

BAB 1

PENDAHULUAN

1.1. Latar Belakang Masalah

Ekstraksi Informasi adalah pengolahan data pada suatu bidang ilmu yang mana dibutuhkan pengambilan data-data dan informasi yang ada didalamnya, yaitu dapat berupa teks dan juga beragam data yang bertujuan supaya informasi yang tersajikan lebih beraturan. Ekstraksi informasi juga merupakan suatu bidang ilmu untuk pengolahan bahasa alami dengan cara mengubah teks tidak terstruktur menjadi sebuah informasi dalam bentuk yang lebih terstruktur [1].

Pendekatan yang dapat digunakan pada kasus ekstraksi informasi salah satunya menggunakan machine learning, karena proses pengembangannya tidak memerlukan banyak waktu, namun diperlukan model data latih yang besar. Ekstraksi informasi memiliki banyak sekali komponen didalamnya dan memiliki format yang beragam, dengan demikian dibutuhkan suatu sistem dalam pembuatan aturan ekstraksi yang beragam pula, serta dapat menghasilkan akurasi yang tinggi.

Penelitian sebelumnya yang terkait dengan dokumen karya tulis ilmiah pernah dilakukan dengan menggunakan sistem berbasis aturan untuk mendapatkan identitas yang terkandung pada dokumen karya tulis ilmiah skripsi yaitu diantaranya cover, judul, dan abstrak. Berdasarkan pengujian yang dilakukan pada 3 buah dokumen diperoleh akurasi 100% [12]. Sedangkan untuk pengujian 50 buah dokumen yang disertakan dengan tahun 2013, diperoleh akurasi rata-rata sebesar 57% [12]. Pada penelitian dengan kasus yang sama tetapi dengan menggunakan algoritma yang berbeda, berdasarkan pengujian menggunakan algoritma LVQ, didapatkan hasil akurasi rata-rata dari 40 buah dokumen sebesar 39% [3].

Penelitian dengan kasus ekstraksi informasi pada karya tulis ilmiah menggunakan GHMM belum pernah diterapkan sama sekali, baik di indonesia ataupun diluar negeri. Pada jurnal berbahasa inggris dan pada kasus diluar ekstraksi informasi GHMM pernah diterapkan yaitu seperti pada web informasi ekstraksi dengan pengujian sebuah

halaman website dengan akurasi tertinggi yang didapatkan yaitu 85,5% [2]. Pada kasus pengenalan Gen manusia dalam DNA di dapatkan akurasi sebesar 85% [8].

Dengan demikian, penelitian ini akan membangun suatu sistem ekstraksi informasi menggunakan metode GHMM pada karya tulis ilmiah yang bertujuan untuk menguji tingkat akurasi yang dapat dihasilkan dengan menggunakan algoritma GHMM, dengan batasan-batasan yaitu mengambil cover dan abstrak dari dokumen karya tulis ilmiah skripsi pada Fakultas Teknik dan Ilmu Komputer, Program Studi Teknik Informatika, UNIKOM.

1.2. Rumusan Masalah

Berdasarkan dari latar belakang masalah yang ada, diketahui bahwa belum diterapkannya metode GHMM pada kasus ekstraksi informasi pada karya tulis ilmiah skripsi di indonesia sehingga belum diketahui akurasi yang dihasilkan jika menggunakan algoritma GHMM.

1.3. Maksud dan Tujuan

Maksud dari penelitian ini yaitu untuk membangun suatu sistem ekstraksi informasi pada karya tulis ilmiah dengan menggunakan metode GHMM.

Dan adapun tujuan dari dibuatnya penelitian ini adalah :

1. Mengukur akurasi dari penggunaan metode GHMM pada kasus ekstraksi informasi karya tulis ilmiah skripsi.
2. Menerapkan metode GHMM pada kasus ekstraksi informasi karya tulis ilmiah skripsi.

1.4. Batasan Masalah

Adapun beberapa batasan masalah yang ada dari penelitian ini yang sekiranya perlu dibatasi adalah :

1. Data Masukan

Data masukan yang digunakan dibagi menjadi 2 bagian yaitu.

a. Data Training

- Data yang digunakan yaitu dokumen karya tulis ilmiah skripsi yang ada di Universitas komputer Indonesia (UNIKOM).
- Format dari file masukannya berupa .csv.
- Format dari file .csv berisi dokumen sampul dan abstrak
- Data yang digunakan untuk training 40 buah dokumen dari tahun 2011 sampai dengan tahun 2018.

b. Data testing

- Format dari file testing berupa .txt
- Data yang digunakan yaitu dokumen karya tulis ilmiah skripsi pada Program Studi Teknik Informatika yang ada di Universitas Komputer Indonesia (UNIKOM).
- Jumlah pengujian untuk data sampul dan abstrak dari skripsi dibutuhkan sebanyak 40 buah dokumen dari tahun 2011 sampai dengan tahun 2018.

2. Proses

- a. Tahapan Preprocessing memiliki rangkaian alur yang terdiri dari tokenisasi kata, dan ekstraksi fitur.
- b. Terdapat 15 macam ekstraksi fitur yang digunakan.
- c. Pengklasifikasian dari kasus ekstraksi informasi karya tulis ilmiah ini menggunakan Algoritma Generalized Hidden Markov Model dan untuk pengujiannya menggunakan viterbi.

3. Data Keluaran

Adapun hasil dari data keluaran oleh sistem yaitu dibagi menjadi 2 bagian. Bagian yang terdapat pada lembar sampul yang berjumlah 9 kategori, dan juga yang terdapat pada lembar abstrak yang berjumlah 7 kategori. Lembar sampul dan abstrak tersebut merupakan hasil dari sistem ekstraksi informasi yang meliputi token-token dan kategori.

1.5. Metode Penelitian

Metode yang digunakan bersifat kuantitatif [7]. Penelitian ini dilakukan berdasarkan prosedur yang telah direncanakan sebelumnya sehingga terdiri dari tahapan-tahapan kegiatan. Pada metode kuantitatif diperlukannya sumber data dan pengumpulan data untuk membangun suatu sistem yang telah direncanakan. Tahapan yang dilakukan disini adalah studi literatur, pengumpulan data-data, pembangunan perangkat lunak dan pengujian



Gambar 1.1 Blok Diagram Alur Kerja Penelitian

1.5.1. Studi Literatur

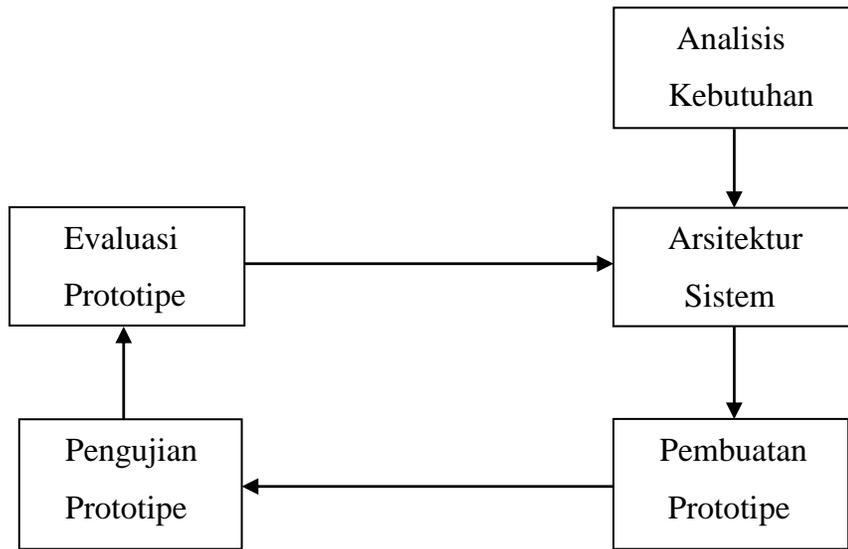
Pada penelitian ini penulis menggunakan studi literatur berupa dokumen yang digunakan sebagai acuan dari penelitian dan juga berupa jurnal-jurnal yang berkaitan tentang ekstraksi informasi, buku, dan informasi yang tersaji pada halaman website dan sumber lain yang memungkinkan untuk dijadikan bahan yang berhubungan dengan penelitian yang sedang dikerjakan.

1.5.2. Pengumpulan Dataset

Dataset yang akan dikumpulkan disini adalah dari dokumen karya tulis ilmiah skripsi yang berasal dari Program Studi Teknik Informatika di Universitas Komputer Indonesia (UNIKOM) yang mana akan diambil data dari sampul (cover) dan abstrak. Dari data sampul dan abstrak tersebut akan dikumpulkan sebanyak 40 dokumen yang akan digunakan sebagai data training dan juga akan diambil 40 dokumen yang akan digunakan sebagai data testing dan dari tahun yang berbeda-beda.

1.5.3. Pembangunan Perangkat Lunak Ekstraksi Informasi

Metode penelitian pada pembangunan perangkat lunak ekstraksi informasi ini menggunakan model prototipe [2]. Penggunaan metode prototipe dikarenakan setiap tahapan yang telah dibuat dapat dievaluasi kembali dan dirancang kembali yang nantinya tahapan yang ada dapat di ulang kembali jika tidak sesuai dengan yang diharapkan pada sistem, dan sehingga ketika diuji didapatkan sistem yang sudah benar atau sesuai. Berikut adalah gambar dari penggunaan model prototipe pada sistem yang akan dibuat.



Gambar 1.2 Blok Diagram Model Prototipe

Cakupan dari model Prototipe adalah sebagai berikut :

a. Analisis Kebutuhan

Menentukan kebutuhan analisis yang diperlukan dalam membuat sistem pada ekstraksi informasi, yaitu seperti dokumen yang diperlukan, cover dan abstrak dari dokumen karya tulis ilmiah skripsi. Menganalisis juga segala kebutuhan baik itu kebutuhan perangkat lunak, dan perangkat keras.

b. Arsitektur Sistem

Membuat arsitektur atau rancangan yang sesuai dengan kebutuhan pada proses pelatihan dan pengujian. Meliputi pembuatan DFD, diagram konteks, dan juga perancangan antarmuka dari sistem.

c. Pembuatan Prototipe

Melakukan pembuatan prototipe untuk preprocessing, training dan nantinya untuk dilanjutkan ke pengujian.

d. Pengujian Prototipe

Menguji sistem yang digunakan dengan metode GHMM, untuk mendapatkan hasil nilai akurasi.

e. Evaluasi Prototipe

Pada tahap ini akan dilakukan evaluasi prototipe terhadap sistem. Evaluasi dilakukan apabila hasil dari pengujian fungsionalitas sistem dan akurasi yang didapat belum sesuai dengan apa yang diharapkan, sehingga sistem perlu kembali disesuaikan dengan algoritma GHMM.

1.5.4. Pengujian

Pengujian yang dilakukan yaitu terhadap fungsionalitas sistem ekstraksi informasi dengan menggunakan Black-Box, dan untuk menguji tingkat akurasi yang dihasilkan dari hasil perhitungan akurasi.

1.6. Sistematika Penulisan

Sistematika penulisan laporan akhir penelitian ini disusun untuk memberikan gambaran umum tentang penelitian yang dilakukan. Adapun Sistematika penulisan tugas akhir ini pada setiap BAB nya adalah sebagai berikut.

BAB 1 PENDAHULUAN

Bab ini berisi penjelasan mengenai latar belakang masalah, rumusan masalah, maksud dan tujuan, batasan masalah, metode penelitian, serta sistematika penulisan yang berkaitan dengan penggunaan algoritma GHMM pada kasus ekstraksi informasi karya tulis ilmiah.

BAB 2 TINJAUAN PUSTAKA

Bab ini membahas landasan teori yang digunakan untuk menganalisis masalah dan mengolah data penelitian yaitu ekstraksi informasi pada karya tulis ilmiah skripsi menggunakan metode GHMM, serta teori pendukung yang berhubungan dengan teori ekstraksi informasi, dan juga mengacu pada teori serta penjelasan dari sumber yang

telah dibuat sebelumnya seperti ekstraksi fitur, tokenisasi, bahasa pemrograman yang digunakan, perangkat-perangkat pendukung lainnya, dan algoritma GHMM.

BAB 3 ANALISIS DAN PERANCANGAN SISTEM

Bab ini menjelaskan mengenai analisis dan perancangan yang akan dibangun mulai dari analisis data masukan, analisis metode GHMM yang digunakan pada karya tulis ilmiah, analisis kebutuhan, analisis sistem, preprocessing, ekstraksi fitur, fungsional dan nonfungsional, pemodelan sistem, serta perancangan antarmuka.

BAB 4 IMPLEMENTASI DAN PENGUJIAN

Bab ini merupakan penjelasan implementasi metode GHMM pada sistem ekstraksi informasi karya tulis ilmiah, setiap proses dari preprocessing sampai ekstraksi fitur yang ada pada lembar cover dan abstrak dan juga sampai pada pengujian dari cover dan abstrak yang telah diklasifikasikan.

BAB 5 KESIMPULAN DAN SARAN

Bab ini berisi hasil kesimpulan dari penelitian yang telah dilakukan dan juga berisi saran yang dapat berguna bagi peneliti selanjutnya yang ingin mengimplementasikan metode GHMM pada kasus-kasus lainnya ataupun yang ingin mengembangkan selanjutnya.