

# KLASIFIKASI JENIS BUKU BERDASARKAN JUDUL DAN SINOPSIS MENGGUNAKAN METODE NAÏVE BAYES CLASSIFIER (STUDI KASUS: STMIK KADIRI)

**Bonifacius Vicky Indriyono<sup>1)</sup>, Ema Utami<sup>2)</sup> Andi Sunyoto<sup>3)</sup>**

<sup>1</sup>Program Pascasarjana Magister Teknik Informatika STMIK AMIKOM Yogyakarta

Jl. Ring Road Utara Condong Catur, Depok, Sleman, Yogyakarta

Telp : (0274) 884 201-204, Fax : (0274) 884 208

E-mail : [bonifaciusvicky@gmail.com](mailto:bonifaciusvicky@gmail.com)<sup>1</sup>

---

## **Abstract**

*A lot of it foots up library material as book available at library each year cause library officer strucks a snag in do classification process binds books. Such too which faced by library officer at STMIK Kadiri to self that as this research object. Type determination or book category initially been done manually namely selectively registers type or available book category deep system which long time. Process that done by it a lot of evoke fault because is done up on officer science of read title and also processed book synopsis one on eventually cause book not lies on rack and also category that necessarily so goes hard with library user to look for needful library material.*

*Up on about problem that therefore in this research is built one application gets desktop based utilize Delphi compiler where in that application load stemming's step with Stemmer's Porter to indonesian and classification phase with Naïve Bayes Classifier. To know Naïve Bayes accuration, therefore done by test-driving several times with training's data amount and testing is different every category. Test-driving result of 900 subdivided document as 100 books every subdivided category deep training's data and testing's data in observational it concludes that in height classification accuration is gotten with many its document purpose as training's data conversely accuration becomes little with progressively at least document purpose as training's data.*

## **Abstrak**

*Banyaknya jumlah bahan pustaka berupa buku-buku baru yang tersedia di perpustakaan setiap tahun menyebabkan petugas perpustakaan mengalami kesulitan dalam melakukan proses klasifikasi buku. Demikian pula yang dihadapi petugas perpustakaan di STMIK Kadiri Kediri yang menjadi obyek penelitian ini. Penentuan jenis atau kategori buku pada awalnya dilakukan secara manual yakni dengan memilih daftar jenis atau kategori buku yang tersedia dalam sistem yang lama. Proses yang dilakukan ini banyak menimbulkan kesalahan karena dilakukan berdasarkan pada pengetahuan petugas dari membaca judul maupun sinopsis buku yang diproses yang pada akhirnya menyebabkan buku-buku tidak berada pada rak maupun kategori yang semestinya sehingga menyulitkan pengguna perpustakaan untuk mencari bahan pustaka yang diperlukan. Berdasarkan pada permasalahan tersebut maka dalam penelitian ini dibangun sebuah aplikasi berbasis desktop menggunakan compiler Delphi dimana dalam aplikasi tersebut memuat tahapan stemming dengan Porter Stemmer untuk bahasa Indonesia dan tahap klasifikasi dengan Naïve Bayes Classifier. Untuk mengetahui tingkat akurasi Naïve Bayes, maka dilakukan uji coba beberapa kali dengan jumlah data training dan testing yang berbeda tiap kategori. Hasil uji coba dari 900 dokumen yang terbagi menjadi 100 buku tiap kategori yang terbagi dalam data training dan data testing dalam penelitian ini menyimpulkan bahwa tingginya akurasi klasifikasi didapatkan dengan banyaknya penggunaan dokumen sebagai data training dan sebaliknya akurasi akan menurun dengan semakin sedikitnya penggunaan dokumen sebagai data training.*

**Kata kunci:** klasifikasi, naïve bayes classifier, porter stemmer, testing, learning

## 1. PENDAHULUAN

Perpustakaan merupakan media penyedia informasi yang di dalamnya terdapat aktivitas pengumpulan, pengolahan, pengawetan, pelestarian dan penyajian serta penyebaran informasi. Perpustakaan sebagaimana yang ada dan berkembang sekarang telah dipergunakan sebagai salah satu pusat informasi, sumber ilmu pengetahuan, penelitian, rekreasi, pelestarian khasanah budaya bangsa, serta memberikan berbagai layanan jasa lainnya. Menurut Sutarno N.S, (2006),” Perpustakaan adalah suatu ruangan, bagian dari gedung/bangunan, atau gedung tersendiri, yang berisi buku-buku koleksi, yang disusun dan diatur sedemikian rupa, sehingga mudah untuk dicari dan dipergunakan sewaktu-waktu diperlukan oleh pembaca”. Sebagian besar masyarakat beranggapan bahwa perpustakaan merupakan tempat tumpukan buku tanpa mengetahui pasti ciri dan fungsi perpustakaan.

Dalam pengelolaan perpustakaan khususnya perpustakaan kampus, masih banyak dijumpai berbagai kendala. Beberapa kendala yang dihadapi didalam perpustakaan adalah masalah pengolahan materi perpustakaan yang meliputi : katalogisasi, klasifikasi, penentuan tajuk subjek, dan penjajaran.

Dari beberapa kendala diatas, masalah klasifikasi menjadi penting karena menyangkut pada proses pencarian bahan pustaka yang dilakukan oleh pengguna perpustakaan dalam hal ini mahasiswa maupun dosen. Kemudahan pencarian bahan pustaka sangat ditentukan oleh bagaimana klasifikasi buku-buku tersebut diatur dengan baik.

Untuk menyelesaikan masalah diatas, maka dalam penelitian ini digunakan metode *Naïve Bayes Classifier* (NBC) dengan pertimbangan bahwa aturan dalam NBC lebih mudah dipahami, proses pengkodean yang sederhana apabila diimplementasikan dan lebih cepat dalam proses perhitungan. Selain metode NBC, peneliti juga menggunakan algoritma *Porter Stemmer* untuk bahasa Indonesia dengan pertimbangan karena penggunaan algoritma ini tidak bergantung pada kamus kata dasar melainkan menggunakan algoritma berbasis aturan dalam pencarian kata dasarnya.

Dari uraian diatas, maka dapat diambil rumusan masalah sebagai berikut :

- 1). Bagaimanakah implementasi penggunaan metode *Naïve Bayes Classifier* dan *Porter stemmer* dalam proses klasifikasi jenis buku?,
- 2). Bagaimanakah pengaruh dari penggunaan *stemming Porter stemmer* pada implementasi

klasifikasi jenis buku ?,

- 3). Bagaimanakah tingkat akurasi dari metode *Naïve Bayes Classifier* dalam melakukan klasifikasi jenis buku berdasarkan judul dan sinopsis?.

Sedangkan batasan masalah dalam penelitian ini antara lain :

- 1). Dokumen yang digunakan dalam penelitian ini adalah data buku fisik dan bentuk file dengan judul serta sinopsisnya berbahasa Indonesia,
- 2). Data buku uji dalam bentuk file yang digunakan berxtention .doc/.docx, .rtf dan .txt. beberapa tujuan yang ingin dicapai dalam penelitian ini diantaranya adalah untuk mengetahui seberapa besar tingkat akurasi penerapan metode *Naïve Bayes Classifier* (NBC) dalam proses klasifikasi jenis buku dan untuk mengetahui peranan metode *porter stemmer* untuk bahasa Indonesia pada proses klasifikasi jenis buku.

Sebelum melakukan proses klasifikasi dengan metode NBC dan *porter stemmer* untuk bahasa Indonesia dalam penelitian ini, beberapa tahapan telah dilakukan mulai dari studi literatur, kajian pustaka, pengumpulan data-data yang diperlukan, memahami teori *Naïve Bayes Classifier* dan *Porter Stemmer* untuk bahasa Indonesia sampai dengan pada proses perancangan serta implementasi terhadap kedua metode tersebut.

### 1.1 Kajian Pustaka

Kajian pustaka ini dilakukan dengan tujuan untuk melihat dan mempelajari sistem yang telah dilakukan oleh peneliti sebelumnya yang berkaitan dengan topik penelitian sekaligus sebagai bahan referensi untuk pembangunan sistem didalam penelitian ini.

Beberapa penelitian terdahulu tentang klasifikasi denganmetode *Naïve Bayes Classifier* yang sudah pernah dilakukan peneliti sebelumnya antara lain penelitian yang dilakukan oleh :

- 1). Selvia Lorena Br Ginting dan Reggy Pasya Trinanda (2011) dengan judul penelitian : “ Teknik Data Mining Menggunakan Metode Bayes Classifier Untuk Optimalisasi Pencarian Pada Aplikasi Perpustakaan ”. Dalam penelitian ini menjelaskan proses pencarian data buku perpustakaan dan proses pengklasifikasian menggunakan metode *Naïve Bayes Classifier* yang diimplementasikan dalam sebuah aplikasi. Hasil dari penelitian ini menyimpulkan bahwa hasil pencarian dan klasifikasi lebih akurat.
- 2). Supriyanto, Alun Sujjada dan Siti Widyayanti (2011) dengan judul penelitian : “Otomasi

katalog perpustakaan melalui klasifikasi kategori buku dengan menggunakan metode naïve bayes". Dalam penelitian ini memaparkan pemanfaatan Naïve Bayes Classifier untuk otomatisasi katalog dan menyimpulkan bahwa metode Naïve Bayes berjalan baik untuk proses klasifikasi.

3). Cahyo Darujati dan Agustinus Bimo Gumelar (2012) dalam penelitian yang berjudul "Pemanfaatan Teknik *Supervised* Untuk Klasifikasi Teks Bahasa Indonesia". Penelitian ini membahas tentang pengembangan dan pemanfaatan aplikasi untuk mengklasifikasi teks bahasa Indonesia secara terpandu (*supervised*), bahasa Indonesia secara terpandu (*supervised*), dengan menerapkan metode naïve bayes.

4). Bambang Kurniawan, Syahril Effendi dan Opim Salim Sitompul (2012) dalam penelitian yang berjudul "Klasifikasi Konten Berita Dengan Metode *Text Mining*". Dalam penelitian ini akan dibangun sebuah sistem berbasis *web* dimana sistem tersebut dapat mengklasifikasikan berita secara otomatis.

5). Amir Hamzah (2012) dengan penelitian yang berjudul "Klasifikasi Teks Dengan Naïve Bayes Classifier (Nbc) Untuk Pengelompokan Teks Berita Dan Abstract Akademis". Penelitian ini mengkaji kinerja NBC untuk kategorisasi teks berita dan teks akademis.

## 1.2 Pengertian Klasifikasi

Menurut Agus Mulyanto, (2009), klasifikasi adalah proses menemukan model atau fungsi yang menjelaskan atau membedakan konsep/kelas data, dengan tujuan dapat memperkirakan kelas dari suatu objek dan merupakan proses awal pengelompokan data. Menurut Walpole, E. R., Myers, R. H., (1995), klasifikasi adalah salah satu tugas yang penting dalam data mining.

Dalam klasifikasi, sebuah pengklasifikasi dibuat dari sekumpulan data latih dengan kelas yang telah di tentukan sebelumnya. Performa pengklasifikasi biasanya diukur dengan ketepatan (atau tingkat galat), sedangkan menurut Han & Kamber, (2006) proses klasifikasi dibagi menjadi dua tahap, yaitu tahapan pembelajaran dan pengujian.

Pada tahap pembelajaran, sebagian data yang telah diketahui kelasnya (data latih) digunakan untuk membuat model klasifikasi. Tahap pengujian menguji data uji dengan model klasifikasi untuk mengetahui akurasi model klasifikasi tersebut. Jika akurasi cukup maka model tersebut dapat digunakan untuk

memprediksi kelas data yang belum diketahui. Manfaat dari klasifikasi ini menurut Eryono, (1999), manfaat yang dapat diambil dari proses klasifikasi dokumen/bahan pustaka adalah : memudahkan dalam mengadakan perimbangan koleksi yg dimiliki, memudahkan penelusuran terhadap bahan pustaka menurut subyek dan memudahkan dalam membuat bibliografi menurut pokok masalah.

## 1.3 Porter Stemmer Untuk Bahasa Indonesia

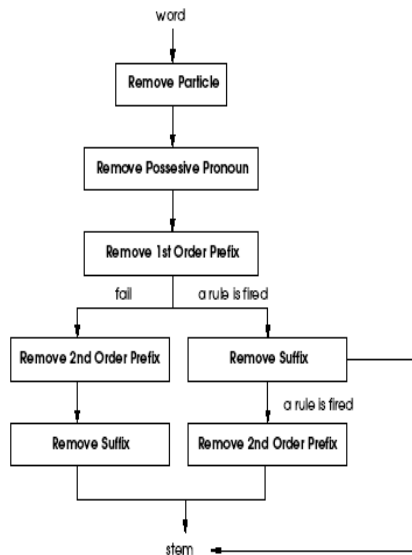
Algoritma stemmer berbahasa Indonesia tanpa menggunakan kamus diperkenalkan oleh Vega VB dan Bressan S, dengan menghilangkan imbuhan-imbuhan pada kata-kata berbahasa Indonesia berimbuhan. Selain Vega, Tala juga memperkenalkan porter like stemmer untuk bahasa Indonesia.

Dalam penelitian yang dilakukan oleh Fadilah Z. Tala dijelaskan, untuk korpus (data yang dipakai sebagai sumber bahan penelitian) yang berkembang dan dalam jumlah yang besar, ketergantungan pada kamus akan menurunkan kemampuan sistem dalam jangka panjang (Tala, 2003). Tala lebih memilih menggunakan komputasi dalam pencarian kata dasar dengan menggunakan algoritma berbasis aturan.

*Porter Stemmer* untuk bahasa Indonesia biasa disebut dengan istilah Stemmer Bahasa Indonesia Tala karena dikembangkan oleh Fadillah Z. Tala pada tahun 2003 dimana implementasi metode ini berdasarkan pada *English Porter Stemmer* yang dikembangkan oleh W.B. Frakes pada tahun 1992. Stemmer Tala sama sekali tidak menggunakan kamus sebagai acuan, seperti halnya stemmer yang lainnya (stemmer Ahmad, Vega dan Jelita).

Dalam algoritma stemmer Tala, terdapat 5 langkah utama dengan 3 langkah awal dan 2 langkah pilihan. Langkah-langkah dalam algoritma tersebut diperlihatkan seperti pada gambar 1:

Pada gambar 1 terlihat beberapa langkah '*removal*'. Langkah *removal* tersebut disusun berdasarkan aturan yang tertera pada tabel 1 sampai dengan tabel 5 dibawah ini :



Gambar 1. Proses Stemming Algoritma Tala (Tala, 2003)

Tabel 1 Inflectional particles (Tala, 2003)

Suffix	Replace	Additional Condition	Examples
kah	NULL	NULL	Bukukah → buku
lah	NULL	NULL	Adalah → ada
pun	NULL	NULL	Bukupun → buku

Tabel 2. Inflectional possessive pronouns (Tala, 2003)

Suffix	Replace	Additional Condition	Examples
ku	NULL	NULL	Bukuku → buku
mu	NULL	NULL	Bukumu → buku
nya	NULL	NULL	Bukunya → buku

Tabel 3. First order of derivational prefixes (Tala, 2003)

Suffix	Replace	Additional Condition	Examples
meng	NULL	NULL	Mengukur → ukur
meny	S	V...	Menyapu → sapu
men	NULL	NULL	Menduga → duga
meny	NULL	NULL	Menuduh → uduh
mem	P	V...	Memilah → pilah
mem	NULL	NULL	Membaca → baca
me	NULL	NULL	Merusak → rusak

peng	NULL	NULL	Pengukur → ukur
peny	S	V...	Penyapu → sapu
pend	NULL	NULL	Penduga → duga
penud	NULL	NULL	Penuduh → uduh

Tabel 4. Second order of derivational prefixes (Tala, 2003)

Suffix	Replace	Additional Condition	Examples
Ber	NULL	NULL	Berlari → lari
Bel	NULL	ajar	Belajar → ajar
Be	NULL	K* er...	Bekerja → kerja
Per	NULL	NULL	Perjelas → jelas
Pel	NULL	ajar	Pelajar → ajar
Pe	NULL	NULL	Pekerja → kerja

Tabel 5. Derivational suffixes (Tala, 2003)

Suffix	Replace	Additional Condition	Examples
Kan	NULL	Prefix ∉ {ke, peng}	Tarikkan → tarik;(men g)ambilkan → ambil
An	NULL	Prefix ∉ {di, meng, ter}	Makanan → makan;(per )janjian → janji
I	NULL	V K...c1c1, c1≠i,c2≠i and Prefix ∉ {ber, ke, peng}	Tandai → tanda;(men )dapati→d apat;Pantai → panta

#### 1.4 Pengertian Naïve Bayes Classifier

Naïve Bayes merupakan *classifier* sederhana yang didasarkan pada penerapan teorema Bayes. Kelebihan teknik ini adalah mampu mengklasifikasikan dokumen dengan tepat serta mudah dalam pengimplementasiannya Thabtah, (2009).

Sedangkan menurut Olson dan Delen, (2008), menjelaskan *Naïve Bayes* untuk setiap kelas keputusan, menghitung probabilitas dengan syarat bahwa kelas keputusan adalah benar, mengingat vektor informasi obyek. Algoritma



*Naïve Bayes* ini mengasumsikan bahwa atribut obyek adalah independen. Probabilitas yang terlibat dalam memproduksi perkiraan akhir dihitung sebagai jumlah frekuensi dari "master" tabel keputusan. Metode *Naïve Bayes* atau *Naïve Bayes Classifier* (NBC) adalah salah satu metode yang digunakan untuk klasifikasi teks. NBC menggunakan teori probabilitas sebagai dasar teori.

Dalam bukunya, Han, J. dan Kamber, M. , (2001), menyatakan: "*Bayesian classifiers* mempunyai tingkat kecepatan dan akurasi yang tinggi ketika diaplikasikan dalam database yang besar".

### 1.5 Algoritma Naïve Bayes Classifier

Pada metode *Naïve Bayes*, sebuah dokumen (teks) direpresentasikan sebagai kumpulan kata-kata (*bag of words*), dimana tiap-tiap kata dalam dokumen tersebut diasumsikan tidak bergantung satu sama lain Schneider , (2005).

Algoritma *Naive Bayes Classifier* merupakan algoritma yang digunakan untuk mencari nilai probabilitas tertinggi untuk mengklasifikasi data uji pada kategori yang paling tepat (Feldman, R & Sanger, J. 2007). Dalam penelitian ini yang menjadi data uji adalah dokumen buku dalam bentuk fisik dan bentuk file. Ada dua tahap pada klasifikasi dokumen. Tahap pertama adalah pelatihan terhadap dokumen yang sudah diketahui jenisnya. Sedangkan tahap kedua adalah proses klasifikasi untuk dokumen yang belum diketahui jenisnya.

Dasar dari teorema *Naïve Bayes* yang dipakai dalam pemrograman adalah rumus bayes yaitu sebagai berikut (Han and Kamber, 2006) :

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (1)$$

dimana  $P(H|X)$  merupakan probabilitas  $H$  di dalam  $X$  atau dengan bahasa lain  $P(H|X)$  adalah persentase banyaknya  $H$  di dalam  $X$ ,  $P(X|H)$  merupakan probabilitas  $X$  di dalam  $H$ ,  $P(H)$  merupakan probabilitas prior dari  $H$  dan  $P(X)$  merupakan probabilitas prior dari  $X$ .

Dalam algoritma *naïve bayes classifier* setiap dokumen direpresentasikan dengan pasangan atribut " $x_1, x_2, x_3, \dots, x_n$ " dimana  $x_1$  adalah kata pertama,  $x_2$  adalah kata kedua dan seterusnya.

Sedangkan  $V$  adalah himpunan kategori dokumen. Pada saat klasifikasi algoritma akan mencari probabilitas tertinggi dari semua kategori dokumen yang diujikan ( $V_{MAP}$ ), dimana

persamaannya adalah sebagai berikut (Lin, S. 2008) :

$$V_{MAP} = \underset{V_j \in V}{argmax} (P(x_1, x_2, x_3, \dots, x_n | V_j) P(V_j)) \quad (2)$$

Persamaan diatas dapat disederhanakan menjadi berikut (Feldman, R & Sanger, J. 2007) :

$$V_{MAP} = \underset{V_j \in V}{argmax} \prod_{i=1}^n (P(x_i | V_j) P(V_j)) \quad (3)$$

Keterangan :

$V_j$  = Kategori buku  $j = 1, 2, 3, \dots, n$ .

$P(x_i | V_j)$  = Probabilitas  $x_i$  pada kategori  $V_j$

$P(V_j)$  = Probabilitas dari  $V_j$

Proses klasifikasi dilakukan dengan membuat model probabilistik dari dokumen *training*, yaitu dengan menghitung nilai  $p(W_k | C)$ . untuk  $W_{kj}$  diskrit dengan  $W_{kj} \in V = \{V_1, V_2, V_3, \dots, V_m\}$  maka  $p(W_k | C)$  dicari untuk seluruh kemungkinan nilai  $W_{kj}$  dan didapatkan dengan melakukan perhitungan berikut (Mitchell, 2005):

$$p(W_k = W_{kj} | c) = \frac{D_b(W_k = W_{kj}, c)}{D_b(c)} \quad (4)$$

$$p(c) = \frac{D_b(c)}{|D|} \quad (5)$$

dengan  $D_b(W_k = W_{kj}, C)$  adalah fungsi yang mengembalikan jumlah dokumen  $b$  pada kategori  $c$  yang memiliki nilai kata  $W_k = W_{kj}$ ,  $D_b(C)$  adalah fungsi yang mengembalikan jumlah dokumen  $b$  yang memiliki kategori  $c$  dan  $|D|$  adalah jumlah seluruh dokumen *training*.

Persamaan  $D_b(W_k = W_{kj}, C)$  seringkali dikombinasikan dengan *Laplacian Smoothing* untuk mencegah persamaan mendapatkan nilai 0, yang bisa mengganggu hasil klasifikasi secara keseluruhan. Oleh karena itu, persamaan  $D_b(W_k = W_{kj}, C)$  dapat dituliskan sebagai berikut (Mitchell, 2005) :

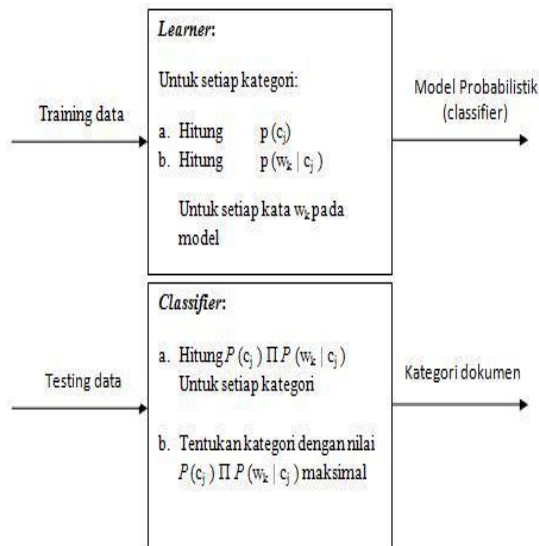
$$p(w_k = w_{kj} | c) = \frac{D_b(w_k = w_{kj}, c) + 1}{D_b(c) + |V|} \quad (6)$$

dengan  $|V|$  merupakan jumlah kemungkinan nilai dari  $W_{kj}$ .

### 1.6 Tahapan Proses Klasifikasi Naïve Bayes

Dalam metode *Naïve Bayes Classifier* terdapat dua tahap utama yakni tahap *Learning* (Pembelajaran) dan tahap *Classify*

(Pengklasifikasian). Alur tahapan klasifikasi dengan *Naïve Bayes* diperlihatkan pada gambar 2 dibawah ini :

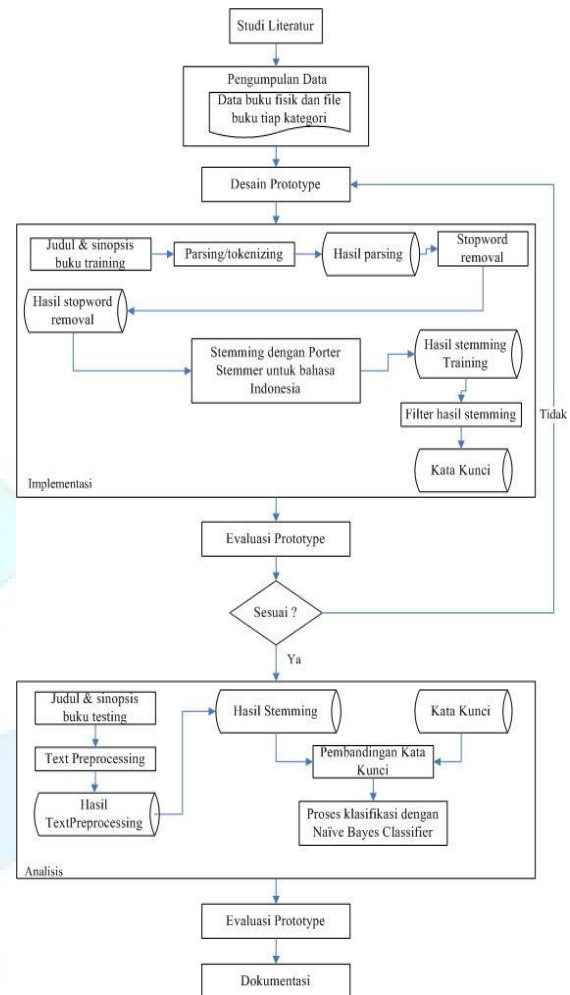


Gambar 2. Proses Klasifikasi dengan *Naïve Bayes* (Mitchell, 2005)

Gambar 2 diatas merupakan alur proses klasifikasi *Naïve Bayes Classifier* yang dimulai dari tahap *Learning* yaitu membentuk kata dasar/*Vocabulary*, menghitung probabilitas pada setiap kategori  $P(v_j)$  dan menentukan frekuensi setiap kata ( $w_k$ ) pada setiap kategori  $P(w_k|v_j)$ . Selanjutnya tahap *Classifier* dimana dalam tahap ini dilakukan proses menghitung  $P(v_j)\prod P(w_k | v_j)$  untuk setiap kategori dan menentukan kategori dengan nilai  $P(v_j)\prod P(w_k | v_j)$  maksimal.

### 1.7 Alur Penelitian

Alur penelitian adalah tahapan proses yang dilakukan oleh peneliti dalam penelitian tentang klasifikasi jenis buku berdasarkan judul dan sinopsis. Adapun alur penelitian yang dilakukan oleh peneliti, diperlihatkan pada gambar 3 dibawah ini :



Gambar 3. Alur Penelitian

Gambar 3 diatas menjelaskan tahapan proses penelitian yang dijalankan dalam penelitian ini. Tahapan proses dimulai dari studi literatur dimana peneliti mengumpulkan bahan-bahan pustaka yang mendukung penelitian, mempelajari teori-teori *Naïve Bayes Classifier* dan *porter stemmer* bahasa Indonesia.

Proses selanjutnya adalah pengumpulan data yang akan dijadikan sebagai bahan uji coba. Data yang dikumpulkan berupa data primer dimana peneliti berinteraksi langsung dengan petugas perpustakaan dan mengajukan pertanyaan tentang proses awal klasifikasi buku. Selain data primer, didapatkan pula data sekunder untuk mendukung data primer.

Data sekunder berupa pengumpulan bahan-bahan pustaka berupa buku fisik dan data buku dalam bentuk file. Buku fisik diambil berasal dari 9 kategori dimana masing-masing kategori diambil 100 buah buku. Dari 100 buah buku tiap kategori tersebut akan dibagi menjadi data *training* dan *testing* untuk dilakukan uji coba dalam beberapa skenario untuk mengukur tingkat akurasi *Naïve Bayes Classifier*. Akurasi

hasil klasifikasi ditentukan dengan persamaan (7) dibawah ini :

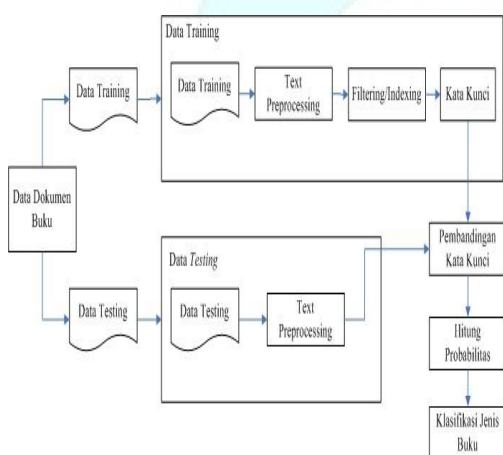
$$Akurasi = \frac{\text{Jumlah kategori benar}}{\text{Jumlah dokumen uji}} \times 100\% \quad (7)$$

## 2. HASIL DAN PEMBAHASAN

Pada bagian hasil dan pembahasan ini, akan diuraikan langkah-langkah analisis dan skenario uji coba dari data dokumen yang telah dikumpulkan sebagai berikut :

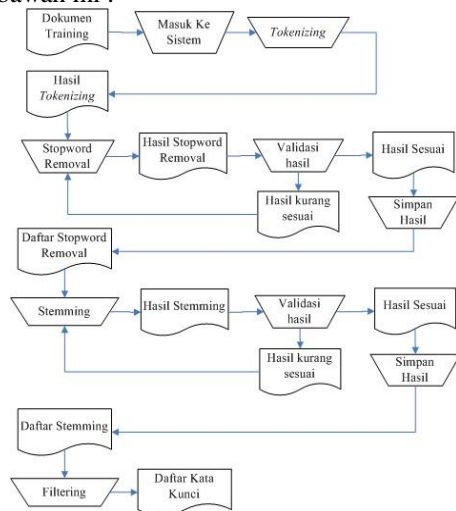
### 2.1 Analisis dan Rancangan Sistem

Secara umum, jalannya proses klasifikasi dengan *Naïve Bayes Classifier* yang dilakukan dalam penelitian ini diperlihatkan pada gambar 4 dibawah ini :



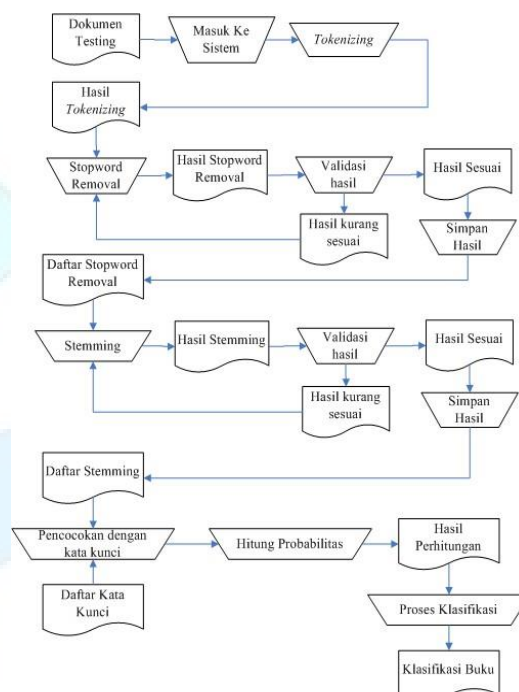
Gambar 4. Gambaran Umum Proses Klasifikasi Naïve Bayes

Dapat dilihat dalam gambar 4 tersebut bahwa dalam metode *Naïve Bayes Classifier* terdapat 2 tahapan penting yaitu tahap pembelajaran (*training*) dan klasifikasi (*classifier/testing*). Alur tahap pembelajaran yang dilakukan dalam penelitian ini diperlihatkan pada gambar 5 dibawah ini :



Gambar 5. Bagan tahap pembelajaran/Training

Gambar 5 diatas menunjukkan bagaimana jalannya proses pembelajaran yang dilakukan terhadap dokumen data *training*. Proses pembelajaran dari tersebut diakhiri dengan *filtering* dari hasil *stemming* untuk dijadikan sebagai kata kunci dari tiap kategori. Tahap selanjutnya yaitu klasifikasi (*classifier*). Dalam penelitian ini, tahap klasifikasi diperlihatkan seperti pada gambar 6 dibawah ini:



Gambar 6. Bagan tahap klasifikasi/classify

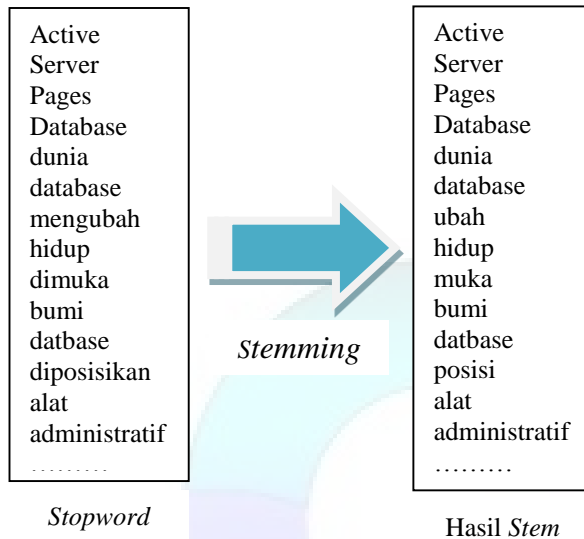
Gambar 6 diatas menunjukkan proses uji coba terhadap data dokumen buku *testing*. Dalam proses ini, terdapat tahap pencocokan kata kunci tiap kategori buku *training* dengan hasil *stem* yang dilakukan terhadap data *testing*. Hasil pencocokan nanti akan dihitung nilai probabilitas kemunculannya sehingga akan diketahui jenis buku baru dari data *testing* tersebut.

### 2.2 Skenario Uji Coba

Setelah dilakukan analisis dan rancangan sistem, maka selanjutnya dilakukan uji coba proses klasifikasi dengan beberapa skenario uji coba. Pada bagian ini, hklasifikasi buku baru didapatkan dari hasil pencocokan kata kunci data *training* dengan hasil *stem* pada data *testing*.

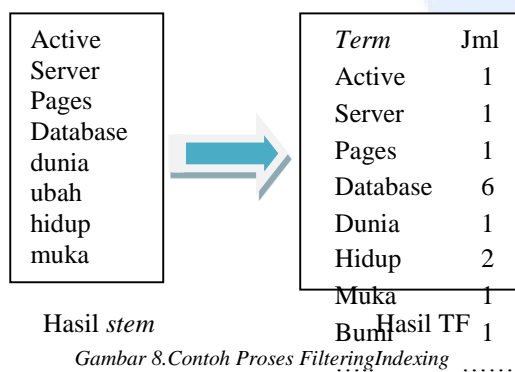
Proses *Stemming* dilakukan dengan menggunakan algoritma *Porter Stemmer for Bahasa Indonesia* dimana fungsi dari algoritma ini antara lain (Tala, 2003) dengan mengembalikan kata menjadi kata dasarnya, akan sangat mengurangi macam kata yang perlu diperiksa dan dibandingkan dengan tabel Stopword maupun tabel *Keyword*. Hal ini akan mempercepat proses perbandingan kata serta

mengurangi isi tabel *Stopword* dan tabel *Keyword*. Ilustrasi dari proses *stemming* dengan *porter stemmer for bahasa Indonesia* ini diperlihatkan pada gambar 7 dibawah ini :



Gambar 7. Contoh proses stemming dengan porter stemmer

Dari hasil *stemming* seperti diperlihatkan pada gambar 7 diatas, maka akan dilakukan pencocokan dengan kata kunci yang diperoleh dari proses *filtering/indexing* hasil *stemming* data *training*. Adapun contoh proses *filtering/indexing* diperlihatkan pada gambar 8 dibawah ini :



Gambar 8. Contoh Proses FilteringIndexing .....

Hasil pencocokan antara daftar kata kunci seperti pada gambar 8 diatas dengan hasil *stemming* nanti akan dihitung nilai probabilitas kemunculannya dengan menggunakan persamaan (4) dan (5). Skenario pertama dilakukan dengan menggunakan 100 buah buku tiap kategori, dimana dari 100 buku tersebut dibagi 90 sebagai data *training* dan 10 untuk data *testing*. Hasil uji coba skenario pertama ini diperlihatkan pada tabel 6 dibawah ini :

Tabel 6. Tabel hasil uji coba klasifikasi skenario 1

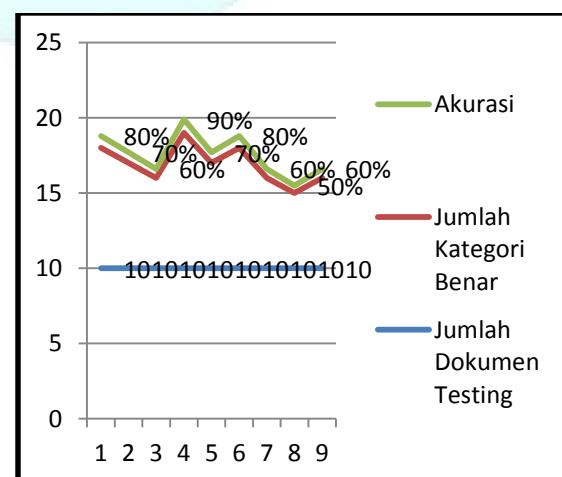
Kategori Buku	Dok. Training	Dok. Testing	Jml. Sesuai
Database	90	10	8
Ms. Office	90	10	7
Pemrograman Web	90	10	6
Delphi	90	10	9
Desain Grafis	90	10	7
Java	90	10	8
Internet	90	10	6
SPSS	90	10	5
3DsMax	90	10	6

Dalam tabel 6 terlihat jumlah buku yang sesuai dengan klasifikasi. Untuk menghitung tingkat akurasi, dipergunakan persamaan (7) dan diperoleh nilai akurasi seperti pada tabel 7 dibawah ini :

Tabel 7. Tabel nilai akurasi uji coba skenario 1

Kategori Buku	Dok. Training	Dok. Testing	Akurasi
Database	90	10	80%
Ms. Office	90	10	70%
Pemrograman Web	90	10	60%
Delphi	90	10	90%
Desain Grafis	90	10	70%
Java	90	10	80%
Internet	90	10	60%
SPSS	90	10	50%
3DsMax	90	10	60%

Grafik akurasi klasifikasi pada skenario pertama diperlihatkan seperti pada gambar 9 dibawah ini:



Gambar 9. Grafik Akurasi Uji Coba Skenario Pertama

Setelah uji coba skenario pertama dijalankan, selanjutnya dilakukan uji coba skenario kedua. Pada skenario kedua ini, porsi dokumen *training* dan dokumn *testing* diubah-ubah tidak memiliki porsi tetap seperti pada skenario pertama. Uji

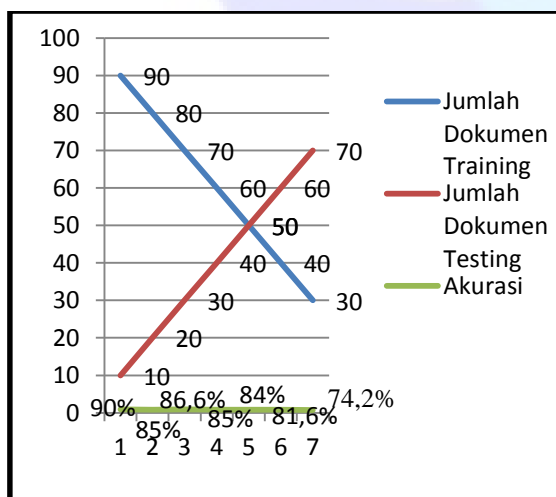


coba dilakukan terhadap 100 dokumen tiap kategori. Tabel 8 dibawah ini memperlihatkan contoh hasil uji coba skenario kedua.

Tabel 8. Tabel nilai akurasi uji coba skenario 2

Jumlah Dokumen Training	Jumlah Dokumen Testing	Akurasi
90	10	90 %
80	20	85 %
70	30	86,6 %
60	40	85 %
50	50	84 %
40	60	81,6 %
30	70	74,2 %

Dari tabel 8 diatas dapat terbaca bahwa akurasi mencapai nilai prosentase tinggi dengan semakin banyaknya penggunaan dokumen sebagai dokumn *training* dan sebaliknya akurasi akan menurun dengan semakin sedikitnya penggunaan dokumen *training*. Grafik akurasi uji coba skenario kedua diperlihatkan pada gambar 10 dibawah ini :



Gambar 10. Grafik Akurasi Uji Coba Scenarion Kedua

## 2. IMPLEMENTASI

Pada bagian ini, rancangan sistem dan interface mulai di implementasikan kedalam sistem yang baru. Implementasi dimulai dari bagaimana data buku dimasukkan kedalam sistem. Gambar 11 dibawah ini menunjukkan form yang digunakan untuk memasukkan data buku.

Gambar 11. Form input data buku

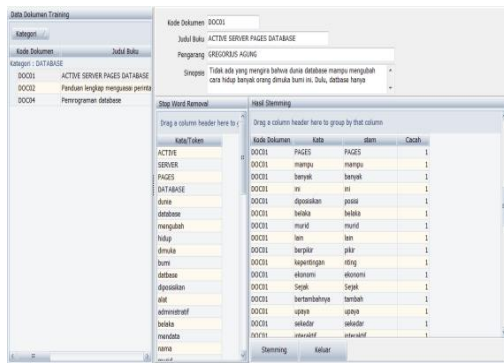
Setelah data buku dimasukkan, berikutnya dilakukan proses *parsing/tokenizing*. Gambar 12 dibawah ini menunjukkan form yang digunakan untuk melakukan proses *parsing/tokenizing*.

Gambar 12. Form untuk proses parsing/tokenizing

Dari proses *parsing/tokenizing*, kemudian dilanjutkan dengan proses *stopword removal* yakni menghilangkan kata-kata yang tidak bermakna. Gambar 9 dibawah ini menunjukkan form yang digunakan untuk melakukan proses *stopword removal*.

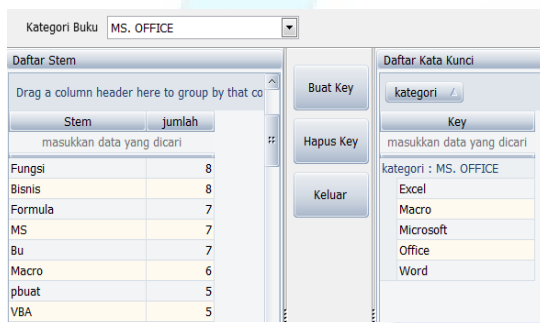
Gambar 13. Form untuk proses stopwords removal

Proses berikutnya, hasil *stopword removal* kemudian di *stem*. Tujuannya adalah selain untuk menghilangkan tanda baca dan simbol-simbol, juga berfungsi untuk membentuk kata dasar. Proses ini menggunakan aturan dari algoritma Tala. Gambar 14 dibawah ini menunjukkan form yang digunakan untuk melakukan proses *stemming* algoritma Tala.



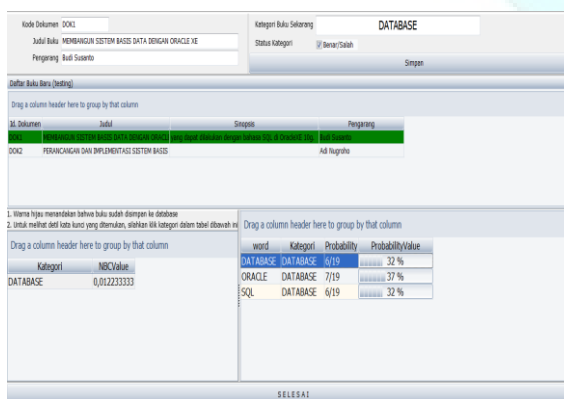
Gambar 14. Form untuk proses stemming

Tahap selanjutnya adalah pembentukan kata kunci. Kata kunci ini didapatkan secara otomatis dengan mencocokkan daftar tabel kata kunci dengan daftar kata hasil *stemming* data *training*. Gambar 15 dibawah ini menunjukkan form yang digunakan untuk proses pembentukan kata kunci.



Gambar 15. Form untuk proses pembentukan kata kunci

Setelah semua proses dilalui, maka implementasi terakhir adalah menentukan kategori/jenis buku baru. Gambar 16 dibawah ini memperlihatkan proses penentuan kategori buku baru dengan perhitungan Naïve Bayes Classifier.



Gambar 16. Form untuk proses klasifikasi NBC

Dari gambar 16 diatas dapat dilihat proses penentuan jenis buku baru berdasarkan pada perhitungan probabilitas dengan Naïve Bayes Classifier. Hasil akhir pada gambar 16, buku yang berjudul “Membangun Sistem Database Dengan Oracle” termasuk jenis buku Database dengan nilai probabilitas seperti yang tertera di gambar.

### 3. SIMPULAN DAN SARAN

Berdasarkan pada rumusan masalah dan hasil penelitian, maka dapat ditarik kesimpulan sebagai berikut :

1. Metode *Naïve Bayes Classifier* dan *Porter Stemmer* dapat diimplementasikan dalam sebuah *prototype* aplikasi klasifikasi jenis buku. Aplikasi ini dibangun dan dilengkapi dengan aturan-aturan sesuai dengan tahap-tahap dalam *Naïve Bayes Classifier* dan aturan-aturan *stemming* menurut *porter stemmer* untuk bahasa Indonesia (*Stemmer Tala*).
2. *Porter Stemmer* untuk bahasa Indonesia memiliki pengaruh yang signifikan terhadap proses klasifikasi. Hal ini dikarenakan algoritma *stemmer* ini dapat digunakan untuk membentuk kata dasar yang sebenarnya sehingga memudahkan dan mempercepat dalam proses perbandingan antara kata kunci dengan teks dokumen uji.
3. Pada hasil uji coba skenario kedua dapat disimpulkan bahwa tingginya tingkat akurasi penggunaan dokumen sebagai data *training* dan sebaliknya akurasi akan menurun dengan semakin sedikitnya penggunaan data *training*.

Untuk penelitian selanjutnya yang membahas tentang penelitian yang serupa, maka dapat diberikan saran-saran sebagai berikut :

1. Penggunaan algoritma *porter stemmer* lebih memudahkan pengguna karena tidak dibutuhkan adanya kamus kata dasar akan tetapi pada pelaksanaan eksekusinya, masih banyak dijumpai hasil *stemming* yang kurang sesuai. Diharapkan nanti, untuk penelitian selanjutnya masalah kurang maksimalnya hasil *stemming* dengan *porter stemmer* ini dapat diatasi.
2. Penentuan kata kunci masih dilakukan dengan proses pemilihan *term* yang terbanyak dan sesuai dengan kategori, sehingga masih dimungkinkan dijumpai banyak kekeliruan. Harapan untuk penelitian selanjutnya, dikembangkan sebuah teknik untuk penentuan kata kunci yang sesuai.
3. Data yang digunakan dalam penelitian ini masih kurang banyak, sehingga belum bisa maksimal dalam menentukan keakuratan *Naïve Bayes*. Untuk penelitian selanjutnya agar ditambah data uji cobanya sehingga benar-benar dihasilkan nilai akurasi yang maksimal.

### 4. DAFTAR RUJUKAN

- Darujati, C; Gumelar, B.A, 2012. *Pemanfaatan Teknik Supervised Untuk Klasifikasi Teks Bahasa Indonesia*, JURNAL LINK, 16(1), ISSN 1858 – 4667, Februari.
- Eryono, K., 1993, *Penentuan Tajuk dan Bentuk Tajuk Nama Orang*. Jakarta: Universitas Terbuka.
- Feldman, R dan Sanger, J. 2007. *The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press.
- Frakes, W. B. dan Baeza, R., 1992. *Information Retrieval Data Structure and Algorithms*, New Jersey: Prentice-Hal.
- Hamzah , A, 2012. *Klasifikasi Teks Dengan Naïve Bayes Classifier (NBC) Untuk Pengelompokan Teks Berita Dan Abstract Akademis*, Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Periode III, ISSN:1979-911X
- Hadi W., Thabtah F., Hawari, AL.S., Ababneh J., 2008. *Naive Bayesian and KNearest Neighbour to Categorize Arabic Text Data*, In proceedings of the European Simulation and Modeling Conference, Le Havre, France.
- Han, J., Kamber, M., 2006. *Data mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann Publishers Inc. ISBN:1-55860-489-8.
- Kurniawan B; Effendi S; Sitompul, O.S, 2012. *Klasifikasi Konten Berita Dengan Metode Text Mining*, *Jurnal Dunia Teknologi Informasi*, 1(1), pp.14-19.
- Lin, S. 2008. A document classification and retrieval system for R&D in semiconductor industry-A hybrid approach *Expert System* 18, 2:4753-4764.
- Lorena Br, G.S, Trinanda, P.R, 2011, *Teknik Data Mining Menggunakan Metode Bayes Classifier Untuk Optimalisasi Pencarian Pada Aplikasi Perpustakaan*, *Jurnal Teknologi & Informasi UNIKOM*.
- Mulyanto, A., 2009. *Sistem Informasi Konsep dan Aplikasi*. Cetakan I. Yogyakarta: Pustaka Pelajar.
- Mitchell, T.M, 2005. *Machine Learning*, New York: McGraw-Hill.
- Olson, D. dan Shi, Y., 2008. *Introduction to Business Data Mining*. Jakarta: Penerbit Salemba Empat.
- Samodra, J; Sumpeno, S; Hariadi, M, 2009. *Klasifikasi Dokumen Teks Berbahasa Indonesia dengan Menggunakan Naïve Bayes*, Seminar Nasional Electrical, Informatics And It's Educations.
- Schneider, M.K, 2005. *Techniques for Improving the Performance of Naïve Bayes for Text Classification*, Proceedings of CICLing, pp.682-693.
- Sedarmayanti ; Hidayat, S., 2002. *Metodologi Penelitian*, Bandung: Mandar Maju.
- Sutarno, N.S., 2006. *Perpustakaan dan Masyarakat*, Bogor: Sagung Seto.
- Supriyanto; Sujjada, A; Widyayanti S, 2011. *Otomasi Katalog Perpustakaan Melalui Klasifikasi Kategori Buku Dengan Menggunakan Metode Naïve Bayes*, *Jurnal Antartika*, 1(1).
- Tala, F.Z, 2003, *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*, thesis, Institute for Logic Language and Computation Universiteit van Amsterdam TheNetherlands.
- Vega, V.B. dan Bressan S., 2004. *Stemming Indonesian without a dictionary*, Ganome Institute dan National University of Singapore.
- Walpole, E. R., Myers, R. H., 1995. *Ilmu Peluang dan Statistika untuk Insinyur dan Ilmuan*, Edisi ke-4. Bandung, ITB.