

Metoda *Naïve Bayes Classifier* dan Penggunaannya pada Klasifikasi Dokumen

Samuel Natalius / 18209031

Program Studi Sistem dan Teknologi Informasi

Sekolah Teknik Elektro dan Informatika

Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia

storelius@gmail.com

Abstract—This Paper, written in Bahasa Indonesia, presents a brief description about the *Naïve Bayes Classifier* method and its use for document classification. The description includes the explanation about the Bayes theorem as the fundamental theorem for the *Naïve Bayes* theorem and the example of the *Naïve Bayes Classifier* use for classification. Furthermore, it also describes about advantages and disadvantages of the method use and provides the real example of the document classification using *Naïve Bayes Classifier* in the present information technology's trends.

Index Terms— Document classification, *Naïve Bayes*, Spam filtering,

I. PENDAHULUAN

Proses transfer informasi pada jaman modern ini telah sampai kepada era elektronik, ditandai dengan semakin digunakannya teknologi berupa komputer dan jaringan internet sebagai sarana utama penyampaian informasi. Seiring waktu, informasi yang beredar melalui teknologi tersebut semakin banyak seiring dengan semakin banyaknya dokumen yang tersimpan dan bertransmisi di komputer dan jaringan internet. Hal ini mendukung diperlukannya suatu proses klasifikasi terhadap dokumen-dokumen.

Klasifikasi adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek (Agus Mulyanto 2009). Klasifikasi merupakan proses awal dari pengelompokan data. Proses klasifikasi dokumen merupakan proses yang sangat penting dalam bidang sistem informasi, khususnya dalam proses penambangan data (*data mining*) untuk memperoleh pengetahuan bisnis (*business knowledge*).

Ada beragam teknik klasifikasi dokumen, di antaranya adalah *Naïve Bayes classifier*, *Decision Trees*, dan *Support Vector Machines*. Pada makalah ini penulis membatasi pengkajian pada metoda *Naïve Bayes classifier*. Metoda *Naïve Bayes classifier* ini merupakan metoda yang paling populer digunakan dalam pengklasifikasian dokumen sekarang ini, khususnya dalam

penyaringan spam (*Spam filtering*).

II. NAÏVE BAYES CLASSIFIER

A. Teorema *Naïve Bayes*

Naïve Bayes Classifier merupakan sebuah metoda klasifikasi yang berakar pada teorema Bayes. Ciri utama dari *Naïve Bayes Classifier* ini adalah asumsi yang sangat kuat (naïf) akan independensi dari masing-masing kondisi/kejadian. Sebelum menjelaskan *Naïve Bayes Classifier* ini, akan dijelaskan terlebih dahulu Teorema Bayes yang menjadi dasar dari metoda tersebut.

Pada teorema Bayes, bila terdapat dua kejadian yang terpisah (misalkan A dan B), maka teorema Bayes dirumuskan sebagai berikut:

$$P(A|B) = \frac{P(A)}{P(B)} P(B|A) \quad (1)$$

Teorema Bayes sering pula dikembangkan mengingat berlakunya hukum probabilitas total, menjadi seperti berikut:

$$P(A|B) = \frac{P(A)P(B|A)}{\sum_{i=1}^n P(A_i|B)} \quad (2)$$

dimana $A_1 \cup A_2 \cup \dots \cup A_n = S$

Untuk menjelaskan teorema *Naïve Bayes*, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok bagi sampel yang dianalisis tersebut. Karena itu, teorema Bayes di atas disesuaikan sebagai berikut:

$$P(C|F_1, \dots, F_n) = \frac{P(C)P(F_1, \dots, F_n|C)}{P(F_1, \dots, F_n)} \quad (3)$$

Dimana variabel C merepresentasikan kelas, sementara variabel $F_1 \dots F_n$ merepresentasikan karakteristik-karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi. Maka rumus tersebut menjelaskan bahwa

peluang masuknya sampel dengan karakteristik tertentu dalam kelas C (*posterior*) adalah peluang munculnya kelas C (sebelum masuknya sampel tersebut, seringkali disebut *prior*), dikali dengan peluang kemunculan karakteristik-karakteristik sampel pada kelas C (disebut juga *likelihood*), dibagi dengan peluang kemunculan karakteristik-karakteristik sampel secara global (disebut juga *evidence*). Karena itu, rumus (3) dapat pula ditulis secara sederhana sebagai berikut:

$$Posterior = \frac{prior \times likelihood}{evidence} \quad (4)$$

Nilai *evidence* selalu tetap untuk setiap kelas pada satu sampel. Nilai dari *Posterior* tersebut yang nantinya akan dibandingkan dengan nilai-nilai *Posterior* kelas lainnya untuk menentukan ke kelas apa suatu sampel akan diklasifikasikan.

Penjabaran lebih lanjut rumus Bayes tersebut dilakukan dengan menjabarkan $P(F_1, \dots, F_n | C)$ menggunakan aturan perkalian, menjadi sebagai berikut:

$$\begin{aligned} P(F_1, \dots, F_n | C) &= P(F_1 | C) P(F_2, \dots, F_n | C, F_1) \\ &= P(F_1 | C) P(F_2 | C, F_1) P(F_3, \dots, F_n | C, F_1, F_2) \\ &\quad \vdots \\ P(F_1, \dots, F_n | C) &= P(F_1 | C) P(F_2 | C, F_1) \dots P(F_n | C, F_1, F_2, \dots, F_{n-1}) \end{aligned} \quad (5)$$

Dapat dilihat bahwa hasil penjabaran tersebut menyebabkan semakin banyak dan semakin kompleksnya faktor-faktor syarat yang mempengaruhi nilai probabilitas, yang hampir mustahil untuk dianalisa satu-persatu. Akibatnya, perhitungan tersebut menjadi sulit untuk dilakukan.

Di sinilah digunakan asumsi independensi yang sangat tinggi (naïf), bahwa masing-masing petunjuk ($F_1, F_2 \dots F_n$) saling bebas (independen) satu sama lain. Dengan asumsi tersebut, maka berlaku suatu kesamaan sebagai berikut:

$$P(F_i | F_j) = \frac{P(F_i \cap F_j)}{P(F_j)} = \frac{P(F_i)P(F_j)}{P(F_j)} = P(F_i)$$

untuk $i \neq j$, sehingga

$$P(F_i | C, F_j) = P(F_i | C) \quad (6)$$

Dari persamaan di atas dapat disimpulkan bahwa asumsi independensi naïf tersebut membuat syarat peluang menjadi sederhana, sehingga perhitungan menjadi mungkin untuk dilakukan. Selanjutnya, penjabaran $P(F_1, \dots, F_n | C)$ dapat disederhanakan menjadi seperti berikut:

$$P(F_1 \dots F_n | C) = P(F_1 | C) P(F_2 | C) \dots P(F_n | C)$$

$$P(F_1 \dots F_n | C) = \prod_{i=1}^n P(F_i | C) \quad (7)$$

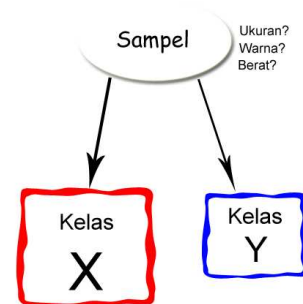
Dengan kesamaan di atas, persamaan teorema Bayes dapat dituliskan sebagai berikut:

$$\begin{aligned} P(C | F_1 \dots F_n) &= \frac{1}{P(F_1, F_2, \dots, F_n)} P(C) \prod_{i=1}^n P(F_i | C) \\ P(C | F_1 \dots F_n) &= \frac{P(C)}{Z} \prod_{i=1}^n P(F_i | C) \end{aligned} \quad (8)$$

Persamaan di atas merupakan model dari teorema *Naïve Bayes* yang selanjutnya akan digunakan dalam proses klasifikasi dokumen. Adapun Z merepresentasikan *evidence* yang nilainya konstan untuk semua kelas pada satu sampel.

B. Klasifikasi dengan *Naïve Bayes classifier*

Klasifikasi adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu obyek (Agus Mulyanto 2009). Oleh karena itu, kelas yang ada tentulah lebih dari satu. Penentuan kelas dari suatu dokumen dilakukan dengan cara membandingkan nilai probabilitas suatu sampel berada di kelas yang satu dengan nilai probabilitas suatu sampel berada di kelas yang lain.



Gambar 1. Ilustrasi contoh proses klasifikasi

Dengan persamaan teorema *Naïve Bayes* yang telah diturunkan di subbab A, kita mendapatkan nilai $P(C | F_1 \dots F_n)$, yaitu nilai peluang suatu sampel dengan karakteristik $F_1 \dots F_n$ berada dalam kelas C , atau dikenal dengan istilah *Posterior*. Umumnya kelas yang ada tidak hanya satu, melainkan lebih dari satu.

Sebagai contoh, ahli statistik ingin mengklasifikasikan sampel kucing ke dalam jenis kelaminnya. Oleh karena itu, terdapat dua kelas yaitu jantan dan betina. Suatu sampel kucing akan diklasifikasikan ke dalam satu kelas saja, entah itu jantan atau betina, dengan melihat petunjuk-petunjuk yang ada (misalnya berat badan, panjang ekor, dll).

Penentuan kelas yang cocok bagi suatu sampel dilakukan dengan cara membandingkan nilai *Posterior*

untuk masing-masing kelas, dan mengambil kelas dengan nilai *Posterior* yang tinggi. Secara matematis klasifikasi dirumuskan sebagai berikut:

$$C_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i=1}^n P(f_i|c) \quad (9)$$

dengan c yaitu variabel kelas yang tergabung dalam suatu himpunan kelas C .

Dapat dilihat bahwa rumusan di atas tidak memuat nilai *Evidence* (Z). Hal ini disebabkan karena *evidence* memiliki nilai yang positif dan tetap untuk semua kelas sehingga tidak mempengaruhi perbandingan nilai *Posterior*. Karena itu, faktor Z ini dapat dihilangkan. Perlu menjadi perhatian pula bahwa metoda *Naïve Bayes classifier* ini dapat digunakan bila sebelumnya telah tersedia data yang dijadikan acuan untuk melakukan klasifikasi.

Sebagai contoh, terdapat dua kelompok merek sepatu (X dan Y), dimana terdapat 3 petunjuk yang digunakan misalnya warna sepatu (merah, hitam), bahan sepatu (kulit, sintetis) dan model sepatu (Tali, Velkro). Sementara itu, terdapat pula 6 data seperti di bawah ini:

| Warna | Bahan | Model | Jenis |
|-------|----------|--------|-------|
| Merah | Kulit | Tali | X |
| Hitam | Kulit | Tali | X |
| Merah | Sintetis | Velkro | Y |
| Hitam | Kulit | Velkro | Y |
| Hitam | Sintetis | Tali | Y |
| Hitam | Sintetis | Velkro | X |

Tabel 1. Contoh data untuk klasifikasi metoda *Naïve Bayes classifier*

Bila terdapat sampel sepatu Hitam, Sintetis, Tali (tidak ada pada data di atas), klasifikasi dapat dilakukan dengan menggunakan *Naïve Bayes classifier*. Pertama-tama harus dicari terlebih dahulu *Posterior* X dan Y untuk sampel tersebut.

$$\begin{aligned} P(X) &= 3/6 = 0.5 & P(Y) &= 0.5 \\ P(\text{Hitam}|X) &= 2/3 = 0.66 & P(\text{Hitam}|Y) &= 1/3 = 0.33 \\ P(\text{Sintetis}|X) &= 1/3 = 0.33 & P(\text{Sintetis}|Y) &= 2/3 = 0.66 \\ P(\text{Tali}|X) &= 2/3 = 0.66 & P(\text{Tali}|Y) &= 1/3 = 0.33 \end{aligned}$$

$$\begin{aligned} \text{Posterior } X &= P(X) P(\text{Hitam}|X) P(\text{Sintetis}|X) P(\text{Tali}|X) \\ &= 0.5 \times 0.66 \times 0.33 \times 0.66 = 0.072 \end{aligned}$$

$$\begin{aligned} \text{Posterior } Y &= P(Y) P(\text{Hitam}|Y) P(\text{Sintetis}|Y) P(\text{Tali}|Y) \\ &= 0.5 \times 0.33 \times 0.66 \times 0.33 = 0.034 \end{aligned}$$

Karena *Posterior* $X > \text{Posterior } Y$, maka sampel sepatu tersebut bermerek X .

III. KLASIFIKASI DOKUMEN

A. Metoda *Naïve Bayes classifier* untuk Klasifikasi Dokumen

Secara umum teknik klasifikasi dokumen sama seperti klasifikasi pada umumnya. Hal yang membedakan adalah karakteristik yang ditinjau. Pada klasifikasi secara umum (misalnya benda fisik), karakteristik yang dapat ditinjau merupakan karakteristik fisik yang beragam seperti ukuran, warna, bahan, dan lain-lain. Pada klasifikasi dokumen, karakteristik semacam itu tidak dapat ditemukan, karena umumnya dokumen hanya terdiri dari data-data literal (tulisan). Karena itu harus terdapat asumsi mengenai karakteristik yang ditinjau agar metoda *Naïve Bayes classifier* dapat digunakan dalam klasifikasi dokumen.

Asumsi yang diambil dalam pengklasifikasian dokumen ini adalah dokumen dipandang sebagai kumpulan kata-kata yang saling bebas (independen) dan proses klasifikasi dokumen dilakukan dengan pengecekan kata-kata yang menyusun informasi di dalam dokumen tersebut. Penentuan kelas dari dokumen sampel dilakukan dengan cara menghitung besarnya peluang kata-kata pada dokumen suatu kelas yang muncul pada dokumen sampel yang dianalisis. Jadi, kata-kata dalam dokumen (w_i) berlaku seperti petunjuk-petunjuk yang telah dijelaskan sebelumnya (F_i), dan gabungan dari kata-kata tersebut menghasilkan suatu dokumen (D). Probabilitas sebuah kelas mengandung suatu dokumen merupakan produk dari probabilitas kata-kata dari dokumen tersebut yang terdapat pada kelas. Dengan kata lain:

$$P(D|C) = \prod_i P(w_i|C) \quad (10)$$

Persamaan teorema *Naïve Bayes* untuk klasifikasi dokumen menjadi sebagai berikut:

$$\begin{aligned} P(C|D) &= \frac{P(C)}{P(D)} P(D|C) \\ &= \frac{P(C)}{P(D)} \prod_i P(w_i|C) \end{aligned} \quad (11)$$

dimana $P(C|D)$ menyatakan kemungkinan suatu dokumen diklasifikasikan pada kelas C .

B. Aplikasi

Aplikasi nyata yang sangat populer dari penggunaan metoda *Naïve Bayes classifier* untuk klasifikasi dokumen dalam bidang teknologi informasi adalah penyaringan spam pada layanan surat elektronik (*spam filtering*). Spam adalah penyalahgunaan sistem pesan elektronik (termasuk media penyiaran dan sistem pengiriman digital) untuk mengirim berita iklan dan keperluan lainnya secara massal^[7]. Penyaringan surat elektronik spam (selanjutnya disebut “spam” saja) perlu dilakukan agar hanya informasi

yang relevan saja yang tersampaikan dan menghindari diterimanya informasi yang tidak berguna, apalagi jika informasi tersebut merugikan penerimanya.

Salah satu langkah penting dalam penyaringan spam pada layanan surat elektronik adalah proses klasifikasi surat elektronik, yang akan memasukkan surat elektronik yang masuk (sampel) ke dalam salah satu dari dua kategori, yaitu kategori spam atau bukan spam. Ada banyak cara yang dapat digunakan untuk mendeteksi apakah suatu surat elektronik dikategorikan spam atau tidak:

1. Mendeteksi alamat pengirim.

Apabila alamat pengirim pernah dikategorikan sebagai spam sebelumnya (oleh pengguna layanan surat elektronik tertentu), maka surat elektronik yang dikirim melalui alamat tersebut otomatis terdeteksi sebagai spam. Cara ini memiliki banyak kelemahan, salah satunya adalah alamat surat elektronik yang begitu banyak di internet sehingga mendeteksi spam hanya dengan melihat alamat pengirim tidaklah efektif. Terlebih lagi, sekarang ini sering terjadi pengiriman spam menggunakan alamat surat elektronik kerabat (oleh *automatic sender*) yang dengan mudah menipu perangkat lunak pendeteksi *spam*.

2. Mendeteksi isi surat elektronik.

Cara kedua ini menggunakan klasifikasi kata per kata untuk mengklasifikasikan surat elektronik sebagai spam atau bukan. Metoda klasifikasi yang paling populer digunakan saat ini adalah metoda *Naïve Bayes classifier*. Cara ini didasarkan kepada statistik bahwa kebanyakan spam menggunakan pemilihan kata yang hampir sama untuk isi surat elektroniknya, misalnya menjurus kepada merek barang tertentu, atau kepada nama seseorang. Dengan menghitung kecocokan kata-kata dalam suatu surat elektronik dengan *library* kata-kata yang dikategorikan spam atau bukan spam, maka perangkat lunak dapat menentukan apakah suatu surat elektronik dimasukkan ke dalam kategori spam atau tidak. Kata-kata yang tersimpan di *library* dapat berasal dari surat-surat elektronik sebelumnya yang telah dinyatakan spam secara manual oleh pengguna layanan surat elektronik.

Ketika sebuah surat elektronik dikirimkan ke alamat pengirim tertentu, layanan surat elektronik akan mengimplementasikan perangkat lunak untuk mendeteksi apakah surat elektronik tersebut dikategorikan ke dalam spam atau tidak. Algoritma perangkat lunak tersebut didasarkan kepada persamaan (10) yang disesuaikan untuk kelas yang ada, yaitu “spam” dan “bukan spam”.

Nilai probabilitas surat elektronik tertentu diklasifikasikan pada kelas “spam” dirumuskan sebagai berikut. Misalkan variabel S merepresentasikan kelas “spam”:

$$P(S|D) = \frac{P(S)}{P(D)} \prod_i P(w_i|S) \quad (12)$$

Sementara itu, untuk kelas “bukan spam” yang

merupakan negasi dari kelas “spam”:

$$P(\bar{S}|D) = \frac{P(\bar{S})}{P(D)} \prod_i P(w_i|\bar{S}) \quad (13)$$

Penentuan kelas bagi surat elektronik yang dianalisis tersebut dilakukan dengan cara mencari nilai probabilitas yang paling maksimum antara kedua nilai yang didapat. Bila didapat bahwa $P(S|D) > P(\bar{S}|D)$ maka surat elektronik tersebut masuk ke dalam kelas “spam”. Sebaliknya, bila $P(S|D) < P(\bar{S}|D)$ maka surat elektronik tersebut masuk ke dalam kelas “bukan spam”.

Perlu diperhatikan bahwa jumlah kata-kata yang terdapat pada setiap surat elektronik berbeda-beda. Hal ini mengindikasikan bahwa faktor w_i bersifat kuantitatif (dapat dihitung), bukan kualitatif seperti contoh pada subbab 2B, dan karenanya memiliki kurva distribusi beserta data statistik seperti rata-rata dan simpangan baku kemunculan suatu kata dalam sebuah surat elektronik.

Katakanlah bahwa kata “Hubungi” diduga sebagai kata yang menjurus kepada spam, dan besar kemunculan suatu kata “Hubungi” dalam sebuah surat elektronik memiliki distribusi peluang berbentuk distribusi normal, maka dalam menganalisis suatu surat elektronik yang masuk, nilai probabilitas untuk $P(\text{“Hubungi”}|S)$ dihitung dengan metode distribusi normal baku.

$$Z = \frac{X - \mu}{\sigma} \quad (14)$$

Dengan X adalah jumlah kata “Hubungi” yang ada di dalam surat elektronik, μ adalah rata-rata kemunculan kata “Hubungi” dalam sebuah surat elektronik (didapat dari data-data sebelumnya), dan σ adalah simpangan baku. Bila nilai Z yang didapat mendekati 0, artinya nilai X mendekati nilai rata-rata, yang juga mengindikasikan bahwa ada tingkat kecocokan yang tinggi antara surat elektronik yang dianalisis dengan surat-surat elektronik lainnya, yang telah diklasifikasikan sebagai spam. Maka, nilai probabilitas $P(\text{“Hubungi”}|S)$ tinggi. Sebaliknya, bila nilai Z jauh dari 0, maka nilai probabilitas $P(\text{“Hubungi”}|S)$ rendah.

IV. KELEBIHAN DAN KEKURANGAN

Kelebihan dari penggunaan *Naïve Bayes classifier* dalam klasifikasi dokumen dapat ditinjau dari prosesnya yang mengambil aksi berdasarkan data-data yang telah ada sebelumnya. Oleh karena itu, klasifikasi dokumen dengan metode ini dapat dipersonalisasi, maksudnya adalah proses klasifikasi dokumen dapat disesuaikan sesuai dengan sifat dan kebutuhan masing-masing orang.

Keuntungan ini secara nyata diperlihatkan dalam contoh *spam filtering* yang telah dicontohkan sebelumnya. Pernyataan suatu surat elektronik adalah spam atau tidak berbeda-beda bergantung pada subyek pembacanya yang

berbeda-beda. Suatu surat elektronik yang diklasifikasikan spam oleh satu orang mungkin diklasifikasikan bukan spam oleh orang lain, dan begitu pula sebaliknya. Dengan klasifikasi cara *Naïve Bayes classifier*, pengklasifikasian spam otomatis ini dapat disesuaikan dengan masing-masing orang sehingga meminimalisasi aksi salah pengklasifikasian secara personal.

Kekurangan dari metoda *Naïve Bayes classifier* ini adalah banyaknya celah untuk mengurangi keefektifan metoda ini dan akibatnya meloloskan dokumen ke dalam kelas tertentu padahal jelas-jelas dokumen tersebut tidak layak berada di kelas tersebut. Dalam kasus *spam filtering*, kelemahan ini banyak digunakan oleh *spammers* berpengalaman untuk meloloskan *spam* ke dalam kelas bukan spam (menganggap surat elektronik bukan spam padahal sebenarnya adalah spam : Galat tipe II).

Banyak cara yang dapat dilakukan, misalnya dengan memasukkan kata-kata yang asing dituliskan sehingga perangkat lunak tidak dapat melakukan pengecekan, atau dengan memasukkan banyak kata yang sebenarnya sering digunakan oleh surat elektronik non-spam agar pengguna secara manual mendeteksi sebagai spam dan untuk selanjutnya perangkat lunak akan mendeteksi surat elektronik dengan kata-kata non-spam tersebut sebagai spam serta memperkecil nilai probabilitas kata-kata spam (memanfaatkan *false positive*/galat tipe I).

Cara lain adalah dengan memanfaatkan media gambar untuk menyampaikan spam. Hal ini didasarkan kepada metoda *Naïve Bayes classifier* yang dirancang hanya untuk mendeteksi kata-kata dan bukan gambar. Akibatnya, perangkat lunak tidak mampu untuk menganalisis gambar dan akhirnya mengklasifikasikan spam tersebut ke dalam kelas bukan spam.

V. KESIMPULAN

Metoda *Naïve Bayes classifier* merupakan metoda klasifikasi yang berdasar kepada teorema bayes, sebuah teorema yang terkenal di dalam bidang ilmu probabilitas. Selain itu, metoda ini turut didukung oleh ilmu statistika khususnya dalam penggunaan data petunjuk untuk mendukung keputusan pengklasifikasian.

Metoda ini sangat luas dipakai dalam berbagai bidang, khususnya dalam proses klasifikasi dokumen. Klasifikasi ini merupakan salah satu teknik dalam *data mining* yang merupakan kegiatan penunjang dalam bidang sistem informasi.

Seperti halnya metoda-metoda lain, metoda *Naïve Bayes classifier* ini tidaklah 100% sempurna. Ada banyak kelebihan dan kekurangan dari metoda ini, yang dapat menjadi dasar bahan kajian lebih lanjut untuk mendapatkan atau mengembangkan metoda klasifikasi lain, yang dapat bekerja dengan lebih efektif dan efisien, serta mengurangi jumlah titik kelemahan yang dapat disalahgunakan oleh orang lain.

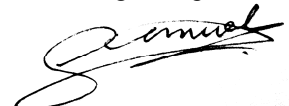
DAFTAR PUSTAKA

- [1] Agus Mulyanto, "Sistem Informasi Konsep & Aplikasi". Cetakan I, Yogyakarta: Pustaka Pelajar, Desember 2009, hal. 204-206.
- [2] Naïve Bayes Example. http://jmvidal.cse.sc.edu/talks/Bayesian_learning/nbex.xml. Tanggal Akses: 13 Desember 2010.
- [3] Otodidak Teknologi Informasi & Komunikasi: *Data Mining*. <http://visilubai.wordpress.com/2010/04/28/data-mining/>. Tanggal Akses: 13 Desember 2010.
- [4] Salmon Run: Document Classification using Naïve Bayes. <http://sujitpal.blogspot.com/2007/04/document-classification-using-naive.html>. Tanggal Akses : 13 Desember 2010.
- [5] Wikipedia: *Bayesian spam filtering*. http://en.wikipedia.org/wiki/Bayesian_spam_filtering. Tanggal Akses : 13 Desember 2010.
- [6] Wikipedia: *Naïve Bayes classifier*. http://en.wikipedia.org/wiki/Naive_Bayes_classifier. Tanggal Akses : 13 Desember 2010.
- [7] Wikipedia Indonesia: Spam. <http://id.wikipedia.org/wiki/Spam>. Tanggal Akses : 16 Desember 2010

PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 29 April 2010



Samuel Natalius / 18209031