Implementasi dan Analisis Kesamaan Semantik Antar Kata Bahasa Indonesia Menggunakan Metode GloVe

Ramanti Dwi Indrapurasih[1], M. Arif Bijaksana[2],Indra Lukmana Sardi[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[4]S1 Teknik Informatika
[1]ramantidwi@students.telkomuniversity.ac.id, [2]arifbijaksana@telkomuniversity.ac.id,
[3]indraluk@telkomuniversity.ac.id

Abstrak

Kesamaan semantik adalah salah satu pengukuran yang ada pada *text mining* untuk mencari nilai kesamaan makna antar kata. Kesamaan semantik ini dapat diaplikasikan pada berbagai aplikasi. Pengukuran ini dilatarbelakangi dimana saat ini komputer belum dapat menyamakan persepsi manusia terkait penilaian kesamaan antar kata. Maka dari itu dalam tugas akhir kali ini membahas mengenai kesamaan semantik antar kata bahasa Indonesia dengan menggunakan metode GloVe . Metode GloVe adalah suatu model untuk *unsupervised learning* pada representasi kata yang mengungguli model lainnya di *word analogy, word similarity,* dan *named entity recognition*. Dengan inputan berupa corpus Wikipedia Bahasa Indonesia dan skor yang dihasilkan dihitung nilai korelasinya menggunakan *correlation pearson* dengan membandingkan skor hasil gold standard dari WordSim-353, SimLex-999 dan Miller Charles. Hasil dari penelitian tugas akhir ini merupakan nilai korelasi antara metode GloVe dengan *gold standard* SimLex-999, WordSim353, dan Miller Charles. Pada penelitian tugas akhir ini menghasilkan nilai korelasi pada *gold standard* dengan nilai korelasi yang didapatkan sebesar 0.1165 untuk Miller Charles, 0.2280 untuk SimLex-999 dan 0.2849 untuk WordSim-353.

Kata kunci : Text mining, Kesamaan Semantik, GloVe

Abstract

Semantic similarity is one of the text mining's measurement to find the value of the similarity between word's meaning. This semantics similarity can be applied in various applications. The measurement's background is caused where the computer not able yet to equate human's perspective related to measurement of the similarity between words. Therefore, this thesis will discuss about semantics similarity between words in Bahasa Indonesia by using GloVe method. GloVe method is a model for unsupervised learning on words representation that surpass another models in word analogy, word similarity and named entity recognition. With the input of a Wikipedia corpus of Bahasa Indonesia and the correlation value from resulted score is calculated with correlation pearson by comparing it with gold standard score from WordSim-353, SimLex-999, and Miller Charles. The final result from this thesis produce a correlation value in gold standard with the obtained correlation value is 0.1165 for Miller Charles, 0.2280 for SimLex-999 and 0.2849 for WordSim-353.

Keywords: Text mining, Semantics Similarity, GloVe

1. Pendahuluan

1.1 Latar Belakang

Dewasa ini teknologi berkembang sangat pesat, salah satu bidangnya yaitu teknologi informasi. Berbagai cara untuk mendapatkan informasi yang ada, seperti mendapatkan informasi seberapa besar nilai kesamaan semantik antar sepasang kata dalam suatu dokumen. Kesamaan semantik memiliki peran penting dalam beberapa *task* dari *Natural Language Processing* dan beberapa bidang terkait seperti *text classification, document clustering, text summarization,* dan lain sebagainya [4]. Kesamaan semantik adalah *task* pada *Natural Language Processing* (NLP) untuk mengukur kesamaan atau kedekatan antara pasangan kata secara semantik. Pasangan kata dikatakan memiliki kesamaan semantik jika pasangan tersebut memiliki makna atau konsep yang sama. Penelitian ini di dasari dimana komputer belum dapat menyamakan persepsi manusia terkait penilaian dari makna pasangan kata yang memiliki kesamaan semantik. Atas dasar ini lah,tugas akhir dibentuk agar komputer dapat menghasilkan informasi seberapa besar nilai kesamaan semantik antar sepasang kata dalam suatu dokumen. Untuk mencari atau menghitung kesamaan semantik antar kata ada beberapa metode yang bisa digunakan salah satunya menggunakan

metode berbasis vektor. GloVe merupakan suatu metode *unsupervised learning* pada representasi kata yang mengungguli model lainnya di *word analogy, word similarity,* dan *named entity recognition* [8]. Sebagian besar metode vektor bergantung pada jarak atau sudut di antara vektor pasangan kata sebagai metode utama untuk mengevaluasi kualitas representasi kata tersebut. Penggunaan metode berbasis vektor memberikan kinerja mutakhir untuk kesamaan semantik dan metode berbasis vektor dengan mudah diadaptasi di tiga basis pengetahuan yang berbeda (WordNet, Wikipedia, dan Wiktionary) [11]. Hal ini menunjukan bahwa metode berbasis vektor dapat menjadi pilihan yang baik dan metode Glove umumnya menghasilkan nilai kesamaan yang tinggi. Dalam metode berbasis vektor ini untuk mengukur kesamaan semantik antar sepasang kata menggunakan fungsi kesamaan kosinus. Fungsi kesamaan kosinus adalah fungsi yang umumnya digunakan pada sebagian besar metode berbasis vektor untuk menghitung nilai kesamaan antar dua vektor kata. Pada penelitian tugas akhir ini diimplementasikan metode GloVe untuk mengukur kesamaan antar pasangan kata menggunakan korpus Wikipedia bahasa Indonesia dan skor yang dihasilkan akan dihitung korelasinya menggunakan *correlation pearson* dengan membandingkan skor *gold standard* dari WordSim353, Sim-Lex999 dan Miller Charles.

1.2 Topik dan Batasannya

Pada penelitian ini memiliki rumusan masalah sebagai berikut :

1. Bagaimana mengimplementasikan kesamaan semantik sepasang kata dalam bahasa Indonesia dengan menggunakan metode GloVe?
2. Bagaimana hasil Analisis nilai korelasi yang dihasilkan menggunakan metode GloVe dengan perbandingan dataset *gold standard* WordSim-353,SimLex-999, dan Miller Charles?

Adapun batasan dari penelitian ini adalah korpus Wikipedia bahasa Indonesia sebagai dataset yang digunakan dalam pengukuran kesamaan semantik. Korpus Wikipedia ini berasal dari hasil *crawling* bapak Herry Sujaini, beliau adalah seorang dosen Informatika pada Universitas TanjungPura dan dataset yang digunakan sebagai perbandingan untuk melihat sistem yang dibangun sudah baik atau belum menggunakan dataset *gold standard* WordSim-353, SimLex-999, dan Miller Charles. Dataset *Gold standard* ini berasal dari penelitian sebelumnya menggunakan metode PMI [7].

1.3 Tujuan

Berdasarkan rumusan masalah yang telah disampaikan, berikut tujuan penelitian ini adalah

1. Menerapkan metode GloVe untuk mengimplementasikan kesamaan semantik antar sepasang kata dalam bahasa Indonesia.
2. Menganalisis nilai korelasi kesamaan semantik pasangan kata dalam bahasa Indonesia menggunakan metode GloVe dengan perbandingan dataset *gold standard* WordSim-353, SimLex-999, dan Miller Charles.

1.4 Organisasi Tulisan

Organisasi tulisan pada penelitian ini terdiri atas : Bagian 1 membahas mengenai latar belakang, rumusan masalah dan tujuan dari penelitian. Bagian 2 membahas studi terkait mengenai teori. Bagian 3 membahas sistem yang dibangun untuk metode GloVe. Bagian 4 menjelaskan hasil dan analisis sistem. Bagian 5 menjelaskan kesimpulan serta saran.

2. Studi Terkait

2.1 *Semantic Similarity*

Semantic similarity merupakan task pada *Natural Language Processing* (NLP) untuk mengukur kesamaan / keterkaitan antara pasangan kata secara semantik. *Semantic similarity* juga merupakan metode yang digunakan untuk menghitung kemiripan yang diliat dari makna [9]. Kesamaan semantik memiliki peran penting dalam beberapa *task* dari *Natural Language Processing* dan beberapa bidang terkait seperti *text classification, document clustering, text summarization,* dan lain sebagainya [4].

2.2 *Word Embeddings*

Word embeddings adalah representasi kata yang dinyatakan dengan setiap kata memiliki vektor yang mewakili makna dari kata tersebut. Dimensi yang digunakan beragam. Model *word embeddings* didesain berdasarkan hipotesis berdistribusi, kata dengan arti yang mirip cenderung memiliki *word embedding* yang sama [3]. *Word embeddings* juga dapat menangkap semantik dan sintaksis kata dari korpus besar yang tidak berlabel [6]. Nilai *similarity* yang dihasilkan *word embeddings* berkisar antara -1 sampai 1, dengan 1 sebagai nilai *similarity* tertinggi [2].

2.3 GloVe

GloVe merupakan representasi kata untuk menghasilkan *word embeddings* oleh Stanford Univerity. GloVe merupakan suatu motode *unsupervised learning* pada representasi kata yang mengungguli model lainnya di *word analogy, word similarity,* dan *named entity recognition* [8]. *Unsupervised learning* adalah suatu pendekatan yang tidak memiliki data latih sehingga datanya berasal dari data yang ada dengan cara mengkelompokan data tersebut menjadi beberapa bagian. GloVe menggunakan kumpulan teks yang pada tugas ini, kumpulan teks menggunakan korpus Wikipedia yang nantinya kumpulan teks akan dibangun *vocabulary* dan setiap kata pada *vocabulary* menghasilkan vektor yang berjumlah ratusan dimensi. Algoritma ini memasukkan peluang munculnya kata dalam suatu *window* (persekitaran kata) kedalam perhitungannya. Model Glove ditetapkan berdasarkan :

$$w_i^T + \vec{w}_k + b_i + \vec{b}_k = log(X_{i_k}) \tag{1}$$

dimana w adalah vektor kata, $\vec{w}$ adalah vektor kontek kata,b_i dan b_k adalah bias skalar untuk kata utama dan kontek kata. X adalah matriks kemunnculan dimana X_{i_k} mempresentasikan jumlah berapa kali kata $_k$ muncul di kontek kata $_i$. f(X_{i_k}) fungsi bobot. Perhitungan X_{i_k} didapatkan dengan cara mengumpulan statistik kemunculan kata dalam bentuk matrik kemunculan x. setiap elemen matriks X_{i_k} mewakili seberapa sering kata muncul dalam konteks kata j, dimana konteks kata merupakan kumpulan kata yang terdiri atas kata- kata yang berada di sebelum dan sesudah kata i sebanyak *windows size* yang diberikan.Pembobotan kata untuk setiap kata dalam konteks kata dengan cara $\frac{1}{distance}$, *distance* disini dihitung dengan cara panjang kontek kata - posisi kata tersebut. Untuk menghitung nilai f(X_{i_k}) dilakukan dengan menggunakan persamaan dibawah ini:

$$f(X_{i_k}) = \begin{cases} (\frac{X_{i_k}}{xmax})^{\alpha}; if X_{i_k} < xmax \\ 1; lainnya \end{cases} \tag{2}$$

Model GloVe diatas memperkenalkan fungsi pembobotan ke dalam fungsi *cost* yang memberikan model seperti di bawah ini:

$$J = \sum_{i,k=1}^{V} f(X_{i_k})(w_i^T \vec{w} j + b_i + b_k - logX_{i_k})^2 \tag{3}$$

Pada model GloVe terdapat parameter yang digunakan antara lain xmax, alpha, dan iterasi. Nilai paramater yang digunakan pada tugas akhir ini adalah untuk xmax sebesar 100 dan alpha sebesar 3/4 merunjuk pada paper GloVe: *Global Vectors for Word Representation* dengan menggunakan dataset wikipedia bahasa Inggris dengan nilai xmax 100 memberikan performansi yang baik walaupun dengan dimensi vektor yang kecil [8] dan penggunaan alpha 3/4 berdasarkan penelitian sebelumnya yang menunjukkan performansi yang baik pada [5]. Untuk parameter iterasi metode Glove dapat menggunakan nilai iterasi yang beragam dan semakin besar nilai iterasi akan menghasilkan nilai performansi yang lebih baik [8]. Pada tugas akhir ini menggunakan iterasi sebesar 50, iterasi diberikan berdasarkan parameter penelitian paper rujukan yang memberikan penggunaan 50 iterasi untuk dimensi vektor yang kurang dari 300 dan 100 iterasi untuk dimensi vektor lebih dari 300 [8]. Besaran parameter lainnya seperti *windows size* dan dimensi vektor digunakan berbagai besaran yang nilainya akan di analisis pada penelitian tugas akhir ini karena parameter dimensi vektor dan windows size adalah parameter yang kuat untuk mempengaruhi nilai kesamaan, ketika nilai kesamaan diukur menggunakan *cosine similarity* [2].

2.4 Word2Vec

Word2Vec adalah representasi vektor kata yang dibangun oleh Mikolov dari Google. Word2Vec juga merupakan metode yang menghasilkan *word embeddings*. Pada Word2Vec menerapkan model Skip-Gram dan model CBOW (*Continous Bag-of-Words*) [5]. Model Skip-Gram menggunakan proyeksi vektor kata-kata konteks untuk memprediksi vektor kata target, sedangkan model CBOW memprediksi vektor kata- kata yang ada dikonteks dengan diberikan vektor kata tertentu [5]. Sama halnya dengan metode GloVe, Word2Vec mendapatkan nilai *similarity* dengan menggunakan *cosine similarity*.

2.5 Korelasi Pearson

Korelasi *pearson* merupakan evaluasi hasil perhitungan keterkaitan semantik dilakukan dengan menghitung korelasi antara skor akhir dari sistem dan *gold standard*. Korelasi pearson digunakan untuk mengukur hubungan 2 variabel dengan menghasilkan hasil yang bersifat kuantitatif / skor. Korelasi *pearson* menghasilkan nilai korelasi antara range -1 sampai 1 [1] Adapun rumus korelasi *pearson* adalah :

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}} \tag{4}$$

dimana n adalah banyaknya pasangan data x dan y, $\sum x$ adalah total dari jumlah variabel x, $\sum y$ adalah total jumlah variabel y, $\sum x^2$ adalah kuadrat dari total jumlah variabel x, $\sum y^2$ adalah kuadrat dari total jumlah variabel y, dan $\sum xy$ adalah jumlah hasil perkalian variabel x dan variabel y.

2.6 Korpus Wikipedia Bahasa Indonesia

Korpus Wikipedia berbahasa Indonesia adalah data yang bersumber dari berbagai tulisan terbitan Wikipedia yang berbahasa Indonesia. Korpus Wikipedia yang digunakan pada tugas akhir ini berasal dari 1159 artikel yang beragam dengan jumlah kata sebesar 504.240 kata. Pemilihan korpus Wikipedia menjadi korpus penelitian tugas akhir ini karena korpus Wikipedia menyediakan berbagai kumpulan artikel berbahasa Indonesia dan pada paper [3] ruang lingkup korpus lebih penting dari pada ukuran korpus dimana menggunakan korpus domain signifikan meningkatkan kinerja untuk *task* yang diberikan. Contoh potongan korpus Wikipedia dapat dilihat pada lampiran 1.

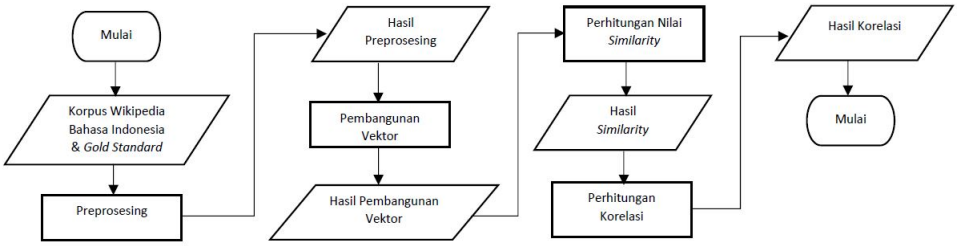
2.7 Gold Standard

Gold Standard merupakan suatu nilai / skor yang dihasilkan dari pendapat sekumpulan manusia yang dijadikan acuan dalam proses pengukuran similaritas maupun keterkaitan semantik antara pasangan kata dalam skala tertentu. *Gold Standard* ditujukan untuk mengetahui seberapa besar korelasi hasil skor yang dikeluarkan sistem terhadap relevansi kata yang diuji [10]. *Gold standard* yang digunakan untuk kesamaan semantik seperti Simlex999, WordSim353, RG65, YP130,Miller Charles dan AG203. Penelitian tugas akhir menggunakan dataset *gold standard* WordSim353, SimLex999, dan Miller Charles. Contoh potongan dataset *gold standard* yang digunakan pada lampiran 3-5.

3. Sistem yang Dibangun

3.1 Alur Sistem

Pada penelitian tugas akhir ini dibangun sistem untuk menghitung nilai kesamaan semantik antar pasangan kata. Gambaran umum alur sistem dapat dilihat pada gambar 1.



Gambar 1. Rancangan Sistem

Gambar 1 memperlihatkan alur sistem yang digunakan untuk menghitung nilai *semantic similarity* menggunakan metode GloVe. Dengan tahapan awal adalah masukan dataset korpus dan *gold standard*. Korpus yang digunakan adalah korpus Wikipedia bahasa Indonesia dan *gold standard* yang digunakan adalah Miller Charles, SimLex-999, dan WordSim-353. Data masukan korpus di preprosesing, hasil preprosesing dibangun vektor setiap kata yang ada di korpus, setelah itu dihitung *similarity*, hasil dari perhitungan *similarity* dan nilai *gold standard* dihitung nilai korelasi yang nantinya hasil tersebut didapatkan sebagai nilai evaluasi sistem.

3.2 Preprosesing

Preprosesing adalah tahap awal yang penting sebelum masuk ke tahap utama. Preprosesing dilakukan menjadikan data menjadi terstruktur sehingga mudah dipahami oleh sistem. Preprosesing dilakukan pada korpus Wikipedia bahasa Indonesia. Preprosesing yang dilakukan pada korpus adalah *case folding* dan *stopwords removal*. *Case folding* adalah suatu proses pemerataan data dengan cara mengubahnya ke dalam *lower case*. *Stopwords removal* adalah suatu proses menghilangkan kata-kata yang tidak penting biasanya kata tersebut bersifat umum dan jumlah kemunculannya besar. Contoh *stopwords removal* dapat dilihat pada Tabel 1. Alur preprosesing dapat dilihat pada lampiran 6.

Tabel 1. Contoh penggunaan *stopwords*

Input	Output
Pembatasan barang impor di Indonesia	pembatasan barang impor Indonesia

3.3 Pembangunan Vektor

Pembangunan vektor ini dilakukan menggunakan metode GloVe. Tahap pembangunan vektor sebagai berikut tahap awal mencari kata unik yang terdapat pada hasil korpus Wikipedia bahasa Indonesia yang telah di propresesing. Setelah itu, sistem akan mencari nilai matrik *co-occurrences* dimana matrik ini sebagai penganti korpus Wikipedia. Hasil dari matriks *co-occurrences* akan di set model GloVe dengan persamaan (1). Hasil set model tersebut akan di hitung nilai *cost function* dengan persamaan (3) yang nantinya akan menghasilkan 2 buah vektor. Kemudian 2 vektor yang terbentuk akan di normalisasikan untuk memperoleh vektor akhir. Alur pembangunan vektor dapat dilihat pada lampiran 7.

3.4 Perhitungan *Similarity*

Pada tahap perhitungan *similarity* masukan yang digunakan adalah pasangan kata yang terdapat pada *gold standard*. Pasangan kata tersebut akan dibangun vektor yang terbentuk kemudian dihitung nilai *cosine similarity*. Hasil akhir adalah nilai *similarity* yang dihasilkan oleh pasangan kata *gold standard*. ALur proses perhitungan *similarity* dapat dilihat pada lampiran 8.

3.5 Perhitungan Korelasi

Perhitungan korelasi dilakukan untuk mengetahui apakah sistem sudah baik atau belum. Pada tahap ini dilakukan pada dataset *gold standard*. Nilai yang dihasilkan oleh metode GloVe nantinya akan dibandingkan dengan metode Word2Vec untuk mengetahui apakah metode GloVe dapat diterapkan dan mengetahui nilai *similarity* yang terbentuk apakah sudah baik atau belum. Word2Vec digunakan sebagai metode pembanding dikarenakan GloVe dan Word2Vec memiliki kesamaan yaitu menggunakan *word embeddings* sebagai representasi kata. Alur perhitungan korelasi dari awal adalah masukan data hasil perhitungan *similarity* korpus dan nilai dataset *gold standard*, hitung nilai keduanya menggunakan *pearson correlation* dengan persamaan (4) untuk menghasilkan nilai korelasinya. Alur proses perhitungan korelasi dapat dilihat pada lampiran 9.

4. Evaluasi

4.1 Hasil Pengujian

Pada tugas akhir ini, akan dicari nilai kesamaan antara dua kata, dua kata tersebut berasal dari pasangan kata yang berada di dataset *gold standard*. Nilai kesamaan semantik dihasilkan dari hasil *cosine similarity* antar dua kata. Hasil kesamaan semantik inilah yang akan dicari nilai korelasinya dengan nilai yang terdapat pada *gold standard*. Hasil pengujian ini berdasarkan dari nilai korelasi terbaik dari metode GloVe dan nilai korelasi metode lain yang dihasilkan dengan parameter yang sama dengan metode GloVe. Parameter yang digunakan adalah 100 dimensi vektor, *windows size* 10.

Tabel 2. Hasil Dataset Wikipedia Bahasa Indonesia dengan Menggunakan 2 Metode

Metode	Miller Charles	SimLex999	WordSim353
GloVe	**0.1165**	**0.2280**	**0.2849**
Word2Vec	0.0444	0.0926	0.0925

Pasangan Kata	Gold Standard	Glove	Word2Vec
ayam,perjalanan	0.08	-0.325	-0.825
kaca,pesulap	0.11	0.066	0.842
senar,senyum	0.13	-0.593	-0.831

(a) Miller Charles

Pasangan Kata	Gold Standard	Glove	Word2Vec
baru,kuno	0.23	0.053	0.337
menyusut,tumbuh	0.23	0.205	0.851
kotor,sempit	0.3	-0.238	0.542

(b) SimLex-999

Pasangan Kata	Gold Standard	Glove	Word2Vec
biarawan,budak	0.54	0.485	0.462
persediaan,jaguar	0.62	0	0
persediaan,hidup	0.88	0.278	0.982

(c) WordSim-353

Gambar 2. Sampel Nilai Kesamaan Semantik Berdasarkan Metode

Pada Tabel 2 memperlihatkan nilai korelasi dari metode GloVe dan metode Word2Vec. Pada metode GloVe nilai korelasi yang dihasilkan pada tiga *gold standard* menghasilkan nilai korelasi yang lebih baik daripada nilai korelasi Word2Vec. Pada Gambar 2 memperlihatkan sampel perbedaan nilai *similarity* yang terbentuk oleh 2 metode beserta nilai *similarity gold standard*. Nilai *similarity* yang terbentuk oleh metode GloVe mendekati nilai *similarity* pada *gold standard* daripada nilai *similarity* Word2Vec. Untuk nilai *similarity* yang dihasilkan oleh Word2Vec menghasilkan nilai *similarity* yang lebih tinggi dari nilai *gold standard* dapat dilihat pada gambar 2a,2b,2c. Walaupun ada juga nilai *similarity* yang mendekati *gold standard* tetapi nilai *similarity* GloVe lebih mendekati *gold standard*. Nilai *similarity* yang dihasilkan oleh metode juga dapat menghasilkan nilai 0 yang dapat dilihat pada pasangan kata (persediaan,jaguar) pada Gambar2c karena pasangan kata tersebut tidak menghasilkan nilai kesamaan semantik. Perbedaan nilai *similarity* pada metode dengan nilai *similarity gold standard* menyebabkan nilai korelasi yang terbentuk. Berdasarkan hasil korelasi yang terbentuk metode GloVe dapat diterapkan pada korpus Wikipedia bahasa Indonesia dengan nilai korelasi 0.1165 untuk *gold standard* Miller Charles, nilai korelasi 0.2280 untuk *gold standard* SimLex-999, dan nilai korelasi 0.2849 untuk *gold standard* WordSim-353.

4.2 Analisis Nilai Korelasi Kesamaan Semantik Berdasarkan *Windows Size*

Pada proses ini dilakukan implementasi sistem untuk mengetahui nilai korelasi terhadap ketiga dataset *gold standard* yang sama-sama terdiri dari dua pasang kata berdasarkan *windows size* yang digunakan, Dimensi vektor 100 digunakan pada implementasi 3 *windows size* dengan iterasi 50.

Tabel 3 memperlihatkan nilai korelasi metode GloVe terhadap tiga *gold standard* berdasarkan *windows size* yang digunakan. Nilai korelasi yang terbentuk berasal dari nilai *similarity* GloVe dengan nilai *similarity gold standard*. Nilai *similarity* yang terbentuk oleh metode GloVe menghasilkan adanya nilai *similarity* yang negatif yang menunjukkan bahwa nilai *similarity* yang terbentuk berlawanan arah dengan nilai *similarity gold standard* dapat dilihat pada gambar 3a,3b. Pada gambar 3 menunjukan hasil nilai *similarity* semakin besar ketika *windows size* yang digunakan semakin besar. Semakin besar nilai *windows size* maka semakin banyak kontek kata yang terbentuk sehingga memperbanyak kemunculan pasangan kata yang terbentuk. Semakin nilai *similarity* yang dihasilkan

Tabel 3. Hasil Nilai Korelasi GloVe Terhadap *Gold Standard* Berdasarkan Besar*Windows Size*

Windows Size	Miller Charles	SimLex999	WordSim353
2	0.0013	0.1980	0.1262
6	0.0031	0.2204	0.2703
10	**0.1165**	**0.2280**	**0.2849**

meningkat maka semakin meningkat juga nilai korelasinya jika nilai korelasi tersebut sejajar dengan peningkatan atau penurunan nilai *similarity* pada *gold standard*. Berdasarkan hasil korelasi metode GloVe yang terbesar pada percobaan di atas adalah menggunakan *windows size* 10 dengan nilai korelasi 0.1165 untuk *gold standard* Miller Charles, nilai korelasi 0.2280 untk *gold standard* SimLex-999, dan nilai korelasi0.2849 untuk *gold standard* WordSim-353.

Pasangan Kata	Gold Standard	WS2	WS6	WS10
ayam,perjalanan	0.08	-0.207	-0.211	-0.325
kaca,pesulap	0.11	0.052	0.052	0.066
senar,senyum	0.13	-0.076	-0.291	-0.593

(a) Miller Charles

Pasangan Kata	Gold Standard	WS2	WS6	WS10
baru,kuno	0.23	-0.268	0.020	0.053
menyusut,tumbuh	0.23	-0.428	0.083	0.205
kotor,sempit	0.3	-0.080	-0.173	-0.238

(b) SimLex-999

Pasangan Kata	Gold Standard	WS2	WS6	W10
biarawan,budak	0.54	0.215	0.306	0.485
persediaan,jaguar	0.62	0	0	0
persediaan,hidup	0.88	0.153	0.265	0.278

(c) WordSim-353

Gambar 3. Sampel Nilai Kesamaan Semantik Berdasarkan *Windows Size*

4.3 Analisis Nilai Korelasi Kesamaan Semantik Berdasarkan Dimensi Vektor

Pada proses ini dilakukan implementasi sistem untuk mengetahui nilai korelasi terhadap ketiga dataset *gold standard* yang sama-sama terdiri dari dua pasang kata berdasarkan dimensi vektor yang digunakan. Pada analisis ini bertujuan untuk mengetahui nilai korelasi yang terbaik dari hasil kesamaan semantik berdasarkan dimensi vektor yang digunakan. Parameter lain yang digunakan pada analisis ini adalah dengan menggunakan *windows size* 10 dan iterasi 50.

Tabel 4. Hasil korelasi dari nilai kesamaan semantik *gold standard* dengan nilai kesamaan semantik metode GloVe

Dimensi Vektor	Miller Charles	SimLex999	WordSim353
50	0.0673	0.2256	0.2471
100	**0.1165**	**0.2280**	**0.2849**
300	0.0899	0.2275	0.2848

Pasangan Kata	Gold Standard	DV50	DV100	DV300
ayam,perjalanan	0.08	-0.303	-0.325	-0.271
kaca,pesulap	0.11	0.050	0.066	0.062
senar,senyum	0.13	-0.555	-0.593	-0.570

(a) Miller Charles

Pasangan Kata	Gold Standard	DV50	DV100	DV300
baru,kuno	0.23	0.044	0.053	0.049
menyusut,tumbuh	0.23	0.203	0.205	0.200
kotor,sempit	0.3	-0.142	-0.238	-0.143

(b) SimLex-999

Pasangan Kata	Gold Standard	DV50	DV100	DV300
biarawan,budak	0.54	0.457	0.485	0.377
persediaan,jaguar	0.62	0	0	0
persediaan,hidup	0.88	0.168	0.278	0.240

(c) WordSim-353

Gambar 4. Sampel Nilai Kesamaan Semantik Berdasarkan Dimensi Vektor

Tabel 4 memperlihatkan nilai korelasi yang dihasilkan dari nilai kesamaan semantik terhadap ketiga dataset Miller Charles, SimLex-999 dan WordSim-353 berdasarkan dimensi vektor yang digunakan. Hasil korelasi berasal dari nilai *similarity* yang dihasilkan. Pada Gambar 4 memperlihatkan tabel nilai *similarity* yang dihasilkan berdasarkan dimensi vektor. Nilai dimensi vektor yang berbeda menyebabkan nilai *similarity* yang dihasilkan berbeda juga. Pada penggunaan dimensi vektor 100 mengalami peningkatan nilai *similarity* terhadap nilai *similarity* dengan menggunakan 50 dimensi vektor. Tetapi mengalami penurunan pada nilai *similarity* yang dihasilkan pada 10 dimensi vektor, sampel nilai kesamaan semantik dapat dilihat pada gambar 4a,4b,4c. Dan nilai *similarity* yang dihasilkan pada percobaan 10 dimensi vektor, nilainya mendekati nilai *similarity gold standard*. Hasil nilai *similarity* inilah yang menyebabkan nilai korelasi yang terbentuk pada Tabel 4. Berdasarkan hasil percobaan diatas maka dapat disimpulkan bahwa nilai dimensi yang cocok dan menghasilkan nilai korelasi tertinggi adalah penggunaan 100 dimensi vektor. Nilai korelasi ini disebabkan karena nilai *similarity* pada 100 dimensi vektor menghasilkan nilai *similarity* yang mendekati nilai *similarity* tiga *gold standard* yang digunakan.

5. Kesimpulan dan Saran

5.1 Kesimpulan

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

1. Sistem yang dibangun dapat mengimplementasikan perhitungan kesamaan semantik antar kata dengan metode GloVe pada pasangan kata yang berasal dari *gold standard* WordSim353, Miller Charles dan SimLex-999.

2. Implementasi penelitian yang menghasilkan nilai korelasi terbaik adalah dengan menggunakan parameter 100 dimensi vektor, *windows size*, dan iterasi sebesar 50 memperoleh nilai korelasi pada Miller Charles sebesar 0.1165, nilai korelasi pada SimLex-999 sebesar 0.2280, dan nilai korelasi WordSim-353 sebesar 0.2849.
3. Parameter yang mempegaruhi nilai korelasi kesamaan semantik adalah dimensi vektor dan *windows size*. Pada dimensi vektor yang sama, jika semakin tinggi nilai *windows size* maka semakin tinggi nilai kesamaan semantik yang dihasilkan.

5.2 Saran

Adapun saran untuk penelitian tugas akhir ini kedepannya adalah sebagai berikut :

1. Gunakan korpus lain untuk menguji metode GloVe untuk mencari nilai kesamaan semantik disarankan kata-kata yang di *gold standard* terdapat di korpus yang digunakan
2. Gunakan dataset *gold standard* lain sebagai dataset yang menghasilkan nilai kesamaan semantik berdasarkan pasangan kata yang ada pada *gold standard*.

Daftar Pustaka

[1] N. Chok. *Pearson's versus spearman's and kendall's correlation coefficient for continuous data: University of Pittsburgh*. PhD thesis, Master Thesis, 2010.

[2] Á. Elekes, M. Schäler, and K. Böhm. On the various semantics of similarity in word embedding models. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries*, pages 139–148. IEEE Press, 2017.

[3] S. Lai, K. Liu, S. He, and J. Zhao. How to generate a good word embedding. *IEEE Intelligent Systems*, 31(6):5–14, 2016.

[4] R. Mihalcea, C. Corley, C. Strapparava, et al. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, volume 6, pages 775–780, 2006.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[6] T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, 2013.

[7] P. G. B. S. Parta. Implementasi dan analisis keterkaitan semantik berbahasa indonesia dengan pendekatan pointwise mutual information. *Telkom University*, 2017.

[8] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

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

[10] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281, 2007.

[11] T. Zesch and I. Gurevych. The more the better? assessing the influence of wikipedia's growth on semantic relatedness measures. In *LREC*, 2010.