Jurnal CoreIT, Vol.1, No.2, Desember 2015
ISSN: 2460-738X (Cetak)

# Pengelompokan Dokumen Menggunakan *Winnowing Fingerprint* dengan Metode *K-Nearest Neighbour*

**Suwanto Sanjaya[1], Ersad Alfarisy Absar[2]**
[1,2]Teknik Informatika, UIN Sultan Syarif Kasim Riau
Jl. H.R. Soebrantas no. 155 KM. 18 Simpang Baru, Pekanbaru 28293
suwantosanjaya@uin-suska.ac.id[1], rsdalfa@gmail.com[2]

***Abstrak* – Text mining dapat didefinisikan sebagai suatu proses menggali informasi oleh seorang user yang berinteraksi dengan sekumpulan dokumen menggunakan tools analisis yang merupakan komponen-komponen dalam data mining. Dalam text mining dikenal beberapa metode untuk klasifikasi teks, salah satunya adalah K-Nearest Neightbour (KNN). KNN adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Pada penelitian ini akan dilakukan klasifikasi terhadap dokumen teks menggunakan metode KNN berdasarkan winnowing fingerprint. Winnowing adalah algoritma yang biasa digunakan untuk mendeteksi kesamaan isi suatu dokumen teks dengan cara memecah kalimat yang ada pada dokumen teks menjadi beberapa karakter sepanjang k-grams dan menghasilkan output berupa kumpulan nilai hash yang disebut fingerprint. Penelitian ini mencoba untuk menjadikan fingerprint sebagai ciri suatu dokumen teks lalu mengelompokkan dokumen teks berdasarkan ciri tersebut. Proses klasifikasi diawali dengan mengumpulkan dokumen latih yang akan dijadikan sebagai acuan dalam pengelompokan dokumen. Dokumen latih tersebut diproses dengan metode winnowing untuk mendapatkan ciri dari dokumen tersebut. Dokumen uji yang ingin dikelompokkan juga harus melewati proses winnowing, setelah fingerprint didapat maka dilanjutkan dengan proses klasifikasi menggunakan metode KNN. Dari hasil pengujian terhadap 10 dokumen uji didapat nilai akurasi pengelompokan 80%.**

***Kata Kunci – Fingerprint, K-Nearest Neightbour, Klasifikasi, Text Mining, Winnowing.***

## PENDAHULUAN

### A. Latar Belakang

Pesatnya perkembangan teknologi informasi menyebabkan banyak informasi tidak lagi disimpan dalam bentuk berkas-berkas dokumen yang diletakkan di dalam lemari melainkan informasi disimpan dalam bentuk *softcopy* berupa dokumen digital. Hal ini mempermudah dalam penyimpanan dan dalam memperbanyak suatu dokumen.

Saat ini, jumlah dan keanekaragaman dokumen teks terus bertambah sehingga menyebabkan penumpukan dokumen. Dokumen yang tersebar dan tidak terkoordinasi dengan baik akan menyulitkan *user* dalam mendapatkan informasi yang diinginkan. Sebagai contoh, *user* ingin mencari materi kuliah dan jurnal yang berhubungan dengan tema penelitian dibidang komputer. Sehingga *user* melakukan pencarian berdasarkan nama dokumen yang diduga memiliki hubungan dengan tema tugas akhir tersebut. Setelah melakukan pencarian, beberapa dokumen yang didapat tidak sesuai dengan yang diharapkan. Beberapa dokumen tersebut berisi hal yang tidak memiliki hubungan dengan tema penelitian yang dicari. Salah satu solusi yang dapat digunakan untuk mengatasi masalah ini adalah dengan menggunakan sistem yang mampu mengklasifikasikan dokumen teks berdasarkan kesamaan isi dokumen. Teknik klasifikasi adalah suatu proses untuk mengelompokkan sejumlah data ke dalam kelas-kelas tertentu yang sudah diberikan berdasarkan kesamaan sifat dan pola yang terdapat dalam data-data tersebut[1]. Menggunakan teknik klasifikasi ini, dokumen dapat dikelompokkan berdasarkan kesamaan isi yang terdapat dalam dokumen tersebut dengan membandingkan dengan kelompok-kelompok dokumen yang telah ada.

Penelitian yang berhubungan tentang pengelompokan dokumen sebelumnya telah diteliti I Wayan Surya Priantara dari Institut Sebelas Maret pada tahun 2011. Pada tulisannya Wayan menggunakan algoritma *k-mean++* sebagai metode pengelompokan dokumen dalam bahasa inggris. Dalam penelitiannya tersebut dijelaskan bahwa algoritma *winnowing* telah memenuhi kebutuhan sebuah algoritma pendeteksian kesamaan dokumen yaitu, dalam melakukan pencocokan terhadap dokumen tidak terpengaruh oleh spasi, jenis huruf, tanda baca dan karakter lainnya[2]. Hal ini sering disebut

dengan *whitespace insensitivity*. Selain itu dengan mengunakan algoritma *winnowing* untuk menghasilkan *fingerprint* suatu dokumen akan mempercepat proses pengelompokan dokumen karena dokumen dikelompokkan berdasarkan *fingerprint* yang dihasilkan. Sedangkan metode *k- nearest neighbour* telah diteliti sebelumnya oleh Widia Nur Diana dari Universitas Brawijaya pada tahun 2011. Penelitian yang dilakukan berhubungan dengan pengelompokan dokumen teks berita berbahasa indonesia. Dari penelitian tersebut disimpulkan bahwa dengan penggunaan metode k-*nearest neighbour* untuk pengelompokan dapat mengurangi jumlah data latih secara efektif dan mengurangi kompleksitas perhitungan[3].

B. Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, maka dapat ditarik sebuah rumusan masalah yaitu, "Bagaimana merancang dan membangun sebuah aplikasi pengelompokan dokumen berdasarkan kesamaan isi dokumen dengan menggunakan *winnowing* sebagai penghasil *fingerprint* dan metode *k-nearest neighbour* untuk mengelompokan dokumen*"*.

C. Batasan Masalah

Agar penulisan dalam tugas akhir ini lebih terarah maka diberikan batasan masalah yaitu:

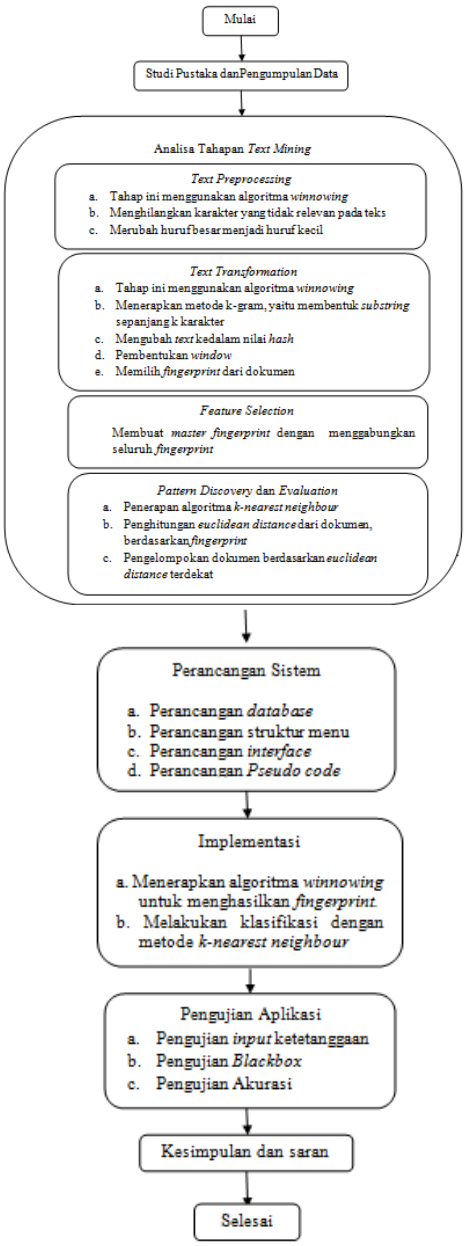
a. Dokumen yang diuji berupa dokumen teks digital yang mempunyai format plaintext (txt dan dtxt).
b. Teks yang digunakan berupa huruf latin dan berbahasa Indonesia.
c. Perhitungan *distance* menggunakan *euclidean distance.*
d. Penentuan *input* nilai ketetanggaan ditentukan oleh pengguna.
e. Koleksi dokumen latih yang digunakan dalam percobaan diambil dari artikel dan tulisan yang kelompoknya telah didefinisikan sebelumnya

D. Tujuan

Tujuan yang ingin dicapai dalam pembuatan penelitian ini, yaitu melakukan klasifikasi terhadap dokumen teks berdasarkan *fingerprint winnowing* dengan menggunakan metode *k-nearest neighbour*.

E. Metodologi Penelitian

Tahapan penelitian yang dilakukan dapat dilihat pada Gambar 1:




Gambar 1. Metodologi Penelitian

## LANDASAN TEORI

A. *Text Mining*

*Text mining* adalah *salah* satu bidang khusus dari *data mining*. Sesuai dengan buku *The Text Mining Handbook*, *text mining* dapat didefinisikan sebagai suatu proses menggali informasi oleh seorang *user* yang berinteraksi dengan sekumpulan dokumen menggunakan tools analisis yang merupakan komponen-komponen dalam *data mining* yang salah satunya adalah kategorisasi.

B. Ruang Lingkup *Text Mining*

*Text mining* merupakan suatu proses yang melibatkan beberapa area teknologi. Namun secara umum proses-proses pada *text mining* mengadopsi proses *data mining*. Bahkan beberapa teknik dalam proses *text mining* juga menggunakan teknik-teknik *data mining*. Ada empat tahap proses pokok dalam *text mining*, yaitu pemrosesan awal terhadap teks (*text preprocessing*), transformasi teks (*text transformation*), pemilihan fitur (*feature selection*), dan penemuan pola (*pattern discovery*)[4]. Adapun penjelasan tentang tahapan pemrosesan adalah sebagai berikut:

a. *Text Preprocessing*

*Text preprocessing* merupakan tahapan awal pada *text mining*. *Preprocessing* merupakan suatu proses untuk menghilangkan bagian-bagian yang tidak diperlukan atau pembersihan teks yang dilakukan untuk mengubah data data berkulitas yaitu data yang telah memenuhi persyaratan untuk dieksekusi pada sebuah algoritma. Bentuk pembersihan teks ini seperti menghilangkan spasi, tanda baca, mengubah huruf kapital menjadi huruf kecil dan menghilangkan karakter-karakter yang tidak relevan lainnya *Text preprocessing* ini juga diterapkan pada algoritma untuk mendeteksi kesamaan isi dokumen seperti algoritma *winnowing.* Ada beberapa hal yang dilakukan pada tahap *preprocessing* pada algoritma *winnowing*, seperti: (1) *Case Folding*, yaitu membuang seluruh karakter-karakter yang tidak relevan seperti: tanda baca, spasi dan juga karakter lain, sehingga nantinya hanya karakter yang berupa huruf atau angka yang akan diproses lebih lanjut. (2) *Filtering*, Tahap lanjutan dari case holding yaitu pengambilan kata penting hasil dari *case holding*, contoh kalimat "Teknik Informatika adalah salah satu jurusan yang terdapat di Fakultas Sains dan Teknologi" melalui tahapan *filtering* akan terbentuk "Teknik Informatika adalah salah satu jurusan yang terdapat difakultas sains dan teknologi".

b. *Text Transformation*

Setelah dilakukan *preprocessing* maka akan di lanjutkan dengan *text transformation.* Pada *text transformation* akan dilakukan lanjutan proses dari algoritma *winnowing.*

Algoritma *Winnowing* merupakan algoritma yang digunakan dalam mendeteksi kesamaan dokumen termasuk bagian-bagian kecil yang mirip dalam dokumen yang berjumlah banyak. Input dari algoritma ini adalah dokumen teks yang diproses sehingga menghasilkan output berupa kumpulan nilai- nilai *hash*. Kumpulan-kumpulan nilai *hash* tersebut selanjutnya disebut *fingerprint*. *Fingerprint* inilah yang dijadikan dasar pembanding antara file-file teks yang telah dimasukkan dan digunakan dalam deteksi penjiplakan[5].

Setelah tahapan *preprocessing* dilakukan, berikut langkah-langkah lanjutan dalam penerapan Algoritma *Winnowing*:

1. Metode K-gram, yaitu membentuk *substring* sepanjang *k* karakter dari sebuah *string*. Sebagai contoh: "Teknik Informatika adalah salah satu jurusan yang terdapat di Fakultas Sains dan Teknologi"

   Memotong *string* sepanjang *k.* misalnya nilai *k* = 7, dari kalimat diatas, sehingga diperoleh hasil sebagai berikut:

   ```
   Tekniki eknikin knikinf nikinfo
   ikinfor kinform informa ......
   ```

2. Perhitungan nilai *hash*, yaitu fungsi yang menerima masukan *string* yang panjangnya sembarang dan mengkonversinya menjadi *string* keluaran yang panjangnya tetap (umumnya berukuran jauh lebih kecil daripada ukuran *string* semula). Berikut persamaan dari metode *hash*:

   $$H_{(c1....ck)} = c_1 * b^{(k-1)} + c_2 * b^{(k-2)} + \cdots + c_{(k-1)} * b^k + c_k$$

   Keterangan:
   c: nilai *ascii* karakter (desimal)
   b**:** basis (bilangan prima)
   k: banyak karakter (indeks karakter)

   Contoh perhitungan dengan basis prima=2:

   $H_{(tekniki)}$ = ascii(t) * $2^{(6)}$ + ascii(e) * $2^{(5)}$+ascii(k) * $2^{(4)}$ + ascii(n) * $2^{(3)}$ + ascii(i) * $2^{(2)}$ + ascii(k) * $2^{(1)}$ + ascii(i) * $2^{(0)}$
   = 116 * 64 + 101 * 32 + 107 * 16 + 110 * 8 + 105 * 4 + 107 * 2 + 105 * 1
   = 7424 + 3232 + 1712 + 880 + 420 + 214 + 105
   = 13987

3. Membentuk *window*, yaitu membagi nilai hash yang terbentuk kedalam beberapa window. Contoh window =4:

   ```
   [13987  13236  13646  13707]
   [13236  13646  13707  13448]
   [13646  13707  13448  13565]
   ```

4. Pemilihan *fingerprint*, yaitu nilai hash minimum dari setiap window yang telah terbentuk.

   ```
   [13987 13236 13646 13707]
   [13236 13646 13707 13448]
   [13646 13707 13448 13565]
   ```

c. *Feature Selection*

*Feature selection* merupakan tahap lanjut dari pengurangan dimensi pada proses transformasi teks. Walaupun tahap sebelumnya sudah melakukan penghapusan kata-kata yang tidak deskriptif (*stopwords*), namun tidak semua kata-kata di dalam dokumen memiliki arti penting. Oleh karena itu, untuk mengurangi dimensi, pemilihan hanya dilakukan terhadap kata-kata yang relevan yang benar-benar merepresentasikan isi dari suatu dokumen. Ide dasar dari pemilihan fitur adalah menghapus kata-kata yang kemunculannya hanya di suatu dokumen tertentu.

Algoritma yang digunakan pada *text mining*, biasanya tidak hanya melakukan perhitungan pada dokumen saja, tetapi juga pada *feature*. Empat macam *feature* yang sering digunakan adalah: (1) *Character*, merupakan komponan individual, bisa huruf, angka, karakter spesial dan spasi, merupakan *block* pembangun pada level paling tinggi pembentuk semantik *feature*, seperti kata, *term* dan *concept*. (2) *Words*. (3) *Terms* merupakan *single word* dan *multiword phrase* yang terpilih secara langsung dari *corpus*. (4) *Concept*, merupakan *feature* yang di- *generate* dari sebuah dokumen secara manual, *rule-based*, atau metodologi lain.

d. *Pattern Discovery / Data Mining*

Ada beberapa teknik yang dimiliki data *mining* berdasarkan tugas yang bisa dilakukan [6], yaitu:

1. Deskripsi
   Para peneliti/analis biasanya mencoba menemukan cara untuk mendeskripsikan pola dan trend yang tersembunyi dalam data.
2. Estimasi
   Estimasi mirip dengan klasifikasi, kecuali variabel tujuan yang lebih ke arah numerik daripada kategori,
3. Prediksi
   Prediksi memiliki kemiripan dengan estimasi dan klasifikasi. Hanya saja, prediksi hasilnya menunjukkan sesuatu yang belum terjadi (mungkin terjadi dimasa depan). Misalnya, ingin diketahui prediksi harga beras tiga bulan yang akan datang.
4. Klasifikasi
   Dalam klasifikasi *variable*, tujuan bersifat kategorik. Misalnya, kita akan mengklasifikasikan pendapatan dalam tiga kelas, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah.
5. *Clustering*
   *Clustering* lebih kearah pengelompokan *record*, pengamatan, atau kasus dalam kelas yang memiliki kemiripan. Sebuah *cluster* adalah kumpulan *record* yang memiliki kemiripan satu dengan yang lain dan memiliki ketidak miripan dengan *record* dalam *cluster* yang lain.
6. Asosiasi
   Mengidentifikasi hubungan antara berbagai peristiwa yang terjadi pada satu waktu. Pendekatan asosiasi tersebut menekankan sebuah kelas masalah yang dicirikan dengan analisis keranjang pasar.

Klasifikasi adalah suatu proses untuk mengelompokkan sejumlah data ke dalam kelas-kelas tertentu yang sudah diberikan berdasarkan kesamaan sifat dan pola yang terdapat dalam data-data tersebut. Secara umum, proses klasifikasi dimulai dengan diberikannya sejumlah data yang menjadi acuan untuk membuat aturan klasifikasi data. Data-data ini biasa disebut dengan *training sets*, dari *training sets* tersebut kemudian dibuat suatu model untuk mengklasifikasikan data. Model tersebut kemudian digunakan sebagai acuan untuk mengklasifikasikan data-data yang belum diketahui kelasnya yang disebut dengan *test set* atau data latih[7].

Klasifikasi (*classification*) adalah metode data *mining* yang dapat digunakan untuk proses pencarian sekumpulan model (fungsi) yang dapat menjelaskan dan membedakan kelas-kelas data atau konsep, yang tujuannya supaya model tersebut dapat digunakan memprediksi objek kelas yang labelnya tidak diketahui atau dapat memprediksi kecenderungan data-data yang akan muncul di masa depan. Metode klasifikasi juga bertujuan untuk melakukan pemetaan data ke dalam kelas yang sudah didefinisikan sebelumnya berdasarkan pada nilai atribut data[1].

*K-Nearest Neighbour* sangat sering digunakan dalam klasifikasi dengan tujuan dari algoritma ini adalah untuk mengklasifikasi objek baru berdasarkan atribut dan *training samples*[8]. Algoritma *K-Nearest Neighbour* (K-NN atau KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. KNN termasuk algoritma *supervised learning* yaitu hasil dari query instance yang baru, diklasifikasikan berdasarkan mayoritas dari kategori pada KNN.

Dekat atau jauhnya tetangga biasanya dihitung berdasarkan jarak Euclidean dengan rumus umum sebagai berikut:

$$\mathbf{d} = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$$

Keterangan
d = Jarak
a = Data uji / testing

b = Sampel data
i = Variabel data
n = Dimensi data

e. *Confusion Matrix*

Percobaan dari penelitian dievaluasi dengan pengukuran akurasi. Pengukuran dilakukan dengan menggunakan tabel klasifikasi yang bersifat prediktif, disebut juga dengan *Confusion Matrix*[9].

Tabel 1. *Confusion Matrix* [9]

| | | Prediksi | |
|---|---|---|---|
| | | Sakit | Tidak |
| Aktual (Sebenarnya) | Sakit | TP | FN |
| | Tidak | FP | TN |

Keterangan:
TP (True Positive) : Jumlah prediksi yang benar dari data yang sakit.
FP (False Positive) : Jumlah prediksi yang salah dari data yang tidak sakit.
FN (False Negative): Jumlah prediksi yang salah dari data yang sakit.
TN (True Negative) : Jumlah prediksi yang benar dari data yang tidak sakit

$$Accuraccy = \frac{(TP+TN)}{(ATP+FP+TN+FN)}$$

## ANALISA DAN IMPLEMENTASI

A. Analisis Masalah

Jumlah dan keanekaragaman dokumen digital saat ini terus bertambah. Dokumen yang tersebar menyebabkan sulit untuk mencari dokumen yang mempunyai tema yang sama atau mempunyai kesamaan isi. Bila melakukan pencarian secara manual menggunakan nama *file* dokumen, sering didapat dokumen yang berisi hal yang tidak berhubungan. Hal itu disebabkan nama *file* dokumen belum tentu menggambarkan isi dari dokumen tersebut. Oleh karena itu, sangat penting untuk mengelola dan mengelompokkan dokumen. Dokumen yang telah dikelompokkan berdasarkan isi karakter yang ada didalamnya akan mempermudah dalam proses pencarian dokumen.

B. Analisis Kebutuhan Data

Data merupakan bahan mentah yang akan diolah untuk menghasilkan sebuah informasi. Beberapa data yang dibutuhkan pada pembuatan aplikasi klasifikasi dokumen ini adalah sebagai berikut:

a. Data Koleksi Dokumen Latih

Dokumen latih adalah data masukan yang berisi kumpulan dokumen teks yang telah mempunyai kelompok. Selanjutnya setiap dokumen latih akan diproses ke dalam tahapan *text mining* hingga menghasilkan *output* berupa *fingerprint*. *Fingerprint* ini yang akan dijadikan acuan sebagai klasifikasi dokumen.

b. Dokumen Uji

Dokumen uji adalah dokumen yang akan ditentukan kelompoknya berdasarkan *fingerprint* dokumen latih sebagai acuan.

c. Konfigurasi *Winnowing*

Konfigurasi *winnowing* adalah penentuan nilai- nilai yang digunakan dalam algoritma *winnowing*. Berdasarkan penelitian [11] konfigurasi dengan nilai *k-gram*=8, bilangan basis prima=3 dan jumlah *window*=8 menghasilkan ciri dokumen yang lebih baik, untuk itu seluruh proses algoritma *winnowing* yang ada akan menggunakan konfigurasi tersebut.

d. Data Nilai K Tetangga Terdekat

Nilai K Tetangga terdekat adalah data masukan berupa angka bernilai ganjil, yang digunakan sebagai penentuan kelompok.

C. Batasan Implementasi

Batasan implementasi pada aplikasi pengelompokan dokumen teks ini adalah sebagai berikut:

a. Bahasa pemrograman yang digunakan adalah PHP dan *database mySQL*.
b. Masukkan dokumen teks yang dapat dideteksi aplikasi ini adalah dokumen yang memiliki format *plain text* (.txt dan .dtxt).
c. Dalam proses pencarian *fingerprint*, menggunakan nilai *k-gram*=8, bilangan prima=3, dan ukuran *window* =8 [11].

D. Lingkungan Implementasi

Lingkungan implementasi adalah lingkungan dimana aplikasi ini dikembangkan. Lingkungan implementasi aplikasi *clustering* dokumen teks ini terdiri dari 2 lingkungan, yaitu lingkungan perangkat keras dan lingkungan perangkat lunak. Berikut ini merupakan spesifikasi lingkungan tersebut:

1. Perangkat Keras
   a. Processor : *Intel Core* i3 *CPU* 2.3 Hz
   b. Memory : 4 GB
   c. Harddisk : 500 GB
2. Perangkat Lunak
   a. Sistem Operasi : *Windows* 7 *Ultimate*
   b. Browser : *Google Chrome*
   c. Bahasa Pemrograman : *PHP*
   d. *Database* : *MySQL*
   e. Tools Perancangan : *Notepad*++.

E. Implementasi Antarmuka Aplikasi

Tahapan ini merupakan tahap implementasi aplikasi dari hasil analisa yang telah diperoleh dan mengimplementasikan hasil perancangan *interface* yang telah dibuat. Pada aplikasi ini memiliki empat menu, yaitu menu beranda, menu data latih, menu kelompok dan menu proses. Berikut ini merupakan implementasi aplikasi pengelompokan dokumen sesuai dengan menu yang ada pada aplikasi.

1. Implementasi *Interface* Menu Beranda

Halaman beranda merupakan halaman yang berisi tentang definisi K-*Nearest Neighbour* dan juga keterangan tambahan yang diperlukan. Berikut adalah tampilan antarmuka menu beranda:


Gambar 2. Interface Menu Beranda

2. Implementasi *Interface* Menu Pelatihan

Halaman ini berisi tampilan dokumen latih yang ada dalam *database* aplikasi. Tampilannya berupa nama *file* dan nama kelompok. Selain itu juga terdapat tombol *input* data, yang berfungsi untuk menambahkan dokumen latih kedalam aplikasi. Berikut adalah tampilan menu pelatihan:


Gambar 3. *Interface* Menu Pelatihan


Gambar 4. *Interface Input* Dokumen Latih

3. Implementasi *Interface* Proses Pengujian

Halaman ini berisikan form untuk proses pengujian dokumen. Disini terdapat informasi yang dibutuhkan untuk memproses dokumen uji, seperti *input* dokumen, id_aturan serta nilai ketetanggan yang diperlukan untuk pengelompokan dokumen uji. Berikut adalah tampilan menu pengujian:


Gambar 5. *Interface* Menu Pengujian


Gambar 6. *Interface* Urutan Dokumen Terdekat Sebanyak K


Gambar 7. *Interface* Kesimpulan Pengujian

F. Pengujian

Sistem akan dilakukan untuk mengetahui apakah sistem yang dibangun sudah sesuai dengan analisa dan tujuan dari dibangunnya sistem ini. Untuk mengetahui hasil dari sistem ini apakah sudah sesuai dengan yang diharapkan, dilakukan penghitungan *similarity* dengan *input*an dokumen yang diuji.

Adapun rencana pengujian yang dilakukan adalah sebagai berikut:

1. Pengujian *Blackbox* untuk menguji apakah program telah sesuai dengan tujuan dan fungsi dengan hasil yang didapat.
2. Menguji hasil klasifikasi dokumen uji dengan mengubah jumlah dokumen latih dan nilai kedekatan (nilai k tetangga terdekat). Koleksi dokumen latih yang digunakan berjumlah 12 dan 25 dokumen.

Berdasarkan beberapa pengujian yang telah dilakukan, diperoleh kesimpulan sebagai berikut:

a. *Fingerprint winnowing* dapat dijadikan

sebagai ciri untuk mengelompokkan suatu dokumen. Hal ini dapat terlihat dari hasil pengujian yaitu dokumen yang memiliki kelompok yang sama akan memiliki nilai kedekatan yang hampir sama pula.

b. *K-Nearest Neightbour* (KNN) sangat bergantung dengan jumlah data latih. Semakin banyak dokumen latih yang ada, maka akan semakin baik KNN dalam melakukan pengelompokkan. Hal ini dapat terlihat dari pengujian I dan II yaitu jumlah dokumen latih yang lebih banyak akan memperbesar keakuratan pengelompokkan. Selain itu jumlah nilai k tetangga terdekat juga sangat berpengaruh, jumlah nilai k harus disesuaikan dengan jumlah dokumen latih agar menghasilkan kelompok yang sesuai.
c. Selain kelemahan yang telah disebutkan, metode klasfikasi ini mempunyai kelemahan yakni bila dokumen uji yang tidak berhubungan dengan dokumen latih dimasukkan, maka tetap akan mendapatkan nilai dan dikelompokkan.
d. Pada pengujian akurasi didapat akurasi yaitu 80%. Adanya kelompok yang tidak relevan dipengaruhi oleh beberapa faktor yang telah disebutkan diatas.
e. Pada pengujian *blackbox* didapat hasil telah sesuai dengan keluaran yang diharapkan.

## KESIMPULAN DAN SARAN

Kesimpulan yang dapat diambil dari penelitian ini adalah sebagai berikut:

1. Pada pengelompokan dokumen uji, jumlah data latih dan parameter *K* dari KNN sangat berpengaruh terhadap hasil pengelompokan. Semakin banyak dokumen latih akan dapat menghasilkan pengelompokan yang lebih baik. Namun, semakin banyak dokumen latih menyebabkan waktu proses menjadi lebih lama disebabkan harus membandingkan dokumen latih ke semua dokumen uji yang ada.
2. Panjang karakter dalam dokumen mempengaruhi nilai *similarity*, hal ini disebabkan sistem mebaca frekuensi setiap *fingerprint* yang terbentuk.
3. Berdasarkan pengujian akurasi terhadap 10 dokumen, persentase akurasi yang didapat adalah 80%. Hal ini disebabkan ada kelompok yang tidak revelan. Kelompok yang tidak relevan dipengaruhi oleh beberapa faktor seperti: nilai *k-gram*, nilai k tetangga terdekat, dan panjang dokumen.

Pada penelitian selanjutnya dapat menggunakan algoritma lain seperti algoritma *biword winnowing* untuk ekstraksi ciri *fingerprint*, sehingga diharapkan menghasilkan ciri dokumen yang lebih baik. Selain itu, metode klasifikasinya juga dapat dimodifikasi menggunakan metode lain seperti naive bayes dan sebagainya.

## REFERENSI

[1] Han, J & Kamber, M. 2006. *Data Mining Concepts and Techniques*. San Fransisco Morgan Kaufmann Publishers.

[2] Priantara, I Wayan Surya., Diana Puspitasari., Umi Laili Yuhana. 2011. *Implementasi Deteksi Penjiplakan Dengan Algoritma Winnowing Pada Dokumen Terkelompok*. Jurusan Teknik Informatika, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh November Surabaya.

[3] Diana, Widia Nur., Achmad Ridok., M. Tanzil Furqon. 2011. *Penerapan Algoritma Improved K-Nearest Neighbors Untuk Pengkategorian Dokumen Teks Berita Berbahasa Indonesia*. Jurusan Matematika Program Studi Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Brawijaya.

[4] Even,Y. & Zohar. 2002. *Introduction to Text Mining.* Automated Learning Group National Center For Supercomputing Aplications. University of Illionis.

[5] Schleimer, Saul., Daniel S. Wilkerson, dan Alex Aiken. 2003. Winnowing : *Local Algorithms for Document Fingerprint*. San diego: In Proceedings Of The ACM SIGMOD International Conference On Management Of Data.

[6] Kusrini, & Luthfi, Emha. 2009. *Algoritma*
*Data Mining.* Yogyakarta:Penerbit Andi.

[7] Rifqi, Maharani., Shaufiah. 2011. *Analisis dan Implementasi Klasifikasi Data Mining Menggunakan Jaringan Syaraf Tiruan dan Evolution Strategis.* Institut Teknologi Telkom Bandung.

[8] Larose, Daniel T. 2005. *Discovering Knowledge in Data: An Introduction to Data Mining*. United States of America: John Wiley & Sons, Inc.

[9] Xhemali, D., Hinde, C.J. & Stone, R.G. 2009. *Naive Bayes vs Decision Trees vs Neural Networks in the Classification of Training Web Pages*. International Journal of Computer Science Issues.