Implementasi dan Analisis Kesamaan Semantik pada Bahasa Indonesia dengan Metode berbasis Vektor

Implementation and Analysis of Semantic Similarity on Bahasa Indonesia by Vector-based Method

Rhesa Fauzan Hermawan[1], Ade Romadhony, S.T., M.T.[2], Said Al Faraby, S.T., M.Sc.[3]

[1,2,3,] Prodi S1 Teknik Informatika, Fakultas Informatika, Universitas Telkom Bandung

[1]rhesafauzanhermawan@gmail.com, [2]aderomadhony@telkomuniversity.ac.id,
[3]saidalfaraby@telkomuniversity.ac.id.

Abstrak

Kesamaan semantik adalah tugas untuk memperkirakan kekuatan hubungan semantik antara unit bahasa atau konsep, dalam hal ini kesamaan makna yang dimiliki oleh sepasang kata. Kesamaan semantik pada kata bahasa Indonesia dapat diukur dengan menggunakan basis pengetahuan seperti Kamus Besar Bahasa Indonesia. Metode berbasis vektor merupakan salah satu metode yang dapat digunakan untuk mengukur kesamaan semantik. Pada Tugas Akhir ini diimplementasikan kesamaan semantik pada pasangan kata bahasa Indonesia dengan menggunakan metode berbasis vektor, pembobotan tf-idf, dan perhitungan kesamaan kosinus, Kamus Besar Bahasa Indonesia sebagai basis pengetahuan, dan dataset yang digunakan dibuat berdasarkan referensi dari SimLex999 dan Rubenstein-goodenough sebanyak 180 pasang kata, gold standard yang didapat berdasarkan hasil kuesioner terhadap 31 orang responden. Hasil penelitian yang telah dilakukan, didapatkan nilai korelasi terbaik sebesar 0.5416 dengan menambahkan definisi sinonim dalam pengujian. Parameter terbaik yang mempengaruhi nilai kesamaan semantik pada penelitian ini adalah dengan menambahkan definisi dari sinonim tanpa stopword removal.

Kata Kunci: gold standard, Kamus Besar Bahasa Indonesia, kesamaan semantik, kesamaan kosinus, metode berbasis vektor, tf-idf.

Abstract

Semantic similarity is a task to estimate the strength of the semantic relationship between language units or concepts, in this case the common meaning possessed by the pair of words. The semantic similarity of the Indonesian words can be measured by using a knowledge base such as Big Indonesian Dictionary. The vector-based method is one of the methods for calculating semantic similarities. In this final project, I implement semantic similarity of Indonesian word pairsusing vector-based method, tf-idf weighting, and cosine similarity calculation.Using Big Indonesian Dictionary as knowledge base and golden standard based on SimLex999 and Rubenstein-goodenough consist of 180 pairs of words, we conduct the experiment.To evaluate the semantic similarity scores produced by the system, we ask 31 respondents to give a similarity score on the golden standard. We get the best correlation value of 0.5416 by adding definition of synonym in test. The best parameter that influences semantic equality value in this research is by adding the synonym without stopword removal.

Keywords: Big Indonesian Dictionary, cosine similarity, golden standard, semantic similarity, tf-idf, vector-based method.

1. Pendahuluan

Setiap orang pasti memiliki penilaian yang berbeda mengenai seberapa besar kesamaan makna dari sepasang kata bahasa Indonesia. Pencarian informasi mengenai kesamaan makna atau semantik pada kata bahasa Indonesia merupakan hal yang baru. Keterbatasan sumber data menjadi tantangan tersendiri dalam penelitian ini, namun kendala tersebut masih dapat dilalui dengan memanfaatkan kamus bahasa Indonesia yang ada.

Kesamaan semantik adalah proses yang digunakan untuk memperkirakan kekuatan hubungan semantik antara unit bahasa, konsep atau contoh seperti mengetahui kemiripan antara ayah dan bapak, melalui deskripsi numerik yang diperoleh sesuai dengan perbandingan informasi pendukung makna atau menggambarkan sifat [1]. Kesamaan antara dua kata atau kalimat berupa angka menggambarkan kedekatan makna antara kedua kata atau kalimat tersebut. Perhitungan kesamaan digunakan dalam berbagai keperluan, misalnya untuk melakukan pencarian informasi di Internet, klasifkasi dokumen dalam arsip, dan kegiatan menganalisis informasi di dunia maya (*data analysis*) [12]. Salah satu metode yang digunakan untuk mengukur nilai kesamaan semantik adalah dengan menggunakan metode berbasis vektor.

Metode berbasis vektor merupakan salah satu metode yang dapat digunakan untuk mengukur kesamaan semantik dari sepasang dokumen. Berdasarkan definisi kata yang didapat dari Kamus Besar Bahasa Indonesia (KBBI), dibentuk vektor yang merepresentasikan kata tersebut. Dengan menggunakan fungsi kosinus yang terdapat pada metode berbasis vektor, dapat diukur seberapa besar kesamaan semantik dari sepasang kata berdasarkan sudut vektornya.

Fungsi kesamaan kosinus adalah kesamaan antara dua teks yang berasal dari nilai kosinus antara dua vektor text kata [2]. Fungsi ini merupakan fungsi yang umum digunakan pada metode berbasis vektor. Untuk mencerminkan betapa pentingnya sebuah kata dalam sebuah kumpulan dokumen atau korpus [13]. maka digunakanlah pembobotan term frequency-inverse document frequency (tf-idf) sebagai faktor bobot.

Pada studi kasus kali ini akan dilakukan implementasi dalam bentuk aplikasi untuk mengukur kesamaan dari sepasang kata bahasa Indonesia, dengan metode berbasis vektor, dihitung menggunakan fungsi kosinus, melalui pembobotan tf-idf, menggunakan basis pengetahuan KBBI.

2. Dasar Teori

2.1 Semantik

Semantik adalah cabang dari linguistik (ilmu bahasa) yang mempelajari arti atau makna yang terkandung pada suatu bahasa, kode, atau jenis representasi lain, atau yang lebih dikenal dengan pembelajaran tentang makna. Kata semantik itu sendiri menunjukkan berbagai ide dari populer yang sangat teknis. Hal ini sering digunakan dalam bahasa sehari-hari untuk menandakan suatu masalah pemahaman yang datang ke pemilihan kata atau konotasi [17]. Semantik berbeda dengan sintaks, studi tentang kombinatorik unit bahasa (tanpa mengacu pada maknanya), dan pragmatik, studi tentang hubungan antara simbol-simbol bahasa, makna, dan pengguna bahasa [18].

2.2 Kesamaan Semantik

Kesamaan semantik atau semantic similarity digunakan untuk mengidentifikasi konsep-konsep yang memiliki kesamaan "karakteristik". Meskipun manusia tidak tahu definisi formal keterkaitan antara konsep, kesamaan semantik bisa menilai keterkaitan antara mereka [3]. Kesamaan semantik adalah metrik yang didefinisikan di atas seperangkat dokumen atau kata, dimana gagasan jarak antara keduanya didasarkan pada kemiripan makna atau konten semantiknya dibandingkan dengan kesamaan yang dapat diperkirakan mengenai representasi sintaksis mereka. Kesamaan semantik juga merupakan alat matematika yang digunakan untuk memperkirakan kekuatan hubungan semantik antara unit bahasa, konsep atau contoh, melalui deskripsi numerik yang diperoleh sesuai dengan perbandingan informasi yang mendukung maknanya atau menggambarkan sifatnya [1].

2.3 Metode Berbasis Vektor

Metode berbasis vektor adalah salah satu metode yang sering digunakan dalam penelitian semantik, salah satunya yaitu untuk mengukur kesamaan semantik pada dokumen.Vektor sendiri merupakan sebuah objek geometri yang memiliki besar/panjang dan arah/sudut. Dalam hal ini, metode berbasis vektor memanfaatkan karakteristik vektor dalam melakukan perhitungan kesamaan semantik dimana dua buah vektor dikatakan sama apabila keduanya memiliki panjang dan arah yang sama. Metode ini memiliki fitur yang dapat digunakan untuk merepresentasikan data [5]. Dalam perbandingan dua buah vektor, metode berbasis vektor menggunakan komponen vektor sebagai pembanding. Hal ini bertujuan untuk mengetahui seberapa besar nilai kesamaan dari dua buah vektor yang berbeda.

2.4 Metode Berbasis Gloss

Fungsi kesamaan kosinus adalah salah satu fungsi yang dapat digunakan pada metode berbasis vektor. Kesamaan antara dua dokumen dapat diturunkan dengan menghitung nilai kosinus antara dua vektor kata pada dokumen [10]. Sebuah dokumen dapat direpresentasikan sebagai vektor kata yang mana dimensi vektor tersebut mengacu pada kata yang terdapat pada dokumen [2]. Fungsi kesamaan kosinus dapat dilihat pada Persamaan 2-1 berikut [8]:

$$sim_{cos}(di, dj) = \frac{\overrightarrow{di}.\overrightarrow{dj}}{\left|\left|\overrightarrow{di}\right|\right|\left|\left|\overrightarrow{dj}\right|\right|} \quad (2\text{-}1)$$

Dapat dilihat pada Persamaan 2-1. Dimana $\overrightarrow{di}$ dan $\overrightarrow{dj}$ adalah sebuah vektor, lalu $\overrightarrow{di}.\overrightarrow{dj}$ adalah komponen vektor yang sama yang dimiliki vektor 1 dan vektor 2, kemudian ||d|| adalah panjang dari vektor tersebut dimana semua nilai komponen pada vektor tersebut dikuadratkan, kemudian dijumlah, dan diakarkan

2.5 TF-IDF

TF-IDF adalah statistik numerik yang dimaksudkan untuk mencerminkan betapa pentingnya sebuah kata dalam dokumen di dalam sebuah kumpulan kata atau korpus [13]. TF-IDF sering digunakan sebagai pembobotan dalam pencarian informasi, penambangan teks, dan pemodelan pengguna [19]. Saat ini, TF-IDF adalah salah satu skema pembobotan bobot yang paling populer. Misalnya, 83% sistem rekomendasi berbasis teks dalam domain perpustakaan digital menggunakan TF-IDF [15].

2.5.1 *Term Frequency* (TF)

Term Frequency (TF) adalah frekuensi dari kemunculan sebuah term dalam dokumen yang bersangkutan. Semakin besar jumlah kemunculan suatu term (TF tinggi) dalam dokumen, semakin besar pula bobotnya atau akan memberikan nilai kesesuaian yang semakin besar [16]. Berikut Persamaan 2-2 dari TF:

$$TF = 1 + log_{10}(f_{t,d}),\ f_{t,d} > 0 \mid TF = 0,\ f_{t,d} = 0 \quad (2\text{-}2)$$

Contoh: satu tambah satu sama dengan dua.
Jumlah kata "satu" = 2, maka nilai TF = 1 + $\log_{10}(2)$ = 1,3.

2.5.2 *Inverse Document Frequency* (IDF)

Inverse Document Frequency (IDF) merupakan sebuah perhitungan dari bagaimana term didistribusikan secara luas pada koleksi dokumen yang bersangkutan. IDF menunjukkan hubungan ketersediaan sebuah term dalam seluruh dokumen. Semakin sedikit jumlah dokumen yang mengandung term yang dimaksud, maka nilai IDF semakin besar [16]. Berikut Persamaan 2-3 dari IDF:

$$idf_f = log_{10}(\frac{N}{df_f}) \quad (2\text{-}3)$$

Dimana:
N = Jumlah dokumen yang ada.
df = Jumlah dokumen dimana terdapat kemunculan kata.

Contoh: kata "makan" muncul di 12 dokumen, dan jumlah dokumen yg ada sebanyak 33, maka nilai IDF dari kata "makan" adalah:
IDF = $log_{10}(33/12)$ = 0.4393

2.6 Kamus Besar Bahasa Indonesia

Kamus Besar Bahasa Indonesia adalah kamus ekabahasa resmi bahasa Indonesia yang disusun oleh Badan Pengembangan dan Pembinaan Bahasa dan diterbitkan oleh Balai Pustaka. Kamus ini menjadi acuan tertinggi bahasa Indonesia yang baku, karena kamus ini merupakan kamus bahasa Indonesia terlengkap dan yang paling akurat yang pernah diterbitkan oleh penerbit yang memiliki hak paten dari pemerintah Republik Indonesia yang dinaungi oleh Kementerian Pendidikan dan Kebudayaan Indonesia. [24].

2.7 Kateglo

Kateglo adalah aplikasi dan layanan web sumber dan isi terbuka untuk kamus, tesaurus, dan glosarium bahasa Indonesia. Namanya diambil dari akronim unsur layanannya: ka(mus), te(saurus), dan glo(sarium). Untuk kamus, kateglo mengambil data berdasarkan dari KBBI daring tiga dan disediakan API untuk mengambil data tersebut.

2.8 Korelasi Pearson

Korelasi pearson merupakan salah satu teknik statistik yang digunakan untuk mengukur kekuatan hubungan dari 2 buah variabel dan juga dapat mengetahui bentuk dari hubungan 2 buah variabel tersebut dengan hasil yang bersifat kuantitatif. Nilai korelasi berada dalam range -1≤ X ≤ 1 [14].

Berikut adalah rumus Persamaan 2-4 dari korelasi pearson:

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{\{n\Sigma x^2 - (\Sigma x)^2\}\ \{n\Sigma y2 - (\Sigma y)2\}}} \quad (2\text{-}4)$$

Dimana:
n = banyaknya pasangan data X dan Y.
Σx = total jumlah variabel X.
Σy = total jumlah variabel Y.
Σx^2 = kuadrat dari total jumlah variabel X.
Σy^2 = kuadrat dari total jumlah variabel Y.

Σxy = perkalian dari total jumlah variabel X dan Y.

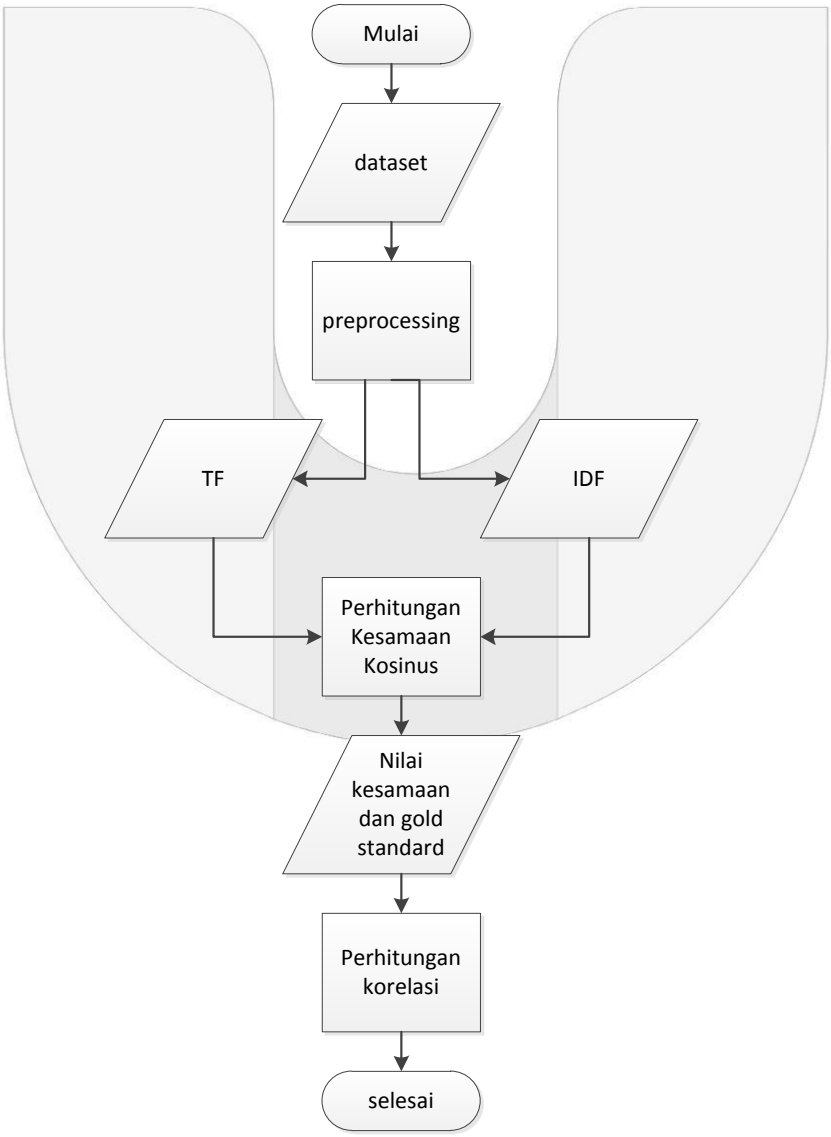
Berikut kriteria hubungan dari korelasi dapat dilihat pada Tabel 1 [14].

Tabel 1 Kriteria hubungan korelasi

r	Kriteria Hubungan
0	Tidak ada korelasi
0-0.5	Korelasi lemah
0.5-0.8	Korelasi sedang
0.8-1	Korelasi kuat/erat
1	Korelasi sempurna

2.9 Gambaran Umum Sistem

Sistem yang akan dibangun bertujuan untuk mengukur nilai kesamaan semantik dari pasangan kata bahasa Indonesia. Sistem ini memanfaatkan algoritma pembentukan representasi vektor kata yang berasal dari program yang sudah ada sebelumnya[1]. Nilai kesamaan semantik dihasilkan dengan menggunakan metode basis vektor melalui perhitungan kesamaan kosinus dan penggunaan basis pengetahuan KBBI untuk pengambilan definisi dan sinonim kata. Kemudian nilai kesamaan semantik akan dibandingkan dengan *gold standard*. Evaluasi dilakukan dengan menghitung nilai korelasi antara nilai kesamaan semantik dengan *gold standard*. Untuk lebih jelasnya berikut alur sistem secara umum yang dibuat dapat dilihat pada flowchart di bawah.



Gambar 1 Flowchart perancangan sistem

[1]Pangestu, Chandra. 2016. "Analisis dan Implementasi Keterkaitan Semantik dengan Metode berbasis Vektor". Jurnal Eproc. Universitas Telkom.

Dari Gambar 1 Flowchart perancangan sistem, gambaran umum sistem sebagai berikut:

1. Sistem memanggil dataset berupa file txt yang berisi kata pertama, kata kedua, dan nilai *gold standard* sebanyak 180 pasangan kata, yang kemudian semua kata tersebut digunakan untuk memanggil definisi dari kata tersebut dan sinonimnya. Kemudian melalui proses *preprocessing* untuk dijadikan nilai *inverse document frequency* (idf). Nilai idf kemudian disimpan kedalam file txt untuk digunakan pada proses selanjutnya.
2. Kemudian dataset dipanggil kembali, kali ini setiap sepasang kata dengan nilai *gold standard* untuk memanggil definisi dan sinonim. Kemudian di*preprocessing* untuk mendapatkaan nilai *term frequency* (tf) dari masing-masing komponen pada pasangan kata.
3. Setelah itu, nilai tf dari masing-masing komponen pada kata dikalikan dengan nilai idf dari komponen tersebut yang telah disimpan pada proses sebelumnya. Kemudian nilai tf-idf yang dihasilkan digunakan untuk perhitungan kosinus untuk masing-masing pasangan kata.
4. Setelah semua perhitungan pasangan kata selesai. Kemudian dilakukan perhitungan korelasi menggunakan rumus korelasi pearson dengan *gold standard* dan nilai kesamaan sebagai parameternya.

2.10 Pengumpulan dan Pembuatan Data

Pada tahapan ini dilakukan pengumpulan data dari basis pengetahuan dan pembuatan data pasangan kata bahasa Indonesia dan *gold standard*.

2.10.1 Data Kamus Besar Bahasa Indonesia

Basis pengetahuan yang digunakan pada penelitian ini adalah KBBI. Data yang diambil berupa definisi dari kata dan sinonimnya yang diambil melalui API kateglo.com.

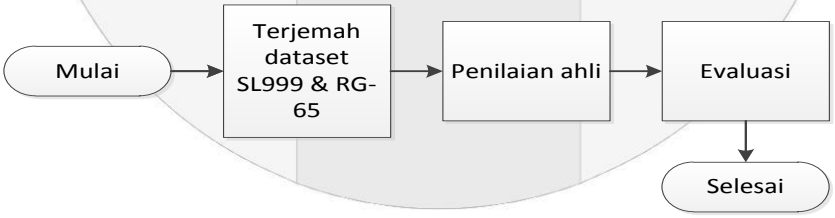
Jenis KBBI yang tersedia pada kateglo adalah versi daring tiga (tahun 2006). Untuk sekarang, KBBI sudah memasuki versi daring lima (tahun 2016) dimana jumlah lema yang dimiliki sebanyak 127.036 lema, lebih banyak daripada daring tiga yang hanya memiliki 78.000 lema [24]. Hal ini mengakibatkan ada beberapa kata pada dataset yang harus dievaluasi.

Kemudian definisi yang diambil pada KBBI merupakan definisi pertama/utama dari kata. Hal ini dikarenakan tidak semua kata memiliki definisi yang bermakna sama. Contohnya sebagai berikut :

Kata “Makan”
Definisi: (1) memasukkan makanan pokok ke dalam mulut serta mengunyah dan menelannya
(7) menyerang, mematikan, mengambil (dalam permainan catur).

2.10.2 Data Pasangan Kata Bahasa Indonesia

Data pasangan kata bahasa Indonesia dikumpulkan secara manual oleh penulis dengan jumlah 180 pasang kata berdasarkan data yang diperoleh dari dataset yang dibuat berdasarkan kemiripan kata yaitu*Simlex999*[2] dan dataset *Rubenstein-goodenough*[3] yang diterjemahkan dari bahasa Inggris ke dalam bahasa Indonesia, yang kemudian digolongkan tingkat kemiripan oleh domain ahli dan dievalusi kembali. Berikut Gambar 2 alur pembuatan data pasangan kata bahasa Indonesia pada penelitian ini:



Gambar 2 Alur pembuatan dataset.

Tahapan dalam pembuatan dataset seperti Gambar 2 adalah sebagai berikut:

1. Dataset *Simlex999* dan *Rubenstein-goodenough-65* di terjemahkan dari bahasa Inggris ke dalam bahasa Indonesia.
2. Dataset *Simlex999* yang berjumlah 999 pasangan kata diambil sebanyak 666 pasang kata yang berbentuk kata benda, sedangkan dataset *Rubenstein-goodenough-65* dipakai seluruhnya sebanyak 65 pasang kata.
3. Untuk dataset *Rubenstein-goodenough-65* dikonsultasikan kepada domain ahli yaitu Dosen bahasa Indonesia Telkom University untuk diberi label tingkat kesamaan pada pasangan kata.
4. Selanjutnya adalah dengan mengevaluasi dataset tersebut.

[2] https://www.cl.cam.ac.uk/~fh295/simlex.html
[3] Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. Communications of the ACM, 8(10):627-633

5. Pertama, menghilangkan pasangan kata jika salah satu kata tidak terdapat pada KBBI daring tiga. Terdapat 6 pasangan kata yang tidak ada seperti "mobil|penyihir", "kaca|penyihir", "anak|penyihir", "bijak|penyihir", "pesulap|penyihir" dan "ayam jantan|ayam jago" pada dataset *Rubenstein-goodenough-65* sehingga jumlah dataset menjadi 59 pasang kata.
6. Kedua, mengambil sample dari dataset *Rubenstein-goodenough-65* dari kesamaan tinggi, sedang dan rendah berdasarkan hasil dari penilaian yang telah dilakukan oleh Dosen bahasa Indonesia masing-masing sebanyak 6 pasang kata dari jumlah total dataset sebanyak 59 pasang kata yang ada.
7. Ketiga, mengambil sampel dari dataset *Simlex999* berdasarkan nilai *gold standard* yang sudah ada pada dataset tersebut, dari yang tinggi, sedang, dan rendah sebanyak 170 pasangan kata dari total sebanyak 666 pasangan kata.
8. Keempat, menormalisasi pasangan kata yang terdapat pada kedua dataset tersebut. Hal ini dikarenakan ada pasangan kata yang sama setelah di terjemahkan dari bahasa Inggris ke bahasa Indonesia seperti pasangan kata "infeksi|penyakit", "anak|laki-laki", "kabut|awan", dan "batang|kayu". Kemudian pasangan kata yang sama yang terdapat pada dataset *Simlex999* dan *Rubenstein-goodenough-65*, seperti pasangan kata "pantai|pesisir". Jumlah pasangan kata yang dinormalisasi ada sebanyak 16 pasang kata menjadi 8 pasang kata.

Setelah melalui tahapan diatas, didapat 180 pasang kata bahasa Indonesia dari 1064 pasang kata hasil terjemahan bahasa Inggris ke bahasa Indonesia yang digunakan sebagai dataset pada penelitian Tugas Akhir ini..

2.10.3 Gold Standard

Data *gold standard* dibuat berdasarkan intuisi manusia sebagai domain ahli. Data tersebut dapat diperoleh dengan melakukan *crowdsourcing* melalui kuesioner dengan kriteria sebagai berikut:

- Jumlah minimal responden sebanyak 30 orang [7].
- Latar pendidikan minimal lulusan SMA atau sederajat.

Berdasarkan hasil kuesioner yang diperoleh, didapat sebanyak 31 responden dengan latar pendidikan lulusan SMA atau sederajat hingga Sarjana yang kemudian diambil hasil penilaian responden untuk diolah menjadi nilai *gold standard*. Berikut rumus untuk menghitung nilai *gold standard* berdasarkan pada hasil kuesioner:

$$Gold\ Standard = \frac{rata2\ nilai\ gold\ standard}{5} \text{(3-1)}$$

Nilai *gold standard* dinormalisasi pada rentang 0-1 untuk menyesuaikan dengan hasil dari perhitungan kesamaan kosinus.

2.10.3.1 Data Kuesioner

Data kuesioner merupakan data pasangan kata bahasa Indonesia berdasarkan indeks tematik yang diberikan kepada responden untuk diberikan nilai *gold standard* pada masing-masing pasangan kata. Pada kuesioner ini, responden diminta untuk mengisikan identitas diri dan kolom penilaian kesamaan semantik.

Nilai *gold standard* merupakan nilai kesamaan semantik pasangan kata bahasa Indonesia yang diisi oleh responden mengikuti aturan kriteria yang ada. Responden diminta untuk menilai pasangan kata yang ada seusai dengan kriteria dan nilai yang telah ditetapkan berdasarkan penilaian masing-masing.

Tabel 2 Skala nilai kuesioner

Nilai	Kesamaan makna kata	Contoh
0	Tidak mirip	mobil – ikan
1	Hanya memiliki keterkaitan/kesamaan rendah	selang – kebun
2	Memilik makna bagian	kepala – wajah
3	Memiliki makna turunan	mahluk hidup - hewan
4	Memiliki makna sinonim	mempelai - pengantin
5	Memiliki makna yang sama	saya – aku

Tabel 2 merupakan skala penilaian pasangan kata bahasa Indonesia berdasarkan jenis makna yang terkandung pada pasangan kata tersebut [25]. Untuk skala dibuat berdasarkan referensi yang ada yang kemudian disesuaikan dengan pengujian pada pasangan kata [26].

2.10.3.2 Evaluasi Kuesioner

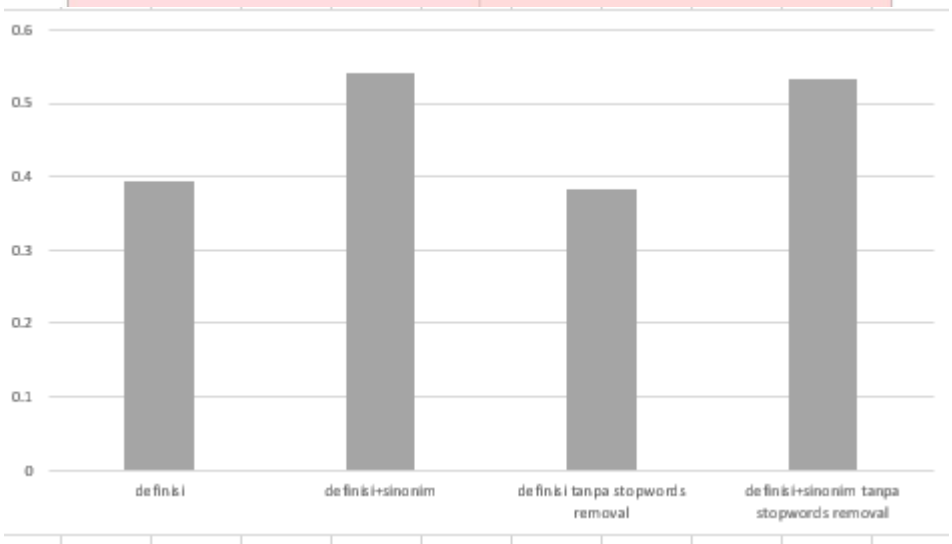
Data *gold standard* yang diperoleh berdasarkan hasil kuesioner yang telah dilakukan kemudian dievaluasi sebagai berikut:

- Mengambil nilai dari responden dengan latar belakang pendidikan lulusan SMA atau sederajat.
- Mengolah data nilai *gold standard* untuk mendapatkan range nilai dan untuk perhitungan korelasi.

3. Pembahasan

3.1 Analisis Nilai Kesamaan Pasang Kata Bahasa Indonesia Berdasarkan pada Nilai Korelasi.

Dari empat percobaan yang telah dilakukan, didapat nilai korelasi yang hasilnya ditampilkan dengan grafik pada Gambar 3 di bawah:



Gambar 3 Grafik hasil korelasi

Kemudian kedua hasil perhitungan korelasi disimpan dalam Tabel 3 dengan tampilan sebagai berikut:

Tabel 3 Hasil perhitungan korelasi metode berbasis vektor

Unit Linguistik	Nilai Korelasi
Definisi kata	0.3936
Definisi kata + Sinonim	0.5416
Definisi kata tanpa *stopword removal*	0.3819
Definisi kata + Sinonim tanpa *stopword removal*	0.5344

Dapat dilihat pada Tabel 3 hasil perhitungan korelasi, dari percobaan yang telah dilakukan untuk mendapatkan nilai korelasi terbaik dimana hasil korelasi yang paling tinggi dari semua percobaan yang ada merupakan hasil yang paling baik. Dimana dengan menggunakan definisi dari kata dan sinonim didapatkan nilai korelasi yang paling tinggi yaitu sebesar 0.5416 yang merupakan korelasi dengan hubungan kekuatan sedang.

Semakin banyak *overlapping* komponen vektor pada pasangan kata akan semakin besar nilai kesamaan semantik yang dihasilkan, sebaliknya semakin sedikit *overlapping* komponen vektor pada pasangan kata akan semakin kecil nilai kesamaan semantik yang dihasilkan.

Penambahan definisi dari sinonim dalam percobaan dapat mempengaruhi nilai kesamaan semantik dari pasangan kata. Hal ini dikarenakan setiap kata memiliki jumlah dan sinonim yang berbeda sehingga berpengaruh pada perhitungan kesamaan kosinus untuk menghitung kesamaan semantik pada penelitian ini.

3.2 Analisis Parameter Terbaik Yang Mempengaruhi Nilai Kesamaan Semantik

Untuk menganalisis parameter terbaik yang mempengaruhi nilai kesamaan semantik, dimana parameter terbaik merupakan nilai terkecil dari selisih nilai kesamaan semantik dengan *gold standard*. Oleh karena itu, akan digunakan nilai rata-rata selisih setiap nilai *gold standard* dengan nilai dari sistem.

Tabel 4-3 Hasil perhitungan selisih nilai rata-rata gold standard dengan nilai sistem

Unit Linguistik	Nilai Rata-Rata Selisih Nilai *Gold Standard* dengan Nilai dari Sistem
Definisi Kata	0.2977
Definisi + Sinonim	0.2444

Definisi kata tanpa *stopword removal*	0.3042
Definisi+sinonim tanpa *stopword removal*	0.2429

Dari Tabel 4-3 hasil perhitungan nilai rata-rata selisih nilai *gold standard* dengan nilai dari sistem. Maka parameter terbaik yang mempengaruhi nilai kesamaan semantik adalah dengan menggunakan definisi dari kata dan sinonim tanpa melakukan *stopword removal* dimana didapat nilai selisih paling kecil dari semua percobaan yang ada yaitu sebesar 0.2429.

4. Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan dapat disimpulkan bahwa:

- Metode berbasis vektor dapat di implementasikan untuk menilai kesamaan semantik pasangan kata bahasa Indonesia dengan menggunakan rumus kesamaan kosinus dan tf-idf sebagai pembobot.
- Nilai korelasi pearson dengan nilai terbaik yang didapatkan adalah dengan menambahkan pengaruh definisi sinonim dari kedua kata yang dibandingkan yaitu dengan nilai 0.5416 yang memiliki nilai korelasi sedang.
- Parameter terbaik yang mempengaruhi nilai kesamaan semantik adalah dengan menambahkan definisi sinonim dan tanpa melakukan stopword removal dengan nilai selisih dengan gold standard yang paling kecil yaitu sebesar 0.2429.

Daftar Pustaka

[1] Harispe S., Ranwez S. Janaqi S., Montmain J., 2015. "Semantic Similarity from Natural Language and Ontology Analysis". *Synthesis Lectures on Human Language Technologies* 8:1: 1–254

[2] Rahutomo F., Kitasuka T., Aritsugi M., 2012. *"Semantic Cosine Similarity".*

[3] Slimani T. "Description and Evaluation of Semantic similarity Measures Approaches". Taif University.

[4] Hikaruyuuki. "Kamus Kata Dasar dan Stopword List Bahasa Indonesia", (online), (http://static.hikaruyuuki.com/wp-content/uploads/stopword_list_tala.txt. diakses 8 Febuari 2017).

[5] Pangestu, Chandra. 2016. "Analisis dan Implementasi Keterkaitan Semantik dengan Metode berbasis Vektor". *Jurnal Eproc*. Universitas Telkom.

[6] Miller G.A., Beckwith R., Fellbaum C., Gross D. and Miller K.."WordNet: An on-line lexical database," *International Journal of Lexicography*, vol. 3, 1990, pp. 235−244.

[7] Anonim. "Menentukan Ukuran Sampel Sederhana", (online). (http://teorionline.net/menentukan-ukuran-sampel-menurut-para-ahli. diakses 11 Agustus 2017).

[8] Madylova A., Oguducu S. G., 2009. "A Taxonomy based Semantic Similarity of Documents using the Cosine Measure". In *Computer and Information Sciences*, 2009. ISCIS 2009. 24th International Symposium. pp. 129 – 134.

[9] Mitchell J., Lapata M., 2008. "Vector-based Models of Semantic Composition". Proceedings of ACL-08: HLT, pages 236–244

[10] G. Salton and C. Buckley, "Term-weighting Approaches in Automatic Text Retrieval," *Information Processing and Management*, vol.24 no.5, 1988, pp.513−523.

[11] Alexander, Alvin."indonesianstemmer.java", (Online), (http://alvinalexander.com/java/jwarehouse/lucene/contrib/analyzers/common/src/java/org/apache/lucene/analysis/id/IndonesianStemmer.java.shtml diakses 10 Januari 2017)

[12] A. Islam and D. Inkpen, "Semantic text similarity using corpus-based word similarity and string similarity," ACM Trans. Knowl. Discov. from Data, vol. 2, no. 2, p. 10, 2008.

[13] Rajaraman, A.; Ullman, J. D. (2011). "Data Mining". *Mining of Massive Datasets*. pp. 1–17.

[14] Elektronika, Teknik. "Pengertian Analisis Korelasi Sederhana Rumus Pearson", (Online), (http://teknikelektronika.com/pengertian-analisis-korelasi-sederhana-rumus-pearson diakses 20 April 2017).

[15] Breitinger, Corinna; Gipp, Bela; Langer, Stefan (2015-07-26). "Research-paper recommender systems: a literature survey". *International Journal on Digital Libraries*. 17 (4): 305–338.

[16] Aditnya. 2016. "Pembobotan kata atau term weighting TF-IDF", (Online), (https://informatikalogi.com/term-weighting-tf-idf/ diakses 20 April 2017).

[17] Wikipedia. "Semantik", (Online), (https://id.wikipedia.org/wiki/Semantik . Diakses 21 April 2017).

[18] Kitcher, Philip; Salmon, Wesley C. (1989). *Scientific Explanation. Minneapolis*, MN: University of Minnesota Press. p. 35.

[19] Wikipedia. 2012. "Tf-idf",(Online), (https://en.wikipedia.org/wiki/Tf-idf. Diakses 21 April 2017).

[20] A. Ballatore; M. Bertolotto; D.C. Wilson (2014). "An evaluative baseline for geo-semantic relatedness and similarity". *GeoInformatica*. 18:4: 747–767.

[21] Wikipedia. "Sinonim", (online), (https://id.wikipedia.org/wiki/Sinonim. diakses 21 April 2017).

[22] Wikipedia. "Korelasi", (online), (https://id.wikipedia.org/wiki/Korelasi. diakses 21 April 2017).

[23] Wikipedia. "Application programming interface", (online), (https://en.wikipedia.org/wiki/Application_programming_interface. diakses 21 April 2017).

[24] Wikipedia. "Kamus Besar Bahasa Indonesia", (online), (https://id.wikipedia.org/wiki/Kamus_Besar_Bahasa_Indonesia. diakses 21 April 2017).
[25] T. Husni; S. Atiqa. 2015. "Efektivitas Algoritma Semantik dengan Keterkaitan Kata dalam Mengukur Kemiripan Teks Bahasa Indonesia", *Khazanah informatika*. Vol. 1 No. 1 ISSN:2477-698X
[26] Maulana, Wahyu. 2016. "Pengukuran Kesamaan Semantik pada Potongan Ayat Alquran dengan Pendekatan Word Alignment". *Skripsi*. Telkom University.