

BAB 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Informasi merupakan sekumpulan fakta yang ada dan diolah menjadi bentuk data sehingga bisa berguna dan dapat digunakan oleh siapapun yang membutuhkan data tersebut sebagai pengetahuan seseorang. Ekstraksi Informasi merupakan proses pengambilan, menguraikan dan memisahkan informasi dari teks yang tidak terstruktur, semi struktur menjadi terstruktur [1]. Ekstraksi informasi bisa berbeda-beda tergantung dari apa yang dibutuhkan dan tidak hanya terbatas pada sebuah dokumen saja.

Pada penelitian sebelumnya telah dilakukan ekstraksi informasi pada dokumen teks skripsi dengan menggunakan metode *rule-based*. Akan tetapi jika dilakukan pengujian dengan dokumen teks skripsi antara tahun 2013 hingga 2017 nilai rata-rata akurasi yang didapat hanya sekitar 56% [2], karena berbasis *rule-based* jika template dokumen berubah maka nilai akurasi dari hasil ekstraksi informasi akan menurun karena sistem yang dibuat tidak dapat mengenali data yang berbeda, dan harus melakukan perubahan atau penambahan aturan baru.

Penelitian setelahnya, pada kasus yang sama telah menggunakan machine learning untuk melakukan ekstraksi informasi pada makalah karya tulis ilmiah dengan menggunakan metode *Support Vector Machine* (SVM) [3], dan ekstraksi informasi dokumen karya tulis menggunakan metode algoritma *Learning Vector Quantization* (LVQ) dengan melakukan pengujian pada masing-masing 40 dokumen karya tulis ilmiah (sampul dan abstrak) tahun 2011 hingga 2018 yang memperoleh akurasi rata-rata 78% [4].

Dalam penelitian ini akan menggunakan metode *Hidden Markov Model* (HMM) karena dari beberapa referensi journal dan paper mengatakan metode ini memiliki nilai akurasi yang baik dan tinggi dalam beberapa topik penelitian seperti *image recognition, gesture, classifier, part-of-speech tagging, speech recognition* [5] dan *named entity recognition* [6] sehingga peneliti ingin menerapkan algoritma

ini untuk kasus ekstraksi informasi pada dokumen karya tulis ilmiah menggunakan *Hidden Markov Model (HMM)*.

Berdasarkan uraian tersebut maka pada penelitian ini akan dibangun sistem ekstraksi informasi dokumen karya tulis ilmiah menggunakan *Hidden Markov Model*.

1.2 Identifikasi Masalah

Berdasarkan latar belakang masalah yang diuraikan diatas, permasalahan yang diidentifikasi yaitu kurangnya nilai akurasi dari sistem ekstraksi informasi dokumen karya tulis ilmiah yang telah dibuat sebelumnya [4].

1.3 Maksud dan Tujuan

Berdasarkan identifikasi masalah yang diteliti, maksud dari penelitian ini yaitu membangun sistem ekstraksi informasi dokumen karya tulis ilmiah menggunakan *Hidden Markov Model*.

Dan tujuan dari penelitian ini adalah untuk mengetahui dan meningkatkan nilai akurasi sistem ekstraksi informasi dokumen karya tulis ilmiah menggunakan *Hidden Markov Model*.

1.4 Batasan Masalah

Agar penelitian tidak meluas, maka penulis memberi batasan masalah sebagai berikut:

1. Data Masukan yang digunakan meliputi:
 - a. Masukan data latih berekstensi *.csv [4].
 - b. Masukkan data uji berekstensi *.txt [4].
 - c. *dataset* hanya lembar sampul dan abstrak dari Skripsi Program Studi Teknik Informatika, Universitas Komputer Indonesia.
 - d. Lembar sampul dan abstrak skripsi masing-masing sebanyak 40 dokumen dari tahun 2011 sampai 2018.
2. Proses

- a. *Preprocessing* lembar sampul dan abstrak dilakukan tokenisasi kata, angka dan yang mengandung simbol.
 - b. Ekstraksi fitur yang digunakan sebanyak 15 fitur meliputi: Initcaps, Allcaps, Containsdigit, Alldigit, Containsdots, Lowercase, Punctuation, Eightdigit, Word, Line_Start, Line_In, Line_End, Person, Organization, Year [4].
 - c. Algoritma HMM:
 - i) Menentukan kondisi awal (π).
 - ii) Mencari probabilitas transisi dan emisi.
 - iii) Algoritma viterbi untuk mencari nilai optimum.
3. Data Keluaran
- a. Data keluaran dari program sistem ekstraksi berupa hasil akurasi dari hasil perbandingan token-kelas data asli dan data hasil pengujian dengan kelas sebanyak 16 kelas yang terdiri dari Judul penelitian (sampul), Jenis penelitian, Kalimat pengajuan, Penulis (sampul), NIM (sampul), Prodi, Fakultas, Universitas, Tahun, Judul halaman (abstrak), Judul Penelitian (abstrak), *Other*, Penulis (abstrak), NIM (abstrak), Isi Abstrak dan Kata kunci.

1.5 Metode Penelitian

Pada penelitian ini terdapat enam tahap, yaitu identifikasi masalah, studi literatur, pengumpulan data, pembangunan perangkat lunak ekstraksi informasi, pengujian sistem ekstraksi informasi dan kesimpulan dari penelitian ini. Alur penelitian dapat dilihat pada Gambar 1.1.



Gambar 1.1 Diagram Alur Penelitian

1.5.1 Identifikasi Masalah

Identifikasi masalah, merupakan proses pengamatan terhadap penelitian yang sudah dilakukan sebelumnya untuk mendefinisikan kebutuhan sistem yang akan dibangun dan tujuan yang akan dicapai.

1.5.2 Studi Literatur

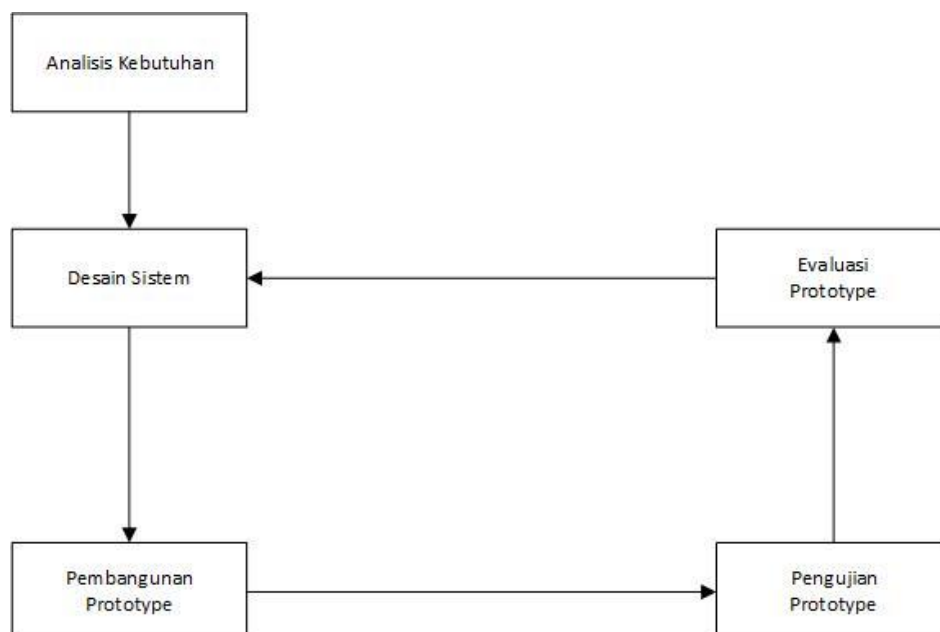
Studi literatur yaitu mencari referensi yang sesuai dengan pembahasan yang ada pada jurnal, paper, buku-buku dan informasi terkait yang ada pada penelitian dengan topik ekstraksi informasi dan algoritma HMM.

1.5.3 Pengumpulan Data

Pada tahap ini dilakukan pengumpulan data dari studi literatur dan *dataset* untuk pengujian maupun pelatihan dari lembar sampul dan abstrak pada skripsi program studi teknik informatika, universitas komputer indonesia.

1.5.4 Pembangunan Perangkat Lunak

Metode pembangunan perangkat lunak pada sistem ekstraksi informasi dokumen karya tulis ilmiah menggunakan model *prototype*. Metode ini digunakan agar proses pengembangan dapat berjalan dengan baik dan jika hasil belum sesuai maka dapat dilakukan perbaikan [7]. Berikut adalah tahapan pada model *prototype* ini dapat dilihat pada Gambar 1.2.



Gambar 1.2 Diagram Model *Prototype*

Penjelasan dari model *prototype* di atas adalah sebagai berikut.

1. Analisis Kebutuhan

Analisis kebutuhan dilakukan terhadap data masukan berupa sampul dan abstrak dokumen karya tulis ilmiah Program Studi Teknik Informatika, Universitas Komputer Indonesia.

2. Desain Sistem

Desain sistem *prototype* adalah tahap perancangan antarmuka, diagram konteks dan DFD.

3. Pembangunan *Prototype*

Pembangunan *prototype* adalah tahap mengimplementasikan hasil dari desain *prototype*, yang meliputi proses *preprocessing*, proses *training* dan *testing*.

4. Pengujian *Prototype*

Pada tahap pengujian *prototype*, fungsional dari sistem akan diuji serta hasil akurasi dari algoritma HMM.

5. Evaluasi *Prototype*

Pada tahap Evaluasi ini jika hasil dari pengujian *prototype* belum sesuai dengan yang diharapkan, maka akan dilakukan perbaikan.

1.5.5 Pengujian

Pada tahap pengujian, sistem ekstraksi informasi akan diperiksa fungsionalitasnya dan nilai akurasi yang didapat.

1.5.6 Kesimpulan Penelitian

Merupakan hasil akhir dan rangkuman dari penelitian, meliputi besarnya akurasi yang didapat dengan algoritma *HMM* pada sistem ekstraksi informasi dokumen karya tulis ilmiah.

1.6 Sistematika Penulisan

Sistematika penulisan disusun untuk memberikan gambaran umum pada setiap bab. Sistematika penulisan laporan penelitian ini sebagai berikut.

BAB 1 PENDAHULUAN

Bab ini membahas tentang latar belakang masalah, identifikasi masalah, maksud dan tujuan, batasan masalah, metode penelitian, serta sistematika penulisan dalam penelitian ini.

BAB 2 LANDASAN TEORI

Bab ini membahas mengenai teori dasar yang berhubungan dan digunakan untuk menganalisis masalah dalam penelitian, dalam penelitian ini membahas mengenai ekstraksi informasi, algoritma *HMM*, tokenisasi, ekstraksi fitur, bahasa pemrograman *php* dan perangkat lunak pendukung seperti *xampp* dan visual studio code.

BAB 3 ANALISIS DAN PERANCANGAN SISTEM

Bab ini membahas secara teknis mengenai tahap-tahap yang dilakukan pada sistem ekstraksi informasi dokumen karya tulis ilmiah menggunakan *HMM*, yang meliputi analisis masalah, analisis sistem, analisis data masukan, analisis *preprocessing*, ekstraksi fitur, algoritma *HMM*, fungsional sistem, dan perancangan antarmuka sistem ekstraksi informasi.

BAB 4 IMPLEMENTASI DAN PENGUJIAN SISTEM

Bab ini membahas mengenai proses analisis dan perancangan sistem ekstraksi informasi dokumen karya tulis ilmiah menggunakan *HMM*, yang meliputi proses ekstraksi informasi lembar sampul dan abstrak sampai dengan hasil klasifikasi, perancangan antarmuka, serta pengujian *dataset* terhadap klasifikasi yang dihasilkan oleh proses ekstraksi informasi.

BAB 5 KESIMPULAN DAN SARAN

Bab ini membahas mengenai hasil yang diperoleh dari penelitian yang telah selesai dilakukan pada sistem ekstraksi informasi dokumen karya tulis ilmiah menggunakan *Hidden Markov Model* dan saran untuk penelitian selanjutnya.

