

ABSTRAK

Semantic Textual Similarity merupakan salah satu task yang berada pada ranah *Natural Language Processing*. Task yang dilakukan meliputi penghitungan skor similaritas makna dari pasangan teks yang diberikan. Dalam penelitian ini, teks berupa pasangan kata yang berasal dari tiga jenis dataset yang berbeda yaitu *Simlex-999*, *WordSim-353* dan *Rubenstein & Goodenough*. Dari hasil penghitungan skor ini, didapatkan nilai similaritas yang merupakan tahapan awal dari implementasi pengembangan penelitian NLP selanjutnya.

Penghitungan skor dilakukan dengan menggunakan bantuan library *JWNL*, *lexical database WordNet*, dengan menggunakan metode-metode yang terdapat pada IC seperti *Resnik*, *Lin* dan *Jiang Conrath*. Metode IC yang dipilih kemudian diterapkan pada lingkungan sistem yang berbeda, seperti melakukan pencarian nilai IC dengan menggunakan metode pengembangan penghitungan nilai hyponym dan dilakukan perbandingan terhadap penghitungan normal frekuensi, pengujian dengan dan tanpa penandaan sense, serta menganalisis perbedaan hasil pengujian berdasarkan karakteristik kata POS NOUN dan VERB. Dengan mengimplementasikan pengembangan metode tersebut, dihasilkan bahwa metode Lin meraih nilai korelasi tertinggi diantara metode IC lainnya sebesar 85,5% pada dataset R&G, dan penghitungan nilai IC dengan frekuensi lebih baik dengan menggunakan penandaan sense, dan sebaliknya pada hyponym. Korelasi kata yang memiliki POS NOUN pun lebih tinggi dengan 59.8 % dibandingkan dengan VERB.

Kata kunci: *Semantic Textual Similarity*, *Information Content*, *Lin*, *WordNet*, *SimLex-999*, jumlah *hyponym*