**PENGEMBANGAN SISTEM NAMED ENTITY RECOGNITION BAHASA INDONESIA DENGAN PENDEKATAN TRANSFER LEARNING**

**Laporan Tugas Akhir I**

**Disusun sebagai syarat kelulusan mata kuliah**

**IF4091/Tugas Akhir I dan Seminar**

**Oleh**

**<NAMA MAHASISWA>**

**NIM : <NIM MAHASISWA>**



**PROGRAM STUDI TEKNIK INFORMATIKA**

**SEKOLAH TEKNIK ELEKTRO & INFORMATIKA**

**INSTITUT TEKNOLOGI BANDUNG**

**<Bulan> <Tahun>**

**PENGEMBANGAN SISTEM NAMED ENTITY RECOGNITION BAHASA INDONESIA DENGAN PENDEKATAN TRANSFER LEARNING**

**Laporan Tugas Akhir I**

**Oleh**

**JOSHUA ADITYA KOSASIH**

**NIM : 13514012**

**Program Studi Teknik Informatika**

Sekolah Teknik Elektro dan Informatika

Institut Teknologi Bandung

Bandung, 3 Januari 2018

Mengetahui,

Pembimbing,

Dr. Masayu Leylia Khodra, S.T. M.T.

NIP. 19760429 200812 2 001

**<CONTOH JUDUL: PENGEMBANGAN SISTEM INFORMASI DI PUSKESMAS DAERAH PEDESAAN>**

**Laporan Tugas Akhir I**

**Oleh**

**<NAMA MAHASISWA>**

**NIM : <NIM MAHASISWA>**

**Program Studi Teknik Informatika**

Sekolah Teknik Elektro dan Informatika

Institut Teknologi Bandung

Bandung, <tanggal>

Mengetahui,

|  |  |
| --- | --- |
| Pembimbing I,  <Nama Pembimbing-II>  NIP. <NIP Pemb II> | Pembimbing II,  <Nama Pembimbing-II>  NIP. <NIP Pemb II> |

DAFTAR ISI

[DAFTAR ISI iv](#_Toc406869913)

[DAFTAR LAMPIRAN vi](#_Toc406869914)

[DAFTAR GAMBAR vii](#_Toc406869915)

[DAFTAR TABEL viii](#_Toc406869916)

[BAB I PENDAHULUAN 1](#_Toc406869917)

[I.1 Latar Belakang 1](#_Toc406869918)

[I.2 Rumusan Masalah 1](#_Toc406869919)

[I.3 Tujuan 2](#_Toc406869920)

[I.4 Batasan Masalah 2](#_Toc406869921)

[I.5 Metodologi 2](#_Toc406869922)

[I.6 Jadwal Pelaksanaan Tugas Akhir 2](#_Toc406869923)

[BAB II STUDI LITERATUR 3](#_Toc406869924)

[II.1 Contoh Subbab 3](#_Toc406869925)

[II.1.1 Contoh Subbab 3](#_Toc406869926)

[BAB III <RENCANA PENYELESAIAN MASALAH> 5](#_Toc406869927)

[DAFTAR PUSTAKA 6](#_Toc406869928)

DAFTAR LAMPIRAN

[Lampiran A. Contoh Judul Lampiran 7](#_Toc406869964)

[A.1 Contoh Judul Anak Lampiran 7](#_Toc406869965)

DAFTAR GAMBAR

[Gambar II.1. Tahapan konstruksi koleksi retorik kalimat 3](#_Toc406855725)

DAFTAR TABEL

[Tabel II.1. Pengelompokan *Tag* MARC-21 4](#_Toc406855720)

# PENDAHULUAN

Bab Pendahuluan secara umum yang dijadikan landasan kerja dan arah kerja penulis tugas akhir, berfungsi mengantar pembaca untuk membaca laporan tugas akhir secara keseluruhan.

## Latar Belakang

Latar Belakang berisi dasar pemikiran, kebutuhan atau alasan yang menjadi ide dari topik tugas akhir. Tujuan utamanya adalah untuk memberikan informasi secukupnya kepada pembaca agar memahami topik yang akan dibahas. Saat menuliskan bagian ini, posisikan anda sebagai pembaca – apakah anda tertarik untuk terus membaca?

## Rumusan Masalah

Rumusan Masalah, berisi masalah utama yang dibahas dalam tugas akhir. Rumusan masalah yang baik memiliki struktur sebagai berikut:

1. Penjelasan ringkas tentang kondisi/situasi yang ada sekarang terkait dengan topik utama yang dibahas Tugas Akhir.
2. Pokok persoalan dari kondisi/situasi yang ada, dapat dilihat dari kelemahan atau kekurangannya. **Bagian ini merupakan inti dari rumusan masalah**.
3. Elaborasi lebih lanjut yang menekankan pentingnya untuk menyelesaikan pokok persoalan tersebut.
4. Usulan singkat terkait dengan solusi yang ditawarkan untuk menyelesaikan persoalan.

Penting untuk diperhatikan bahwa persoalan yang dideskripsikan pada subbab ini akan dipertanggungjawabkan di akhir pelaksanaan tugas akhir apakah terselesaikan atau tidak.

## Tujuan

Tuliskan tujuan utama dan/atau tujuan detil yang akan dicapai dalam pelaksanaan tugas akhir. Fokuskan pada hasil akhir yang ingin diperoleh setelah tugas akhir diselesaikan, terkait dengan penyelesaian persoalan pada rumusan masalah. Penting untuk diperhatikan bahwa tujuan yang dideskripsikan pada subbab ini akan dipertanggungjawabkan di akhir pelaksanaan tugas akhir apakah tercapai atau tidak.

## Batasan Masalah

Tuliskan batasan-batasan yang diambil dalam pelaksanaan tugas akhir. Batasan ini dapat dihindari (tidak perlu ada) jika topik/judul tugas akhir dibuat cukup spesifik.

## Metodologi

Tuliskan semua tahapan yang akan dilalui selama pelaksanaan tugas akhir. Tahapan ini spesifik untuk menyelesaikan persoalan tugas akhir. Tahapan studi literatur tidak perlu dituliskan karena ini adalah pekerjaan yang harus Anda lakukan selama proses pelaksanaan tugas akhir.

## Jadwal Pelaksanaan Tugas Akhir

Tuliskan rencana kegiatan dan jadwal (dirinci sampai per minggu) mulai dari awal pelaksanaan Tugas Akhir I s.d. sidang tugas akhir berikut *milestones* dan *deliverables* yang harus diberikan. Jadwal ini dapat dibantu dengan membuat sebuah tabel *timeline*.

# STUDI LITERATUR

Bab Studi Literatur digunakan untuk mendeskripsikan kajian literatur yang terkait dengan persoalan tugas akhir. Tujuan studi literatur adalah:

1. menunjukkan kepada pembaca adanya *gap* seperti pada rumusan masalah yang memang belum terselesaikan,
2. memberikan pemahaman yang secukupnya kepada pembaca tentang teori atau pekerjaan terkait yang terkait langsung dengan penyelesaian persoalan, serta
3. menyampaikan informasi apa saja yang sudah ditulis/dilaporkan oleh pihak lain (peneliti/Tugas Akhir/Tesis) tentang hasil penelitian/pekerjaan mereka yang sama/mirip/erat kaitannya dengan persoalan tugas akhir.

## Contoh Subbab

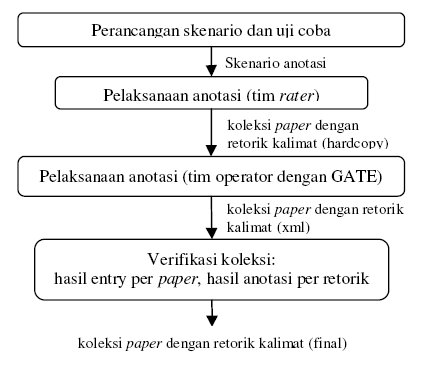
Contoh subbab

### Contoh Subbab

Contoh subbab.

#### Contoh Gambar dan Tabel

Contoh gambar adalah sebagaimana terlihat pada Gambar I.1.



Gambar II.1. Tahapan konstruksi koleksi retorik kalimat

Contoh tabel adalah sesuai yang terlihat pada Tabel I.1.

Tabel II.1. Pengelompokan *Tag* MARC-21

| Nomor *Tag* | Keterangan |
| --- | --- |
| 0xx | Kendali informasi, nomor dan kode |
| 1xx | Main entry |
| 2xx | Judul, edisi, cetakan |
| 3xx | Deskripsi fisik |
| 4xx | Pernyataan seri |
| 5xx | Catatan |
| 6xx | Subject added entries |
| 7xx | Added entries selain dari subject added entries |
| 8xx | Series added entry |
| \9xx | Definisi lokal |

# <RENCANA PENYELESAIAN MASALAH>

## Analisis persoalan

Tujuan dari tugas akhir ini adalah untuk membuat model NER bahasa Indonesia yang baik menggunakan pendekatan transfer learning. Model NER yang baik perlu untuk dapat mengenali dan mengklasifikasikan setiap kemunculan named-entity dalam teks dengan tepat. Berbagai penelitian telah dilakukan untuk meningkatkan kinerja dari model NER bahasa Indonesia. Namun seperti telah dibahas pada bab 2, hasil kinerja dari model NER bahasa Indonesia (berdasarkan skor f-1) masih kurang memuaskan, terutama jika dibandingkan dengan kinerja NER bahasa Inggris yang banyak berkembang.

### Persoalan Named Entity Recognition bahasa Indonesia

Beberapa penelitian NER bahasa Indonesia berfokus pada cara untuk mengatasi sedikitnya data latih. Terdapat penelitian yang memakai data online dari Wikipedia bahasa Indonesia (Luthfi, Distiawan, & Manurung, 2014). Namun salah satu permasalahan dari penelitian tersebut adalah kurang baiknya data berlabel yang dipakai, dimana pelabelan yang dilakukan secara otomatis pada artikel-artikel Wikipedia tersebut tidak tepat. Terdapat pula penelitian yang menggunakan metode semi-supervised learning (Leonandya, Distiawan, & Praptono, 2015) (Aryoyudanta, Adji, & Hidayah, 2016). Metode ini cukup baik dilakukan untuk mengatasi data latih yang sedikit, namun kesalahan pelabelan dapat terbawa ke iterasi-iterasi selanjutnya sehingga diperlukan mekanisme koreksi pelabelan data.

|  |  |  |
| --- | --- | --- |
| Penelitian | Jumlah NE/jenis pada data | Skor f-1 |
|  |  |  |
|  |  |  |
|  |  |  |

^ Apakah perlu dibuat tabel perbandingan?

### Analisa pendekatan transfer learning

Selain cara diatas, terdapat cara lain untuk mengatasi sedikitnya data latih, yaitu menggunakan pendekatan transfer learning (Wang & Zheng, 2015). Namun, saat ini belum pernah dilakukan penelitian menggunakan pendekatan transfer learning untuk membangun model NER bahasa Indonesia ataupun model-model task NLP lainnya. Dalam melakukan transfer learning, terdapat tiga buah persoalan yang perlu dijawab, what to transfer, how to transfer, dan when to transfer.

Sebelum menjawab what to transfer, perlu diperhitungkan setting transfer manakah yang tepat untuk membuat model NER bahasa Indonesia pada Tugas Akhir ini. Telah dibahas sebelumnya pada bab 2, terdapat tiga buah setting transfer yang dapat dikategorikan berdasarkan jenis perbedaan data dan ketersediaan data berlabel. Data sumber yang dipakai adalah korpus POS tag bahasa Indonesia dan data target yang dipakai adalah korpus NER tag bahasa Indonesia. Detil dari data-data tersebut akan dibahas pada subbab berikutnya. Kedua data tersebut memiliki dimensi fitur serta distribusi probabilitas yang serupa. Perbedaan kedua data tersebut terletak pada dimensi label , dimana data POS tag diberi label part-of-speech dan data NER tag diberi label named-entity. Data NER tag bahasa Indonesia yang berlabel juga tersedia, walaupun dengan jumlah yang sedikit. Dengan demikian, pada Tugas Akhir ini akan dipakai setting transfer induktif.

Setelah mengetahui setting transfer yang akan dipakai, induktif, berikutnya adalah menentukan bagian pengetahuan mana yang akan ditransfer. Seperti telah dijelaskan pada bab 2, terdapat empat pendekatan transfer learning berdasarkan bagian pengetahuan yang akan ditransfer. Berdasarkan penelitian oleh Yang, bagian yang cocok untuk ditransfer antar aplikasi/task adalah representasi fitur (Yang, Salakhutdinov, & Cohen, 2017). Bagian pengetahuan yang lain seperti parameter dan relational-knowledge kurang sesuai untuk ditransfer.

Selanjutnya adalah how to transfer. Untuk dapat melakukan feature-representation transfer, diperlukan model yang dapat menangkap representasi fitur dari data. Untuk bentuk model yang sesuai pada jenis persoalan ini juga sudah dieksplorasi (Yang, Salakhutdinov, & Cohen, 2017). Pada model tersebut digunakan model recurent neural network untuk mengekstraksi fitur dari data masukan menjadi representasi fitur yang akhirnya menjadi input bagi model klasifikasi CRF diatasnya. Karena task sumber dan target berbeda, model klasifikasi CRF pada sumber tidak ikut ditransfer untuk dipakai pada klasifikasi data target.

Mengenai when to transfer, penelitian Yang telah menghasilkan kesimpulan bahwa transfer antar task (POS tag ke NER tag) dapat meningkatkan kinerja (Yang, Salakhutdinov, & Cohen, 2017).

### Analisa kebutuhan data

Dengan pendekatan transfer learning, model dapat memanfaatkan data POS tag sebagai data sumber sebelum dilatih menggunakan data NER tag sebagai data target. Ketersediaan data POS tag bahasa Indonesia dan data NER tag bahasa Indonesia menjadi vital dalam tugas akhir ini.

Data POS tag bahasa Indonesia diambil dari Universal Dependencies (UD) (Nivre, et al., 2016). Pada UD versi 2, korpus POS bahasa Indonesia memiliki 121923 token yang telah diberi label POS secara manual.

Bentuk

Data NER tag bahasa Indonesia diambil dari github milik

Analisis data POS & NER

Contoh input, contoh output

Referensi arsitektur lain

## Rancangan solusi

Gambaran sistem

Jelasin per step

Contoh2 input output per step

Tujuan utama penulisan bab ini adalah untuk menguraikan rencana penyelesaian masalah tugas akhir yang akan dieksekusi secara utuh pada saat pelaksanaan Tugas Akhir II. Bab ini merupakan bab penutup Laporan Tugas Akhir I yang dapat dipandang sebagai bab yang akan menjembatani perpindahan ke proses pelaksanaan Tugas Akhir II. Pengembangan lebih lanjut dari bab ini dapat menjadi bagian dari bab Deskripsi Solusi pada Laporan Tugas Akhir.

DAFTAR PUSTAKA

Balabanovic, M. (1998). *Learning to surf: Multi-agent systems for adaptive web page recommendation*. Doctoral dissertation, Stanford University, Menlo Park, CA: Department of Computer Science.

McKusick, K.B., & Langley, P. (1991). Constraints on tree structure in concept formation. Prosiding *The 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval*, 206-214. New York, NY:ACM Press.

Mitchell, T.M. (1997). *Machine Learning*. New York, NY: McGraw-Hill.

Pazzani, M., & Billsus, D. (1997). Learning and revising user profiles: The identification of interesting web sites. *Machine Learning*, *27*, 313-331.

1. Contoh Judul Lampiran
   1. Contoh Judul Anak Lampiran

Contoh anak lampiran