Implementasi dan Analisis Kesamaan Semantik Antar Kata Bahasa Indonesia Menggunakan Metode Pointwise Mutual Information Max

I Gusti Ayu Chandra Devi[1], Moch. Arif Bijaksana[2], Indra Lukmana Sardi[3]
[1;2;3]Fakultas Informatika, Universitas Telkom, Bandung
[4]S1 Teknik Informatika
[1]yuchanevi@students.telkomuniversity.ac.id, [2]arifbijaksana@telkomuniversity.ac.id,
[3]indraluk@telkomuniversity.ac.id

Abstrak

Pencarian informasi sudah menjadi bagian dari kebutuhan manusia, terutama pencarian informasi meng-gunakan bahasa sehari – hari. Salah satu contohnya adalah Bahasa Indonesia. Dalam melakukan pencarian informasi yang efektif, diperlukan kecerdasan yang sama antara komputer dan manusia dalam mengolah informasi. Manusia terbantu dalam pencarian informasi karena manusia dapat mengolah kata yang di-gunakan dalam pencarian informasi. Manusia memiliki pengetahuan tentang hubungan satu kata dengan kata lainnya, sedangkan komputer tidak dapat mengetahuinya karena komputer tidak mengetahui sense dari satu kata tersebut. Agar komputer memiliki kecerdasan yang sama, dibutuhkan pencarian nilai ke-saman semantik(semantic similarity) antar kata. Berdasarkan ide tersebut, metode similarity yang dipilih untuk mencari nilai similarity antar kata Bahasa Indonesia adalah metode PMImax yang merupakan tu-runan dari metode PMI. Metode PMImax dipilih karena metode ini dapat menghasilkan nilai similarity berdasarkan kemuculan suatu kata di dalam suatu korpus. Metode ini juga menghasilkan nilai similarity yang baik saat diterapkan dalam Bahasa Inggris. Sehingga penelitian ini menguji apakah metode PMImax dapat diterapkan dalam pencarian nilai similarity dalam Bahasa Indonesia, dan seberapa baik metode ini saat diterapkan. Dengan menggunakan korelasi pearson hasil penelitian ini menunjukkan bahwa, metode PMImax cukup baik diterapkan dalam mencari nilai similarity dalam kata – kata Bahasa Indonesia diban-dingkan dengan metode PMI dan Word2Vec. Nilai korelasi yang dihasilkan, 0,26 pada Miller and Charles, 0,33 pada Simlex-999 dan 0,52 pada WordSim-353 Similarity.

Kata kunci : PMImax, PMI, Kesamaan Semantik, Kesamaan Semantik Antar Kata

Abstract

Searching for information is part of people's needs, specially in using colloquial. For example Bahasa. In searching for information effectively, human and computers need to have the same knowledge in processing the information. People can easily get the information, because people know how to process the word they need. They have knowledge about how one word relates to another words, but computers can't do that because computers don't know any sense of the words. Therefore, computers need to find similarity value for each words. Based on the idea, similarity's method that is choosen for calculating semantic similarity value between two words in Bahasa is PMImax that is a derivative from PMI method. This method was chosen because this method can give similarity value based on the words co-occurrence in a corpus. This method also gave a good result in English words. This study examines if this method can be implemented in Bahasa for calculating similarity value, and also examines how good this method in the implemantation. Using pearson correlation, the result of this study is PMImax gave good results when it is implemented in Bahasa compared to PMI and Word2Vec method. The correlation's scores are 0,26 in Miller and Charles, 0,33 in SimLex-999, 0,52 in WordSim-353 Similarity.

Keywords: PMImax, PMI, Semantic Similarity, Semantic Similarity Between words

1. Pendahuluan

1.1 Latar Belakang

Pencarian informasi sudah menjadi bagian dari kebutuhan manusia. Dalam melakukan pencarian informasi yang efektif, diperlukan kecerdasan yang sama antara komputer dan manusia dalam mengolah informasi. Manusia terbantu dalam pencarian informasi karena manusia dapat mengolah kata yang digunakan dalam pencarian infor-masi. Namun, manusia tidak dapat melakukan pencarian informasi di dalam sumber yang banyak dengan cepat. Sehingga membutuhkan komputer untuk membantu dalam pencarian informasi. Manusia memiliki pengetahuan tentang hubungan satu kata dengan kata lainnya, sedangkan komputer tidak dapat mengetahuinya karena kompu-ter tidak mengetahui sense dari satu kata tersebut. Agar komputer memiliki kecerdasan yang sama, dibutuhkan

pencarian nilai kesamaan semantik (semantic similarity) antar kata. Kesamaan semantik (semantic similarity) me-rupakan suatu metode untuk menghitung suatu kemiripan dari makna [8]. Sedangkan kesamaan semantik kata (word semantic similarity) merupakan sebuah perhitungan untuk menghitung seberapa mirip makna dari sepasang kata [2]. Terdapat dua pendekatan yang berlaku dalam perhitungan kesamaan semantik, yaitu berdasarkan pada penggunaan tesaurus (contohnya WordNet) dan statistik dari korpus yang besar [2]. Pendekatan dengan menggu-nakan korpus dengan task pencarian kata bersinonim dalam tes TOEFL memiliki akurasi lebih dari 80%.

Metode Pointwise Mutual Information (PMI) muncul sebagai metode yang terkenal untuk menghitung seman-tic similarity menggunakan korpus [2]. Perhitungan word semantic similarity menggunakan metode PMI dalam kata Bahasa Indonesia mendapatkan nilai akurasi yang masih lemah sebesar 0,256 pada penelitian [4]. Hal ini dapat terjadi karena metode PMI menghasilkan nilai similarity yang jauh lebih tinggi dibandingkan dengan nilai similarity dari gold standard. Sedangkan mengacu pada penelitian [9], pengunaan metode Pointwise Mutual Infor-mation Max (PMI max) dalam kata Bahasa Inggris mendapatkan nilai akurasi baik yaitu dengan nilai 0,665 yang disebabkan karena nilai similarity dari metode PMImax dapat mendekati nilai similarity dari gold standard.

Pada Tugas Akhir ini, akan dilakukan pencarian nilai semantic similarity untuk kata Bahasa Indonesia meng-gunakan metode PMI max. Dalam pencarian nilai tersebut, digunakan korpus Wikipedia Bahasa Indonesia. Untuk melihat tingkat akurasi dari metode yang digunakan, akan dilakukan pencarian nilai korelasi antara nilai semantic similarity yang dihasilkan metode PMI max dengan nilai yang ada pada gold standard. Perhitungan korelasi yang akan digunakan pada Tugas Akhir ini adalah korelasi pearson sedangkan Gold Standard yang digunakan adalah Simlex-999, WordSim-353 Similarity, dan Miller and Charles.

Hasil dari Tugas Akhir ini adalah untuk melihat apakah metode PMImax dapat diterapkan untuk mencari nilai similarity antar kata Bahasa Indonesia, serta mengevaluasi seberapa baik metode PMI max dalam penerapannya dilihat dari nilai korelasi.

1.2 Topik dan Batasannya

Rumusan masalah dari Tugas Akhir ini adalah metode PMI masih menghasilkan nilai korelasi yang lemah yaitu sebesar 0,256 saat mencari nilai similarity dalam kata - kata Bahasa Indonesia. Sedangkan batasan masalah dari Tugas Akhir ini adalah:

1. Dalam Tugas Akhir ini digunakan Korpus Wikipedia.
2. Tugas Akhir yang dikerjakan menggunakan gold standard WordSim 353, Simlex-999, Miller and Charles yang telah diterjemahkan mengacu pada penelitian sebelumnya.
3. Nilai variabel p dan q diambil dari penelitian sebelumnya, menggunakan nilai terbaik.

1.3 Tujuan

Tujuan dari penelitian ini adalah:

1. Menerapkan metode PMImax untuk mencari nilai similarity dalam kata - kata Bahasa Indonesia.
2. Menganalisis hasil korelasi Pearson, untuk melihat seberapa baik metode PMImax dalam mencari nilai semantic similarity dalam kata - kata Bahasa Indonesia berdasarkan hasil korelasi.

1.4 Organisasi Tulisan

Urutan penulisan laporan ini adalah sebagai berikut: Bagian 2 menunjukkan teori terkait dengan tugas akhir ini. Pembangunan sistem untuk metode PMImax dan Korelasi Pearson akan dijelakan di bagian 3. Pada bagian 4 akan dijelaskan mengenai hasil dan analisis sistem. Dan kesimpulan serta saran akan dijelakan pada bagian 5.

2. Studi Terkait

2.1 Semantic Similarity

Semantic similarity adalah suatu metode untuk menghitung suatu kemiripan dari makna [8]. Sedangkan kesa-maan semantik antar kata (word semantic similarity) merupakan sebuah perhitungan untuk menghitung seberapa mirip makna dari sepasang kata dengan melihat sinonim yang memiliki nilai tertinggi [2]. Terdapat beberapa kategori perhitungan dalam semantic similarity [8], yaitu Structure-based yang merupakan perhitungan semantic similarity menggunakan sebuah fungsi di dalam struktur hirarki ontology, information content yang merupakan perhitungan dengan mencari banyaknya kemunculan terms dari suatu koleksi dokumen yang diberikan, Feature-Based yaitu perhitungan dua terms dengan melihat gloss di wordnet, dan Hybrid Measure yang merupakan ga-bungan perhitungan - perhitungan sebelumnya. Tugas akhir ini, menggunakan jenis Information Content [6].

2.2 Pointwise Mutual Inoformation

PMI muncul sebagai metode yang terkenal dalam perhitungan word similarity [2]. PMI membutuhkan statistik sederhana antara 2 kata yaitu, frekuensi munculnya 2 buah kata secara individu (marginal frequency) dan secara bersamaan (co-occurrence frequency) [2]. Metode ini memiliki kekurangan yaitu menghasilkan nilai semantic similarity yang tinggi pada pasangan kata dengan probabilitas kemunculan rendah [2].

2.3 Pointwise Mutual Information max

Dalam pencarian nilai semantic similarity, terdapat beberapa pendekatan yang dapat digunakan, salah satu-nya adalah pendekatan PMImax. PMImax merupakan modifikasi dari metode PMI [2]. PMImax mengestimasi nilai maksimum korelasi antara dua kata, yaitu korelasi antara kedua makna terdekat [2]. Untuk melihat korelasi antara kedua makna terdekat, pada metode ini, akan dicari nilai dari makna pada setiap kata [2]. Sehingga diha-rapkan metode ini dapat memperbaiki kekurangan metode PMI, dan mencari similarity tanpa melihat probabilitas kemunculan kata. Contoh saat mencari kemiripan kata televisi dan kamera. Metode ini akan mencari banyaknya kemunculan pasangan kata ini secara individu serta secara bersamaan pada korpus yang digunakan. Kemunculan kata digunakan untuk melihat nilai sense dari kata. Lalu, dari nilai sense yang dihasilkan, dapat dilihat kemiripan dari kata televisi dan kamera.

$$\text{PMImax} = \log \frac{f\,d(w1,w2) \quad \frac{e^k}{N} \left(f\,w1 \,.\, f\,w2 \; \frac{f\,w1}{yw1} : \frac{f\,w2}{yw2} \right) : N}{\frac{f\,w1}{yw1} \; \frac{f\,w2}{fw2}} \tag{1}$$

Persamaan (1), merupakan rumus dari pendekatan PMImax. f d(w1; w2) merupakan nilai kemunculan kata 1(w1), dan kata 2(w2) secara bersamaan dalam suatu korpus. e^k merupakan konstanta dengan nilai tetapan 10. N merupakan jumlah kata dimiliki suatu korpus. f w1, dan f w2 merupakan kemunculan kata 1, dan kata 2 secara individu. yw1, dan yw2 merupakan nilai sense (makna) dari kata 1(w1) dan kata 2(w2). Pencarian nilai sense, dapat dilihat pada persamaan (2). Dimana nilai q dan nilai p merupakan sebuah variabel. q adalah variabel yang memiliki range nilai -6 sampai 10 dengan kelipatan 1. p adalah variabel yang memiliki range nilai 0 sampai 10 dengan kelipatan 0,5.

$$yw = \frac{(\log(f\,w) + q)^p}{(\log(700) + q)^p} \tag{2}$$

2.4 Metriks Evaluasi

Metriks evaluasi bertujuan untuk mengevaluasi kinerja dari suatu sistem yang telah dibangun. Salah satu cara untuk mengevaluasi sistem adalah dengan menghitung korelasi dari dua buah variabel. Salah satu perhitungan ko-relasi yang digunakan untuk menghitung dua buah variabel yang berhubungan secara linier dengan menggunakan korelasi pearson [1]. Korelasi pearson memiliki range nilai dari -1 sampai +1 [1]. Korelasi negatif merupakan ko-relasi dimana salah satu nilai variabel terus meningkat, sedangkan nilai variabel lain menurun [1]. Korelasi negatif memiliki nilai r < nol. Korelasi positif merupakan korelasi dimana nilai kedua variabel meningkat atau menurun secara bersamaan [1]. Korelasi positif memiliki nilai r > nol. Sedangkan korelasi yang memiliki nilai r = 0, disebut tidak memiliki korelasi [1]. Tidak memiliki korelasi karena disebabkan arah nilai dari kedua variabel tidak beraturan [1]. Rumus korelasi pearson dapat dilihat pada persamaan (3). x merupakan variabel 1 dan y merupakan variabel 2.

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}} \tag{3}$$

2.5 Korpus

Korpus adalah sebuah kumpulan teks dalam ukuran besar yang menjadi sebuah dasar dalam analasis linguistik [7]. Bentuk jamak dari korpus adalah corpora. Terdapat beberapa corpora yang terkenal seperti British National Corpus (BNC), COBUILD/Birmingham Corpus, IBM/Lancaster Spoken English Corpus [7]. Korpus dapat digo-longkan menjadi 2 jenis yaitu korpus terbuka dan korpus yang tertutup. Korpus terbuka merupakan suatu korpus yang di dalamnya tidak mengandung data atau informasi secara spesifik mengenai suatu bidang [7]. Sedangkan korpus tertutup merupakan korpus yang di dalamnya mengandung data atau informasi secara spesifik mengenai suatu bidang [7]. Dalam bentuknya korpus diambil dari artikel, buku sejarah, puisi, lagu, dll [7].

2.6 Gold Standard

Gold Standard adalah sebuah sumber yang digunakan untuk mengevaluasi model distribusi semantic [3]. Go-ld Standard telah menjadi standar yang telah terpercaya dimana dapat digunakan sebagai referensi yang akurat dan dapat diandalkan [3]. Gold standard yang paling sering digunakan dalam evaluasi nilai similarity adalah WordSim-353 [3]. Terdapat juga gold standard lain yang biasa digunakan yaitu Miller and Charles, Rubenstein-Goodenough, MEN, dan Simlex-999 [2][3].

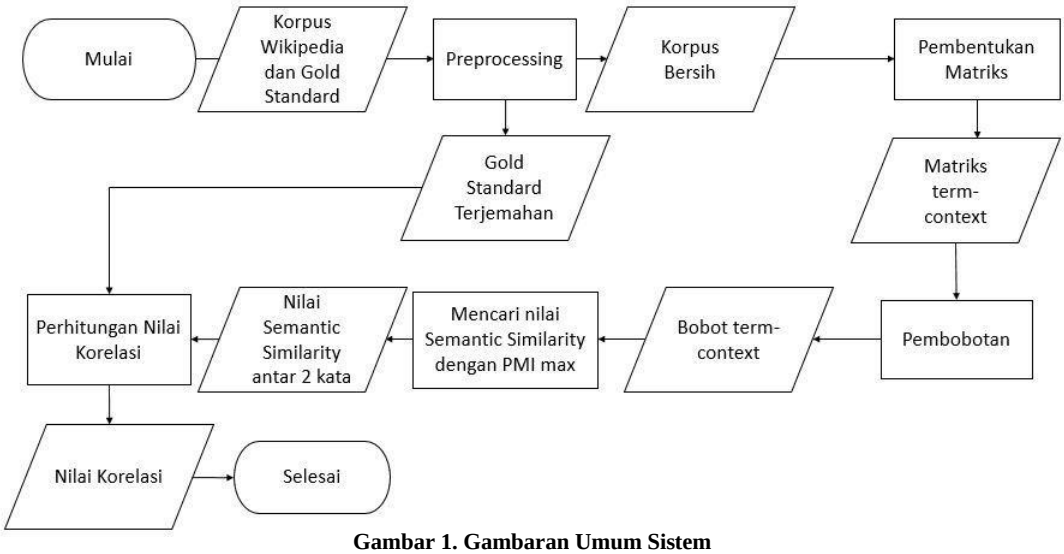
2.7 Tesaurus Bahasa Indonesia

Tesaurus Bahasa Indonesia merupakan suatu buku yang menjadi sumber informasi yang di dalamnya berisi kata - kata yang saling berkaitan maknanya [5]. Penyusunan tesaurus pada awalnya dilakukan berdasarkan tema, namun untuk memudahkan pengguna dalam pencarian kata, penyusunan tesaurus dirubah menjadi berdasarkan abjad [5]. Fungsi dari tesaurus sendiri adalah untuk membantu pengguna dalam menjelaskan gagasan yang ingin disampaikan menggunakan kata - kata yang berbeda namun makna yang sama [5].

3. Alur Sistem Kesamaan Semantik Antar Kata

Pada Tugas Akhir ini, dibangun sistem untuk menghitung semantic similarity berupa sebuah aplikasi dengan menggunakan bahasa pemrograman python. Sistem menggunakan 2 jenis dataset inputan yaitu korpus dan gold

standard dalam .txt untuk mencari nilai semantic similarity. Korpus yang digunakan adalah korpus Wikipedia Bahasa Indonesia, dan gold standard yang digunakan adalah Miller and Charles, Simlex-999, dan Wordsim-353 Similarity. Output akhir dari sistem yang dibangun adalah nilai Similarity dari metode PMI max dalam bentuk .csv, serta nilai korelasi dari korelasi pearson dalam .txt. Alur gambaran umum sistem terdapat pada Gambar 1.

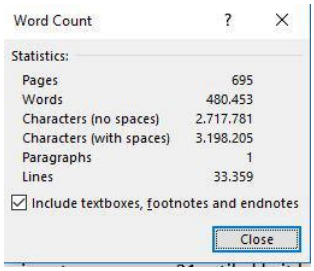


Gambar 1. Gambaran Umum Sistem

3.1 Dataset

3.1.1 Korpus Wikipedia Bahasa Indonesia

Korpus Wikipedia Bahasa Indonesia adalah sebuah dataset yang diambil dari artikel pilihan Wikipedia dalam Bahasa Indonesia [4]. Korpus yang digunakan pada Tugas Akhir ini diambil dari hasil crawling oleh Bapak Herry Sujaini yang merupakan seorang dosen Program Studi Teknik Informatika di Universitas Tanjungpura. Korpus ini berguna untuk pembuatan kata - kata unik serta pembuatan matriks term - context. Korpus Wikipedia memiliki karakteristik seperti yang ditampilkan pada Gambar 2.



Gambar 2. Karakteristik Korpus

3.1.2 Gold standard

Terdapat tiga Gold Standard yang digunakan yaitu Miller and Charles, Simlex-999, dan Wordsim-353 Simi-larity. Ketiga Gold Standard tersebut disimpan dalam format file .csv dan telah diterjemahkan ke dalam Bahasa Indonesia yang diambil dari penelitian [4]. Gold Standard ini sudah memiliki pasangan kata dan skor untuk seti-ap pasangan kata. Gold Standard digunakan untuk pencarian nilai semantic similarity dan juga dalam pencarian nilai korelasi untuk sistem. Dalam mencari nilai semantic similarity, Gold Standard yang digunakan, adalah Gold Standard tanpa skor similarity. Lalu hasil pengujian tadi akan menghasilkan nilai semantic similarity dari metode PMImax yang disimpan dalam file .csv. Sedangkan, untuk pencarian nilai korelasi, digunakan skor Gold Standard yang sudah ada, dan dikorelasikan dengan nilai semantic similarity dari PMImax menggunakan koreasi pearson. Hasil korelasi ini disimpan dalam file berbentuk .txt.

Pada Tabel 3a, ditampilkan contoh dataset Miller and Charles yang telah diterjemahkan beserta skor simi-laritynya dengan range skor 0 sampai 4. Pada Tabel 3b, ditampilkan potongan dataset simlex-999 yang telah diterjemahkan beserta skor similaritynya dengan range skor 0 sampai 10. Pada Tabel 3c, ditampilkan potongan dataset Wordsim-353 yang telah diterjemahkan beserta skor similaritynya dengan range skor 0 sampai 10.

3.2 Preprocessing

Prepocessing merupakan proses penting yang dilakukan sebelum melakukan pemrosesan utama. Preproces-sing dilakukan untuk mempermudah sistem memahami inputan yang diberikan. Preprocessing dilakukan terhadap kedua dataset yaitu pada korpus (Lampiran 8a) dan gold standard (Lampiran 8b). Pada korpus, prepocessing yang

No	Kata 1	Kata 2	Skor
1	mobil	automobil	3,92
2	pantai	darat	3,7
3	sulap	ahli	3,5
4	perapian	kompor	3,11
5	makanan	buah	3,08
6	burung	derek	2,97

(a)

No	Kata 1	Kata 2	Skor
1	tua	baru	1,58
2	pintar	cerdas	9,2
3	keras	sulit	8,77
4	senang	riang	9,55
5	keras	mudah	0,95
6	cepat	cepat	8,75

(b)

No	Kata 1	Kata 2	Skor
1	harimau	kucing	7,35
2	harimau	harimau	10
3	pesawat	mobil	5,77
4	melatih	mobil	6,31
5	televisi	radio	6,77
6	media	radio	7,42

(c)

Gambar 3. (a) Sample Dataset Miller and Charles,(b) Simlex-999,(c) Wordsim-353 Similarity

dilakukan yaitu stemming dan stopword removal. Stemming merupakan suatu proses pengubahan kata berimbuhan menjadi kata dasar. Kata - kata yang berimbuhan yang dimaksud, yaitu berimbuhan di depan, di belakang maupun di tengah kata. Contoh stemming dapat dilihat pada Tabel 1. Korpus hasil stemming ini akan dilanjutkan dengan proses stopword removal. Stopword removal merupakan proses menghilangkan kata - kata umum yang biasanya muncul dalam jumlah besar dan dianggap tidak memiliki makna. Contoh kata - kata umum seperti ”dan” , ”dari”, ”di”. Contoh stopword removal dapat dilihat pada Tabel 2. Setelah semua proses preprocessing terselesaikan, maka akan menghasilkan korpus hasil preprocessing yang siap digunakan.

Sedangkan untuk gold standard, preprocessing yang dilakukan adalah penerjemahan dari Bahasa Inggris ke Bahasa Indonesia. Ini dilakukan, karena gold standard yang ada, masih menggunakan Bahasa Inggris, sedangkan pada Tugas Akhir ini pencarian nilai semantic similarity dilakukan untuk kata Bahasa Indonesia. Contoh hasil gold standard terjemahan dapat dilihat pada Gambar 3a, 3b, dan 3c.

Tabel 1. Contoh Proses Stemming

No	Input	Output
1	lapisan	lapis
2	melingkupi	lingkup
3	permukaan	muka
4	mempunyai	punya

Tabel 2. Contoh Proses Stopword Removal

No	Input	Output
1	atmosfer adalah lapisan gas yang melingkupi sebuah planet	atmosfer lapisan gas melingkupi sebuah planet
2	di bumi atmosfer terdapat dari ketinggian 0 km	bumi atmosfer ketinggian 0 km

3.3 Pembentukan Matriks

Pembentukan matriks dilakukan untuk membantu proses pencarian nilai bobot. Dalam proses pembuatan ma-triks, digunakan kata unik yang diambil dari dataset korpus Wikipedia Bahasa Indonesia yang telah dipreprocessing. Proses pencarian kata unik tersebut dilakukan dengan menggunakan proses tokenisasi. Dari proses pencarian kata unik tersebut, dihasilkan kata - kata unik. Setelah mendapatkan kata - kata unik, maka sistem akan membentuk matriks term - term dari kata - kata unik tersebut. Proses pembentukan matriks dapat dilihat pada Lampiran 9.

3.4 Pembobotan

Proses pembobotan di sini dilakukan untuk mendapatkan nilai co-occurance atau kemunculan antara kata 1 dan kata 2 secara bersamaan. Dalam proses pembobotan, dibutuhkan matriks term - context yang telah dibentuk pada proses sebelumnya. Kata - kata unik yang membentuk matriks term - context akan dicari nilai bobotnya dengan mencari nilai kemunculan kata - kata tersebut pada korpus Wikipedia Bahasa Indonesia menggunakan window size berukuran 7 dan 25. Alur proses dapat dilihat pada Lampiran 10.

3.5 Perhitungan PMImax

Dalam sistem, akan dilakukan proses perhitungan PMImax yang bertujuan untuk mencari nilai semantic simi-larity. Dengan mendapatkan nilai semantic similarity, dapat dilihat seberapa erat hubungan 2 buah kata. Inputan sistem adalah berupa dataset gold standard tanpa skor similarity. Format file dataset ini adalah .txt. Setiap pasang-an kata akan dicek nilai bobotnya. Nilai bobot dari proses pembobotan berguna untuk memenuhi nilai frekuensi kata 1 dan kata 2 pada persamaan PMImax. Selanjutnya sistem akan menghitung total jumlah kata yang ada pada korpus. Dilanjutkan dengan menghitung nilai sense pada kata 1 dan kata 2 dengan menggunakan rumus sense pada persamaan (2). Setelah semua proses dilakukan dan variabel telah terpenuhi, sistem akan langsung menghitung nilai semantic similarity menggunakan PMImax. Dari proses perhitungan nilai semantic similarity akan dihasilkan nilai semantic similarity antar sepasang kata. Alur proses dapat dilihat pada Lampiran 11.

3.6 Perhitungan korelasi

Pada proses perhitungan korelasi, dibutuhkan 2 data yaitu data skor semantic similarity pada sistem dan skor similarity pada gold standard. Kedua data tersebut akan dilakukan perhitungan korelasi menggunakan korelasi pearson. Dari proses perhitungan korelasi, dihasilkan nilai korelasi dari kedua data. Nilai korelasi ini, berguna untuk mengevaluasi sistem yang dibangun, apakah sistem yang dibangun cukup baik untuk mencari nilai semantic similarity antar kata Bahasa Indonesia. Alur sisem dapat dilihat pada Lampiran 12.

4. Hasil Pengujian dan Analisis

Pada bagian ini, akan dijelaskan mengenai hasil pengujian serta analisis dari hasil similarity yang dihasilkan metode PMImax dalam kata - kata Bahasa Indonesia.

4.1 Hasil Pengujian Penerapan Metode PMImax

Tujuan utama dari penelitian Tugas Akhir ini, untuk melihat apakah metode PMImax dapat menghasilkan nilai semantic similarity dalam kata - kata Bahasa Indonesia serta mengevaluasi seberapa baik metode tersebut. Di dalam pegujiannya, metode PMImax berhasil mendapatkan hasil similarity untuk kata - kata dalam Bahasa Indonesia. Untuk mengevaluasi seberapa baik penerapan metode ini, hasil similarity yang diperoleh dari metode ini akan dibandingkan dengan skor similarity pada gold standard menggunakan korelasi pearson (persamaan(3)).

Tabel 3. Perbandingan Nilai Korelasi antara PMImax dengan PMI dan Word2Vec dalam 3 Gold Standard

No	Metode	Miller and Charles	Simlex-999	Wordsim-353
1	PMImax	**0,26**	**0,33**	**0,52**
2	PMI	0,09	0,01	0,16
3	Word2Vec	-0,1	-0,01	0,24

Dari Tabel 3 dapat dilihat nilai evaluasi dari metode PMImax yang diterapkan. Nilai korelasi yang dihasilkan pada ketiga gold standard masih memiliki nilai < 1. Hal ini disebabkan karena nilai similarity yang dihasilkan sistem masih memiliki banyak perbedaan dengan nilai similarity pada gold standard. Beberapa pasangan kata tidak memiliki nilai similarity setelah dicari menggunakan metode PMImax. Namun pada Gold Standard, pasangan tersebut memiliki nilai similarity. Hal ini menyebabkan nilai korelasi yang dihasilkan masih < 1. Namun, nilai korelasi dapat ditingkatkan dengan menghapus nilai similarity yang memperoleh nilai 0.

Namun, walaupun hasil korelasi masih < 1, metode PMImax menghasilkan nilai korelasi positif pada ketiga gold standard. Hal ini menunjukkan metode ini dapat dievaluasi menggunakan ketiga gold standard tersebut.

Saat dibandingkan dengan metode PMI dan Word2Vec, metode PMImax memberikan hasil nilai korelasi ter-baik. Metode ini menghasilkan nilai korelasi tertinggi yaitu 0,52 yang menginterpretasikan bahwa korelasi yang dihasilkan adalah korelasi sedang. Hal ini menunjukkan metode PMImax baik diterapkan untuk mencari nilai similarity pada kata - kata Bahasa Indonesia.

4.2 Analisis Hasil Nilai Similarity dari Metode PMImax

Pada bagian ini akan dijelaskan mengenai analisis hasil nilai similarity yang dihasilkan dari metode PMImax. Korpus yang digunakan dalam pencarian nilai semantic similarity adalah korpus Wikipedia Bahasa Indonesia. Sedangkan terdapat 4 dataset yang digunakan yaitu Miller and Charles, Simlex-999, WordSim-353 Similarity dan Tesaurus Bahasa Indonesia. Dataset tesaurus ditambahkan untuk memperkuat analisis hasil nilai similarity pada metode PMImax. Untuk dapat menganalisis pola dari nilai similarity yang dihasilkan, digunakan dua jenis ukuran variabel window size yaitu 7 dan 25.

4.2.1 Pengaruh Kemunculan Kata Secara Individu Terhadap Hasil Similarity

Gambar 4a, 4b, 4c dan 4d merupakan potongan nilai similarity serta jumlah kemunculan kata pada korpus. Da-pat dilihat pada Gambar 4, terdapat beberapa pasang kata yang memiliki nilai similarity baik meningkat maupun tetap. Namun, ada beberapa pasang kata yang tidak memiliki nilai similarity. Pada Miller and Charles sebanyak 9 pasang kata dari 19 pasang kata, pada Simlex-999 sebanyak 541 dari 946 pasang kata, pada WordSim-353 Si-milarity sebanyak 71 dari 162 pasangan kata, dan pada Tesaurus Bahasa Indonesia sebanyak 18 dari 49 pasang kata tidak memiliki nilai similarity. Dari Tabel dapat dilihat bahwa kemunculan kata sangat mempengaruhi nilai similarity suatu pasang kata. Saat salah satu kata tidak terdapat pada korpus, maka pasangan kata tersebut tidak memiliki nilai similarity. Contohnya pada pasangan kata ”perapian” dan ”kompor” pada Tabel 4a. Kata ”pera-pian” tidak terdapat pada korpus, sehingga membuat pasangan kata tersebut tidak memiliki nilai bobot serta nilai similarity. Namun, seperti pada Tabel 4b, pada pasangan kata ”senang” dan ”riang” kedua kata tersebut muncul di dalam korpus. Namun, karena kemunculuan masing - masing kata tersebut tidak banyak, sehingga membuat pasangan kata tersebut tidak memiliki nilai bobot, namun tetap memiliki nilai similarity walaupun nilainya ti-dak besar. Kemunculan kata ini juga memengaruhi hasil nilai similarity pada dataset Tesaurus Bahasa Indonesia. Dengan mengetahui Tesaurus berisi sekumpulan kata bersinonim, seharusnya setiap pasangan kata pada dataset ini memiliki nilai similarity. Namun, dapat dilihat pada Tabel 4d, beberapa pasangan kata tidak memiliki nilai similarity, ini disebabkan karena metode PMImax menghasilkan nilai similarity berdasarkan kemunculan kata - kata pada korpus. Dan beberapa kata tersebut tidak muncul di dalam korpus. Dalam pengujian, ditemukan bahwa penyebab dari tidak munculnya kata di dalam korpus karena topik pada korpus tidak mengandung kata - kata pada gold standard. Selain itu juga karena dilakukan stemming pada korpus sehingga, kata - kata seperti kata kerja yang terdapat pada dataset gold standard tidak muncul pada korpus. Contoh kata - kata kerja adalah ”meminta” dan ”mengaku” yang terdapat pada dataset gold standard simlex-999 (terlampir).

Gold Standard 1 - ws 7				Gold Standard 1 - ws 25			
I	II	III	IV	I	II	III	IV
automobil	1	1	3,55	automobil	1	3	4,43
mobil	297			mobil	297		
pantai	219	14	3,66	pantai	219	35	4,49
darat	420			darat	420		
sulap	2	0	2,28	sulap	2	0	2,28
ahli	150			ahli	150		
perapian	0	0	0	perapian	0	0	0
kompor	1			kompor	1		
makanan	0	0	0	makanan	0	0	0
buah	2472			buah	2472		
burung	67	0	0	burung	67	0	0
derek	0			derek	0		

(a)

Gold Standard 2 - ws 7				Gold Standard 2 - ws 25			
I	II	III	IV	I	II	III	IV
tua	279	0	0	tua	279	0	0
baru	0			baru	0		
pintar	17	2	4,23	pintar	17	5	5,06
cerdas	28			cerdas	28		
keras	143	1	2,54	keras	143	6	3,63
sulit	153			sulit	153		
senang	35	0	2,29	senang	35	0	2,29
riang	2			riang	2		
keras	143	5	3,29	keras	143	16	4,25
mudah	246			mudah	246		
cepat	406	468	6,72	cepat	406	652	7,05
cepat	406			cepat	406		

(b)

Gold Standard 3 - ws 7				Gold Standard 3 - ws 25			
I	II	III	IV	I	II	III	IV
harimau	7	0	2,29	harimau	7	0	2,29
kucing	17			kucing	17		
harimau	7	9	6,51	harimau	7	17	7,14
harimau	7			harimau	7		
pesawat	1080	1	1,12	pesawat	1080	5	2,07
mobil	297			mobil	297		
melatih	0	0	0	melatih	0	0	0
mobil	297			mobil	297		
televisi	108	16	5,03	televisi	108	34	5,75
radio	61			radio	61		
media	140	1	2,8	media	140	8	4,27
radio	61			radio	61		

(c)

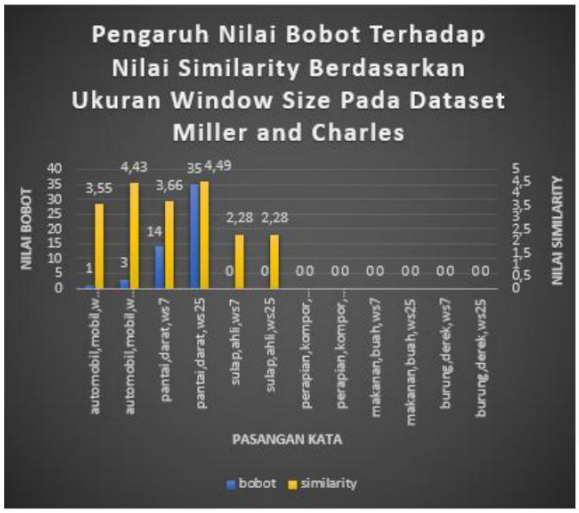
Tesaurus - ws 7				Tesaurus - ws 25			
I	II	III	IV	I	II	III	IV
abadi	39	1	3,56	abadi	39	1	3,56
awet	17			awet	17		
abjad	123	18	5,12	abjad	123	40	5,9
aksara	55			aksara	55		
acara	140	0	2,27	acara	140	0	2,27
agenda	4			agenda	4		
adik	28	0	0	adik	28	0	0
adimas	0			adimas	0		
adopsi	67	0	0	adopsi	67	0	0
mengang	0			mengangkat	0		
aduk	4	0	2,3	aduk	4	0	2,3
baur	4			baur	4		

(d)

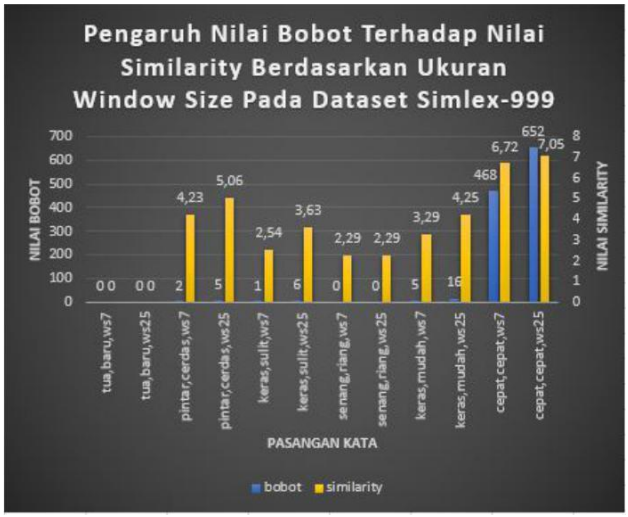
Gambar 4. (a) Sample Hasil Kemunculan Kata dan Nilai Similarity Dataset Miller and Charles,(b) Simlex-999,(c) Wordsim-353 Similarity,(d) Tesaurus Bahasa Indonesia. Ket. Tabel: (I) Pasangan Kata; (II) Nilai Kemunculan Kata Secara Individu di Dalam Korpus; (III) Nilai Kemunculan Pasangan Kata Secara Ber-samaan di Dalam Korpus; (IV) Hasil Nilai Similarity dari Sistem

4.2.2 Pengaruh Nilai Bobot Terhadap Nilai Similarity

Peningkatan nilai similarity sangat dipengaruhi oleh peningkatan nilai bobot pada setiap pasangan kata. De-ngan meningkatnya nilai bobot, nilai similarity juga ikut meningkat. Meningkatnya nilai bobot, dikarenakan oleh ukuran window size. Dapat dilihat pada Gambar 5a, 5b, 6a dan 6b. Nilai bobot meningkat saat ukuran window size diubah. Perubahan window size dapat memperluas pencarian kemunculan pasangan kata pada korpus. Semakin tinggi ukuran window size semakin luas pencarian yang dapat dilakukan. Saat pencarian pasangan kata semakin luas, maka nilai bobot meningkat. Saat nilai bobot meningkat, nilai similarity juga meningkat.

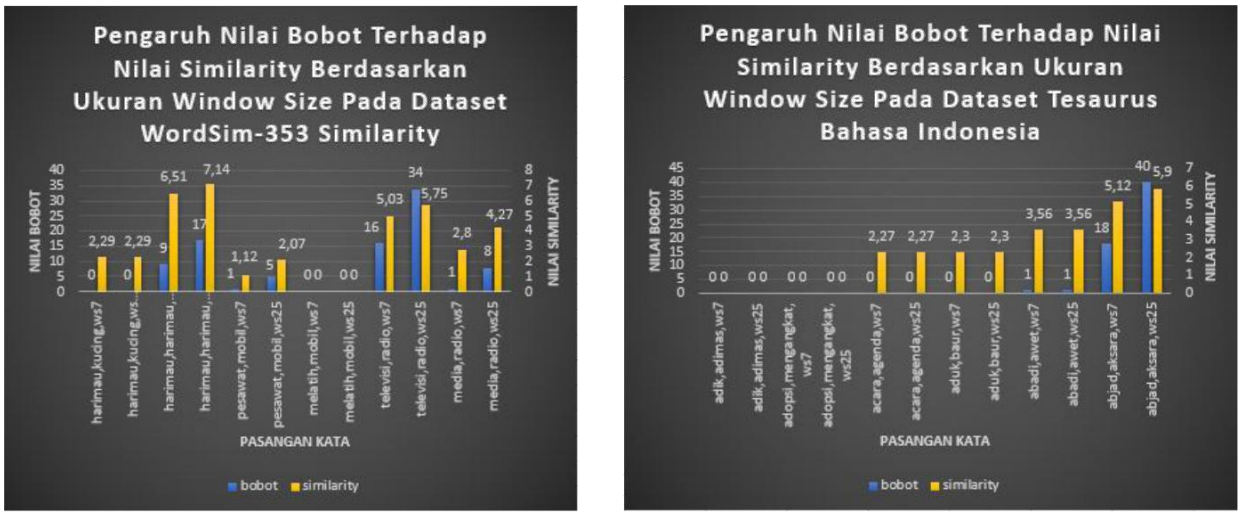


(a)



(b)

Gambar 5. (a) Pengaruh Nilai Bobot Terhadap Nilai Similarity pada Dataset Miller and Charles, (b) Simlex-999



(a) (b)

Gambar 6. (a) Pengaruh Nilai Bobot Terhadap Nilai Similarity pada Dataset WordSim-353 Similarity,(b) Tesaurus Bahasa Indonesia

5. Kesimpulan dan Saran

Dari pengujian dan analisis yang telah dilakukan pada penelitian ini, dapat disimpulkan bahwa :

1. Dengan hasil nilai korelasi tertinggi yang dihasilkan metode PMImax sebesar 0,52 yaitu korelasi sedang, dapat disimpulkan bahwa metode ini dapat diterapkan dalam mencari nilai semantic similarity dalam Bahasa Indonesia.
2. Metode PMImax baik diterapkan untuk mencari nilai similarity dalam kata - kata Bahasa Indonesia diban-dingkan dengan metode PMI dan Word2Vec.
3. Metode PMImax menghasilkan nilai korelasi positif pada gold standard Miller and Charles, Simlex-999 dan WordSim-353 sehingga metode ini dapat dievaluasi menggunakan ketiga gold standard ini.
4. Dalam metode PMImax, kemunculan kata secara individu di dalam korpus sangat menentukan apakah pa-sangan kata tersebut memiliki nilai similarity atau tidak.
5. Dalam metode PMImax kemunculan kata secara bersamaan(bobot) dipengaruhi oleh ukuran window size dan banyaknya kemunculan suatu kata secara individu di dalam suatu korpus.
6. Peningkatan nilai kemunculan kata secara bersamaan (bobot) memengaruhi peningkatan nilai similarity yang dihasilkan metode PMImax.

Sedangkan saran yang dapat diberikan untuk penelitian selanjutnya adalah :

1. Dalam penelitian selanjutnya, disarankan untuk mencoba tidak melakukan stemming pada korpus untuk menyesuaikan kata - kata yang terdapat pada gold standard terutama kata kerja.
2. Sangat disarankan untuk menggunakan korpus yang memiliki kata - kata dan topik yang sama dengan gold standard, saat menggunakan metode PMImax sebagai metode pencarian nilai similarity.

Daftar Pustaka

[1] N. S. Chok. Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data. PhD thesis, University of Pittsburgh, 2010.

[2] L. Han, T. Finin, P. McNamee, A. Joshi, and Y. Yesha. Improving word similarity by augmenting pmi with estimates of word polysemy. IEEE Transactions on Knowledge and Data Engineering, 25(6):1307–1322, 2013.

[3] F. Hill, R. Reichart, and A. Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. Computational Linguistics, 41(4):665–695, 2015.

[4] P. G. B. S. Parta. Implementasi dan Analisis Keterkaitan Semantik Berbahasa Indonesia dengan Pendekatan Pointwise Mutual Information. Telkom University, 2017.

[5] T. Redaksi. Tesaurus bahasa indonesia pusat bahasa. Pusat Bahasa, Departemen Pendidikan Nasional, 2008.

[6] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007, 1995.

[7] E. Septiandri. Rancang bangun aplikasi information retrieval untuk mengkoleksi data paralel korpus teks bahasa inggris–bahasa indonesia. Jurnal Sistem dan Teknologi Informasi (JustIN), 3(2), 2015.

[8] T. Slimani. Description and evaluation of semantic similarity measures approaches. arXiv preprint arXiv:1310.8059, 2013.

[9] I. M. D. Yoga. Implementasi dan Analisis Keterkaitan Semantik Antar Kata Menggunakan Pointwise Mutual Information max dengan Estimasi dari Kata Polisemi. Telkom University, 2016.