentimen klasifikasi yang kompleks.

Karena ketersediaan teks dalam bentuk digital menjamur dan meningkatnya kebutuhan untuk mengakses dengan cara yang fleksibel, klasifikasi teks menjadi tugas dasar dan penting. Meskipun sederhana, algoritma naïve bayes merupakan algoritma populer untuk klasifikasi teks (Ye, Zhang, & Law, 2009). Akan tetapi, masalah utama untuk

klasifikasi teks adalah dimensi tinggi dari ruang fitur. Hal ini sangat sering karena domain teks memiliki beberapa puluhan ribu fitur. Kebanyakan dari fitur ini tidak relevan dan bermanfaat bagi klasifikasi teks. Bahkan beberapa fitur mungkin mengurangi akurasi klasifikasi. Selain itu, sejumlah besar fitur dapat memperlambat proses klasifikasi (Chen et al., 2009).

Tingkatan lain yang umumnya ditemukan dalam pendekatan klasifikasi sentimen adalah seleksi fitur. Seleksi fitur bisa membuat pengklasifikasi baik lebih efisien dan efektif dengan mengurangi jumlah data yang dianalisa, maupun mengidentifikasi fitur yang sesuai untuk dipertimbangkan dalam proses pembelajaran (Moraes, Valiati, & Neto, 2013). Menurut John, Kohavi, dan Pfleger dalam Chen, ada dua jenis utama metode seleksi fitur dalam *machine learning*: wrapper dan filter. Wrapper menggunakan akurasi klasifikasi dari beberapa algoritma sebagai fungsi evaluasinya (Chen et al., 2009). Metode filter terdiri dari *document frequency*, *mutual information*, information gain, dan *chi-square*. Information gain sering lebih unggul dibandingkan yang lain. Information gain mengukur berapa banyak informasi kehadiran dan ketidakhadiran dari suatu kata yang berperan untuk membuat keputusan klasifikasi yang benar dalam *class* apapun. Informastion gain adalah salah satu pendekatan filter yang sukses dalam pengklasifikasian teks (Uysal & Gunal, 2012).

Sementara itu, menurut Hu (Hu & Hu, 2005), adaboost adalah algoritma yang ide dasarnya adalah untuk memilih dan menggabungkan sekelompok pengklasifikasi lemah untuk membentuk klasifikasi yang kuat. Adaboost adalah algoritma yang iteratif menghasilkan pengklasifikasi dan kemudian menggabungkan mereka untuk membangun klasifikasi utama (Kim, Hahn, & Zhang, 2000). Algoritma adaboost iteratif bekerja pada klasifikasi naïve bayes dengan bobot normal dan mengklasifikasikan masukan yang diberikan ke dalam kelas yang berbeda dengan beberapa atribut (Korada, Kumar, & Deekshitulu, 2012). Adaboost dirancang khusus untuk klasifikasi. Adaboost adalah algoritma pembelajaran yang dapat digunakan untuk meningkatkan akurasi untuk setiap pembelajaran algoritma yang lemah. Algoritma adaboost digunakan untuk meningkatkan akurasi lemah klasifikasi naïve bayes.

Pada penelitian ini menggunakan algoritma naïve bayes disertai information gain sebagai metode seleksi fitur dan metode adaboost sebagai teknik untuk memperbaiki tingkat klasifikasi yang diterapkan untuk mengklasifikasikan teks pada komentar dari *review* suatu restoran untuk meningkatkan akurasi analisa sentimen.

## 2 PENELITIAN TERKAIT

Ada beberapa penelitian yang menggunakan algoritma naïve bayes sebagai pengklasifikasi, metode adaboost, atau information gain sebagai seleksi fitur dalam klasifikasi teks analisa sentiment pada *review*, diantaranya:Penelitian yang dilakukan oleh Zhang, Ye, Zhang, dan Li mengenai analisa sentimen pada *review* restoran yang ditulis dalam bahasa Canton (Zhang et al., 2011b). Ulasan diambil dari situs www.openrice.com yang terdiri dari 1500 *review* positif dan 1500 *review* negatif. Dua penutur asli dilatih untuk ulasan ini dan didapatkanlah *review* yang sesuai dan digunakan untuk proses klasifikasi terdiri dari 900 *review* positif dan 900 *review* negatif. Sebagai langkah awal, peneliti melakukan seleksi fitur dengan cara mensubtitusi kalimat yang memiliki makna yang sama. Setelah subtitusi selesai, peneliti mengkombinasikan kata sifat dengan n-grams untuk melihat sentimen dalam teks. Algoritma *featureselection* yang digunakan adalah information gain. *Classifier* yang digunakan adalah support vector machine dan naïve bayes.

Sementara itu, penelitian yang dilakukan oleh Kang, Yoo, dan Han mengenai analisa sentimen pada *review* restoran. Sekitar 70.000 dokumen dikumpulkan dari pencarian situs restoran (Kang et al., 2012). Sisi positif dan negatif dikumpulkan dan diklasifikasikan sebelum sentimen analisis dilakukan. Ulasan yang terpilih adalah 5700 *review* positif dan 5700 *review* negatif. Untuk *textprocessing*, peneliti menggunakan *tokenization* dan melakukan pemilihan *review* yang mencakup kata-kata sentimen terkait dengan *review* restoran menggunakan n-grams. *Classifier* yang digunakan adalah support vector machine dan naïve bayes.

Dan penelitian yang dilakukan oleh He dan Zhou mengenai analisa sentimen pada *review* film, buku, DVD, dan barang elektronik (He & Zhou, 2011). Untuk *review* film diambil dari website IMDB dan *review* buku, DVD, dan barang elektronik diperoleh dari www.amazon.com sebanyak 100 *review* positif dan 1000 *review* negatif. Ulasan berisi peringkat terstruktur (bintang) dan teks. Untuk *textprocessing*, peneliti menggunakan *tokenization* dan melakukan pemilihan *review* yang mencakup kata-kata sentimen terkait dengan review menggunakan pengklasifikasi Lexicon Labeling, Heuristic Labeling, Self-labeled instance, Self-learned Features, dan Oracle Labeling. *Classifier* yang digunakan adalah naïve bayes dan support vector machine.

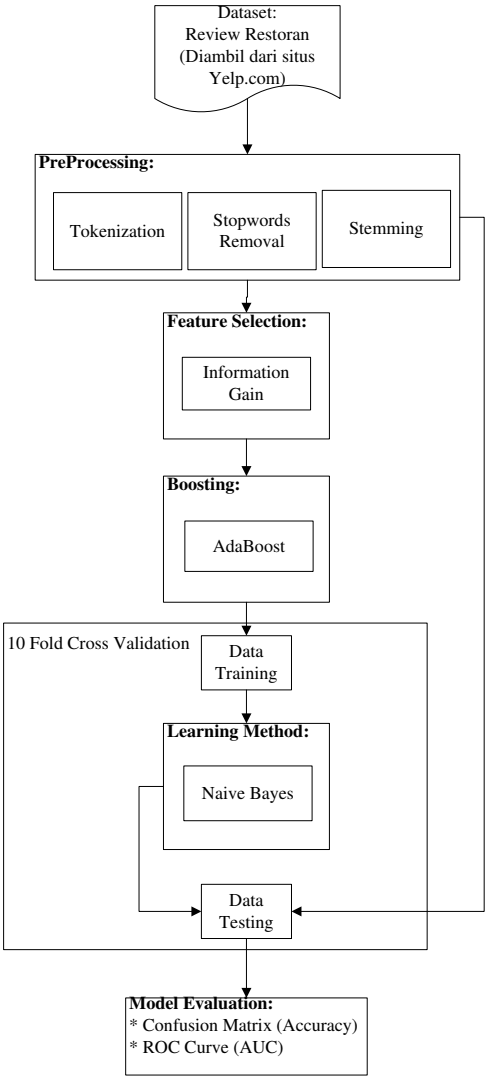
## 3 METODE YANG DIUSULKAN

Penelitian ini menggunakan data review restoran yang berada di New York, yang diambil dari situs http://www.yelp.com/nyc. *Review* restoran yang digunakan hanya 200 *review* restoran 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.

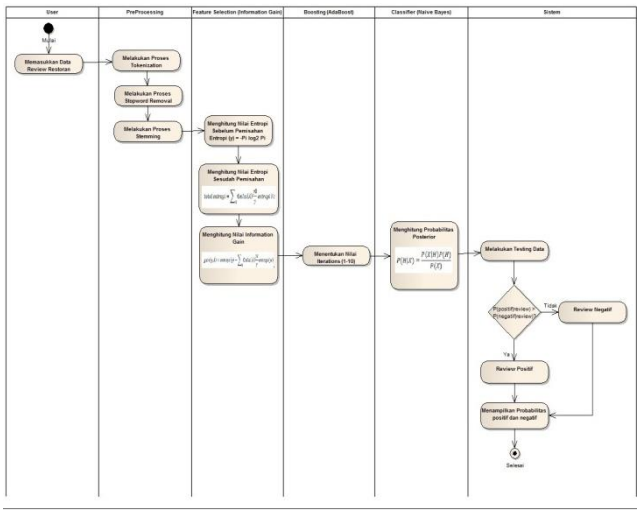
*Pre processing* yang dilakukan, diantaranya adalah:

a. *Tokenization*
   Dalam proses tokenization ini, semua kata yang ada di dalam tiap dokumen dikumpulkan dan dihilangkan tanda bacanya, serta dihilangkan jika terdapat simbol atau apapun yang bukan huruf
b. *StopwordsRemoval*
   Dalam proses ini, kata-kata yang tidak relevan akan dihapus, seperti kata the, of, for, with yang merupakan kata-kata yang tidak mempunyai makna tersendiri jika dipisahkan dengan kata yang lain dan tidak terkait dengan dengan kata sifat yang berhubungan dengan sentimen.
c. *Stemming*
   Dalam proses ini kata-kata akan dikelompokkan ke dalam beberapa kelompok yang memiliki kata dasaryang sama, seperti *drug*, *drugged*, dan *drugs* di mana kata dasar dari semuanya adalah kata *drug*.

*Feature selection* yang peneliti usulkan adalah metode dengan jenis filter, yakni information gain dan metode *boosting* yaitu adaboost, yang digunakan secara integrasi agar akurasi algoritma naïve bayes dapat meningkat. Penelitian ini nantinya menghasilkan akurasidan nilai AUC. Lihat Gambar 1 untuk model yang diusulkan secara detail dan ringkas, sementara itu Gambar 2 adalah model yang diusulkan berbentuk *activity diagram*.




Gambar 1. Model yang Diusulkan



Gambar 2.*Activity Diagram* Model yang Diusulkan

## 4 HASIL PENELITIAN

Proses eksperimen ini menggunakan aplikasi *RapidMiner* 5.2. Untuk pengujian model dilakukan menggunakan dataset *review* restoran. Spesifikasi komputer yang digunakan untuk eksperimen ini dapat dilihat pada Tabel 1.

Tabel 1 Spesifikasi Komputer yang Digunakan

| *Processor* | *Intel(R) Celeron(R) CPU 874 @1.10GHz* |
|---|---|
| Memori | 4.00 GB |
| Harddisk | 320 GB |
| Sistem Operasi | *Microsoft Windows 7* |
| Aplikasi | *RapidMiner 5.2* |

Proses klasifikasi di sini adalah untuk menentukan sebuah kalimat sebagai anggota *class* positif atau *class* negatif berdasarkan nilai perhitungan probabilitas dari rumus bayes yang lebih besar. Jika hasil probabilitas kalimat tersebut untuk *class* positif lebih besar dari pada *class* negatif, maka kalimat tersebut termasuk ke dalam *class* positif. Jika probabilitas untuk *class* positif lebih kecil dari pada *class* negatif, maka kalimat tersebut termasuk ke dalam *class* negatif. Penulis hanya menampilkan 10 dokumen dari keseluruhan 200 data training dan 4 kata yang berhubungan dengan sentimen dan yang paling sering muncul, yaitu *bad*, *good*, *delicious* dan *disappoint*. *Bad*, muncul sebanyak 21 kali yaitu dalam *review* positif sebanyak 4 kali dan *review* negatif sebanyak 17 kali. *Good*, muncul sebanyak 91 kali yaitu dalam *review* positif sebanyak 50 kali dan *review* negatif sebanyak 41 kali. *Delicious*, muncul sebanyak 25 kali yaitu dalam *review* positif sebanyak 22 kali dan *review* negatif sebanyak 3 kali. *Disappoint*, muncul sebanyak 33 kali yaitu dalam *review* positif sebanyak 3 kali dan *review* negatif sebanyak 30 kali. Kehadiran kata di dalam suatu dokumen akan diwakili oleh angka 1 dan angka 0 jika kata tersebut tidak muncul di dalam dokumen.

Tabel 2 Hasil Klasifikasi Teks

| **Dokumen Ke-** | ***Bad*** | ***Delicious*** | ***Good*** | ***Dissapoint*** | ***Class*** |
|---|---|---|---|---|---|
| 1 | 0 | 2 | 3 | 0 | Positif |
| 2 | 0 | 1 | 1 | 0 | Positif |
| 3 | 0 | 2 | 1 | 0 | Positif |
| 101 | 1 | 0 | 3 | 2 | Negatif |
| 102 | 2 | 1 | 1 | 1 | Negatif |
| 103 | 1 | 0 | 4 | 0 | Negatif |

Probabilitas bayes yang dijabarkan adalah probabilitas untuk dokumen ke 103.

4. Hitung probabilitas bersyarat (*likelihood*) dokumen ke 103 pada *class* positif dan negatif.
   Untuk *class* positif:
   P(103|positif)=P(bad=1|positif)xP(delicious=0|positif)xP(good=4|positif) x P(dissapoint=0|positif)

$$P(103|positif) = \frac{0}{6} x \frac{5}{6} x \frac{5}{6} x \frac{0}{6}$$

   =0 x 0,833 x 0,833 x 0
   =0

   P(103|negatif)=P(bad=1|negatif)xP(delicious=0|negatif)xP(good=4|negatif) x P(dissapoint=0|negatif)

$$P(103|negatif) = \frac{4}{5} x \frac{1}{5} x \frac{8}{5} x \frac{3}{5}$$

   =0,8 x 0,2 x 1,6 x 0,6
   =0,1536

5. Probabilitas prior dari *class* positif dan negatif dihitung dengan proporsi dokumen pada tiap *class*:

$$\text{P(positif)} = \frac{3}{6} = 0{,}5$$
$$\text{P(negatif)} = \frac{2}{6} = 0{,}333$$

6. Hitung probabilitas posterior dengan memasukkan rumus Bayes dan menghilangkan penyebut P(103):

$$\text{P(positif|103)} = \frac{(0)(0{,}5)}{P(103)} = 0$$
$$\text{P(negatif|103)} = \frac{(0{,}1536)(0{,}333)}{P(103)} = 0{,}0511488$$

Berdasarkan probabilitas diatas, maka dapat disimpulkan bahwa dokumen ke 103 termasuk dalam *class* negatif, karena P(positif|103) lebih kecil dari pada P(negatif|103).

Dari sebanyak 200 data *review* restoran yaitu 100 *review* positif dan 100 *review* negatif, sebanyak 89 data diprediksi sesuai yaitu negatif, dan sebanyak 11 data diprediksi negatif tetapi ternyata positif, 57 data diprediksi sesuai yaitu positif dan 43 data diprediksi positif tetapi ternyata negatif. Hasil yang diperoleh dengan menggunakan algoritma NB adalah nilai *accuracy* = 73.00% seperti pada tabel 3 dan AUC = 0.500, seperti pada Gambar 3.

Tabel 3*ConfusionMatrix* Algoritma NB

| ***Accuracy*: 73.00% +/- 9.34% (mikro: 73.00%)** | | | |
|---|---|---|---|
| | ***truenegative*** | ***truepositive*** | ***classprecision*** |
| ***Prediction negative*** | **89** | **43** | **67.42%** |
| ***Prediction positive*** | **11** | **57** | **83.82%** |
| ***classrecall*** | **89.00%** | **57.00%** | |

Nilai accuracy dari *confusionmatrix* tersebut adalah sebagai berikut:

$$accuracy = \frac{(TN + TP)}{(TN + FN + TP + FP)}$$
$$accuracy = \frac{(89 + 57)}{(89 + 11 + 57 + 43)}$$
$$accuracy = \frac{146}{200} = 0.73 = 73.00\%$$


Gambar 3 Grafik Area *UnderCurve* (AUC) Algoritma Naïve Bayes

Penelitian ini menggunakan metode filter yaitu information gain, dimana data yang diolah diberikan bobot dari information gain untuk meningkatkan akurasi algoritmanaïve bayes. Penelitian ini menggunakan operator *selectbyweight* dengan memilih parameter *weightrelation*=top k, dan k=10. Dimana nanti akan dihasilkan 10 atribut teratas. 10 atribut yang terpilih akan ditampilkan bobotnya masing-masing, untuk lebih jelasnya dapat dilihat pada Tabel 4

Tabel 4. Sepuluh Fitur Teratas dan Bobotnya

| **Atribut** | **Bobot** |
|---|---|
| overpr | 0.575 |
| want | 0.576 |
| review | 0.593 |
| favorit | 0.708 |
| amaz | 0.713 |
| delici | 0.713 |
| good | 0.767 |
| definit | 0.911 |
| disappoint | 1 |

Bobot diatas adalah bobot yang sudah di-*generate* oleh operator *selectbyweight*. Karena hasilnya masih ada angka 0, maka atribut yang ditampilkan bobotnya dari masing-masing dokumen hanya yang mempunyai bobot 1. Diantara 10 atribut diatas, hanya kata dissapoint yang memiliki bobot 1. Tabel 5 menunjukkan atribut tersebut didalam dokumen dalam bentuk *vector*.

Tabel 5 Atribut Dalam Bentuk *Vector*

| **No** | **Dokumen Ke-** | ***Dissapoint*** | ***Class*** |
|---|---|---|---|
| 1 | 12 | 2 | Negatif |
| 2 | 17 | 2 | Negatif |
| 3 | 28 | 2 | Negatif |
| 4 | 96 | 2 | Negatif |
| 5 | 23 | 1 | Negatif |
| 6 | 64 | 1 | Positif |
| 7 | 76 | 1 | Positif |
| 8 | 149 | 1 | Positif |
| 9 | 64 | 1 | Positif |
| 10 | 76 | 1 | Positif |

1. Cari nilai entropi sebelum pemisahan:
   y berisi 200 data dengan 100 keputusan positif dan 100 keputusan negatif.
   $Entropy(y) = -Pilog2Pi$
   $Entropy(y) =$ entropi[100,100]
   $= -\frac{100}{200}\log2\left(\frac{100}{200}\right) - \frac{100}{200}\log2\left(\frac{100}{200}\right) = 1$

2. Cari nilai entropi setelah pemisahan:
   Untuk atribut *dissapoint*,
   Nilai (positif)=[0,1]
   y=[100,100]
   y0=[97,70]
   y1=[3,30]

   a. *dissapoint* = 0

$$\text{entropy}[97,70] = -\frac{97}{167}\log2\left(\frac{97}{167}\right) - \frac{70}{157}\log2\left(\frac{70}{167}\right) = 0{,}29032$$

b. *dissapoint* = 1

$$\text{entropy}[3,30] = -\frac{3}{33}\log2\left(\frac{3}{33}\right) - \frac{30}{33}\log2\left(\frac{30}{33}\right) = 0{,}1323$$

3. Cari nilai information gain

$$\text{gain}(y, A) = entropi(y) \sum_{0} \in nilai(A)\frac{yc}{y} entropi(yc)$$

$$= \text{entropi}(y) - \frac{167}{200} entropi(y0) - \frac{3}{200} entropi(y1)$$

$$= 1 - \left(\frac{167}{200}\right) 0{,}29032 - \frac{33}{200} 0{,}1323 = 0{,}73335$$

Pengukuran dengan *confusion matrix* di sini akan menampilkan perbandingan dari hasil akurasi model naïve bayes sebelum ditambahkan seleksi fiturinformation gain dan metode adaboost yang bisa dilihat pada Tabel 6 dan setelah ditambahkan seleksi fiturinformation gain dan metode adaboost yang bisa dilihat pada Tabel 7.

Tabel 6 *ConfusionMatrix* Algoritma NaïveBayes Sebelum Penambahan Seleksi Fitur Information Gain dan Metode Adaboost

| **Accuracy: 70.00% +/- 8.66% (mikro: 70.00%)** | | | |
|---|---|---|---|
| | ***true negative*** | ***true positif*** | ***class precision*** |
| ***Prediction negative*** | **89** | **49** | **64.49%** |
| ***Prediction positive*** | **11** | **51** | **82.26%** |
| ***class recall*** | **89.00%** | **51.00%** | |

Tabel 7 *ConfusionMatrix* Algoritma NaïveBayes Sesudah Penambahan Seleksi Fitur Information Gain dan Metode Adaboost

| ***Accuracy*:99.50%** | | | |
|---|---|---|---|
| | ***true negative*** | ***true positif*** | ***class precision*** |
| ***Prediction negative*** | 99 | 0 | 100% |
| ***Prediction positive*** | 1 | 100 | 99.01% |
| ***class recall*** | 99.00% | 100.00% | |

$$akurasi = \frac{100 + 99}{100 + 0 + 1 + 99} = \frac{199}{200} = 0.995 = 99.50\%$$

Hasil pengujian *confusion matrix* di atas diketahui bahwa menggunakan algoritma naïve bayes mempunyai akurasi hanya 70.00% sedangkan algoritma naïve bayes dengan seleksi fiturinformation gain dan metode adaboost memiliki tingkat akurasi yang lebih tinggi yaitu 99.50%. Akurasi naik 29.50% dari yang sebelumnya.

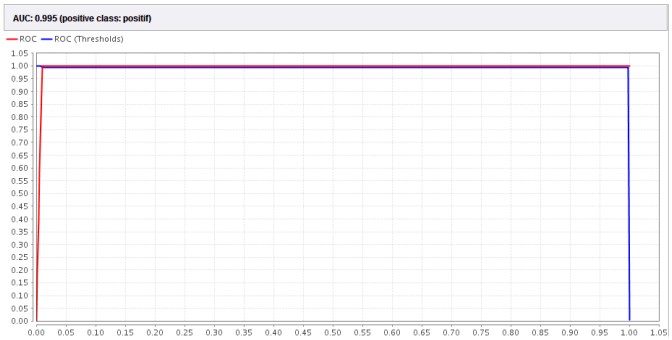
Grafik ROC akan membentuk garis dimana garis tersebut menunjukkan hasil prediksidari model klasifikasi yang digunakan. Apabila garis tersebut berada di atas diagonal grafik maka hasil klasifikasi bernilai baik (*good classification*), sedangkan garis yang berada di bawah diagonal grafik menghasilkan nilai klasifikasi yang buruk (*poor classification*). Garis yang menempel pada sumbu Y menunjukkan grafik tersebut menunjukkan klasifikasi yang baik (Gorunescu, 2011).

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

1. Nilai AUC 0.90-1.00 = *excellent classification*
2. Nilai AUC 0.80-0.90 = *good classification*
3. Nilai AUC 0.70-0.80 = *fair classification*
4. Nilai AUC 0.60-0.70 = *poor classification*
5. Nilai AUC 0.50-0.60 = *failure*

Berikut adalah tampilan kurva ROC yang akan dihitung nilai AUC-nya. Gambar 4 adalah kurva ROC untuk model naïve bayes sebelum menggunakan metode adaboost dan seleksi fitur IG dan gambar 5 adalah kurva ROC untuk model naïve bayes setelah menggunakan metode adaboost dan seleksi fitur information gain.


Gambar 4 Kurva AUC untuk Algoritma NaïveBayes Sebelum Menggunakan Seleksi Fitur Information Gain dan Metode Adaboost




Gambar 5 Kurva AUC Untuk Algoritma NaïveBayes Sesudah Menggunakan Seleksi Fitur Information Gain dan Metode Adaboost

Grafik diatas yang menunjukkan bahwa algoritma naïve bayes hanya memiliki nilai AUC 0.500 yang artinya *failure* (gagal) dibandingkan dengan algoritma naïve bayes yang menggunakan seleksi fitur information gain dan metode adaboost yang memiliki nilai AUC 0.995 yang artinya *excellent classification*.

Pada pengujian ini, menggunakan klasifikasi naïve bayes, algortima *feature selection* information gain dan teknik boosting yaitu metode adaboost. Pengujian dilakukan uji coba dengan melakukan optimalisasi perulangan (*iterations*). Tabel 8 adalah hasil dari percobaan yang telah dilakukan untuk penentuan nilai *accuracy* dan AUC.

Tabel 8 Pengujian Indikator

| *Iterations* | *Accuracy* | AUC |
|---|---|---|
| 1 | 80,50% | 0,805 |
| 2 | 80,50% | 0,850 |
| 3 | 80,50% | 0,870 |
| 4 | 80,50% | 0,878 |
| 5 | 80,50% | 0,882 |
| 6 | 81,50% | 0,887 |
| 7 | 80,50% | 0,890 |
| 8 | 80,50% | 0,890 |
| 9 | 80,50% | 0,890 |
| 10 | 80,50% | 0,890 |

Dari semua *iterations* yang di uji, *accuracy* tertinggi adalah pada saat *iterations* = 6, yaitu nilai *accuracy* = 81,50% dan nilai AUC = 0.887.

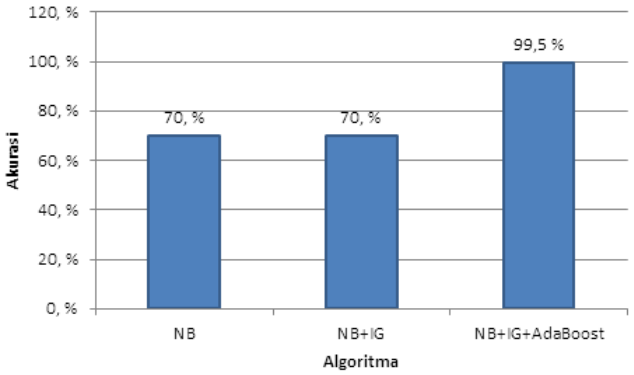
Dengan memiliki model klasifikasi teks pada *review*, pembaca dapat dengan mudah mengidentifikasi mana *review* yang positif maupun yang negatif. Dari data *review* yang sudah ada, dipisahkan menjadi kata-kata, lalu diberikan bobot pada masing-masing kata tersebut. Dapat dilihat kata mana saja yang berhubungan dengan sentimen yang sering muncul dan mempunyai bobot paling tinggi. Dengan demikian dapat diketahui *review* tersebut positif atau negatif.

Dalam penelitian ini, menunjukkan seberapa baik model yang terbentuk. Tanpa menggunakan metode seleksi fitur, algoritma naïve bayes sendiri sudah menghasilkan akurasi sebesar 70.00% dan nilai AUC 0.500. Akurasi tersebut masih kurang akurat, sehingga perlu ditingkatkan lagi menggunakan seleksi fitur yaitu information gain dan teknik *boosting* yaitu metode adaboost. Setelah menggunakan metode adaboost dan information gain, akurasi algoritma naïve bayes meningkat menjadi 99.50% dan nilai AUC 0.995. seperti yang bisa dilihat pada Tabel 9.

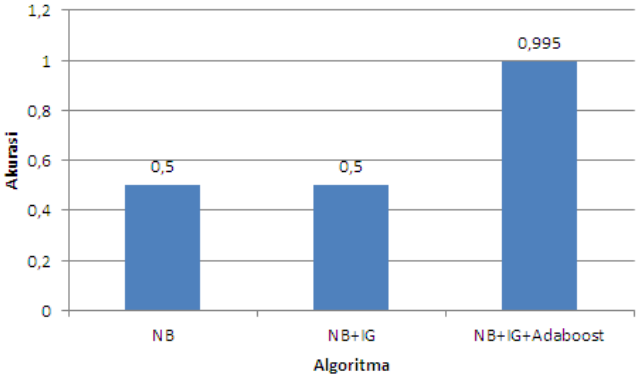
Tabel 9 Perbandingan Model Algoritma Naïve Bayes Sebelum dan Sesudah Menggunakan Seleksi Fitur Information Gain dan Metode Adaboost

| | **Algoritma Naive Bayes** | **Algoritma Naive Bayes Information Gain + AdaBoost** |
|---|---|---|
| Sukses prediksi positif | 51 | 100 |
| Sukses prediksi negatif | 89 | 99 |
| Akurasi model | 70.00% | 99.50% |
| AUC | 0.500 | 0.995 |

Berdasarkan hasil evaluasi di atas diketahui bahwa algoritma naïve bayes yang menggunakan seleksi fitur information gain dan metode *boosting* adaboost, mampu meningkatkan tingkat akurasi *review* restoran. Gambar 6 memperlihatkan tingkat akurasi yang meningkat dalam bentuk sebuah grafik. Sedangkan Gambar 7 memperlihatkan nilai AUC.




Gambar 6 Grafik Akurasi Algoritma NaïveBayes Sebelum dan Sesudah Menggunakan Seleksi Fitur Information Gain dan Metode Adaboost




Gambar 7 Grafik Nilai AUC Algoritma NaïveBayes Sebelum dan Sesudah Menggunakan Seleksi Fitur Information Gain dan Metode Adaboost

Penelitian mengenai ulasan *review* terkadang kedapatan perbedaan antara ulasan yang dibuat konsumen secara online dengan editor ulasan demi menarik konsumen untuk lama restoran tersebut (Zhang et al., 2010). Dalam beberapa kasus, ditemukan sentimen campuran, di mana sentimen positif memiliki ungkapan yang sama dengan sentimen negatif. Kasus seperti itu, aspek polaritas domain tidak bisa hanya dianggap sebatas positif atau negatif. Salah satu cara alternatif adalah untuk menetapkan nilai disetiap kemungkinan aspek polaritas yang sama yang diungkapkan dalam review tersebut. (Zhu, Wang, Zhu, Tsou, & Ma, 2011). Penelitian ini menunjukkan bahwa naïve bayes memerlukan sejumlah dukungan untuk meningkatkan akurasi tingkat klasifikasi. Tinggi atau tidaknya tingkat akurasi tergantung oleh jumlah fitur yang ada (Zhang et al., 2011b). Menurut hasil, kinerja naïve bayes meningkat dengan bantuan algortima adaboost dan menghasilkan hasil yang akurat dengan mengurangi kesalahan misklasifikasi dengan meningkatkan iterations (Korada, Kumar, & Deekshitulu, 2012). Metode information gain menunjukkan hasil yang memuaskan dalam filtering sebuah istilah (Moraes, Valiati, & Neto, 2013).

Dari pengolahan data yang sudah dilakukan dengan metode *boosting* yaitu adaboost dan seleksi fitur yaitu information gain, terbukti dapat meningkatkan akurasi algoritma naïve bayes. Data *review* restoran dapat diklasifikasi dengan baik ke dalam bentuk positif dan negatif.

## 5 KESIMPULAN

Naïve bayes merupakan salah satu pengklasifikasi yang mengklasifikasikan suatu teks, salah satu contoh yakni review restoran. Naïve bayes sangat sederhana dan efisien,

juga sangat populer digunakan untuk klasifikasi teks dan memiliki performa yang baik pada banyak domain.

Pengolahan data yang dilakukan ada 3 tahap, yakni naïve bayes, naïve bayes dan information gain, dan naïve bayes, information gain, dan adaboost. Dan ternyata, jika hanya naïve bayes saja yang digunakan, akurasi hanya mencapai 70% dan AUC=0,500. Sama halnya jika naïve bayes disertai dengan information gain, akurasi yang dicapaipun hanya 70% dan AUC=0,500, itu membuktikan bahwa information gain tidak mempengaruhi akurasi terhadap naïve bayes. Akan tetapi, jika naïve bayes dan information gain disertai pula dengan adaboost, akurasi meningkat 29,5% menjadi 99,5% dan AUC=0,995.

## REFERENCES


Ali, W., Shamsuddin, S. M., & Ismail, A. S. (2012). Intelligent Naïve Bayes-based approaches for Web proxy caching. *Knowledge-Based Systems*, *31*, 162–175.

Bauer, E. (1999). An Empirical Comparison of Voting Classification Algorithms : Bagging , Boosting , and Variants. *Machine Learning Research*, *139*, 105–139.

Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications*, *36*, 5432–5435.

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

He, Y., & Zhou, D. (2011). Self-training from labeled features for sentiment analysis. *Information Processing & Management*, *47*, 606–616.

Huang, J., Rogers, S., & Joo, E. (2013). Improving Restaurants. *Information System*, 1–5.

Kang, H., Yoo, S. J., & Han, D. (2012a). Expert Systems with Applications Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems With Applications*, *39*(5), 6000–6010.

Kang, H., Yoo, S. J., & Han, D. (2012b). Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications*, *39*, 6000–6010.

Korada, N. K., Kumar, N. S. P., & Deekshitulu, Y. V. N. H. (2012). Implementation of NBian Classifier and Ada-Boost Algorithm Using Maize Expert System. *International Journal of Information Sciences and Techniques*, *2*, 63–75.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Toronto: Morgan and Claypool

Muthia, D. A. (2013). Analisis Sentimen Pada Review Buku Menggunakan Algoritma. *Sistem InformasiSistem Informasi*, 1–9.

Reyes, A., & Rosso, P. (2012). Making objective decisions from subjective data : Detecting irony in customer reviews. *Decision Support Systems*, *53*, 754–760.

Sharma, A., & Dey, S. (2012). A Comparative Study of Feature Selection and Machine Learing Techniques for Sentiment Analysis. *Information Search and Retrieval*, 1–7

Wang, R. (2012). Adaboost for Feature Selection, Classification and Its Relation with SVM, A Review. *Physics Procedia*, *25*, 800–807.

Wayan, N. (2013). Naïve Bayes Classifier Dan Support Vector Machines Untuk Sentiment Analysis. *Sistem Informasi*, 2–4.

Wu, X. (2009). The Top Ten Algorithms in Data Mining. Boca Raton: Taylor and Francis

Zhang, & Gao, F. (2011). An Improvement to NB for Text Classification. *Procedia Engineering*, *15*, 2160–2164.

Zhang, Ye, Q., Zhang, Z., & Li, Y. (2011). Sentiment classification of Internet restaurant reviews written in Cantonese. *Expert Systems with Applications*, *38*, 7674–7682.

Zhang, Z., Ye, Q., Law, R., & Li, Y. (2010). The impact of e-word-of-mouth on the online popularity of restaurants: A comparison of consumer reviews and editor reviews. *International Journal of Hospitality Management*, *29*, 694–700.


## BIOGRAFI PENULIS

**Lila Dini Utami**. Lahir pada tanggal 28 Juni 1988 di Jakarta. Memperoleh gelar Sarjana Komputer (S.Kom) dari STMIK Nusa Mandiri Jakarta (Jurusan Sistem Informasi) pada tahun 2011. Serta memperoleh gelar M.Kom dari Pascasarjana STMIK Nusa Mandiri pada tahun 2014 (Jurusan Ilmu Komputer).

**Romi Satria Wahono**. Memperoleh Gelar B.Eng dan M.Eng pada bidang ilmu komputer di Saitama University, Japan, dan Ph.D pada bidang sofwater engineering di Universiti Teknikal Malaysia Melaka. Menjadi pengajar dan peneliti di Fakultas Ilmu Komputer, Universitas Dia Nuswantoro. Merupakan pendiri dan CEO PT. Brainmatics, sebuah perusahaan yang bergerak di bidang pengembangan software. Minat penelitian pada bidang software engineering dan machine learning. Profesional member dari asosiasi ilmiah ACM, PMI dan IEEE Computer Society.