BAB I
PENDAHULUAN

1.1 Latar Belakang

Natural Language Processing (NLP) adalah area penelitian dan pengaplikasan yang mengekplorasi bagaimana caranya sebuah komputer dapat digunakan dan memanipulasi berupa teks atau suara bahasa alami sehingga berguna untuk beragam hal [1]. NLP mempunyai tujuan untuk melakukan proses pembuatan model komputasi dari bahasa, sehingga dapat terjadi interaksi antara manusia dengan komputer dengan perantara bahasa alami. Secara umum NLP memodelkan pengetahuan terhadap fonetik dan fonologi, morfologi, sintaksis, semantik dan pragmatik pada bahasa [2].

Identifikasi Parafrasa merupakan sebuah percabangan dari studi NLP. Parafrasa menurut Harimurti Kridalaksana merupakan istilah linguistik yang berarti pengungkapan kembali suatu konsep dengan cara lain dalam bahasa yang sama, namun tanpa mengubah maknanya [3]. Laman berita merupakan contoh nyata, sebuah berita yang memiliki informasi dapat ditulis ulang dengan informasi yang sama dan konsep berbeda, walaupun kedua berita tersebut memiliki elemen leksikal yang berbeda atau mungkin memiliki struktur sintaksis yang berbeda namun memiliki makna yang sama dapat disebut sebagai parafrasa [4]. Dalam mengenali suatu bentuk parafrasa dapat dilakukan oleh evaluasi manusia, namun evaluasi parafrasa oleh manusia membutuhkan biaya yang besar dan waktu yang lebih lama, hal ini bisa menjadi masalah besar untuk *developer* [5].

Automatic metric adalah sebuah mesin evaluasi otomatis yang menggunakan *metric* yang dapat digunakan sebagai ukuran identifikasi parafrasa. Pada beberapa *automatic metric* menggunakan fitur-fitur yang dapat digunakan sebagai ekstraksi lingual sehingga menghasilkan nilai (*score*) yang dapat digunakan sebagai ukuran parafrasa dua buah kalimat yang dibandingkan [6] Pada penelitian ini penulis menggunakan tiga algoritma *automatic metric* yaitu BLEU, METEOR, Damerau-Levensthein *Edit Distance* yang menguji data *testing* yang sama. Kemudian hasil dari masing-masing *automatic metric* akan dihitung nilai akurasinya dengan menggunakan *F1-measure*. Data *set* yang digunakan berasal dari data Twitter yang berupa *corpus* yang sudah di-*format* ke dokumen ber-ekstensi *.txt*.

Berdasarkan latar belakang yang telah disampaikan, dilakukan sebuah penelitian Analisis dan Implementasi algoritma BLEU, METEOR, Damerau-Levensthein *Edit Distance* untuk mengidentifikasi parafrasa, selanjutnya hasil dari identifikasi oleh *automatic metric* akan evaluasi dengan hasil korelasi masing-masing algoritma dengan *human judgement* dengan melakukan melakukan perhitungan nilai akurasi dan F1. Dengan melakukan perhitungan nilai akurasi dan F1 dapat menjadi pertimbangan untuk melihat

performansi *metric* dimana dapat melihat seberapa akurat *automatic metric* jika dibandingkan dengan penilaian manusia.

1.2 Perumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan di atas, berikut rumusan masalah yang telah ditetapkan pada tugas akhir ini :

1. Bagaimana mengimplementasikan algortima BLEU, Meteor dan Damereu-levensthein *Edit distance* dalam mengidentifikasi *parafrasa* dua buah *tweet*?
2. Bagaimana memperoleh nilai akurasi dan F1 dari algoritma BLEU, METEOR dan Damereu-levensthein *Edit distance* pada identifikasi parafrasa dua buah *tweet*?
3. Bagamaina membandingkan algortima BLEU, METEOR dan Damereu-levensthein *Edit distance* dalam mengidentifikasi *parapfrasa* dua buah *tweet*?

1.3 Batasan Masalah

Adapun batasan masalah yang dijadikan acuan dalam tugas akhir ini :

1. Algoritma yang akan dikaji yaitu BLEU, METEOR dan Damereu-levensthein *Edit distance*.
2. Membandingkan hasil dari tiga algoritma yaitu BLEU, METEOR dan *Edit distance*.
3. Menggunakan *corpus* berupa *tweet* dalam Bahasa inggris.
4. Dataset yang digunakan dari SemEval 2015 task 1.

1.4 Tujuan

Tujuan pembuatan tugas akhir dengan judul “Analisis dan Implementasi identifikasi Parafrasa *Tweet* Menggunakan Algoritma BLEU, Meteor dan *Edit distance*” diantaranya yaitu :

1. Mengetahui implementasi algortima BLEU, Meteor dan Damereu-levensthein *Edit distance* dalam mengidentifikasi *parafrasa* dua buah *tweet*?
2. Mengetahui nilai akurasi dan F1 dari algoritma BLEU, METEOR dan Damereu-levensthein *Edit distance* pada identifikasi parafrasa dua buah *tweet*.
3. Mengetahui perbandingan algortima BLEU, METEOR dan Damereu-levensthein *Edit distance* dalam mengidentifikasi *parapfrasa* dua buah *tweet*?

1.5 Manfaat Penelitian

Dengan penilitian yang dilakukan diharapkan mendapatkan hasil dari pengaplikasian metode-metode yang tercantum yang selanjutnya bisa diukur seberapa efektifnya metode yang digunakan tersebut. Dalam penilitian ini penulis mengharapkan dari penilitian identifikasi parafrasa ini dapat memberikan sebuah solusi agar dapat mengidentifikasi pengguna robot dalam Twitter yang melakukan *spamming* dengan *tweet* yang sama.

1.6 Metodologi Penyelesaian Masalah

Dalam tugas akhir ini, digunakan metodologi penelitian yang dapat dijabarkan sebagai berikut :

1. Studi literatur
   Dalam pengumpulan bahan penilitian penulis mengumpulkan referensi dari berbagai jurnal, e-book, artikel, diskusi, internet dan sumber-sumber lainnya. Termasuk pengumpulan data-data pendukung dan melakukan analisa sistem yang diperlukan dalam tahap implementasi.
2. Konsultasi dan diskusi
   Tahap ini dilakukan dengan melakukan konsultasi dan diskusi dengan dosen pembimbing tugas akhir dan senior atau teman yang mengambill topik tugas akhir yang sama.
3. Analisis perancangan dan implementasi
   Merancang sistem dengan menggunakan metode yang telah ditentukan kemudian diimplementasikan dalam aplikasi untuk mendeteksi apakah kedua kalimat sama atau tidak.
4. Uji coba sistem
   Uji coba terhadap aplikasi perangkat lunak yang telah dibuat dan menganalisa performansi dari algoritma yang digunakan.
5. Penulisan laporan
   Tahap ini dilakukan penyusunan laporan sebagai proses pendokumentasian terhadap proses dan hasil dari pelaksanaan tugas akhir ini

1.7 Sistematika Penulisan

Adapun Sistematika penulisan laporan tugas akhir ini disusun sesuai rencana berikut :

BAB I Pendahuluan

Pada bab pertama ini menjelaskan gambaran secara umum tentang tentang tugas akhir yang dikerjakan mengenai latar belakang masalah, batasan masalah, perumusan masalah, tujuan penelitian, manfaat penelitian, metodologi penelitian dan sistematika penulisan.

BAB II Landasan Teori

Pada bab kedua ini akan dijelaskan dasar-dasar teori yang mendukung penyelesaian tugas akhir, khususnya teori yang berkaitan dengan masalah penelitian sehingga dapat mendukung pengimplementasian algoritma serta pengujiannya.

BAB III Perancangan

Pada bab ketiga ini dijelaskan perancangan yang diperlukan berdasarkan mekanisme dan batasan yang digunakan, sehingga dapat diimplementasikan serta direalisasikan *software* yang akan dibuat sesuai dengan yang diinginkan.

BAB IV Analisis Dan Pengujian

Pada bab keempat ini akan dijelaskan hasil pengujian dari perancangan dan implementasinya mengenai parameter- parameter yang dibutuhkan sesuai metode yang digunakan dalam sistem yang telah dibangun pada bab sebelumnya.

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.