1. Pendahuluan

1.1 Latar Belakang

Model semantic distribusional merupakan metode yang digunakan untuk menyelesaikan masalah dalam level *similarity*. Model komposisi semantik distribusional merupakan *extend* dari model semantic distribusional yang digunakan untuk menyelesaikan masalah dalam level *entailment.* Dilihat dari hal ini dengan sedikit pengembangan pada model semantic distribusional yang pada umumnya untuk menyelesaikan masalah dalam level *similarity* apakah model komposisi semantik distribusional dapat optimal dalam menyelesaikan masalah dalam level *entailment.*

Textual entailment adalah bentuk keterkaitan / relasi semantik yang didasarkan pada hubungan sebab-akibat dilihat dari makna berdasarkan keseluruhan isi kalimat secara lengkap [2]. Setiap kalimat secara tekstual dapat dicari keterkaitannya melalui model komposisi semantik distribusional, dimana hasil semantik dari setiap kata dalam kalimat-kalimat tersebut secara terdistribusi dapat memberikan nilai probabilitas yang dapat digunakan untuk menentukan nilai bobot apakah sebuah kalimat memiliki keterkaitan dengan kalimat lainnya [2].

Model komposisi semantik distribusional pemanfaatannya banyak digunakan dalam Pemrosesan Bahasa Alami (PBA) atau juga disebut *Natural Language Processing* (NLP) [3]. Pada pengaplikasiannya dapat menggunakan beberapa *tool NLP* yang dapat digunakan untuk mendapatkan hasil yang baik dan cepat, salah satunya adalah menggunakan WordNet. Wordnet dapat dikategorikan sebagai sistem *lexical database* Bahasa Inggris yang saat ini populer digunakan. Sistem *lexical database* pada WordNet sendiri adalah kumpulan data (kata) yang menyimpan informasi relasi semantik antar synset (satuan dalam WordNet yang mengartikan sebagai sinonim set) [4].

Dalam penelitian ini, diimplementasikan model komposisi semantik distibusional pada sejumlah dataset dari berbagai sumber yang menyimpan informasi tentang *textual entailment* dan membandingkan dengan hasil penelitian dalam bentuk evaluasi untuk mengukur seberapa besar nilai *accuracy* yang didapatkan.

1.2 Perumusan Masalah

Terdapat beberapa perumusan masalah yang dapat diambil terkait latar belakang diatas.

1. Bagaimanakah mengimplementasikan model komposisi semantik distribusional pada dataset dengan kalimat lengkap yang berpasangan untuk mendapatkan hasil *textual entailment*?
2. Bagaimanakah menentukan hasil evaluasi perbandingan model komposisi semantik distribusional untuk mendapatkan hasil *textual entailment* dalam penelitian dengan hasil *textual entailment* yang tertera pada dataset sesuai sumber awalnya?

1.3 Tujuan

Berdasarkan perumusan masalah diatas, tujuan dari Tugas Akhir ini ialah.

1. Mengimplementasikan model komposisi semantik distribusional untuk mendapatkan hasil *textual entailment*.
2. Mendapatkan hasil evaluasi perbandingan model komposisi semantik distribusional untuk mendapatkan hasil *textual entailment* dalam penelitian dibandingkan dengan hasil *textual entailment* yang tertera pada dataset sesuai sumber awalnya.

1.4 Batasan Masalah

Batasan masalah pada Tugas Akhir ini ialah.

1. WordNet yang digunakan dalam penelitian adalah WordNet versi 2.1 yang digunakan pada sistem operasi Microsoft Windows yang dikeluarkan dan dikembangkan oleh Princeton University
2. Dataset yang digunakan untuk mendapatkan hasil *textual entailment* yang digunakan dalam penelitian adalah dataset yang diterbitkan SICK (*Sentences Involving Compositional Knowledge*) yang terdiri dari 10.000 pasang kalimat lengkap yang juga pernah digunakan oleh SemEval 2014
3. Label hasil *textual entailment* yang akan digunakan terdiri dari *entailment*, *contradiction* dan *neutral*.

1.5 Metodologi Penyelesaian Masalah

Metodologi penyelesaian masalah yang digunakan dalam penelitian ini adalah sebagai berikut:

1. Studi Literatur
   Pada tahap ini dilakukan proses pembelajaran terhadap beberapa sumber seperti jurnal, artikel, buku, dan internet yang dapat mendukung penyelesaian tugas akhir ini.
2. *Pre-processing* Data
   Data mentah yang berasal dari dataset SICK (*Sentences Involving Compositional Knowledge*) harus dilakukan *pre-processing* untuk menghilangkan *noise* dan mendapatkan data yang diinginkan, dan memberi label pada setiap kata yang ditemukan dalam kalimat sebagai input untuk tahap ekstraksi fitur. *Pre-processing* yang dilakukan antara lain:
   a. *Lemmatization*, merupakan proses merubah kata menjadi bentuk kata dasarnya.
   b. *Stop Word Removal*, menghilangkan kata-kata yang sering muncul dan dianggap tidak memiliki arti penting seperti kata sambung. Sebagai contoh, kata “dan”, “atau”, “jika”, “karena”, dan lain-lainnya.
   c. *POS (Part-of-Speech) Tagging*, merupakan proses pelabelan atau pemberian tag pada setiap kata yang diwakili dengan dengan simbol huruf yang memberikan informasi tentang kategori kata (kata benda / kata kerja / kata sifat, dll.)

3. Ekstraksi Fitur
   Ekstraksi fitur digunakan untuk mencari dan mengambil kata yang mengandung ciri (fitur) yang mewakili kalimat. Fitur ini digunakan sebagai input pada model komposisi semantik distribusional. Ekstraksi fitur ini menggunakan hasil *POS tagging* yang didapatkan dari tahap *pre-processing* sebelumnya. Sedangkan metode yang digunakan dalam tahap ekstraksi fitur adalah dengan menggunakan metode *knowledge pattern*.
4. Perancangan Sistem
   Tahap perancangan sistem pada tugas akhir ini di gambarkan menggunakan *flowchart* (diagram alir). Tahap perancangan sistem juga membantu penulis dalam mematangkan rancangan fungsionalitas pada sistem.
5. Implementasi
   Implementasi dilakukan bersamaan dengan pengumpulan data dan implementasi sesuai dengan perancangan sistem yang telah ditentukan sebelumnya.
6. Pengujian dan Analisis
   Pada tahap ini dilakukan pengujian terhadap sistem yang telah dibuat agar tidak terjadi *error* dan analisis terhadap hasil yang telah diperoleh.
7. Pembuatan Laporan
   Pembuatan laporan dilakukan setelah hasil dari analisis yang dilakukan sebelumnya telah didapatkan.

1.6 Sistematika Penulisan

Tugas akhir dengan judul "**Evaluasi Model Komposisi Semantik Distribusional Pada Kalimat Lengkap Melalui Tekstual Entailment**" ini disusun dengan sistematika penulisan sebagai berikut:

1. Pendahuluan
   Bab 1, menjelaskan latar belakang masalah yang diangkat, perumusan masalah, tujuan, batasan masalah dan metodologi penyelesaian masalah serta sistematika penulisan dari tugas akhir ini.
2. Tinjauan Pustaka
   Bab 2 ini, menjelaskan teori-teori yang mendukung dan terkait dalam tugas akhir ini.
3. Perancangan Sistem
   Bab 3 ini, memuat gambaran umum sistem, *flowchart pre-processing* data, *flowchart* proses ekstraksi fitur, *flowchart* tahapan model komposisi semantik distribusional, spesifikasi sistem dan fungsionalitas sistem serta validasi sistem.
4. Pengujian dan Analisis
   Bab 4, memuat pengujian sistem, metrik evaluasi yang digunakan, hasil pengujian dan analisis.
5. Kesimpulan dan Saran
   Bab 5 ini, menjelaskan kesimpulan hasil dari penelitian tugas akhir yang dilakukan dan saran yang membangun untuk kedepannya.