BAB II

LANDASAN TEORI

2.1 Semantic Similarity

Semantic similarity adalah gagasan yang dapat mengukur tingkat kemiripan dari kata yang dibandingkan. Perbandingan dapat terjadi dalam bentuk kata, kalimat pendek, dan dokumen. *Semantic similarity* memiliki peran yang penting dalam *Natural Language Processing* dan bidang lain yang berhubungan dengan *text classification, document clustering, text summarization,* dll. *Semantic similarity* juga merupakan pedoman matematika yang digunakan untuk membandingkan kekuatan dari perbedaan kata yang dibandingkan. Contohnya, mengetahui perbedaan antara sepeda dengan motor atau perbedaan dari mobil dan kuda. (Jatkina, dkk. 2019).

2.2 Natural Language Processing

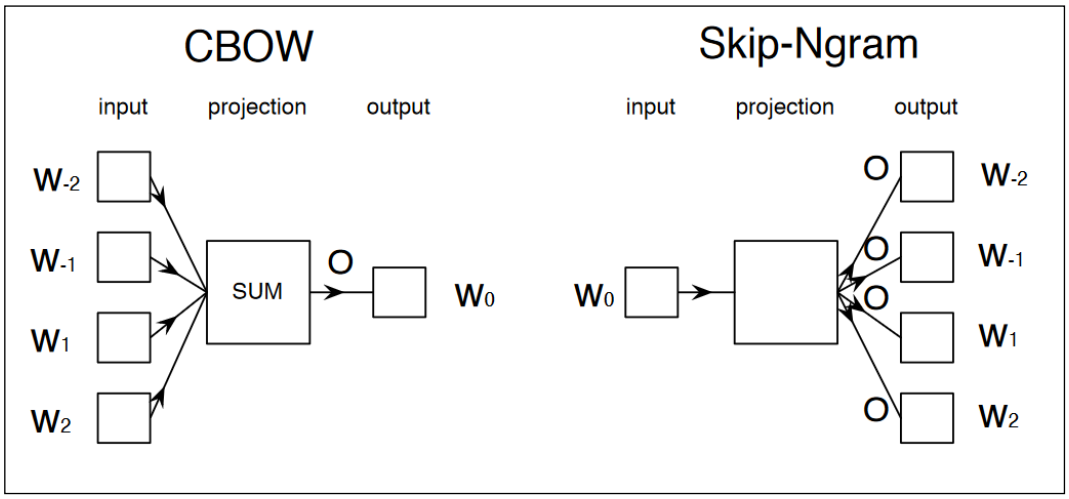
Natural language adalah bahasa yang digunakan oleh manusia. *Natural Language Processing* (NLP) mengikat semua yang dibutuhkan sebuah komputer untuk mengerti *natural language* dan juga menghasilkan *natural language*. *Natural Language Processing* merupakan salah satu bidang dari kepintaran semu dan ilmu bahasa, dikhususkan agar komputer dapat mengerti pernyataan atau kata yang ditulis dalam bahasa manusia yang biasa digunakan untuk berkomunikasi. (Chopra, dkk. 2013)

2.3 Word2Vec

Word2Vec adalah algoritma yang mengambil input berupa kumpulan tulisan yang banyak, lalu memproduksi ruang vektor yang memiliki ratusan

dimensi. Dengan setiap kata ditempatkan di dekat kata-kata lainnya yang memiliki arti yang mirip. (Mikolov, 2013)

Algoritma Word2Vec memiliki 2 jenis yaitu: *Continuous Bag of Words* (CBOW) dan Skip-Gram. CBOW merupakan cara memprediksi sebuah kata yang berhubungan dengan *input* yang berupa kumpulan dari beberapa kata. Sedangkan Skip-Gram memprediksi kata-kata yang berhubungan dengan satu *input*. Parameter *"window size"* digunakan untuk membatasi banyaknya kata yang dicari (Ling, 2015).



Gambar 2.1 Proses CBOW dan Skrip-Gram (Ling, 2015)

Word2Vec dapat mencari kemiripan dari kata dapat dengan *cosine similarity equation*. *Cosine similarity* merupakan perhitungan dari kemiripan dua vektor n-dimensi dengan melihat nilai sudut *cosine* dari dua kata. (Jatkina, dkk. 2019) Rumus dari *cosine similarity* adalah:

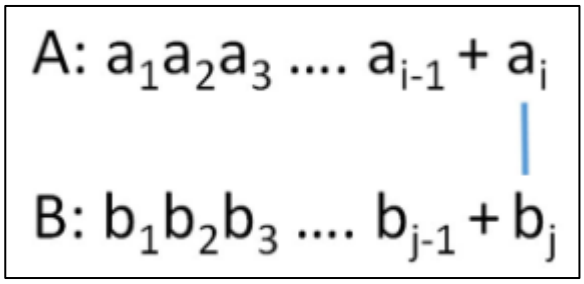
$$\text{similarity} = \cos\theta = \frac{\bar{x} \bullet \bar{y}}{\|\bar{x}\|\|\bar{y}\|} \quad (1)$$

Di mana:

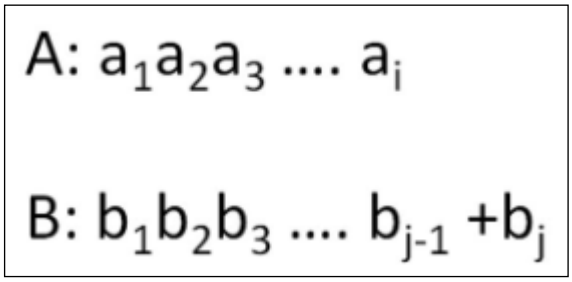
$\bar{x} \bullet \bar{y}$: Perkalian titik x dan y.
$\|\bar{x}\|$: Panjang vektor x.
$\|\bar{y}\|$: Panjang vektor y.

2.4 Damerau-Levenshtein Distance

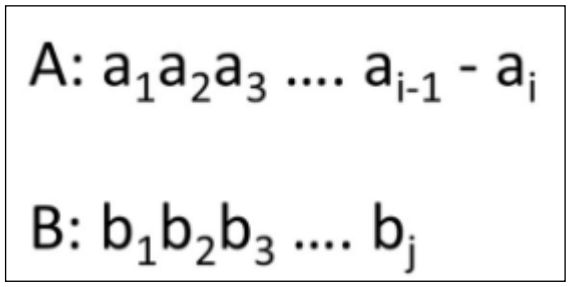
Damerau-Levenshtein Distance (DLD) adalah proses penggantian *string* A menjadi *string* B dengan angka penyuntingan paling sedikit. DLD mempunyai 4 tipe pernyuntingan yaitu: mengganti karakter a dari kata A menjadi karakter lainnya (*substitution*), menambahkan karakter ke kata A (*insertion*), menghapus karakter pada kata A (*deletion*), dan menukar dua karakter yang berdekatan (*transposition*). (Zhao dan Sahni, 2017). Setiap tipe penyuntingan yang dilakukan oleh DLD harus memiliki kondisi yang digambarkan pada Gambar 6.2 sampai Gambar 6.5.



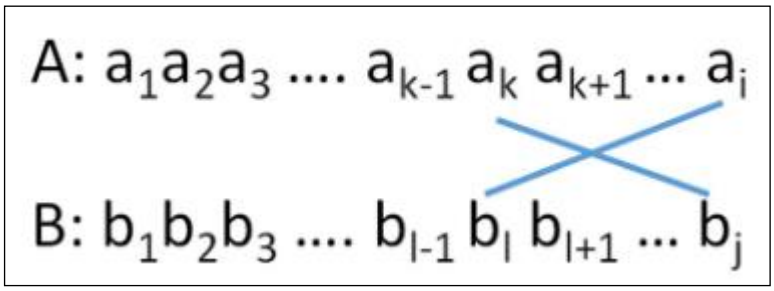
Gambar 2.2 Kondisi proses *substitution* (Zhao dan Sahni, 2017)



Gambar 2.3 Kondisi proses *insertion* (Zhao dan Sahni, 2017)



Gambar 2.4 Kondisi proses *deletion* (Zhao dan Sahni, 2017)



Gambar 2.5 Kondisi proses *transposition* (Zhao dan Sahni, 2017)

2.5 fastText

FastText adalah *library* yang dikeluarkan oleh Facebook yang merupakan pengembangan dari Word2Vec, di mana setiap kata digambarkan sebagai jumlah dari vektor dan setiap vektor menggambarkan sebuah potongan dari kata. Tujuan dari pengunaan metode *training* tersebut adalah untuk mendeteksi adanya penggunaan kata-kata yang jarang dipakai, atau *slang.* (Athiwaratkun dkk. 2018)

Pada setiap *keyword* yang di-*input*, *FastText* akan memotong kata menjadi n-gram karakter dan menambahkan sebuah batas yang digambarkan "<" sebagai batas awal dan ">" sebagai batas akhir. (Athiwaratkun dkk. 2018)