BAB I
PENDAHULUAN

1.1 Latar Belakang

Seiring dengan pesatnya perkembangan teknologi internet, peningkatan jumlah teks elektronik pun terjadi secara pesat. Begitu pun dengan perkembangan signifikan yang terjadi pada dokumen-dokumen ilmiah seperti jurnal ilmiah, tesis, disertasi, dan karya-karya ilmiah lainnya [17]. Oleh karena itu, kebutuhan akan pemrosesan informasi yang efektif dan efisien baik dari segi waktu maupun akurasi pemrosesan informasi pada dokumen-dokumen ilmiah tersebut juga meningkat.

Ekstraksi kata kunci merupakan salah satu cara untuk memproses informasi pada suatu dokumen teks. Kata kunci adalah kumpulan kata penting dalam sebuah dokumen yang memberikan gambaran mengenai isi dari dokumen tersebut [17]. Proses ekstraksi kata kunci pada dasarnya bertujuan untuk mendapatkan gambaran singkat mengenai isi suatu dokumen yang direpresentasikan oleh kumpulan kata kunci tanpa harus memproses dokumen tersebut secara keseluruhan.

Karakteristik kata kunci pada tiap dokumen berbeda-beda dan cenderung dipengaruhi beberapa aspek. Salah satu aspek tersebut ialah aspek linguistik yang melandasi penulisan dokumen tersebut. Kata kunci pada dokumen berbahasa Indonesia memiliki karakteristik tersendiri seperti halnya kata kunci pada dokumen berbahasa lain. Kata kunci pada bahasa Indonesia umumnya berada pada awal dan akhir paragraf [7]. Hal ini berbeda dengan kata kunci pada bahasa Turki yang umumnya berada pada bagian tengah paragraf [17]. Selain itu, perbedaan afiliasi suatu dokumen juga dapat mempengaruhi perbedaan karakteristik kata kunci. Sebagai contoh, dokumen-dokumen jurnal ilmiah yang berasal dari disiplin ilmu eksakta memungkinkan adanya kata kunci yang terdiri dari karakter numerik sebagaimana halnya dokumen-dokumen jurnal non eksakta memungkinkan keberadaan kata kunci yang merupakan sebuah frasa. Perbedaan-perbedaan karakteristik kata kunci inilah yang menjadi tantangan dalam proses ekstraksi kata kunci.

Conditional Random Field (CRF) merupakan model probabilistik yang memetakan distribusi probabilitas kondisional dari kumpulan keluaran berupa label yang mungkin untuk suatu sekuens data (observasi) [6]. Dengan memetakan distribusi probabilitas kondisional secara keseluruhan dibandingkan memetakan distribusi probabilitas kondisional pada tiap state, CRF dapat mengoptimalkan proses-proses yang berinteraksi dengan sekuens data. Hal ini membuat CRF banyak diimplementasikan pada proses-proses yang berkaitan dengan segmentasi, pengenalan pola, dan pelabelan sekuens data.

Proses ekstraksi kata kunci pada dasarnya dapat dipandang sebagai proses pelabelan pada *term-term* yang terdapat dalam suatu sekuens *term* yakni dokumen. Ekstraksi kata kunci dilakukan dengan mengekstrak *term-term* yang

berlabel kata kunci. Dengan demikian CRF dapat diimplementasikan untuk memodelkan proses ekstraksi kata kunci dengan menganalogikan proses ekstraksi kata kunci sebagai proses pelabelan sekuens *term* pada suatu dokumen.

1.2 Perumusan Masalah

Permasalahan pada Tugas Akhir ini dirumuskan ke dalam beberapa hal yakni:

1. Bagaimana model CRF dapat mengekstraksi kata kunci dari suatu dokumen teks?
2. Bagaimana menentukan kumpulan parameter pendukung optimal pada model CRF untuk dapat mengekstraksi kata kunci seakurat mungkin?
3. Bagaimana pengaruh jumlah data latih, pemilihan fitur, dan jumlah fitur yang dilibatkan terhadap akurasi ekstraksi kata kunci yang dihasilkan oleh model CRF?
4. Bagaimana pengaruh eliminasi *stopwords* terhadap performansi model CRF dalam mengekstraksi kata kunci?
5. Bagaimanakah pengaruh homogenitas kelompok dokumen terhadap performansi ekstraksi kata kunci dengan model CRF?

Adapun batasan masalah pada pengerjaan Tugas Akhir ini adalah sebagai berikut:

1. Model CRF yang digunakan adalah *Linear-chained CRF Model*.
2. Dokumen teks yang digunakan sebagai data latih dan data uji pada Tugas Akhir ini adalah abstrak jurnal ilmiah berbahasa Indonesia.
3. Jurnal Ilmiah yang digunakan pada Tugas Akhir ini ialah jurnal ilmiah pada bidang Kedokteran dan Kesehatan Masyarakat.
4. *Term* yang digunakan pada data latih dan data uji adalah kata.
5. Atribut *term* yang digunakan untuk mengekstraksi kata kunci dibatasi hanya pada atribut statistik *term* tersebut.

Selain itu, dugaan awal yang akan penulis uji pada Tugas Akhir ini yaitu:

1. CRF dapat mengekstraksi kata kunci secara umum tanpa membatasi ukuran kata kunci (*n-gram keywords*).
2. CRF dapat mengekstraksi kata kunci pada dokumen dengan tingkat homogenitas yang berbeda-beda dengan performansi yang relatif sama.

1.3 Tujuan

Tujuan dari Tugas Akhir ini adalah sebagai berikut:

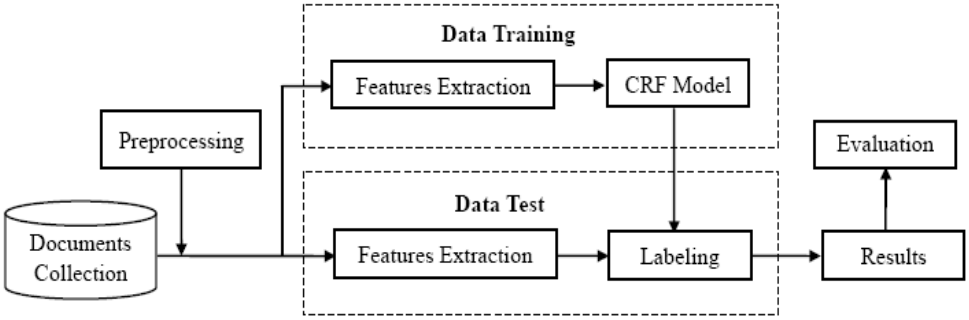
1. Mengimplementasikan CRF untuk mengekstraksi kata kunci pada jurnal ilmiah berbahasa Indonesia.
2. Menganalisis pengaruh jumlah data latih, pemilihan dan jumlah fitur yang digunakan, serta keterlibatan eliminasi *stopwords* terhadap akurasi ekstraksi kata kunci yang dihasilkan oleh model CRF.
3. Menganalisis performansi CRF dalam mengekstraksi kata kunci pada dokumen dengan tingkat homogenitas yang berbeda.

1.4 Metodologi Penyelesaian Masalah

Metodologi yang digunakan dalam memecahkan masalah di atas adalah dengan menggunakan langkah-langkah berikut:

1. Studi literatur
   Pencarian referensi dan sumber-sumber yang berhubungan dengan implementasi CRF pada berbagai bidang khususnya pada proses pelabelan dan segmentasi yang relevan dengan topik ekstraksi kata kunci.
2. Pengumpulan data
   Pengumpulan data berupa dokumen-dokumen jurnal ilmiah berbahasa Indonesia pada bidang kedokteran dan kesehatan masyarakat yang akan menjadi input pada proses pelatihan dan pengujian.
3. Analisis dan perancangan sistem
   Melakukan analisis dan perancangan terhadap sistem yang dibangun, menganalisis metode yang akan digunakan untuk menyelesaikan permasalahan, termasuk menentukan *platform* bahasa atau alat bantu pemrograman yang digunakan. Selain itu juga dilakukan analisis mengenai relasi antara parameter-parameter yang menjadi input terhadap kecenderungan perilaku sistem pada proses ekstraksi kata kunci sekaligus mengetahui seberapa signifikan parameter-parameter tersebut berpengaruh baik secara langsung maupun tidak langsung dalam proses ekstraksi kata kunci.
4. Implementasi dan pembangunan sistem
   Secara umum prosedur yang dilakukan pada penelitian ini adalah :
   1. Proses pengumpulan dan pengelompokan dokumen jurnal ilmiah. Dokumen-dokumen jurnal ilmiah tersebut akan dikelompokkan sebagai data latih dan data uji. Bagian yang akan diperhitungkan hanyalah bagian abstrak dari dokumen-dokumen latih tersebut.
   2. Melakukan *cleaning* yaitu pemisahan paragraf, kalimat, dan kata yang terdapat pada bagian abstraksi dokumen latih. Selain itu dilakukan *punctuation removal* yakni pembersihan tanda baca agar tanda baca tersebut tidak dikategorikan sebagai *term*. Kata kunci yang berasal dari pakar pada abstrak dokumen tersebut disimpan.
   3. Melakukan ekstraksi fitur-fitur dari tiap *term* berupa kata-kata pada abstraksi dokumen latih.
   4. Melakukan estimasi parameter dengan memaksimalkan *log-likelihood* yakni suatu besaran yang merepresentasikan seberapa besar tingkat akomodasi parameter oleh data latih. Estimasi parameter bertujuan untuk mendapatkan parameter fitur yang paling optimal untuk membangun model distribusi kondisional CRF yang paling optimal juga.
   5. Melakukan pengujian terhadap dokumen uji berdasarkan model distribusi kondisional CRF yang telah dihasilkan dengan cara melabeli tiap kata pada dokumen uji dengan label yang bersesuaian berdasarkan model distribusi kondisional tersebut.

Secara umum, prosedur ekstraksi kata kunci dengan CRF dapat digambarkan seperti gambar di bawah [20] :



Gambar 1 Proses Ekstraksi Kata Kunci dengan CRF Model [20]

5. Pengujian dan analisis

   Pengujian dan analisis dilakukan oleh penulis tanpa melibatkan pihak lain, meliputi:

   a. Melakukan perubahan jumlah dokumen latih untuk mengetahui pengaruh jumlah dokumen latih terhadap performansi model CRF yang dihasilkan dalam mengekstraksi kata kunci.
   b. Membandingkan data latih dengan eliminasi dan tanpa eliminasi *stopwords* untuk mengetahui pengaruh keberadaan *stopwords* terhadap performansi model CRF yang dihasilkan dalam mengekstraksi kata kunci.
   c. Membandingkan perubahan baik pendekatan ekstraksi maupun jumlah fitur yang dilibatkan untuk mempengaruhi seberapa besar fitur mempengaruhi akurasi ekstraksi kata kunci baik.
   d. Membandingkan performansi CRF dalam mengekstraksi kata kunci pada dokumen Kedokteran dengan dokumen Kesehatan Masyarakat yang memiliki tingkat homogenitas konten yang berbeda.
   e. Pengukuran akurasi ekstraksi kata kunci melalui perhitungan *precision, recall,* dan *F-measure* dengan membandingkan kata kunci yang dihasilkan dengan CRF dan kata kunci yang dihasilkan oleh pakar.

6. Pengambilan kesimpulan dan penyusunan laporan Tugas Akhir.