PENGGUNAAN WORD EMBEDDING UNTUK BILINGUAL INFORMATION RETRIEVAL BAHASA INGGRIS-BAHASA INDONESIA

Skripsi

oleh:

RICHARD LOIS SETIAWAN
71200594

PROGRAM STUDI INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
2024

PENGGUNAAN WORD EMBEDDING UNTUK BILINGUAL INFORMATION RETRIEVAL BAHASA INGGRIS-BAHASA INDONESIA

Skripsi

Diajukan kepada Program Studi Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana
Sebagai Salah Satu Syarat dalam Memperoleh Gelar
Sarjana Komputer

Disusun oleh

RICHARD LOIS SETIAWAN
71200594

PROGRAM STUDI INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
2024

PERNYATAAN KEASLIAN SKRIPSI

Saya menyatakan dengan sesungguhnya bahwa skripsi dengan judul:

PENGGUNAAN WORD EMBEDDING UNTUK BILINGUAL INFORMATION RETRIEVAL BAHASA INGGRIS-BAHASA INDONESIA

yang saya kerjakan untuk melengkapi sebagian persyaratan menjadi Sarjana Komputer pada pendidikan Sarjana Program Studi Informatika Fakultas Teknologi Informasi Universitas Kristen Duta Wacana, bukan merupakan tiruan atau duplikasi dari skripsi kesarjanaan di lingkungan Universitas Kristen Duta Wacana maupun di Perguruan Tinggi atau instansi manapun, kecuali bagian yang sumber informasinya dicantumkan sebagaimana mestinya.

Jika dikemudian hari didapati bahwa hasil skripsi ini adalah hasil plagiasi atau tiruan dari skripsi lain, saya bersedia dikenai sanksi yakni pencabutan gelar kesarjanaan saya.

Yogyakarta, 20 Juni 2024

RICHARD LOIS SETIAWAN
71200594

HALAMAN PERSETUJUAN

Judul Skripsi	:	PENGGUNAAN WORD EMBEDDING UNTUK BILINGUAL INFORMATION RETRIEVAL BAHASA INGGRIS-BAHASA INDONESIA
Nama Mahasiswa	:	RICHARD LOIS SETIAWAN
N I M	:	71200594
Matakuliah	:	Skripsi (Tugas Akhir)
Kode	:	TI0366
Semester	:	Genap
Tahun Akademik	:	2023/2024

Telah diperiksa dan disetujui di
Yogyakarta,
Pada tanggal 20 Juni 2024

Dosen Pembimbing I	Dosen Pembimbing II
Lucia Dwi Krisnawati, Dr. Phil.	Aditya Wikan Mahastama, S.Kom., M.Cs.

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI SKRIPSI/TESIS/DISERTASI UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademika Universitas Kristen Duta Wacana, saya yang bertanda tangan di bawah ini:

Nama	: RICHARD LOIS SETIAWAN
NIM	: 71200594
Program studi	: INFORMATIKA
Fakultas	: FAKULTAS TEKNOLOGI INFORMASI
Jenis Karya	: Skripsi

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Kristen Duta Wacana **Hak Bebas Royalti Noneksklusif** (*None-exclusive Royalty Free Right*) atas karya ilmiah saya yang berjudul:

"PENGGUNAAN WORD EMBEDDING UNTUK BILINGUAL INFORMATION RETRIEVAL BAHASA INGGRIS-BAHASA INDONESIA"

beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti/Noneksklusif ini Universitas Kristen Duta Wacana berhak menyimpan, mengalih media/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat dan mempublikasikan tugas akhir saya selama tetap mencantumkan nama kami sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Yogyakarta
Pada Tanggal : 26 Juni 2024

Yang menyatakan

(RICHARD LOIS SETIAWAN)
NIM.71200594

HALAMAN PENGESAHAN

PENGGUNAAN WORD EMBEDDING UNTUK BILINGUAL INFORMATION RETRIEVAL BAHASA INGGRIS-BAHASA INDONESIA

Oleh: RICHARD LOIS SETIAWAN / 71200594

Dipertahankan di depan Dewan Penguji Skripsi
Program Studi Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana - Yogyakarta
Dan dinyatakan diterima untuk memenuhi salah satu syarat memperoleh gelar
Sarjana Komputer
pada tanggal 11 Juni 2024

Yogyakarta, 20 Juni 2024
Mengesahkan,

Dewan Penguji:

1. Lucia Dwi Krisnawati, Dr. Phil.
2. Aditya Wikan Mahastama, S.Kom., M.Cs.
3. Yuan Lukito, S.Kom., M.Cs.
4. Gloria Virginia, S.Kom., MAI, Ph.D.

Pekan

(Restyandito,S.Kom.,MSIS.,Ph.D.)

Ketua Program Studi

(Joko Purwadi, S.Kom., M.Kom.)

Karya sederhana ini dipersembahkan

kepada Tuhan, Keluarga Tercinta,

dan Kedua Orang Tua

Lakukan segala sesuatu seperti untuk Tuhan bukan untuk manusia

Anonim

Through hard work, perseverance and a faith in God, you can live your dreams.

(Ben Carson)

KATA PENGANTAR

Segala puji dan syukur kepada Tuhan yang maha kasih, karena atas segala rahmat, bimbingan, dan bantuan-Nya maka akhirnya Skripsi dengan judul Penggunaan Word Embedding Untuk Bilingual Information Retrieval Bahasa Inggris – Bahasa Indonesia ini telah selesai disusun.

Penulis memperoleh banyak bantuan dari kerja sama baik secara moral maupun spiritual dalam penulisan Skripsi ini, untuk itu tak lupa penulis ucapkan terima kasih yang sebesar-besarnya kepada:

1. Tuhan Yesus Kristus yang maha kasih,
2. Orang tua, Andika Setiawan dan Anna Novia yang selama ini telah memberikan dukungan secara moral dan material serta mendoakan penulis sehingga penulis mampu menyelsaikan skripsi ini,
3. Bapak Restyandito, S.Kom, MSIS., Ph.D. selaku Dekan FTI, yang telah memberikan dukungan dan bimbingan dalam menyelesaikan skripsi ini,
4. Bapak Joko Purwadi, S.Kom, M.Kom. selaku Kaprodi Informatika, yang telah memberikan arahan dan kesempatan untuk mengeksplorasi topik ini,
5. Ibu Dr. Phil. Lucia Dwi Krisnawati, S.S., M.A. selaku Dosen Pembimbing 1, yang telah berbagi ilmu, waktunya, dan kesabaran dalam membimbing penulis selama proses penulisan skripsi,
6. Bapak Aditya Wikan Mahastama, S.Kom., M.Cs. , selaku Dosen Pembimbing 2 yang telah memberikan ilmu dan kesabaran dalam membimbing penulis,
7. Ibu Maria Nila, Ibu Andhika Galuh, Ibu Agata Filiana, Mas Silvanus Satno, dan Mba Grace Shinta yang selalu senantiasa mendukung dan menolong penulis ketika ada kesulitan serta mendengar keluh kesah penulis,
8. Ko Niko, Ci Cepi, serta keluarga tercinta yang selalu memberikan dukungan, semangat, dan doa dalam setiap langkah perjalanan penulis,
9. Teman-teman Informatika UKDW 2020 yang telah memberikan dukungan selama pengerjaan skripsi,

10. Marco Juan, Vincensia Eugene, Anagracia Audrey, Gracielle Austin, Maria Angeline, Belinda Patricia, Sheren Cindy, Christina Michelle, dan kawan-kawan lainnya yang selalu memberikan dukungan, semangat, dan hiburan kepada penulis selama pengerjaan skripsi ini,
11. Lain-lain yang telah mendukung moral, spiritual, dan dana untuk belajar selama ini.

Laporan proposal/skripsi ini tentunya tidak lepas dari segala kekurangan dan kelemahan, untuk itu segala kritikan dan saran yang bersifat membangun guna kesempurnaan skripsi ini sangat diharapkan. Semoga skripsi ini dapat bermanfaat bagi pembaca semua dan lebih khusus lagi bagi pengembangan ilmu komputer dan teknologi informasi.

Yogyakarta, 20 Juni 2024

Richard Lois Setiawan

DAFTAR ISI

PERNYATAAN KEASLIAN SKRIPSI iii
HALAMAN PERSETUJUAN iv
HALAMAN PENGESAHAN v
HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS SECARA ONLINE UNIVERSITAS KRISTEN DUTA WACANA YOGYAKARTA...**Error! Bookmark not defined.**
KATA PENGANTAR ix
DAFTAR ISI xi
DAFTAR TABEL xiv
DAFTAR GAMBAR xv
DAFTAR LISTING xvii
INTISARI xviii
ABSTRACT xx
BAB I PENDAHULUAN 1
1.1 Latar Belakang Masalah 1
1.2 Perumusan Masalah 2
1.3 Batasan Masalah 2
1.4 Tujuan Penelitian 2
1.5 Manfaat Penelitian 3
1.6 Metodologi Penelitian 3
1.7 Sistematika Penulisan 4
BAB II TINJAUAN PUSTAKA DAN DASAR TEORI 6
2.1 Tinjauan Pustaka 6
2.2 Landasan Teori 8
2.2.1 Information Retrieval 8
2.2.2 Bilingual Information Retrieval 10
2.2.3 Word Embedding 10
2.2.4 Preprocessing 12

2.2.5 Universal Sentence Encoder (USE) 13
2.2.6 Facebook AI Similarity Search (FAISS)..... 17
2.2.7 Evaluasi 19
BAB III METODE PENELITIAN..... 22
3.1 Analisis Kebutuhan Sistem 22
3.2 Perancangan Penelitian..... 23
3.2.1 Studi literatur..... 23
3.2.2 Pengumpulan dan Pengolahan Data..... 23
3.2.3 Pengembangan Sistem 24
3.2.4 Evaluasi 26
3.2.5 Penulisan Laporan..... 26
3.2.6 Pembangunan Data Latih dan Data Uji..... 26
3.3 Perancangan Antar muka Pengguna..... 28
3.4 Perancangan Pengujian Sistem..... 29
BAB IV IMPLEMENTASI DAN PEMBAHASAN 30
4.1 Implementasi Awal..... 30
4.1.1 Pengumpulan Data 30
4.1.2 Persiapan Data..... 32
4.1.3 Antarmuka..... 36
4.2 Implementasi Sistem 36
4.2.1 Pembuatan Dataset 36
4.2.2 Prapemrosesan..... 38
4.2.3 Pembuatan Word Embedding 39
4.2.4 Pembuatan Indeks 40
4.2.5 Pencarian Dokumen 42
4.3 Pengujian dan Analisis 42

4.3.1 Skenario Pertama: Kueri Bahasa Indonesia dan Mengambil 10 Dokumen 44
4.3.2 Skenario Kedua: Kueri Bahasa Indonesia dan Mengambil 20 Dokumen 48
4.3.3 Skenario Ketiga: Kueri Bahasa Inggris dan Mengambil 10 Dokumen 52
4.3.4 Skenario Keempat: Kueri Bahasa Inggris dan Mengambil 20 Dokumen 56
4.3.5 Skenario Kelima: Kueri Campuran Bahasa Indonesia dan Inggris serta Mengambil 10 Dokumen 59
4.3.6 Skenario Keenam: Kueri Campuran Bahasa Indonesia dan Inggris serta Mengambil 20 Dokumen 62
4.3.7 Analisis Seluruh Hasil Pengujian 65
BAB V KESIMPULAN DAN SARAN 73
5.1 Kesimpulan 73
5.2 Saran 74
DAFTAR PUSTAKA 75
LAMPIRAN A KODE SUMBER PROGRAM 79
LAMPIRAN B KARTU KONSULTASI DOSEN 1 104
LAMPIRAN C KARTU KONSULTASI DOSEN 2 105

DAFTAR TABEL

Tabel 2.1 Contoh Penerapan Punctuation Removal 13
Tabel 2.2 Contoh Penerapan Case Folding 13
Tabel 3.1 Hasil Pencarian Kueri 26
Tabel 3.2 Profil Data Latih 27
Tabel 4.1 Hasil Precision Seluruh Skenario 65
Tabel 4.2 Hasil Recall Seluruh Skenario 68
Tabel 4.3 Hasil F1-Score Seluruh Skenario 70

DAFTAR GAMBAR

Gambar 2.1 Ilustrasi one-hot vectors dan word embedding vectors (Jiao & Zhang, 2021) 11
Gambar 2.2 Contoh Heat Map Textual Similarity Dengan USE 16
Gambar 3.1 Flowchart Alur Penelitian 23
Gambar 3.2 Contoh Data Korpus 23
Gambar 3.3 Blok Diagram Sistem 24
Gambar 3.4 File CSV Ground Truth 27
Gambar 3.5 File CSV Kueri Pencarian 28
Gambar 3.6 Rancangan Antar Muka Pengguna 28
Gambar 4.1 Contoh Dokumen dari Wikipedia 31
Gambar 4.2 Contoh Dokumen yang Telah Diubah Formatnya 31
Gambar 4.3 Data Frame Kueri Pencarian 32
Gambar 4.4 Hasil Pelabelan 35
Gambar 4.5 Halaman Antarmuka Sistem 36
Gambar 4.6 Hasil Prapemrosesan 38
Gambar 4.7 Hasil Vektor Word Embedding 39
Gambar 4.8 Contoh Kueri Pencarian Dalam 2 Bahasa 43
Gambar 4.9 Hasil Pencarian Tiap Kueri 44
Gambar 4.10 Hasil Pengujian Indeks IVF Pada Skenario Pertama 45
Gambar 4.11 Hasil Pengujian Seluruh Indeks Untuk Skenario Pertama 46
Gambar 4.12 Hasil Pengujian Indeks IVF Pada Skenario Kedua 49
Gambar 4.13 Hasil Pengujian Seluruh Indeks Untuk Skenario Kedua 50
Gambar 4.14 Hasil Pengujian Indeks IVF Pada Skenario Ketiga 53
Gambar 4.15 Hasil Pengujian Seluruh Indeks Untuk Skenario Ketiga 54
Gambar 4.16 Hasil Pengujian Indeks IVF Pada Skenario Keempat 56
Gambar 4.17 Hasil Pengujian Seluruh Indeks Untuk Skenario Keempat 57
Gambar 4.18 Hasil Pengujian Indeks IVF Pada Skenario Kelima 59

Gambar 4.19 Hasil Pengujian Seluruh Indeks Untuk Skenario Kelima 60
Gambar 4.20 Hasil Pengujian Indeks IVF Pada Skenario Keenam 62
Gambar 4.21 Hasil Pengujian Seluruh Indeks Untuk Skenario Keenam 63
Gambar 4.22 Grafik Precision Untuk Berbagai Indeks Dalam Enam Skenario .. 66
Gambar 4.23 Grafik Recall Untuk Berbagai Indeks Dalam Enam Skenario 68
Gambar 4.24 Grafik F1-Score Untuk Berbagai Indeks Dalam Enam Skenario .. 70

DAFTAR LISTING

Listing 4.1 Listing Pembuatan CSV Data 33
Listing 4.2 Pembacaan File Data Latih 34
Listing 4.3 Pembuatan Dataset 37
Listing 4.4 Pseudocode Prapemrosesan 38
Listing 4.5 Pembuatan Word Embedding 39
Listing 4.6 Pembuatan Indeks 41
Listing 4.7 Fungsi Pencarian 42

INTISARI

PENGGUNAAN WORD EMBEDDING UNTUK BILINGUAL INFORMATION RETRIEVAL BAHASA INGGRIS-BAHASA INDONESIA

Oleh

RICHARD LOIS SETIAWAN

71200594

Dalam era digital, informasi memainkan peran penting yang mendominasi berbagai aspek kehidupan. Namun, banyak informasi penting seperti jurnal, artikel, dan publikasi penelitian tersedia dalam Bahasa Inggris. Informasi ini sering kali sulit diterjemahkan dengan akurat ke Bahasa Indonesia. Untuk itu referensi asli dalam bahasa Inggris tetap menjadi kebutuhan utama. Kesulitan dalam menemukan informasi dalam bahasa Inggris sering kali disebabkan oleh keterbatasan dalam menuliskan kata kunci, sehingga lebih praktis menuliskannya dalam bahasa Indonesia. Penelitian ini bertujuan mengatasi tantangan ini dengan pendekatan *Bilingual Information Retrieval* (BIR) dan menggunakan teknik *word embedding* untuk merepresentasikan data.

Data yang digunakan bersumber dari Wikipedia, terdiri dari 150 dokumen dalam bahasa Indonesia dan Inggris. Vektor *word embedding* dibuat menggunakan model Universal Sentence Encoder (USE) buatan Google. Vektor ini kemudian di indeks menggunakan Facebook AI Similarity Search (FAISS). Untuk melakukan pencarian maka dibentuk juga vektor *embedding* dari kueri yang kemudian dibandingkan dengan vektor di indeks menggunakan berbagai model indeks yang disediakan dari FAISS. Sistem yang telah dibuat dievaluasi menggunakan metrik pengukuran *precision*, *recall*, dan *f1-score*.

Word embedding berhasil diterapkan dalam pengembangan sistem BIR dengan. Model indeks FAISS yang digunakan ada 6 yaitu FlatIP (*inner product*), FlatL2 (*euclidean distance*), *inverted file* (IVF IP), *Hierarchical Navigable Small World* (HNSW), *product quantization* (PQ), dan IVF PQ. Sistem dievaluasi enam skenario pengujian dengan kueri berbagai bahasa dan jumlah dokumen yang diambil berbeda. Indeks FlatIP dan Flat L2 menunjukkan performa yang serupa dengan nilai *precision*, *recall*, dan *F1-Score* yang lebih rendah di semua skenario pengujian. Indeks IVF dengan 20 kluster memiliki mendapatkan *precision* dan *recall* yang tinggi. Indeks HNSW memiliki performa serupa dengan FlatIP dan FlatL2, namun dengan kueri berbahasa Inggris perofrmanya lebih baik. Indeks PQ memiliki performa terburuk dari semua skenario. Kombinasi indeks IVF IP 20 dan PQ menunjukkan performa yang baik di semua skenario.

Kata-kata kunci : *word embedding, bilingual information retrieval, universal sentence encoder, FAISS*

ABSTRACT

BILINGUAL INFORMATION RETRIEVAL ENGLISH-INDONESIA USING WORD EMBEDDING

By

RICHARD LOIS SETIAWAN

71200594

In the digital era, information plays a crucial role that dominates various aspects of life. However, much important information such as journals, articles, and research publications is available in English, often difficult to accurately translate into Indonesian. Hence, original references in English remain essential. Difficulty in finding information in English is often due to limitations in writing keywords, making it more practical to write them in Indonesian. This study aims to address these challenges using Bilingual Information Retrieval (BIR) approach and employing word embedding techniques for data representation.

The data used is sourced from Wikipedia, comprising 150 documents in both Indonesian and English. Word embedding vectors are created using Google's Universal Sentence Encoder (USE). These vectors are then indexed using Facebook AI Similarity Search (FAISS). For searching, query embedding vectors are also formed and compared with indexed vectors using various FAISS index models. The system is evaluated using precision, recall, and F1-score metrics.

Word embedding has been successfully applied in developing the BIR system. Six FAISS index models were evaluated: FlatIP (inner product), FlatL2 (euclidean distance), inverted file (IVF IP), Hierarchical Navigable Small World (HNSW), product quantization (PQ), and IVF PQ. The system was tested across six different scenarios with queries in various languages and different numbers of documents retrieved. FlatIP and FlatL2 indices show similar performance with

lower precision, recall, and F1-scores across all testing scenarios. IVF with 20 clusters achieves high precision and recall. HNSW performs similarly to FlatIP and FlatL2 but shows better performance with English queries. PQ exhibits the poorest performance across all scenarios. The combination of IVF IP 20 and PQ demonstrates good performance across all scenarios.

Keywords : word embedding, bilingual information retrieval, universal sentence encoder, FAISS

BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Dalam era digital, informasi memegang peranan yang penting dan mendominasi berbagai aspek kehidupan. Internet telah menjadi salah satu wadah utama tempat beredarnya informasi. Namun, sebagian besar informasi ini tersedia dalam Bahasa Inggris (W3Techs, 2023). Hal serupa terjadi pada pengetahuan yang tersedia melalui jurnal, artikel, dan publikasi penelitian, di mana sekitar 75% dari konten tersebut tersedia dalam Bahasa Inggris (Curcic, 2023). Bahkan, inovasi teknologi yang terus bermunculan sering kali diperkenalkan melalui bahasa Inggris.

Informasi, jurnal, dan artikel yang awalnya ditulis dalam bahasa Inggris seringkali mengalami proses terjemahan ke bahasa Indonesia. Namun, informasi yang telah diterjemahkan ini tidak selalu akurat dan bisa menjadi sumber kebingungan. Oleh karena itu, informasi yang tertulis dalam bahasa Inggris dibutuhkan sebagai referensi asli. Selain itu, masalah lainnya adalah ketika kita ingin mencari informasi yang tersedia dalam bahasa Inggris, namun kesulitan untuk menuliskan kata kunci dalam bahasa Inggris. Kata kunci yang terlintas dalam ingatan seringkali adalah dalam bahasa Indonesia karena lebih familiar. Dalam konteks *information retrieval* tradisional, kedua masalah ini tidak dapat terselesaikan (Rahmanda dkk., 2019).

Penelitian ini mencoba menyelesaikan beberapa masalah mengenai pencarian informasi dalam dua bahasa yaitu bahasa Inggris dan bahasa Indonesia. *Bilingual information retrieval* (BIR) atau biasa disebut *cross-lingual information retrieval* (CLIR) merupakan sebuah pendekatan lain dari *information retrieval*. BIR dapat memberikan luaran dokumen dari bahasa target sedangkan kueri yang diberikan pengguna berupa bahasa sumber (Litschko dkk., 2022). Bahasa sumber pada penelitian ini adalah bahasa Indonesia, sedangkan bahasa targetnya bahasa Inggris. Penelitian ini mencoba memanfaatkan *word embedding* yang sudah

populer dan banyak digunakan (Hambarde & Proenca, 2023). Perbedaan penelitian ini dengan yang lain adalah bahasa yang digunakan sebagai bahasa sumber dan bahasa target serta model *embedding* yang digunakan.

1.2 Perumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, berikut ini merupakan beberapa rumusan masalah dalam penelitian ini.

1. Apakah penggunaan *word* embedding sesuai untuk *bilingual information retrieval bahasa Inggris – bahasa Indonesia*?
2. Bagaimana hasil evaluasi *precision*, *recall*, dan F1-*Score* yang dihasilkan dari *bilingual information retrieval* dengan menggunakan *word embedding*?

1.3 Batasan Masalah

Penelitian ini memiliki batasan masalah sebagai berikut.

1. Bahasa yang digunakan sebagai korpus dan masukan pengguna adalah bahasa Indonesia dan bahasa Inggris.
2. Jumlah dokumen yang digunakan sebagai korpus sejumlah 150 dokumen untuk tiap bahasa.
3. Format dokumen yang dijadikan korpus adalah *.txt.
4. *Preprocessing* pada teks yang akan digunakan adalah *punctuation removal* dan *case folding*.
5. Isi dari dokumen didapatkan dari laman https://www.wikipedia.org/.
6. Evaluasi yang digunakan adalah *precision*, *recall*, dan F1-score.

1.4 Tujuan Penelitian

Penelitian ini bertujuan untuk mengembangkan sistem *bilingual information retrieval* dengan menggunakan *word embedding*. Sistem mampu memberikan luaran dokumen yang relevan dalam kedua bahasa yang digunakan. Tujuan lainnya mengevaluasi keefektifan penggunaan *word embedding* dalam sistem *bilingual information retrieval* antara bahasa Inggris dan bahasa Indonesia.

1.5 Manfaat Penelitian

Berdasarkan latar belakang dan tujuan penelitian, penelitian ini dapat memberikan manfaat berikut.

1. Penelitian ini membantu peneliti untuk bisa mengembangkan sistem *bilingual information retrieval* dengan menggunakan *word embedding* yang tepat dan baik. Peneliti mampu memahami konsep *word embedding* dan *bilingual information retrieval* secara baik dan cara mengimplementasikannya dalam sistem.
2. Hasil penelitian ini dapat menjadi wawasan baru dalam pembuatan pencarian informasi lintas bahasa terutama bahasa Inggris dan bahasa Indonesia. Penelitian dapat menjadi referensi baru ketika para akademisi akan mengembangkan sistem yang serupa.
3. Pengguna secara umum dapat merasakan manfaat penelitian ini berupa pengaksesan informasi yang lebih mudah. Pencarian informasi tidak terbatas pada bahasa yang digunakan sebagai masukan dari pengguna saja. Waktu yang dihabiskan oleh pengguna dalam melakukan pencarian akan lebih efisien.

1.6 Metodologi Penelitian

Metode penelitian yang dilakukan pada penelitian ini adalah sebagai berikut.

1. Studi Literatur

Penelitian dimulai dengan studi literatur untuk memperdalam teori terkait penelitian yang akan dilakukan. Literatur yang dipelajari antara lain terkait *bilingual information retrieval*, *word embedding*, Universal Sentence Encoder (USE), dan Facebook AI Similarity Search (FAISS).

2. Pengumpulan dan Pengolahan Data

Penelitian ini membutuhkan data berupa dokumen bahasa Indonesia dan bahasa Inggris. Dokumen-dokumen tersebut didapatkan dari laman https://www.wikipedia.org/. Proses pengumpulan data dilakukan dengan pencarian konten dengan judul yang sama dalam kedua bahasa, diikuti dengan

pengecekan semantik kata-kata dalam konten tersebut. Jika ditemukan kata-kata yang tidak memiliki padanan langsung, kata-kata dalam bahasa Inggris dijadikan sebagai acuan utama dan diterjemahkan secara manual ke dalam bahasa Indonesia. Hasilnya adalah data yang telah disesuaikan semantiknya dalam kedua bahasa. Data dari Wikipedia disimpan dalam format file teks (.txt).

3. Pembangunan Sistem

Proses utama dalam penelitian ini merupakan pembangunan sistem *bilingual information retrieval*. Data yang telah selesai diolah akan masuk ke dalam tahap preproses seperti *punctuation removal* dan *case folding*. Data tersebut kemudian dibuat *word embedding*nya dengan menggunakan USE. Vektor *embedding* dibuat indeksnya dengan menggunakan FAISS untuk mempermudah melakukan pencarian.

4. Evaluasi

Sistem yang telah selesai dibangun perlu dilakukan evaluasi untuk mengukur performanya. Indikator yang digunakan untuk evaluasi pada penelitian ini adalah *precision*, *recall*, dan F1-Score.

5. Penulisan Laporan

1.7 Sistematika Penulisan

Laporan skripsi ini disusun dengan sistematika bagian pertama, terdiri dari empat bab:

- BAB I berisi latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, metodologi penelitian, dan sistematikan penulisan laporan penelitian.
- BAB II berisi mengenai tinjauan pustaka yang digunakan dalam mendukung penelitian ini dan teori-yang menjadi dasar penelitian.
- BAB III berisi rancangan penelitian yang digunakan, pengumpulan dan pengolahan data, pengembangan aplikasi, dan evaluasi. Bab ini akan menjelaskan tahapan-tahapan yang dilakukan dalam penelitian dari awal hingga akhir.

- BAB IV akan menguraikan hasil implementasi dari penelitian yang telah dilakukan, serta membahas hasil evaluasi yang didapatkan dari penelitian. Bab ini akan memberikan wawasan yang lebih mendalam tentang implementasi aplikasi dan hasilnya.
- BAB V akan merangkum kesimpulan yang diambil dari hasil penelitian. Selain itu, bab ini akan memberikan saran-saran yang relevan untuk penelitian selanjutnya atau pengembangan lebih lanjut pada topik ini.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Penelitian ini berhasil mengembangkan sistem *bilingual information retrieval* (BIR) dengan menggunakan *word embedding*. Sistem telah mampu memberikan luaran dokumen yang relevan dalam kedua bahasa yang digunakan. Berdasarkan hasil yang didapatkan dari pengujian yang telah dilakukan, penggunaan *word embedding* terbukti sesuai dan bermanfaat untuk sistem BIR. *Word embedding* memungkinkan kata-kata direpresentasikan dalam bentuk vektor dengan tetap mempertahankan konteks dan makna semantik. Hal ini sangat penting untuk mengatasi perbedaan bahasa dan membantu dalam menemukan informasi yang lebih sesuai dan relevan. *Embedding* yang dihasilkan untuk kueri bahasa Inggris memberikan hasil yang lebih baik dibandingkan kueri dalam bahasa Indonesia. Ini terjadi karena model USE lebih banyak dilatih dalam bahasa Inggris dibanding bahasa Indonesia.

Kinerja sistem diuji dengan menggunakan metrik *precision*, *recall*, dan *F1-Score* pada 6 model indeks dari FAISS. Indeks FlatIP dan FlatL2 menunjukkan performa yang serupa dengan nilai *precision*, *recall*, dan *F1-Score* yang lebih rendah di semua skenario. Indeks IVF, terutama IVF IP dengan 20 kluster menunjukkan performa terbaik dalam beberapa skenario dengan *precision* dan *recall* yang tinggi. Indeks HNSW memiliki performa serupa dengan FlatIP dan FlatL2 untuk kueri berbahasa Indonesia tetapi menunjukkan peningkatan *recall* yang signifikan untuk kueri berbahasa Inggris, meskipun *precision*nya tetap rendah. Indeks PQ memiliki performa terburuk dibandingkan dengan indeks lainnya, meskipun unggul dalam mengurangi dimensi data dan penggunaan memori. Kombinasi indeks IVF IP 20 PQ menunjukkan performa yang baik di semua skenario, menggabungkan keunggulan IVF IP 20 dalam menangkap dokumen relevan dengan efisiensi pengurangan dimensi data dari PQ.

5.2 Saran

Berdasarkan hasil penelitian yang telah dilakukan, berikut beberapa saran yang dapat diajukan:

1. Jumlah data latih yang digunakan dalam penelitian ini dapat ditingkatkan. Dengan penambahan jumlah data latih, kita dapat menguji apakah performa BIR akan meningkat.
2. Sistem yang telah dibuat dapat diuji dengan menggunakan model *word embedding* yang berbeda seperti BERT dan ELMo.
3. Sistem yang telah dibuat belum dapat digunakan oleh publik. Oleh karena itu, pada penelitian selanjutnya sistem dapat dipublikasikan ke dalam server publik agar dapat diakses secara luas.

DAFTAR PUSTAKA

Abka, A. F., Azizah, K., & Jatmiko, W. (2022). Transformer-based Cross-Lingual Summarization using Multilingual Word Embeddings for English - Bahasa Indonesia. *International Journal of Advanced Computer Science and Applications*, *13*(12). https://doi.org/10.14569/IJACSA.2022.0131276

Adriani, M., Hayurani, H., & Sari, S. (2008). Indonesian-English Transitive Translation for Cross-Language Information Retrieval. Dalam *LNCS* (Vol. 5152). http://www.lemurproject.org/

Almeida, F., & Xexéo, G. (2019). *Word Embeddings: A Survey*. http://arxiv.org/abs/1901.09069

Anggastya, D. A. (2019). *Perbandingan Frequency Based dan Prediction Based dalam Model Support Vector Machine (SVM) Multiclass (Studi Kasus: Teks Berita)*. Politeknik Negeri Bandung.

Ayudita, I. M., Adikara, P. P., & Indriati. (2018). Sistem Pencarian Jurnal Ilmiah Cross Language dengan Metode Vector Space Model (VSM). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, *2*(12), 6837–6841. http://j-ptiik.ub.ac.id

Azmi, M., Ali, E., & Saputra Wijaya, Y. (2020). Perbandingan Boolean Model Dan Vector Space Model Dalam Pencarian Dokumen Teks. *Jurnal Teknologi Informasi & Komunikasi*, *11*, 268–277. https://doi.org/10.31849/digitalzone.v11i2.4168CCS

Bhattacharya, P., Goyal, P., & Sarkar, S. (2016). Using Word Embeddings for Query Translation for Hindi to English Cross Language Information Retrieval. *Computacion y Sistemas*, *20*(3), 435–447. https://doi.org/10.13053/CyS-20-3-2462

Bintana, R. R., Fatichah, C., & Purwitasari, D. (2018). Pencarian Question-Answer Menggunakan Convolutional Neural Network Pada Topik Agama Berbahasa Indonesia. *Jurnal ULTIMATICS*, *10*(1), 57–64. https://doi.org/10.31937/ti.v10i1.842

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). *Enriching Word Vectors with Subword Information*. http://arxiv.org/abs/1607.04606

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. St., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., & Kurzweil, R. (2018). *Universal Sentence Encoder*. http://arxiv.org/abs/1803.11175

Curcic, D. (2023, Juni 1). *Number of Academic Papers Published Per Year*. https://wordsrated.com/number-of-academic-papers-published-per-year/

Das, M., Kamalanathan, S., & Alphonse, P. J. A. (2020). *A Comparative Study on TF-IDF Feature Weighting Method and its Analysis using Unstructured Dataset*.

Dayton, F. (2023). *Vector Similarity Search: A Deeper Dive*.

Hambarde, K. A., & Proenca, H. (2023). *Information Retrieval: Recent Advances and Beyond*. https://doi.org/10.1109/ACCESS.2023.3295776

Hirst, G., Lin, J., Dyer, C., Wong, K.-F., Li, W., Xu, R., & Zhang, Z.-S. (2010). *Cross-Language Information Retrieval Jian-Yun*.

Hofmann, T. (1999). *Probabilistic Latent Semantic Analysis*.

Jiang, Z., El-Jaroudi, A., Hartmann, W., Karakos, D., & Zhao, L. (2020). *Cross-lingual Information Retrieval with BERT*. http://arxiv.org/abs/2004.13005

Jiao, Q., & Zhang, S. (2021). A Brief Survey of Word Embedding and Its Recent Development. *IAEAC 2021 - IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference*, 1697–1701. https://doi.org/10.1109/IAEAC50856.2021.9390956

Johnson, J., Douze, M., & Jégou, H. (2017). *Billion-scale similarity search with GPUs*. http://arxiv.org/abs/1702.08734

Krisnawati, L. D., & Mahastama, A. W. (2023). *Penggunaan Word Embedding Dalam Temu Kembali Dokumen Sumber*.

Lashkari, A. H., Mahdavi, F., & Ghomi, V. (2009). A boolean model in information retrieval for search engines. *Proceedings - 2009 International Conference on Information Management and Engineering, ICIME 2009*, 385–389. https://doi.org/10.1109/ICIME.2009.101

Litschko, R., Glavaš, G., Ponzetto, S. P., & Vulić, I. (2018). *Unsupervised Cross-Lingual Information Retrieval using Monolingual Data Only*.

Litschko, R., Vulić, I., Ponzetto, S. P., & Glavaš, G. (2022). On cross-lingual retrieval with multilingual text encoders. *Information Retrieval Journal*, *25*(2), 149–183. https://doi.org/10.1007/s10791-022-09406-x

Li, Y., & Yang, T. (2018). Word embedding for understanding natural language: A survey. Dalam *Guide to Big Data Applications* (Vol. 26, hlm. 83–104). Springer Nature.

Lomeli, M. (2023, Mei 25). *Faiss Indexes*. https://github.com/facebookresearch/faiss/wiki/Faiss-indexes.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*.

Nurdin, A., Aji, B. A. S., Bustamin, A., & Abidin, Z. (2020). PERBANDINGAN KINERJA WORD EMBEDDING WORD2VEC, GLOVE, DAN FASTTEXT PADA KLASIFIKASI TEKS. *Jurnal TEKNOKOMPAK*, *14*(2), 74.

Nurrohmat, M. A., & Azhari. (2019). Sentiment Analysis of Novel Review Using Long Short-Term Memory Method. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *13*(3), 209. https://doi.org/10.22146/ijccs.41236

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. http://www.aclweb.org/anthology/D14-1162

Rahmanda, R. A., Adriani, M., & Tanaya, D. (2019). Cross Language Information Retrieval Using Parallel Corpus with Bilingual Mapping Method. *2019 International Conference on Asian Language Processing (IALP)*, 222–227. https://doi.org/10.1109/IALP48816.2019.9037705

Roshdi, A., & Roohparvar, A. (2015). Review: Information Retrieval Techniques and Applications. Dalam *International Journal of Computer Networks and Communications Security* (Vol. 3, Nomor 9). www.ijcncs.org

Savittri, S. A., Amalia, A., & Budiman, M. A. (2021). A relevant document search system model using word2vec approaches. *Journal of Physics: Conference Series*, *1898*(1). https://doi.org/10.1088/1742-6596/1898/1/012008

Schwarz, M., Chapman, K., & Häussler, B. (2022). *Multilingual Medical Entity Recognition and Cross-lingual Zero-Shot Linking with Facebook AI Similarity Search*. https://icd.who.int/en

Suhartono, D. (2014). *Probabilistic Latent Semantic Analysis (PLSA) untuk Klasifikasi Dokumen Teks Berbahasa Indonesia*.

Vulić, I., & Moens, M. F. (2015). Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. *SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 363–372. https://doi.org/10.1145/2766462.2767752

W3Techs. (2023, September 15). *Usage statistics of content languages for websites*. https://w3techs.com/technologies/overview/content_language

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G. H., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., & Kurzweil, R. (2019). *Multilingual Universal Sentence Encoder for Semantic Retrieval*. http://arxiv.org/abs/1907.04307

Ye, X., Shen, H., Ma, X., Bunescu, R., & Liu, C. (2016). From word embeddings to document similarities for improved information retrieval in software engineering. *Proceedings - International Conference on Software Engineering, 14-22-May-2016*, 404–415. https://doi.org/10.1145/2884781.2884862