BAB I
PENDAHULUAN

Bab ini berisi tentang latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, metode penelitian, dan sistematika penulisan.

1.1 Latar Belakang Masalah

Al Quran adalah firman Allah yang diturunkan kepada Nabi Muhammad SAW. Al Quran diturunkan dalam bahasa Arab. Bahasa Arab merupakan salah satu bahasa tertua di dunia. Bahasa Arab merupakan bahasa yang lengkap dan sempurna bila dibandingkan dengan bahasa-bahasa yang lain. Kesempurnaan dan kelengkapannya itulah merupakan keistimewaan baginya. Banyak orang menganggap bahasa Arab itu rumit, komplek, sukar dan lain sebagainya, terutama di kalangan pelajar dan mahasiswa [1]. Dalam pengucapan saja terkadang masih memiliki kesalahan yang sangat fatal. Apalagi bahasa Arab yang digunakan dalam suatu program menggunakan bahasa mesin atau komputer.

Part Of Speech (POS) *tagging* adalah suatu proses yang memberikan label kelas kata secara otomatis yang berupa kata kerja (*verb*), kata benda (*noun*), kata sifat (*adjectives*), kata keterangan (*adverb*) dan lain sebagainya pada tiap kata dalam suatu kalimat. Dalam bahasa Arab, ada tiga kategori POS utama, yaitu kata benda, kata kerja dan partikel [2]. POS *tagging* (pelabelan kelas kata) merupakan salah satu bagian yang sangat penting dalam bidang *Natural Language Processing* (NLP) seperti *summarization text*, *Speech Recognition* (SR), *Question Answering* (QA) dan *Informarion Retrieval* (IR). Melakukan palabelan POS secara manual membutuhkan waktu yang lama dan biaya yang mahal karena memerlukan ahli bahasa. Oleh karena itu mengembangkan POS *tagging* secara otomatis merupakan kebutuhan yang mendesak [3].

Masalah utama dalam POS *tagging* antara lain kata ambigu dan kata *Out Of Vocabulary* (OOV) [4]. Kata ambigu merupakan kata yang memiliki sifat berbeda jika ditempakan pada konteks yang berbeda. Sedangkan kata OOV merupakan kata yang ada dalam *corpus* uji namun tidak ada dalam *corpus* pelatihan, hal ini akan menyebabkan masalah *sparse* data.

Masalah ambiguitas dalam bahasa bisa diselesaikan dengan adanya POS *tagging corpus*. *Corpus* merupakan sumber daya yang mendasar berupa sejumlah besar sampel teks bahasa alami terstruktur untuk komputasi linguistik. POS *tagging corpus* merupakan *dataset* berisi informasi tentang kategori sintaksis kata (tata bahasa) yang membantu peneliti tanpa perlu keterampilan linguistik yang kuat. *Corpus* untuk POS *tagging* bahasa Arab dianggap relatif miskin dibandingkan dengan bahasa lain dikarenakan kompleksitas bahasa dan biaya pengembangan sumber daya, dan beberapa *corpus* dikembangkan untuk penggunaan pribadi bukan untuk umum. Kurangnya *corpus* untuk POS *tagging* bahasa Arab merupakan masalah penting dalam mengembangkan POS *tagging* bahasa Arab. Beberapa *corpus* untuk POS *tagging* bahas Arab yaitu, *Quranic corpus tagset* yang terdiri dari 6.236 kalimat (ayat) dengan total token 77.430 kata terdiri dari 44 *tag*, KALIMAT *multipurpose Arabic corpus* yang terdiri dari 18.167.183 kata dan 33 *tag*, Khoja *corpus* terdiri dari 50.000 kata, Abumalloh *corpus* terdiri dari 25.000 kata, dan NEMLAR *corpus* terdiri dari 500.000 kata [17]. Karena sifatnya yang tidak konvensional, POS *tagging* bahasa Arab bukanlah tugas yang mudah, dan kurangnya literatur dalam POS *tagging* bahasa Arab ini adalah masalah penting dalam mengembangkan POS *taggers* Arab. Bahasa Arab juga memiliki tingkat ambiguitas yang tinggi karena berbagai alasan, seperti penghilangan huruf vokal dan kesamaan huruf tetap dengan huruf induk atau akar. Analisis morfologis biasanya mempengaruhi tingkat analisis lain yang lebih tinggi seperti analisis sintaksis dan semantik [5].

Seiring dengan perkembangan zaman, banyak metode-metode baru yang muncul untuk menyelesaikan masalah POS *tagging*. Diantara metode baru yang ada, yaitu metode POS *tagging* berbasis aturan, metode POS *tagging* berbasis probabilistik, metode POS *tagging* evolusioner, dan metode POS *tagging* hibrida yaitu gabungan dari dua metode atau lebih. Kelemahan utama dari pendekatan POS *tagging* sistem berbasis aturan adalah pekerjaan yang melelahkan dalam mengkode aturan secara manual, membutuhkan latar belakang linguistik, dan sistem ini tidak kuat karena harus dirancang ulang sebagian atau seluruhnya ketika terjadi perubahan pada domain atau dalam bahasa. Sedangkan, kekurangan dari

pendekatan POS *tagging* probabilistik yaitu terjadinya *Out Of Vocabulary* (OOV) yang disebabkan nilai probabilitas transisi nol [19] [20].

POS *tagging* telah secara luas dipelajari dan dikembangkan untuk bahasa Arab. Selama beberapa tahun terakhir beberapa upaya telah dilakukan pada Arab POS *tagging* menggunakan pendekatan yang berbeda. Beberapa jurnal dan proyek penelitian diusulkan untuk mengembangkan POS *tagging* bahasa Arab misalnya Al-Taani dan Al-Rub [8] mengusulkan pendekatan berbasis aturan untuk penandaan teks Arab non-disuarakan, Ali dan Jarray [9] melakukan pendekatan evolusioner untuk *tag* teks Arab berdasarkan algoritma genetika pertama kali, Hadni et al. [10] mengusulkan pendekatan *hybrid* menggabungkan penelitian Al-Taani dan Al-Rub [8] dengan pendekatan *Hidden Markov Model* (HMM), Mahafdah et al. [11] mengusulkan penggunaan K-*Nearest Neighbor* (KNN) dan pengklasifikasi *Naive Bayes* (NB) untuk Arab POS *tagging*, Aliwy [12] memperkenalkan pendekatan penandaan berdasarkan teknik *master-slave*, Ababou dan Mazroui [13] mengusulkan pendekatan statistik gabungan untuk POS *tagging* bahasa Arab, sebuah model bahasa dibuat oleh Zeroual dan Lakhouaja [14] untuk mengadopsi pohon *tagger* untuk menentukan *tag* dan teks *lemmatizing* Arab, Othmane et al. [15] mengusulkan pendekatan evolusioner POS *tagging* Arab berdasarkan algoritma *ant colony*, Albared et al. [16] mengusulkan sebuah pendekatan untuk *tag* teks Arab menggunakan *Hidden Markov Model* (HMM).

Tabel 1 Merangkum literatur sebelumnya

Sumber	Pendekatan	Ukuran *dataset*	Akurasi
[8]	Berbasis aturan	2.355 kata	94%
[9]	Algoritma genetika	45.000 kata	94,5%
[10]	Hibrida	18.000.000 kata	98%
[11]	KNN + NB	77.430 kata	98,3%
[12]	*Master-slave*	45.000 kata	90%
[13]	Hibrida	500.000 kata	94%
[14]	Pohon keputusan	78.121 kata	92,6%
[15]	Koloni semut	300.000 kata	97,3%
[16]	HMM	18.167.183 kata	95,8%

Terdapat banyak metode untuk menyelesaikan masalah *Out Of Vocabulary* (OOV) pada POS *tagging*. Pada penelitian ini, metode yang digunakan untuk menyelesaikan masalah POS *tagging* adalah metode pendekatan evolusioner menggunakan algoritma *Bee Colony Optimization* (BCO). Sebagian besar penelitian evolusi bergantung pada probabilitas kondisional seperti N-*gram* dan *Hidden Markove Model* (HMM) untuk melatih sistem. Karena model HMM mengalami kesulitan dalam memperkirakan akurasi probabilitas transisi yang disebabkan terbatasnya jumlah data pelatihan [17] sehingga menyebabkan masalah OOV, akibatnya metode baru dikembangkan untuk menghindari masalah yang HMM hadapi dalam memperkirakan akurasi probabilitas transisi dengan menggunakan algoritma BCO. Algoritma ini menggunakan teknik pembobotan baru yang menetapkan nilai transisi berdasarkan pada kejadian *tag* berikutnya tidak secara probabilitas. Ruang pencarian dikurangi dengan hanya menggunakan *tag* yang berlaku untuk setiap kata, juga lebah diperbolehkan untuk mengeksplorasi ruang pencarian dengan *multithreaded* yaitu ada beberapa nomor lebah mengeksplorasi ruang pencarian dan mencari urutan terbaik dari *tag* pada saat yang sama.

Matematika merupakan bidang ilmu yang dapat diapliaksikan ke dalam cabang ilmu-ilmu lain. Setiap permasalahan yang terjadi di dunia nyata ternyata dapat diselesaikan dengan matematika, yaitu dengan membuat model matematika dari permasalahan tersebut. Pada penelitian ini pendekatan POS *tagging* evolusioner memodelkan masalah POS *tagging* sebagai masalah optimasi dengan menggunakan representasi graf berbobot dan berarah.

Representasi graf merupakan graf yang di proses dengan program komputer. Graf berbobot dan berarah G = (V, E) digunakan untuk mewakili kalimat dimana V adalah satu set *node* mewakli semua label kelas kata yang mungkin dalam kalimat dan E adalah seperangkat busur yang menghubungkan *node* satu ke *node* berikutnya. Untuk mengevaluasi akurasi model yang dibangun, digunakan salah satu teknik validasi silang yaitu *k-fold cross validation*.

1.2 Rumusan Masalah

Berdasarkan latar belakang tersebut, skripsi ini memiliki rumusan masalah, antara lain:

1. Apa saja faktor permasalahan dalam proses POS *tagging* otomatis?
2. Bagaimana cara membangun POS *tagging* bahasa Arab menggunakan algoritma *Bee Colony Optimization* (BCO) untuk menyelesaikan masalah nilai transisi nol?
3. Seberapa besar tingkat keakurasian yang dihasilkan dari model yang dibangun untuk menyelesaikan POS *tagging* bahasa Arab?
4. Apa saja faktor yang mempengaruhi hasil POS *tagging* bahasa Arab menggunakan algoritma *Bee Colony Optimization* (BCO)?
5. Apa keunggulan algoritma *Bee Colony Optimization* (BCO)?

1.3 Batasan Masalah

Pada tugas akhir ini terdapat beberapa batasan masalah, batasan masalah yang digunakan diantaranya yaitu:

1. *Dataset* yang digunakan adalah Al Quran yang sudah ditransliterasi berasal dari Quran *corpus*. 150 kalimat sempurna sederhana digunakan untuk kategori *dataset* mudah, 50 kalimat dengan S/P/O/K lebih dari satu digunakan untuk kategori *dataset* sedang, dan 50 ayat Al Quran pilihan digunakan untuk kategori *dataset* sulit.
2. *Dataset* akan dibagi kedalam set pelatihan dan set pengujian.
3. *Dataset* akan dibagi kedalam 5 partisi sama rata.
4. Metode yang digunakan untuk POS *tagging* bahasa Arab adalah metode POS *tagging* evolusioner menggunakan algoritma *Bee Colony Optimization* (BCO).
5. Model yang digunakan untuk menyelesaikan permasalahan POS *tagging* adalah representasigraf berbobot dan berarah G = (V, E).
6. Metode yang digunakan untuk mengevaluasi model adalah *k-fold cross validation*.
7. Metode yang digunakan untuk menghitung bobot transisi merupakan metode baru tidak secara probabilistik.

1.4 Tujuan Penelitian

Adapun tujuan dan manfaat dari tugas akhir ini antara lain:

1. Menganalisa faktor-faktor yang menjadi permasalahan dalam proses POS *tagging* otomatis.
2. Membangun POS *tagging* bahasa Arab menggunakan algoritma *Bee Colony Optimization* (BCO) untuk menyelesaikan masalah nilai transisi nol.
3. Mengetahui tingkat keakurasian dari model yang dibangun untuk menyelesaikan POS *tagging* bahasa Arab.
4. Menganalisa faktor-faktor yang mempengaruhi hasil POS *tagging* bahasa Arab menggunakan algoritma *Bee Colony Optimization* (BCO).
5. Mengetahui keunggulan algoritma *Bee Colony Optimization* (BCO).

1.5 Metode Penelitian

Metode yang ditempuh oleh penulis dalam menyelesaikan tugas akhir ini adalah sebagai berikut:

1. Skripsi

Tahap skripsi merupakan tahap penulis mengumpulkan data dan informasi serta memahami materi mengenai masalah POS *tagging* menggunakan algoritma *Bee Colony Optimization* (BCO).

2. Percobaan

Pada tahap ini penulis melakukan beberapa percobaan dengan menggunakan salah satu teknik validasi silang yaitu *k-fold cross validation* dan membagi *dataset* ke dalam 5 partisi sama rata.

Percobaan yang akan dilakukan diantaranya, percobaan klasifikasi sederhana dengan *dataset* sebanyak 150 kalimat sempurna sederhana (*simple* sentence) yang bersumber dari Al Quran, percobaan klasifikasi sedang dengan *dataset* sebanyak 50 kalimat sempurna dengan S/P/O/K lebih dari satu (ada anak kalimat), dan percobaan klasifikasi ayat lengkap dengan *dataset* sebanyak 50 ayat Al Quran pilihan.

Model akan dievaluasi dengan cara menjadikan data latih sebagai data uji dengan total percobaan sebanyak satu kali, membagi *dataset* menjadi 80% set pelatihan dan 20% set pengujian dengan total percobaan sebanyak lima kali, membagi *dataset* menjadi 60% set pelatihan dan 40% set pengujian dengan total percobaan sebanyak sepuluh kali, membagi *dataset* menjadi 40% set pelatihan dan 60% set pengujian dengan total percobaan sebanyak sepuluh kali, dan membagi *dataset* menjadi 20% set pelatihan dan 80% set pengujian dengan total percobaan sebanyak lima kali, kemudian dilakukan analisis hasil dan ketepatan dari metode yang digunakan pada masing-masing kelas *dataset.*

1.6 Sistematika Penulisan

Berdasarkan sistematika penulisan, skripsi ini terdiri atas empat bab ditambah dengan daftar pustaka, dimana setiap bab nya memiliki beberapa subbab.

BAB I PENDAHULUAN

Bab ini berisi beberapa hal tentang pendahuluan diantaranya berupa latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, metode penelitian, dan sistematika penulisan dari masalah yang dikaji.

BAB II LANDASAN TEORI

Bab ini berisi teori-teori yang melandasi pembahasan utama yang berkaitan dan menunjang dalam penulisan skripsi, seperti *Part Of Speech* (POS) *tagging*, masalah POS *tagging*, algoritma *Bee Colony Optimization* (BCO), dan *k-fold cross validation.*

BAB III *PART OF SPEECH* (POS) *TAGGING* BAHASA ARAB MENGGUNAKAN ALGORITMA *BEE COLONY OPTIMIZATION* (BCO) PADA DATA AL QURAN

Bab ini berisi pembahasan tentang tahap pengambilan *dataset*, kemudian tahap transliterasi dan *text preprocessing*, lalu dilakukan teknik klasifikasi yaitu membagi *dataset* menjadi set pelatihan yang akan digunakan untuk membangun model dan set pengujian yang akan digunakan untuk menentukan kekaurasian model. Model yang digunakan adalah representasi graf berbobot dan berarah, penelitian ini menerapkan algoritma *Bee Colony Optimization* (BCO).

BAB IV ANALISIS *PART OF SPEECH* (POS) *TAGGING* BAHASA ARAB MENGGUNAKAN ALGORITMA *BEE COLONY OPTIMIZATION* (BCO) PADA DATA AL QURAN

Bab ini berisi pemaparan mengenai analisis hasil percobaan POS *tagging* bahasa Arab menggunakan algoritma *Bee Colony Optimization* (BCO) yang sudah dilakukan pada bab sebelumnya pada beberapa kategori *dataset* yang berasal dari teks Al Quran dengan menggunakan salah satu teknik validasi silang yaitu *k-fold cross validation*.

BAB V PENUTUP

Bab ini berisi simpulan sebagai hasil dari rumusan masalah yang telah dipaparkan serta berisi saran untuk penelitian selanjutnya sebagai pengembangan dari topik permasalahan bersangkutan.

DAFTAR PUSTAKA