BAB I

PENDAHULUAN

Pada bagian ini berisi bahasan tentang latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, ruang lingkup penelitian, dan sistematika penulisan.

1.1 Latar Belakang Masalah

Data-data berupa teks yang terus menerus meningkat penggunaannya menyebabkan data mencapai jumlah yang besar. Data tersebut akan memakan waktu yang sangat lama apabila dikelola secara manual, oleh karenanya para ahli mengembangkan bidang ilmu untuk mengelola data teks secara otomatis yang disebut *text mining* atau penambangan teks. *Text mining* bertujuan untuk mengambil informasi dari data teks yang diproses sebelumnya. Teknik yang dapat digunakan untuk mengelola data yang berjumlah besar adalah teknik *classification*. *Classification* merupakan suatu pendekatan sistematis yang berguna untuk membangun model kategori dari dataset yang diberikan. Salah satu penggunaan teknik *classification* digunakan pada penelitian ini yaitu *part of speech tagging.*

Part of speech (POS) *tagging* merupakan bagian dari *Natural Language Processing* (NLP). POS tagging adalah proses memberikan tag atau label kelas kata pada sebuah kata dalam dokumen secara otomatis. Menetapkan kelas kata (*part of speech*) untuk setiap kata dapat dilakukan secara manual akan tetapi sangat memakan waktu dan membutuhkan biaya yang mahal karena dalam pengerjaannya memerlukan ahli bahasa. Itulah sebabnya POS *tagging* menjadi salah satu masalah yang dipelajari dengan baik pada bidang NLP [1]. Hasil penelitian POS tagging dapat digunakan untuk dasar penelitian NLP lainnya, seperti: *Text Summarization, Machine Translation, Information Retrieval, Language Generator,* dan *Question and Answering.*

Masalah yang terjadi pada proses POS *tagging* adalah ambiguitas kata yaitu kata yang memiliki lebih dari satu kemungkinan kelas kata dan OOV (*Out Of Vocabulary*) yaitu kata-kata yang tidak terdapat dalam data pelatihan. Metode yang dapat digunakan untuk menyelesaikan masalah POS *tagging* adalah metode *Hidden Markov Model* (HMM). HMM merupakan model matematika yang sering digunakan untuk POS tagging, dikarenakan HMM melihat urutan dari suatu kejadian. Urutan kejadian pada POS *tagging* adalah kata pada sebuah kalimat. HMM termasuk kedalam metode berbasis probabilitas (probabilitas transisi dan probabilitas emisi) untuk menentukan tag pada sebuah kata secara otomatis. Perhitungan probabilitas transisi pada metode HMM dapat dilakukan dengan *N-gram* meliputi *Unigram*, *Bigram*, *Trigram*, dan *Quadgram*. *Unigram* menghitung probabilitas transisi pada tag itu sendiri, sedangkan POS *tagging* memerlukan probabilitas urutan tag sebelumnya. Oleh karena itu penelitian ini tidak menggunakan *Unigram*. Menghitung probabilitas transisi untuk *Bigram* yaitu berdasarkan tag sebelumnya, *Trigram* berdasarkan 2 tag sebelumnya sedangkan *Quadgram* berdasarkan 3 tag sebelumnya. Dikarenakan penelitian menggunakan metode HMM *Bigram* telah banyak dilakukan seperti pada [2], [3] dan probabilitas transisi pada *Quadgram* akan menghasilkan kombinasi yang terlalu banyak. Oleh karena itu, pada penelitian ini menggunakan metode HMM *Trigram* untuk digunakan dalam POS *tagging* bahasa Arab.

Metode HMM sering dikombinasikan dengan algoritma Viterbi. Algoritma Viterbi merupakan algoritma *dynamic programming* untuk menemukan barisan rangkaian tersembunyi (*viterbi path*) yang paling maksimal dari suatu barisan rangkaian pengamatan kejadian terutama dalam lingkup HMM.

Data bersumber dari *quran corpus* dikategori menggunakan POS tagging untuk didapatkan tag yang sesuai untuk setiap kata dari data yang diberikan. Karena data berbahasa Arab, POS *tagging* bahasa Arab bukanlah tugas yang mudah. Bahasa Arab memiliki tingkat ambiguitas yang tinggi, seperti penghilangan huruf vokal dan kesamaan huruf tetap dengan huruf induk atau akar. Karena bahasa Arab mempunyai keistimewaan di bidang tata bahasa di samping keistimewaannya yang lain, maka bahasa Arab dianggap rumit, komplek, sukar dan lain sebagainya, terutama di kalangan pelajar dan mahasiswa.

Penelitian tentang *part of speech tagging* telah banyak dilakukan sebelumnya, baik untuk dokumen berbahasa Indonesia, bahasa Inggris dan bahasa Arab. Metode yang digunakan untuk menyelesaikan masalah *part of speech tagging* juga berbeda-beda. *Part of speech tagging* terhadap dokumen bahasa Indonesia dengan mengimplementasikan *Brill Tagger* [2]. Dataset yang digunakan adalah artikel berita berjumlah 100 dokumen, 80% untuk proses *training* dan sisanya untuk proses *testing*. Nilai akurasi yang diperoleh sebesar 99,75%. POS *tagging* menggunakan *Hidden Markov Model* (HMM) dan algoritma Viterbi untuk dokumen berbahasa Inggris dengan nilai akurasi 96,35% [3]. *Part of speech tagging* bahasa Inggris menggunakan metode probabilistik, yaitu *Maximum Entropy Model* dengan nilai akurasi 96,6% [4]. POS *tagging* dengan metode HMM dan *rule based* untuk teks bahasa Indonesia dengan nilai akurasi 92,2% [5]. Penelitian yang dilakukan oleh Imad Zeroual, Abdelhak Lakhouaja dan Rachid Belahbib yang berjudul *Towards a standard Part of Speech tagset for the Arabic language* yang menjelaskan POS *tagging* bahasa Arab dengan metode probabilistik, HMM *Trigram* untuk pelatihan data dan algoritma Viterbi untuk pengujian data. Data yang digunakan berupa data teks bahasa Arab yaitu NEMLAR dan Al-Mus'haf, dengan nilai akurasi yang didapat untuk NEMLAR 93,55% dan Al-Mus'haf sebesar 96,11% [6].

Dalam penelitian ini metode yang digunakan adalah metode probabilistik, yang berdasarkan pada model Markov tersembunyi kedua [7]. Metode ini disebut metode HMM *Trigram*, yang merupakan perluasan dari metode HMM. Metode HMM *Trigram* akan digunakan pada sebuah sistem untuk dapat membantu proses POS *tagging* terhadap dokumen bahasa Arab seperti *quran corpus.*

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan sebelumnya, didapatkan rumusan masalah sebagai berikut:

1. Pendekatan berbasis aturan untuk masalah POS *tagging* menggunakan aturan buatan tangan oleh ahli bahasa, akan ada beberapa kata dalam teks input yang tidak dapat ditangani oleh aturan buatan tangan yang menyebabkan terjadinya OOV (*Out Of Vocabulary).*

2. Pada HMM *Bigram* saat menghitung probabilitas transisi pada kalimat yang jumlahnya banyak terdapat kesalahan dalam penentuan nilai transisinya.
3. Pada perhitungan probabilitas transisi HMM *Trigram* yaitu mengamati POS tag dengan 2 tag sebelumnya sedangkan pada probabilitas transisi HMM *Bigram* hanya mengamati POS tag dengan 1 tag sebelumnya saja.

1.3 Batasan Masalah

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

1. *Dataset* yang digunakan diperoleh dari https://corpus.quran.com/, dimana data telah melalui tahap transliterasi.
2. Data yang digunakan berjumlah 250 kalimat, dengan 150 kalimat sempurna sederhana untuk kategori sederhana, 50 kalimat dengan S/P/O/K lebih dari satu untuk kategori sedang (ada anak kalimat) , dan 50 ayat pilihan dalam Al-Qur'an yang digunakan untuk kategori ayat lengkap.
3. Metode yang digunakan untuk *part of speech tagging* adalah metode HMM *Trigram*.
4. Metode yang digunakan untuk variasi percobaan adalah *k-fold cross validation*.
5. Bahasa pemrograman yang digunakan adalah Python.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah di atas, tujuan dari penelitian ini adalah:

1. Membangun POS *tagging* bahasa Arab menggunakan metode HMM *Trigram* pada data Al-Qur'an untuk menyelesaikan permasalahan OOV (*Out Of Vocabulary)*.
2. Mendapatkan tingkat keakurasian dari metode HMM *Trigram* untuk menyelesaikan POS *tagging* pada data Al-Qur'an.
3. Mengetahui faktor-faktor yang mempengaruhi hasil POS *tagging* bahasa Arab menggunakan metode HMM *Trigram.*

1.5 Metode Penelitian

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

1. Studi Literatur

Tahap studi literatur merupakan tahap penulis mengumpulkan data dan memahami materi yang terkait dengan penelitian ini, yaitu mengenai *part of speech tagging* menggunakan beberapa metode, salah satunya metode HMM *Trigram* yang berdasarkan *Hidden Markov Model* (HMM) orde dua.

2. Penelitian

Pada tahap penelitian penulis menganalisa metode HMM *Trigram* terhadap penentuan *part of speech* atau kelas kata pada dokumen, kemudian memberi label kelas kata pada setiap kata dalam dokumen. Selanjutnya membandingkan hasil dari pelabelan menggunakan metode HMM *Trigram* dengan pelabelan pada quran korpus, untuk mendapatkan keakuratan metode untuk *part of speech tagging* pada data Al-Qur'an.

1.6 Sistematika Penulisan

Sistematika penulisan skripsi ini terdiri atas lima bab serta daftar pustaka dimana dalam setiap bab terdapat beberapa subbab.

BAB I PENDAHULUAN

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

BAB II LANDASAN TEORI

Bab ini berisi tentang teori-teori dasar yang berkaitan dengan penulisan skripsi, diantaranya *Text Mining, Text Classification, Text Preprocessing, Natural Language Processing,* bahasa Arab, pembagian kelas kata atau *part of speech, Part Of Speech* (POS) *Tagging, N-gram,* metode HMM *Trigram,* algoritma Viterbi dan metode variasi percobaan menggunakan *k-fold cross validation*.

BAB III *PART OF SPEECH* (POS) *TAGGING* MENGGUNAKAN METODE HMM *TRIGRAM* PADA DATA AL-QUR'AN

Bab ini berisi pembahasan dari penelitian yang dilakukan mulai dari pengambilan *dataset*, lalu tahap *text preprocessing*, kemudian melakukan *part of speech tagging*, dengan membagi *dataset* menjadi data *training* dan data *testing*, melakukan variasi percobaan dengan metode *k-fold cross validation* untuk mendapatkan akurasi yang baik.

BAB IV ANALISIS HASIL *PART OF SPEECH* (POS) *TAGGING* MENGGUNAKAN METODE HMM *TRIGRAM* PADA DATA AL-QUR'AN

Bab ini berisi pemaparan mengenai analisis hasil *part of speech tagging* menggunakan metode HMM *Trigram* yang sudah dilakukan pada *dataset* yang terbagi kedalam tiga kategori, yaitu sederhana, sedang (ada anak kalimat) dan ayat lengkap.

BAB V PENUTUP

Bab ini berisi tentang kesimpulan atas penelitian yang dilakukan sebagai hasil dari rumusan masalah yang sebelumnya telah dipaparkan dan beberapa saran pengembangan kedepannya.

DAFTAR PUSTAKA