BAB 1
PENDAHULUAN

1.1. Latar Belakang Masalah

Karangan prosa yang panjang mengandung rangkaian cerita kehidupan seseorang dengan orang di sekelilingnya dengan menonjolkan watak dan sifat setiap pelaku [1]. Novel merupakan karya sastra yang sangat digemari terutama di Indonesia. Banyak rangkaian cerita di dalamnya yang juga melibatkan banyak tokoh, baik antagonis maupun protagonis. Banyaknya tokoh dalam cerita-cerita tersebut membuat sedikit banyaknya kesulitan terhadap pembaca, karena terkadang keliru menafsirkan siapa yang sedang dibahas di dalam cerita tersebut. Untuk mengatasi hal di atas, diperlukan *Coreference Resolution* untuk mengidentifikasi kesetaraan antar entitas, seperti kata ganti maupun yang merupakan kata benda seseorang. Proses ini dilakukan untuk menemukan apakah dua ekspresi dalam *natural language* merujuk ke entitas yang sama [2].

Tujuan *Coreference Resolution* adalah untuk mengidentifikasi kesetaraan antara entitas dan antara *tag*, yaitu kata ganti dan entitas yang diidentifikasi dalam tahap mengidentifikasi entitas bernama [3]. Penelitian terhadap *coreference resolution* sudah banyak diteliti dalam pelbagai bahasa seperti bahasa Indonesia [2][4][5], bahasa China [6], maupun bahasa lainnya. Penelitian *coreference resolution* sebelumnya pernah dilakukan terhadap kasus bahasa Indonesia dengan menggunakan metode *Support Vector Machine* yang dilakukan oleh Ayu [2], penelitian yang dilakukan oleh Syifa [4], dan penelitian yang dilakukan oleh Dwi [5]. Ketiga penelitian ini menggunakan metode *SVM* dengan *kernel Linear*. Keuntungan metode *Support Vector Machine* (*SVM*) adalah memiliki *course of dimensionality*, yaitu jika metode "terpaksa" menggunakan data latih terbatas karena kendala tertentu [2].

Penelitian mengenai *Coreference Resolution* untuk Bahasa Indonesia dengan menggunakan *Support Vector Machine* (*SVM*) yang sudah dilakukan dengan

fokus untuk pronomina persona *plural* atau kata ganti orang *plural* [4] maupun singular

[2]. Pada penelitan Ayu menghasilkan nilai akurasi rata-rata sebesar 50.14% [2]. Kemudian, penelitian oleh Syifa Muhammad Husni dengan nilai akurasi yang dihasilkan adalah 61.41% [4]. Penelitian terakhir yang dilakukan oleh Dwi menghasilkan nilai akurasi sebesar 63.85% dengan akurasi tertingginya sebesar 83.3% dan akurasi terendahnya sebesar 45.4% [5].

Pada peneltian yang dilakukan oleh Dwi, terdapat kendala ketika pendeteksian kata ganti kepemilikan seperti -ku, -mu, dan -nya. Sebagai contoh, kata "misalnya" dan kata "sepenuhnya" yang menggunakan imbuhan akhiran –nya terdeteksi sebagai kata ganti kepemilikan yang merujuk kepada orang, namun sebenarnya bukan sebagai kata ganti kepemilikan yang merujuk kepada orang. Dan tidak maksimal dalam pendeteksian entitas orang sebagai satu kesatuan, sehingga token yang seharusnya adalah satu entitas orang dengan satu token, terdeteksi menjadi satu entitas orang dengan beberapa token atau bahkan terdeteksi sebagai entitas yang berbeda padahal sejatinya adalah satu entitas yang sama, atau bahkan dapat terdeteksi tidak sebagai entitas orang. Performa dari POS Tag juga masih kurang maksimal dikarenakan data latih untuk model POS Tag masih relatif sedikit, sehingga menimbulkan misklasifikasi pada pemberian tag. Penelitian tersebut tidak mendeteksi entitas yang terdapat dalam data latih maupun data uji dan penelitian tersebut menggunakan *library* sastrawi yang menggunakan algoritma *stemming* dari Nazief dan Adriani serta *Enchanced Confix Stripping* [5]. Modifikasi proses *stemming* pada tahap *preprocessing* yang dilakukan oleh Dwi masih belum maksimal, sehingga masih terdapat banyak misklasifikasi pada pendeteksian kata ganti berimbuhan.

Berdasarkan pemaparan di atas, maka penelitian ini akan menggunakan metode *Support Vector Machine* dengan *kernel RBF* dengan memaksimalkan proses pendeteksian –ku, -mu, dan –nya pada tahap *preprocessing* sebagai kata ganti kepemilikan yang merujuk kepada entitas orang, memaksimalkan kinerja POS Tagger agar memberikan tag yang lebih sesuai, dan memanfaatkan *Named Entity* untuk menghasilkan nilai akurasi *Coreference Resolution* dengan kasus novel Bahasa Indonesia.

1.1. Rumusan Masalah

Berdasarkan pemaparan pada latar belakang masalah di atas, maka timbul beberapa pertanyaan baru terhadap seberapa besar nilai akurasi yang dapat diperoleh sistem klasifikasi *coreference resolution* dengan kasus kata kepunyaan dalam teks novel Bahasa Indonesia dengan menggunakan metode *SVM*. Pertama, *kernel* yang digunakan oleh *SVM* diubah menjadi *kernel RBF*. Kedua, menggunakan pendeteksian *named entity* pada tahap *preprocessing* data yang dilatih spesifik terhadap teks novel Bahasa Indonesia. Ketiga, menggunakan *POS Tagger* yang lebih spesifik dilatih untuk menangangi kasus teks novel Bahasa Indonesia. Keempat, mengevaluasi dan menganalisa kembali proses *stemming* sehingga memiliki presisi yang lebih tinggi dan tepat untuk men-*stem* setiap kata menjadi akar katanya.

1.2. Maksud dan Tujuan

Adapun maksud dari penelitian ini adalah membangun prototipe sistem *coreference resolution* dengan menggunakan metode *SVM* dengan *kernel RBF*. Tujuan dari penelitian ini, yaitu untuk mengetahui nilai akurasi dari *coreference resolution* menggunakan metode *SVM* dengan *kernel RBF* dalam kasus Bahasa Indonesia.

1.3. Batasan Masalah

Pada penelitian ini, beberapa batasan masalah yang akan diteliti adalah sebagai berikut:

a. Masukan

Berikut adalah batasan-batasan terhadap data masukan.

1. Jenis dokumen masukan adalah dokumen yang berekstensi *Text Documents* (.txt).
2. Dokumen yang digunakan untuk tahap *training* merupakan teks dari novel Bahasa Indonesia.

1. Dokumen masukan untuk *training* bersumber dari potongan teks novel karya sastrawan Triyanto Triwikromo yang berjudul "Surga Sungsang".
2. Dokumen yang digunakan untuk tahap *testing* merupakan teks dari novel Bahasa Indonesia.

a. Proses

Berikut adalah batasan-batasan masalah untuk proses.

1. Sistem hanya akan mendeteksi kata yang berupa kata benda, nama seseorang, kata ganti orang pertama, kata ganti orang kedua, dan kata ganti orang ketiga.
2. Tahap *preprocessing* yang dilakukan adalah *Tokenizing, POS Tagging, Named Entity Detection, Case Folding, Filtering,* dan *Stemming*.
3. Tahap ekstraksi fitur menggunakan 7 jenis fitur, yaitu *i-pron, j-pron, i-propernoun, j-propernoun, Distance of String, Distance of Words,* dan *String Match.*
4. Pasangan kata yang diperoleh hanya pasangan yang salah satu katanya memiliki *tag POS* berupa *NNP* dan salah satunya lagi memiliki *tag POS* berupa *NNP* atau *PRP*. Kata yang memiliki *tag POS* berupa *NNP* pun harus memiliki label entitas berupa *"PERSON"*.
5. Menggunakan metode *Support Vector Machine* (*SVM*) dengan *Kernel Radial Basis Function* (*RBF*).
6. Klasifikasi dilakukan menggunakan *library Support Vector Machine* dari *Python*, yaitu *scikit-learn*.

b. Keluaran

Berikut adalah batasan-batasan terhadap data keluaran.

1. Pasangan kata dari hasil klasifikasi, *coreference* atau bukan.
2. Hasil pengujian menghasilkan nilai akurasi.

1.1. Metodologi Penelitian

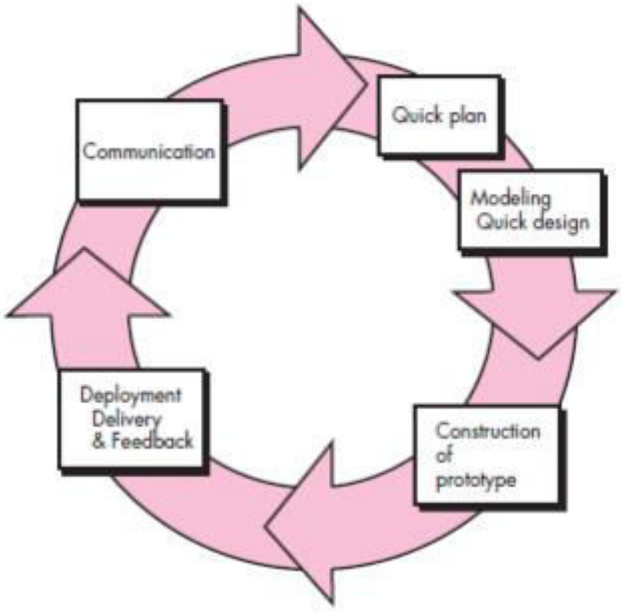
Metodologi penelitian merupakan suatu proses yang digunakan untuk memecahkan suatu masalah yang logis, dimana memerlukan data-data untuk mendukung terlaksananya suatu penelitian. Metodologi yang digunakan pada penelitian ini adalah metode deskriptif. Metode deskriptif adalah suatu metode dalam meneliti status, suatu objek, suatu set kondisi atau sesuatu. Metode ini memiliki tujuan untuk membuat deskripsi atau gambaran secara sistematis, faktual dan akurat mengenai fakta-fakta, sifat-sifat serta hubungan antar fenomena yang diselidiki. Pada penelitian ini, langkah pertama yang dilakukan adalah mengumpulkan literatur yang terkait dengan *coreference resolution* dari berbagai kasus dan Bahasa. Kemudian, fokus selanjutnya adalah mengumpulkan serta membaca literatur terkait *coreference resolution* pada kasus Bahasa Indonesia. Lalu, mengumpulkan serta membaca literatur terkait metode yang cocok diterapkan untuk *coreference resolution*. Mengumpulkan data-data selain literatur yang dibutuhkan, seperti data untuk tahap *training* dan *testing*. Kemudian, yang terakhir setelah mendapatkan seluruh literatur yang dibutuhkan dan data-data lain yang dibutuhkan, maka pembangungan perangkat lunak dengan *model prototype*.

1.1.1. Metode Pengumpulan Data

Metode pengumpulan data yang digunakan pada penelitian ini adalah Studi Literatur. Teknik pengumpulan data dengan cara mengumpulkan dari beberapa literatur, *paper*, jurnal dan buku yang berkaitan dengan *Coreference Resolution* dan *Support Vector Machine* baik berbentuk cetak maupun elektronik. Pengumpulan data juga termasuk dengan mengumpulkan data-data masukan untuk tahap *training* dan *testing* model.

1.1.2. Metode Pembangungan Perangkat Lunak

Model pembangunan perangkat lunak yang digunakan dalam penelitian ini adalah *Model Prototype*. Tahapan-tahapan pada model *prototype* menurut Roger S. Pressman dapat dilihat pada penjelasan dan Gambar 1.1 berikut.



Gambar 1.1 Skema Prototype [7]

1. *Communication*

   Mendefinisikan objektif secara keseluruhan dan mengidentifikasi kebutuhan dari *coreference resolution*. Pada penelitian ini, kebutuhan dari *coreference resolution* didapatkan melalui studi literatur dari penelitian-penelitian terkait sebelumnya.

2. *Quick Plan*

   Merupakan tahapan menentukan rencana yang akan dilakukan setelah mengetahui kebutuhan apa saja yang diperlukan. Pada penelitian ini, dilakukan perancangan terhadap pengolahan data latih dan data uji, tahap *preprocessing*, dan tahap ekstraksi fitur.

3. *Modelling Quick Design*

   Merupakan tahap desain yang dikerjakan setelah kebutuhan selesai dikumpulkan secara lengkap. Pada penelitian ini, *modelling* yang akan dibangun, yaitu model *SVM*.

4. *Construction of Prototype*

   Merupakan tahap desain program yang diterjemahkan ke dalam kode-kode dengan menggunakan bahasa pemrograman yang sudah

ditentukan dan *testing* yang disesuaikan dengan *modelling* yang telah selesai dibuat.

1. *Deployment Delivery and Feedback*

   Prototype dievaluasi oleh pengguna dan digunakan untuk memperbaiki persyaratan perangkat lunak yang akan dikembangkan. Apabila *prototype* sudah sesuai dengan kebutuhan, maka perangkat lunak dapat diterima dan penelitianpun selesai. Jika terdapat suatu revisi yang harus dilakukan ataupun tidak sesuai dengan kebutuhan, maka kembali lagi ke proses *communication* dan melakukan revisi yang dibutuhkan.

1.1. Sistematika Penulisan

Dalam penulisan dokumen skripsi ini dilakukan secara bertahap, yang bertujuan agar dapat memahami isi dokumen skripsi ini secara keseluruhan. Untuk itu penulisan dokumen skripsi ini terdiri dari beberapa bab yang menjelaskan secara rinci hasil penelitian. Isi pokok dari dokumen skripsi ini adalah:

BAB I PENDAHULUAN

Berisi tentang latar belakang, identifikasi masalah, maksud dan tujuan, batasan masalah, metodologi penelitian, manfaat penelitian, ruang lingkup penelitian, dan sistematika penulisan.

BAB II LANDASAN TEORI

Bab ini berisikan pemaparan seluruh teori-teori dari referensi dan tinjauan yang digunakan kepada objek penelitian yang dilakukan sebagai penunjang. Dan juga teori yang berupa pengertian dan definisi yang diambil dari referensi yang berkaitan dengan penyusunan dokumen skripsi ini.

BAB III ANALISIS DAN PERANCANGAN

Bab ini berisikan seluruh pembahasan pada penelitian ini.

BAB IV IMPLEMENTASI DAN PENGUJIAN

Bab ini berisi tentang implementasi dari hasil analisis dan perancangan yang sudah dilakukan sebelumnya dalam bentuk aplikasi. Aplikasi digunakan untuk melatih model dan menampilkan hasil uji dari tahap *training* yang sudah dibuat.

BAB V PENUTUP

Berisi kesimpulan yang sudah diperoleh dari hasil penulisan dokumen skripsi ini dan saran untuk pengembangan penelitian lebih lanjut.

DAFTAR PUSTAKA

LAMPIRAN