

**IMPLEMENTASI WORD EMBEDDING UNTUK
MENGANALISA KESAMAAN SEMANTIK ANTAR KATA
DENGAN METODE WORD2VEC**

Proposal Skripsi

Proposal ini disusun untuk memenuhi persyaratan
skripsi

Oleh :

Raka Rasell
32160036



Program Studi Teknik Informatika
Fakultas Teknologi Dan Desain
Universitas Bunda Mulia

Jakarta

2019

IDENTITAS PENYUSUN

Nama : Raka Rasell

NIM : 32160036

Program Studi : Teknik Informatika

Kelas : 7 MTI-1 (Malam)

Angkatan : 2016

Semester : 7 (Tujuh)

No. Telepon : +62 856 8692 835

E-mail : rrrraka@gmail.com

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Kesamaan semantik yaitu merupakan suatu pengukuran untuk mencari nilai yang menyatakan tingkat kesamaan atau kedekatan secara semantik, kalimat, atau teks. Sepasang kata dinyatakan semantik apabila memiliki kesamaan dari sisi makna atau konsep. Pasangan kata dikatakan memiliki kesamaan semantik jika kata tersebut memiliki makna atau konsep yang sama. Penelitian ini didasari dimana komputer belum dapat menyamakan persepsi manusia terkait penilaian dari makna pasangan kata yang memiliki kesamaan semantik. [1]

Pada tugas akhir ini dilakukan perhitungan nilai kesamaan semantik antara dua buah kata dengan berbasis vektor. Metode yang digunakan adalah *Word2vec*, karena metode ini dapat memproses kedekatan vektor kata-kata dan dinilai memiliki nilai performa yang baik.[2] *Word2vec* yaitu metode yang dapat menghitung nilai vektor hubungan antara sepasang kata atau lebih. Input dari *Word2vec* yaitu berupa korpus, sedangkan outputnya berupa vektor kata yang selanjutnya dapat menghasilkan nilai kesamaan semantik yang dihasilkan dari perhitungan *cosine similarity*.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, maka rumusan masalah yang akan dibahas adalah:

1. Bagaimana mengimplementasi metode *Word2vec* untuk menganalisa semantik kata?

2. Bagaimana tingkat akurasi semantik kata dengan menggunakan metode Word2vec?

1.3 Tujuan dan Manfaat

Tujuan dan manfaat adalah untuk menganalisis faktor faktor yang mempengaruhi semantik kata dan mengetahui korelasi antar kata.

1.3.1 Tujuan

Tujuan yang hendak dicapai pada penelitian saya adalah sebagai berikut:

1. Dapat mengimplementasi metode Word2vec untuk semantik kata.
2. Dapat menganalisa korelasi antar kata dengan menggunakan metode Word2vec.

1.3.2 Manfaat

Manfaat yang hendak dicapai dalam penelitian ini adalah sebagai berikut:

1. Mengetahui korelasi dan makna antar kata.

1.4 Ruang Lingkup

Untuk memastikan agar penelitian yang dilakukan tidak menyimpang dari pokok pembahasan, maka penulis telah menetapkan batasan-batasan dalam penelitian ini, yaitu:

1. Data yang akan digunakan merupakan data teks berbahasa Indonesia.
2. Data berasal dari Wikipedia

1.5 Metodologi Perancangan Sistem

1. Analisis Pengumpulan Data

Metode yang digunakan untuk mengumpulkan data (yang selanjutnya disebut corpus) ialah dengan men-*download* semua artikel berbahasa Indonesia yang akan digunakan untuk *training* data. Corpus tersebut didapat dari *dump* Wikipedia. Untuk *download* sebuah Corpus terdapat dua cara:

1. <https://dumps.wikimedia.org/idwiki/latest/> dengan nama idwiki-latest-pages-articles.xml.bz2
2. <https://dumps.wikimedia.org/idwiki/<timestamp>> dengan nama idwiki-<timestamp>-pages-articles.xml.bz2

2. Metode Perancangan Aplikasi

Perancangan aplikasi yang dibuat dengan metode Word2vec.

3. Pemodelan Sistem

Pemodelan sistem pada implementasi ini menggunakan python karena sintaks yang sedikit dan python mendukung untuk proses word embedding.

1.6 Sistematika penulisan

1. BAB 1 PENDAHULUAN

Bab 1 berisi Latar Belakang Masalah, Rumusan Masalah, Tujuan dan Manfaat, Ruang Lingkup, Metodologi Penelitian, dan Sistematika Penulisan

2. BAB 2 LANDASAN TEORI

Bab 2 berisi teori yang akan menjadi dasar untuk digunakan dalam menyelesaikan permasalahan yang sudah dijelaskan pada bab sebelumnya

3. BAB 3 ANALISIS DAN PERANCANGAN

Bab 3 berisi tentang analisis kebutuhan yang diperlukan dalam menyelesaikan permasalahan yang sudah dijelaskan pada bab sebelumnya

4. BAB 4 IMPLEMENTASI

Bab 4 berisi penerapan teori ke dalam bahasa pemrograman dan pengujian terhadap pengklasifikasian yang akan dilakukan

5. BAB 5 SIMPULAN DAN SARAN

Bab 5 berisi tentang simpulan dari penelitian ini dan saran untuk memperbaiki penelitian

2. BAB 2

LANDASAN TEORI

2.1 *Machine Learning*

Machine learning digunakan untuk mengajarkan mesin bagaimana menangani data dengan lebih efisien. Terkadang setelah melihat data, kami tidak dapat menafsirkan pola atau mengekstrak informasi dari data. Dalam hal ini, kami menerapkan pembelajaran mesin. Dengan banyaknya kumpulan data yang tersedia, permintaan akan pembelajaran mesin meningkat. Banyak industri dari kedokteran hingga militer menerapkan pembelajaran mesin untuk mengekstraksi informasi yang relevan [4].

Pada *machine learning*, umumnya terdapat 2 macam *learning* untuk mengklasifikasi suatu objek, yaitu:

a. Supervised Learning

Supervised Learning adalah metode yang membutuhkan bantuan eksternal. *Dataset input* dibagi menjadi *dataset* train dan test. *Dataset* kereta memiliki variabel keluaran yang perlu diprediksi atau diklasifikasikan. Semua metode mempelajari semacam pola dari *dataset* pelatihan dan menerapkannya pada *dataset* uji untuk prediksi atau klasifikasi [4].

Terdapat 3 metode *supervised learning* yang paling sering digunakan, diantaranya:

- i. Decision Tree adalah jenis pohon yang atribut kelompok dengan menyortir mereka berdasarkan nilai-nilai mereka.

Decision Tree digunakan terutama untuk tujuan klasifikasi.

Setiap pohon terdiri dari simpul dan cabang. Setiap node mewakili atribut dalam grup yang akan diklasifikasikan dan setiap cabang mewakili nilai yang dapat diambil node .

ii. Naïve Bayes, digunakan untuk tujuan pengelompokan dan klasifikasi. Arsitektur yang mendasari Naïve Bayes tergantung pada probabilitas kondisional. Ini menciptakan pohon berdasarkan kemungkinan mereka terjadi. Pohon pohon ini juga dikenal sebagai Bayesian Network.

iii. Support Vector Machine (SVM), digunakan untuk klasifikasi. SVM bekerja berdasarkan prinsip perhitungan margin. Pada dasarnya, menarik *margin* antara kelas. *Margin* ditarik sedemikian rupa sehingga jarak antara *margin* dan kelas adalah maksimum dan karenanya, meminimalkan kesalahan klasifikasi.

b. Unsupervised Learning

Unsupervised Learning mempelajari beberapa fitur dari data. Ketika data baru diperkenalkan, ia menggunakan fitur yang dipelajari sebelumnya untuk mengenali kelas data. Ini terutama digunakan untuk pengelompokan dan pengurangan fitur.

Terdapat 2 metode *unsupervised learning* yang paling sering digunakan, diantaranya:

i. K-Means Clustering, atau *grouping* adalah jenis teknik *unsupervised learning* yang ketika dimulai, membuat kelompok secara otomatis. Barang-barang yang memiliki

karakteristik serupa diletakkan di *cluster* yang sama. Metode ini disebut *k-means* karena ia menciptakan *kcluster* yang berbeda. *Mean* dari nilai-nilai dalam *cluster* tertentu adalah pusat dari *cluster* itu.

- ii. Principal Component Analysis atau PCA, dimensi data dikurangi untuk membuat perhitungan lebih cepat dan lebih mudah. Contoh pada data 2D, Ketika data sedang plot dalam grafik, itu akan memakan dua sumbu. PCA

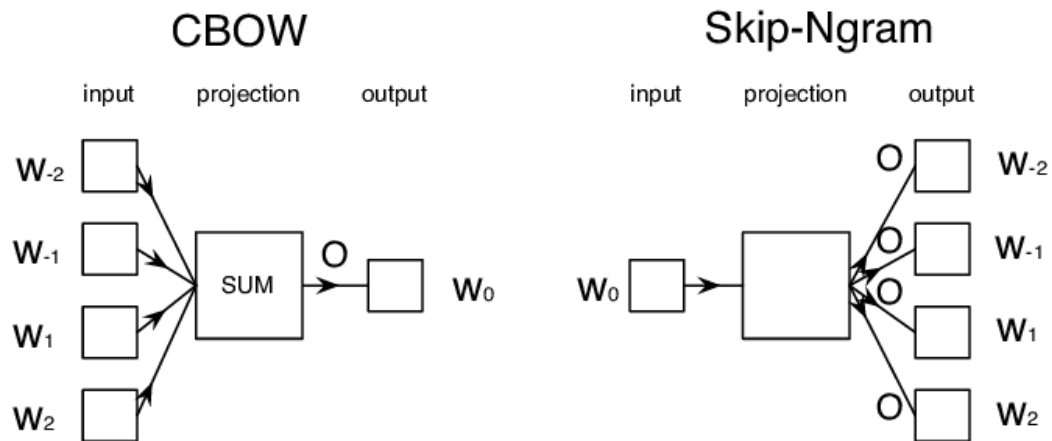
diterapkan pada data, data kemudian akan menjadi 1D.

2.2 *Word Embedding*

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

2.3 **Word2vec**

Word2vec merupakan suatu alat yang baru dikembangkan oleh Thomas Mikolov. Word2vec dapat mengolah kata-kata dari dataset yang sangat besar dalam waktu yang relatif singkat dengan nilai akurasi yang lebih baik dibandingkan dengan alat yang pernah ada sebelumnya. Cara kerja alat ini yaitu dengan mengambil korpus teks sebagai input, lalu menghasilkan representasi vektor setiap kata yang ada pada korpus teks tersebut sebagai output. File vektor yang dihasilkan dapat digunakan untuk penelitian pada pemrosesan bahasa alami dan aplikasi pembelajaran mesin. Vektor kata tersebut juga dapat digunakan untuk mengukur jarak kedekatan antar vektor kata yang lain. Word2vec memiliki dua arsitektur pemodelan yang dapat digunakan untuk merepresentasikan vektor kata, arsitektur tersebut yaitu *continous bag-of-word(CBOW)* dan *Skip-gram*.



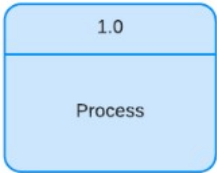
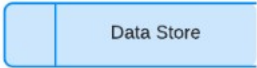


2.4 Data Flow Diagram (DFD)

Data Flow Diagram (DFD) awalnya dikembangkan oleh Chris Gane dan Trish Sarson pada tahun 1979 yang termasuk dalam *Structured System Analysis and Design Methodology* (SSAMD). Sistem yang dikembangkan berbasis pada dekomposisi fungsional dari sebuah sistem. Edward Yourdon dan Tom DeMarco memperkenalkan metode yang lain pada tahun 1980an di mana mengubah persegi dengan sudut lengkung (pada DFD Chris Gane dan Trish Sarson) dengan lingkaran untuk menotasikan.

DFD Edward Yourdon dan Tom DeMarco populer digunakan sebagai model analisis sistem perangkat lunak DFD adalah suatu model logika data atau proses yang dibuat untuk menggambarkan dari mana asal data dan kemana tujuan data yang keluar dari sistem dimana data disimpan proses apa yang menghasilkan data tersebut. Berdasarkan pengertian tersebut dapat disimpulkan bahwa *data flow diagram* (DFD) merupakan suatu proses data yang bergambarkan dari mana asal data dan kemana tujuan data tersebut keluar dari sistem dan disimpan.

Pada penelitian kali ini penulis menggunakan *symbol* DFD milik *Gane/Sarson*.

Tabel 2.1 Data Flow Diagram [9]



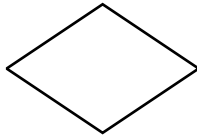
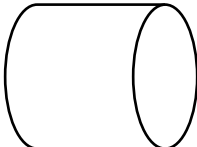
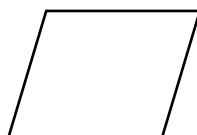

No	Simbol	Nama	Keterangan
1		<i>Process</i>	Memproses <i>Input</i> menjadi <i>Output</i>
2		<i>Data Store</i>	Tempat penyimpanan data
3		<i>External Entity</i>	Pihak penerima (tujuan) atau sumber data
4		<i>Data Flow</i>	Aliran data dari <i>Entity</i> ke <i>Entity</i>

2.5 Flowchart Diagram

Flowchart adalah representasi secara simbolik dari suatu metode atau prosedur untuk menyelesaikan suatu masalah, dengan menggunakan *flowchart* akan memudahkan pengguna melakukan pengecekan bagianbagian yang terlupakan dalam analisis masalah, disamping itu

flowchart juga berguna sebagai fasilitas untuk berkomunikasi antara pemrogram yang bekerja dalam tim suatu proyek [10].

Tabel 2.2 Flow Chart Diagram [11]

No	Simbol	Nama	Keterangan
1		<i>Terminator</i>	Penanda awal dan akhir dari <i>flowchart</i> .
2		<i>Process</i>	Menandakan aktivitas pemrosesan dari <i>input</i> menjadi <i>output</i> .
3		<i>Decision</i>	Menandakan adanya percabangan / pilihan berdasarkan suatu kondisi.
4		<i>Direct Access Storage</i>	Menandakan akses ke media penyimpanan untuk menyimpan atau membaca <i>file</i> .
5		<i>Input / Output</i>	Menandakan adanya interaksi dari <i>user</i> ke dalam proses dan dari proses ke <i>user</i> .
6		<i>Flow Line</i>	Menandakan arah aliran aktifitas.

3 DAFTAR PUSTAKA

- [1] M. . Nabila Nanda Widyastuti, Arif Bijaksana, Ir., M.Tech., Ph.D, Indra Lukmana Sardi, S.T., “Analisis Word2vec untuk Perhitungan Kesamaan Semantik antar Kata,” *e-Proceeding Eng.*, vol. Vol.5, No., no. 3, pp. 7603–7612, 2018.
- [2] K. R. Prilianti and K. Kunci, “Aplikasi Text Mining untuk Automasi Penentuan Tren Topik Skripsi dengan Metode K-Means Clustering,” vol. 2, no. 1, pp. 1–6, 2014.
- [3] D. S. Indraloka and B. Santosa, “Penerapan Text Mining untuk Melakukan Clustering Data Tweet Shopee Indonesia,” *J. Sains dan Seni ITS*, vol. 6, no. 2, pp. 6–11, 2017, doi: 10.12962/j23373520.v6i2.24419.
- [4] R. S. Putra, “Analisis sentimen twitter dengan klasifikasi naïve bayes menggunakan seleksi fitur mutual information dan inverse document frequency riky sutriadi putra,” 2017, doi: 10.1093/bjsw/bcm026.
- [5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *1st Int. Conf. Learn. Represent. ICLR 2013 - Work. Track Proc.*, pp. 1–12, 2013.