MENENTUKAN ENTITAS DARI PEMBERITAAN MEDIA DARING MENGGUNAKAN HIDDEN MARKOV MODEL UNTUK METODE POS TAGGING

SKRIPSI

Disusun Oleh:

Husain Abdul Aziz

14.11.7723

PROGRAM SARJANA
PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA

2018

MENENTUKAN ENTITAS DARI PEMBERITAAN MEDIA DARING MENGGUNAKAN HIDDEN MARKOV MODEL UNTUK METODE POS TAGGING

SKRIPSI

untuk memenuhi sebagian persyaratan
mencapai gelar Sarjana
pada Program Studi Informatika

Disusun Oleh:

Husain Abdul Aziz

14.11.7723

PROGRAM SARJANA
PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA

2018

PERSETUJUAN

SKRIPSI

MENENTUKAN ENTITAS DARI PEMBERITAAN MEDIA DARING MENGGUNAKAN HIDDEN MARKOV MODEL DAN METODE POS TAGGING

yang dipersiapkan dan disusun oleh

Husain Abdul Aziz

14.11.7723

telah disetujui oleh Dosen Pembimbing Skripsi

pada tanggal 20 Juli 2017

Dosen Pembimbing,

Hartatik, S.T., M.Cs.

NIK. 190302232

PENGESAHAN

SKRIPSI

MENENTUKAN ENTITAS DARI PEMBERITAAN MEDIA DARING MENGGUNAKAN HIDDEN MARKOV MODEL UNTUK METODE POS TAGGING

yang dipersiapkan dan disusun oleh

Husain Abdul Aziz

14.11.7723

telah dipertahankan di depan Dewan Penguji
pada tanggal 23 Agustus 2018

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
Mardhiya Hayaty, S.T., M.Kom **NIK. 190302108**	
Erni Seniwati, S.Kom, M.Cs **NIK. 190302231**	
Yuli Astuti, M.Kom **NIK. 190302146**	

Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 10 September 2018

DEKAN FAKULTAS ILMU KOMPUTER,

Krisnawati, S.Si, M.T.
NIK. 190302038

PERNYATAAN

Saya yang bertandatangan dibawah ini menyatakan bahwa, skripsi ini merupakan karya saya sendiri (ASLI), dan isi dalam skripsi ini tidak terdapat karya yang pernah diajukan oleh orang lain untuk memperoleh gelar akademis di suatu institusi pendidikan tinggi manapun, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis dan/atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Segala sesuatu yang terkait dengan naskah dan karya yang telah dibuat adalah menjadi tanggungjawab saya pribadi.

Yogyakarta, 5 September 2018

Husain Abdul Aziz

NIM. 14.11.7723

MOTTO

”Man Jadda Wa Jada”

Anon

“Kerahkan hati, pikiran, dan jiwamu kedalam aksimu yang paling kecil sekalipun. Inilah rahasia kesuksesaan.”

Swarmi Sivanda

“Betapa bodohnya manusia, dia menghancurkan masa depan kini sambil mengkawatirkan masa depan Tapi menangis dimasa depan dengan mengingat masa lalunya.”

Ali bin Abi Thalib

PERSEMBAHAN

Alhamdulilah segala puji syukur atas berkat rahmat dan karunia Allah SWT yang telah memberikan kemudahan dan kelancara bagi penulis dalam menyelesaikan skripsi ini dengan sebaik-baiknya. Skripsi ini penulis persembahkan untuk:

1. Kedua orang tua penulis yang tercinta, Bapak Mustofa dan Ibu Sunarni yang telah memberikan dukungan terbesar, menguatkan, dan menyemangati penulis dalam suka maupun duka.
2. Adik-adik penulis, Lina Azizah Fathin dan Azizah Qoidatu Fariha yang selalu memberikan dukungan pada penulis.
3. Ibu Hartatik selaku dosen pembimbing yang telah memberikan bimbinganya untuk penulisan skripsi ini.
4. Seluruh teman-teman kelas 14-S1.TI-02, terimakasih untuk waktu, kebersamaan, susah, senang, canda, tawa dan dukungan yang diberikan kepada penulis.
5. Bapak Kun Budiharta dan Bapak Rahadian P. Paramita, yang telah membimbing penulis dalam melakukan penelitian ini.
6. Untuk Wulan, Aldo, Samuel, Mas Marji, Eka, Sigit, Nuzul, Ariyo, Joko, Palupy, Arief, serta masih banyak lagi yang belum disebutkan. Terima kasih atas semangat, dukungan dan doa yang diberikan kepada penulis.

KATA PENGANTAR

Puji syukur atas kehadirat Allah SWT yang telah melimpahkan rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi dengan judul "Menentukan Entitas dari Pemberitaan Media Daring Menggunakan Hidden Markov Model dan Metode POS Tagging" dengan sebaik-baiknya. Tidak lupa shalawat serta salam penulis haturkan kepada junjungan besar Nabi Muhammad SAW.

Dengan selesainya skripsi ini, maka penulis mengucapkan terima kasih kepada:

1. Allah SWT yang telah memberikan petunjuk dan membantu seluruh proses dalam penulisan skripsi ini.
2. Bapak M. Suyanto, Prof., Dr., M.M selaku Ketua Universitas Amikom Yogyakarta.
3. Ibu Hartatik, S.T, M.Cs selaku dosen pembimbing yang telah membantu dan memberi arahan dengan sabar kepada penulis sehingga skripsi ini dapat diselesaikan dengan sebaik-baiknya.
4. Ibu, Bapak, Adik-adik penulis yang selalu setia memberikan semangat, dan doa kepada penulis, sehingga skripsi ini dapat selesai seperti yang diharapkan.
5. Bapak Kun Budiharta yang telah memberikan bimbingan dan dukungan dalam penyusunan skripsi ini.
6. Bapak Rahadian P. Paramita telah memberikan bimbingan dalam penyusunan skripsi ini.
7. Wahyu Setiawan selaku alumni AMIKOM Yogyakarta yang membantu dalam penelitian skripsi ini.

8. Khotimah Tri Wulandari yang selalu sabar dalam membantu, memotivasi dan memberikan dukungan dalam pembuatan skripsi ini.

9. Febyola Aldo Brilyansyah yang telah membantu dan memberikan dukungan kepada penulis.

10. Para Dosen dan Staff Universitas Amikom Yogyakarta yang telah membantu memberikan ilmu pengetahuan, pengalaman, dan motivasi selama proses perkuliahan.

11. Seluruh teman-teman kelas 14-S1.TI-02 dan semua pihak yang telah membantu dan tidak dapat disebutkan satu persatu.

Penulis menyadari bahwa dalam penyusunan skripsi ini masih banyak kekurangan. Besar harapan penulis untuk kritik, saran, bimbingan dan arahan menuju perbaikan dalam skripsi ini. Dan semoga skripsi ini dapat bermanfaat bagi penulis maupun pembaca.

Yogyakarta, 5 September 2018

Penulis

Husain Abdul Aziz

NIM. 14.11.7723

DAFTAR ISI

PERSETUJUAN ii
PENGESAHAN iii
PERNYATAAN iv
MOTTO v
PERSEMBAHAN vi
KATA PENGANTAR vii
DAFTAR ISI ix
DAFTAR TABEL i
DAFTAR GAMBAR ii
INTISARI iv
ABSTRACT v
BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Rumusan Penelitian 2
1.3 Batasan Penelitian 3
1.4 Tujuan Penelitian 3
1.5 Manfaat Penelitian 3
1.6 Metodologi Penelitian 4
1.6.1 Tahapan Pengumpulan Data 4
1.6.2 Tahapan Penelitian 4
1.7 Sistematika Penulisan 5
BAB II LANDASAN TEORI 7
2.1 Tinjauan Pustaka 7

2.2 Text Mining 11
2.3 Scraping 11
2.4 Pre-processing 12
2.5 N-Gram 13
2.6 POS Tag 13
2.7 Hidden Markov Model (HMM) 15
2.8 Named Entity Recognition (NER) 20
2.9 Confussion Matrix 21
2.10 Flowchart 23
BAB III PERANCANGAN 25
3.1 Analisis Masalah 25
3.2 Solusi yang Dipilih 26
3.3 Analisis Kebutuhan 26
3.3.1 Analisis Kebutuhan Fungsional 26
3.3.2 Analisis Kebutuhan Non-Fungsional 27
3.4 Hidden Markov Model 28
3.4.1 Scraping & Cleaning Data 28
3.4.2 Tokenization 29
3.4.3 Anotasi 32
3.4.4 HMM POS Tagger 34
3.4.5 Entity Extraction 36
3.5 Named Entity Recognition 38
3.5.1 Preparing Data 38
3.5.2 Anotasi 40
3.5.3 Training and Testing 43

3.6 Perancangan Aplikasi 43
3.7 Perancangan Interface 46
3.7.1 Perancangan Halaman Utama 46
3.7.2 Perancangan Halaman Kategori 46
3.7.3 Perancangan Halaman Hasil NER 47
3.7.4 Perancangan Halaman Statistik 48
3.7.5 Perancangan Halaman Today Entity 48
3.7.6 Perancangan Halaman Repository 49
BAB IV IMPLEMENTASI 50
4.1 Implementasi Sistem 50
4.2 Pembuatan Database 50
4.1 Implementasi Pengumpulan Data Google News 51
4.2 Implementasi Perancangan Fungsi dan Sistem 53
4.2.1 Perancangan Fungsi dan Sistem HMM Post Tagger 53
4.2.2 Perancangan Fungsi dan Sistem Named Entity Recognition 58
4.5 Pembuatan Interface 62
4.6 Evaluasi Program 66
BAB V KESIMPULAN 70
5.1 Kesimpulan 70
5.2 Saran 70
DAFTAR PUSTAKA 72

DAFTAR TABEL

Tabel 2.1 Tabel Perbandingan Penelitian 9
Tabel 2.2 Tabel *Tagset* Bahasa Indonesia 14
Tabel 2.3 Tabel *Confusion Matrix* 21
Tabel 2.4 Tabel Penjelasan *Flowchart* 23
Tabel 3.1 Tabel Proses *Sentence Tokenization* 30
Tabel 3.2 Tabel *Word Tokenization* 31
Tabel 3.3 Tabel Hasil Anotasi 32
Tabel 3.4 Ilustrasi *Matrix Graf* Kata 35
Tabel 3.5 Tabel Proses *Entity Extraction* 38
Tabel 3.6 Tabel Daftar Label Entitas 40
Tabel 3.7 Tabel Proses Anotasi Entitas 41
Tabel 4.1 Tabel Percobaan 5.000 Kalimat 67
Tabel 4.2 Tabel Percobaan 10.000 Kalimat 67
Tabel 4.3 Tabel Percobaan 15.000 Kalimat 68
Tabel 4.4 Tabel Percobaan 20.000 Kalimat 68
Tabel 4.5 Tabel Hasil Percobaan Terbaik 69

DAFTAR GAMBAR

Gambar 2.1 Ilustrasi Contoh Kasus *Markov Chain* 16
Gambar 3.1 *Flowchart* Proses *Scraping* Berita..... 28
Gambar 3.2 Flowchart Proses *Cleaning* Berita..... 28
Gambar 3.3 *Flowchart* Proses *Scrapping & Cleaning* Data 29
Gambar 3.4 *Flowchart* Proses *Tokenization* 30
Gambar 3.5 *Flowchart* Proses Anotasi POS *Tag*..... 32
Gambar 3.6 Gambar Kemungkinan Perpindahan Antar Label Kata..... 35
Gambar 3.7 Lintasan Untuk Label Kata Terbaik 36
Gambar 3.8 *Flowchart* Proses *Entity Extraction* 37
Gambar 3.9 *Flowchart* Proses *Formating* Data *Prodigy* 39
Gambar 3.10 *Flowchart* Proses Anotasi NER 40
Gambar 3.11 *Flowchart* Proses Keseluruhan Sistem..... 43
Gambar 3.12 *Flowchart* Proses Keseluruhan Sistem..... 45
Gambar 3.13 Rancangan Tampilan Halaman Utama..... 46
Gambar 3.14 Rancangan Tampilan Halaman Kategori 47
Gambar 3.15 Rancangan Tampilan Halaman Hasil NER..... 47
Gambar 3.16 Rancangan Tampilan Halaman Statistik 48
Gambar 3.17 Rancangan Tampilan Halaman Entitas 49
Gambar 3.18 Rancangan Halaman Repository 49
Gambar 4.1 *Query* Pembuatan *Database*..... 50
Gambar 4.2 *Que*ry Pembuatan Tabel *"template"*..... 51
Gambar 4.3 *Query* Pembuatan Tabel Kategori..... 51
Gambar 4.*4 Script* untuk *Scraping Content*..... 52
Gambar 4.5 *Script* untuk Mengambil *Dataset* 54
Gambar 4.6 Isi *file corpus.txt* Sebelum Anotasi 54
Gambar 4.7 Isi *file corpus.tx*t Setelah Anotasi..... 55
Gambar 4.8 *Script* untuk Menyiapkan *Dataset* HMM..... 55
Gambar 4.9 *Script* untuk *Training* model HMM 56
Gambar 4.10 Hasil *Testing* model HMM..... 56

Gambar 4.11 *Script* untuk Menyimpan Model HMM 56
Gambar 4.12 *Script* untuk Membuka Model HMM 57
Gambar 4.13 *Script* untuk *Entity Extraction* 57
Gambar 4.14 *Script* untuk Menyiapkan *Dataset* NER 58
Gambar 4.15 Isi *file gold_corpus.txt* dengan format *Prodigy* 58
Gambar 4.16 *Command* untuk Menjalankan *Prodigy* 59
Gambar 4.17 Tampilan *Prodigy* untuk proses Anotasi 59
Gambar 4.18 *Command* untuk Menyimpan *Gold Corpus* 59
Gambar 4.19 Isi dari *file* gold_corpus.json 60
Gambar 4.20 *Script* untuk *Training* NER 61
Gambar 4.21 *Script* untuk *Testing* model NER 62
Gambar 4.22 Hasil NER 62
Gambar 4.23 Tampilan Halaman Utama 63
Gambar 4.24 Tampilan Halaman Kategori 63
Gambar 4.25 Tampilan Halaman hasil NER 64
Gambar 4.26 Tampilan Halaman Statistik 65
Gambar 4.27 Tampilan Halaman *Today Entity* 65
Gambar 4.28 Tampilan Halaman *Repository* 66

INTISARI

Seiring dengan berkembangnya zaman, tidak dapat dipungkiri bahwa kebutuhan manusia akan suatu informasi yang cepat dan akurat semakin meningkat. Dengan adanya teknologi internet, informasi yang dibutuhkan dapat dengan sangat cepat sampai ke para pembaca. Persebaran berita pun semakin luas jangkauanya, mulai dari sumber berita yang dikota maupun desa yang bisa dengan mudah untuk diakses. Pembaca berita pun bervariasi dengan latar belakang yang berbeda-beda, sehingga akan terjadi perbedaan dalam menerima suatu informasi dari berita yang sedang terjadi.

Metode *POS Tagging* (*Part-of-Speech Tagging*), merupakan sebuah metode untuk memberikan label kelas kata pada suatu kata sehingga akan diketahui keterangan dari masing-masing kata. *Hidden Markov Model* (HMM), adalah suatu model statistik yang terdiri dari dua bagian state yang saling terkait. Bagian yang dapat diamati adalah *observed state*, sedangkan bagian yang tersembunyi disebut *hidden state*. Pada metode *POS Tagging*, *observed state* adalah urutan kata sedangkan *hidden state* adalah urutan tag atau label.

Tujuan dari perancangan aplikasi berbasis *web* ini adalah untuk mendapatkan berita dari pemberitaan media daring dan memberikan label untuk setiap kata pada beritanya, sehingga bisa diketahui entitas yang dibicarakan dalam masing-masing berita. Penulis bertujuan untuk dapat membantu pengguna mengetahui siapa dan apa saja yang dibicarakan dalam suatu berita, tanpa harus membaca keseluruhan dari isi beritanya.

Kata Kunci: *Hidden Markov Model, POS Tagging, Named Entity Recognition*

ABSTRACT

Along with the development of the times, it cannot be denied that human needs for fast and accurate information are increasing. With the advent of internet technology, the information needed can very quickly reach the readers. The spread of news is even wider, ranging from news sources in the city and villages that can be easily accessed. News readers also vary with different backgrounds, so there will be differences in receiving information from the news that is happening.

POS Tagging method (Part-of-Speech Tagging), is a method to labeling the part of speech on a word so that it will knowed the description of each word. Hidden Markov Model (HMM), is a statistical model consisting of two interrelated state parts. The observable part is the observed state, while the hidden part is called the hidden state. In POS Tagging method, observed state is word order while hidden state is sequence of tag.

The purpose of this web application design is to get the news from the online news media and to labeling every word in the news so that any entity can be found in each news. The author aims to be able to help users know who and what is discussed in a news, without having to read the entire content of the news.

Keyword: *Hidden Markov Model, POS Tagging, Named Entity Recognition*