Perancangan Aplikasi Pengklasifikasian Dokumen Dengan Algoritma Nearest Neighbor

Susiana[1], Riyadi J. Iskandar[2], Tony Darmanto[3]
Teknik Informatika, STMIK Widya Dharma, Pontianak
e-mail: [1]susihuang18@gmail.com, [2]riyadi@stmik-widyadharma.ac.id,[3]tony@stmik-widyadharma.ac.id

Abstract

The variation of language that using in document-composing imply a necessity of language-based classification. Therefore, the writer design a language-based document classifier application using The Nearest Neighbor Algorithm which is combined with Euclidean Distance. In classifying the document, the procedures of the research are about training the sample-document and testing the document which is going to be classified. The testing are done in documents which are written in Indonesian, English, and combination of the two languages. From this research, the writer concludes that the result of feature's utilization (which are the relative frequency of each character in the document) as a reference in Indonesian and English written document classification by using the Nearest Neighbor algorithm and Euclidean Distance is already good.

Keywords— *Document, Nearest Neighbor, Euclidean Distance*

Abstrak

Penggunaan jenis bahasa yang beragam dalam penulisan dokumen menyebabkan perlunya suatu proses klasifikasi berdasarkan jenis bahasa penulisan. Hal ini mendorong penulis merancang suatu aplikasi pengklasifikasian dokumen dengan menggunakan algoritma Nearest Neighbor yang dikombinasikan dengan jarak Euclidean. Dalam melakukan klasifikasi dokumen, penulis menggunakan prosedur yang terbagi atas pelatihan terhadap dokumen sampel dan pengujian dokumen yang akan diklasifikasi. Pengujian dilakukan terhadap sejumlah dokumen bahasa Indonesia, bahasa Inggris dan dokumen dengan kombinasi kedua bahasa tersebut. Dari penelitian yang telah dilakukan, penulis mengambil kesimpulan bahwa hasil penggunaan fitur berupa frekuensi relatif setiap huruf pada dokumen sebagai acuan klasifikasi dokumen bahasa Indonesia dan dokumen bahasa Inggris dengan menggunakan algoritma Nearest Neighbor dan jarak Euclidean sudah baik.

Kata kunci*— Dokumen, Nearest Neighbor, Jarak Euclidean*

1. PENDAHULUAN

Di masa kini, teknologi informasi memegang peranan penting dalam berbagai aspek kehidupan manusia. Perkembangan teknologi informasi meningkat seiring dengan penggunaannya oleh manusia dalam mempermudah memecahkan berbagai persoalan yang ditemui dalam kehidupan sehari-hari. Salah satu contoh pemanfaatan teknologi informasi adalah di bidang pemrosesan kata.

Pemrosesan kata berbasis komputer sejauh ini telah menghasilkan dokumen yang tak terhitung jumlahnya. Dokumen-dokumen tersebut dapat berupa informasi berita, buku, karya ilmiah, dan sebagainya. Selain keberagaman informasi yang ditampung, dokumen juga ditulis dalam jenis bahasa yang beragam. Ini dikarenakan hampir semua negara di dunia telah memanfaatkan teknologi informasi untuk melakukan pemrosesan kata.

Penggunaan jenis bahasa yang beragam dalam penulisan dokumen menyebabkan perlunya suatu proses identifikasi jenis bahasa penulisan dan kemudian dilakukan pengelompokan dokumen berdasarkan jenis bahasa penulisannya. Hal ini dapat memudahkan pencarian dan pengolahan informasi dokumen pada suatu penyedia layanan informasi misalnya perpustakaan digital yang menampung dokumen dalam jumlah yang sangat banyak.

2. METODE PENELITIAN

2.1 Bentuk Penelitian dan Teknik Pengumpulan Data

Bentuk penelitian dan teknik pengumpulan data yang digunakan adalah :

a. Rancangan Penelitian
   Dalam penelitian ini penulis menggunakan Rancangan Penelitian Deskriptif, penulis menjelaskan langkah-langkah perancangan aplikasi pengklasifikasian dokumen.
b. Metode Pengumpulan Data
   Metode pengumpulan data yang digunakan adalah metode studi literatur yaitu dengan mengumpulkan dan mempelajari literatur-literatur yang berkaitan dengan objek penelitian ini.
c. Metode Pengembangan Aplikasi
   Metode yang digunakan untuk melakukan klasifikasi dokumen adalah dengan melakukan pelatihan terhadap dokumen sampel dan pengujian dokumen yang akan diklasifikasi. Algoritma klasifikasi yang digunakan adalah algoritma Nearest Neighbor.
d. Teknik Analisis Data
   Teknik analisis data yang digunakan dalam penelitian ini adalah Unified Modelling Language (UML).
e. Teknik Perancangan Aplikasi
   Untuk perancangan aplikasi ini, penulis menggunakan Matrix Laboratory (MATLAB) versi R2013a.

2.2 Landasan Teori

Klasifikasi merupakan penyusunan kelompok atau golongan menurut kaidah atau standar yang ditetapkan. Manfaat dari klasifikasi dokumen adalah untuk memudahkan dalam pengorganisasian dokumen, terutama dokumen dalam jumlah yang besar. Pencarian akan lebih mudah dilakukan jika dokumen dalam keadaan terorganisir.

2.2.1 Dokumen

Dokumen memuat informasi yang menjadi bukti suatu hal. Dokumen adalah informasi terekam, termasuk data dalam sistem komputer, yang dibuat atau diterima oleh organisasi atau perorangan dalam transaksi kegiatan atau melakukan tindakan sebagai bukti aktivitas tersebut[1].

2.2.2 Klasifikasi

Klasifikasi berasal dari bahasa Latin yaitu *classis* yang artinya pengelompokkan benda yang sama serta memisahkan benda yang tidak sama[2]. Klasifikasi adalah pengelompokkan fakta berdasarkan atas ciri atau kriteria tertentu[3].

2.2.3 *Plain Text*

Plain *text* merupakan standar dokumen teks yang berisi rangkaian teks yang tidak terformat. *Plain teks* tidak didukung oleh format teks seperti pengaturan *style text*, baik tebal, miring, ataupun garis bawah dan pengaturan *style font*, baik jenis maupun ukuran font untuk teks-teks tertentu. *Plain text* dapat disimpan dalam beragam ekstensi seperti .log, .readme, dan .asc. Namun ekstensi yang paling sering dan umum digunakan adalah .txt. [4]

2.2.4 Algoritma Nearest Neighbor

Salah satu algoritma yang dapat digunakan untuk melakukan klasifikasi adalah algoritma *Nearest Neighbor*. Algoritma *Nearest Neighbor* adalah pendekatan untuk mencari kasus dengan menghitung kedekatan antara kasus baru dengan kasus lama, yaitu berdasarkan pada pencocokan bobot dari sejumlah fitur yang ada[5]. Algoritma *Nearest Neighbor* mengelompokkan suatu contoh dengan menetapkan kelas berdasarkan suatu contoh terdekat yang telah diketahui kelasnya, sebagai ukuran jarak [6].

Berikut adalah penjelasan cara algoritma *nearest neighbor* bekerja :

a. Ambil sebuah objek dengan fitur d (tetapi tidak diketahui kelompok keanggotaannya)
b. Hitung jarak dari objek ini terhadap setiap objek yang terdapat pada kumpulan pelatihan (yang telah diketahui kelompok keanggotaannya)
c. Perhatikan tetangga terdekat pada kumpulan pelatihan
d. Perhatikan bagaimana tetangga terdekat tersebut diklasifikasi. Ini menjadi (prediksi) klasifikasi dari objek.[7]

Fitur yang biasanya disebut juga dengan variabel atau atribut, digunakan sebagai acuan dalam melakukan klasifikasi. Fitur yang akan digunakan pada penelitian ini adalah distribusi huruf yang terdapat pada dokumen.

Penentuan kemiripan dokumen sampel (dokumen latih) dan dokumen yang akan diuji dilakukan dengan menggunakan Jarak Euclidean (*Euclidean Distance*).

2.2.5 Jarak Euclidean

Jarak digunakan untuk menentukan tingkat kesamaan (*similarity degree*) atau ketidaksamaan (*disimilarity degree*) dua vektor. Tingkat kesamaan berupa suatu nilai dan berdasarkan nilai tersebut dua vektor itu akan dikatakan mirip atau tidak. Jarak Euclidean (*Euclidean Distance*) adalah metrika yang paling sering digunakan untuk menghitung kesamaan 2 vektor[8]. Jarak Euclidean adalah suatu pengukur perbedaan yang dapat digunakan pada berbagai jenis variabel [9].

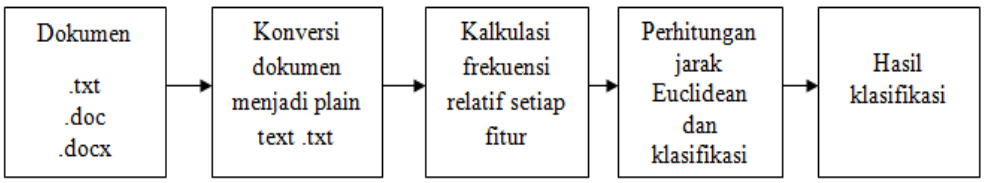
Jarak Euclidean umumnya digunakan untuk mengukur jarak antara dua hal. Jarak Euclidean antara $x = (x_1,x_2,...,x_D)$ and $y = (y_1,y_2,...,y_D)$ adalah:[10]

$$d_E(\mathbf{x},\mathbf{y}) = \| \mathbf{x} - \mathbf{y} \| = \sqrt{\sum_{i=1}^{D} (x_i - y_i)^2}$$

3. HASIL DAN PEMBAHASAN

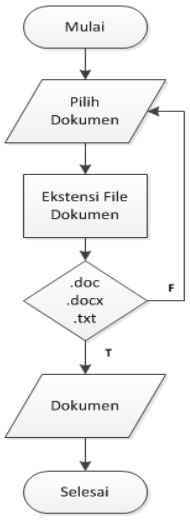
3.1 Strategi Pemecahan Masalah

Dalam merancang aplikasi pengklasifikasian dokumen penulis menggunakan strategi pemecahan masalah yang terdiri atas sejumlah langkah yang ditampilkan pada gambar berikut:



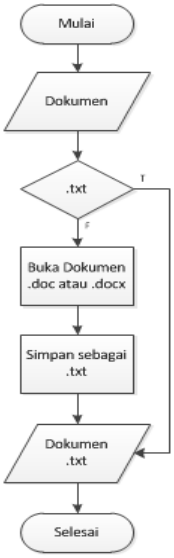
Gambar 1 Gambaran Umum Pemecahan Masalah

Langkah pertama yang dilakukan sebelum melakukan klasifikasi adalah memilih dokumen yang akan diproses. Dokumen yang dapat digunakan sebagai dokumen masukan adalah dokumen dengan ekstensi file .txt, .doc, dan .docx. *Flowchart* berikut menampilkan alur logika yang digunakan pada langkah ini:



Gambar 2 Flowchart Proses Memasukkan Dokumen

Dokumen masukan yang bukan merupakan *plain text* (berektensi .txt), yakni dokumen .doc ataupun .docx akan diubah ke dalam bentuk *plain text* terlebih dahulu agar dapat diproses lebih lanjut. Setelah dilakukan konversi, akan diperoleh *output* berupa dokumen dalam bentuk *plain text* berekstensi .txt. Alur logika untuk proses konversi dokumen masukan menjadi bentuk *plain text* ditampilkan pada *flowchart* berikut:



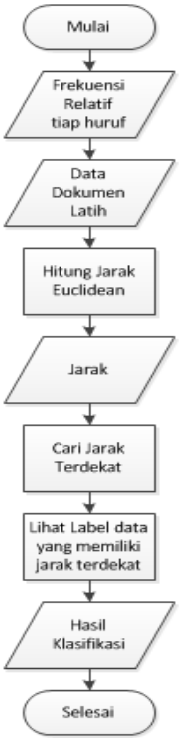
Gambar 3 Flowchart Konversi Dokumen Menjadi Plain Text

Setelah didapatkan dokumen berupa *plain text* dari tahap sebelumnya, akan dilakukan pembacaan terhadap isi dokumen dan perhitungan frekuensi relatif tiap fitur yang digunakan (frekuensi relatif huruf a-z). Sebelum menghitung jumlah frekuensi relatif tiap huruf, terlebih dahulu dilakukan penyaringan isi dokumen. Dengan menggunakan huruf sebagai fitur untuk melakukan klasifikasi, karakter bukan huruf yang terdapat pada dokumen akan dibuang dan kemudian dilakukan pengubahan semua karakter huruf pada dokumen menjadi bentuk *lower case* (huruf kecil).

Gambar 4 Flowchart Kalkulasi Frekuensi Relatif Fitur

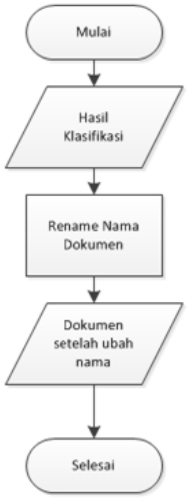
Perhitungan jarak antara setiap fitur pada data latih dan data baru akan dilakukan dengan acuan berupa hasil perhitungan dari tahap sebelumnya. Jarak dihitung dengan menggunakan rumus Jarak Euclidean. Setelah didapatkan hasil perhitungan jarak, akan dicari data latih yang memiliki jarak terdekat

terhadap data baru. Data baru akan dimasukkan dalam kelas yang sama dengan kelas yang ditempati oleh data latih yang memiliki jarak terdekat.



Gambar 5 Flowchart Perhitungan Jarak Euclidean Dan Klasifikasi

Pada tahap akhir, setelah didapatkan hasil klasifikasi berupa jenis bahasa dokumen yang diproses, nama dokumen yang telah diklasifikasi dapat diubah dengan nama baru yang mengandung keterangan jenis bahasa dokumen tersebut.



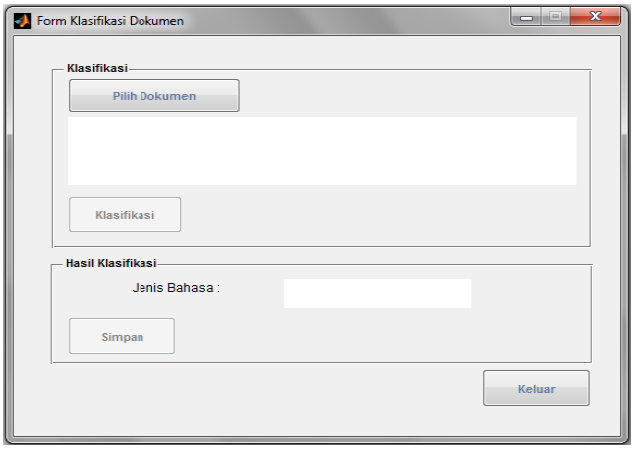
Gambar 6 Flowchart Hasil Klasifikasi

3.2 Tampilan Form Aplikasi

Form tambah data latih digunakan untuk menambah dokumen latih. Form tersebut ditampilkan pada gambar berikut ini:

Gambar 7 Form Tambah Data Latih

Untuk melakukan klasifikasi dokumen, pengguna mengakses form klasifikasi dokumen dan memilih dokumen yang ingin diklasifikasi. Pada form klasifikasi dokumen akan ditampilkan hasil klasifikasi berupa bahasa penulisan dokumen.



Gambar 8 Form Tambah Klasifikasi Dokumen

3.3 Pengujian Aplikasi

Pada tahap pengujian, digunakan sejumlah data uji yang berbeda untuk masing-masing kasus uji. Data uji yang digunakan terdiri dari dokumen .txt, .doc. dan .docx.

3.2.1 Kasus uji 1 - Menguji kemampuan mengklasifikasi dokumen berbahasa Indonesia

Pada kasus ini, pengujian dilakukan untuk mengetahui kemampuan aplikasi dalam mengklasifikasi dokumen berbahasa Indonesia. Data uji yang digunakan dan hasil pengujian ditampilkan pada tabel berikut:

Tabel 1 Data Uji dan Hasil Pengujian Kasus Uji 1

No	Nama Dokumen	Ekstensi File	Bahasa Penulisan	Hasil Klasifikasi (Bahasa)
1	2012-1-00061-IF Bab 1	.doc	Indonesia	Indonesia
2	2012-1-00061-IF Bab 2	.doc	Indonesia	Indonesia
3	I_Abs1	.doc	Indonesia	Indonesia
4	I_Abs2	.doc	Indonesia	Indonesia
5	I_Abs3	.txt	Indonesia	Indonesia
6	I_Jurnal1	.doc	Indonesia	Indonesia
7	I_Jurnal2	.doc	Indonesia	Indonesia
8	PCD-0	.docx	Indonesia	Indonesia
9	PCD-1	.docx	Indonesia	Indonesia
10	PCD-2	.docx	Indonesia	Indonesia

3.2.2 Kasus uji 2 - Menguji kemampuan mengklasifikasi dokumen berbahasa Inggris

Pengujian dilakukan untuk mengetahui kemampuan aplikasi dalam mengklasifikasi dokumen berbahasa Inggris. Data uji dan hasil pengujian ditampilkan pada tabel berikut:

Tabel 2 Data Uji dan Hasil Pengujian Kasus Uji 2

No	Nama Dokumen	Ekstensi File	Bahasa Penulisan	Hasil Klasifikasi (Bahasa)
1	ProposalBackground	.docx	Inggris	Inggris
2	Summary	.txt	Inggris	Inggris
3	Summary2	.txt	Inggris	Inggris
4	E_Jurnal1	.doc	Inggris	Inggris
5	2012-2-00437-IG Bab2001	.doc	Inggris	Inggris
6	2012-2-00437-IG Bab1001	.doc	Inggris	Inggris
7	2012-2-00409-IG Bab2001	.doc	Inggris	Inggris
8	2012-1-00134-IG Bab2001	.doc	Inggris	Inggris
9	2012-2-00409-IG Bab1001	.doc	Inggris	Inggris
10	2012-1-00134-IG Bab1001	.doc	Inggris	Inggris

3.2.3 Kasus uji 3 – Menguji kemampuan mengklasifikasi dokumen berbahasa Indonesia dan Inggris.

Pengujian dilakukan untuk mengetahui kemampuan aplikasi dalam mengklasifikasi dokumen yang dominan salah satu bahasa, bahasa Indonesia atau bahasa Inggris. Data uji dan hasil pengujian ditampilkan pada tabel di bawah ini:

Tabel 3 Data Uji dan Hasil Pengujian Kasus Uji 3

No	Nama Dokumen	Ekstensi File	Dominan Bahasa	Hasil Klasifikasi (Bahasa)
1	Jurnal1	.doc	Indonesia	Indonesia
2	Jurnal2	.doc	Indonesia	Indonesia
3	Jurnal3	.doc	Indonesia	Indonesia
4	Jurnal4	.doc	Indonesia	Indonesia
5	BabII	.docx	Indonesia	Indonesia
6	Soal1	.docx	Inggris	Inggris
7	Soal2	.docx	Inggris	Inggris
8	Soal3	.doc	Inggris	Inggris
9	Soal4	.docx	Inggris	Indonesia
10	Teori2	.txt	Inggris	Inggris

4. KESIMPULAN

a. Penggunaan fitur berupa frekuensi relatif huruf yang terdapat pada dokumen untuk mengklasifikasi dokumen berbahasa Indonesia dan Inggris menghasilkan hasil klasifikasi yang cukup baik.
b. Algoritma Nearest Neighbor yang dikombinasikan dengan rumusan jarak Euclidean dalam penentuan jarak (similarity) antar kasus dapat diterapkan dalam prosedur klasifikasi dokumen dengan baik.
c. Kemampuan aplikasi yang dirancang dalam mengklasifikasi dokumen-dokumen yang ditulis dalam bahasa Indonesia atau bahasa Inggris sudah memuaskan.
d. Kemampuan aplikasi yang dirancang dalam mengklasifikasi dokumen-dokumen yang ditulis dalam bahasa Indonesia dan bahasa Inggris (dengan catatan bahwa komposisi penggunaan kata salah satu bahasa lebih mendominasi penulisan dokumen) sudah cukup memuaskan.

5. SARAN

a. Aplikasi ini hanya dirancang untuk melakukan klasifikasi dokumen yang menggunakan bahasa Indonesia, bahasa Inggris dan gabungan kedua bahasa saja, sehingga masih terbatas dalam hal jangkauan klasifikasi. Dalam hal ini, jika aplikasi ingin dikembangkan lebih lanjut oleh pemrogram berikutnya, pengembangan program dapat difokuskan pada perluasan jangkauan bahasa yang dapat diklasifikasi.

b. Aplikasi ini hanya mendukung pengklasifikasian dokumen-dokumen yang memiliki ekstensi file .txt, .doc, dan .docx saja, sehingga jika ingin diteliti lebih lanjut dapat dilakukan pengembangan pada penambahan jangkauan file yang didukung oleh aplikasi.
c. Penggunaan aplikasi oleh pengguna pemula dapat dimulai dengan terlebih dahulu memahami petunjuk pengoperasian yang telah diuraikan, sehingga proses klasifikasi menggunakan aplikasi dapat berjalan dengan baik.
d. Dokumen yang dimasukkan dalam proses-proses yang berhubungan dengan klasifikasi oleh aplikasi haruslah dokumen yang valid (sesuai kebutuhan dan spesifikasi yang telah diuraikan) agar hasil yang diinginkan dapat diperoleh.

DAFTAR PUSTAKA

[1] Sukoco, Badri Munir. (2007). *Manajemen Administrasi Perkantoran Modern*. Erlangga. Jakarta.

[2] Darmono. (2007). *Perpustakaan Sekolah*. Grasindo. Jakarta

[3] Widjono Hs. (2007). *Bahasa Indonesia Mata Kuliah Pengembangan Kepribadian di Perguruan Tinggi*. Grasindo. Jakarta.

[4] Jubilee Enterprise. (2010). *Rahasia Manajemen File.* Elex Media Komputindo. Jakarta.

[5] Kusrini dan Emha Taufiq Luthfi. (2009). *Algoritma Data Mining*. Andi. Yogyakarta.

[6] Bruyne, Steven De. (2010). *Process Data and Classifier Models for Accessible Supervised Classification Problem Solving*. VUBPRESS. Brussels.

[7] Ledolter, Johannes. (2013). *Data Mining and Business Analytics with R*. John Wiley Sons, Inc. New Jersey.

[8] Putra, Darma. (2010). *Pengolahan Citra Digital*. Andi. Yogyakarta.

[9] Drennan, Robert D. (2009). *Statistics for Archaeologists*. Springer. New York.

[10] Barbakh, Wesam Ashour, Ying Wu, dan Colin Fyfe. (2009). *Non-Standard Parameter Adaption for Exploratory Data Analysis. Springer*. Verlan Berlin Neidelberg.