APLIKASI KLASIFIKASI DOKUMEN MENGGUNAKAN METODA NAÏVE BAYSIAN

Marvin Chandra Wijaya[1)], Semuil Tjiharjadi[2)]
[1,2)]Jurusan Sistem Komputer, Universitas Kristen Maranatha Bandung
Jl. Suria Sumantri 65, Bandung 022 - 2012186
e-mail : marvinchw@gmail.com

Abstrak
Suatu makalah yang diterima atau didapatkan oleh sebuah institusi atau perorangan melalui berbagai sumber merupakan makalah dengan berbagai macam klasifikasi. Misalnya suatu perpustakaan perlu memilah-milah dokumen atau makalah yang diterima ke dalam berbagai kategori. Sebagai contoh suatu makalah dapat merupakan salah satu dari kateogri berikut : Komputer, Elektro / elektronika, Teknik Sipil, Teknik Industri, Ekonomi dan kedokteran.
Banyak metode yang dapat digunakan untuk melakukan proses klasifikasi terhadap data, salah satu yang akan digunakan adalah dengan menggunakan metode Naïve Bayessian. Dengan menggunakan metoda ini perlu adanya pembelajaran terlebih dahulu sebelum program dapat digunakan untuk melakukan klasifikasi dokumen / makalah. Pertama-tama program diberi sejumlah input berupa data dokumen / makalah yang sudah diklasifikasikan terlebih dahulu. Setelah cukup banyak data makalah / dokumen untuk setiap kategorinya, maka program sudah siap digunakan untuk melakukan klasifikasi dokumen.

***Keyword** : Klasifikasi Dokumne, Naïve Baysian*

1. PENDAHULUAN

Pada saat ini sudah banyak sekali dokumen-dokumen yang dipublish di internet atau pun menggunakan berbagai media lainnya. Orang-orang atau instusi seperti perpustakaan yang membutuhkan dokumen dapat mencari dokumen tersebut dari mana saja. Setelah dicari dokumen-dokumen tersebut, maka dokumen tersebut perlu dipilah-pilah (diklasifikasikan) agar lebih teroganisir dengan baik penyimpan dokumen tersebut.

2. TINJAUAN PUSTAKA

2.1 Algoritma Naive Bayesian

Pemrograman Klasifikasi dokumen yang dipakai untuk mengkategorisasikan dokumen ini menggunakan algoritma yang disebut *naive* Bayesian. *Naive* disini bermakna bahwa untuk setiap bahasa, kata-kata yang muncul dianggap bermakna tunggal.

2.2 Rumus Bayes

Dasar dari teorema *naive* Bayesian yang dipakai dalam pemrograman adalah rumus Bayes:
P (A | B) = (P (B | A) * P (A)) / P (B)
Peluang kejadian A sebagai B ditentukan dari peluang B saat A, peluang A, dan peluang B. Pada pengaplikasiannya nanti rumus ini berubah menjadi :
P (Ci | D) = (P (D | Ci) * P (Ci)) / P (D)
Dimana pada rumus ini:

- P (Ci | D) adalah peluang dokumen D pada kategori Ci.
- P (D | Ci) adalah peluang pada kategori Ci, kata pada dokumen D muncul pada kategori tersebut.
- P (Ci) adalah peluang dari kategori yang diberikan, dibandingkan dengan kategori-kategori lainnya yang dianalisa.
- P (D) adalah peluang dari dokumen tersebut secara spesifik. Pada pengembangannya, P (D) dapat dihilangkan karena nilainya tetap, sehingga saat dibandingkan dengan tiap kategori, nilai ini dapat dihapus.

2.3 Pengaplikasian Naive Bayesian

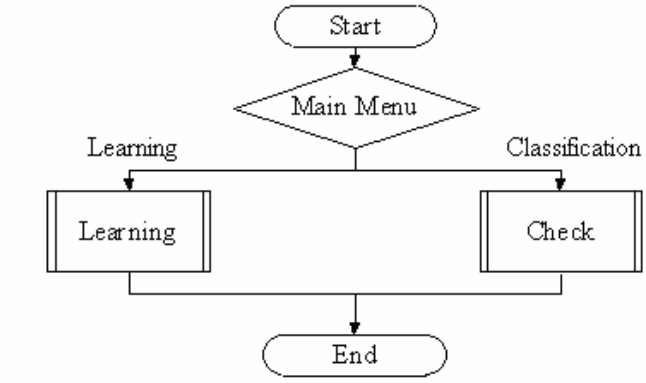
Pada pengaplikasiannya didalam program, rumus akan berubah menjadi:
P (*Komputer* | D) = (P (D | Komputer) * P (Komputer))
P (Elektro | D) = (P (D | Elektro) * P (Elektro))
P (Teknik Industri | D) = (P (D | Teknik Industri) * P (Teknik Industri))

P (Teknik Sipil | D) = (P (D | Teknik Sipil) * P (Teknik Sipil))
P (Ekonomi | D) = (P (D | Ekonomi) * P (Ekonomi))
P (Kedokteran | D) = (P (D | Kedokteran) * P (Kedokteran))

Dapat dilihat seperti pada keterangan P (D), nilai peluang dokumen secara spesifik dianggap sama sehingga dihilangkan terlebih dahulu dari awal pemrograman agar program berjalan lebih efisien. Dengan melihat rumusan di atas, ini berarti bahwa peluang dokumen D pada *Komputer, Elektro / elektronika, Teknik Sipil, Teknik Industri, Ekonomi dan kedokteran* adalah peluang pada kategori *Komputer, Elektro / elektronika, Teknik Sipil, Teknik Industri, Ekonomi dan kedokteran*, kata pada dokumen D muncul pada kategori tersebut dikalikan dengan peluang kategori *Komputer, Elektro / elektronika, Teknik Sipil, Teknik Industri, Ekonomi dan kedokteran* tersebut. Setelah ini didapatkan, hasilnya dijumlahkan untuk tiap kata yang terdapat pada dokumen (untuk masing-masing peluang, *Komputer, Elektro / elektronika, Teknik Sipil, Teknik Industri, Ekonomi dan kedokteran*). Sehingga akhirnya diperoleh enam buah nilai, peluang dokumen D sebagai *Komputer, Elektro / elektronika, Teknik Sipil, Teknik Industri, Ekonomi dan kedokteran*. Diakhir program, keenam nilai ini dibandingkan. Nilai peluang yang lebih tinggi menandakan dokumen D tersebut sebagai kategori tersebut.

3. METODE PENELITIAN

Program ini terbagi menjadi lima bagian blok diagram alir. Menu utama yang dapat memanggil *subprogram Learning* dan *Check,* dimana *learning* proses adalah bagian input awal sebagai pembelajaran *database* dan *check* adalah bagian pengklasifikasian dokumen nantinya. Bagian *Input database* (*learning*) akan terbagi dalam enam buah diagram alir, masing-masing untuk tiap *database*. Berikut adalah diagram alir yang dipakai dalam pembuatan program
Gambar dibawah adalah *flowchart* untuk *Menu* utama program.



Gambar 1. *Flowchart Menu* Utama

4. HASIL DAN PEMBAHASAN

Langkah yang pertama adalah dengan memasukkan terlebih dahulu data untuk pemebelajaran, masing-masing diberi input 5 buah dokumen (Komputer, Teknik Elektro, Teknik Sipil, Teknik Industri, Ekonomi dan Kedokteran).
Pada studi kasus dipergunakan masing-masing 5 buah dokumen lain yang berbeda (Komputer, Teknik Elektro, Teknik Sipil, Teknik Industri, Ekonomi dan Kedokteran).

Tabel 1.
Studi Kasus Pertama

Dokumen	Hasil Pengklasifikasian	Nilai kebenaran
Komputer 1	Komputer	Benar
Komputer 2	Teknik Elektro	Salah
Komputer 3	Komputer	Benar
Komputer 4	Komputer	Benar
Komputer 5	Teknik Elektro	Salah
Teknik Elektro 1	Teknik Elektro	Benar
Teknik Elektro 2	Teknik Elektro	Benar
Teknik Elektro 3	Teknik Elektro	Benar
Teknik Elektro 4	Komputer	Salah
Teknik Elektro 5	Teknik Elektro	Benar
Teknik Sipil 1	Teknik Sipil	Benar
Teknik Sipil 2	Teknik Sipil	Benar
Teknik Sipil 3	Teknik Sipil	Benar
Teknik Sipil 4	Teknik Sipil	Benar
Teknik Sipil 5	Teknik Industri	Salah
Teknik Industri 1	Teknik Industri	Benar
Teknik Industri 2	Teknik Industri	Benar
Teknik Industri 3	Teknik Sipil	Salah
Teknik Industri 4	Teknik Industri	Benar
Teknik Industri 5	Teknik Industri	Benar
Ekonomi 1	Ekonomi	Benar
Ekonomi 2	Ekonomi	Benar
Ekonomi 3	Ekonomi	Benar
Ekonomi 4	Ekonomi	Benar
Ekonomi 5	Ekonomi	Benar
Kedokteran 1	Kedokteran	Benar
Kedokteran 2	Kedokteran	Benar
Kedokteran 3	Kedokteran	Benar
Kedokteran 4	Kedokteran	Benar
Kedokteran 5	Kedokteran	Benar

Dari 30 data yang diujikan ada 5 yang salah. Persentase kesalahannya adalah

$$\text{Kesalahan} = \frac{5}{30} \text{ X } 100\% = 16{,}67\%$$

Kemudian dilakukan penambahan database dengan melakukan penambahan pembelajaran yaitu dengan menambahkan 5 buah dokumen (Komputer, Teknik Elektro, Teknik Sipil, Teknik Industri, Ekonomi dan Kedokteran).
Pada studi kasus kedua dipergunakan masing-masing 5 buah dokumen yang tadi diujikan pada studi kasus yang pertama (Komputer, Teknik Elektro, Teknik Sipil, Teknik Industri, Ekonomi dan Kedokteran).

Tabel 2.
Studi Kasus Pertama

Dokumen	Hasil Pengklasifikasian	Nilai kebenaran
Komputer 1	Komputer	Benar
Komputer 2	Komputer	Benar
Komputer 3	Komputer	Benar
Komputer 4	Komputer	Benar
Komputer 5	Teknik Elektro	Salah
Teknik Elektro 1	Teknik Elektro	Benar
Teknik Elektro 2	Teknik Elektro	Benar
Teknik Elektro 3	Teknik Elektro	Benar
Teknik Elektro 4	Teknik Elektro	Benar
Teknik Elektro 5	Teknik Elektro	Benar
Teknik Sipil 1	Teknik Sipil	Benar
Teknik Sipil 2	Teknik Sipil	Benar
Teknik Sipil 3	Teknik Sipil	Benar
Teknik Sipil 4	Teknik Sipil	Benar
Teknik Sipil 5	Teknik Sipil	Benar
Teknik Industri 1	Teknik Industri	Benar
Teknik Industri 2	Teknik Industri	Benar
Teknik Industri 3	Teknik Industri	Benar
Teknik Industri 4	Teknik Industri	Benar
Teknik Industri 5	Teknik Industri	Benar
Ekonomi 1	Ekonomi	Benar
Ekonomi 2	Ekonomi	Benar
Ekonomi 3	Ekonomi	Benar
Ekonomi 4	Ekonomi	Benar
Ekonomi 5	Ekonomi	Benar
Kedokteran 1	Kedokteran	Benar
Kedokteran 2	Kedokteran	Benar
Kedokteran 3	Kedokteran	Benar
Kedokteran 4	Kedokteran	Benar
Kedokteran 5	Kedokteran	Benar

Dari 30 data yang diujikan ada 1 yang salah. Persentase kesalahannya adalah

$$\text{Kesalahan} = \frac{1}{30} \text{ X } 100\% = 3{,}3\ \%$$

5. KESIMPULAN

Berikut ini adalah kesimpulan yang diperoleh dari hasil percobaan :

1. Program untuk mengidentifikasi sebuah dokumen, sehingga dapat terklasifikasi sebagai dokumen komputer, elektro, teknik sipil, teknik industri, ekonomi atau kedokteran telah berhasil dibuat.
2. Tingkat keberhasilan dari program klasifikasi dokumen ini dipengaruhi berdasarkan banyaknya jenis dan variasi kata pada keenam database, serta jumlah kata yang sama pada keenam database.
3. Pada dokumen Komputer dan Teknik Elektro masih ada sedikit kesalahan, dikarenakan cukup banyak kata-kata yang sama digunakan pada kedua bidang ilmu tersebut.

6. DAFTAR PUSTAKA

1. Graham, John-Cumming; 2005; *Build Your Own Naïve Bayesian Spam Filter*; The Spammers' Compendium.
2. Halvorson, Michael; 2000; *Step by Step Microsoft Visual Basic 6.0 prof.*; Alih Bahasa : Adi Kurniadi, PT. Elex Media Komputindo, Jakarta.
3. Pamungkas; 2002; *Tip & Trik Microsoft Visual Basic 6.0*; cetakan keempat, PT. Elex Media Komputindo, Jakarta.
4. Sulaiman, Agus; 2007; *Koneksi Database Dengan ADODC*; Jakarta.
5. http://www.jgc.org, 5 Maret 2007.