BAB I
PENDAHULUAN

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

1.1 Latar Belakang Masalah

Al-Qur'an merupakan firman Allah SWT yang diturunkan kepada Nabi Muhammad SAW melalui malaikat Jibril secara berangsur-angsur. Al-Qur'an merupakan kitab yang menjadi pedoman hidup umat manusia agar bisa membedakan antara yang hak atau yang bathil, keta'atan atau kemaksiatan, berpahala atau berdosa. Al-Qur'an menggunakan bahasa Arab, salah satu hikmah Al-Qur'an diturunkan menggunakan bahasa Arab adalah agar manusia mampu berfikir akan kedalaman isinya, karena tanpa menguasai bahasa itu niscaya manusia tidak akan mendapatkan pemahaman yang maksimal akan isinya. Allah berfirman dalam Surat Yusuf ayat 2 yang berbunyi :

إِنَّا أَنزَلْنَاهُ قُرْآناً عَرَبِيّاً لَّعَلَّكُمْ تَعْقِلُونَ(2)

Artinya :"Sesungguhnya kami menurunkannya (Al-Qur'an) dengan berbahasa Arab, agar kamu memahaminya." (QS. Yusuf :2)

Kemudian dalam surat Asy-Syura' ayat 7 yang berbunyi :

وَكَذَٰلِكَ أَوْحَيْنَآ إِلَيْكَ قُرْءَانًا عَرَبِيًّا لِّتُنذِرَ أُمَّ ٱلْقُرَىٰ وَمَنْ حَوْلَهَا وَتُنذِرَ يَوْمَ ٱلْجَمْعِ لَا رَيْبَ فِيهِ ۚ فَرِيقٌ فِى ٱلْجَنَّةِ وَفَرِيقٌ فِى ٱلسَّعِيرِ

Artinya : "Demikianlah kami wahyukan kepadamu Al-Quran dalam bahasa Arab, supaya kamu memberi peringatan kepada Ummul Qura (Penduduk Mekkah) dan penduduk (Negeri-negeri) sekelilingnya serta memberi peringatan (pula) tentang hari berkumpul (kiamat) yang tidak ada keraguan kepadanya. Segolongan masuk surga, dan segolongan masuk Jahannam." (QS. Asy Syura' : 7)

Dan dalam surat Az-Zukhruf: 1-3 yang berbunyi:

حم (1) وَالْكِتَابِ الْمُبِينِ (2) إِنَّا جَعَلْنَاهُ قُرْآنًا عَرَبِيًّا لَعَلَّكُمْ تَعْقِلُونَ (3)

Artinya: "Haa Miim. Demi Kitab (Al Quran) yang menerangkan. Sesungguhnya Kami menjadikan Al Quran dalam bahasa Arab supaya kamu memahami(nya)."(QS. Az-Zukhruf :1-3).

Menurut Imam Ibnu Katsir dalam Tafsirnya yang berjudul *Tafsir al-Qur'an al-Adzim* menjelaskan bahwa

وذلك لأن لغة العرب أفصح اللغات وأبينها وأوسعها، وأكثرها تأدية للمعاني التي تقوم بالنفوس؛ فلهذا أنزل أشرف الكتب بأشرف اللغات، على أشرف الرسل، بسفارة (8)أشرف الملائكة، وكان ذلك في أشرف بقاع الأرض، وابتدئ إنزاله في أشرفشهور السنة وهو رمضان، فكمل من كل الوجوه

"Yang demikian itu (bahwa Al-Qur'an diturunkan dalam bahasa Arab) karena bahasa Arab adalah bahasa yang paling fasih, jelas, luas, dan maknanya lebih mengena lagi cocok untuk jiwa manusia. Oleh karena itu kitab yang paling mulia diturunkan (Al-Qur'an) kepada rasul yang paling mulia (Muhammad shollallohu 'alaihi wa sallam), dengan bahasa yang termulia (bahasa Arab), melalui perantara malaikat yang paling mulia (Jibril), ditambah diturunkan pada dataran yang paling muia diatas muka bumi (tanah Arab), serta awal turunnya pun pada bulan yang paling mulia (Ramadhan), sehingga Al-Qur'an menjadi sempurna dari segala sisi." [Tafsirul Qur'an Al-Adzim 4/366].

Khalifah Umar Bin Khattab R.A, menegaskan bahwa bahasa Arab adalah bagian dari agama, beliau berkata :

تعلموا العربية فإنها من دينكم

"*Pelajarilah bahasa Arab, sesungguhnya ia bagian dari agama kalian."*[Iqtidha' shiratal mustaqim 527-528 jilid I, tahqiq syaikh Nasir Abdul Karim Al-'Aql].

Menurut Syekh Manna al-Qattan dalam *Mabahis fi Ulum al-Qur'an* menjelaskan bahwa Al-Qur'an menggunakan bahasa Arab maka syarat utama bagi seorang penafsir Al-Qur'an (*mufassir*) harus mengetahui kaidah bahasa Arab, dan memahami dasar-dasarnya serta mengetahui rahasia yang terkandung di dalamnya seperti ilmu Nahwu, Sharaf, Ma'ani, Bayan, Badi', dan ilmu Mantik. Salah satu ilmu yang harus dikuasai dalam memahami bahasa Arab yaitu ilmu *Nahwu*. Ilmu *Nahwu* adalah salah satu bidang ilmu tata bahasa Arab yang

mempelajari tentang bagaimana menentukan kedudukan satu kalimat dari segi *i'rob* nya (Ahmad al- Hasyimi: 1354 H). Dalam ilmu ini membahas kaidah-kaidah bahasa Arab untuk mengetahui bentuk kata dan keadaan-keadaannya ketika masih satu kata (*Mufrod*) atau ketika sudah tersusun (*Murokkab*). Ruang lingkup pembahasan ilmu *Nahwu* meliputi, اسم (kata benda), فعل (kata kerja), حرف (huruf)[1].

Menurut Imam Asy-Syafi'i bahwa :

من تبَحَرَّ فى النحو اهتدى إلى كل العلوم

"Siapa yang menguasai ilmu Nahwu, dia dimudahkan untuk memahami seluruh ilmu."[Syadzarat ad-Dzahab, hlm. 1/321].

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 setiap 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 aplikasi *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]. Selain itu, para pelajar bahasa Arab seperti santri-santri di pondok pesantren biasanya memiliki sedikit kesulitan dalam menerjemahkan kitab kuning ataupun bahasa Arab lainnya karena diharuskan mengerti terlebih dahulu struktur dari bahasa Arab nya. Oleh karena itu, dengan berkembangannya ilmu pengetahuan dan teknologi yang pesat, manusia dituntut kemampuannya dalam membuat ataupun membangun suatu program yang dapat menyelesaikan masalah tersebut. POS *tagging* dengan menggunakan metode *Hidden Markov Model* dapat menyelesaikan masalah tersebut.

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. Salah satunya pendekatan berbasis aturan. Kelemahan utama dari pendekatan 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. Melihat kelemahan yang dimiliki pendekatan berbasis aturan para peneliti membuat suatu model atau sistem POS *tagging* melalui pendekatan *probabilistic* setiap kata dan kelas kata memiliki nilai peluang yang akan menentukan kecocokan suatu kata pada kelas katanya, ini akan lebih memudahkan dan mengefektifkan kinerja dalam penentuan struktur kata bahasa arab dibandingan dengan POS *tagging* berbasis aturan [3, 4].

Penelitian tentang *Part of Speech Tagging* menggunakan *Hidden Markov Model* dan beberapa metode lainnya telah banyak dilakukan oleh peneliti. Andrew Fireman dkk melakukan penelitian *Part of Speech Tagging* menggunakan *Rule Based*[5], kemudian Mona Diab dkk melakukan penelitian *Part of Speech Tagging* menggunakan SVM[6], Muhammad Hjouj dkk melakukan penelitian *Part of Speech Tagging* dengan menggunakan *Rule based* dan *Name Entity Recognition* pada data Arab[7]. Kadim dkk [8] mengusulkan dua model untuk *Hidden Markov Model* berbasis *Part of Speech Tagging* yang bekerja secara pararel menggunaan korpus Nemlar Arab. Ba-Alwi dkk [9] melakukan studi perbandingan antara pendekatan *POS Tagger* statistik berbasis morfem dan berbasis kata yaitu HMM *POS tagger based prefix guessing,* HMM *POS tagger based linear interpolation guessing,* dan *TnT tagger,* mereka menggunakan korpus linguistik beranotasi yang disebut Al-Qur'an Arab. Zeroual dkk [9] menyajikan tagger *Part of Speech* (POS) probabilistik untuk teks Arab berdasarkan *Hidden Markov Model* (HMM) yang disebut Tree Tagger. Hadni dkk [10] memperkenalkan tagger POS *hybrid* dengan 33 tag-set based *Hidden Markov Model* (HMM) dan *Rule-Based* menggunakan *Holy Quran Corpus*. Aajmi dkk [11] mempersentasikan metode *POS tagger* baru berdasarkan *Hidden Markov Model* (HMM) untuk mengekstraksi bobot kata dengan menghilaangkan prefix dan sufix yang dilampirkan pada sebuah kata. Köprü dkk [12] menemukan HMM berbasis *POS tagger* yang cepat dan lugas untuk teks Arab.

Penelitian tentang *Part of Speech Tagging* pada data bahasa Arab masih jarang dilakukan khusus nya pada data Al-Qur'an. Oleh karenanya penulis tertarik untuk meneliti tentang *Part of Speech Tagging* pada data Al-Qur'an dengan

menggunakan *Hidden Markov Model.* Kemudian meskipun penelitian *Part of Speech Tagging* menggunakan *Hidden Markov Model* telah banyak dilakukan dan telah banyak pengembangan penelitian, tetapi penulis masih ingin mengeksplorasi metode ini dengan berbagai percobaan dan ingin mengetahui pola struktur kata pada data Al-Qur'an. Pada penelitian ini penulis mencoba menggunakan 44 tag-set perincian dari nomina (*Isim*), verba (*Fi'il*), dan partikel (*Harf*), kemudian peneliti mencoba melakukan percobaan dengan membagi kedalam tiga kategori dataset yaitu untuk kategori pertama menggunakan dataset mudah yang struktur kalimatnya hanya SPO/K , kemudian dataset sedang yaitu struktur kalimatnya yang terdiri dari SPO/K terdapat lebih dari satu kali, dan terakhir dataset sulit yaitu satu ayat penuh.

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. Karena bahasa Arab mempunyai keistimewaan di bidang tata bahasa di samping keistimewaannya yang lain, maka banyak orang menganggap bahasa Arab itu rumit, komplek, sukar dan lain sebagainya, terutama di kalangan pelajar dan mahasiswa[13].

Karena sifatnya yang tidak konvensional, POS *tagging* bahasa Arab bukanlah tugas yang mudah, dan kurangnya literatur dalam POS *tagging* bahasa Arab 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 morfologi biasanya mempengaruhi tingkat analisis lain yang lebih tinggi seperti analisis sintaksis dan semantic[13].

Masalah utama dalam POS tagging antara lain kata ambigu dan kata *Out Of Vocabulary* (OOV)[14]. Kata ambigu merupakan kata yang memiliki sifat berbeda jika ditempatkan 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. Salah satu metode dari POS *tagging* berbasis stokastik yaitu *Hidden Markov Model* bisa menyelesaikan masalah tersebut .

Matematika merupakan bidang ilmu yang dapat diaplikasikan ke dalam cabang ilmu-ilmu lain. Setiap permasalahan yang terjadi di dunia nyata ternyata

dapat diselesakan dengan matematika, yaitu dengan membuat model matematika dari permasalahan tersebut. Pendekatan POS *tagging* stokastik memodelkan masalah POS *tagging* sebagai masalah probabilitas dengan menggunakan *Hidden Markov Model* dan representasi graf berbobot dan berarah.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang sudah dijelaskan sebelumnya, maka penulis merumuskan masalah yang diteliti pada skripsi ini sebagai berikut:

1. Pada POS *tagging* berbasis aturan dalam prosesnya dibutuhkan ahli linguistik untuk membuat aturan tata bahasanya, dan juga membutuhkan waktu dan biaya yang sangat besar dalam pembuatan aturannya.
2. Pada POS *tagging* berbasis aturan seringkali terjadi kesalahan *Out of Vocabulary* (OOV) yaitu suatu kata yang tidak ada di data latih, tetapi ada saat pengujian kata yang mengakibatkan kesalahan dalam penentuan tag.

1.3 Batasan Masalah

Agar penulisan skripsi ini tidak terlalu luas, maka penulis akan membatasi masalah POS *tagging* ini pada:

1. *Dataset* yang digunakan adalah data Al-Qur'an yang sudah ditransliterasi yang diambil dari *Qur'an.corpus*. terdiri dari 200 ayat dan terbagi kedalam 3 kategori struktur kata yaitu: sederhana (SPO/K) 150 ayat, kompleks (SPO/K terdapat lebih dari satu kali) 50 ayat, satu ayat penuh (tata bahasa arab, sintaksis dan morfologi untuk setiap kata di seluruh kategori data).
2. Dataset akan dibagi kedalam set pelatihan dan set pengujian
3. Dataset akan dibagi kedalam 5 partisi sama rata
4. Metode yang digunakan untuk mengevaluasi model adalah *k-fold cross validation.*

1.4 Tujuan dan Manfaat Penelitian

Berdasarkan latar belakang masalah dan rumusan masalah yang telah dipaparkan di atas, terdapat beberapa tujuan yang ingin dicapai oleh penulis dalam melakukan penelitian pada Skripsi ini antara lain:

1. Sebagai implementasi konsep wahyu memandu ilmu Fakultas Sains dan Teknologi UIN Sunan Gunung Djati Bandung, dimana Al-Qur'an sebagai sumber utama ilmu pengetahuan yang digunakan dalam objek penelitian ini kemudian diintegrasikan dengan teknologi masa sekarang.
2. Membangun model POS *tagging* bahasa Arab menggunakan *Hidden Markov Model* untuk menyelesaikan permasalahan POS *tagging* berbasis aturan dan permasalahan *Out of Vacabulary* (OOV).
3. Menganalisa faktor-faktor yang mempengaruhi hasil kinerja POS *tagging* bahasa Arab menggunakan *Hidden Markov Model*.

Adapun manfaat dari penulisan skripsi ini, diantaranya sebagai berikut:

1. Memberikan pemahaman tentang cara menentukan label kelas kata dengan menggunakan *Hidden Markov Model*.
2. Memudahkan proses POS *tagging* dengan Menggunakan *Hidden Markov Model*.
3. Untuk pengembangan lebih lanjut bisa menjadi suatu aplikasi yang dapat memudahkan umat islam maupun penuntut ilmu seperti santri dalam belajar bahasa Arab khususnya dalam bidang ilmu Nahwu maupun dalam mempelajari Al-Quran dengan melihat pola-pola struktur kata pada setiap kata dalam Al-Qur'an.

1.5 Metode Penelitian

1. Studi Literatur

Tahap Studi Literatur merupakan tahap penulis mengumpulkan dan memahami materi yang terkait dengan *Part of Speech tagging* dan metode *Part of Speech tagging* berbasis *probabilistic* yaitu *Hidden Markov Model.*

2. Percobaan

Pada tahap penelitian penulis melakukan pengembangan dengan menambahkan label kelas kata untuk mengetahui keefektifan model *Hidden Markov Model* dengan menggunakan 44 label kelas kata. Lalu untuk menguji kelayakan metode *Hidden Markov Model* ini menggunakan tiga kategori data Al-Qur'an serta dilakukan variasi percobaan dengan 31 kali percobaan dan juga dievaluasi dengan *k-fold cross validation*. kemudian menganalisis hasil dari *Part*

of Speech Tagging menggunakan *Hidden Markov Model* untuk mengetahui pola-pola data dan mengetahui penyebab dari faktor-faktor yang mempengaruhi hasil tersebut.

1.6 Sistematika Penulisan

Berdasarkan sistematika penulisannya, studi literatur ini terdiri atas empat bab serta daftar pustaka, dimana dalam setiap bab terdapat beberapa subbab.

BAB I PENDAHULUAN

Bab ini berisi tentang latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, ruang lingkup penelitian, dan sistematika penelitian.

BAB II LANDASAN TEORI

Bab ini memaparkan tentang landasan teori yang menunjang Skripsi ini seperti *text mining*, Natural Language Proccessing, *Part of Speech tagging*, *text processing*, *Hidden Markov Model*, dan Algoritma Viterbi.

BAB III PROSES *PART OF SPEECH TAGGING* DENGAN MENGGUNAKAN *HIDDEN MARKOV MODEL* PADA DATA AL-QUR'AN

Bab ini memaparkan tentang penelitian yang dilakukan mulai dari pengumpulan data, *text processing*, Pelatihan Data dengan *Hidden Markov Model*, pengujian data dengan Algoritma Viterbi, dan terakhir Evaluasi.

BAB IV ANALISIS HASIL *PART OF SPEECH TAGGING* DENGAN MENGGUNAKAN *HIDDEN MARKOV MODEL*

Bab ini berisi pemaparan mengenai analisis hasil *part of speech tagging* menggunakan *Hidden Markov Model* yang sudah dilakukan pada *dataset* yang terbagi kedalam tiga kategori, yaitu mudah, sedang dan sulit.

BAB V : PENUTUP

Bab ini merupakan intisari dari bab-bab sebelumnya yang terdiri kesimpulan dan saran untuk pengembangan penelitian yang lebih baik.