BAB 1
PENDAHULUAN

1.1 Latar Belakang

Natural Language Processing (NLP) merupakan sebuah teknik yang berfungsi untuk menganalisis dan merepresentasikan bahasa manusia secara otomatis dengan mempelajari model matematis dan komputasi dari berbagai macam aspek bahasa dan pengembangan pada sistem yang luas. NLP digunakan untuk mengambil struktur gramatikal. NLP membangun *output* berdasarkan aturan yang ada pada bahasa yang dijadikan objek pemrosesan [1].

Contoh pemanfaatan NLP adalah pada deteksi plagiarisme, *information retrieval, text summarization, question answering, machine translation*. Pada kasus deteksi plagiarisme salah satu proses yang diperlukan adalah proses pengenalan parafrasa. Parafrasa sendiri merupakan pengungkapan kembali suatu tuturan dari sebuah tingkatan atau macam Bahasa menjadi yang lain tanpa mengubah pengertian; Parafrasa dapat diartikan juga sebagai penguraian kembali suatu teks dalam bentuk yang lain, dengan maksud untuk dapat menjelaskan makna yang tersembunyi [2]. Parafrasa digunakan oleh seorang penulis untuk menjelaskan sesuatu menggunakan pendekatan yang berbeda namun mengandung pesan yang sama.

Hal yang membuat proses pengenalan parafrasa penting adalah perlunya mesin untuk membedakan secara otomatis frasa-frasa yang berbeda bentuk namun memiliki makna yang sama. Misalnya pada kalimat “penyebab kebakaran hutan”, seharusnya komputer akan mengenali bahwa kalimat tersebut serupa dengan kalimat “sumber kebakaran hutan”. Pada pengenalan parafrasa bahasa Indonesia terdapat prefiks, sufiks, infiks, dan konfiks pada struktur bahasa sehingga sulit untuk menyocokan kata yang berkaitan.

Untuk menghadapi permasalah diatas maka dibutuhkan sebuah proses yang dinamakan identifikasi parafrasa. Identifikasi parafrasa adalah proses untuk mengenali ungkapan dari sepasang kalimat apakah keduanya memiliki arti sama atau tidak. Pendekatan yang dilakukan untuk mengidentifikasi parafrasa adalah melakukan *preprocessing* yang bertujuan untuk meningkatkan kualitas data, *preprocessing* terdiri dari 3 tahap yaitu *tokenization, non-alphanumerical removal*, dan *stemming.* Algoritma *stemming* yang digunakan untuk *preprocessing* dataset parafrasa bahasa Indonesia adalah algoritma Nazief-Adriani karena memiliki performansi terbaik untuk dataset bahasa Indonesia [3]. Data hasil *preprocessing* tersebut lalu dilakukan proses *feature extraction* yang bertujuan untuk membangun fitur-fitur baru dari data set tersebut. Fitur yang pertama adalah fitur sintaktik yang merupakan hasil dari perhitungan jarak antara dua kalimat, perhitungan jarak tersebut menggunakan metode *Normalized Levensthein Distance.* Fitur yang kedua adalah fitur semantik, fitur ini menghitung kemiripan pasangan kalimat berdasarkan pohon semantik, perhitungan jarak semantik dilakukan dengan menggunakan metode Wu *and* Palmer. Setelah dilakukan ekstraksi fitur, dataset tersebut dibagi menjadi dua bagian yaitu data *training* dan data *testing.* Setelah data selesai dibagi, maka dilakukan diskritisasi nilai fitur dengan *clustering* menggunakan metode K-

Means. Metode yang digunakan untuk melatih *classifier* adalah *Bayesian Networks*. Perhitungan parameter yang digunakan *classifier* ini adalah MAP(*Maximum A Posteriori*) dan *Multinomial Distribution Probability*.

Bayesian networks merupakan suatu metode pemodelan data berbasis probabilitas yang merepresentasikan suatu himpunan variabel dan *conditional dependency*-nya melalui suatu *Directed Acyclic Graph*(DAG) [4]. Ada empat alasan mengapa mengambil *bayesian networks* sebagai *classifier*, pertama *bayesian networks* dapat menangani dataset yang tidak lengkap, kedua *bayesian networks* memungkinkan proses *learning* mengenai hubungan sebab-akibat, yang ketiga *bayesian networks* sejalan dengan teknik *bayesian* statistik yang memfasilitasi kombinasi antara data dan *domain knowledge*, yang terakhir adalah *bayesian networks* menyediakan jalan yang efisien untuk menghindari data yang bersifat *over fit* [5].

1.2 Perumusan Masalah

Permasalahan yang dirumuskan dalam penelitian ini adalah:

a. Bagaimana *pre-processing* untuk dataset Bahasa Indonesia?
b. Bagaimana menggunakan *Bayesian Networks* sebagai *classifier* untuk mengenali parafrasa berbahasa Indonesia?
c. Bagaimana cara mengekstraksi fitur dari pasangan kalimat sehingga menjadi sebuah fitur baru?
d. Pembagian data *training* dan data *testing* seperti apa yang terbaik bagi *classifier Bayesian Networks.*

Pada penelitian ini terdapat beberapa batasan masalah untuk menghindari meluasnya materi pembahasan. Adapun batasan masalah dalam penelitian ini adalah sebagai berikut.

a. Identifikasi hanya berlaku untuk parafrasa pada dataset dan dokumen berbahasa Indonesia.
b. Hasil identifikasi dibedakan menjadi dua yaitu *true* jika merupakan parafrasa dan *false* jika bukan merupakan parafrasa.

1.3 Tujuan

Berdasarkan perumusan masalah, maka tujuan dari penelitian ini adalah:

a. Melakukan *pre-processing* pada *dataset* parafrasa Bahasa Indonesia.
b. Menggunakan *classifier Bayesian Networks* sebagai *classifier* untuk mengenali parafrasa berbahasa Indonesia.
c. Melakukan ekstraksi fitur dari pasangan kalimat sehingga menjadi fitur baru.
d. Mencari pembagian data *training* dan data *testing* terbaik bagi *classifier Bayesian Networks.*

1.4 Metodologi Penelitian

Adapun tahapan metode yang digunakan pada Tugas Akhir ini adalah sebagai berikut:

a. Studi literatur dan identifikasi masalah
   Tahap ini adalah tahap mencari, membaca, dan mempelajari berbagai *paper*, jurnal, buku, maupun *website* yang berhubungan dengan studi kasus. Dengan tahap ini, penulis mempelajari mengenai *teori-teori* pengenalan parafrasa dan *classifier* model yang bisa digunakan untuk menyelesaikan tugas akhir ini.

b. Pengumpulan data
   Tahap ini merupakan tahap pencarian sumber data dan penyusunan dataset parafrasa Bahasa Indonesia yang dapat berupa kumpulan kata, frase maupun kalimat. Data-set inilah yang akan dijadikan data *training*, data *validation* maupun data *testing.*

c. Analisis dan perancangan
   Tahap ini adalah tahap menspesifikasikan kebutuhan yang diperlukan pada proses pembuatan sistem.

d. Implementasi sistem dan pengujian
   Tahap ini adalah tahap mengimplementasikan hal-hal yang sudah dibuat di tahap analisis dan perancangan. Lalu dilakukan pengujian terhadap aplikasi yang telah dibuat dan memastikan kelayakan aplikasi tersebut apakah sudah memenuhi semua point dari tujuan pembuatan tugas akhir ini.

e. Analisis hasil pengujian
   Proses ini merupakan proses evaluasi terhadap hasil pengujian yang telah dilakukan dengan melakukan analisis terhadap hasil yang didapatkan apakah sesuai dengan yang diharapkan.

f. Penyusunan laporan
   Tahap ini dilakukan pelaporan terhadap hasil analisis dan pengujian terhadap topik tugas akhir ke dalam bentuk buku beserta dengan dokumentasi yang diperlukan seperti referensi, sistem yang sudah diimplementasikan dan hasil penelitian yang didapatkan.

1.5 Sistematika Penelitian

Guna memahami lebih jelas laporan Skripsi ini, dilakukan dengan cara mengelompokkan materi menjadi beberapa sub bab dengan sistematika penulisan sebagai berikut:

a. BAB 1 PENDAHULUAN
   Bab ini menjelaskan tentang informasi umum mengenai penelitian yang dilakukan, yaitu latar belakang penelitian, perumusan masalah, tujuan penelitian, metodologi penelitian, dan sistematika penelitian.

b. BAB II TINJAUAN PUSTAKA
   Bab ini berisikan teori yang diambil dari beberapa kutipan buku, yang berupa pengertian dan definisi. Bab ini juga menjelaskan konsep identifikasi parafrasa, konsep dasar *preprocessing*, konsep dasar ekstraksi fitur, konsep dasar *classifier*, dan konsep dasar evaluasi performansi sistem.

c. BAB III PERANCANGAN SISTEM
   Bab ini menjelaskan perancangan sistem yang diusulkan dengan menggunakan *Flow Chart*. Pada bab ini dijelaskan proses *preprocessing* data, ekstraksi fitur, diskritisasi data, pembagian porsi data, pembangunan *classifier*, dan evaluasi performansi sistem. Pada bab ini dijelaskan pula mengenai lingkungan pengembangan sistem.

d. BAB IV EVALUASI DAN ANALISIS
   Bab ini menjelaskan mengenai dataset seperti apa yang akan digunakan dalam proses pengujian. Selain itu dijelaskan mengenai scenario pengujian seperti apa yang akan dilakukan, dan bagaimana hasil dari pengujian yang dilakukan berdasarkan skenario pengujian yang ada.

e. BAB V KESIMPULAN
   Bab ini berisi kesimpulan dan saran yang berkaitan dengan hasil analisa sistem. Kesimpulan menjelaskan apakah sistem yang dibuat sudah sudah sesuai dengan tujuan penelitian.