PENGENALAN TULISAN DAN EKSTRAKSI INFORMASI PADA CITRA ABSTRAK SKRIPSI MENGGUNAKAN *SUPPORT VECTOR MACHINE* DAN *RULES BASED SYSTEM*

Muhammad Fajry Hamzah[1], Galih Hermawan[2]

[1,2] Teknik Informatika – Universitas Komputer Indonesia
Jl. Dipatiukur No 112-116 Bandung, 40132
E-mail : fajryhamzah@gmail.com[1], galih.hermawan@email.unikom.ac.id[2]

ABSTRAK

Kemajuan teknologi telah banyak mempengaruhi berbagai aspek dalam kehidupan manusia, salah satunya adalah dalam pengelolaan informasi. Kebanyakan orang lebih memilih menyimpan informasi-informasinya dalam bentuk digital. Selain memiliki banyak kelebihan dibanding dengan informasi cetak, informasi digital juga akan memberi kemudahan dalam mengakses dan memanipulasi informasi tersebut.

Salah satu dokumen yang biasa diperlukan untuk diambil informasinya adalah laporan skripsi mahasiswa. Informasi yang lengkap tentang laporan skripsi mahasiswa sebenarnya bisa didapatkan hanya dari halaman abstraknya saja. Tetapi terkadang, abstrak skripsi yang tersedia hanya memiliki *hardcopy*-nya saja, sehingga pustakawan harus memasukan identitas pada suatu dokumen dengan cara mengisi data-data yang diperlukan kedalam sistem. Metode *Support Vector Machine* dapat digunakan untuk mengenali tiap karakter untuk mengubah tulisan pada citra menjadi tulisan karakter digital sehingga nantinya bisa dikenali tiap bagian pada abstrak skripsi dengan menggunakan metode *Rule Based*.

Hasil dari penelitian didapatkan hasil keakurasian sebesar 5,47% untuk pengenalan tulisan pada citra abstrak skripsi. Sedangkan untuk tingkat akurasi pengenalan karakter menggunakan SVM itu sendiri mencapai 54.30%. Rendahnya tingkat akurasi pengenalan pada citra abstrak dipengaruhi oleh proses segmentasi yang kurang mampu menyelesaikan masalah yang ada pada pemisahan karakter. Sedangkan untuk pengkategorian informasi dengan menggunakan Rule Based memiliki akurasi sebesar 99%.

Kata kunci : *Support Vector Machine*, *Rule Based*, abstrak skripsi, pengenalan citra tulisan, Ekstraksi Informasi

1. PENDAHULUAN

Kemajuan teknologi telah banyak mempengaruhi berbagai aspek dalam kehidupan manusia, salah satunya adalah dalam pengelolaan informasi. Kebanyakan orang lebih memilih menyimpan informasi-informasinya dalam bentuk digital. Selain memiliki banyak kelebihan dibanding dengan informasi cetak, informasi digital juga akan memberi kemudahan dalam mengakses dan memanipulasi informasi tersebut. Salah satu contoh dokumen yang diperlukan informasinya adalah laporan skripsi. Informasi lengkap pada laporan skripsi sebenarnya bisa didapatkan hanya dari halaman abstraknya saja seperti judul laporan, nama penulis, NIM penulis, isi ringkasan laporan dan kata kunci pada laporan, Tetapi terkadang untuk dokumen skripsi lama, abstrak skripsi yang tersedia hanya memiliki hardcopy-nya saja sehingga pustakawan harus secara manual mengubah abstrak tersebut menjadi bentuk teks digital untuk nantinya dimasukan kepada sistem [1]. Hal ini akan menjadi masalah ketika dokumen yang harus di konversikan berjumlah banyak sehingga akan memakan waktu dan tenaga. Selain itu, terjadinya *human error* ketika proses pengkonversian dan pengkategorian informasi bisa saja terjadi dikarenakan berbagai faktor. Masalah tersebut dapat ditangani dengan cara mengubah informasi yang ada pada citra abstrak skripsi menjadi softcopy dalam bentuk teks digital dan memperoleh informasi dalam dokumen tersebut secara otomatis sehingga permasalahan bisa diminimalisir.

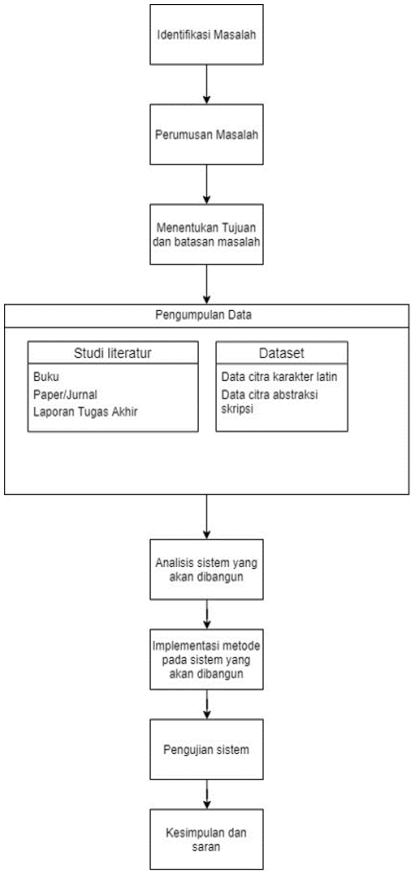
Penelitian tentang pengenalan tulisan sudah banyak dilakukan, tetapi jarang ada penelitian yang mengenali tulisan pada dokumen penuh. Ada beberapa metode yang bisa digunakan untuk pengenalan karakter tulisan pada suatu citra [2], seperti dengan menggunakan metode *Artificial Neural Network* (ANN) [3], *K-Nearest Neighbour* (KNN) dan *Support Vector Machine* (SVM) [4]. Penelitian sebelumnya mencoba membandingkan performa dari ANN dan SVM [5] dengan akurasi terbaik didapatkan dari metode SVM dengan hasil sebesar 94,43%, penelitian lain juga mencoba membandingkan metode KNN dan SVM [6] dengan SVM mendapatkan akurasi terbaik sebesar 93,13%. Jadi bisa disimpulkan bahwa metode Support Vector Machine mengklasifikasi karakter lebih baik sehingga menjadikan alasan kenapa penelitian ini akan menggunakan metode SVM untuk mengenali karakter tulisan pada citra abstrak skripsi. Selain itu

penelitian tentang ekstraksi informasi pada laporan skripsi khususnya laporan skripsi unikom sudah pernah dilakukan, dengan metode *Rules Based* hasil yang didapatkan sangat baik [1] sehingga metode yang sama akan digunakan pada penelitian ini dengan *rules* yang sedikit berbeda dikarenakan bedanya masukan yang digunakan, sehingga *rules* yang dipakai pada penelitian ini akan meminimalisir kesalahan pengekstraksian dikarenakan hasil dari pengenalan tulisan yang kurang baik.

2. ISI PENELITIAN

2.1 Metode Penelitian

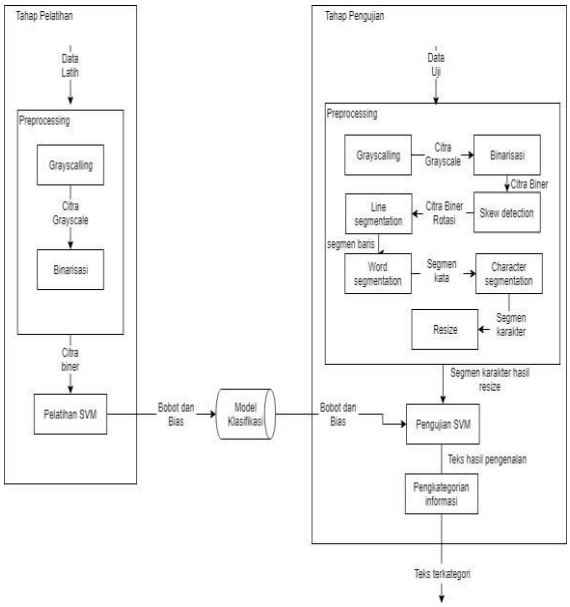
Penelitian ini memiliki 8 tahap alur kerja penelitian, diantaranya adalah identifikasi dan perumusan masalah, menentukan tujuan dan batasan, pengumpulan data, analisis sistem yang akan dibangun, implementasi sistem, pengujian dan tahap akhir membuat kesimpulan dan saran. Berikut adalah alur tahapan pada penelitian.



Gambar 1. Alur Tahapan Penelitian

2.2 Gambaran Sistem

Sistem yang dibangun dibagi dalam dua bagian, yaitu bagian pelatihan dan bagian pengujian, berikut adalah gambaran dari sistem.



Gambar 2. Gambaran Sistem

pada dua bagian tersebut terdapat proses-proses yang berbeda. Pada bagian pelatihan, data latih adalah sebuah citra tiap karakter alphabet, nomor dan karakter simbol yang sering muncul pada halaman abstrak skripsi. Pada bagian pelatihan terdapat proses *preprocessing* yang berisikan proses *grayscalling* dan binarisasi yang dilanjutkan dengan proses pelatihan model SVM yang nantinya akan disimpan dalam media penyimpanan. Bagian kedua adalah tahap pengujian, data uji adalah citra dari halaman abstrak skripsi yang bisa didapatkan dengan cara di *scan* ataupun foto. Terdapat beberapa proses pada bagian ini, diantaranya adalah proses *preprocessing* yang berisikan proses *grayscalling,* binarisasi, *skew correction*, *line segmentation, word segmentation, character segmentation* dan *resize*. Setelah proses *preprocessing* dilakukan, maka karakter yang sudah disegmentasi dan di*resize* akan dijadikan masukan pada proses pengujian SVM dengan menggunakan model SVM yang sudah dilatih pada tahap pelatihan. Setelah semua karakter dikenali pada proses pengujian SVM, proses selanjutnya adalah mengekstraksi informasi tersebut menjadi kategori-kategori yang sudah ditentukan, seperti judul skripsi, nama penulis, NIM penulis, isi abstrak dan kata kunci pada abstrak.

2.3 Data Latih

Data latih yang digunakan pada penelitian ini adalah citra karakter dengan *font times new roman* berupa alphabet,nomor dan beberapa karakter simbol yang sering muncul pada halaman abstrak dengan berbagai variasi seperti rotasi dan blur sehingga total data yang digunakan dalam pelatihan adalah 300 data. Berikut adalah contoh dari data latih yang digunakan.

Gambar 3. Data Latih

2.4 *Grayscale*

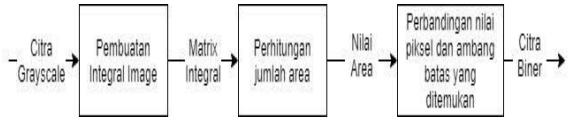
Proses *grayscale* mengubah citra yang berwarna menjadi abu-abu. Berikut adalah rumus yang digunakan dalam proses *grayscale*.

$$Y' = 0{,}299R' + 0{,}587G' + 0{,}114B' \quad (1)$$

Semua piksel pada citra akan dihitung menggunakan rumus diatas sehingga nilai citra yang tadinya mempunyai lebih dari 1 *channel* warna berubah menjadi 1 *channel* warna dengan nilai pada *range* 2^8.

2.5 Binarisasi

Binarisasi pada penelitian menggunakan metode *Bradley-Roth Adaptive Thresholding* [7]. Metode Bradley-Roth menggunakan integral image sebagai citra masukan. *Integral image* adalah hasil penjumlahan dari area yang ditentukan, teknik ini mampu mempercepat dalam mendapatkan hasil perhitungan ambang. Berikut adalah diagram blok pada proses binarisasi.

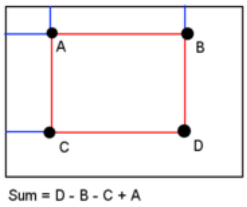


Gambar 4. Diagram Blok Binarisasi

Pembuatan *integral image* dilakukan dengan persamaan berikut.

$$I(x,y) = \sum_{\substack{x' \le x \\ y' \le y}} i(x', j') \quad (2)$$

Kemudian nilai area akan dihitung dari wilayah seperti berikut.



Gambar 5. Area acuan binarisasi

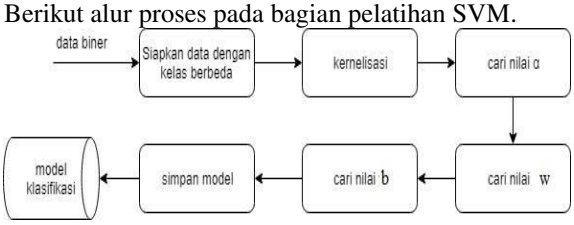
$$i(x,y) = I(D) - I(B) - I(C) + I(A) \quad (3)$$

Sehingga nantinya nilai area tersebut akan dibandingkan dengan nilai piksel dengan persentase yang sudah ditentukan, jika perhitungan memenuhi syarat maka piksel akan diubah menjadi piksel hitam. Berikut persamaan perbandingan antara nilai piksel dan intensitas warna pada area.

$$\left(nilai_{piksel} * jumlah_{pikselarea}\right) < \left(sum * \frac{100-15}{100}\right) \quad (4)$$

2.6 Pelatihan SVM

Support Vector Machine (SVM) adalah sistem pembelajaran yang menggunakan ruang hipotesis berupa fungsi – fungsi linier dalam sebuah fitur yang berdimensi tinggi [8] dan dilatih dengan menggunakan algoritma pembelajaran yang didasarkan pada teori optimasi. Pada proses training SVM dilakukan untuk menemukan vektor α, nilai w dan konstanta b untuk mendapatkan hyperplane terbaik. Dalam proses pelatihan, dibutuhkan satu set input-output data atau dalam kasus ini dibutuhkan citra karakter tulisan dan nama karakter tersebut. Berikut alur proses pada bagian pelatihan SVM.



Gambar 6. Alur Pelatihan SVM

Citra yang sudah dibinarisasi kemudian akan diubah menjadi bentuk vektor yang nantinya akan digunakan pada pelatihan ataupun pengujian SVM. Format yang digunakan adalah sebagai berikut:

[nilai_piksel_1,nilai_piksel_2,...,nilai_piksel_n]

SVM adalah salah satu algoritma yang termasuk dalam supervised learning, oleh karena itu dibutuhkan satua buah vektor lagi yang berisikan nama kelas dari vektor yang akan dilatih. Jumlah dari vektor kelas (y) adalah = N data latih yang digunakan. Pelatihan SVM membutuhkan minimal 2 data latih dengan kelas yang berbeda. Langkah pertama dalam pelatihan adalah kernelisasi vektor masukan. Kernel yang digunakan bisa berbagai macam, seperti contoh adalah kernel Linear. Berikut persamaan dari kernel linear.

$$K(x_i, x) = x_i^T x \quad (5)$$

Langkah selanjutnya adalah dengan mencari nilai w dan b tetapi sebelumnya harus mencari nilai dari alpha terlebih dahulu. Nilai dari alpha hasil pengubahan menggunakan *Lagrange Multiplier* bisa diselesaikan dengan menggunakan *Quadratic Programming* dengan menyelesaikan persamaan pada *dual form*-nya. Penyelesaian masalah *Quadratic Programming* bisa dilakukan menggunakan algoritma optimasi seperti *Sequntial Minimal Optimization*(SMO) [9]. Pertama-tama nilai alpha dicari dengan menggunakan persamaan berikut:

$$a_2^{new} = a_2 + \frac{y_2(E_1 - E_2)}{n} \quad (6)$$

Dimana nilai n dan E bisa didapatkan dari persamaan berikut:

$$n = K(x1,x1) + K(x2,x2) - 2K(x1,x2) \quad (7)$$
$$E_i = w.X + b = -1 \quad (8)$$
$$E_2 = w.X + b = 1 \quad (9)$$

Dalam SMO penyelesaian alpha dilakukan pada dua alpha dalam satu iterasi sehingga membutuhkan

sebuah batas. Apabila y1 tidak sama dengan y2 maka persamaan batas sebagai berikut:

$L = \max(0, a_2 - a_1)$ (10)

$H = \min(C, C + a_2 - a_1)$ (11)

Dan jika y1 = y2, maka persamaan batas menjadi:

$L = \max(0, a_2 + a_1 - C)$ (12)

$H = \min(C, a_2 + a_1)$ (13)

Setelah nilai alpha yang dipilih ditemukan, maka dilakukan pengecekan agar nilai alpha tidak keluar dari *boundary*.

$$\alpha_2^{\text{new,clipped}} = \begin{cases} H & \text{if} \quad \alpha_2^{\text{new}} \geq H; \\ \alpha_2^{\text{new}} & \text{if} \quad L < \alpha_2^{\text{new}} < H; \\ \text{L} & \text{if} \quad \alpha_2^{\text{new}} \leq L. \end{cases}$$

Setelah itu nilai dari alpha pertama adalah sebagai berikut:

$a_1^{new} = \text{a}_1 + s(a_2 - a^{new\,clipped}{}_2)$ (14)

Setelah seluruh nilai *support vector* (alpha) sudah teroptimisasi, maka nilai dari *w* bisa didapatkan dengan persamaan berikut.

$\text{w} = \sum_{i=1}^{n} a_i y_i x_i$ (15)

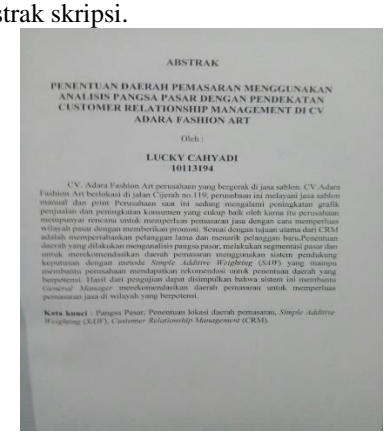
Dan nilai *b* dengan mensubstitusikan *w, x dan y* pada persamaan berikut.

$b = y - w.x$ (16)

Proses diatas akan dilakukan pada semua karakter pelatihan dengan pendekatan *One vs One* sehingga pada penelitian ini, jumlah model yang akan tercipta adalah $75(75-1)/2 = 2775$ buah model.

2.7 Data Pengujian

Pada tahap ini, citra abstrak skripsi akan digunakan, berikut adalah salah satu contoh citra dari abstrak skripsi.



ABSTRAK

PENENTUAN DAERAH PEMASARAN MENGGUNAKAN ANALISIS PANGSA PASAR DENGAN PENDEKATAN CUSTOMER RELATIONSHIP MANAGEMENT DI CV ADARA FASHION ART

Oleh :

LUCKY CAHYADI
10113194

CV. Adara Fashion Art perusahaan yang bergerak di jasa sablon. CV.Adara Fashion Art berlokasi di jalan Cijerah no.119, perusahaan ini melayani jasa sablon manual dan print Perusahaan saat ini sedang mengalami peningkatan grafik penjualan dan peningkatan konsumen yang cukup baik oleh karna itu perusahaan mempunyai rencana untuk memperluas pemasaran jasa dengan cara memperluas wilayah pasar dengan memberikan promosi. Sesuai dengan tujuan utama dari CRM adalah mempertahankan pelanggan lama dan menarik pelanggan baru,Penentuan daerah yang dilakukan menganalisis pangsa pasar, melakukan segmentasi pasar dan untuk merekomendasikan daerah pemasaran menggunakan sistem pendukung keputusan dengan metode *Simple Additive Weighting* (SAW) yang mampu membantu perusahaan mendapatkan rekomendasi untuk penentuan daerah yang berpotensi. Hasil dari pengujian dapat disimpulkan bahwa sistem ini membantu *General Manager* merekomendasikan daerah pemasaran untuk memperluas pemasaran jasa di wilayah yang berpotensi.

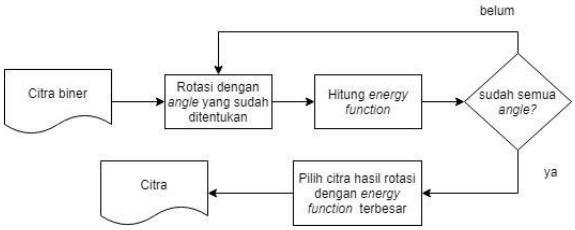
Kata kunci : Pangsa Pasar, Penentuan lokasi daerah pemasaran, *Simple Additive Weighting* (SAW), *Customer Relationship Management* (CRM).

Gambar 7. Citra Abstrak

Citra abstrak akan diubah menjadi *grayscale* dan dibinarisasi menggunakan metode yang sama pada tahap pelatihan.

2.8 *Skew Correction*

Skew correction adalah proses dimana mengubah citra dengan rotasi untuk memperbaiki jenis citra yang mempunyai kemiringan. Kemiringan citra bisa mengakibatkan pendeteksian tulisan yang ada pada citra menjadi sulit kenali. Pendeteksian dan pengoreksian tingkat kemiringan citra dilakukan dengan menggunakan metode horizontal profile projection dimana pengkoreksian akan dicoba dengan berbagai angle dan angle dengan jumlah energy function maksimal yang akan dijadikan angle citra terbaru [10]. Berikut adalah alur yang digunakan pada proses *skew correction*:



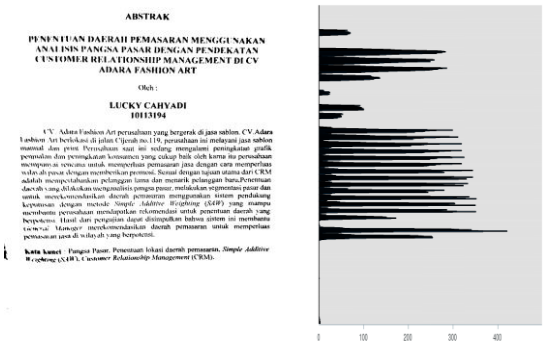
Gambar 8. Alur *Skew Correction*

Berikut adalah persamaan dalam menentukan nilai *energy function*:

$$h = \left(area_{top} - area_{bottom}\right)^2 \quad (17)$$

2.9 *Line Segmentation*

Tahap selanjutnya adalah melakukan proses Line Segmentation yang digunakan untuk memisahkan tulisan tiap baris. Metode yang digunakan pada Line Segmentation adalah metode Horizontal Profile Projection yang bekerja dengan menghitung jumlah piksel berwarna hitam pada citra secara horizontal (sumbu x) [11]. Berikut adalah contoh proyeksi dari piksel hitam pada citra abstrak:



ABSTRAK

PENENTUAN DAERAH PEMASARAN MENGGUNAKAN ANALISIS PANGSA PASAR DENGAN PENDEKATAN CUSTOMER RELATIONSHIP MANAGEMENT DI CV ADARA FASHION ART

Oleh :

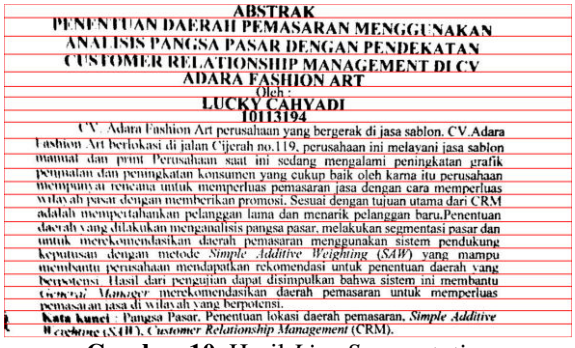
LUCKY CAHYADI
10113194

CV. Adara Fashion Art perusahaan yang bergerak di jasa sablon. CV.Adara Fashion Art berlokasi di jalan Cijerah no.119, perusahaan ini melayani jasa sablon manual dan print Perusahaan saat ini sedang mengalami peningkatan grafik penjualan dan peningkatan konsumen yang cukup baik oleh karna itu perusahaan mempunyai rencana untuk memperluas pemasaran jasa dengan cara memperluas wilayah pasar dengan memberikan promosi. Sesuai dengan tujuan utama dari CRM adalah mempertahankan pelanggan lama dan menarik pelanggan baru,Penentuan daerah yang dilakukan menganalisis pangsa pasar, melakukan segmentasi pasar dan untuk merekomendasikan daerah pemasaran menggunakan sistem pendukung keputusan dengan metode *Simple Additive Weighting* (SAW) yang mampu membantu perusahaan mendapatkan rekomendasi untuk penentuan daerah yang berpotensi. Hasil dari pengujian dapat disimpulkan bahwa sistem ini membantu *General Manager* merekomendasikan daerah pemasaran untuk memperluas pemasaran jasa di wilayah yang berpotensi.

Kata kunci : Pangsa Pasar, Penentuan lokasi daerah pemasaran, *Simple Additive Weighting* (SAW), *Customer Relationship Management* (CRM).

Gambar 9. Citra Biner dan Proyeksinya

Pemotongan citra perbaris dilakukan dengan melihat *gap* kosong yang ada pada setiap baris sehingga hasil dari *line segmentation* bisa dilihat pada gambar berikut:



ABSTRAK

PENENTUAN DAERAH PEMASARAN MENGGUNAKAN ANALISIS PANGSA PASAR DENGAN PENDEKATAN CUSTOMER RELATIONSHIP MANAGEMENT DI CV ADARA FASHION ART

Oleh :

LUCKY CAHYADI
10113194

CV. Adara Fashion Art perusahaan yang bergerak di jasa sablon. CV.Adara Fashion Art berlokasi di jalan Cijerah no.119, perusahaan ini melayani jasa sablon manual dan print Perusahaan saat ini sedang mengalami peningkatan grafik penjualan dan peningkatan konsumen yang cukup baik oleh karna itu perusahaan mempunyai rencana untuk memperluas pemasaran jasa dengan cara memperluas wilayah pasar dengan memberikan promosi. Sesuai dengan tujuan utama dari CRM adalah mempertahankan pelanggan lama dan menarik pelanggan baru,Penentuan daerah yang dilakukan menganalisis pangsa pasar, melakukan segmentasi pasar dan untuk merekomendasikan daerah pemasaran menggunakan sistem pendukung keputusan dengan metode *Simple Additive Weighting* (SAW) yang mampu membantu perusahaan mendapatkan rekomendasi untuk penentuan daerah yang berpotensi. Hasil dari pengujian dapat disimpulkan bahwa sistem ini membantu *General Manager* merekomendasikan daerah pemasaran untuk memperluas pemasaran jasa di wilayah yang berpotensi.

Kata kunci : Pangsa Pasar, Penentuan lokasi daerah pemasaran, *Simple Additive Weighting* (SAW), *Customer Relationship Management* (CRM).

Gambar 10. Hasil *Line Segmentation*

2.10 *Word Segmentation*

Setelah tiap baris ditemukan, langkah selanjutnya adalah melakukan proses word segmentation yang bekerja untuk memisahkan tiap kata pada tiap baris yang sudah ditemukan. Metode yang dilakukan untuk memisahkan tiap kata adalah dengan menggunakan metode *Vertical Profile Projection*. Metode yang dilakukan hampir sama dengan metode untuk menemukan baris hanya saja pada *vertical profile projection* perhitungan jumlah piksel dilakukan secara vertikal atau dalam sumbu y. Berikut adalah contoh proyeksi pada *word segmentation:*

Gambar 11. Proyeksi Pada Pemisahan Kata

Rentang ambang batas yang digunakan untuk pemisah kolom adalah dihitung dari 1/3 dari nilai tinggi citra baris. Nilai tersebut didapatkan dari penelitian sebelumnya [12] tetapi nilai tersebut tidak akan digunakan jika nilai tersebut melebihi nilai spasi maksimal yang ada citra , maka nilai yang dipakai adalah 80% dari nilai spasi maksimal yang ada pada suatu baris. Berikut adalah hasil dari proses *word segmentation*:

Gambar 12. Hasil *Word Segmentation*

2.11 *Character Segmentation*

Character segmentation dilakukan untuk memisahkan tiap karakter pada citra hasil segmentasi kata. Segmentasi yang dilakukan untuk tiap karakter dilakukan dengan menggunakan metode yang sama dengan segmentasi kata yaitu metode vertical profile projection yang menghitung jumlah piksel hitam pada sumbu y (vertikal). Pemisahan untuk tiap karakter, dilakukan dengan memanfaatkat gap yang ada pada tiap kata. Berikut adalah hasil dari proses *character segmentation*:

Gambar 13. Hasil *Character Segmentation*

Hasil dari pemotongan karakter nantinya akan direpresentasikan sebagai matriks seperti berikut:

1	1	1	1	1	1	1	1	0	0	1	1	1	1	1
1	1	1	1	1	1	1	1	0	0	0	1	1	1	1
1	1	1	1	1	1	1	0	0	0	0	1	1	1	1
1	1	1	1	1	1	1	0	0	0	0	1	1	1	1
1	1	1	1	1	1	0	0	0	0	0	0	1	1	1
1	1	1	1	1	1	0	0	0	0	0	0	1	1	1
1	1	1	1	1	0	0	0	0	0	0	0	0	1	1
1	1	1	1	1	0	0	0	1	0	0	0	0	1	1
1	1	1	1	1	0	0	1	1	0	0	0	0	0	1
1	1	1	1	1	0	0	1	1	1	0	0	0	0	1
1	1	1	1	0	0	1	1	1	1	0	0	0	0	1
1	1	1	1	0	0	0	0	0	0	0	0	0	0	1
1	1	1	0	0	0	0	0	0	0	0	0	0	0	1
1	1	1	0	0	1	1	0	1	1	1	0	0	0	1
1	1	0	0	0	1	1	1	1	1	1	1	0	0	0
1	1	0	0	1	1	1	1	1	1	1	1	0	0	0
1	0	0	0	1	1	1	1	1	1	1	1	0	0	0

Gambar 14. Matriks Karakter ‘A’

2.12 *Resize*

Hasil dari segmentasi karakter akan menghasilkan ukuran yang berbeda-beda. Oleh karena itu maka dibutuhkan sebuah metode untuk mengubah ukuran dari citra yang telah disegmentasi. Pengubahan ukuran dilakukan untuk menyesuaikan dengan ukuran citra yang dipakai untuk pelatihan pengklasifikasi. Metode yang digunakan untuk pengubahan ukuran adalah dengan menggunakan metode *Nearest Neighbor*. Nearest Neighbor merupakan metode interpolasi paling sederhana dan cepat dengan memindahkan ruang yang kosong dengan piksel yang berdekatan (the nearest neighboring pixel) pada saat pengecilan atau pembesaran skala gambar [13]. Berikut adalah contoh dari hasil *resize* menggunakan ukuran 15x15 px.

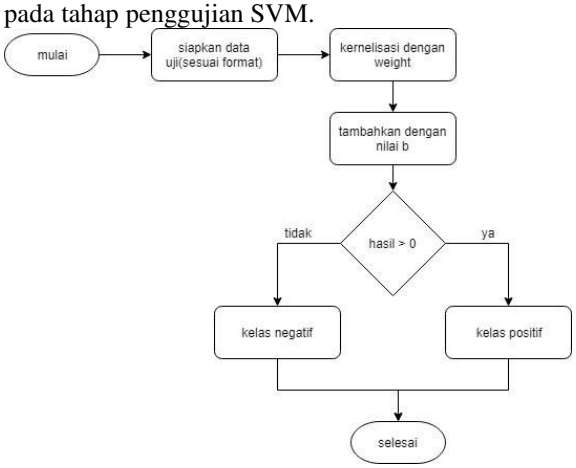
1	1	1	1	1	1	1	1	0	0	1	1	1	1	1
1	1	1	1	1	1	1	1	0	0	0	1	1	1	1
1	1	1	1	1	1	1	0	0	0	0	1	1	1	1
1	1	1	1	1	1	0	0	0	0	0	1	1	1	1
1	1	1	1	1	1	0	0	0	0	0	1	1	1	1
1	1	1	1	1	0	0	0	0	0	0	1	1	1	1
1	1	1	1	1	0	0	0	0	0	0	0	1	1	1
1	1	1	1	0	0	1	1	0	0	0	0	0	1	1
1	1	1	1	0	0	1	1	1	0	0	0	0	1	1
1	1	1	1	0	1	1	1	1	0	0	0	0	0	1
1	1	1	1	0	0	0	0	0	0	0	0	0	0	1
1	1	1	0	0	0	0	0	0	0	0	0	0	0	1
1	1	1	0	1	1	0	1	1	1	0	0	0	0	1
1	1	0	0	1	1	1	1	1	1	1	0	0	0	0
1	0	0	0	1	1	1	1	1	1	1	0	0	0	0

Gambar 15. Matriks Hasil *Resize*

2.13 Penggujian SVM

Setelah karakter tulisan telah disegmentasi, langkah selanjutnya adalah tahap klasifikasi. Pengklasifikasian dilakukan untuk mengenali karakter apa yang ada pada citra. Nilai dari citra hasil segmentasi akan diubah menjadi sebuah format vektor masukan ke model SVM (X) yang sudah di training. Format vektor yang digunakan terdiri dari NxN tergantung pada ukuran citra pada saat model dilatih dan nilai diambil dari tiap piksel citra yang

sudah disegmentasi sebelumnya. Berikut adalah alur pada tahap penggujian SVM.



Gambar 16. Alur pada tahap penggujian SVM

Vektor masukan terlebih dahulu di kernelisasi dengan fungsi kernel yang sama pada tahap pelatihan, kemudian nilai w dan b yang sudah didapatkan pada saat pelatihan akan disubstitusikan dengan persamaan berikut.

$$w.x_i + b \quad (18)$$

Suatu vektor akan dinyatakan berada pada kelas positif jika hasil prediksi dari persamaan (18) > 0. Perhitungan diatas akan dilakukan pada semua model SVM yang sudah dilatih dan keputusan akhir akan diambil dari maksimal vote model SVM yang berhasil mengklasifikasi masukan tersebut sebagai suatu kelas.

2.14 Pengkategorian Informasi

Tahap terakhir adalah pengambilan bagian-bagian informasi yang terkandung pada hasil klasifikasi tersebut menggunakan *Rules Based*. Informasi yang diambil adalah judul, nama pembuat, NIM pembuat, isi abstrak dan keyword yang ada. Untuk mengambil informasi yang ada pada, akan dibuatkan sebuah rules set yang berguna sebagai aturan yang dipakai untuk mendapatkan informasi yang ada pada hasil pengenalan tulisan. Aturan yang dipakai adalah berupa kata kunci dan posisi baris. Berikut adalah aturan yang dipakai untuk mengidentifikasi informasi pada abstrak skripsi.

Tabel 1. Tabel *Rules*

Bagian	Aturan
Judul	1. Set baris awal ke baris kedua 2. Set baris akhir ke baris sebelum baris yang hanya mempunyai kata "oleh" atau "by" atau baris yang mempunyai 2 kata dan kata kedua hanya memiliki 1 karakter 3. Jika sampai baris 6 tidak ditemukan, maka baris akhir di set ke baris sebelum baris dengan jumlah karakter paling sedikit 4. Ambil tulisan dari baris awal sampai baris akhir yang sudah di set
Nama	1. Set baris nama ke dua baris setelah baris akhir judul 2. Ambil tulisan dari baris tersebut
NIM	1. Set baris NIM ke baris setelah nama penulis 2. Ambil tulisan dari baris tersebut
Kata kunci	1. Periksa tulisan dari baris paling akhir 2. Jika ditemukan kata "kata kunci" atau "keyword", ambil tulisan dari baris akhir 3. Jika tidak, cek dari dua baris terakhir 4. Jika tidak ditemukan maka jumlah karakter pada dua baris tersebut akan dihitung dan dibandingkan 5. Jika jumlah baris terakhir < baris kedua terakhir, maka batas awal diset ke baris kedua terakhir 6. Jika tidak maka batas akhir = batas akhir
Isi	1. Set baris awal ke baris setelah NIM 2. Set baris akhir ke baris sebelum baris awal bagian kata kunci 3. Ambil tulisan dari baris awal sampai baris akhir yang sudah diset

Setelah semua bagian teridentifikasi, maka sistem telah menyelesaikan setiap prosesnya.

2.15 Hasil Pengujian

Pengujian dilakukan dengan menggunakan 40 citra abstrak skripsi yang didapatkan dari hasil foto dan *scan.* Pengujian terlebih dahulu dilakukan pada 6 sampel data untuk mencari parameter SVM yang terbaik untuk nantinya akan digunakan dalam pengujian keseluruhan. Berikut adalah parameter yang digunakan.

Tabel 2. Tabel parameter SVM

Kernel	Parameter	
	C	Gamma (γ)
Linear	1	-
	10	-
	100	-
RBF	1	Scale
	100	Scale
	1	1
	100	1

Pengujian juga dilakukan dengan model yang dilatih dengan 3 ukuran berbeda, yaitu 10x10,15x15,20x20 px. Selain itu, pengujian juga akan dilakukan hanya pada model SVM saja untuk mendapatkan akurasi dari model yang sebenarnya tanpa terganggu proses lainnya dengan menggunakan 225 citra karakter yang akan digunakan untuk menguji model tersebut. Dan pada pengujian terakhir, pengkategorian

informasi akan dilakukan untuk mendapatkan tingkat akurasi ekstraksi informasi menggunakan *Rules Based.*

2.15.1 Hasil Pengujian Sampel

Dari hasil pengujian sampel, didapatkan bahwa model yang dilatih menggunakan citra berukuran 20x20 melampaui kemampuan model yang dilatih dengan citra ukuran lain. Berikut hasil uji dari model tersebut.

Tabel 3. Tabel Uji Sampel

Model		Akurasi	
Kernel	Parameter	*Case Sensitive*	*Case Insensitive*
Linear	C=1	**16.4%**	**17.61%**
	C=10	**16.4%**	**17.61%**
	C=100	**16.4%**	**17.61%**
RBF	C=1,Gamma=scale	15.86%	17%
	C=100,Gamma=scale	16.05%	17.22%
	C=1,Gamma=1	0.2%	1.96%
	C=100,Gamma=1	0.37%	2.12%

Dari hasil pengujian, maka parameter yang akan digunakan pada pengujian keseluruhan data adalah Model yang dilatih dengan citra ukuran 20x20px dan berparameter kernel linear dengan nilai C = 100.

2.15.2 Hasil Pengujian Keseluruhan

Dengan menggunakan model dan parameter terbaik pada hasil pengujian sampel, berikut adalah nilai keakurasian dari keseluruhan data pengujian.

$$Accuracy_{(case\ sensitive)} = 5{,}02\%$$
$$Accuracy_{(case\ insensitive)} = 5{,}47\%$$

2.15.3 Hasil Pengujian Model SVM

Untuk mendapatkan nilai akurasi dari model SVM yang lebih akurat, maka pengujian model SVM dilakukan secara terpisah. Dari hasil pengujian sampel, didapatkan bahwa model yang dilatih menggunakan citra berukuran 20x20 melampaui kemampuan model yang dilatih dengan citra ukuran lain. Berikut hasil uji dari model tersebut.

Tabel 4. Tabel Uji SVM

Model		Akurasi	
Kernel	Parameter	*Case Sensitive*	*Case Insensitive*
Linear	C=1	**46.61%**	**54.30%**
	C=10	**46.61%**	**54.30%**
	C=100	**46.61%**	**54.30%**
RBF	C=1,Gamma=scale	46.15%	52.94%
	C=100,Gamma=scale	45.25%	51.58%
	C=1,Gamma=1	1.36%	2.71%
	C=100,Gamma=1	1.36%	2.71%

Maka didapatkan hasil terbaik dari pengujian menghasilkan nilai akurasi terbaik sebesar 46.61% untuk case sensitive dan 54.30% untuk case insensitive.

2.15.4 Hasil Pengujian Ekstraksi Informasi

Pengujian dilakukan dengan membandingkan bagian kategori pada citra sebenarnya dengan bagian kategori yang ditemukan oleh sistem. Berikut adalah hasil dari pengujian pengkategorian informasi pada citra abstrak.

Tabel 5. Tabel Uji Ekstraksi Informasi

Kode Abstrak	Terdeteksi sesuai?				
	Judul	Nama	Nim	Isi	Keyword
F1	ya	ya	ya	ya	ya
F2	ya	ya	ya	ya	ya
F3	ya	ya	ya	tidak	tidak
F4	ya	ya	ya	ya	ya
F5	ya	ya	ya	ya	ya
F6	ya	ya	ya	ya	ya
S1	ya	ya	ya	ya	ya
S2	ya	ya	ya	ya	ya
S3	ya	ya	ya	ya	ya
S4	ya	ya	ya	ya	ya
S5	ya	ya	ya	ya	ya
S6	ya	ya	ya	ya	ya
S7	ya	ya	ya	ya	ya
S8	ya	ya	ya	ya	ya
S9	ya	ya	ya	ya	ya
S10	ya	ya	ya	ya	ya
S11	ya	ya	ya	ya	ya
S12	ya	ya	ya	ya	ya
S13	ya	ya	ya	ya	ya
S14	ya	ya	ya	ya	ya
S15	ya	ya	ya	ya	ya
S16	ya	ya	ya	ya	ya
S17	ya	ya	ya	ya	ya
S18	ya	ya	ya	ya	ya
S19	ya	ya	ya	ya	ya
S20	ya	ya	ya	ya	ya
S21	ya	ya	ya	ya	ya
S22	ya	ya	ya	ya	ya
S23	ya	ya	ya	ya	ya
S24	ya	ya	ya	ya	ya
S25	ya	ya	ya	ya	ya
S26	ya	ya	ya	ya	ya
S27	ya	ya	ya	ya	ya
S28	ya	ya	ya	ya	ya

S29	ya	ya	ya	ya	ya
S30	ya	ya	ya	ya	ya
S31	ya	ya	ya	ya	ya
S32	ya	ya	ya	ya	ya
S33	ya	ya	ya	ya	ya
S34	ya	ya	ya	ya	ya

Sehingga didapatkan nilai akurasi sebagai berikut:

$$\text{Accuracy} = \frac{Hasil\ benar}{Banyak\ data} x\ 100\% = 99\ \%$$

3 PENUTUP

Kesimpulan yang dapat diambil dari penelitian pengenalan karakter tulisan pada dokumen abstrak skripsi dengan metode Support Vector Machine maka diperoleh akurasi terbaik dari pengenalan karakter tulisan sebesar 5,02% untuk case sensitive dan 5,47% untuk case insensitive dengan model SVM terbaik berkernel linear dengan nilai parameter C=100 dan dilatih dengan citra berukuran 20x20px, sedangkan untuk tingkat akurasi pengenalan karakter menggunakan SVM itu sendiri mencapai 46.61% untuk case sensitive dan 54.30% untuk case insensitive. Rendahnya tingkat akurasi pengenalan pada citra abstrak dipengaruhi oleh proses segmentasi yang kurang mampu menyelesaikan masalah yang ada pada pemisahan karakter. Sedangkan untuk pengkategorian informasi dengan menggunakan Rule Based System memiliki akurasi sebesar 99%.

Saran untuk pengembangan pengenalan tulisan dari citra abstrak skripsi ini agar menjadi lebih baik adalah sebagai berikut:

1. Diperlukan metode segmentasi citra tulisan untuk kasus broken character dan touching character.
2. Menggunakan algoritma pengenalan lainnya seperti Artificial Neural Network ataupun algoritma lainnya sebagai pembanding.
3. Menggunakan metode lain pada tahap preprocessing agar dapat memisahkan karakter uppercase dan lower case.

DAFTAR PUSTAKA

[1] D. Mustaqwa dan I. Nelly, "Implementasi Ekstraksi Informasi Pada Dokumen Teks Skripsi Menggunakan Metode Rule Based," Universitas Komputer Indonesia, 2017.

[2] I. M. Raden Sofian Bahri, "Perbandingan Algoritma Template Matching dan Feature Extraction pada Optical Character Recognition," *J. Komput. dan Inform.*, vol. 1, no. 1, hal. 29, 2012.

[3] R. Anugrah dan K. B. Y. Bintoro, "Latin Letters Recognition Using Optical Character Recognition to Convert Printed Media Into Digital Format," *J. Elektron. dan Telekomun.*, vol. 17, no. 2, hal. 56, 2017.

[4] P. A, "A Comparative Study of Optical Character Recognition for Printed and Handwritten Tamil Text," *Int. J. Eng. Res. Appl.*, vol. 7, no. 8, hal. 56–60, 2017.

[5] M. R. Phangtriastu, J. Harefa, dan D. F. Tanoto, "Comparison between Neural Network and Support Vector Machine in Optical Character Recognition," *Procedia Comput. Sci.*, vol. 116, hal. 351–357, 2017.

[6] B. El Kessab, C. Daoui, B. Bouikhalene, dan R. Salouan, "A Comparative Study between the Support Vectors Machines and the K-Nearest Neighbors in the Handwritten Latin Numerals Recognition," *Int. J. Signal Process. Image Process. Pattern Recognit.*, vol. 8, no. 2, hal. 325–336, 2015.

[7] T. Kalaiselvi, "A Comparative Study On Thresholding Techinques For Gray Image Binarization," *Int. J. Adv. Res. Comput. Sci.*, vol. 8, no. 7, hal. 1168–1172, Agu 2017.

[8] M. F. Soleh dan K. K. Purnamasari, "Implementasi Metode Support Vector Machine (Svm) Dan Zoning Untuk Pengenalan Tulisan Tangan Pada Kasus Pengecekan Jawaban Ujian," Universitas Komputer Indonesia, 2018.

[9] J. C. Platt, "Sequential Minimal Optimization : A Fast Algorithm for Training Support Vector Sequential Minimal Optimization : A Fast Algorithm for Training Support Vector Machines," *Microsoft Res.*, vol. 98, no. Technical Report, hal. 14, 2013.

[10] B. Jain dan M. Borah, "A Comparison Paper on Skew Detection of Scanned Document Images Based on Horizontal and Vertical," *Int. J. Sci. Res. Publ.*, vol. 4, no. 6, hal. 4–7, 2014.

[11] A. Septiarini, "Segmentasi Karakter Menggunakan Profil Proyeksi," *Inform. Mulawarman J. Ilm. Ilmu Komput.*, vol. 7, no. 2, hal. 66–69, 2012.

[12] N. Priyanka, S. Pal, dan R. Mandal, "Line and Word Segmentation Approach for Printed Documents," *Int. J. Comput. Appl. IJCA ,Special Issue RTIPPR*, no. 1, hal. 30–36, 2010.

[13] S. Safinaz, "An Efficient Algorithm for Image Scaling with High Boost Filtering," *Int. J. Sci. Res. Publ.*, vol. 4, no. 5, hal. 1–9, 2014.