

1. Pendahuluan

1.1 Latar Belakang

Pada era data digital ini terdapat banyak data berbentuk teks yang digunakan untuk berbagai keperluan, Oleh karena itu perlu diadakan pengolahan data untuk mendapat informasi yang diperlukan dari data teks tersebut salah satu jenis pengolahan data teks tersebut adalah keterkaitan semantik yang menghitung keterkaitan antar kata yang diolah. Salah satu contoh kata yang memiliki keterkaitan semantik yang tinggi adalah kata apel dengan kata buah yang memiliki keterkaitan tinggi karena apel merupakan salah satu jenis buah, oleh sebab itu kedua kata

Keterkaitan antar kata ini memiliki kegunaan yang banyak diterapkan untuk *natural language processing* (NLP), *information retrieval* (IR), dan *artificial intelligence*, termasuk penanganan disambiguasi word sense [1], deteksi malapropism [2], *paraphrase recognition* [3], dan *image and document retrieval* [4].

Salah satu jenis metode penghitungan keterkaitan semantik adalah *Pointwise Mutual Information* (PMI) yang dapat menghitung nilai *similarity* yang akurat kepada keterkaitan antar kata yang melibatkan kata yang jarang muncul dalam suatu dokumen atau database teks. Yang dalam penggunaannya tidak membutuhkan hipotesis distribusional. Banyak varian muncul dari *Pointwise Mutual Information* (PMI) yang memiliki kelebihan dan kekurangan masing-masing, salah satu varian dari PMI ini adalah *Positive Pointwise Mutual Information Cosine* (PPMIC) yang mengembangkan rumus dasar PMI untuk mendapatkan hasil yang sesuai dengan target.

PPMIC yang menggunakan PMI sebagai alat untuk pemberian bobot dan cosine sebagai alat penghitungan *similarity*. Yang Bullinaria dan Levy membuktikan dengan mendapatkan hasil kolerasi TOEFL sinonim dengan korpus BNC sebesar 80% [5].

Pada tugas akhir ini penulis akan mengimplementasikan pendekatan PPMIC dalam bentuk aplikasi untuk mengukur keterkaitan semantik antara sepasang kata dengan beberapa konteks kata sederhana menggunakan dataset yang disediakan oleh *WordSim353*, *Simlex-999*, dan *Miller and Charles* yang merupakan *gold standard* atau nilai acuan yang dinilai berdasarkan persepsi manusia dan membandingkan hasil pengukuran yang dihasilkan oleh aplikasi penulis dengan 3 *gold standard* tadi.

1.2 Perumusan Masalah

Terdapat beberapa perumusan masalah yang dapat diambil terkait latar belakang diatas.

1. Bagaimanakah mengimplementasikan kesamaan semantik pada sepasang kata dalam Bahasa Inggris dengan pendekatan *Positive Pointwise Mutual Information Cosine*?
2. Bagaimanakah hasil analisis dari implementasi kesamaan semantik antar kata berbahasa Inggris dengan pendekatan *Positive Pointwise Mutual Information Cosine* dan nilai korelasinya dengan *WordSim353*, *Simlex-999*, dan *Miller and Charles*.

1.3 Tujuan

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

1. Menerapkan pengukuran *Positive Pointwise Mutual Information Cosine* untuk mendapatkan nilai kesamaan semantik pada sepasang kata dengan beberapa konteks kata sederhana berbahasa Inggris.
2. Melakukan analisis terhadap hasil implementasi semantik antar kata berbahasa Inggris dengan pendekatan *Positive Pointwise Mutual Information Cosine* dan nilai korelasinya dengan *WordSim353*, *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 *WordSim353*, *Simlex-999*, dan *Miller and Charles*.
2. Dataset yang digunakan ialah *Brown Corpus* karena mengandung data kata mencapai 1 juta dari 15 kategori teks.
3. *WordSim353*, *Simlex-999*, dan *Miller and Charles* merupakan *Gold Standar* yang digunakan untuk membandingkan nilai keterkaitan yang dihasilkan oleh sistem.
4. Sistem hanya mengenali Bahasa Inggris sebagai masukan katanya karena menggunakan dataset *Brown Corpus*

1.5 Metodologi Penyelesaian Masalah

Metodologi yang digunakan 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. *Preprocessing Data*
Data mentah yang berasal dari dataset *Brown Corpus* harus dilakukan *preprocessing* untuk menghilangkan *noise* dan mendapatkan data yang diinginkan. *Preprocessing* yang dilakukan antara lain:
 - a. *Lemmatization*, merupakan proses merubah kata menjadi bentuk kata dasarnya.
 - b. *Stop Word Removal*, menghilangkan kata-kata yang sering muncul dan dianggap tidak memiliki arti penting seperti kata sambung. Sebagai contoh, kata “dan”, “atau”, “jika”, “karena”, dan lain-lainnya.
 - c. *Case Folding*, merupakan proses untuk mengubah semua huruf kapital menjadi huruf kecil untuk menghindari adanya masalah *case sensitive*.
3. Perancangan Sistem

Tahap perancangan sistem pada tugas akhir ini di gambarkan menggunakan *flowchart*. *Flowchat* perancangan sistem sesuai dengan pendekatan yang diinginkan yaitu PPMIC. Tahap perancangan sistem juga membantu penulis untuk mematangkan rancangan fungsionalitas pada sistem.

4. Implementasi

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

5. Pengujian dan Analisis

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

6. Pembuatan Laporan

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

1.6 Sistematika Penulisan

Tugas akhir dengan judul “**Second Order Co-ocurrence 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.