

# Klasifikasi Data Forum dengan menggunakan Metode *Naïve Bayes Classifier*

Aida Indriani

Program Studi Teknik Informatika, STMIK PPKIA Tarakanita Rahmawati  
Jl. Yos Sudarso 8 Tarakan, Indonesia  
ditamaniez2014@gmail.com

**Abstract**—Klasifikasi adalah suatu proses pengkategorian yang dilakukan terhadap sekumpulan dokumen. Klasifikasi sangat penting untuk kemudahan pengguna dalam melakukan pencarian dokumen. Proses klasifikasi diawali dengan membagi koleksi dokumen menjadi 2 bagian utama yaitu data latih dan data uji. Dari data latih, dengan menggunakan metode tertentu diperoleh model klasifikasi yang nantinya akan digunakan untuk penentuan kelas terhadap data uji. Banyak sekali metode-metode yang dapat digunakan dalam klasifikasi dokumen, antara lain: *K-NN*, *Rocchio*, *WAK-NN*, *Naïve Bayes Classifier* (NBC), dan lain-lain.

Forum adalah tempat/media untuk melakukan komunikasi atau diskusi. Pada bidang pendidikan, keberadaan forum sangat penting untuk menunjang proses belajar mengajar. Dengan adanya forum, dosen dan mahasiswa bisa saling berdiskusi tanpa harus bertemu langsung. Dalam forum terdapat beberapa kelas, dengan tujuan agar memudahkan dosen dan mahasiswa dalam mencari bahan diskusi. Pendataan kelas yang tepat untuk pertanyaan atau pernyataan yang ditulis oleh pengguna menjadi sangat penting.

Metode klasifikasi yang digunakan untuk mengklasifikasikan data forum secara otomatis pada penelitian ini yaitu NBC. NBC adalah metode klasifikasi yang berdasarkan probabilitas dan Teorema Bayes. Dengan NBC, diperoleh hasil akurasi kebenaran pengklasifikasian terhadap 15 data uji sebesar 73%. Pada penelitian ini, pengukuran efektifitas klasifikasi dengan menggunakan Confusion Matrix dengan cara menentukan nilai TP, TN, FP, dan FN.

**Keywords** – Naïve Bayes Classifier, Forum Engine, Confusion Matrix

## I. INTRODUCTION

Seiring dengan perkembangan teknologi, cara manusia berkomunikasi atau berdiskusi tidak harus selalu bertatap muka atau bertemu secara langsung dengan orang-orang. Salah satunya adalah dengan menyediakan sebuah fasilitas (media). Salah satunya adalah dengan menciptakan sebuah komunitas maya. Komunitas maya diciptakan untuk saling berkomunikasi diantara para pengguna internet dengan menggunakan teknologi yang menggunakan platform internet.

Ada berbagai macam jenis dari komunitas maya, salah satunya adalah forum berbasis web yang sering dikenal dengan *Forum Engine* (mesin forum) [1]. Forum adalah sebuah sarana komunikasi bagi penggunanya. Biasanya, didalam forum itu sendiri terdapat beberapa kategori forum dengan tujuan memudahkan pengguna dalam berinteraksi [2]. Saat ini, banyak sekali forum engine yang biasa ditemukan melalui internet, antara lain: [www.indoforum.org](http://www.indoforum.org), [www.kaskus.us](http://www.kaskus.us), [www.bluefame.com](http://www.bluefame.com), [www.kafegaul.com](http://www.kafegaul.com), dan masih banyak lagi.

Dalam dunia pendidikan sebuah komunikasi sangat penting sebagai penunjang pembelajaran, baik komunikasi antara dosen dan mahasiswa maupun antar mahasiswa itu sendiri. Sampai saat ini, komunikasi atau diskusi yang biasa dilakukan masih banyak yang melalui tatap muka atau bertemu secara langsung. Diskusi yang dilakukan melalui tatap muka masih terdapat kekurangan-kekurangannya, antara lain: sibuk atau padatnnya jadwal dosen sehingga membuat mahasiswa harus antri atau menunggu jika ada sesuatu yang ingin didiskusikan, tentu saja ini akan membuang waktu, segan atau takutnya mahasiswa untuk bertemu dengan dosen secara langsung, dan lain sebagainya.

Salah satu cara untuk merubah gaya berdiskusi antara dosen dan mahasiswa maupun antar mahasiswa itu sendiri yaitu dengan membuat sebuah media komunikasi yaitu forum. Didalam forum itu sendiri terdapat beberapa kelas (kategori) dengan tujuan untuk memudahkan mahasiswa dalam mencari informasi terkait dengan tugas-tugas matakuliah maupun dalam pembuatan skripsi/tugas akhir. Untuk menghasilkan pengkategorian data forum yang tepat yaitu dengan melakukan proses klasifikasi terhadap pertanyaan atau pernyataan yang dituliskan oleh pengguna. Dalam penelitian ini, penulis menggunakan metode Naïve Bayes Classifier untuk melakukan klasifikasi otomatis berdasarkan pertanyaan atau pernyataan yang dimasukkan oleh pengguna. Untuk mengukur akurasi klasifikasi data forum, penulis menggunakan Confusion Matrix.

## II. THEORITICAL BASIS

### 2.1 Pre-Processing

*Pre-processing* dalam proses klasifikasi dokumen digunakan untuk membangun sebuah *index* dari koleksi dokumen. *Index* adalah himpunan *term* yang menunjukkan isi atau topik yang dikandung oleh dokumen. Pembuatan *inverted*

*index* harus melibatkan konsep *linguistic processing* yang bertujuan meng-ekstrak term-term penting dari dokumen yang dipresentasikan sebagai *bag-of-words*. Ekstraksi term biasanya melibatkan tiga operasi utama, antara lain:

1. Pemisahan rangkaian term (*tokenization*). *Tokenization* adalah tugas memisahkan deretan kata di dalam kalimat, paragraf atau halaman menjadi token atau potongan kata tunggal atau *termmedword*. Tahapan ini juga menghilangkan karakter-karakter tertentu seperti tandabaca dan mengubah semua *token* ke bentuk huruf kecil (*lower case*).
2. Penghapusan *stop-words*. *Stop word* didefinisikan sebagai term yang tidak berhubungan (*irrelevant*) dengan subyek utama dari database meskipun katatersebut sering kali hadir di dalam dokumen. Contoh *stop words* adalah ada, adalah, adanya, adapun, agak, dll.
3. *Stemming*. Kata-kata yang muncul di dalam dokumen sering mempunyai banyak varian morfologik. Karena itu, setiap kata yang bukan *stop-words* direduksi ke *stemmed word (term)* yang cocok yaitu kata tersebut distem untuk mendapatkan bentuk akarnya dengan menghilangkan awalan atau akhiran. Dengan cara ini, diperoleh kelompok kata yang cocok dimana kata-kata di dalam kelompok tersebut merupakan varian sintaktis dari satu samalain dan dapat menghimpun hanya satu kata per kelompok. Sebagai contoh, kata terdaftar, pendaftaran, berbagi-pakai, term umum daftar, dan dapat diperlakukan sebagai bentuk lain dari kata ini [3].

## 2.2 Naïve Bayes Classifier (NBC)

Klasifikasi dokumen adalah bidang penelitian dalam perolehan informasi yang mengembangkan metode untuk menentukan atau mengkategorikan suatu dokumen ke dalam satu atau lebih kelompok yang telah dikenal sebelumnya secara otomatis berdasarkan isi dokumen. Klasifikasi dokumen bertujuan untuk mengelompokkan dokumen yang tidak terstruktur ke dalam kelompok-kelompok yang menggambarkan isi dari dokumen. Dokumen dapat berupa dokumen teks seperti artikel berita [4]. Pada bagian ini membahas tentang bagaimana meng-klasifikasikan data forum diskusi sekitar matakuliah maupun tugas akhir/skripsi. Pada klasifikasi terdapat 2 (dua) proses yaitu model pelatihan dan model pengujian. Pada model pelatihan terdapat suatu kumpulan data yang memiliki kelas yang telah ditetapkan (data latih). Kumpulan data latih digunakan untuk membentuk model klasifikasi. Pada model pengujian, model klasifikasi yang sudah terbentuk diuji dengan data yang lainnya (data uji) untuk mengetahui akurasi dari model klasifikasi tersebut. Model klasifikasi yang dihasilkan dapat digunakan untuk menentukan kelas dari data yang baru [5].

NBC adalah metode klasifikasi yang berdasarkan probabilitas dan *Teorema Bayesian* dengan asumsi bahwa setiap variable X bersifat bebas (*independence*). Dengan kata lain, NBC mengansumsikan bahwa keberadaan sebuah atribut (*variable*) tidak ada kaitannya dengan keberadaan atribut (*variable*) yang lain [6]. Metode NBC menempuh dua tahap dalam proses klasifikasi teks, yaitu tahap pelatihan dan tahap

klasifikasi. Pada tahap pelatihan dilakukan proses analisis terhadap sampel dokumen berupa pemilihan *vocabulary*, yaitu kata yang mungkin muncul dalam koleksi dokumen sampel yang sedapat mungkin dapat menjadi representasi dokumen. Selanjutnya adalah penentuan probabilitas prior bagi tiap kategori berdasarkan sampel dokumen. Pada tahap klasifikasi ditentukan nilai kategori dari suatu dokumen berdasarkan term yang muncul dalam dokumen yang diklasifikasi [7]. Perhitungan perbandingan antara term pada data testing dengan setiap kelas yang ada dengan menggunakan persamaan 1.

$$P(a_i|v_j) = \frac{n_c + mp}{n + m} \quad (1)$$

dimana:

$n$  = jumlah term pada data latih dimana  $v = v_j$

$n_c$  = jumlah term dimana  $v = v_j$  dan  $a = a_i$

$p$  = probabilitas setiap kelas dalam data latih

$m$  = jumlah term pada data uji

sedangkan untuk menentukan klasifikasi pada data uji, digunakan persamaan 2 [8].

$$V_{nb} = \operatorname{argmax}_{v_j \in V} P(v_j) \prod_i P(a_i|v_j) \quad (2)$$

## 2.3 Confusion Matrix

Confusion matrix adalah sebuah tabel yang menyatakan jumlah data uji yang benar diklasifikasikan dan jumlah data uji yang salah diklasifikasikan. Contoh confusion matrix untuk klasifikasi biner ditunjukkan pada tabel 1.

Tabel 1  
Confusion Matrix untuk klasifikasi biner

		Kelas Prediksi	
		1	0
Kelas Sebenarnya	1	TP	FN
	0	FP	TN

Keterangan untuk tabel 1 dinyatakan sebagai berikut:

*True Positive (TP)*, yaitu jumlah dokumen dari kelas 1 yang benar dan diklasifikasikan sebagai kelas 1.

*True Negative (TN)*, yaitu jumlah dokumen dari kelas 0 yang benar diklasifikasikan sebagai kelas 0.

*False Positive (FP)*, yaitu jumlah dokumen dari kelas 0 yang salah diklasifikasikan sebagai kelas 1.

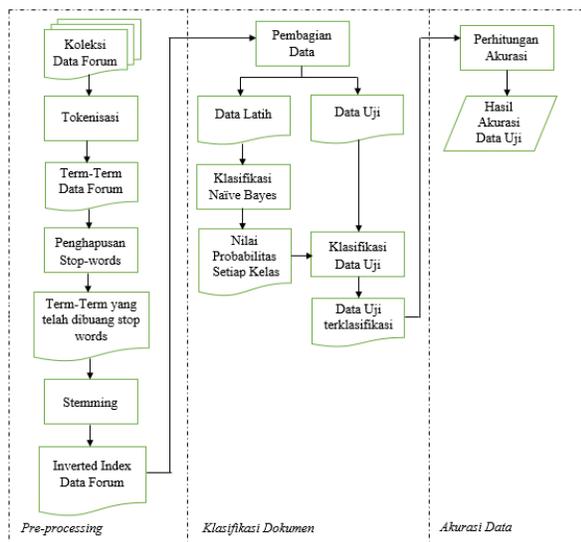
*False Negative (FN)* yaitu jumlah dokumen dari kelas 1 yang salah diklasifikasikan sebagai kelas 0.

Perhitungan akurasi dinyatakan dalam persamaan 3 [9].

$$\text{Akurasi} = \frac{TP + TN}{TP + FN + FP + TN} \times 100\% \quad (3)$$

### III. ARCHITECTURE SYSTEM

Arsitektur sistem digunakan untuk menggambarkan sistem kerja yang digunakan pada proses analisa dan implementasi, adapun arsitektur sistem dari keseluruhan sistem yang digunakan ditunjukkan pada Gambar 1.



Gambar.1 Architecture System

Tahap pertama dimulai dari pre-processing koleksi dokumen data forum yang dimulai dari proses tokenisasi yaitu pemisahan rangkaian term menjadi bentuk token atau potongan kata tunggal dan di dalam proses tokenisasi juga menghilangkan tanda baca dan merubah token menjadi huruf kecil, kemudian dilanjutkan dengan penghapusan stop-words yaitu membuang term yang tidak penting dalam koleksi dokumen, diakhiri dengan proses stemming yaitu mengubah term kembali menjadi bentuk kata dasar. Pada tahap pre-processing juga akan menghasilkan sebuah inverted index yaitu matriks antara term dan koleksi data forum.

Tahap kedua yaitu Inverted index yang telah dihasilkan dari proses pre-processing dibagi menjadi 2 bagian yaitu data latih dan data uji. Dari data latih, dengan menggunakan NBC dilakukan perhitungan untuk menghasilkan nilai probabilitas setiap kelas. Nilai probabilitas setiap kelas yang dihasilkan digunakan untuk melakukan klasifikasi data uji sehingga menghasilkan data uji yang terklasifikasi.

Tahap terakhir dari pengklasifikasian data forum ini yaitu mengukur prosentasi ketepatan (akurasi) dalam menentukan klasifikasi dari data uji yang diperoleh dari perhitungan pada tahap kedua. Untuk mengukur akurasi klasifikasi data uji digunakan sebuah confusion matrix yaitu dengan menentukan berapa nilai TP, TN, FP dan FN.

### IV. RESULT AND ANALYSIS

#### 1. Pre-processing

Pre-processing dilakukan terhadap sekumpulan koleksi data forum yang telah memiliki kelas masing-masing yang

dilakukan secara manual. Berikut adalah contoh koleksi data forum sesuai dengan kelas masing-masing ditunjukkan pada tabel 2.

TABEL 2. KOLEKSI DATA FORUM

No.	Data Forum	Kelas
D1	Bagaimana cara membuat Galeri Image pada Eclipse	Android
D2	Ada yang tau gak cara membuat koneksi pada Delphi dengan MySQL	Delphi
D3	Saya kesulitan dalam membuat Mailmerge pada Ms.Word, bagaimana caranya ya?	Office
D4	Membuat fungsi Sum pada Ms.Excel seperti apa ya?	Office
D5	Bagaimana cara menghilangkan warning pada Android	Android
D6	Bagaimana membuat form cetak dengan Delphi	Delphi

Pada tabel 2 terdapat 6 contoh data forum yang mempunyai kelas masing-masing yaitu ada 2 data forum yang mempunyai kelas "Android", 2 data mempunyai kelas "Delphi" dan 2 data mempunyai kelas "Office". Dari data forum pada tabel 2, dilakukan tahapan pertama dalam pre-processing yaitu tokenisasi. Bentuk hasil tokenisasi ditunjukkan pada tabel 3.

TABEL 3. HASIL TOKENISASI DATA FORUM

No.	Data Forum
D1	bagaimana cara membuat galeri image pada eclipse
D2	ada yang tau gak cara membuat koneksi pada delphi dengan mysql
D3	saya kesulitan dalam membuat mailmerge pada ms word bagaimana caranya ya
D4	membuat fungsi sum pada ms excel seperti apa ya
D5	bagaimana cara menghilangkan warning pada android
D6	bagaimana membuat form cetak dengan delphi

Tabel 3 merupakan data forum yang telah melewati tahap pertama dalam proses pre-processing yaitu tokenisasi. Data dipecah menjadi token (term), huruf pada term diubah menjadi huruf kecil dan semua tanda baca dihilangkan. Selanjutnya yaitu masuk kedalam tahapan kedua dalam pre-processing yaitu penghapusan stop-words. Bentuk hasil penghapusan stop-words ditunjukkan pada tabel 4.

TABEL 4. HASIL PENGHAPUSAN STOP-WORDS DATA FORUM

No.	Data Forum
D1	galeri image eclipse
D2	koneksi delphi mysql
D3	kesulitan mailmerge ms word
D4	fungsi sum ms excel
D5	menghilangkan warning android
D6	form cetak delphi

Tabel 4 merupakan data forum yang telah melewati tahap kedua dalam proses pre-processing yaitu penghapusan stop-words. Ada beberapa kata yang dihilangkan, contohnya saja pada data forum yang ke-D1, kata “bagaimana”, “cara”, dan “pada” dihapus. Selain menggunakan kumpulan daftar stop-words yang sering digunakan dalam penghapusan stop-words, penulis juga membuat sekumpulan stop-words yang memiliki arti yang sama dengan daftar stop-words yang biasanya digunakan. Dikarenakan dalam penulisan data forum, bahasa yang digunakan tidak selalu bahasa formal. Bahasa yang digunakan dalam penulisan forum biasanya mengikuti bahasa yang sering digunakan saat ini. Contoh kata-kata yang juga dimasukkan kedalam stop-words ditunjukkan pada tabel 5.

TABEL 5. CONTOH DAFTAR STOP WORDS YANG DITAMBAHKAN

Kata	Kata yang mempunyai arti yang sama
tidak	tdk, gak, g, enggak
kamu	u, kmu
boleh	blh
cara	cr, cra
simpan	smpn, save

Setelah melalui tahapan penghapusan stop-words, langkah terakhir yaitu tahapan stemming yaitu merubah term menjadi bentuk kata dasar dengan menghilangkan awalan dan akhiran dari term. Bentuk hasil stemming data forum ditunjukkan pada tabel 6.

TABEL 6. HASIL STEMMING DATA FORUM

No.	Data Forum
D1	galeri image eclipse
D2	koneksi delphi mysql
D3	sulit mailmerge ms word
D4	fungsi sum ms excel
D5	hilang warning android
D6	form cetak delphi

Pada tabel 6, diperoleh kumpulan kata yang telah di-stemming. Ada beberapa kata yang mengalami proses stemming, contohnya pada data forum ke-D3 yaitu kata “kesulitan” berubah bentuk menjadi “sulit” dengan menghilangkan awalan “ke” dan akhiran “an”. Pada tahapan pre-processing, diperoleh inverted index yaitu matriks antara term dan data. Contoh Inverted index yang terbentuk dari beberapa term dan data ditunjukkan pada tabel 7.

TABEL 7. CONTOH MATRIKS ANTARA BEBERAPA TERM DAN DATA FORUM

Term	D1	D2	D3	D4	D5	D6
galeri	1					
image	1					
eclipse	1					
koneksi		1				
delphi		1				1
....	...	...	...	...	...	...
....	...	...	...	...	...	...
android					1	

form						1
cetak						1

Tabel 7 merupakan matriks hubungan antara term dan data. Kata “delphi” terdapat pada data forum D2 dan D6, begitu juga dengan kata “ms” terdapat pada data forum D3 dan D4. Langkah selanjutnya yaitu melakukan proses klasifikasi dengan NBC.

## 2. Naïve Bayes Classifier (NBC)

Tahapan awal pada proses NBC yaitu dengan menghitung probabilitas masing-masing kelas terhadap keseluruhan data latih. Pada tabel 2, diketahui bahwa data latih sebanyak 6 data dengan jumlah kelas sebanyak 3 kategori yaitu “Android”, “Office”, dan “Delphi”. Berikut adalah perhitungan probabilitas untuk kelas “Android”.

Probabilitas disimbolkan sebagai p.

$$p(\text{Android}) = \frac{\text{Jumlah kelas android}}{\text{Jumlah data latih}}$$

$$p(\text{Android}) = \frac{2}{6} = 0,33$$

Dari perhitungan probabilitas untuk kelas “Android”, diperoleh nilai probabilitas yaitu 0,33. Untuk nilai probabilitas kelas “Office” dan “Delphi” dilakukan perhitungan yang sama seperti pada perhitungan probabilitas kelas “Android”. Dari perhitungan yang dilakukan, diperoleh nilai probabilitas untuk kelas “Office” sebesar 0,33 dan kelas “Delphi” sebesar 0,33.

Langkah berikutnya adalah melakukan peng-klasifikasian data uji dengan menggunakan nilai probabilitas dari setiap kelas yang telah diperoleh. Berikut adalah contoh data forum yang akan diklasifikasikan secara otomatis dengan menggunakan NBC. Data forum uji ditunjukkan pada tabel 8.

TABEL 8. KOLEKSI DATA UJI FORUM

No.	Data Forum	Kelas
D7	Tolong dong, gimana membuat daftar isi secara otomatis dengan Ms. Word.	?
D8	Bingung nich, cara memasukkan gambar pada eclipse dan bagaimana menjalankan aplikasi androidnya.	?

Pada tabel 8, terdapat 2 data uji yang belum memiliki kelas. Untuk tahapan awal, data uji juga melalui tahapan pre-processing seperti yang dilalui oleh data latih. Data uji yang telah melewati tahapan pre-processing, ditunjukkan pada tabel 9.

TABEL 9. PRE-PROCESSING DATA UJI FORUM

No.	Data Forum	Kelas
D7	daftar isi otomatis ms word	?
D8	masuk gambar eclipse jalan aplikasi android	?

Tabel 9 merupakan 2 data forum yang akan diklasifikasikan secara otomatis dengan menggunakan NBC.

Langkah awal yaitu dengan menentukan nilai  $n$ ,  $n_c$ ,  $p$  dan  $m$  untuk setiap kelas yang ada dan untuk setiap term yang terdapat pada data uji, seperti yang telah dijelaskan pada poin 2.3. Berikut adalah contoh nilai-nilai untuk kelas “Android”.

Android	Term “daftar”	Term “isi”	Term “otomatis”
$n = 6$	$n = 6$	$n = 6$	$n = 6$
$n_c = 0$	$n_c = 0$	$n_c = 0$	$n_c = 0$
$p = 0,33$	$p = 0,33$	$p = 0,33$	$p = 0,33$
$m = 5$	$m = 5$	$m = 5$	$m = 5$
Term “ms”	Term “word”		
$n = 6$	$n = 6$		
$n_c = 0$	$n_c = 0$		
$p = 0,33$	$p = 0,33$		
$m = 5$	$m = 5$		

Dari nilai-nilai yang telah diperoleh dan dengan menggunakan persamaan 1 diperoleh perhitungan sebagai berikut:

$$P(\text{Android}|\text{Daftar}) = \frac{0 + 5 \cdot 0,33}{6 + 5} = \frac{1,65}{11} = 0,15$$

$$P(\text{Android}|\text{Isi}) = \frac{0 + 5 \cdot 0,33}{6 + 5} = \frac{1,65}{11} = 0,15$$

Untuk perhitungan  $P(a_i|v_j)$  yang lainnya dilakukan proses yang sama seperti pada perhitungan  $P(\text{Android}|\text{Daftar})$ . Dengan menggunakan persamaan 2, yaitu mencari nilai maksimal dari hasil perkalian nilai probabilitas dan nilai  $P$  untuk setiap kelas, diperoleh hasil sebagai berikut:

$$V(\text{Android}) = 0,33 \cdot 0,15 \cdot 0,15 \cdot 0,15 \cdot 0,15 = 0,000026$$

$$V(\text{Delphi}) = 0,33 \cdot 0,15 \cdot 0,15 \cdot 0,15 \cdot 0,15 = 0,000026$$

$$V(\text{Office}) = 0,33 \cdot 0,13 \cdot 0,13 \cdot 0,13 \cdot 0,28 \cdot 0,21 = 0,000040$$

$$V_{nb} = \text{argmax} (v(\text{android}) | v(\text{delphi}) | v(\text{office}))$$

$$V_{nb} = \text{argmax} (0,000026 | 0,000026 | 0,000040)$$

$$V_{nb} = 0,000040$$

Nilai maksimal yang diperoleh yaitu 0,000040. 0,000040 nilai  $v$  untuk kelas “Office”. Jadi kesimpulan yang diperoleh adalah data uji D7 termasuk kelas “Office”. Untuk data uji D8 dilakukan proses yang sama seperti pada data uji D7.

### 3. Confusion Matrix

Confusion matrix digunakan untuk pengukuran efektifitas klasifikasi. Dengan menggunakan persamaan 3, dilakukan proses perhitungan akurasi untuk 15 data uji. Hasil klasifikasi dengan NBC untuk 15 data uji ditunjukkan pada tabel 10.

TABEL 10. HASIL KLASIFIKASI DATA UJI

No.	Kelas Sebenarnya	Kelas NBC
D7	Office	Office
D8	Android	Android
D9	Office	Android
D10	Delphi	Delphi

D11	Delphi	Delphi
D12	Office	Office
D13	Android	Office
D14	Android	Android
D15	Delphi	Delphi
D16	Office	Delphi
D17	Office	Office
D18	Android	Android
D19	Delphi	Delphi
D20	Android	Delphi
D21	Delphi	Delphi

Dari tabel 10, diperoleh nilai  $TP = 3$ ,  $TN = 8$ ,  $FP = 2$ ,  $FN = 2$ . Langkah berikutnya yaitu dengan menggunakan persamaan 3, dihitung nilai akurasi sebagai berikut:

$$\text{Akurasi} = \frac{3 + 8}{3 + 2 + 2 + 8} \times 100\% = \frac{11}{15} = 73\%$$

Kesimpulan yang diperoleh yaitu, dari 15 data uji diperoleh akurasi kecocokan kelas sebenarnya terhadap kelas prediksi dengan NBC sebesar 73%.

## V. CONCLUSIONS

Dari pembahasan pada bab-bab sebelumnya pada penelitian ini, dapat diambil kesimpulan sebagai berikut :

1. Penggunaan metode NBC dapat digunakan untuk pengklasifikasian otomatis terhadap data forum dengan tingkat akurasi klasifikasi sebesar 73% dengan menggunakan pengukuran efektifitas Confusion Matrix.
2. Untuk pengembangan lebih lanjut, dapat menambahkan fitur-fitur khusus dalam tokenisasi kalimat dengan menggunakan Bi-gram atau gabungan Uni-gram dan Bi-gram sehingga mendapatkan hasil klasifikasi dengan akurasi yang lebih tinggi.

## REFERENCES

- [1] Wikipedia. [http://id.wikipedia.org/wiki/Komunitas\\_maya](http://id.wikipedia.org/wiki/Komunitas_maya). (diakses tanggal 07 Maret 2014).
- [2] Student Forum Community. <http://student.roomforum.com/t15-apa-itu-forum>. (diakses tanggal 07 Maret 2014).
- [3] Husni Ilyas. 2010. Unified Messaging System Information Retrieval & Klasifikasi Teks. Komputasi | Suatu Permulaan Data Mining & IR, pp.6-8.
- [4] Herny Februariyanti dan Eri Zuliarso. 2012. Klasifikasi Dokumen Berita Teks Bahasa Indonesia menggunakan Ontologi. Jurnal Teknologi Informasi DINAMIK Volume 17, No. 1, pp.14-23.
- [5] Jiawei Han dan Micheline Kamber. 2006. Data Mining: Concepts and Technique, 2<sup>nd</sup> ed.(Chapter 6). University of Illionis at Urbana-Champaign.
- [6] Taufik Fuadi Abidin,. Naïve Bayesian Classifier. Bahan Kuliah Data Mining Program Studi Teknik Informatika FMIPA Universitas Syiah Kuala, pp.7.
- [7] Amir Hamzah. 2012. Klasifikasi Teks dengan Naïve Bayes Classifier (NBC) untuk Pengelompokan Teks Berita dan Abstract Akademis. Prosiding Seminar Nasional Apikasi Sains & Teknologi (SNAST) Periode III, p.B269-B277. Yogyakarta.
- [8] Eric Meisner. 2010. Naïve Bayes Classifier Example”.

- [9] Kristina Paskianti. 2011. Klasifikasi Dokumen Tumbuhan Obat menggunakan Algoritma KNN Fuzzy. Thesis Fakultas Matematika dan Ilmu Pengetahuan Alam IPB, Bogor.