BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Ekstraksi informasi adalah suatu proses untuk mengubah informasi tidak terstruktur yang terdapat dalam teks ke dalam data terstruktur [1]. Ekstraksi informasi dilakukan dengan cara mendeteksi komponen-komponen pada suatu dokumen. Dalam kasus ekstraksi informasi biasanya melibatkan dokumen yang memiliki informasi dengan format yang beragam. Dokumen yang memiliki informasi dengan format yang beragam adalah dokumen surat masuk. Surat masuk adalah surat yang diterima oleh suatu instansi yang berasal dari instansi lain dengan tujuan menyampaikan suatu informasi. Dalam dokumen surat masuk terdapat komponen kepala surat, tanggal surat, nomor surat, dan sebagainya yang memiliki format beragam sesuai masing-masing instansi dari pembuat surat. Hal ini menyebabkan tidak adanya aturan baku dalam struktur penulisan komponen surat masuk. Maka untuk dapat mendeteksi setiap informasi pada surat masuk yang beragam sulit dilakukan, karena banyaknya struktur pola–pola komponen yang harus dipelajari pada dokumen surat masuk dengan cara menganalisis pola–pola dokumen yang akan dijadikan fitur yang dapat mengenali komponen surat masuk.

Penelitian mengenai ekstraksi informasi sudah dilakukan oleh Dimas [2] pada dokumen skripsi dengan metode *Rule-Based* yang memperoleh nilai akurasi rata-rata sebesar 57%, hasil ini membuktikan bahwa metode *Rule-Based* mengalami kesulitan ketika mengolah dokumen yang memiliki format yang beragam. Metode *Rule-Base* yang diterapkan dipengaruhi oleh ekstraksi fitur yang berbasis aturan yang mengenali ciri dari skripsi tahun 2017 yang memperoleh nilai akurasi tinggi sebesar 100%, sedangkan ketika pengujian dokumen teks skripsi tahun 2013 yang memiliki perbedaan format nilai akurasi turun. Penelitian selanjutnya yang telah dilakukan oleh Firdamdam [3] pada dokumen jurnal dan abstrak dalam karya tulis ilmiah yang memiliki perbedaan format pada tahun 2013 dan 2017 metode *LVQ* memperoleh hasil akurasi rata-rata 78%. Hasil akurasi yang cukup tinggi ini

ternyata dipengaruhi oleh ekstraksi fitur yang diterapkan sebanyak 15 fitur untuk dapat mendeteksi komponen pada dokumen karya tulis ilmiah. 15 Fitur yang dibuat pada kasus dokumen karya tulis ilmiah ini belum tentu cocok diterapkan pada dokumen yang berbeda, dikarenakan setiap dokumen memiliki pola tersendiri. Pada penelitian yang dilakukan oleh Chandra [4] untuk kasus ekstraksi informasi dengan media surat masuk menggunakan metode *Naive Bayes* memperoleh akurasi 96,96%, nilai akurasi yang tinggi dipengaruhi oleh ekstraksi fitur yang diterapkan sebanyak 11 fitur yang mewakili ciri pada dokumen surat masuk. Pada penelitian di bidang klasifikasi abstrak tesis menggunakan *LVQ* [5] menghasilkan nilai akurasi sebesar 90%. Pada penelitian dibidang klasifikasi [6] yang membandingkan metode *SVM* dengan *LVQ* menghasilkan nilai akurasi 80% untuk *SVM* sedangkan 87% untuk *LVQ* dan menyimpulkan bahwa *LVQ* 11% lebih baik. Dengan nilai akurasi yang cukup tinggi dari literatur penelitian sebelumnya, hal ini melandasi dipilihnya metode *Learning Vector Quantization* untuk surat masuk.

Berdasarkan paparan sebelumnya, dapat disimpulkan bahwa ekstraksi informasi sangat dipengaruhi oleh ekstraksi fitur yang diterapkan untuk mengenali pola–pola yang terdapat pada suatu dokumen. Maka dari itu penelitian ini akan menganalisis fitur dari penelitian sebelumnya yaitu Firdamdam [3] dan Chandra [4] yang akan diterapkan pada dokumen surat masuk dengan menggunakan metode *Learning Vector Quantization* sebagai klasifikasi dalam ekstraksi informasi.

1.2 Identifikasi Masalah

Berdasarkan latar belakang tersebut terdapat identifikasi masalah sebagai berikut :

1. Dibutuhkannya analisis fitur untuk dapat mengenali pola–pola dokumen surat masuk agar dapat mendeteksi komponen surat masuk yang beragam.
2. Belum terukurnya tingkat akurasi metode *Learning Vector Quantization* pada kasus ekstraksi informasi dengan dokumen surat masuk.

1.3 Maksud dan Tujuan

Maksud dari penelitian ini adalah menerapkan hasil analisis fitur pada sistem ekstraksi informasi surat masuk menggunakan algoritma *Learning Vector Quantization*. Adapun tujuan dari penelitian ini adalah mengukur dan menghasilkan akurasi algoritma *LVQ* pada kasus ekstraksi informasi surat masuk.

1.4 Batasan Masalah

Agar pembahasan ini terfokus dalam lingkup masalah yang diinginkan, maka ada batasan masalah yang akan dilakukan. Adapun batasan masalah yang akan dibatasi adalah sebagai berikut.

1. Data Masukan
    a. Data *training* dan *testing* diperoleh dari surat masuk dinas resmi yang didapat dari penelitian sebelumnya yaitu penelitian Chandra [4].
    b. Format awal data surat masuk memiliki format *PDF* hasil konversi *file .docx*.
    c. Data masukan *training* memiliki format *CSV* yang diperoleh dari hasil dari konversi file *PDF* ke *TXT* kemudian dikonversi menjadi format *CSV*, yang berisi 2 kolom yaitu data perbaris dan label kelas surat masuk.
    d. Data masukan *testing* memiliki format *TXT* yang diperoleh dari hasil dari konversi file *PDF* ke *TXT,* yang berisi data baris surat masuk.
    e. Data yang diambil sesuai dengan aturan–aturan penulisan surat resmi yang memiliki kepala surat, tanggal surat, nomor surat, perihal surat, penerima surat, isi surat, pengirim surat, dan tembusan surat.

2. Proses
    a. Tahapan proses pada *preprocessing data training* meliputi deteksi baris, tokenisasi, ekstraksi fitur, dan *training LVQ*.
    b. Tahapan proses pada *preprocessing data testing* meliputi deteksi baris, tokenisasi, ekstraksi fitur, dan *testing LVQ* .
    c. Ekstraksi fitur yang digunakan didapatkan dari hasil analisis dari penelitian sebelumnya yaitu penelitian yang dilakukan oleh Firdamdam Sasmita [3] dan Chandra Ratiwi [4].
3. Keluaran

    Data keluaran yang dihasilkan oleh sistem meliputi token dan kelas dengan kategori pada surat masuk sebanyak 8 kategori yaitu kepala surat, tanggal surat, nomor surat, perihal surat, penerima surat, isi surat, pengirim surat, dan tembusan surat.

1.5 Metode Penelitian

Alur metode pada penelitian ini terdapat empat tahapan, yaitu studi literatur, pengumpulan *dataset* surat masuk, analisis, pembangungan perangkat lunak ekstraksi informasi surat masuk menggunakan metode *Learning Vector Quantization*, pengujian, dan penarikan kesimpulan. Berikut ini blok diagram alur metode penelitian ini dapat dilihat pada Gambar 1.1.

Gambar 1.1 Blok Diagram Alur Metode Penelitian

1.5.1 Studi Literatur

Studi literatur pada penelitian ini dengan cara membaca dan mengumpulkan referensi–referensi dari jurnal penelitian sebelumnya, dan mencari pengetahuan tentang metode *Learning Vector Quantization* dari jurnal dan internet.

1.5.2 Pengumpulan *Dataset* Surat Masuk

Setelah tahapan studi literatur maka langkah selanjutnya yaitu pengumpulan *dataset* surat masuk, *dataset* surat masuk ini digunakan sebagai masukan sistem dan masukan untuk algoritma *Learning Vector Quantization. Dataset* surat masuk yang diperoleh didapatkan dari penelitian sebelumnya yaitu penelitian yang dilakukan oleh Chandra [4] yaitu surat masuk dinas.

1.5.3. Tahap Analisis

Tahapan berikutnya yaitu tahap analisis, pada tahapan ini dilakukan analisis tahapan agar proses ekstraksi informasi dapat dilakukan, adapun analisis tersebut yaitu analisis *preprocessing data training* yang terdiri dari deteksi baris, tokenisasi, ekstraksi fitur dan *training LVQ*, kemudian analisis *preprocessing data testing* yang terdiri dari deteksi baris, tokenisasi, ekstraksi fitur dan *testing LVQ*, dan tahap analisis evaluasi hasil klasifikasi.

1.5.4. Pembangunan Perangkat Lunak Ekstraksi Informasi *LVQ*

Metode yang digunakan dalam pengembangan perangkat lunak adalah metode *prototype* [7], yang meliputi beberapa proses antara lain sebagai berikut.

a. Analisis Kebutuhan

   Pada tahap ini merupakan tahapan yang mendukung kebutuhan suatu sistem agar dapat dibangun. Pada sistem ekstraksi informasi surat masuk dibutuhkan analisis kebutuhan agar sistem yang dibangun sesuai dengan tujuannya. Adapun analisis kebutuhan sistem ekstraksi informasi terdiri dari analisis kebutuhan data, tahap ini akan menjelaskan data yang akan dijadikan masukan *training* dan *testing* untuk sistem yang diterapkan. Kemudian analisis kebutuhan non-fungsional yang terdiri dari analisis kebutuhan pengguna, analisis kebutuhan perangkat keras, dan analisis kebutuhan perangkat lunak. Selanjutnya analisis kebutuhan fungsional yang terdiri dari pemodelan alur data pada sistem ekstraksi informasi menggunakan *DFD*, perancangan antarmuka sistem yang terdiri dari *training* dan *testing*, dan alur dari semua proses pada sistem yang digambarkan dengan *flowchart*.

b. Desain *Prototype*

   Tahap ini merupakan penjabaran dari proses sebelumnya yang didapatkan dari analisis kebutuhan, selanjutnya proses desain *prototype* sistem yang akan dibangun dengan mendesain antarmuka yang diterapkan akan mengalami evaluasi terus menerus sesuai kebutuhan perangkat lunak ekstraksi informasi.

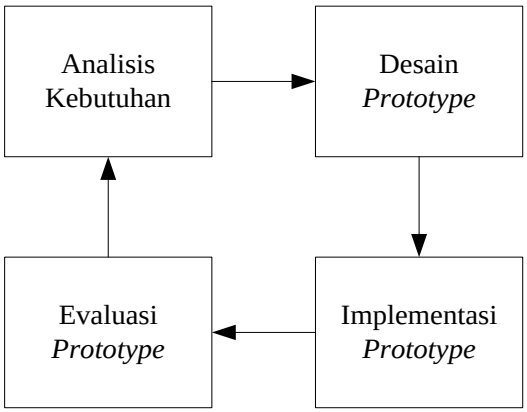
c. Implementasi *Prototype*

Setelah tahap desain sistem selesai, maka tahap selanjutnya adalah implementasi ke dalam bahasa pemrograman sesuai dengan desain sistem yang sebelumnya sudah dibuat.

d. Evaluasi *Prototype*

Setelah program selesai, maka tahap selanjutnya adalah evaluasi atau pengujian terhadap *prototype* sistem yang telah dibuat. Pada tahap ini dilihat apakah masih ada kekurangan atau error pada *prototype*, apabila terdapat *error* atau kekurangan maka akan dicatat kemudian proses kembali lagi ke tahap nomor 1 untuk melakukan perbaikan. Proses berakhir ketika pada tahap evaluasi tidak terdapat lagi error atau kekurangan.

Penggambaran model *prototype* dapat dilihat pada Gambar 1.2 dibawah ini.



Gambar 1.2 Model *Prototyping*

1.5.5. Pengujian

Dalam tahap ini akan dilakukan pengujian terhadap perangkat lunak yang telah dibangun menggunakan metode pengujian *Black Box*. Perhitungan nilai akurasi dilakukan dengan menggunakan perhitungan akurasi.

1.5.6. Penarikan Kesimpulan

Tahap ini akan menyajikan hasi keseluruhan dari hasil penelitian dan nilai akurasi yang dihasilkan dari algoritma *Learning Vector Quantization* dan menghasilkan kesimpulan dan saran terkait pengujian yang telah dilakukan

1.6 Sistematika Penulisan

Sistematika penulisan ini disusun untuk memberikan gambaran umum tentang penelitian yang dijalankan dan dibagi dalam beberapa bab dengan pokok pembahasan sistematika secara umum sebagai berikut :

BAB 1 PENDAHULUAN

Bab ini menguraikan tentang dasar-dasar pemikiran yang berisi tentang latar belakang masalah, rumusan masalah, maksud dan tujuan, batasan masalah, metode penelitian, serta sistematika penulisan.

BAB 2 TINJAUAN PUSTAKA

Pada bab ini Membahas berbagai konsep dasar dan teori-teori yang berkaitan dengan topik penelitian yang dilakukan dan hal-hal yang berguna dalam proses analisis permasalahan serta tinjauan terhadap penelitian-penelitian serupa yang telah pernah dilakukan sebelumnya termasuk sintesisnya. Membahas tentang tinjauan perusahaan dan konsep dasar serta teori-teori yang berkaitan dengan topik penelitian dan yang melandasi pembangunan aplikasi ini.

BAB 3 ANALISIS DAN PERANCANGAN SISTEM

Bab ini berisi tentang analisis dan perancangan aplikasi yang dibangun, meliputi analisis masalah, analisis data masukan, dan analisis sistem, perancangan prosedural serta perancangan antarmuka dan jaringan semantik.

BAB 4 IMPLEMENTASI DAN PENGUJIAN SISTEM

Dalam bab ini menjelaskan tentang implementasi dan pengujian dari perangkat lunak yang dibangun sesuai dengan analisis dan perancangan yang telah dibuat. Implementasi yang akan dibahas adalah implementasi antarmuka. Pada bagian pengujian yang akan dibahas adalah pengujian sistem yang telah dibuat

BAB 5 KESIMPULAN DAN SARAN

Pada bab ini berisi tentang kesimpulan dari penelitian yang telah dilakukan serta memaparkan saran yang dapat membantu dalam penelitian berikutnya ataupun saran untuk tempat penelitian itu sendiri.