BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Dokumen karya tulis ilmiah memiliki banyak informasi dengan format dokumen yang beragam. Hal itu menyebabkan manusia susah mencari setiap komponen satu persatu secara manual sehingga dibutuhkan ekstraksi informasi untuk mengklasifikasi komponen-komponen yang ada pada dokumen karya tulis ilmiah. Ekstraksi informasi adalah suatu proses untuk mengubah informasi tidak terstruktur yang terdapat dalam teks ke dalam data terstruktur. Mendeteksi setiap komponen pada dokumen karya tulis ilmiah sulit dilakukan untuk dokumen yang formatnya beragam, oleh karena itu dibutuhkan pembuatan aturan ekstraksi yang sesuai untuk masing-masing dokumen.

Penelitian sebelumnya yang terkait dengan ekstraksi informasi pada skripsi telah dilakukan oleh Firdamdam [1]. Penelitian dilakukan untuk mengekstraksi cover dan abstrak pada dokumen skripsi. Berdasarkan pengujian untuk 40 dokumen diperoleh akurasi sebesar 78%. Penelitian ini menggunakan algoritma LVQ klasifikasi dan masih menggunakan *ruled based* untuk meningkatkan akurasi hasil klasifikasi yang dilakukan oleh algoritma LVQ. Permasalahan tersebut dapat diatasi dengan menggunakan *full automatic training* tanpa harus membuat aturan perbaikan.

CNN menjadi salah satu algoritma yang memiliki akurasi tinggi pada klasifikasi teks sebagaimana penelitian yang telah dilakukan oleh Zhiquan Wang dengan mengkomparasikan algoritma KNN, SVM dan CNN. Pada penelitian tersebut didapat CNN lebih unggul dengan akurasi 87% [2]. Sama halnya dengan penelitian yang dilakukan oleh March dan Jugal pada ekstraksi informasi untuk mendeteksi frasa Bahasa Inggris yang akan dikelompokkan dengan persamaan katanya dengan mengkomparasikan algoritma RNN, SVM, FCM dan CNN dimana CNN F1-score terunggul yaitu 82,7% [3]. Sedangkan penelitian yang dilakukan oleh Guru dan Ratneshwer untuk menguji performa algoritma *Deep learning* dengan komparasi algoritma ANN, CNN, SOM, LVQ, dan multiLVQ didapat bahwa CNN memiliki akurasi paling tinggi dengan akurasi sebesar 91,07% [4].

Oleh karena itu, CNN memiliki kemungkinan memperoleh akurasi lebih baik dibandingan algoritma lainnya jika diimplementasikan dalam ekstraksi informasi. Dengan demikian pada penelitian ini akan dilakukan penelitian tentang ekstraksi informasi pada dokumen karya tulis ilmiah menggunakan algoritma *Convolutional Neural Network* (CNN).

1.2 Identifikasi Masalah

Berdasarkan latar belakang diatas maka masalah dapat diidentifikasi yaitu penelitian ekstraksi informasi sebelumnya yang menggunakan algoritma LVQ masih menggunakan *ruled based* untuk memperbaiki hasil klasifikasinya.

1.3 Maksud dan Tujuan

Maksud dari penelitian ini adalah membangun suatu sistem ekstraksi informasi dokumen karya tulis ilmiah menggunakan algoritma CNN. Adapun tujuan dari penelitian ini adalah untuk mengukur akurasi algoritma CNN pada kasus ekstraksi informasi dokumen karya tulis ilmiah dengan format yang beragam.

1.4 Batasan Masalah

Batasan masalah pada penelitian ini adalah:

1. Data Masukan

   Data masukan meliputi:

   a. Data *Training*

      1. Format masukan .csv
      2. Terdiri dari lembar sampul dan abstrak (Skripsi Program Studi Teknik Informatika, Universitas Komputer Indonesia).
      3. Jumlah lembar sampul dan abstrak skripsi untuk data *training* sebanyak 60 dokumen dari tahun 2011 sampai dengan 2018.

   b. Data *Testing*

      1. Format file masukan .txt
      2. Terdiri dari lembar sampul dan abstrak (Skripsi Program Studi Teknik Informatika, Universitas Komputer Indonesia).

3. Jumlah lembar sampul dan abstrak skripsi untuk data *testing* sebanyak 20 dokumen dari tahun 2011 sampai dengan 2018.

2. Proses
    1. Setiap data pada komponen lembar sampul dan abstrak ditokenisasi per kata, kumpulan angka, dan simbol.
    2. Ekstraksi fitur yang digunakan sebanyak 15 fitur yaitu *initcaps, allcaps, containsdigit, alldigit, containsdots, lowercase, punctuation, eightdigit, word, line_start, line_in, line_end, person, organization, dan year*.
3. Data Keluaran

Data keluaran yang dihasilkan oleh sistem ekstraksi informasi meliputi token dengan kategori pada lembar sampul sebanyak 9 kategori dan abstrak sebanyak 7 kategori.

1.5 Metode Penelitian

Pada penelitian ini terdapat empat tahapan alur kerja yaitu:

1. Studi Literatur

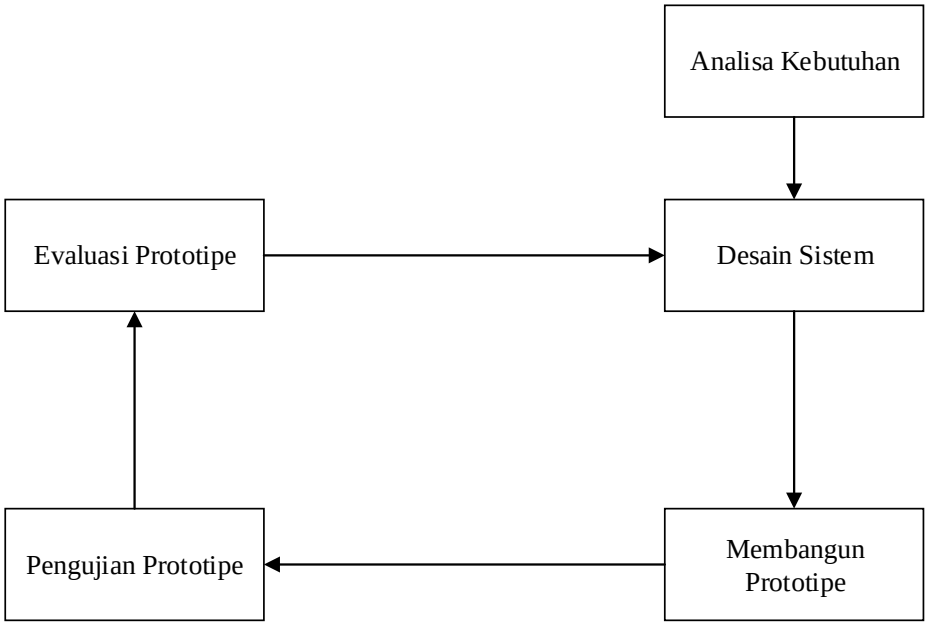
   Studi literatur digunakan untuk mengumpulkan berbagai bahan referensi mengenai Ekstraksi Informasi, NLP dan *Convolutional Neural Network*.

2. Pengumpulan dataset

   Pengumpulan dataset untuk mengumpulkan dokumen karya tulis ilmiah skripsi Program Studi Teknik Informatika, Universitas Komputer Indonesia.

3. Pembangunan perangkat lunak

   Metode pembangunan perangkat lunak dalam penelitian ini adalah menggunakan model *protopyping*. Model *prototyping* dipilih karena perlu diketahuinya keberhasilan sistem dari evaluasi yang dilakukan. Jika *prototipe* belum sesuai dengan yang diharapkan, maka akan dilakukan perbaikan terhadap algoritma CNN yang diimplementasikan pada sistem tersebut.



Gambar 1.1 Diagram Model *Prototyping* [5]

4. Pengujian

   Pengujian dilakukan untuk menghitung nilai akurasi algoritma CNN dengan menguji *epoch* dan data masukan yang telah diseimbangkan oleh *SMOTE*.

1.6 Sistematika Penulisan

Sistematika penulisan penelitian ini disusun untuk memberikan gambaran umum mengenai penelitian yang dikerjakan. Sistematika penulisan penelitian sebagai berikut:

BAB 1 PENDAHULUAN

Bab ini berisi penjelasan mengenai latar belakang permasalahan, identifikasi masalah, maksud dan tujuan, batasan masalah, metodologi penelitian, dan sistematika penulisan ekstraksi informasi karya tulis ilmiah menggunakan algoritma *convolutional neural network*.

BAB 2 TINJAUAN PUSTAKA

Bab ini berisi berbagai konsep dan teori-teori para ahli yang berkaitan dengan topik penelitian ekstraksi informasi dokumen karya tulis ilmiah menggunakan algoritma *convolutional neural network*.

BAB 3 ANALISIS DAN PERANCANGAN SISTEM

Bab ini berisi penjelasan tentang analisis masalah, analisis sistem, data masukan, preprocessing, ekstraksi fitur, algoritma CNN, fungsional dan non fungsional serta perancangan antar muka.

BAB 4 IMPLEMENTASI DAN PENGUJIAN

Bab ini berisi implementasi dan pengujian. Implementasi meliputi implementasi perangkat lunak, implementasi perangkat keras, dan implementasi antarmuka. Pengujian pada bab ini juga berupa pengujian fungsionalitas system dan nilai akurasi ekstraksi informasi karya tulis ilmiah menggunakan algoritma CNN.

BAB 5 KESIMPULAN DAN SARAN

Bab ini berisi hasil dari penelitian ekstraksi informasi dokumen karya tulis ilmiah menggunakan algoritma CNN meliputi kesimpulan dan saran untuk penelitian selanjutnya.