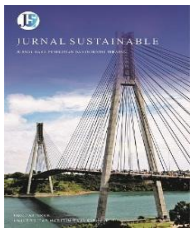


Jurnal Sustainable: Jurnal Hasil Penelitian dan Industri Terapan
Vol. 12, No. 2, hal. 37-41, Oktober 2023
Jurnal Sustainable: Jurnal Hasil Penelitian dan Industri Terapan

ISSN 2615-6334 (Online)
ISSN 2087-5347 (Print)

Implementasi Teknik Word Embedding Untuk Rekomendasi Hasil Pencarian Katalog Online Menggunakan Algoritma WORD2VEC

Raja Azian[1], Nola Ritha[2], Muhamad Radzi Rathomi[3]
[1,2,3]Jurusan Teknik Informatika, Fakultas Teknik dan Teknologi Kemaritiman
Universitas Maritim Raja Ali Haji
[1,2,3]Jl. Politeknik Senggarang, Tanjungpinang 29100, Indonesia
Email Author : 170155201010@student.umrah.ac.id[1], nola.ritha@ umrah.ac.id[2], radzi@gmail.com[3]
*Corresponding Author: nola.ritha@ umrah.ac.id

Abstract— *The purpose of this study is to apply the word2vec algorithm to recommend search results in online catalogs. The reason for taking this title is because, based on the results of observations and the observations of researchers, the data search process, especially in online catalogs, only reaches the syntactic level. So that the results are given only up to the syntactic level. Therefore, researchers utilize the word2vec algorithm, which has the ability to represent words at the semantic level, to carry out a search process where the results of this process are used as alternative search results or recommendations for search results. The data that the researchers used was data on 12,701 book titles in the Raja Ali Haji Maritime University library. To evaluate the recommendations for the search results obtained, the researcher tested the recommendation system for search results using several scenarios, and then the results were measured using a precision test on the k document (P@k). From the results of the precision test measurements on the k document, various results were found. Scenarios 1 and 2 show a fairly high precision value with a value of 0.53 and 0.59, while for testing in scenarios 3, 4, and 5, the resulting precision value is relatively low with a value of 0.50, 0.42, and 0.14.*
Keywords— word2vec, skip-gram, semantic-analysis, information-retrieval

Intisari— Tujuan dari penelitian ini adalah untuk mengaplikasi algoritma word2vec dalam merekomendasikan hasil pencarian pada katalog online. Alasan mengambil judul tersebut karena berdasarkan hasil observasi dan pengamatan peneliti proses pencarian data khususnya pada katalog online hanya sampai pada level sintaksis saja. Sehingga hasil yang diberikan hanya sampai pada level sintaksis saja. Oleh sebab itu, peneliti memanfaatkan algoritma word2vec yang mempunyai kemampuan untuk merepresentasikan kata pada level semantis untuk melakukan proses pencarian dimana hasil dari proses ini dijadikan hasil pencarian alternatif atau rekomendasi hasil pencarian. Data yang peneliti gunakan adalah data judul buku sebanyak 12.701 judul buku yang ada pada perpustakan Universitas Maritim Raja Ali Haji. Untuk mengevaluasi rekomendasi hasil pencarian yang didapat, peneliti menguji system rekomendasi hasil pencarian dengan menggunakan beberapa scenario kemudian hasilnya diukur dengan menggunakan uji presisi pada k document (P@k). Dari hasil pengukuran uji presisi pada k document didapati hasil yang bervariatif. Skenario 1 dan 2 menunjukan nilai presisi yang cukup tinggi dengan nilai sebesar 0.53 dan 0.59, sedangkan untuk pengujian pada scenario 3, 4, dan 5 nilai presisi yang di hasilkan terbilang rendah dengan nilai 0.50, 0.42, 0.14.
Kata kunci— word2vec, skip-gram, semantic-analysis, information-retrieval

I. PENDAHULUAN

Online Public Access Catalog (OPAC) bisa diakses dan digunakan kapan dan dimana saja, hal ini memudahkan pengguna untuk mencari buku-buku yang terdapat pada koleksi suatu perpustakaan [2]. Namun hasil pencarian yang dihasilkan oleh katalog online tidak selalu memberikan hasil yang optimal. Hal ini disebabkan oleh beberapa hal mendasar, diantaranya adalah, pada proses pencarian pengguna tidak selalu menggunakan kata kunci yang sama dengan kata kunci pada dokumen yang tersimpan [10], atau kata kunci yang pengguna gunakan sangat kompleks [9].

Metode-metode melakukan pencarian berbasis kata seperti *Exact String Matching*, *Approximate String Matching*, ataupun *Hybrid String Matching* banyak diusulkan oleh peneliti untuk mendapatkan hasil pencarian yang optimal [5]. Akan tetapi, metode-metode memberikan hasil yang optimal hanya pada tingkat sintaksis saja tetapi tidak pada tingkat semantik. Hal ini menjadi alasan kenapa hasil pencarian yang diberikan tidak optimal ketika kata kunci pencarian yang digunakan kompleks atau ketika kata kunci yang digunakan tidak terdapat pada dokumen yang disimpan.

Word embedding adalah teknik *Natural Language Processing* (NLP) untuk merepresentasikan kata-kata ke dalam bentuk vektor. Vektor hasil *word embedding* juga dikenal dengan *semantic vector space* karena ia bisa merepresentasikan kata baik dari sisi semantik (makna) maupun dari sisi sintaksis (struktur) [1]. Vektor hasil *word embedding* membuka kemungkinan untuk bisa melakukan operasi-operasi NLP pada kata-kata tersebut. Salah satunya adalah analisis semantik [12].

Analisis semantik pada NLP adalah proses untuk mengevaluasi dan mewakili bahasa natural dengan interpretasi yang mirip dengan manusia [11]. Salah satu algoritma word embedding yang terbukti efisien merepresentasikan vektor dari kata baik dari sisi sintaksis maupun semantik adalah *word2vec* [7].

Berdasarkan uraian sebelumnya, untuk mendapatkan hasil pencarian yang optimal, maka proses pencarian sebaiknya tidak berhenti pada tingkat sintaksis tapi dilanjutkan pada tingkat semantik. Sehingga hasil pencarian akan dihasilkan lebih optimal dan diharapkan memudahkan pengguna dalam menemukan informasi yang diinginkan.

Pada jurnal ini membahas mengenai pengaplikasian teknik *word embedding* untuk menghasilkan rekomendasi hasil pencarian dari pecarian katalog *online* (OPAC) dengan menggunakan algoritma *word2vec*.

II. LANDASAN TEORI

A. *Jaringan Syaraf Tiruan*

Jaringan syaraf tiruan adalah model matematis dari proses interpretasi dan menyimpan informasi yang terjadi pada otak manusia [6]. Jaringan syaraf tiruan mampu menyelesaikan permasalahan yang dikira sulit atau tidak mungkin diselesaikan dengan pendekatan matematis tradisional ataupun statistic [8]. Hal ini mungkin karena jaringan syaraf tiruan bisa mengenali pola-pola kompleks pada data beserta hubungannya sebelum menghasilkan satu atau lebih keputusan [4].

B. *Word2vec*

Word2vec adalah salah satu metode pada *word embedding* yang menggunakan pendekatan prediksi berbasis jaringan syaraf tiruan yang kecil [7]. *Word2vec* bisa diimplementasikan dengan salah satu dari dua pendekatan. Pendekatan pertama adalah pendekatan *Continuous Bag Of Words* (CBOW) dan pendekatan kedua adalah pendekatan *Skip Gram* (SG).

Pendekatan dengan model CBOW adalah pendekatan yang memprediksi kata target berdasarkan kata konteks yang diberikan. Model ini memaksimalkan probabilitas kemungkinan sebuah kata berada dalam konteks tertentu.

Sedangkan pendekatan dengan model SG adalah pendekatan yang memprediksi kata konteks berdasarkan kata target yang diberikan. Model ini berfokus memaksimalkan probabilitas kemungkinan kata konteks yang didapatkan berdasarkan kata target yang diberikan.

C. *Cosine Similarity*

Cosine Similarity adalah formula matematis untuk menghitung kesamaan dokumen terlepas dari ukurannya [11]. Kesamaan dihitung berdasarkan sudut kosinus yang terbentuk diantara dua vector dalam ruang multi dimensi.

ISSN 2615-6334 (Online)
ISSN 2087-5347 (Print)

$$similarity = \frac{A \cdot B}{\|A\|\|B\|} \quad (1)$$

Di mana:
$A \ dan \ B$ = komponen vector A dan B

D. Evaluasi

Precision at k documents (P@k) adalah metode evaluasi keefektifan yang bertujuan untuk kepuasan pengguna dengan menyajikan daftar data hingga *k document* yang sudah diurutkan berdasarkan peringkat. P@k dapat memberikan informasi tentang seberapa baik sistem mampu mengidentifikasi dokumen yang relevan terhadap permintaan pencarian yang diberikan [3].

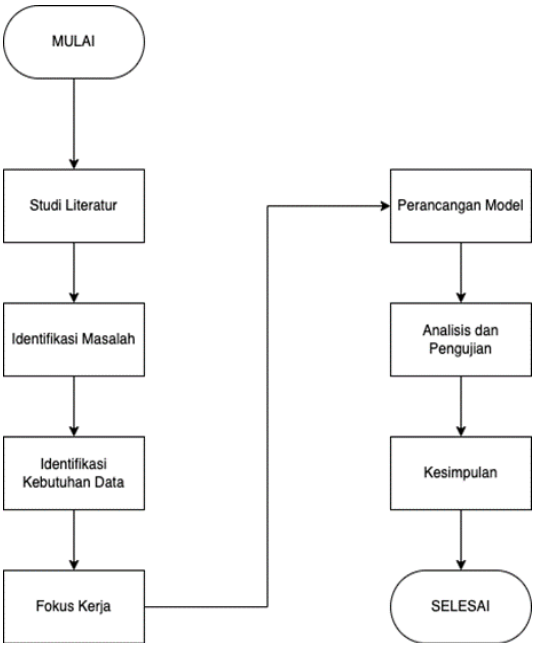
$$P@k = \frac{|\ Res[1..k]\ \cap Rel\ |}{k} \quad (2)$$

dimana:
Rel = Total prediksi relevant
Res [1..*k*]= Total prediksi yang diberikan system sampai dengan *k*

III. METODE PENELITIAN

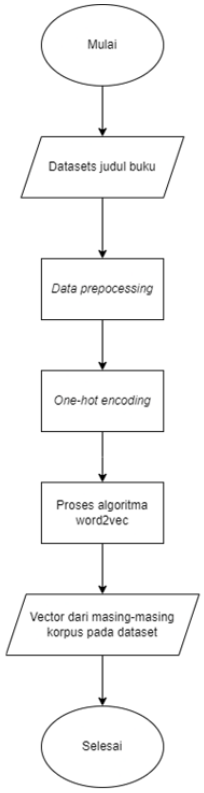
A. Alur Penelitian

Alur penelitian pada penelitian ini dimulai dengan studi literatur, mengidentifikasi masalah, mengidentifikasi kebutuhan data, perancangan model pembelajaran *word2vec*, analisis dan pengujian, dan diakhiri dengan perumusan kesimpulan.



Gambar 1. Alur Penelitian

B. Perancangan Pembelajaran Word2vec

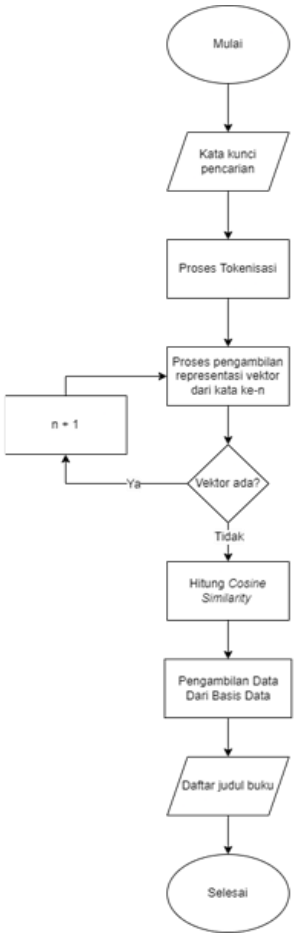


Gambar 2. Perancangan pembelajaran word2vec

Pada Gambar 2. Proses pembelajaran dimulai dengan melakukan *data preprocessing* pada datasets judul buku, proses *preprocessing* ini meliputi proses *tokenization*, *stop word removal*, dan *duplicate word removal*. Kemudian dilanjutkan dengan membuat vector one-hot encoding berdasarkan dari datasets yang sudah dilakukan data preprocessing. Vektor *one-hot encoded* inilah yang akan menjadi *input*-an untuk pembelajaran algoritma *word2vec*, dimana pada akhir pembelajaran akan menghasil vektor-vektor dari setiap kata yang ada pada *datasets*.

C. Perancangan Sistem Rekomendasi Hasil Pencarian

Pada Gambar 3. Proses untuk mendapatkan rekomendasi hasil pencarian dimulai dengan, kata kunci pencarian yang diberikan akan di ubah kedalam bentuk token, token ini digunakan untuk mengetahui apakah token tersebut mempunyai nilai vektor dari hasil proses pembelajaran *word2vec*, jika ada makan vector tersebut akan dilakukan operasi cosine similarity dengan vector-vektor lainnya yang juga merupakan hasil dari pembelajaran *word2vec*.



Gambar 3. Perancangan Sistem Rekomendasi Hasil Pencarian

IV. HASIL DAN PEMBAHASAN

Pengujian rekomendasi hasil pencarian yang dihasilkan oleh sistem dilakukan dengan menggunakan 5 skenario. Skenario yang dimaksud dijelaskan pada Tabel 1.

Tabel 1. Skenario Pengujian

No	Skenario	Kata Kunci Pencarian
1	Skenario 1	Penelitian
2	Skenario 2	Belajar Matematika
3	Skenario 3	Undang Undang
4	Skenario 4	Usaha
5	Skenario 5	Laut Maritim

Hasil dari pengujian masing-masing scenario akan dievaluasi dengan menggunakan uji nilai presisi pada k document (P@k). Proses evaluasi ini dihitung dengan persamaan (2). Hasil dari masing-masing scenario bisa dilihat pada Tabel 2.

Tabel 2. Nilau Uji Presisi pada *k document*

Skenario	Total Rekomendasi Pencarian (*k*)	Total Rekomendasi Relevan	Nilai Presisi pada *k document*
Skenario 1	13	7	0.53
Skenario 2	27	16	0.59
Skenario 3	12	6	0.50
Skenario 4	14	6	0.42
Skenario 5	14	2	0.14

Dalam Skenario 1, terdapat 13 rekomendasi pencarian dengan 7 rekomendasi yang relevan. Nilai presisi pada skenario ini adalah 0.53, yang menunjukkan bahwa model klasifikasi mampu menghasilkan rekomendasi yang relatif akurat dengan tingkat presisi yang cukup baik.

Pada Skenario 2, terdapat 27 rekomendasi pencarian dengan 16 di antaranya relevan. Dalam hal ini, nilai presisi mencapai 0.59, yang menunjukkan peningkatan dalam kualitas rekomendasi dengan adanya peningkatan jumlah rekomendasi yang relevan.

Skenario 3 memiliki 12 rekomendasi pencarian dengan 6 rekomendasi yang relevan. Dalam kasus ini, nilai presisi mencapai 0.50, yang menunjukkan tingkat presisi yang relatif rendah dibandingkan dengan skenario sebelumnya.

Skenario 4 menunjukkan 14 rekomendasi pencarian dengan hanya 6 yang relevan. Nilai presisi pada skenario ini adalah 0.42, yang menunjukkan tingkat presisi yang lebih rendah lagi.

Pada skenario 5, terdapat 14 rekomendasi pencarian dengan hanya 2 yang relevan. Dalam kasus ini, nilai presisi mencapai 0.14, yang menunjukkan tingkat presisi yang sangat rendah.

Berdasarkan nilai presisi pada k document, kualitas rekomendasi pencarian bervariasi dalam setiap skenario. Skenario 1 dan 2 menunjukkan tingkat presisi yang lebih tinggi, sementara skenario 3, 4, dan 5 memiliki tingkat presisi yang lebih rendah.

V. KESIMPULAN

Berdasarkan hasil pengujian dan evaluasi pada beberapa skenario 1, 2, 3, 4, dan 5, dapat ditarik kesimpulan sebagai berikut:

1. Teknik word embedding menggunakan algoritma *word2vec* mampu memberikan rekomendasi hasil pencarian katalog online. Hal ini dibuktikan oleh hasil pengujian pada scenario 1, 2, dan 3 dimana evaluasi nilai

ISSN 2615-6334 (Online)
ISSN 2087-5347 (Print)

presisi pada *k document* yang didapatkan sebesar 0.53, 0.59, dan 0.50 atau untuk setiap hasil pencarian *k document* sistem mampu memberikan setidaknya 50% rekomendasi hasil pencarian yang relevan.

2. Adapun penurunan akurasi dalam skenario-skenario dengan nilai presisi yang rendah dapat disebabkan oleh beberapa faktor. Salah satu diantaranya adalah keterbatasan dalam data latih yang digunakan. Jika dataset yang digunakan untuk melatih model tidak cukup representatif atau tidak mencakup variasi yang memadai dari kata kunci dan judul buku, maka model akan mengalami kesulitan dalam mengklasifikasikan dan menemukan konteks kata dengan akurasi yang tinggi.

VI. SARAN

Berdasarkan hasil pengujian yang menunjukkan variasi dalam kualitas rekomendasi pencarian, terdapat beberapa saran yang dapat diberikan untuk meningkatkan akurasi model dalam memberikan rekomendasi hasil pencarian yang lebih relevan dan bermanfaat.

1. Menggunakan data yang lebih representatif dan mencakup variasi yang memadai dari kata kunci dan judul buku
2. Pemilihan parameter pelatihan yang lebih optimal. Parameter seperti jumlah *epoch*, ukuran jendela, dimensi embedding, dan tingkat pembelajaran (*learning rate*) perlu disesuaikan dengan dataset dan tujuan spesifik dari rekomendasi hasil pencarian.
3. Penentuan kata dalam konteks yang sama menggunakan distributional hypothesis tidak selalu bisa menemukan relasi semantic antara kata, sehingga campur tangan ahli bahasa bisa menjadi pertimbangan untuk mengimprovisasi pemetaan relasi semantic antara korpus kata.
4. Menggunakan pendekatan algoritma yang lebih adaptif. Model word2vec skip-gram yang digunakan dalam penelitian ini memiliki kecenderungan untuk mempelajari hubungan antara kata-kata yang sering muncul bersama. Namun, jika kata kunci pencarian jarang muncul bersama dengan judul buku yang relevan, model akan mengalami kesulitan dalam mengenali pola tersebut.

REFERENSI

[1] Al-Saqqa, S. dkk., (2019). The Use of Word2vec Model in Sentiment Analysis. *Proceedings of the 2019 International Conference on Artificial Intelligence,Robotics and Control*, 39–43.

[2] Azzahra, D. dkk., (2020). Pengembangan Aplikasi Online Public Access Catalog (OPAC) Perpustakaan Berbasis Web Pada STAI Auliaurrasyiddin TEMBILAHAN. *Jurnal Teknologi Dan Sistem Informasi Bisnis*, 2(2), 152–160.

[3] Büttcher, S. dkk., (2010). Information Retrieval: Implementing and Evaluating Search Engines. *MIT Press*.

[4] Di Franco, G. dkk., (2021). Machine learning, artificial neural networks and social research. Quality & Quantity, 55(3), 1007–1025.

[5] Hakak, S. I. dkk., (2019). Exact String Matching Algorithms: Survey, Issues, and Future Research Directions. *IEEE Access*, 7, 69614–69637

[6] Keijsers, N. L. W., (2010). Neural Networks. Dalam Encyclopedia of Movement Disorders (hlm. 257–259). *Elsevier*.

[7] Mikolov, T. dkk., (2013). Efficient Estimation of Word Representations in Vector Space.

[8] Marini, F., (2009). Neural Networks. Dalam Comprehensive Chemometrics (hlm. 477–505). *Elsevier*.

[9] Savittri, S. A. dkk., (2021). A relevant document search system model using word2vec approaches. *Journal of Physics: Conference Series*, 1898(1), 012008.

[10] Silva Fuentes, M. A. dkk., (2019). Semantic Search System using Word Embeddings for query expansion. *2019 IEEE PES Innovative Smart Grid Technologies Conference - Latin America (ISGT Latin America)*, 1–6.

[11] P., S., & Shaji, A. P., (2019). A Survey on Semantic Similarity. *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*, 1–8.

[12] Wang, S. dkk., (2020). A survey of word embeddings based on deep learning. *Computing*, 102(3), 717–740