

1. Pendahuluan

1.1 Latar Belakang

Data digital pada era saat ini terdapat banyak data berbentuk teks yang digunakan untuk berbagai keperluan, oleh sebab itu perlu diadakan pengolahan data untuk mendapat informasi yang diperlukan dari data teks tersebut. Jenis pengolahan data teks tersebut salah satunya adalah keterkaitan semantik yang menghitung keterkaitan antar kata yang diolah. Contoh kata yang memiliki keterkaitan semantik yang tinggi adalah kata baju dengan kata pakaian yang memiliki keterkaitan tinggi karena baju merupakan salah satu jenis pakaian, oleh sebab itu kedua kata memiliki keterkaitan semantik.

Keterkaitan semantik dapat menunjukkan dua konsep yang saling terkait dan berhubungan dimana dengan menggunakan semua hubungan di antara teks, kata dan kalimat. Hubungan diantara mereka disebut dengan *hyponymic/hypernymic*, *meronymic*, dan semua jenis hubungan fungsional termasuk *has-part*, *is-made-of*, dan *s-an-attribute-of*. [1]. Contoh kata yang dapat saling berhubungan satu dengan kata yang lain adalah mobil dan motor yang kemudian akan dibandingkan dengan kata mobil dan rem, kata mobil dan motor memiliki kesamaan bahwa kata tersebut adalah kendaraan namun, kata mobil dan rem memiliki hubungan saling terkaitan satu sama lainnya.

PMI (*Pointwise Mutual Information*) telah dimunculkan sebagai salah satu pengukuran statistik *word similarity* yang tidak berdasarkan pada hipotesis distribusional. Menghitung PMI hanya memerlukan statistik yang sederhana dari dua buah kata, *marginal frequencies* dan *co-occurrence frequency* dalam kumpulan data tulisan [2]. PMI telah banyak diketahui memiliki kecenderungan memberikan skor yang tinggi untuk keterkaitan pada sepasang kata[9].

Tugas akhir ini penulis akan mengimplementasikan pendekatan PMI (*Pointwise Mutual Information*) dalam bentuk aplikasi untuk mengukur keterkaitan semantik antara satu pasang kata dengan beberapa konteks kata sederhana menggunakan dataset bahasa Indonesia Wikipedia dan skor yang dihasilkan akan dihitung nilai korelasinya dengan membandingkan skor hasil PMI dengan Skor *Gold Standard WordSim353*, *Simlex-999*, dan *Miller and Charles* sehingga kedepannya dapat digunakan untuk analisis teks dan pencarian informasi.

1.2 Perumusan Masalah

Terdapat beberapa perumusan masalah yang dapat diambil terkait latar belakang yang dibuat ialah:

1. Bagaimanakah mengimplementasikan keterkaitan semantic pada sepasang kata dalam Bahasa Indonesia dengan *Pointwise Mutual Information*?
2. Bagaimanakah hasil analisis dari implementasi kesamaan semantik antar kata berbahasa Indonesia dengan pendekatan *Pointwise Mutual Information* dan nilai korelasinya dengan *WordSim-353*, *Simlex-999*, dan *Miller and Charles*.

1.3 Tujuan

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

1. Mengukur *Pointwise Mutual Information* untuk mendapatkan nilai kesamaan semantik pada sepasang kata dengan beberapa konteks kata sederhana berbahasa Indonesia.
2. Melakukan analisis terhadap hasil implementasi semantik antar kata berbahasa Indonesia dengan pendekatan *Pointwise Mutual Information* dan nilai korelasinya dengan *WordSim-353*, *Simlex-999*, dan *Miller and Charles*.

1.4 Batasan Masalah

Batasan masalah pada Tugas Akhir ini ialah.

1. Sistem dibuat untuk melakukan pengukuran dan menghasilkan nilai kesamaan semantik dari sepasang kata yang selanjutnya dihitung nilai korelasi yang dimiliki oleh sistem dan *WordSim-353*, *Simlex-999*, dan *Miller and Charles*.
2. Database yang akan digunakan adalah Bahasa Indonesia Wikipedia.
3. *WordSim-353*, *Simlex-999*, dan *Miller and Charles* merupakan *Gold Standar* yang digunakan untuk membandingkan nilai keterkaitan yang dihasilkan oleh sistem.
4. Dalam Tugas Akhir ini system hanya mengenali inputan Bahasa Indonesia saja.

1.5 Rancangan Kegiatan

Rancangan kegiatan yang akan dilakukan untuk menyelaikan Tugas Akhir 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. Pengumpulan Data
Pengumpulan data dilakukan menggunakan aplikasi dan juga internet untuk membuat dataset wikipedia berbahasa Indonesia.
3. Perancangan Sistem
Tahap perancangan sistem pada tugas akhir ini di gambarkan menggunakan *flowchart*. *Flowchat* perancangan sistem sesuai dengan pendekatan yang di inginkan yaitu PMI. Tahap perancangan sistem

juga membantu penulis untuk mematangkan rancangan fungsionalitas pada sistem.

4. Pengujian dan Analisis

Pada tahap ini dilakukan pengujian terhadap sistem yang telah dibuat agar tidak terjadi *error* dan analisis terhadap hasil yang telah diperoleh.

5. Implementasi

Implementasi dilakukan bersamaan dengan pengumpulan data dan implementasi sesuai dengan perancangan sistem yang telah ditentukan sebelumnya.

6. Pembuatan Laporan

Pembuatan laporan dilakukan setelah hasil dari analisis yang dilakukan sebelumnya telah didapatkan.

1.6 Sistematika Penulisan

Tugas akhir dengan judul “**Implementasi dan Analisis Keterkaitan Semantik Berbahasa Indonesia dengan Pendekatan Pointwise Mutual Information**” 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 preprocessing data*, *flowchart* penghitungan bobot kata, 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.