

# Klasifikasi Berita Berbahasa Indonesia menggunakan *Naïve Bayes Classifier*<sup>1</sup>

Yudi Wibisono

Jurusan Pendidikan Matematika FPMIPA UPI

Jl. DR. Setiabudhi 229. Bandung 40154

[yudi@upi.edu](mailto:yudi@upi.edu) , <http://matematika.upi.edu/staff/yudi>

## Abstrak

Pada makalah ini dibahas penggunaan teorema Bayes untuk mengklasifikasikan secara otomatis berita berbahasa Indonesia. Makalah ini juga membahas hasil eksperimen klasifikasi berita yang berasal dari situs [www.kompas.com](http://www.kompas.com) dan menunjukkan bahwa metode Naïve Bayes efektif untuk klasifikasi berita berbahasa Indonesia.

## 1. Pendahuluan

Jumlah sumber berita berbahasa Indonesia yang tersedia di internet semakin besar. Menurut situs <http://dmoz.org/World/Indonesia/Berita/> pada bulan Agustus 2005 terdapat 128 situs berita berbahasa Indonesia. Hal ini menimbulkan aliran informasi berita dalam jumlah besar setiap harinya. Klasifikasi berita secara otomatis, yaitu proses penggolongan suatu berita ke dalam suatu kategori semakin dibutuhkan untuk melakukan analisis berita.

Salah satu metode klasifikasi yang dapat digunakan adalah metode Naïve Bayes yang sering disebut sebagai *Naïve Bayes Classifier* (NBC). Kelebihan NBC adalah sederhana tetapi memiliki akurasi yang tinggi [1,3]. NBC menggunakan teori probabilitas sebagai dasar teori.

Ada dua tahap pada proses klasifikasi teks. Tahap pertama adalah pelatihan terhadap himpunan artikel contoh (*training example*). Sedangkan tahap kedua adalah proses klasifikasi dokumen yang belum diketahui kategorinya.

## 2. Metode Naïve Bayes untuk Klasifikasi Teks

Penggunaan metode naïve Bayes untuk klasifikasi teks telah dibahas oleh [3]. Berikut akan disajikan garis besar metode naïve Bayes untuk klasifikasi teks.

Pada NBC setiap dokumen berita direpresentasikan dalam pasangan atribut  $\langle a_1, a_2, \dots, a_n \rangle$  di mana  $a_1$  adalah kata pertama,  $a_2$  kata kedua dan seterusnya. Sedangkan  $V$  adalah himpunan kategori berita (olahraga, ilmu teknologi dan sebagainya).

Pada saat klasifikasi, pendekatan Bayes akan menghasilkan label kategori yang paling tinggi probabilitasnya ( $v_{MAP}$ ) dengan masukan atribut  $\langle a_1, a_2, \dots, a_n \rangle$

$$v_{MAP} = \arg \max_{v_j \in V} P(v_j | a_1, a_2, \dots, a_n) \quad (2.1)$$

---

<sup>1</sup> Makalah yang disampaikan pada Seminar Nasional Matematika 2005 di Universitas Pendidikan Indonesia, Bandung, tanggal 20 Agustus 2005.

Teorema Bayes menyatakan:

$$P(B | A) = \frac{P(A | B)P(B)}{P(A)} \quad (2.2)$$

Menggunakan teorema Bayes ini, persamaan (2.1) ini dapat ditulis:

$$v_{MAP} = \arg \max_{v_j \in V} \frac{P(a_1, a_2 \dots a_n | v_j)P(v_j)}{P(a_1, a_2 \dots a_n)} \quad (2.3)$$

$P(a_1, a_2 \dots a_n)$  nilainya konstan untuk semua  $v_j$  sehingga persamaan ini dapat ditulis sebagai berikut:

$$v_{MAP} = \arg \max_{v_j \in V} P(a_1, a_2 \dots a_n | v_j)P(v_j) \quad (2.4)$$

Tingkat kesulitan menghitung  $P(a_1, a_2 \dots a_n | v_j)$  menjadi tinggi karena jumlah term  $P(a_1, a_2 \dots a_n | v_j)$  bisa jadi akan sangat besar. Ini disebabkan jumlah term tersebut sama dengan jumlah semua kombinasi posisi kata dikali dengan jumlah kategori.

*Naïve Bayes Classifier* menyederhanakan hal ini dengan mengasumsikan bahwa di dalam setiap kategori, setiap kata independen satu sama lain. Dengan kata lain:

$$P(a_1, a_2 \dots a_n | v_j) = \prod_i P(a_i | v_j) \quad (2.5)$$

Substitusi persamaan ini dengan persamaan 2.4 akan menghasilkan:

$$v_{MAP} = \arg \max_{v_j \in V} P(v_j) \prod_i P(a_i | v_j) \quad (2.6)$$

$P(v_j)$  dan probabilitas kata  $w_k$  untuk setiap kategori  $P(w_k | v_j)$  dihitung pada saat pelatihan.

$$P(v_j) = \frac{|docs_j|}{|Contoh|} \quad (2.7)$$

$$P(w_k | v_j) = \frac{n_k + 1}{n + |kosakata|} \quad (2.8)$$

di mana  $|docs_j|$  adalah jumlah kata pada kategori  $j$  dan  $|Contoh|$  adalah jumlah dokumen yang digunakan dalam pelatihan. Sedangkan  $n_k$  adalah jumlah kemunculan kata  $w_k$  pada kategori  $v_j$ ,  $n$  adalah jumlah semua kata pada kategori  $v_j$  dan  $|kosakata|$  adalah jumlah kata yang unik (*distinct*) pada semua data latihan.

Ringkasan algoritma untuk *Naïve Bayes Classifier* adalah sebagai berikut:

- A. Proses pelatihan. Input adalah dokumen-dokumen contoh yang telah diketahui kategorinya:
1. Kosakata  $\mathcal{B}$  himpunan semua kata yang unik dari dokumen-dokumen contoh
  2. Untuk setiap kategori  $v_j$  lakukan:
    - a.  $\text{Docs}_j \subseteq \mathcal{B}$  Himpunan dokumen-dokumen yang berada pada kategori  $v_j$
    - b. Hitung  $P(v_j)$  dengan persamaan 2.7
    - c. Untuk setiap kata  $w_k$  pada kosakata lakukan:
      - i. Hitung  $P(w_k | v_j)$  dengan persamaan 2.8
- B. Proses klasifikasi. Input adalah dokumen yang belum diketahui kategorinya:
1. Hasilkan vmap sesuai dengan persamaan 2.6 dengan menggunakan  $P(v_j)$  dan  $P(w_k | v_j)$  yang telah diperoleh dari pelatihan

### 3. Hasil Eksperimen

Data yang digunakan untuk eksperimen bersumber dari situs [www.kompas.com](http://www.kompas.com) yang diambil mulai tanggal 29 Juni 2005 sampai dengan 15 Juli 2005. Jumlah dokumen yang digunakan adalah 582. Ada enam kategori berita yaitu: Metro, Kesehatan, Olahraga, Teknologi dan Gaya Hidup.

Pengolahan awal (*preprocessing*) dilakukan dengan menghilangkan semua karakter selain huruf, menjadikan semua kata menjadi huruf kecil dan kemudian menghilangkan 329 kata yang paling sering muncul (*stop words*).

*Stop words* dipilih dengan cara mengambil 500 kata yang paling sering muncul seperti: “yang”, “di”, “dan”, “itu”. Kemudian secara manual dipisahkan kata-kata yang bukan *stop word* walaupun kata tersebut sering muncul (contoh: “spbu”, “bbm”).

Dokumen kemudian dibagi menjadi dua bagian. Bagian pertama berperan sebagai data contoh yang akan digunakan dalam proses pelatihan. Sedangkan bagian kedua digunakan sebagai data pengujian untuk melihat tingkat akurasi.

Akurasi dihitung dengan:

$$\text{Akurasi} = \frac{\text{Jumlah Klasifikasi Benar}}{\text{Jumlah Dokumen Ujicoba}} \times 100\%$$

Tabel 1 memperlihatkan hasil eksperimen dengan berbagai proporsi antara data contoh dan data uji coba

<b>Jumlah Dokumen Contoh</b>	<b>Jumlah Dokumen Uji Coba</b>	<b>Akurasi (%)</b>
524 (90%)	58 (10%)	89.47
407 (70%)	175 (30%)	90.23
291 (50%)	291 (50%)	86.90
175 (30%)	407 (70%)	85.47
58 (10%)	524 (90%)	68.64

**Tabel 1: Akurasi dengan berbagai proporsi dokumen contoh dan dokumen ujicoba**

Dari Tabel 1 terlihat bahwa nilai akurasi NBC tinggi, terutama jika dokumen contoh yang digunakan besar ( $\geq 400$  dokumen). Hal yang menarik untuk diteliti lebih lanjut adalah akurasi masih tetap relatif tinggi walaupun dokumen contoh secara ekstrim dikurangi hanya sebesar 58 dokumen (10%).

#### **4. Kesimpulan**

Berdasarkan hasil eksperimen, NBC terbukti dapat digunakan secara efektif untuk mengklasifikasikan berita secara otomatis. Algoritma NBC yang sederhana dan kecepatannya yang tinggi dalam proses pelatihan dan klasifikasi membuat algoritma ini menarik untuk digunakan sebagai salah satu metode klasifikasi.

#### **REFERENSI**

1. I. Rish: An empirical study of the naive Bayes classifier.
2. Hinrich Schütze: Information Retrieval and Text Mining.
3. Tom M. Mitchell: Machine Learning, McGraw-Hill, 1997.