

**IMPLEMENTASI *NAMED ENTITY RECOGNITION* DALAM
SISTEM OTOMATISASI PEMBERIAN INDEKS DOKUMEN
PADA APACHE SOLR**

Skripsi



oleh:

**EUSEBIA FERNANDA SIAHAAN
71180323**

**PROGRAM STUDI INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA**

2023

**IMPLEMENTASI *NAMED ENTITY RECOGNITION* DALAM
SISTEM OTOMATISASI PEMBERIAN INDEKS DOKUMEN
PADA APACHE SOLR**

Skripsi



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

Disusun oleh

EUSEBIA FERNANDA SIAHAAN

71180323

PROGRAM STUDI INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA

2023

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 : Eusebia Fernanda Siahaan
NIM : 71180323
Program studi : Informatika
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:

**IMPLEMENTASI NAMED ENTITY RECOGNITION DALAM SISTEM
OTOMATISASI PEMBERIAN INDEKS DOKUMEN PADA APACHE SOLR**

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 : Jakarta
Pada Tanggal : 21 Agustus 2023

Yang menyatakan


(Eusebia Fernanda Siahaan)
71180323

PERNYATAAN KEASLIAN SKRIPSI

Saya menyatakan dengan sesungguhnya bahwa skripsi dengan judul:

IMPLEMENTASI *NAMED ENTITY RECOGNITION* DALAM SISTEM OTOMATISASI PEMBERIAN INDEKS DOKUMEN PADA APACHE SOLR

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.

Jakarta, 8 Juni 2023



EUSEBIA FERNANDA SIAHAAN
71180323

HALAMAN PERSETUJUAN

Judul Skripsi : IMPLEMENTASI NAMED ENTITY
RECOGNITION DALAM SISTEM OTOMATISASI
PEMBERIAN INDEKS DOKUMEN PADA
APACHE SOLR

Nama Mahasiswa : EUSEBIA FERNANDA SIAHAAN

N I M : 71180323

Matakuliah : Skripsi (Tugas Akhir)

Kode : TI0366

Semester : Genap

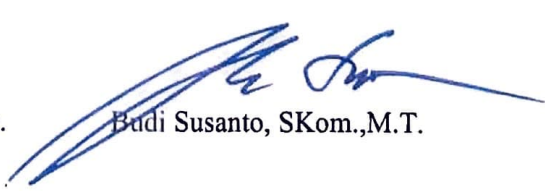
Tahun Akademik : 2022/2023

Telah diperiksa dan disetujui di
Yogyakarta,
Pada tanggal 19 Juli 2023

Dosen Pembimbing I

Dosen Pembimbing II


Gloria Virginia, S.Kom., MAI, Ph.D.


Budi Susanto, SKom., M.T.

HALAMAN PENGESAHAN

IMPLEMENTASI NAMED ENTITY RECOGNITION DALAM SISTEM OTOMATISASI PEMBERIAN INDEKS DOKUMEN PADA APACHE SOLR

Oleh: EUSEBIA FERNANDA SIAHAAN / 71180323

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 21 Juni 2023

Yogyakarta, 19 Juli 2023
Mengesahkan,

Dewan Penguji:

1. Gloria Virginia, S.Kom., MAI, Ph.D.
2. Budi Susanto, SKom.,M.T.
3. I Kadek Dendy S., S.T., M.Eng.
4. Willy Sudiarto Raharjo, S.Kom.,M.Cs.



Dekan



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

Ketua Program Studi



(Gloria Virginia Ph.D.)

**HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI
TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS
SECARA ONLINE
UNIVERSITAS KRISTEN DUTA WACANA YOGYAKARTA**

Saya yang bertanda tangan di bawah ini:

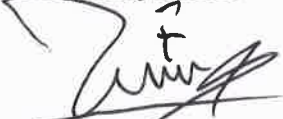
NIM : 71180323
Nama : Eusebia Fernanda Siahaan
Prodi / Fakultas : Teknologi Informasi / Informatika
Judul Tugas Akhir : IMPLEMENTASI *NAMED ENTITY
RECOGNITION* DALAM SISTEM
OTOMATISASI PEMBERIAN INDEKS
DOKUMEN PADA APACHE SOLR

bersedia menyerahkan Tugas Akhir kepada Universitas melalui Perpustakaan untuk keperluan akademis dan memberikan **Hak Bebas Royalti Non Eksklusif** (*Non-exclusive Royalty-free Right*) serta bersedia Tugas Akhirnya dipublikasikan secara online dan dapat diakses secara lengkap (*full access*).

Dengan Hak Bebas Royalti Noneklusif ini Perpustakaan Universitas Kristen Duta Wacana berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk *database*, merawat, dan mempublikasikan Tugas Akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta. Demikian pernyataan ini saya buat dengan sebenar-benarnya.

Jakarta, 8 Juni 2023

Yang menyatakan,



(71180323 – Eusebia Fernanda Siahaan)



Karya sederhana ini dipersembahkan

kepada Tuhan, Keluarga,

Ibunda dan Adik Tercinta.



The loneliest moment in someone's life is when they are watching their whole world fall apart, and all they can do is stare blankly.

(F. Scott Fitzgerald)

KATA PENGANTAR

Segala puji dan syukur kepada Tuhan yang maha kasih, karena atas segala rahmat, bimbingan, dan bantuan-Nya maka akhirnya Skripsi dengan judul Implementasi *Named Entity Recognition* dalam Sistem Otomatisasi Pemberian Indeks Dokumen pada Apache SOLR 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 yang Maha Esa,
2. Ibu dan adik dari penulis yang selama ini telah memberi semangat dan mendoakan penulis,
3. Bapak Restyandito, S.Kom, MSIS., Ph.D selaku Dekan FTI, yang mengatur jalannya pendidikan di FTI,
4. Ibu Gloria Virginia, S.Kom., MAI., Ph.D, selaku Kaprodi Informatika sekaligus Dosen Pembimbing 1, yang telah memberikan ilmunya dan dengan penuh kesabaran membimbing penulis,
5. Bapak Budi Susanto, S.Kom.,M.T., selaku Dosen Pembimbing 2 yang telah memberikan ilmu dan kesabaran dalam membimbing penulis,
6. Grup TDOK dan DFB yang telah mendukung dan membantu.
7. Keluarga tercinta yang selalu menerima penulis apa adanya,
8. Lain-lain yang telah mendukung moral, spiritual, dan dana untuk belajar.

Laporan 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.

Jakarta, 8 Juni 2023

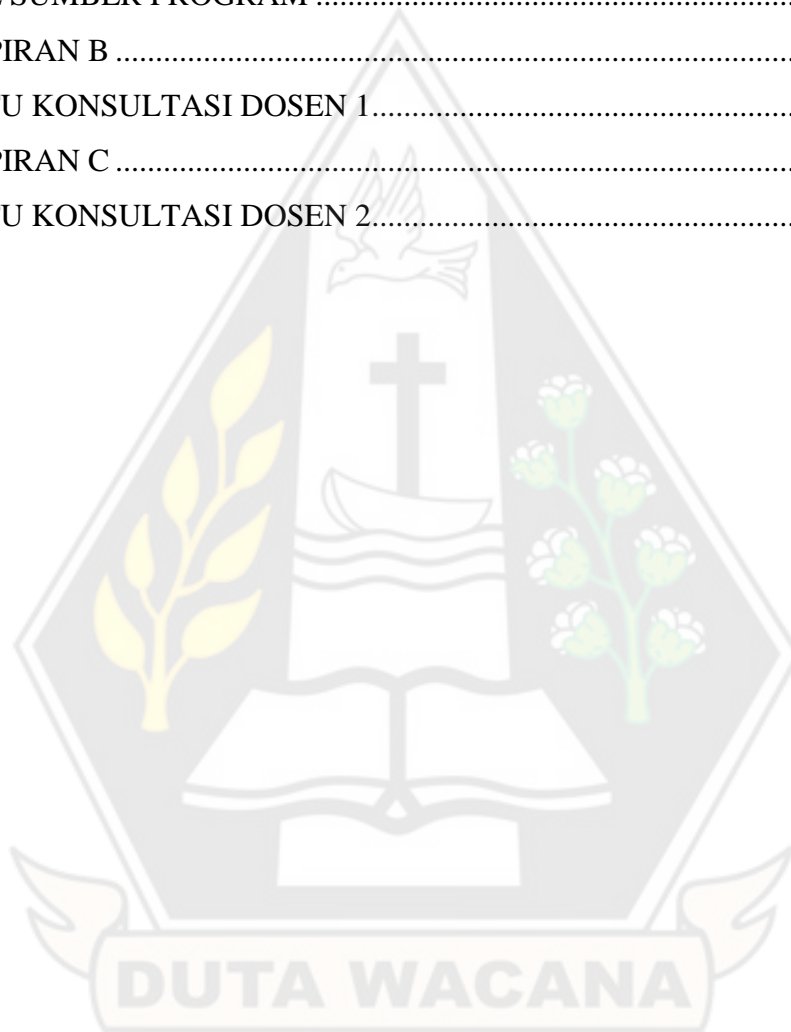
Penulis

DAFTAR ISI

PERNYATAAN KEASLIAN SKRIPSI.....	iii
HALAMAN PERSETUJUAN.....	iv
HALAMAN PENGESAHAN.....	v
HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS SECARA ONLINE.....	vi
KATA PENGANTAR	ix
DAFTAR ISI.....	x
DAFTAR TABEL.....	xiii
DAFTAR GAMBAR	xiv
INTISARI.....	xvi
ABSTRACT.....	xviii
BAB I.....	1
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.....	6
TINJAUAN PUSTAKA DAN DASAR TEORI	6
2.1 Tinjauan Pustaka	6
2.2 Landasan Teori	6
2.2.1 Named Entity Recognition.....	6
2.2.2 Conditional Random Fields & Linear Chain Random Fields	7
2.2.3 Stochastic Gradient Descent	10

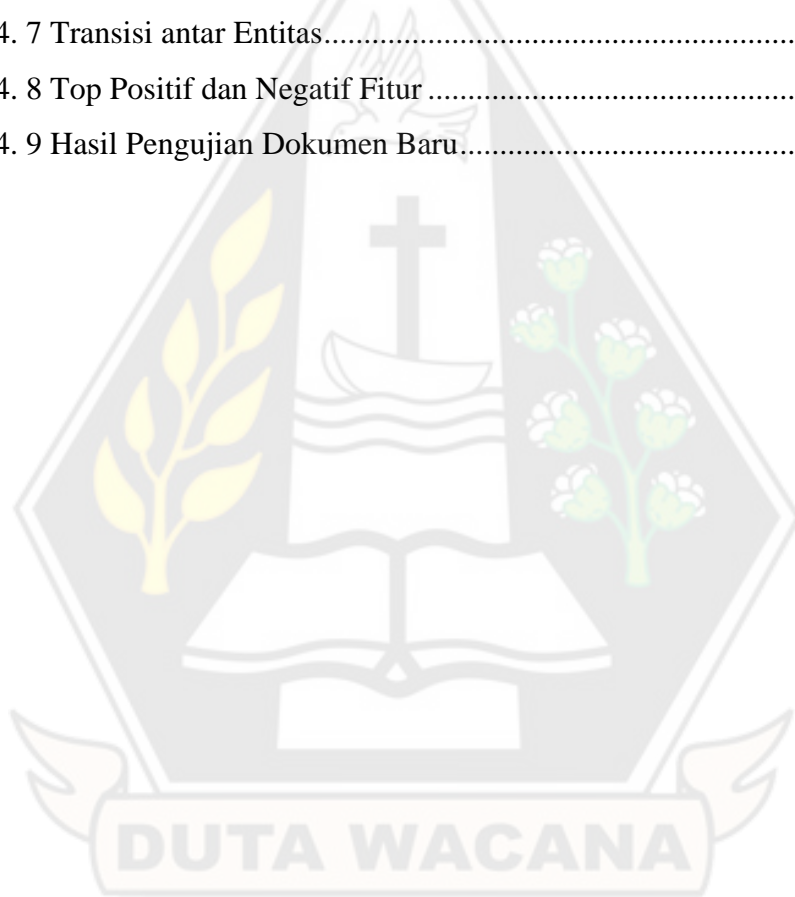
2.2.4	Regulasi L2	11
2.2.5	Precision, Recall & F1-Score	12
2.2.6	Apache SOLR	13
BAB III	14
METODOLOGI PENELITIAN	14
3.1	Analisis Kebutuhan Sistem	14
3.1.1	Kebutuhan Fungsional	14
3.1.2	Kebutuhan Teknis	14
3.2	Perancangan Penelitian.....	14
3.2.1	Pembuatan Data Latih dan Data Uji	15
3.2.2	Pembuatan Model Linear Chain Conditional Random Fields	21
3.2.3	Pembuatan Sistem	21
3.3	Diagram Alir.....	24
3.4	Use Case Diagram	26
3.5	Perancangan Basis Data	27
3.6	Perancangan Antarmuka Pengguna.....	28
3.7	Perancangan Pengujian Sistem.....	30
BAB IV	31
IMPLEMENTASI DAN PEMBAHASAN	31
4.1	Implementasi Awal.....	31
4.1.1	Instalasi Sistem yang Dibutuhkan.....	31
4.1.2	Data Masukan.....	33
4.1.3	Pra-proses	34
4.1.4	Pelatihan dan Pengujian Model.....	38
4.2	Implementasi Sistem	46
4.3	Pengujian dan Analisis	49
4.4	Pembahasan	49
BAB V	51

KESIMPULAN DAN SARAN.....	51
5.1 Kesimpulan.....	51
5.2 Saran.....	51
DAFTAR PUSTAKA	52
LAMPIRAN A.....	53
KODE SUMBER PROGRAM	53
LAMPIRAN B	64
KARTU KONSULTASI DOSEN 1.....	64
LAMPIRAN C	65
KARTU KONSULTASI DOSEN 2.....	65



DAFTAR TABEL

Tabel 4. 1 Hasil POS Tagging	35
Tabel 4. 2 Hasil Pemberian Entitas	36
Tabel 4. 3 Hasil Pengujian Pertama	41
Tabel 4. 4 Tag dan Banyaknya Token Data Pengujian	42
Tabel 4. 5 Hasil Pengujian dengan Data Lama dan Data Baru	42
Tabel 4. 6 Hasil Pengujian Kedua	43
Tabel 4. 7 Transisi antar Entitas	44
Tabel 4. 8 Top Positif dan Negatif Fitur	45
Tabel 4. 9 Hasil Pengujian Dokumen Baru	49



DAFTAR GAMBAR

Gambar 2. 1 Diagram Hidden Markov Model	8
Gambar 2. 2 Diagram <i>Conditional Random Fields</i>	9
Gambar 2. 3 Diagram <i>Linear Chain Conditional Random Fields</i>	9
Gambar 3. 1 Diagram Pembangunan Sistem	15
Gambar 3. 2 Alur Pengumpulan Dokumen.....	15
Gambar 3. 3 Alur Tokenisasi	16
Gambar 3. 4 Alur POS <i>Tagging</i>	18
Gambar 3. 5 Alur Penambahan Entitas	20
Gambar 3. 6 Diagram Sistem	22
Gambar 3. 7 Diagram Alir Penelitian	24
Gambar 3. 8 <i>Use Case Diagram</i>	26
Gambar 3. 9 Basis Data Indeks	27
Gambar 3. 10 Purwarupa Tampilan Utama.....	28
Gambar 3. 11 Tampilan Unggah 1	29
Gambar 3. 12 Tampilan Unggah 2	29
Gambar 4. 1 XAMPP Control Panel.....	32
Gambar 4. 2 Apache SOLR Administrator.....	32
Gambar 4. 3 Hasil Ekstraksi PDF	34
Gambar 4. 4 Data Latih.....	37
Gambar 4. 5 Data Latih CSV	38
Gambar 4. 6 Chart Hasil Pengujian Pertama	40
Gambar 4. 7 Chart Hasil Pengujian Kedua	43
Gambar 4. 8 Transisi Entitas Keseluruhan.....	45
Gambar 4. 9 Tampilan Utama Web 1	46
Gambar 4. 10 Tampilan Utama Web 2	47
Gambar 4. 11 Tampilan Unggah Dokumen 1	47

Gambar 4. 12 Tampilan Unggah Dokumen 2 47
Gambar 4. 13 Tampilan Unggah Dokumen 3 48
Gambar 4. 14 Tampilan Pencarian 48



INTISARI

IMPLEMENTASI *NAMED ENTITY RECOGNITION* DALAM SISTEM OTOMATISASI PEMBERIAN INDEKS DOKUMEN PADA APACHE SOLR

Oleh

EUSEBIA FERNANDA SIAHAAN

71180323

Fakultas Teknologi Informasi (FTI) Universitas Kristen Duta Wacana (UKDW) memiliki suatu repositori untuk menyimpan seluruh dokumen resmi yang telah diterbitkan. Repositori tersebut menggunakan Apache SOLR (Search on Lucene, Rescaled) atau yang biasa disebut SOLR sebagai mesin pencari. SOLR mengalami kesulitan dalam proses pencarian dokumen. Hal ini dikarenakan sebagian besar dokumen yang tersimpan merupakan data tidak terstruktur dan belum memiliki indeks. SOLR akan kesulitan untuk melakukan pencarian berdasarkan isi dari dokumen tersebut, seperti contoh mencari berdasarkan judul atau tanggal pengesahan dokumen dari dokumen tersebut. SOLR sendiri tidak memiliki fitur untuk menentukan indeks secara otomatis dalam dokumen tersebut. Pada kasus ini, *Named Entity Recognition* (NER) dapat membantu pada proses pemberian indeks secara otomatis terhadap dokumen. NER membantu menentukan entitas dari suatu kata atau frasa dimana kata yang sudah mendapat entitas dapat dijadikan sebagai indeks untuk mesin pencari SOLR.

Peneliti menggunakan dokumen resmi yang dikeluarkan dan tersimpan pada repositori FTI UKDW. Isi dari dokumen tersebut akan diberi entitas dengan mengimplementasi NER menggunakan metode *Conditional Random Fields* (CRF). Entitas yang terbentuk akan diubah menjadi indeks dan tersimpan dalam format JSON. Indeks dengan format JSON tersebut akan disimpan pada mesin pencari SOLR yang akan digunakan SOLR dalam proses pencarian kembali.

Penelitian ini, menghasilkan sebuah sistem otomatisasi pemberian indeks dokumen dengan implementasi NER. Sistem tersebut dapat memberikan indeks dokumen dengan mengandalkan entitas yang didapati dari pengujian dengan CRF. Indeks yang didapat dapat digunakan oleh repositori FTI UKDW. Indeks yang telah terbentuk akan dipakai oleh SOLR untuk proses pencarian kembali.

Kata-kata kunci: *Named Entity Recognition, Conditional Random Fields, Apache SOLR.*



ABSTRACT

IMPLEMENTATION OF NAMED ENTITY RECOGNITION IN AUTOMATIC DOCUMENT INDEXING SYSTEM ON APACHE SOLR

By

EUSEBIA FERNANDA SIAHAAN

71180323

The Faculty of Information Technology (FTI) at Duta Wacana Christian University (UKDW) has a repository for storing all official documents that have been published. The repository uses Apache SOLR (Search on Lucene, Rescaled), commonly known as SOLR, as its search engine. However, SOLR encounters difficulties in the document search process. This is because most of the stored documents are unstructured data and do not have any indexes. SOLR struggles to perform searches based on the content of these documents, such as searching by title or date of document approval. SOLR itself lacks the feature to automatically determine indexes within the documents.

In this case, Named Entity Recognition (NER) can help automate the process of providing indexes to the documents. NER assists in identifying entities from words or phrases, where the words that have been assigned entities can be used as indexes for the SOLR search engine.

Researchers utilize the official documents issued and stored in the FTI UKDW repository. The content of these documents will be assigned entities by implementing NER using the Conditional Random Fields (CRF) method. The formed entities will then be transformed into indexes and stored in JSON format. The JSON-formatted indexes will be saved in the SOLR search engine, which will be utilized by SOLR in the subsequent search process.

This research yields an automated system for indexing documents using NER implementation. The system is capable of providing document indexes by relying on the entities obtained through CRF testing. The resulting indexes can be

utilized by the FTI UKDW repository, and SOLR will utilize these indexes for subsequent search processes.

Keywords: *Named Entity Recognition, Conditional Random Fields, Apache SOLR.*



BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Fakultas Fakultas Teknologi Informasi (FTI) Universitas Kristen Duta Wacana (UKDW) memiliki suatu repositori untuk menyimpan seluruh dokumen resmi yang telah diterbitkan mulai dari dokumen kebijakan akademik, dokumen penyelenggaraan akademik sampai dokumen standar operasional. Repositori tersebut menggunakan Apache SOLR (Search on Lucene, Rescaled) atau yang biasa disebut SOLR sebagai mesin pencari dokumen. SOLR mengalami kesulitan dalam proses pencarian dokumen. Hal ini dikarenakan sebagian besar dokumen yang tersimpan belum memiliki indeks. SOLR akan kesulitan untuk melakukan pencarian berdasarkan isi dari dokumen tersebut, seperti contoh mencari berdasarkan judul atau tanggal pengesahan dokumen dari dokumen tersebut. SOLR sendiri tidak memiliki fitur untuk menentukan indeks secara otomatis dalam dokumen tersebut. Setiap dokumen tersebut membutuhkan indeks yang tepat dan membutuhkan waktu cukup lama jika pemberian indeks dilakukan secara manual sehingga diperlukan sistem yang dapat menentukan indeks secara otomatis.

Pada kasus ini, *Named Entity Recognition* (NER) mampu untuk membantu pengembangan sistem otomatisasi pemberian indeks dokumen. NER merupakan langkah awal menuju sebuah ekstraksi informasi untuk mengidentifikasi sebuah entitas yang terdapat dalam sebuah teks ke dalam sebuah kategori yang telah ditentukan sebelumnya. Misalnya nama seseorang, organisasi, lokasi dan sesuatu entitas lain dalam sebuah teks (Rachmad, 2020). Dengan kata lain, NER dapat mengekstrak entitas dari dokumen menjadi indeks yang berguna untuk memudahkan pekerjaan dari SOLR.

Berdasarkan fenomena ini, penulis akan membangun sebuah sistem otomatisasi pemberian indeks dokumen dengan mengimplementasikan NER. Hal pertama yang dilakukan adalah pembuatan model NER menggunakan *Conditional Random Fields* (CRF) menggunakan dokumen-dokumen yang terdapat pada

repositori FTI UKDW. Langkah kedua adalah pembuatan API sebagai sistem otomatisasi pemberian indeks dokumen dengan menggunakan bahasa pemrograman Python. Langkah ketiga adalah pembuatan *website* sederhana sebagai antarmuka untuk pengguna. Hal ini untuk memudahkan pengguna dalam menggunakan sistem dan melakukan pencarian dokumen kembali. Langkah terakhir adalah integrasi API, *website*, dan Apache SOLR menjadi suatu sistem. *Website* yang terintegrasi dengan sistem otomatisasi pemberian indeks dokumen ini diharapkan mampu untuk memenuhi kebutuhan FTI UKDW dalam memberikan indeks dokumen secara otomatis dan dapat melakukan pencarian kembali berdasarkan isi dokumen.

1.2. Perumusan Masalah

Bagaimana implementasi *Named Entity Recognition* (NER) yang tepat untuk pembangunan sistem otomatisasi indeks dokumen pada mesin pencari berbasis Apache SOLR terhadap dokumen Fakultas Teknologi Informasi Universitas Kristen Duta Wacana?

1.3. Batasan Masalah

Agar penelitian dapat fokus terhadap permasalahan yang ada, maka terdapat batasan masalah yang dibuat sebagai berikut:

1. Menggunakan dokumen yang dimiliki oleh Fakultas Teknologi Informasi Universitas Kristen Duta Wacana.
2. Dokumen menggunakan Bahasa Indonesia.
3. Dokumen menggunakan format PDF hasil konversi dari format Word.
4. Menggunakan *library* atau pustaka yang sudah tersedia pada Python.

1.4. Tujuan Penelitian

Tujuan dari penelitian ini adalah menemukan implementasi *Named Entity Recognition* (NER) yang tepat dalam pembangunan sistem otomatisasi pemberian indeks dokumen pada mesin pencari berbasis Apache SOLR terhadap dokumen Fakultas Teknologi Informasi Universitas Kristen Duta Wacana.

1.5. Manfaat Penelitian

Dengan mengimplementasikan *Named Entity Recognition* (NER) secara tepat, repositori Fakultas Teknologi Informasi Universitas Kristen Duta Wacana mampu secara otomatis memberikan indeks dokumen. Diharapkan mesin pencari Apache SOLR dapat secara optimal menemukan dokumen menggunakan indeks.

1.6. Metodologi Penelitian

Untuk membangun sebuah sistem pencarian dokumen pada pencarian repositori dokumen, diperlukan tahapan-tahapan seperti berikut:

1. Membuat data model

Data model dibuat dengan mengumpulkan dokumen-dokumen yang sudah ada pada repositori FTI UKDW. Dokumen yang diambil hanyalah dokumen tidak terstruktur dan memiliki format PDF. Langkah membuat data model sendiri yaitu:

- a. Pengambilan isi dokumen yang kemudian disimpan kedalam format TXT.
- b. Praproses data TXT tersebut dengan cara tokenisasi dan *filtering*. Penulis tidak melakukan *cleaning data* karena huruf kapital, angka dan tanda baca sangat berpengaruh untuk pelabelan.
- c. Token yang sudah didapat akan diberi label secara manual oleh penulis.
- d. Token yang sudah diberi label diubah format dokumen menjadi format CSV.
- e. Token dengan format CSV tersebut kemudian mengalami proses latih menggunakan algoritma *Conditional Random Fields* (CRF).
- f. Keluaran dari proses latih tersebut menghasilkan data model yang akan disimpan dalam format Joblib. Data model tersebut disimpan dan akan dipanggil bila akan melakukan data uji terhadap dokumen baru.

2. Pra-proses dokumen baru.

Dokumen baru yang telah terunggah pada repositori kemudian mengalami pra-proses sama seperti dokumen yang dijadikan data model. Dokumen yang telah terunggah tersebut melalui tahapan tokenisasi dan *filtering*.

3. Pemberian entitas

Hasil dari pra-proses kemudian diuji dengan menggunakan CRF dengan menggunakan data model yang telah tersimpan dalam format Jolib. Hasil dari proses uji tersebut juga disimpan pada data model agar data model dapat lebih kaya lagi akan data uji.

4. Konversi ke indeks dengan format JSON

Dokumen yang telah mendapatkan entitas tadi, kemudian diambil kata dan entitas tersebut untuk dijadikan indeks. Indeks tersebut memiliki format JSON dan akan tersimpan pada SOLR.

1.7. Sistematika Penulisan

Pada penelitian ini, penulis membagi laporan ini menjadi 7 bagian yaitu sebagai berikut:

1. Bab I Pendahuluan

Pendahuluan berisi latar belakang, rumusan masalah, tujuan, manfaat, batasan masalah dan sistematika pembahasan.

2. Bab II Landasan Keputusan

Pada Bagian kedua ini akan memberi penjelasan mengenai landasan kepustakaan yang berisi uraian dan pembahasan tentang teori, konsep, model, metode, atau sistem dari pustaka ilmiah, yang berkaitan dengan *Named Entity Recognition*, *Conditional Random Fields* dan Apache SOLR.

3. Bab III Metode Penelitian

Dalam bagian ini akan membahas metode, teknik atau langkah-langkah yang digunakan dalam penelitian serta logika di balik metode, teknik,

atau langkah-langkah tersebut sesuai dengan konteks penelitian. Akan dijelaskan juga mengapa metode tersebut dipilih.

4. Bab IV Perancangan

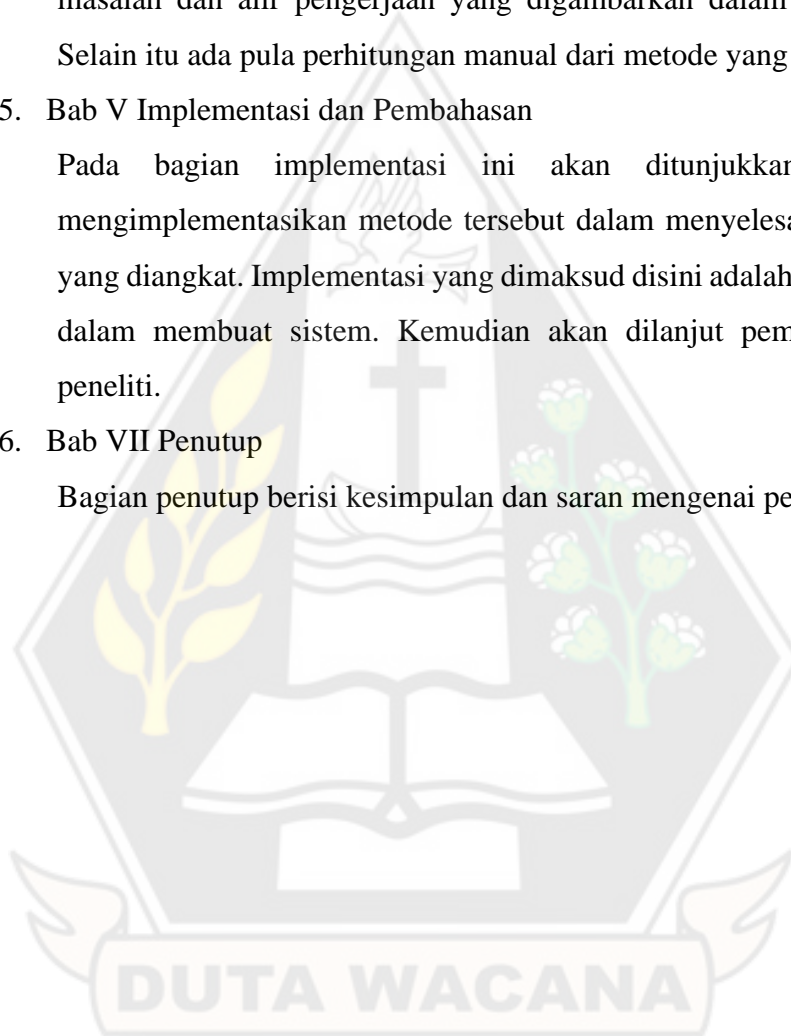
Dalam bagian ini akan menjelaskan mengenai bagaimana menyelesaikan permasalahan ini dengan cara melakukan formulasi masalah dan alir pengerjaan yang digambarkan dalam diagram alir. Selain itu ada pula perhitungan manual dari metode yang dipilih.

5. Bab V Implementasi dan Pembahasan

Pada bagian implementasi ini akan ditunjukkan bagaimana mengimplementasikan metode tersebut dalam menyelesaikan masalah yang diangkat. Implementasi yang dimaksud disini adalah implementasi dalam membuat sistem. Kemudian akan dilanjut pembahasan oleh peneliti.

6. Bab VII Penutup

Bagian penutup berisi kesimpulan dan saran mengenai penelitian ini.



BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Pada penelitian ini, dapat ditarik kesimpulan yaitu implementasi NER dengan CRF berhasil dilakukan meskipun masih jauh dari sempurna. Pada pengujian untuk data baru didapati *F1-Score* sekitar 0,167. Entitas yang dapat dikenali pun sangat terbatas yaitu untuk entitas organisasi (org) seperti kata “UKDW” dan entitas kota (*city*) seperti “Yogyakarta”. Keterbatasan entitas menyebabkan indeks yang diterima Apache SOLR juga terbatas, hal ini menyebabkan pencarian terhadap dokumen juga terbatas.

5.2 Saran

Penelitian ini masih memiliki kekurangan diantaranya dokumen yang dilatih dan diuji masih sangat sedikit. Maka dari itu, saran dari penelitian ini adalah untuk menambahkan dokumen yang akan dilatih dan diuji. Regulasi L2 juga belum mampu untuk mengurangi *overfitting* pada algoritma optimisasi SGD, maka dari itu diperlukan pengujian menggunakan algoritma lain seperti *Limited Memory Broyden Fletcher Goldfarb Shanno (L-BFGS)* yang juga menggunakan regulasi L2 dan juga memakai regulasi L1.

DAFTAR PUSTAKA

- Akbar, A. A., Sabariah, M. K., & Kurniati, A. P. (2015). Peringkasan Multi-dokumen Berita Berbahasa Indonesia menggunakan Conditional Random Fields (CRF). *E-Proceeding of Engineering*, 1586-1597.
- Jaariyah, N., & Rainarli, E. (2017). Conditional Random Fields Untuk Pengenalan Entitas Bernama Pada Teks Bahasa Indonesia. *Komputa : Jurnal Ilmiah Komputer Dan Informatika*, 29-34.
- Patil, N., Patil, A., & Pawar, B. V. (2019). Named Entity Recognition using Conditional Random Fields. *Procedia Computer Science*, 1181-1188.
- Rachmad, D. S. (2020). Review Named Entity Recognition dengan Menggunakan Machine Learning. *Jurnal Sains Dan Informatika*, 28–33.
- Sutton, C., & McCallum, A. (t.thn.). An Introduction to Conditional Random Fields for Relational Learning.
- Wibawa, A. S., & Purwarianti, A. (2016). Indonesian Named-entity Recognition for 15 Classes Using Ensemble Supervised Learning. *Procedia Computer Science*, 221-228.
- Wibisono, Y., & Khodra, M. L. (2018). *Pengenalan Entitas Bernama Otomatis untuk Bahasa Indonesia dengan Pendekatan Pembelajaran Mesin*. Diambil kembali dari <https://doi.org/10.31227/osf.io/vud2p>
- Zhixiang, J., Xiaohui, W., Changyu, C., & Hongjian, S. (2020). Power Entity Recognition Based on Bidirectional Long Short-term Memory and Conditional Random Fields. *Global Energy Interconnection*, 186-192.