BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Ekstraksi informasi merupakan proses pengambilan informasi dari teks tidak terstruktur yang menghasilkan teks terstruktur dan rapi sekaligus mempermudah dalam menemukan informasi dari teks terstruktur [1]. Pengertian lain ekstraksi informasi menurut Piskorsi, ekstraksi informasi bertujuan untuk mengekstraksi sekumpulan data teks untuk mendapatkan fakta-fakta berkaitan dengan kejadian, entitas atau keterhubungan dalam bentuk informasi terstruktur sebagai masukan untuk basis data [2]. Dokumen karya ilmiah memiliki format yang beragam sehingga sulit untuk mendeteksi setiap komponen pada dokumen. Oleh karena itu, dibutuhkan metode untuk mengekstraksi informasi dari dokumen karya ilmiah sehingga aturan ekstraksi informasi sesuai dengan format dokumen karya ilmiah.

Penelitian sebelumnya menggunakan *rule-based* pada dokumen karya tulis ilmiah skripsi berbahasa Indonesia [3]. Pada penelitian tersebut ekstraksi informasi dilakukan dengan sistem berbasis aturan untuk mendeteksi identitas pada dokumen skripsi, diantaranya *cover,* abstrak dan *abstract.* Berdasarkan pengujian 3 buah dokumen pada tahun 2017, akurasi yang diperoleh sebesar 100% sedangkan pengujian terhadap 50 dokumen yang beragam diperoleh rata-rata akurasi 57%. Penurunan akurasi disebabkan sistem yang tidak mampu menangani dokumen dengan format beragam. Oleh karena itu, permasalah tersebut dapat diatasi jika menggunakan *machine learning*.

Pada penelitian sebelumnya, penggunaan *machine learning* di bidang ekstraksi informasi sudah dilakukan pada makalah ilmiah [1]. Pada penelitian tersebut belum ada algoritma *Maximum Entropy Markov Model* (*MEMM*) sehingga dibutuhkan algoritma tersebut. Hasil penelitian sebelumnya oleh Susan Mengel dan Yaoquin Jing penggunaan MEMM untuk ekstraksi struktur data pada halaman web memiliki *error rate* yang rendah 0.14(0.08 dengan 30 halaman web)

[4]. Shurthi S. melakukan studi pengenalan entitas bernama untuk bahasa malaysia menggunakan pos tag TnT dan metode MEMM dengan hasil akurasi 82.5% [5]. A. Nedjo et al. menggunakan MEMM untuk POS Tag otomatis pada bahasa *Oromo* dan menghasilkan akurasi 99.3% pada data latih dan 93.01% pada data uji [6].

Berdasarkan pemaparan tersebut maka dalam peneltian ini digunakan metode MEMM untuk membangun sistem ekstraksi informasi dokumen karya ilmiah berbahasa Indonesia, dengan batasan dokumen yang akan digunakan pada penelitian ini adalah dokumen karya ilmiah skripsi Program Studi Teknik Informatika Universitas Komputer Indonesia.

1.2 Identifikasi Masalah

Berdasarkan penjelasan yang sudah diuraikan pada latar belakang, maka identifikasi masalah dalam penelitian ini adalah dibutuhkan hasil akurasi dari sistem ekstraksi informasi menggunakan MEMM untuk dokumen karya ilmiah.

1.3 Maksud dan Tujuan

Maksud dari penelitian ini adalah membangun sistem ekstraksi informasi dokumen karya ilmiah menggunakan algoritma *Maximum Entropy Markov Model*. Sementara, tujuan dari penelitian ini adalah dapat mengukur akurasi algoritma *Maximum Entropy Markov Model* untuk mendeteksi kategori pada dokumen karya ilimiah yang beragam.

1.4 Batasan Masalah

Batasan masalah yang ada dalam penelitian ini meliputi:

1. Input
    a. Dokumen yang digunakan sebagai masukan adalah hasil *scan* dari sampul dan abstrak skripsi bahasa Indonesia dengan ekstensi *.pdf kemudian dikonversi *file *.csv*.
    b. Dokumen yang digunakan diambil dari dokumen skripsi Program Studi Teknik Informatika UNIKOM secara acak.
    c. Data latih menggunakan 40 dokumen dari tahun 2011 sampai 2018.

d. Data uji menggunakan 40 dokumen dari tahun 2011 sampai 2018.

2. Proses
   a. *Preprocessing* yang akan dilakukan adalah Tokenisasi dan Ekstraksi Fitur.
   b. Ekstraksi fitur berjumlah 15 fitur berdasarkan 3 sub fitur utama yaitu fitur lokal, fitur tata letak dan fitur *named entity*. Fitur lokal terdiri dari 10 fitur, fitur tata letak terdiri dari 3 fitur dan fitur *named entity* terdiri dari 2 fitur [3].
   c. Metode yang digunakan untuk klasifikasi adalah *maximum entropy markov model* dan untuk *decoding* sebagai pengujiannya dengan *viterbi*.
3. Output
   a. Kelas target yang dihasilkan 17 kelas yang terdiri dari judul penelitian (sampul), jenis penelitian, kalimat pengajuan, penulis (sampul), nim (sampul), prodi, fakultas, universitas, kota, tahun, jenis halaman, judul penelitian (abstrak), isi abstrak, penulis (abstrak), nim (abstrak), isi abstrak, *others dan* kata kunci.
   b. Hasil akurasi dengan *f-measure, precision* dan *recall.*

1.5 Metode Penelitian

Metode yang digunakan pada penelitian ini adalah metode penelitian eksperimental [7], karena penelitian ini sesuai dengan ciri metode yaitu memiliki varibel yang berkarakter dan sebab akibat dari masing-masing variabel.

Gambar 1.1 Langkah-langkah Penelitian

1.5.1 Identifikasi Masalah

Menentukan objek penelitian dan permasalahan yang terjadi berdasarkan hasil pengamatan atau pencarian penelitian sebelumnya.

1.5.2 Studi Literatur

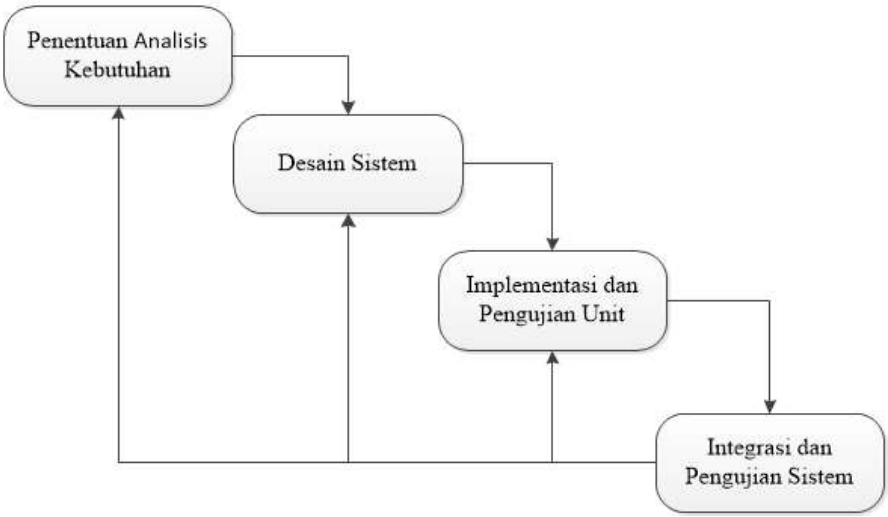
Melakukan pengumpulan informasi seperti jurnal, *paper* dan buku referensi mengenai topik yang akan dijadikan penelitian seperti algoritma *Maximum Entropy Markov Model*, ekstraksi informasi dan karya ilmiah sebagai objek penelitian.

1.5.3 Pengumpulan Data

Pengumpulan data hasil scan halaman depan dan abstrak dari skripsi Teknik Informatika UNIKOM secara acak kemudian disimpan dalam format *.pdf.*

1.5.4 Pembangunan Sistem Ekstraksi Informasi

Metode pembangunan sistem yang digunakan untuk penelitian ini menggunakan model *waterfall sommerville* [8] yang melakukan pendekatan secara sistematis dan berurutan dalam pembangunan perangkat lunak yang dirubah sesuai dengan kebutuhan penelitian meliputi proses sebagai berikut [8].



Gambar 1.2 Diagram Waterfall [9]

a. Penentuan Analisis Kebutuhan

Pada tahap ini dilakukan pengumpulan kebutuhan penelitian secara keseluruhan. Data masukan hasil *scan* halaman sampul dan abstrak skripsi Teknik Informatika dalam bentuk *.pdf* kemudian dikonversi ke format *.csv*. Batasan penggunaan *library* pendukung seperti *Tesseract* dan *pdfBox*, hasil akurasi perhitungan, rumus algoritma *MEMM* sesuai dengan hasil studi literatur untuk diterapkan dalam sistem dan mennetukan perangkat lunak yang digunakan untuk membuat sistem. Analisis yang dilakukan pada penelitian ini adalah analisis data masukan, analisis *preprocessing*, analisis pelatihan dan pengujian *MEMM*, analisis rencana pengujian, analisis ekstraksi informasi melalui alur diagram *flowchart*, analisis kebutuhan perangkat keras, analisis kebutuhan perangkat lunak dan analisis kebutuhan pengguna.

b. Desain Sistem

Perancangan sistem berdasarkan hasil analisis menggunakan perangkat pemodelan sistem *Unified Modeling Language* (UML), perancangan struktur menu, perancangan data, perancangan antar muka dan perancangan jaringan semantik.

c. Implementasi dan Pengujian Unit

Mengimplementasikan perancangan sistem dalam bentuk bahasa pemrograman diantaranya tahap *preprocessing* (tokenisasi), tahap ekstraksi fitur dan tahap pelatihan dengan *maximum entropy markov model*. Pada penelitian ini menggunakan bahasa pemrograman Java.

d. Integrasi dan Pengujian Sistem

Simulator siap digunakan untuk diuji. Pengujian dilakukan sesuai metode yang diterapkan dan fungsi dapat berjalan dengan baik tanpa kesalahan.

1.5.5 Pengujian

Tahap pengujian mengukur nilai akurasi algoritma *MEMM* menggunakan nilai *f-measure*, *precision* dan *recall*.

1.5.6 Penarikan Kesimpulan

Tahap penarikan kesimpulan mengukur nilai akurasi dan fungsionalitas sistem ekstraksi informasi dengan menggunakan metode *black-box*.

1.6 Sistematika Penulisan

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

BAB 1 PENDAHULUAN

Bab ini menguraikan tentang latar belakang dari penelitian ekstraksi informasi karya ilmiah. Maksud dan tujuan dari penelitian ini memberikan informasi yang relevan setelah dilakukan ekstraksi informasi pada karya ilmiah. Batasan Masalah dalam melakukan penelitian, metode penelitian dan sistematika penulisan

BAB 2 TINJAUAN PUSTAKA

Bab ini membahas tentang teori dan metode yang digunakan pada penelitian. Pembahasan dimulai dengan penjelasan ekstraksi informasi, karya ilmiah, metode untuk klasifikasi, tahap pengujian klasifikasi, penjelasan mengenai struktur pemrograman dan bahasa pemrograman yang digunakan.

BAB 3 ANALISIS DAN PERANCANGAN SISTEM

Pada bab ini menjelaskan analisis masalah, analisis sistem yang terdiri dari analisis data masukan analisis tahap preprocessing, ekstraksi fitur, analisis klasifikasi menggunakan algoritma *maximum entropy markov model* dengan fitur informasi fitur lokal, fitur tata letak dan fitur *named entity* serta analisis rencana pengujian. Selain itu dijelaskan analisis kebutuhan sistem terdiri dari analisis kebutuhan fungsional dan non-fungsional.

BAB 4 IMPLEMENTASI DAN PENGUJIAN SISTEM

Bab ini menjelaskan sistem ekstraksi dokumen karya ilmiah menggunakan algoritma *maximum entropy markov model*, yang meliputi proses ekstraksi informasi lembar sampul dan abstrak sampai hasil klasifikasi, perancangan antarmuka, serta pengujian terhadap klasifikasi yang dihasilkan oleh proses ekstraksi informasi.

BAB 5 KESIMPULAN DAN SARAN

Bab ini membahas hasil akhir dari penelitian dan bagaimana tindak lanjut untuk pengembangan penelitian berikutnya.