JURUSAN TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA

USULAN TUGAS AKHIR

1. IDENTITAS PENGUSUL

NAMA : Abdur Rozaq
NRP : 5107 100 701
DOSEN WALI : Yudhi Purwananto, S.Kom. M.Kom.

2. JUDUL TUGAS AKHIR

"Klasifikasi Dokumen Teks Berbahasa Arab Menggunakan Algoritma Naïve Bayes"

3. ABSTRAK

Berkembangnya teknologi informasi mengakibatkan semakin meningkatnya ketersediaan penyampaian dan penyimpanan informasi melalui internet, dimana internet menjadi media publikasi yang sangat populer. Banyaknya informasi digital yang tidak terstruktur sebagai akibat dari perkembangan teknologi informasi, sehingga dibutuhkan suatu cara pengorganisasian atau pengelompokan informasi untuk kemudahan pengaksesannya. Oleh karena itu kategorisasi teks secara otomatis merupakan salah satu solusi untuk masalah tersebut karena dengan signifikan dapat mereduksi biaya kategorisasi secara manual.

Klasifikasi teks menggolongkan data sesuai dengan informasi yang terdapat pada dokumen tersebut. Hal ini mengefisienkan waktu pengumpulan informasi tentang topik yang ingin dipelajari karena semua informasi sudah dikelompokkan berdasarkan topik pembahasan.

Klasifikasi dokumen teks berbahasa arab berbeda dengan bahasa-bahasa lain pada umumnya. Bahasa arab termasuk bahasa sematik dengan sistem penulisan dari kanan ke

---

kiri. Bahasa Arab *modern* mempunyai tiga vokal yaitu *a*, *i* dan *u*. Maka dari itu penggunaan teorema Naïve Bayes bisa diterapkan tanpa harus mengetahui struktur dari bahasa arab itu sendiri.

Klasifikasi dengan metode Naïve Bayes akan diimplementasikan dalam aplikasi ini. Metode Naive Bayes berdasarkan pada teorema Bayes yang melakukan perhitungan probabilitas kategori. Dengan memilih nilai terbesar dari perhitungan probabilitas kategori untuk menentukan kelas suatu dokumen.

4. PENDAHULUAN

4.1 Latar Belakang Masalah

Seiring berkembangnya teknologi akhir-akhir ini semakin meningkatkan ketersediaan informasi dan penyampaiannya melalui internet. Tersimpannya jumlah data yang sangat besar menjadikan proses pengolahan data dan informasi kembali ke masalah utama yaitu *indexing* yang bagus dan kesimpulan isi dokumen. Banyak metode statistik yang berguna untuk kategorisasi informasi sehingga akan lebih mudah dalam proses temu kembali.

Regression Models, Nearest Network Classifier, Decision Tree, Naïve Bayes Classifier adalah beberapa contoh metode statistik yang mengaplikasikan kategorisasi informasi. Dan pada tugas akhir ini diterapkan metode Naïve Bayes yang akan diimplementasikan untuk proses klasifikasi dokumen teks berbahasa arab.

Bahasa arab (**اللغة العربية**) merupakan bahasa semitik, yang terdiri dari 28 huruf dan fitur dasarnya adalah sebagian besar kata dibentuk dengan suatu pola yang dapat dianalisa sampai ke *root*. Pengecualian dari aturan ini adalah kata benda umum dan partikel. Bahasa arab sangat inflektif dengan 85% dari kata-kata bahasa arab berasal dari *root tri-lateral*. Kata benda dan kata kerja berasal dari sekitar 10.000 *root*. Bahasa arab memiliki tiga gender yaitu feminim, maskulin dan neutral, bisa bersifat tunggal, ganda (merepresentasikan dua hal) dan plural.

Sebagai contoh kata **مدرسة** (sekolah) dibentuk dari *root* **درس** (belajar). Awalan dan akhiran dapat ditambahkan ke dalam kata yang telah dibentuk dari *root.* Yang membuat bahasa arab termasuk rumit untuk diproses adalah kata benda dan kata kerja mempunyai awalan. Awalan **ال** akan ditambahkan untuk menunjukkan kata itu sebagai kata benda dan masih banyak lagi konjungsi dan preposisi yang ditambahkan sebagai

awalan pada kata benda dan kata kerja. Hal ini dapat menghalangi pengambilan varian morfologi kata-kata bahasa arab.

Pemilihan metode Naive Bayes ini sesuai percobaan yang pernah dilakukan oleh W. Hadi dkk [3] tentang perbandingan hasil klasifikasi dokumen teks berbahasa arab dengan menggunakan algoritma Naive Bayes dan *K-Nearest Neighbor* yang menunjukkan bahwa algoritma Naive Bayes menghasilkan nilai perhitungan *precision, recall* dan F1 yang lebih tinggi daripada penggunaan algoritma *K-Nearest Neighbor* yang berbasis pada koefisien *Cosine.*

Implementasi metode Naïve Bayes pada aplikasi ini dilakukan untuk proses kategorisasi dokumen teks berbahasa arab dari kitab-kitab yang terdapat di aplikasi tersebut. Aplikasi ini dibuat dengan dasar agar lebih memudahkan pengguna dalam pembelajaran suatu pembahasan dari kitab-kitab yang ada pada aplikasi karena sudah terkategorikan sesuai isi yang terkandung di dalamnya.

4.2 Perumusan Masalah

Ada beberapa permasalahan dalam pembuatan tugas akhir ini adalah sebagai berikut :

1. Bagaimana cara menerapkan preprocessing (*stemming* dan pembuangan *stop word*) dalam proses klasifikasi dokumen teks berbahasa arab?
2. Bagaimana pengimplementasian metode Naive Bayes dalam proses klasifikasi dokumen teks berbahasa arab?
3. Bagaimana perbandingan nilai perhitungan evaluasi antara algoritma Naive bayes dengan algoritma *K-Nearest Neighbor*?

4.3 Batasan Masalah

Dari permasalahan-permasalahan di atas, terdapat beberapa batasan dalam tugas akhir ini :

1. Aplikasi dibangun menggunakan bahasa pemrograman JAVA dengan *library* GWT *(Google Web Toolkit)* dan *database server* MySQL.
2. Dokumen-dokumen yang digunakan sebagai data adalah kitab-kitab dengan pembahasan fiqih yang diambil dari website http://www.shamela.ws/
3. Daftar kata dasar dan *stop word* diambil dari Dr. Sheren Khoja.

4.4 Tujuan Tugas Akhir

Tujuan tugas akhir ini adalah sebagai berikut:

Membuat aplikasi pengklasifikasi dokumen teks berbahasa arab menggunakan metode Naïve Bayes.

4.4 Manfaat Tugas Akhir

Penerapan metode Naïve Bayes dalam aplikasi ini akan memudahkan pengguna dalam pencarian dokumen pada kategori tertentu. Hal ini disebabkan dokumen-dokumen yang ada terlebih dahulu telah dikelompokkan ke dalam beberapa kategori. Sehingga pengguna bisa dengan mudah mencari dokumen dengan pembahasan yang telah disediakan pada aplikasi.

5. TINJAUAN PUSTAKA

Tinjauan pustaka yang digunakan dalam pembuatan tugas akhir ini berasal dari *paper* yang berjudul *" Naïve Bayesian Based on Chi Square to Categorize Arabic Data"* yang ditulis oleh Fadi Thabtah, Mohammad Ali H. Eljinini Mannam Zamzeer dan Wa'el Musa Hadi [1].

Pada *paper* ini dijelaskan tentang metode Naive bayes untuk klasifikasi dokumen teks berbahasa arab. Naïve Bayes merupakan salah satu metode *machine learning* yang menggunakan perhitungan probabilitas. Konsep dasar yang digunakan oleh Naïve Bayes adalah Teorema Bayes, yaitu melakukan klasifikasi dengan menghitung nilai probabilitas $p(C = c_i \mid D = d_j)$, yaitu probabilitas kategori c_i jika diketahui dokumen d_j. Klasifikasi dilakukan untuk menentukan kategori $c \varepsilon C$ dari suatu dokumen $d \varepsilon D$ di mana $C = \{c_1, c_2, c_3, \ldots, c_i\}$ dan $D = \{ d_1, d_2, d_3, \ldots, d_i \}$. Penentuan dari kategori sebuah dokumen dilakukan dengan mencari nilai maksimum dari $p(C = c_i \mid D = d_j)$ pada P = $\{p(C = c_i \mid D = d_j) \mid c \varepsilon C$ dan $d \varepsilon D\}$. Nilai probabilitas $p(C = c_i \mid D = d_j)$ dapat dihitung dengan persamaan :

$$p(C = c_i \mid D = d_j) = \frac{P(C = c_i \cap D = d_j)}{P(D = d_{jj})}$$

$$= \frac{P(D = d_j \mid C = c_i)}{P(D = d_j)}$$

dengan $P(D = d_j \mid C = c_i)$ merupakan nilai probabilitas dari kemunculan dokumen d_j jika diketahui dokumen tersebut berkategori c_i , $p(C = ci)$ adalah nilai probabilitas kemunculan kategori c_i, dan $p(D = d_j)$ adalah nilai probabilitas kemunculan dokumen d_j.

Agar dokumen-dokumen teks berbahasa arab dapat diproses dengan metode Naive Bayes perlu dilakukan *preprocessing*. Dalam tahapan *preprocessing* terdapat beberapa langkah yaitu:

a. Hapus tanda baca, diakritik (tanda pengenal) dan karakter-karakter yang bukan termasuk huruf arab.
b. Hapus kata-kata yang termasuk *stop word*. Daftar kata-kata yang termasuk *stop word* ini diambil dari Khoja *Stemmer* [4].
c. Pengembalian kata-kata ke bentuk dasar *(stemming).*

Dalam sistem temu kembali informasi *stemming* digunakan untuk mereduksi variasi kata yang terbentuk dan mengembalikan ke kata dasar. Metode *stemming* ini dilakukan dengan membuang prefiks dan sufiks agar proses pengembalian ke kata dasar bisa lebih efektif. Setelah itu kata tersebut akan dicocokkan ke *dictionary* kata dasar arab.

Metode Naïve Bayes menganggap sebuah dokumen sebagai kumpulan dari kata-kata yang menyusun dokumen tersebut dan tidak memperhatikan urutan kemunculan kata pada dokumen tersebut. Sehingga perhitungan probabilitas $p(D = d_j \mid C = c_i)$ dapat dianggap sebagai hasil perkalian dari probabilitas kemunculan kata-kata pada dokumen d_j

Untuk mengetahui kualitas dan akurasi hasil dari klasifikasi tersebut digunakan penghitungan *precision,recall* dan perhitungan F1. Untuk menghitung *precision* digunakan rumus:

$$\text{Precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

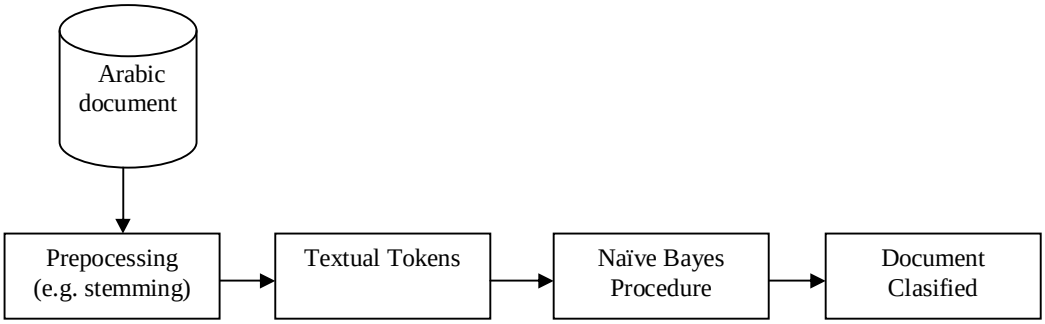
Sedangkan untuk menghitung *recall* :

$$\text{Precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

Dan untuk menghitung F1 :

$$F1 = \frac{2*Precision*Recall}{Precision+Recall}$$

Alur berjalannya sistem ini dapat dilihat seperti pada Gambar 1 yang secara umum menunjukkan bagaimana aplikasi pengklasifikasi dokumen teks berbahasa arab ini bekerja.



Gambar 1. Diagram alur sistem

6. METODE PENELITIAN

Pengerjaan tugas akhir ini terdiri atas beberapa tahapan, yaitu :

1. Studi Literatur

Studi literatur dilakukan untuk mencari pemahaman mengenai hal-hal yang berkaitan dengan proses klasifikasi dokumen teks berbahasa arab. Terdapat beberapa hal yang perlu untuk dipahami pada proses ini, antara lain:

a. *Preprocessing*

Prepocessing merupakan proses *document cleaning* agar dokumen-dokumen yang dijadikan sebagai data siap untuk diproses. *Preprocessing* sendiri memiliki beberapa tahapan, diantaranya penghapusan tanda baca, tanda pengenal, karakter-karakter yang bukan termasuk huruf *hijaiyah*, dan proses *stemming.*

Contoh *Stemming* :

الْفَصْلُ ⟶ فصل ⟵ prefiks

رَمَضَانَ ⟶ رمض ⟵ sufiks

الْمُسْلِمَيْنِ ⟶ سلم ⟵ konfiks

Contoh perhitungan klasifikasi :

b. Naive Bayes

Teorema Naive Bayes yaitu melakukan klasifikasi dengan menghitung nilai probabilitas dokumen dan kategori. Pemberian kategori dari sebuah dokumen dilakukan dengan memilih nilai *c* yang memiliki nilai probabilitas terbesar.

c. Evaluasi Hasil

Perhitungan evaluasi terhadap output yang dihasilkan oleh sistem dilakukan dengan menggunakan perhitungan nilai *precision* dan *recall* serta perhitungan nilai F1.

2. Perancangan Perangkat Lunak

Secara umum terdapat tiga tahapan sebelum menghasilkan output dokumen yang telah terklasifikasi. Tiga tahap tersebut diantaranya adalah *stemming, textual token* dan proses Naive Bayes itu sendiri.

Seperti pada proses temu kembali informasi pada umumnya, *stemming* menjadikan daftar kata pada dokumen kembali ke bentuk dasar, sehingga semua dokumen akan kembali ke bentuk dasar dari kata-kata yang terdapat dalam dokumen tersebut. Pada dokumen teks berbahasa arab, peoses *textual token* berfungsi menghilangkan karakter-karakter yang tidak termasuk dalam huruf *hijaiyah*. Sehingga dokumen yang ada benar-benar hanya terdiri dari kata-kata murni bahasa arab. Proses yang paling menentukan dalam klasifikasi dokumen ini adalah proses perhitungan Naive Bayes. Proses ini menghitung nilai probabilitas dokumen dan kategori. Pemberian kategori dari sebuah dokumen dilakukan dengan memilih nilai *c* yang memiliki nilai probabilitas terbesar.

Dengan kebutuhan yang telah dipelajari dalam proses perancangan sistem di atas, dibuatlah desain antar muka pengguna dari aplikasi tersebut. Desain antar muka pengguna sebagai jembatan bagi user agar semakin mudah dalam pemakain sistem yang akan dibuat.

Pada desain antar muka perangkat lunak ini akan menampilkan hasil klasifikasi dokumen berdasarkan kelas yang sudah ditentukan sebelumnya. Sehingga dengan dikelompokkannya dokumen-dokumen tersebut akan mempermudah user dalam pembelajaran maupun pencarian pembahasan tertentu dalam sebuah dokumen secara manual.

3. Implementasi

Pada tahap implementasi ini aplikasi sudah mulai dibuat secara menyeluruh berdasar pada desain sistem yang sudah dirancang. Bahasa yang digunakan untuk membangun aplikasi ini adalah JAVA dengan menggunakan *library* GWT *(Google Web Toolkit)* dan database yang digunakan adalah MySQL. Aplikasi ini dibangun sebagai aplikasi berbasis web (*web based application).*

4. Uji Coba dan Evaluasi

Setelah implementasi selesai dilakukan, akan dilakukan tahap uji coba dan evaluasi dengan melakukan skenario berdasarkan alur yang telah dibuat. Hasil dari uji coba ini akan dianalisa menggunakan penghitungan *precision, recall* dan perhitungan F1. Pada dasarnya, nilai *precision* dan *recall* berada pada rentang 0-1. Oleh karena itu, suatu sistem dinilai baik ketika dapat menghasilkan nilai *precision* dan *recall* yang mendekati 1.

Hasil ujicoba dan evaluasi akan dibandingkan dengan penggunaan algoritma *K-Nearest Neighbor* yang telah dilakukan pada percobaan sebelumnya.

5. Penyusunan Buku Tugas Akhir

Pada tahap ini dilakukan untuk pembuatan laporan dari semua dasar teori dan metode yang digunakan serta hasil – hasil yang diperoleh selama pengerjaan tugas akhir. Laporan tugas akhir ini akan dibagi menjadi beberapa bab sebagai berikut:

a. Bab I, Pendahuluan, berisi latar belakang, permasalahan, tujuan, batasan permasalahan, metodologi, dan sistematika penulisan.
b. Bab II, Landasan Teori, akan diulas dasar ilmu yang mendukung pembahasan tugas akhir ini.

c. Bab III, Desain Aplikasi, akan dijelaskan desain pembuatan aplikasi klasifikasi dokumen teks berbahasa arab ini.
d. Bab IV, Implementasi dari aplikasi yang telah dibuat, akan dilakukan pembuatan aplikasi yang dibangun dengan komponen-komponen yang telah ada yang sesuai dengan permasalahan dan batasannya yang telah dijabarkan pada bab pertama.
e. Bab V, Uji Coba dan Analisa Hasil, akan dilakukan uji coba berdasarkan parameter-parameter yang ditetapkan, dan kemudian dilakukan analisa terhadap hasil uji coba tersebut.
f. Bab VI, Penutup, berisi kesimpulan yang dapat diambil dari pengerjaan tugas akhir ini beserta saran untuk pengembangan selanjutnya.

6. JADWAL KEGIATAN

Tugas akhir ini diharapkan bisa dikerjakan menurut jadwal sebagai berikut:

No	Tahapan	Bulan															
		Maret				April				Mei				Juni			
1	Studi Literatur																
	a. Preprocessing																
	b. Metode Naïve Bayes																
	c. Evaluasi Hasil																
2	Perancangan Perangkat Lunak																
	a. Perancangan Sistem																
	b. Perancangan Antar Muka Pengguna																
3	Implementasi Perangkat Lunak																
4	Uji Coba dan Evaluasi																
5	Penyusunan Buku Tugas Akhir																

7. DAFTAR PUSTAKA

[1] Eljinini Mohammad Ali H., Hadi Wa'el Musa, Thabtah Fadi, Zamzeer Mannam, 2009. *"Naïve Bayesian Based on Chi Square to Categorize Arabic Data"*. *Communication of the IBIMA Volume 10, 2009 ISSN : 1943-7765*

[2] Khoja, S., Garside, R., Knowles, G., 1999, "Stemming Arabic Text", *Lancaster, UK, Computing Departement, Lancarter University*

[3] Hadi W., Thabtah F., ALHawari S., Ababneh J."Naive Bayesian and K-Nearest Neighbour to Categorize Arabic Text Data", *In proceedings of the European Simulation and Modeling Conference, Le Havre, France, 2008.*

[4] Khoja, S., 2001, "APT: Arabic Partofspeech Tagger," *Proc. of the Student Workshop at NAACL.*

LEMBAR PENGESAHAN

Surabaya, 15 Maret 2010

Mengetahui/Menyetujui

Dosen pembimbing I

Dr. Agus Zainal Arifin, S.Kom., M.Kom.

NIP. 197208091995121001

Dosen pembimbing II

Diana Purwitasari, S.Kom., M.Sc.

NIP. 197804102003122001