BAB 1. PENDAHULUAN

Similaritas antara dua teks atau kalimat merupakan angka yang menggambarkan kedekatan makna antara kedua teks atau kalimat. Perhitungan similaritas digunakan dalam berbagai keperluan, misalnya untuk melakukan pencarian informasi di internet, pencarian dokumen di *harddisk*, klasifikasi dokumen dalam arsip, deteksi plagiasi, dan kegiatan menganalisis informasi di dunia maya (*data analysis*) (Islam dan Inkpen, 2008).

Penerapan algoritma similaritas paling banyak terjadi pada proses pencarian informasi. Algoritma similaritas digunakan untuk mengukur kemiripan makna kata atau frase yang dicari dengan teks yang ada dalam halaman yang ditelusuri. Pencarian informasi tidak cukup dilakukan dengan membandingkan kata atau frase yang dicari dengan kata atau frase yang ada dalam dokumen. Pencarian yang efektif memerlukan analisis mengenai makna kata dan frase yang diinginkan oleh *user* dan penentuan tema dokumen yang ditelusuri. Efektifitas pencarian dapat diperbaiki pula dengan memanfaatkan fitur-fitur yang ada dalam sebuah dokumen, semisal *hyperlink* pada pencarian di sebuah halaman web. Ochoa (2012) menyatakan bahwa analisis *backlink* (banyaknya *link* ke sebuah *website*) yang dipadukan dengan skor similaritas akan menghasilkan daftar hasil pencarian yang mempunyai kemungkinan tinggi mengandung informasi yang diinginkan oleh pengguna.

Penerapan algoritma similaritas dapat membantu proses klasifikasi dengan menentukan *tag* atau kata kunci yang paling tepat untuk sebuah dokumen. Pengklasifikasian kumpulan dokumen diperlukan pada sebuah perpustakaan digital untuk mengelompokkan dokumen dengan subjek yang sama (Boyack, dkk., 2011; Sun dkk., 2010). Algoritma similaritas juga diterapkan dalam proses deteksi plagiasi, yaitu dengan membandingkan dua dokumen atau lebih dan menentukan tingkat kemiripan dari paragraf-paragraf yang ada dalam dokumen (Malcolm dan Lane, 2008). Adapun dalam kegiatan analisis data, algoritma similaritas digunakan untuk mendefinisikan kata yang dicermati beserta kata sejenis untuk dihitung frekuensi kemunculannya dalam berita di dunia maya atau dalam obrolan di situs media sosial.

Similaritas dua buah kalimat dapat ditentukan dengan algoritma similaritas semantik, yaitu algoritma yang memperhatikan makna kata yang menyusun kalimat.

Penentuan similaritas secara semantik lebih akurat daripada perhitungan similaritas berdasarkan pencocokan kata (Mihalcea, Corley & Strapparava, 2006). Namun, penerapan algoritma similaritas semantik untuk teks bahasa Indonesia belum banyak dilakukan karena berbagai kendala di antaranya karena jejaring kata bahasa Indonesia belum tersedia baik secara gratis maupun komersial. Kendala lain adalah belum adanya kumpulan dokumen (atau korpus) berbahasa Indonesia yang diterima sebagai standar untuk melakukan pengujian algoritma similaritas (Asian, Williams & Tahaghoghi, 2005) sehingga penelitian tentang algoritma similaritas menjadi sangat minim.

Uraian beberapa paragraf di atas menunjukkan perlunya upaya observasi terhadap algoritma similaritas semantik pada kalimat bahasa Indonesia untuk dapat digunakan dalam berbagai aplikasi. Oleh karena itu, perlu diupayakan penyusunan basis data pengetahuan (*knowledge*) dalam konteks jejaring kata bahasa Indonesia kemudian mencari algoritma similaritas semantik yang terbaik. Perlu juga dibuat korpus standar berbahasa Indonesia yang sebagai alat uji dalam observasi algoritma similaritas. Yang menjadi pertanyaan kemudian adalah sejauh mana ***'Efektivitas Algoritma Similaritas Semantik Berbasis Jejaring Kata untuk Mengukur Kemiripan Kalimat Bahasa Indonesia?'*** Pertanyaan inilah yang akan dibahas dalam penelitian ini.