

# BAB 1

## PENDAHULUAN

### 1.1. Latar Belakang Masalah

Ekstraksi informasi merupakan suatu bidang ilmu untuk pengolahan bahasa alami, dengan cara mengubah teks tidak terstruktur menjadi informasi dalam bentuk terstruktur [1]. Karena biasanya teks ditulis secara bebas dan tidak terstruktur seperti halnya sebuah dokumen. dibutuhkan fungsi dari ekstraksi informasi yaitu untuk mencari komponen-komponen atau informasi yang ada pada sebuah dokumen. Pada penelitian ini dokumen yang digunakan adalah dokumen karya tulis ilmiah skripsi berbahasa Indonesia. Dalam dokumen karya tulis ilmiah skripsi terdapat beberapa dokumen, diantaranya dokumen cover dan abstrak yang akan diteliti pada penelitian ini.

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%. Sedangkan untuk pengujian 50 buah dokumen yang disertakan dengan tahun 2013, diperoleh akurasi rata-rata sebesar 57% [2]. 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].

Pada penelitian sebelumnya, ekstraksi informasi sudah pernah dilakukan juga pada jurnal berbahasa Inggris menggunakan metode CRF (*Conditional Random Field*). Metode CRF untuk kasus ekstraksi informasi sudah diimplementasikan dalam kasus header dan kutipan dari makalah penelitian. Pada Penelitian tersebut algoritma CRF memperoleh akurasi sebesar 98% [4]. Adapun penelitian sebelumnya, pada jurnal berbahasa Indonesia menggunakan metode CRF dengan kasus *Part of Speech Tagging*. Pada penelitian tersebut di dapat rata-rata akurasi yang dihasilkan adalah 78.20% [5]. Belum adanya penelitian ekstraksi

informasi berbahasa Indonesia menggunakan algoritma CRF sehingga dibutuhkan algoritma tersebut untuk mengetahui akurasi dari kasus ekstraksi informasi dokumen karya tulis ilmiah berbahasa Indonesia.

Berdasarkan pemaparan tersebut penelitian ini akan menggunakan algoritma CRF untuk membangun sistem ekstraksi informasi dokumen karya tulis ilmiah skripsi berbahasa Indonesia, dengan batasan dokumen yang akan digunakan pada penelitian ini adalah dokumen karya tulis ilmiah skripsi Universitas Komputer Indonesia (UNIKOM).

## **1.2. Identifikasi Masalah**

Dari yang sudah dijelaskan pada latar belakang masalah, belum adanya suatu sistem ekstraksi informasi karya tulis ilmiah berbahasa Indonesia yang menggunakan algoritma CRF (*Conditional Random Field*).

## **1.3. Maksud dan Tujuan**

Maksud dari penelitian ini adalah membangun sistem ekstraksi informasi berbahasa Indonesia dengan menggunakan algoritma CRF (*Conditional Random Field*).

Adapun tujuan dari penelitian ini, yaitu :

1. Menyediakan suatu sistem ekstraksi informasi menggunakan algoritma CRF.
2. Dapat mengukur akurasi dari algoritma *Conditional Random Field*.

## **1.4. Batasan Masalah**

Adapun batasan masalah dari penelitian ini sebagai berikut.

### **1. Data Masukan**

Adapun data masukan yang akan digunakan sebagai berikut.

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

- Data yang digunakan untuk training dari dokumen karya tulis ilmiah diambil dari cover dan abstrak sebanyak 40 buah dokumen yang tersedia 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 sampel dan abstrak dari skripsi dibutuhkan sebanyak 40 dokumen dari tahun 2011 sampai dengan 2018.

2. Proses

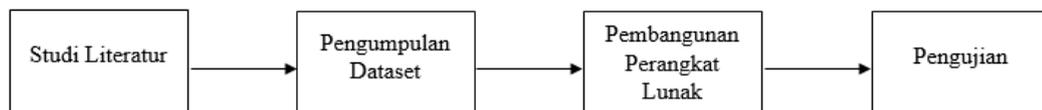
Pada proses penelitian ini ada batasan yang akan digunakan untuk ekstraksi fitur, yaitu 15 fitur [4] [3].

3. Keluaran

Data keluaran yang dihasilkan berupa token dan kategori dari token tersebut. Kategori terbagi menjadi 2 yaitu sampel dengan 9 kategori dan abstrak 7 kategori.

### 1.5. Metode Penelitian

Tahapan metode penelitian yang digunakan dalam penelitian ini adalah metode kuantitatif. Dalam penelitian ini mempunyai 4 langkah, yaitu studi literatur, pengumpulan dataset, pembangunan perangkat lunak dan pengujian dan berikut blok diagram metode penelitian.



**Gambar 1.1 Blok Diagram Metode Penelitian**

### 1.5.1. Studi Literatur

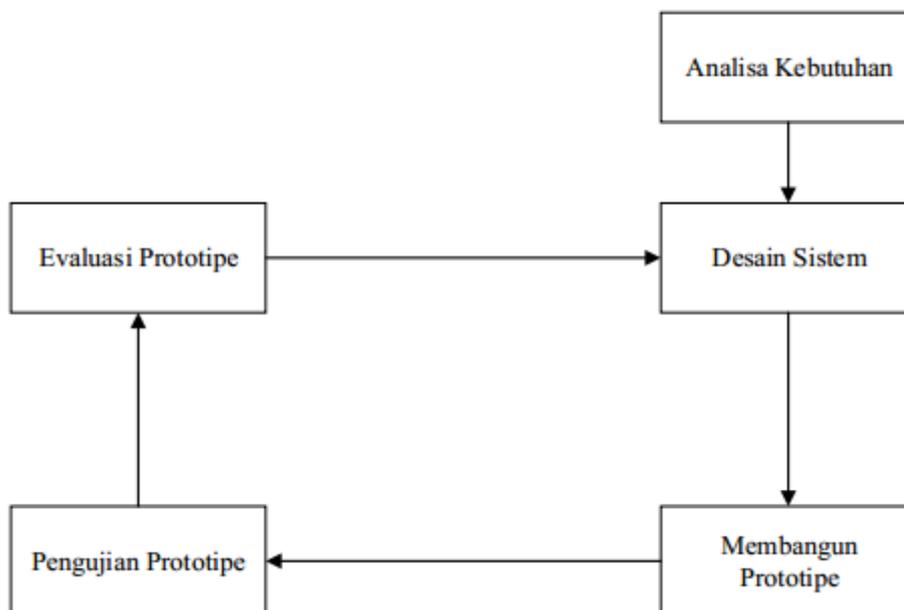
Pada tahap pertama penelitian yaitu melakukan studi literatur. Studi literatur pada penelitian ini menggunakan metode pengumpulan literatur berupa buku dan jurnal. Studi literatur dilakukan untuk mengidentifikasi masalah dan cara penyelesaian

### 1.5.2. Pengumpulan Dataset

Dataset yang digunakan untuk penelitian ini berupa dokumen karya tulis ilmiah Skripsi Universitas Komputer Indonesia diantaranya cover dan abstrak untuk data *training* dan data *testing* pada algoritma CRF.

### 1.5.3. Pembangunan Perangkat Lunak

Pada pembangunan perangkat lunak metode yang digunakan adalah dengan menggunakan model *prototipe*. Berikut model *prototyping* pada sistem ekstraksi informasi menggunakan algoritma CRF.



**Gambar 1.2 Blok Diagram Prototipe**

Berikut penjelasan dari model sistem prototipe.

#### 1. Analisis Kebutuhan

Pada proses ini, untuk membangun sistem ekstraksi informasi menggunakan algoritma CRF yang pertama dilakukan ialah menganalisis kebutuhan seperti

kebutuhan data masukan untuk algoritma CRF, kebutuhan membangun perangkat lunak dan kebutuhan perangkat keras.

## **2. Desain Sistem**

Pada proses desain, untuk pembuatan perancangan perangkat lunak yang dapat diperkirakan. Proses ini berfokus pada pembuatan perancangan antarmuka, *DFD*, dan diagram konteks.

## **3. Membangun Prototipe**

Proses yang dilakukan pada tahap ini adalah pembangunan prototipe meliputi proses *preprocessing*, proses *training* dan proses *testing* pada algoritma CRF untuk pembuatan program.

## **4. Pengujian Prototipe**

Pada proses pengujian prototipe dilakukan terhadap fungsionalitas sistem dan nilai akurasi dari algoritma CRF.

## **5. Evaluasi Prototipe**

Pada tahap ini dilakukan evaluasi prototipe terhadap sistem. Apabila sistem sesuai dengan keinginan maka sistem tersebut dapat digunakan. Akan tetapi jika sistem tersebut tidak sesuai dengan keinginan maka sistem kan disesuaikan kembali dengan algoritma.

### **1.5.4. Pengujian**

Pengujian dilakukan terhadap fungsionalitas sistem ekstraksi informasi dengan menggunakan metode *black-box* dan nilai akurasi menggunakan perhitungan akurasi [3].

## **1.6. Sistematika Penulisan**

Sistematika penulisan disusun untuk memberikan gambaran umum tentang penelitian yang dilakukan. Sistematika penulisan adalah sebagai berikut.

### **BAB 1 PENDAHULUAN**

Pada bab ini berisi penjelasan mengenai latar belakang, rumusan masalah, maksud dan tujuan, batasan masalah, metodologi penelitian, dan

sistematika penulisan untuk ekstraksi informasi dokumen karya ilmiah menggunakan algoritma CRF (*Conditional Random Field*).

## BAB 2 LANDASAN TEORI

Bab ini membahas landasan teori yang digunakan untuk menganalisis masalah dan mengolah data penelitian yaitu teori mengenai ekstraksi informasi dan CRF(Conditional Random field).

## BAB 3 ANALISIS DAN PERANCANGAN

Di bab ini diuraikan proses-proses mengenai ekstraksi informasi pada dokumen karya tulis ilmiah menggunakan metode CRF dan kebutuhan algoritma yang digunakan. analisis dan perancangan yang akan dibangun mulai dari analisis masalah, analisis data masukan, analisis metode CRF pada dokumen karya tulis ilmiah, analisis kebutuhan fungsional dan non fungsional, pemodelan sistem, perancangan antarmuka.

## BAB 4 IMPLEMENTASI DAN PENGUJIAN

Bab ini berisi tentang implementasi dari analisis yang sudah dilakukan pada bab 3 dan pengujian akan dilakukan untuk fungsionalitas dari sistem ekstraksi informasi menggunakan algoritma CRF yang sudah dibangun.

## BAB 5 KESIMPULAN DAN SARAN

Pada Bab ini berisi mengenai hasil kesimpulan dari penelitian yang sudah selesai dilakukan. serta saran yang dapat digunakan untuk pengembangan selanjutnya.