Implementasi dan Analisis Kesamaan Semantik Antar Kata Berbahasa Inggris dengan Metode Positive Pointwise Mutual Information Cosine *Implementation and Analysis Semantic Similarity Between Words in English with the Method of Positive Pointwise Mutual Information Cosine*

KD Krisna Dwipayana[1], Ir. M. Arif Bijaksana, M.Tech.,Ph.D.[2], Mohamad Syahrul Mubarok[3]

Prodi S1 Teknik Informatika, Fakultas Teknik, Universitas Telkom

[1] kdkrisnadwipayana@gmail.com , [2]arifbijaksana@telkomuniversity.ac.id, [3]msyahrulmubarok@telkomuniversity.ac.id

Abstrak

Keterkaitan semantik adalah salah satu jenis pengukuran yang ada pada *text mining* untuk menggambarkan bagaimana hubungan antara kata. Tujuan dari pengukuran keterkaitan semantik ini adalah untuk memperoleh nilai yang merepresentasikan seberapa besar keterkaitannya. *Pointwise Mutual Information* (PMI) merupakan salah satu pengukuran secara statistik untuk keterkaitan semantik yang telah banyak digunakan. Penerapan PMI diketahui mengalami bias untuk sepasang kata dengan frekuensi rendah, hal ini menyebabkan adanya pengembangan berupa varian pada pengukuran PMI untuk menghindari keadaan bias tersebut.

Positive Pointwise Mutual Information Cosine (PPMIC) merupakan salah satu varian yang digunakan dalam tugas akhir ini untuk menghitung keterkaitan semantik. Perhitungan nilai PPMIC dilakukan pada dataset yang didapat dari *brown corpus.* Nilai PPMIC yang didapat oleh aplikasi dihitung korelasinya dengan Word-Sim-353 yang merupakan indeks keterkaitan kata berdasarkan sudut pandang manusia. Hasil dari penelitian pada tugas akhir ini merupakan nilai korelasi antara skor yang dihasilkan sistem dengan *gold standard SimLex-999, WordSim353 dan Miller and Charles* yang akan menghasilkan nilai kolerasi yang akan menunjukan seberapa akurat metode pengukuran PPMIC.

Kata Kunci: Keterkaitan Semantik, Pointwise Mutual Information,Positive Pointwise Mutual Information Cossine.

Abstract

Semantic similiarity is one type of measurement in the text mining to describe how the relationship between words. The purpose of this semantic association measurement is to obtain a value that represents how much the association. Pointwise Mutual Information (PMI) is a statistical measurement of the semantic relationship that has been widely used. Application of PMI known to have a bias for the pair of words with low frequency, this led to the development of a variant form of the PMI measurements to avoid circumstances such bias. Positive pointwise Mutual Information Cosine (PPMIC) is one variant used in this thesis to calculate semantic similiarity.

PPMIC value calculation performed on datasets obtained from brown corpus. PPMIC value obtained by the application of computed correlation with Word-Sim-353 which is an index of words based on the similiarity of human standpoint. The results of the research in this thesis is the correlation between the scores generated by the gold standard system SimLex-999, WordSim353 and Miller and Charles will resulting correlation value that would show how accurate the measurement method PPMIC.

Keywords: Semantic Similiarity, Pointwise Mutual Information, Positive pointwise Mutual Information Cossine*.*

1. Pendahuluan

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 disambiguisasi 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 dapt menghitung nilai *similiarity* yang akurat kepada keterkaitan antar kata yang melibatkan kata yang jarang muncul dalam suatu dokumen atau database teks. Yang dalam pengunaannya 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 pemberiaan bobot dan cosine sebagai alat penghitungan *similiarity*. 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.

2. Dasar Teori

2.1 Pointwise Mutual Information

PMI adalah sebuah teknik perhitungan keterhubungan dari sebuah asosiasi yang digunakan dalam teori informasi dan statistik. Dalam perhitungan linguistik, PMI untuk dua istilah yang diberikan mengindikasikan kemungkinan untuk mencari satu istilah dalam teks dokumen yang mengandung istilah lainnya. Rumus umum dari PMI sendiri adalah :

$$\mathrm{PMI}(c1, c2) = \log \frac{fd(c1, c2) \cdot N}{fc1 \cdot fc2} \quad (1)$$

fd(c1,c2) adalah bahwa konsep c1 dan c2 terjadi secara bersamaan sedangkan fc1, fc2 adalah frekuensi kemunculan masing-masing, N merupakan total kata dalam korpus [5]. Meskipun banyak digunakan, PMI memiliki batasan umum yaitu pertama, PMI mungkin menghasilkan skor negatif atau positif yang mempersulit interpretasi dan tidak ada batasan nilai khusus. Kedua, PMI dikenal memberikan skor tinggi untuk pasangan kata yang kemunculan secara bersamaannya rendah [6].

2.2 Positive Pointwise Mutual Information Cosine

Positive Pointwise Mutual information Cossine adalah variant PMI yang mengabungkan Positive Pointwise mutual information dan Persamaan Cossine untuk mendapatkan hasil kemungkinan terbaik untuk ketrkaitn dan kesamaan kata. Sebelum membahas PPMIC secara keseluruhan kita akan memebahas bagian dari PPMIC sendiri

PPMI merupaka pengembangan dari PMI dengan mengubah semua nilai negative yang didapatkan pada hasil perhitungan PMI menjadi nilai positif yaitu nol, tujuan pengubahan nilai ini adalah untuk memperbesar kereleveanan pmi dengan hasil sesunguhnya.

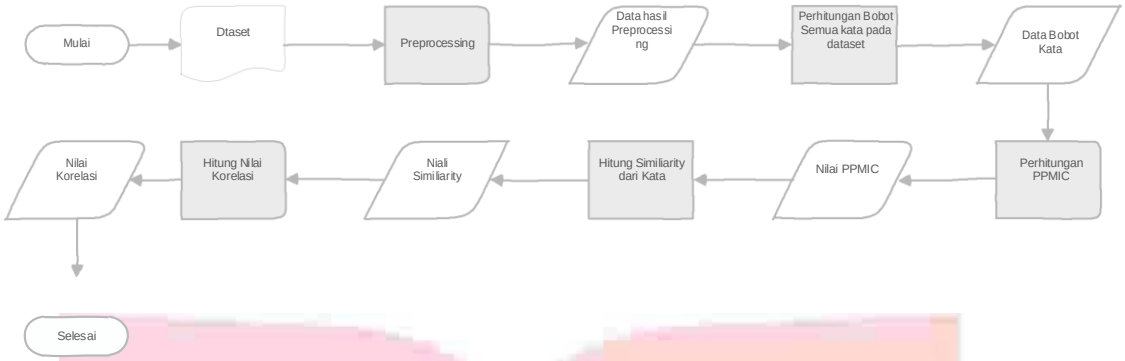
Sedangkan Cosine adalah metode pengukuran yang digunakan untuk kesamaan kata dalam vector di dalam NLP berdasarkan dot produk dari operator dalam linear algebra, yang sering disebut sebagai inner product, berikut adalah persamaan dot produk :

$$\text{dot-product}(\vec{a}, \vec{b}) = \sum_{i=1}^{N} a_i \cdot b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n \quad (2)$$

Yang akan menghasilkan persamaan cosine seperti berikut :

$$\text{cosine}(a, b) = \frac{\sum_{i=1}^{N} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \sqrt{\sum_{i=1}^{n} b_i^2}} \quad (3)$$

Dengan mengabungkan rumus cosine diatas dengan PPMI yang membantu agar range vector yang diahasilkan berjarak 0-1. Berikut contoh penerapan pada data PPMI yang sudah dihitung sebelumnya [7].

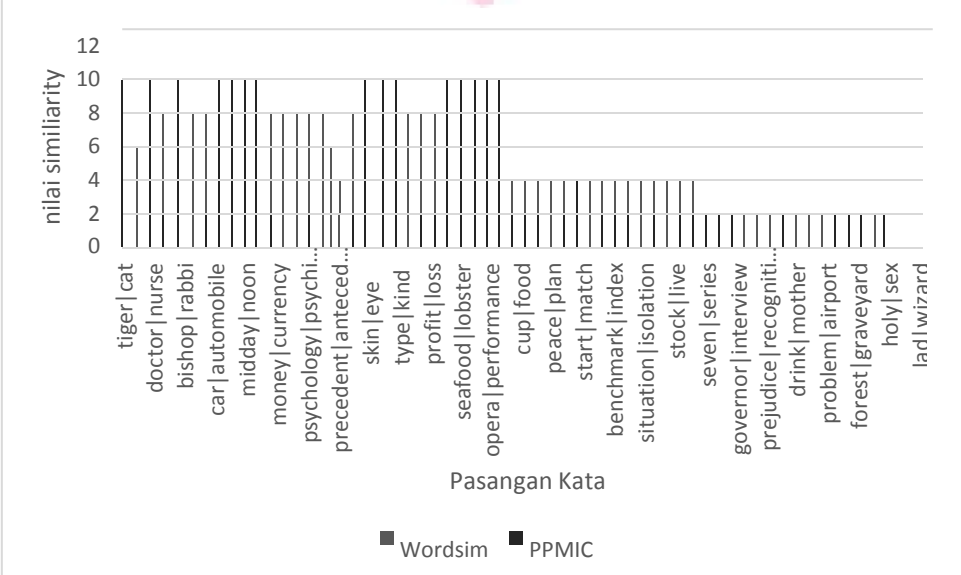


Gambar 1 Flowchart gambaran umum sistem

3. Pembahasan

3.1 Analisis Nilai *Semantic Similarity* Antar Kata PPMIC Berdasarkan Nilai Korelasi Terbaik

Pada analisis ini, akan dicari hubungan keterkaitan antara dua kata berdasarkan korelasi terbaik. Nilai hasil keluaran sistem terhadap seluruh pasangan kata pada dataset *gold standard* WordSim-353, Miller Charles dan Simlex-999 itu sendiri dengan menggunakan perhitungan korelasi Pearson. Pada pengujian ini dipisahkan ke dalam *window size* 11, 20 dan 25.



Gambar 2 Grafik skor PPMIC window size 11 dan skor dataset WordSim-353 semantic similarity

Dapat dilihat pada Gambar 2 yang memperlihatkan bagaimana korelasi antara sistem yang menggunakan *window size* 11 (garis berwarna hitam) dengan dataset *Gold Standard* WordSim-353 *semantic similarity* (garis berwarna abu-abu). Pada Gambar 2 tersebut terlihat bahwa sebagian besar pasangan kata sudah menghasilkan skor yang mendekati *gold standard*,namun bayak kata yang masih mamiliki kolerasi yang rendah karena banyaknya kata pada *gold standard (wordsim-353)* memiliki *co-occourance* yang rendah pada *brown corpus* .

Berikut hasil korelasi PPMIC dengan *window size* 15, 20 dan 25 pada dataset *gold standard* WordSim-353 Simlex-999 dan Miller-Charles dalam bentuk tabel yang diperlihatkan pada Tabel 1.

Tabel 1 Hasil korelasi PPMIC dengan Gold Standard

Nama Gold Stadard	Window size	Nilai Kolerasi
Wordsim-353	11	0.33
	20	0.27
	25	0.24
Simlex-999	11	-0.01
	20	-0.01
	25	-0.02
Miller and Charles	11	0.23
	20	0.21
	25	0.10

Dari hasil pengujian Tabel 1 dapat disimpulkan window size mempengaruhi hasil kolerasi, yang dimana nilai kolerasi yang makin besar menunjukan makin kuatnya hubungan hasil sistem dengan hasil *Gold Standrd*. Didapatnya nilai positif 0.33 sebagai nilai kolerasi yang menunjukan kolerasi yang lemah ayng menurut kolerasi pearson berkisar 0.1-0.5. Sedangkan adanya korelasi minus pada Gold Standard *Simlex-999* menunjukan nilai *similiarity* yang diperoleh sistem memiliki data yang berkorelasi terbalik, atau data yang trennya saling berlawanan. Perbedaan hasil korelasi, dan nilai similiarity yang diperoleh karena faktor window size yang menyebabkan konteks atau kata yang dibandingkan menjadi berbeda yang mempengaruhi perhitungan PPMIC. Dengan kata lain window size sangat mempengaruhi nilai similiarity, yang selanjutnya akan mempengaruhi nilai korelasi yang diperoleh.

4. Kesimpulan

Berdasarkan implementasi dan analisis pengujian yang dilakukan dapat ditarik kesimpulan sebagai berikut :

1. Sistem yang dibangun dapat mengimplementasikan perhitungan keterkaitan semantik antar kata dengan metode PPMIC pada pasangan kata dataset *gold standard* WordSim-353, Miller Charles, dan Simlex-999 dan memperoleh korelasi terbaik pada korelasi Pearson sebesar 0,33 dengan dataset *gold standard* WordSim-353 *semantic similarity*.
2. Parameter yang mempengaruhi nilai korelasi keterkaitan semantik antar kata menggunakan PPMIC adalah konteks kata antar kata yang digunakan dan dibandingkan.
3. Nilai *semantic similarity* antar kata sangat dipengaruhi oleh kemunculan kata tersebut pada korpus serta nilai *Co-Occurence* sepasang kata tersebut. Semakin tinggi *Co-Occurence* sepasang kata, maka akan semakin tinggi skornya.

Daftar Pustaka

[1] P. Resnik, "An Information Based Measure and Its Application to Problems of Ambiguity in Natural Language," dalam *Semantic Similarity in a Taxonomy*, J. Artificial Intelligence Research, 1999, pp. 95-130.

[2] A. Budanitsky , G. Hirst, "Evaluating Wordnet-Based Measures of Lexical Semantic Relatedness," *Computational Linguistics*, vol. 32, no. 1, pp. 13-47, 2006.

[3] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity," *Proc. 21st Nat'l Conf. Artificial Intelligence*, pp. 775-780, 2006.

[4] G. Varelas, E. Voutsakis, P. Raftopoulou, E. Petrakis, and E. Milios, "Semantic Similarity Methods in Wordnet and Their Application to Information Retrieval on the Web," *Proc. ACM Workshop Web Information and Data Management*, 2005.

[5] L. Han, T. Finin, P. McNamee, A. Joshi dan Y. Yesha, "Improving Word Similarity by Augmenting PMI with Estimates of Word Polysemy," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, pp. 1307-1319, 2013.

[6] Role Francois, Nadif Mohamed, "Handling The Impact Of Low Frequency Events On Co-Occurrence Based Measures Of Word Similarity," dalam *KDIR- International Conference on Knowledge Discovery and Information Retrieval*, Paris, 2011.

[7] J. H. M. Daniel Jrafsky, Speech and Language Processing, Prentice-Hall Inc, 2015.