BAB 1
PENDAHULUAN

1.1. Latar Belakang Masalah

Novel merupakan karya sastra yang banyak digemari oleh masyarakat terutama di Indonesia. Di dalam novel mengandung rangkaian cerita kehidupan seseorang dan orang-orang di sekelilingnya [2]. Banyaknya tokoh dalam cerita tersebut, membuat adanya kesulitan untuk membedakan siapa yang sedang dibahas dalam cerita tersebut. Untuk mengatasi hal ini dibutuhkan *Coreference Resolution* untuk mengidentifikasi kesetaraan antar entitas, seperti kata ganti dan entitas yang merupakan kata benda nama seseorang. Proses ini dilakukan untuk menemukan apakah dua ungkapan dalam *natural language* merujuk ke entitas yang sama [1]. *Coreference Resolution* merupakan subtugas dari *Natural Language Processing* (NLP) yang bertugas untuk mengidentifikasi kesetaraan antar entitas.

Coreference resolution sudah banyak diteliti dalam berbagai macam bahasa seperti bahasa Inggris [3], bahasa China [1], bahasa Indonesia [4], bahkan dalam dokumen yang bersifat biomedis [5]. Penelitian mengenai coreference resolution pada bahasa Indonesia sudah pernah dilakukan sebelumnya. Pada penelitiannya menggunakan tiga metode diantaranya Decision Tree, Maximum Entropy dan Association Rules pada kasus berita koran. Penelitian yang dilakukannya menunjukkan bahwa metode decision tree dan maximum entropy membutuhkan batas jumlah dokumen tertentu, sehingga hasil yang diberikan dapat maksimal. Jika jumlah dokumen kurang dari batas, maka terdapat kondisi yang tidak terakomodasi dalam sistem. Sedangkan jika jumlah dokumen yang digunakan terlalu banyak maka akan terdapat *noise* yang tidak perlu dipakai dalam membangun model. Sehingga hasil yang diberikan tidak maksimal. Sebaliknya, metode association rules memberikan hasil yang lebih bagus dengan jumlah dokumen yang lebih banyak. Namun yang perlu diperhatikan adalah jumlah dokumen minimal yang menyebabkan metode ini tidak dapat memberikan hasil yang maksimal. Tidak perlu

memberikan sebanyak mungkin dokumen pelatihan, karena nilai maksimal akan dicapai pada suatu jumlah dokumen tertentu [4].

Metode *Support Vector Machine* (SVM) untuk kasus *coreference resolution* sudah pernah dilakukan sebelumnya, seperti halnya penelitian yang dilakukan oleh Gao Junwei dan para peneliti lainnya yang meneliti tentang "Coreference resolution pada *noun phrase*". Dalam penelitian ini untuk menghasilkan data *train* dan data *test* digunakan metode seperti yang pernah dilakukan oleh Soon et al (2001). Dimana data diklasifikasikan menjadi dua kelas yaitu *coreference* dan *not-coreference*. Dalam penelitiannya, bahasa yang digunakan adalah bahasa China [1]. Metode *Support Vector Machine* (SVM) memiliki kelebihan *course of dimensionality* yaitu jika metode itu "terpaksa" bekerja dengan keterbatasan data latih karena kendala tertentu. Vapnik membuktikan bahwa tingkat generalisasi yang diperoleh metode SVM tidak dipengaruhi oleh dimensi dari *input vector*. Selain itu metode SVM ini bekerja lebih cepat dibandingkan dengan ketiga metode di atas [6]. Namun belum pernah dilakukan penelitian mengenai *coreference resolution* pada bahasa Indonesia dengan menggunakan metode SVM sehingga belum diketahui nilai akurasinya.

Berdasarkan uraian di atas, penelitian ini akan menggunakan dokumen Novel Berbahasa Indonesia sebagai input dan menggunakan metode *Support Vector Machine* (SVM) untuk *Coreference Resolution*.

1.2. Identifikasi Masalah

Identifikasi masalah menurut latar belakang di atas adalah belum adanya nilai akurasi untuk menentukan apakah SVM tepat digunakan pada *coreference resolution* dengan kasus novel berbahasa Indonesia.

1.3. Maksud dan Tujuan

Penelitian ini bermaksud untuk membangun pendeteksi *Coreference Resolution* yang tepat pada dokumen. Adapun tujuan yang ingin dicapai dari penelitian ini adalah untuk mengetahui nilai akurasi *coreference resolution* dengan metode *Support Vector Machine* (SVM).

1.4. Batasan Masalah

Pembatasan ruang lingkup pekerjaan yang akan dilakukan terhadap permasalahan yang ada dalam penelitian adalah sebagai berikut.

a. Masukan

Berikut adalah batasan masalah untuk data masukan.

1. Jenis dokumen teks adalah dokumen yang berformat .txt
2. Dokumen yang digunakan untuk tahap *training* merupakan novel berbahasa Indonesia dan data label sebagai kunci.
3. Dokumen yang digunakan untuk tahap *testing* merupakan novel berbahasa Indonesia.
4. Dokumen masukan untuk *training* bersumber dari novel karya sastrawan Triyanto Triwikromo yang berjudul "Surga Sungsang" [7] dan novel karya sastrawan Donny Dhirgantoro yang berjudul "5 cm" [8].
5. Dokumen masukan untuk *testing* bersumber dari beberapa novel karya sastrawan Indonesia.

b. Proses

Berikut adalah batasan masalah untuk proses.

1. Sistem hanya mendeteksi kata yang berupa kata benda nama seseorang dan kata ganti untuk orang.
2. Tahap p*reprocessing* yang dilakukan adalah *case folding, filtering, tokenizing, stemming,* POS *tagging.*
3. Pada tahap ekstraksi fitur digunakan enam fitur diantaranya adalah *i-pron, i-prop, j-pron, j-prop, distance of words* (DOW), *distance of sentence* (DOS) dan *string_match.*
4. Klasifikasi dilakukan dengan menggunakan *library Support Vector Machine* (LIBSVM).

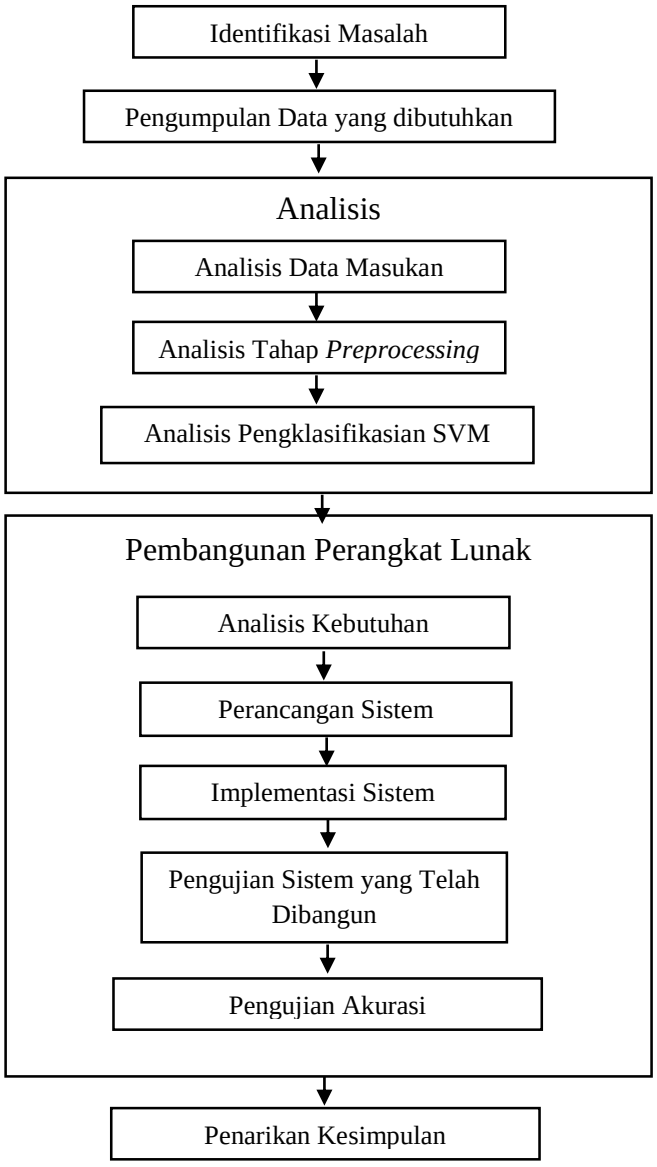
c. Keluaran

Berikut adalah batasan masalah untuk data keluaran.

1. Pasangan kata dengan hasil klasifikasi *coreference* atau bukan *coreference.*
2. Nilai akurasi dari hasil klasifikasi.

1.5. Metodologi Penelitian

Metodologi penelitian yang digunakan dalam penelitian ini adalah metode deskriptif. Karena penelitian ini menggambarkan secara sistematis fakta dan karakteristik objek dan subjek yang diteliti secara tepat [9]. Alur penelitian dapat dilihat pada gambar berikut.



Gambar 1. 1 Skema Metodologi Penelitian Kuantitatif

1.5.1. Metode Pengumpulan Data

Pada penelitian ini metode pengumpulan data yang digunakan adalah studi literatur yang diperoleh dari sumber-sumber tertulis baik yang tercetak maupun elektronik, seperti buku, *e-book*, jurnal, *paper*, dan sumber-sumber lain yang berhubungan dengan *coreference resolution*.

1.5.2. Metode Pembangunan Perangkat Lunak

Metode pembangunan perangkat lunak yang digunakan adalah *Sequential Linear Model* [10]. Metode ini digunakan karena cocok untuk pembangunan sistem *Coreference Resolution* yang diawali dengan analisis dan pemodelan sistem informasi kemudian diakhiri dengan pengujian. Adapun tahapan yang akan dilakukan adalah sebagai berikut.

a. Analisis kebutuhan

Analisis kebutuhan meliputi penentuan perangkat lunak, penentuan perangkat keras dan analisis data masukan yaitu novel berbahasa indonesia untuk mengetahui tahap *preprocessing* yang sesuai.

b. Desain sistem dan perangkat lunak

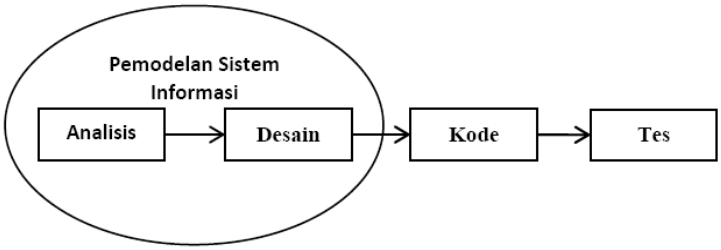
Langkah awal yang dilakukan adalah membuat rancangan sistem secara keseluruhan, meliputi skema *preprocessing*, *training*, dan *testing*.

c. Implementasi

Tahap ini merupakan tahap pembuatan aplikasi sesuai dengan desain yang telah dibuat pada tahap perancangan.

d. Pengujian

Pengujian yang dilakukan meliputi pengujian metode SVM untuk mengetahui pengaruhnya terhadap akurasi *coreference resolution.*



Gambar 1. 2 Model *Sequential Linear Model* [10]

1.6. Sistematika Penulisan

Sistematika penulisan laporan akhir penelitian ini disusun untuk memberikan gambaran umum tentang penelitian yang dijalankan. Sistematika penulisan tugas akhir ini adalah sebagai berikut.

BAB 1 PENDAHULUAN

Bab ini berisi penjelasan mengenai latar belakang masalah, identifikasi masalah, maksud dan tujuan, batasan masalah, metodologi penelitian serta sistematika penulisan dari penelitian ini.

BAB 2 LANDASAN TEORI

Bab ini membahas mengenai teori-teori yang digunakan untuk menganalisis masalah dan teori yang dipakai dalam data penelitian yaitu teori mengenai *coreference resolution, pre-processing* (*case folding, filtering, stemming, tokenizing, POS tagging*), ekstraksi fitur, metode *Support Vector Machine* bahasa pemrograman PHP.

BAB 3 ANALISIS DAN PERANCANGAN

Pada bab ini dijelaskan secara rinci mulai dari analisis masalah, analisis sistem, analisis data masukan, analisis tahap *preprocessing*, analisis klasifikasi dengan menggunakan algoritma *Support Vector Machine*. Selain itu juga dijelaskan analisis kebutuhan fungsional dan non-fungsional serta perancangan menu, antarmuka dan prosedural dari aplikasi yang akan digunakan untuk penerapan dan pengujian akurasi klasifikasi dari metode yang digunakan.

BAB 4 IMPLEMENTASI DAN PENGUJIAN

Pada bab ini ditampilkan implementasi dari hasil analisis dan perancangan yang sudah dilakukan sebelumnya menjadi sebuah aplikasi. Aplikasi digunakan untuk melakukan pengujian akurasi metode SVM. Informasi yang ditampilkan sebagai hasil dari pengujian yaitu persentase akurasi dari metode tersebut.

BAB 5 KESIMPULAN DAN SARAN

Bab ini menjelaskan tentang kesimpulan yang diperoleh dari hasil penelitian ini, serta saran–saran untuk pengembangan penelitian selanjutnya