BAB I PENDAHULUAN

1.1 Latar Belakang

Pesatnya teknologi yang berkembang pada saat ini berbanding lurus dengan tingginya tingkat kebutuhan manusia akan penggunaan teknologi. Hal ini mengakibatkan ketergantungan yang terjadi pada manusia dengan teknologi yang ada. 84% dari 5000 responden bahkan menyatakan bahwa mereka tidak dapat pergi ke luar rumah tanpa membawa *gadget* mereka [1]. Dengan diperolehnya fakta ini, teknologi secara otomatis menciptakan kumpulan data yang besar yang terhasilkan setiap waktunya. Dari berbagai jenis teknologi dan kegiatan yang dilakukan manusia, salah satu jenis data yang sangat populer pada saat ini adalah data berbentuk teks. Tak ingin terbuang dengan percuma, data teks yang dihasilkan lalu dimanfaatkan oleh peneliti pada proses pengembangan ilmu pengetahuan. *Text Mining* merupakan solusi yang tepat untuk pengolahan dokumen teks, sehingga dapat menghasilkan pemanfaatan lebih untuk diterapkan dalam kehidupan.

Semantic Textual Similarity, merupakan salah satu *task* pada *text mining* dalam ranah *Natural Language Processing* untuk melihat kemiripan makna pasangan teks dengan memberikan skor kemiripan antar keduanya. Kegunaan skor ini adalah untuk tahapan awal bagi penerapan penelitian NLP, seperti *question answering* yang ditanamkan pada aplikasi sehingga dapat sistem menjawab pertanyaan yang diajukan oleh manusia secara faktual. Untuk contoh lainnya adalah dalam dunia pendidikan. *Automatic essay grading,* sebuah sistem yang dapat melakukan penilaian secara otomatis dengan melihat ciri-ciri pada teks yang memenuhi kriteria penilaian sehingga nilai dapat dikategorikan. Implementasi lainnya adalah *plagiarism detection* [2], yaitu pendeteksian plagiarisme, dengan melihat kemiripan makna dan struktur kalimat dengan literatur yang telah ada sebelumnya. Juga pencarian informasi yang dilakukan pada mesin pencari, menerapkan prinsip similaritas ini. Dilakukan pencarian pada dokumen-dokumen yang terdapat pada *database* mesin pencari, lalu dicocokkan kemiripannya pada kata yang diinputkan oleh *user*. Sehingga dokumen yang ditampilkan sesuai dengan pencarian yang diinginkan.

Meskipun banyak riset terkait yang telah dilakukan untuk mendapatkan skor similaritas antar teks[18], masih banyak celah pengembangan metode yang dapat dilakukan. *Semantic Similarity Measurement* yang digunakan dalam penghitungan dari kali ini adalah penggunaan metode *Information Content Based* dalam penghitungan pasangan kata. Meskipun satuan teks yang digunakan kecil belum bisa dapat diimplementasikan langsung dalam pengembangan yang disebutkan pada contoh awal, pengujian antar pasangan kata ini merupakan tahapan awal dalam melakukan penelitian kemiripan semantik yang lebih luas. Dengan dilakukannya metode ini beserta pengembangannya, diharapkan skor kemiripan

dari pasangan kalimat tersebut menghasilkan nilai korelasi yang tinggi dengan *Gold Standard* yang sudah ditetapkan.

1.2 Perumusan Masalah

Berdasarkan latar belakang tersebut, rumusan permasalahan yang dibangun adalah sebagai berikut:

1. Bagaimana pencarian skor *Semantic Textual Similarity* dengan menggunakan tiga metode *IC Based (Resnik, Lin,* dan *Jiang Conrath)*?
2. Bagaimana melakukan pencarian skor pada pasangan kata *Semantic Textual Similarity* dengan menggunakan metode *IC* hyponim dan frekuensi?
3. Bagaimana melakukan pencarian skor pada pasangan kata *Semantic Textual Similarity* dengan menggunakan metode *IC* dengan menggunakan dan tanpa menggunakan penandaan sense?
4. Bagaimana pengaruh kondisi dataset dengan POS noun dan verb pada taksonomi WordNet dalam pencarian skor similaritas dengan metode IC yang diterapkan?

1.3 Batasan Masalah

Adapun batasan masalah dalam riset ini adalah sebagai berikut:

1. Data berupa dokumen teks yang berisikan pasangan kata
2. Pasangan kata yang digunakan menggunakan bahasa Inggris, dan hanya menangani kata benda (*noun*) dan kata kerja *(verb)*
3. Skor yang dihasilkan memiliki rentang yang berbeda-beda, disesuaikan dengan Gold Standard yang dimiliki dari tiap dataset
4. Sistem yang dibangun hanya sampai dengan tahapan menghasilkan skor similaritas antar kata

1.4 Tujuan

Tujuan dari riset ini adalah sebagai berkut:

1. Mengetahui penghitungan metode *Information Content* terbaik dibandingkan dengan tiga metode (*Resnik, Lin dan Jiang Conrath*).
2. Mengimplementasikan metode *IC based* dalam penghitungan skor similaritas pasangan kata dengan menggunakan penghitungan nilai hyponym dan frekuensi
3. Mengimplementasikan kedua kondisi dengan dan tanpa *sense-tagged* untuk mengetahui pengaruh pada skor similaritas yang optimal
4. Mengetahui keterkaitan kondisi dataset yang memiliki POS yang berbeda pada taksonomi WordNet dalam menghasilkan skor similaritas yang optimal

1.5 Metodologi Penyelesaian Masalah

Adapun metodologi penyelesaian masalah dalam riset ini adalah:

1. Studi Literatur
   Pada tahap ini dilakukan studi literatur yang terkait dengan studi kasus maupun metode analisis yang digunakan, yaitu *Semantic Textual Similarity*. Berikut dengan informasi karakteristik dataset yang digunakan.
2. Pengumpulan Data
   Pada tahap ini dilakukan pengumpulan beberapa jenis data yang didapatkan yaitu Simlex-999, WordSim-353 dan R&G berupa dokumen teks .txt yang berisikan sekumpulan pasangan kata.
3. Perancangan Sistem
   Dalam tahapan ini dilakukan perancangan sistem yang akan diimplementasikan pada riset ini, sesuai dengan metode yang dipilih, termasuk desain dan perangkat lunak yang digunakan.
4. Pembangunan dan Implementasi Sistem
   Pada tahap ini dibangun sistem yang sebelumnya telah dirancang, dengan menggunakan metode yang telah dipilih. Implementasi pada penelitian tugas akhir ini dibangun menggunakan bahasa pemrograman Java dan IDE Netbeans.
5. Pengujian dan Analisis
   Pada tahap ini dilakukan pengujian studi kasus kemiripan dari beberapa pasangan kata berbahasa inggris sehingga menghasilkan skor. Setelah itu pada hasil skor tersebut dilakukan korelasi dengan *Gold Standard*. Setelah dilakukan pengujian, dilakukan analisis terhadap hasil skor similaritas dan juga hasil korelasi, sehingga kesimpulan dapat ditentukan.