BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Ekstraksi Informasi merujuk pada ekstraksi otomatis dari informasi terstruktur seperti entitas, hubungan antar entitas, dan atribut deskripsi entitas dari sumber yang tidak terstruktur [1]. Ekstraksi informasi dilakukan dengan cara mendeteksi bagian-bagian yang ada pada suatu dokumen, salah satunya adalah dokumen Surat Keputusan.

Penelitian sebelumnya mengenai ekstraksi informasi telah dilakukan oleh Arbi [2] dan Dimas [3] yang menggunakan metode *rule-based* dengan nilai akurasi tertinggi mencapai 100%. Pada penelitian yang dilakukan oleh Arbi [2], ekstraksi informasi dilakukan untuk dokumen teks novel, sedangkan pada penelitian yang dilakukan oleh Dimas [3], dokumen yang digunakan adalah dokumen teks karya tulis skripsi. Walaupun nilai akurasi yang dihasilkan tinggi, namun karena masih menggunakan metode *rule-based*, kesulitan mulai ditemui ketika menghadapi format dokumen yang beragam. Semakin banyak format dokumen yang dihadapi, maka semakin banyak aturan yang harus dibuat. Hal ini dapat diatasi dengan menggunakan *machine learning*.

Penelitian mengenai ekstraksi informasi yang menggunakan *machine learning* telah dilakukan dengan metode Naïve Bayes oleh Chandra [4] dan *Conditional Random Fields* (CRF) oleh Chengzi Chang [5] dan Fuchun Peng [6]. Pada penelitian yang dilakukan oleh Chandra [4], mengangkat kasus dokumen surat masuk mendapatkan akurasi 96,96%. Penelitian lain yang dilakukan oleh Fuchun Peng melakukan perbandingan penggunaan metode *Hidden Markov Models* (HMM), *Support Vector Machine* (SVM) dan *Conditional Random Fields* (CRF)[6]. Pada penelitian tersebut, akurasi tertinggi didapat menggunakan metode CRF dengan nilai 98.3%. Meskipun nilai akurasi yang didapat pada penelitian dengan metode *Naïve Bayes* lebih besar jika dibandingkan dengan penelitian yang menggunakan metode CRF, pengembangan menggunakan metode *Naïve Bayes* dinilai sudah tertinggal. Hal ini melandasi dipilihnya metode *Conditional Random Fields* (CRF) dalam penelitian ini.

Berdasarkan uraian diatas, penelitian ini akan membangun sebuah sistem ekstraksi informasi dengan metode *Conditional Random Fields* dengan mengangkat kasus pada dokumen surat keputusan berbahasa Indonesia. Dengan batasan dokumen yang akan digunakan pada penelitian ini adalah dokumen surat keputusan yang merujuk kepada satu orang.

1.2 Identifikasi Masalah

Berdasarkan latar belakang masalah yang diurakian diatas, masalah yang dapat teridentifikasi yaitu dibutuhkannya hasil akurasi dari sistem ekstraksi informasi dengan menggunakan metode *Conditional Random Fields* untuk dokumen surat keputusan dengan format beragam.

1.3 Maksud dan Tujuan Penelitian

Berdasarkan masalah yang diteliti, maksud dari penelitian ini adalah untuk membangun sistem yang dapat mengekstrak informasi yang ada pada surat keputusan. Sementara tujuan dari penelitian ini adalah mengukur kinerja dari penggunaan metode *Conditional Random Fields* dalam ekstraksi informasi pada kasus Surat Keputusan.

1.4 Batasan Masalah

Adapun batasasn masalah dalam penelitian ini adalah sebagai berikut.

1. Data masukan berupa hasil konversi dari hasil scan dokumen surat keputusan menjadi teks dengan menggunakan *Optical Character Recognition* (OCR) dari ABBY.
2. Data masukan yang digunakan untuk proses *Training* adalah dengan format *file* masukan .csv yang terdiri dari 2 kolom data yaitu hasil konversi dan label per kelas yang dilakukan secara manual.

3. Data masukan yang digunakan untuk proses *Testing* adalah dengan format *file* masukan .csv yang terdiri dari 2 kolom data yaitu hasil konversi dan label awal kelas.
4. Setiap data pada komponen lembar surat keputusan ditokenisasi per kata, kumpulan kata, dan simbol.
5. Ekstraksi fitur merujuk kepada banyaknya kelas yaitu sebanyak 15 kelas.
6. Data keluaran berupa informasi meliputi token dengan label kelas sebanyak 15 kelas. Namun hanya 6 kelas yang kemudian akan diambil, meliputi nomor surat, tentang, jabatan yang menetapkan surat, nama orang yang dituju dalam surat, tanggal penetapan dan nama penetap surat.

1.5 Metode Penelitian

Pada penelitian ini digunakan metode kuantitatif. Alur penelitian ini terdapat lima tahapan, yaitu pengumpulan data, analisis metode CRF, pembangunan perangkat lunak, pengujian dan kesimpulan dan saran.

1.5.1 Metode Pengumpulan Data

Metode pengumpulan data yang digunakan adalah dengan melakukan studi literatur. Studi literatur adalah mengumpulkan data melalui buku-buku, hasil penelitian, jurnal, situs internet, artikel terkait dengan permasalahan yang terjadi.

1.5.2 Metode Tahapan Analisis

Berdasarkan hasil kajian dan evaluasi yang dilakukan setelah tahap pengumpulan data, tahapan analisis pada penelitian ini adalah sebagai berikut.

a. Menganalisis Dokumen Masukan, yaitu menganalisis pola dokumen yang akan digunakan sebagai objek penelitian. Dokumen yang digunakan adalah dokumen surat keputusan yang telah di *scan* dan tersimpan dalam format .jpg.

b. Menganalisis Metode, mulai dari ekstraksi fitur, proses training hingga proses testing sehingga dapat menghasilkan prediksi label kelas untuk suatu data masukan. Metode yang digunakan adalah metode *Conditional Random Fields*.
c. Menganalisis Data Keluaran, yaitu melakukan analisis terhadap nilai akurasi yang didapat dari tahap pengujian.

1.5.3 Metode Pengembangan Perangkat Lunak

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

a. Analisis Kebutuhan

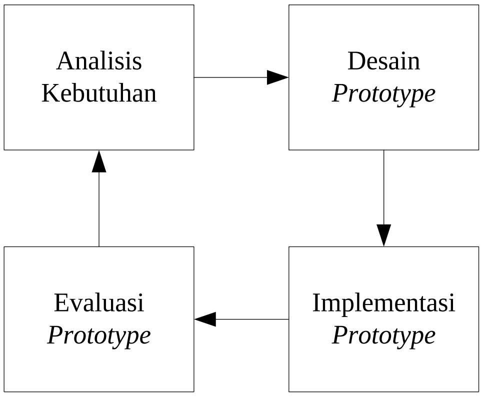
   Pada tahap ini dilakukan analisis kebutuhan sistem yang diperlukan dalam membangun sistem ekstraksi informasi menggunakan CRF. Analisis dilakukan terhadap data masukan, tahapan proses yang harus dilakukan, perangkat lunak, dan perangkat keras yang dibutuhkan.
b. Desain *Prototype*

   Tahap ini merupakan penjabaran dari proses sebelumnya, pada tahap ini ditentukan seperti apa sistem yang nantinya akan dibangun, metode yang digunakan dalam proses 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. *Construction*

   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 dibawah ini.



Gambar 1. 1 Model Prototyping [7]

1.5.4 Metode Pengujian

Dalam tahap ini akan dilakukan pengujian akurasi dari pengaplikasian metode *Conditional Random Fields* pada sistem ekstraksi informasi surat keputusan. Perhitungan nilai akurasi dilakukan dengan menggunakan perhitungan akurasi.

1.5.5 Penarikan Kesimpulan

Berdasarkan pada hasil dari pengujian maka dapat ditarik kesimpulan mengenai hasil dari penelitian yang dilakukan yaitu ekstraksi informasi dengan menggunakan CRF.

1.6 Sistematika Penulisan

Sistematika penulisan disusun untuk memberikan gambaran secara umum mengenai permasalahan dan pemecahannya. Sistematika penulisan tugas akhir ini adalah sebagai berikut.

BAB 1 PENDAHULUAN

Bab ini membahas mengenai latar belakang, identifikasi masalah, maksud dan tujuan, batasan masalah, metode penelitian, serta sistematika penulisan untuk menjelaskan pokok – pokok pembahasannya.

BAB 2 LANDASAN TEORI

Pada bab ini akan menjelaskan mengenai objek dari penelitain, dan teori – teori pendukung yang berhubungan dengan penelitian ini serta perangkat lunak pendukung yang digunakan dalam penelitian 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

Bab ini berisi mengenai hasil dari implementasi dari analisis yang telah dilakukan di BAB 3 dan pengujian akurasi metode CRF yang digunakan untuk ekstraksi informasi pada surat keputusan.

BAB 5 KESIMPULAN DAN SARAN

Bab ini bereisi mengenai kesimpulan yang diperoleh dari hasil pengujian system serta saran yang dapat digunakan untuk pengembangan ekstraksi informasi pada dokumen surat keputusan di penelitian selanjutnya.