BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Suatu cerita memiliki isi, dimana kumpulan kata tertata rapi pada setiap paragraf, berisikan suatu kalimat dirangkai oleh beberapa kata yang akan menyampaikan atau memberitahukan suatu informasi atau suatu gagasan yang berbentuk serangkaian yang saling berkaitan satu sama lain. Fungsi dari paragraf itu sendiri merupakan indikator dimulainya topik baru serta memisahkan gagasan utama yang berbeda. Pemakaian paragraf mempermudah pembaca menguasai teks secara merata. Panjang dari suatu paragraf dapat menentukan kualitas dari bacaan. Kalimat itu sendiri ialah satuan bahasa kata ataupun rangkaian kata yang mampu berdiri sendiri serta mengungkapkan arti yang lengkap ataupun satuan bahasa terkecil yang menyatakan pemikiran yang utuh.

Kalimat yang digunakan dapat dikembalikan ke dalam sebagian kalimat dasar yang sangat terbatas. Dengan perkataan lain, segala kalimat yang kita gunakan berasal dari sebagian pola kalimat dasar saja. Sesuai dengan kebutuhan setiap manusia, kalimat dasar tersebut kita kembangkan, yang pengembangannya itu tentu saja didasarkan pada kaidah yang berlaku. Bersumber pada penjelasan terdahulu, bisa ditarik kesimpulan, kalimat dasar merupakan kalimat yang berisi informasi pokok dalam struktur inti serta belum dihadapkan perubahan. Perubahan itu bisa berbentuk akumulasi faktor semacam penambahan penjelasan kalimat maupun penjelasan subjek, predikat, objek ataupun pelengkap. Kata merupakan hal yang sangat penting dari suatu kalimat, apalagi suatu kata dapat mempengaruhi alur dari jalannya suatu cerita yang akan disampaikan dengan baik.

Kata atau ayat itu sendiri memiliki arti atau berfungsi sebagai subjek, predikat, objek, atau deskripsi dalam kalimat, di antara fungsi gramatikal lainnya. Kata dasar, kata turunan, kata berulang, dan kata majemuk adalah empat kategori di mana kata dapat dibagi. Kata dasar berasal dari bentukan kata yang merupakan derivasi atau imbuhan. Imbuhan yang terdapat pada awalan kalimat (prefiks), tengah (infiks), dan akhiran (sufiks). Sedangkan kata majemuk adalah gabungan

dari beberapa kata dasar yang terpisah untuk menciptakan makna baru, sedangkan kata berulang adalah kata dasar atau bentuk dasar yang mengalami pengulangan secara keseluruhan atau sebagian. Dalam tata bahasa baku bahasa Indonesia, kata memiliki kelas kata yang terbagi menjadi tujuh kategori, yaitu: nomina(kata benda), adjektiva(kata sifat), adverbia(kata keterangan), pronomina(kata ganti), numeralia(kata bilangan) dan kata tugas atau partikel yang terdiri dari preposisi(kata depan), konjungsi(kata sambung), artikula(kata sandang) dan interjeksi(kata seru). Kelas kata ini termasuk dalam metode skripsi yang sedang dikaji oleh penulis disebut dengan *part of speech tagging* yang disingkat menjadi POST.

Teknik pemberian kata dalam teks (korpus) ke kelas kata tertentu berdasarkan definisinya dan kata-kata yang mendahului atau terhubung dengannya dalam frasa, kalimat, atau paragraf dikenal sebagai penandaan bagian ucapan dalam linguistik. Proses pemberian label kelas kata secara otomatis ke setiap kata dalam kalimat dikenal sebagai pelabelan kelas kata. *Part-of-Speech Tagging* merupakan bagian dari *Natural Languange Processing* dalam menentukan kelas kata. Hasil penelitian *Part of Speech Tagging* pada dokumen dapat digunakan sebagai dasar penelitian dalam *Natural Languange Processing* lainnya, seperti: *Language Generator, Information Retrieval, Text Summarization, Question and Answering, dan Machine Translation* [1]. Dalam penelitian ini penandaan kelas kata akan berfokus pada kata berbahasa Arab, karena dataset yang digunakan dalam penelitian ini adalah Al-Qur'an.

Al-Qur'an ini sudah tak asing bagi kaum muslim. Al-Qur'an merupakan kitab suci utama dalam agama Islam, yang umat Islam percaya bahwa kitab ini diturunkan oleh tuhan kepada Nabi Muhammad, kitab ini terbagi dalam beberapa bab dan setiap suratnya terbagi ke dalam beberapa ayat. Al-Qur'an menggunakan bahasa Arab, di mana bahasa Arab merupakan bahasa yang sangat istimewa. Hal itu diketahui dengan kompleksitas yang menghadirkan beberapa tantangan untuk penandaan POS seperti ambiguitas tinggi, ketersebaran data dan keberadaan besar kata-kata yang tidak dikenal. Dengan mengingat hal ini, masalah utama di sini adalah menemukan bagaimana metode yang ada bekerja efisien dalam bahasa

Arab dan bagaimana korpus Al-Qur'an dapat digunakan untuk menghasilkan kerangka kerja yang efisien dalam penandaan POS berbahasa Arab.

Lalu klasifikasi teks di mana proses berdasarkan kata, frasa maupun kombinasinya dengan baik untuk menentukan kategori yang telah ditetapkan sebelumnya. Pengolahan klasifikasi teks melibatkan dua proses utama, yakni pertama ekstraksi fitur yang menjadi kata kunci yang efektif dalam tahap pelatihan atau *training* dan kemudian proses kedua yakni klasifikasi dokumen setelah melalui tahap uji atau *testing* [2]. Proses pengelompokan dokumen atau teks bisa dilakukan dengan dua cara yaitu dengan klasifikasi dan dengan clusterisasi. Klasifikasi dan *clustering* telah dipelajari selama bertahun-tahun oleh para peneliti pencarian informasi, dengan tujuan meningkatkan efektivitas, atau dalam beberapa kasus efisiensi, aplikasi pencarian. Dari perspektif lain, dua tugas ini adalah masalah pembelajaran mesin klasik. Dalam pembelajaran mesin, algoritma pembelajaran biasanya dicirikan sebagai diawasi atau tanpa pengawasan. Dalam pembelajaran diawasi, model dipelajari menggunakan satu set item sepenuhnya berlabel, yang sering disebut *training set*[2].

Menurut Hatta (2013) bahwa klasifikasi teks dokumen juga dapat diterapkan dalam Bahasa Arab, ini dikarenakan Bahasa Arab memiliki morfologi yang lebih kaya dan kompleks daripada bahasa Inggris maupun bahasa Indonesia, dimana dalam teks Bahasa Arab kita dapat mencari bentuk morfologi sebuah kata dari *stem* atau kata dasarnya. *Stemming* merupakan suatu proses menemukan kata dasar dari sebuah kata, dengan menghilangkan semua imbuhan (*affixes*) baik yang terdiri dari awalan (*prefixes*), sisipan (*infixes*), akhiran (*suffixes*) dan kombinasi dari awalan dan akhiran (*confixes*) pada kata turunan. Bentuk kata dapat diubah menjadi kata fundamental dengan menggunakan *stemming*. Setelah beberapa tahapan pada proses *Preprocessing*, dilanjut dengan memahami suatu metode klasifikasi yang pertama yaitu metode *K-Nearest Neighbor*. *K-Nearest Neighbor* (KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Teknik ini sangat sederhana dan mudah diimplementasikan. Mirip dengan teknik

clustering, mengelompokkan suatu data baru berdasarkan jarak data baru itu ke beberapa data tetangga (*neighbor*) terdekat [3].

Lalu metode klasifikasi yang kedua yaitu *Naïve Bayes* atau pengklasifikasi *Bayes* merupakan salah satu pengklasifikasi statistik, dimana pengklasifikasi ini dapat memprediksi probabilitas keanggotaan kelas suatu data *tuple* yang akan masuk ke dalam kelas tertentu, sesuai dengan perhitungan probabilitas. Pengklasifikasi *Bayes* didasari oleh teorema *Bayes* yang ditemukan oleh *Thomas Bayes* pada abad ke-18. Dalam studi perbandingan algoritma klasifikasi telah ditemukan simple *Bayesian* atau yang biasa dikenal dengan *Naïve Bayes classifier. Naïve Bayes classifier* menunjukkan akurasi dan kecepatan yang tinggi bila diterapkan pada database yang besar. Metode ini sering digunakan dalam menyelesaikan masalah dalam bidang mesin pembelajaran karena metode ini dikenal memiliki tingkat akurasi yang tinggi dengan perhitungan sederhana [2].

Di sini penulis mengusulkan eksperimen kombinasi pengklasifikasi kerangka kerja untuk penandaan POS Arab, dengan memilih dua pengklasifikasi probabilistik beragam terbaik yang digunakan dalam bahasa non-Arab: yaitu *K-Nearest Neighbor* (KNN) dan *Naïve Bayes* (NB). Dari penelitian sebelumnya yaitu studi literatur (SL) sudah didapat bahwasannya tingkat akurasi tertinggi diperoleh dengan hasil 94% pada metode *K-Nearest Neighbor* dan 64% diperoleh hasil dari metode *Naïve Bayes.* Peneliti ingin mengetahui bagaimana alur dari suatu program dengan menggunakan dataset bahasa Arab dan mengetahui hasil tingkat akurasi dari masing-masing metode statistik. Setelah mendapatkan hasil dari penelitian sebelumnya penulis melakukan beberapa percobaan lagi dengan metode yang sama ternyata hasilnya berbeda dari yang sebelumnya.

Dengan dataset Al-Qur'an dan dua metode statistik yaitu *K-Nearest Neighbor* memperoleh tingkat akurasi dengan 30,84% dan *Naïve Bayes* memperoleh 54,50%, hasil ini ternyata jauh berbeda dari penelitian sebelumnya. Setelah mendapatkan tingkat akurasi masing-masing metode statistik dengan fitur efektif untuk mendukung proses alur program yang sedang dimulai, lanjut melalui proses pemungutan suara mayoritas yang digunakan sebagai strategi kombinasi untuk mengeksploitasi keunggulan pengklasifikasi. Selain itu, pendalaman studi

telah dilakukan pada daftar fitur yang besar untuk mengeksploitasi fitur yang efektif dan menyelidiki perannya dalam meningkatkan kinerja penanda POS untuk bahasa Arab Al-Qur'an. Oleh karena itu, penelitian ini bertujuan untuk secara efisien mengintegrasikan rangkaian fitur dan algoritma penandaan yang berbeda untuk mensintesis penandaan POS yang lebih akurat. Setelah melakukan tahap pengklasifikasi dengan beberapa metode dilanjutkan dengan menggabungkan *K-Nearest Neighbour* dan *Naive Bayes* menggunakan algoritma voting mayoritas (AVM) atau algoritma strategi kombinasi. Lalu metode ansambel yang menggunakan beberapa algoritma pembelajaran untuk menghasilkan jawaban prediksi yang lebih baik daripada yang dapat dihasilkan oleh algoritma pembelajaran konstituen tunggal, digunakan untuk menemukan algoritma yang memberikan hasil prediksi terbaik jika dibandingkan dengan algoritma lain.

Sebagai umat beragama Islam terdapat data yang merupakan sumber hukum syariat utama dalam agama Islam yang bersumber dari Al-Qur'an yang merupakan firman Allah untuk menjadi pedoman hidup umat manusia agar bisa membedakan antara yang hak atau yang batil, ketaatan atau kemaksiatan, berpahala atau berdosa. Kitab suci umat Islam yang bertindak sebagai pedoman tentang bagaimana manusia harus menjalani kehidupan mereka. Sebagaimana firman Allah *Subhanahu Wa Ta'ala* dalam Qur'an Surah Al-Baqarah ayat 2.

ذٰلِكَ الْكِتٰبُ لَا رَيْبَ ۛ فِيْهِ ۛ هُدًى لِّلْمُتَّقِيْنَۙ

Artinya: *"Kitab (Al-Qur'an) ini tidak ada keraguan padanya: petunjuk bagi mereka yang bertakwa."* (QS. Al-Baqarah: 2)

Maka berdasarkan kitab-kitab Allah yaitu Zabur, Taurat, Injil dan Al-Qur'an menjadi petunjuk bagi manusia untuk bisa beribadah dengan benar. Tanpa menggunakan kitab-kitab Allah sebagai petunjuk, akan mengubah manusia menjadi hamba yang celaka. Fungsi kitab Allah sebagai petunjuk atau pedoman manusia juga dibahas dalam beberapa ayat Al-Qur'an antara lain:

إِنَّ هَٰذَا ٱلْقُرْءَانَ يَهْدِى لِلَّتِى هِىَ أَقْوَمُ وَيُبَشِّرُ ٱلْمُؤْمِنِينَ ٱلَّذِينَ يَعْمَلُونَ ٱلصَّٰلِحَٰتِ أَنَّ لَهُمْ أَجْرًا كَبِيرًا

Artinya: *"Sesungguhnya Al-Qur'an ini memberikan petunjuk kepada (jalan) yang lebih lurus dan memberi kabar gembira kepada orang-orang Mukmin yang mengerjakan amal saleh bahwa bagi mereka ada pahala yang besar."* (QS. Al-Isra: 9)

وَلَقَدْ جِئْنٰهُمْ بِكِتٰبٍ فَصَّلْنٰهُ عَلٰى عِلْمٍ هُدًى وَّرَحْمَةً لِّقَوْمٍ يُّؤْمِنُوْنَ

Artinya: *"Dan sesungguhnya Kami telah mendatangkan sebuah Kitab (Al-Qur'an) kepada mereka yang Kami telah menjelaskannya atas dasar pengetahuan Kami: menjadi petunjuk dan rahmat bagi orang-orang yang beriman."* (QS. Al-A'raf: 52)

وَنَزَّلْنَا عَلَيْكَ الْكِتٰبَ تِبْيَانًا لِّكُلِّ شَيْءٍ وَّهُدًى وَّرَحْمَةً وَّبُشْرٰى لِلْمُسْلِمِيْنَ

Artinya: *"Dan Kami turunkan kepadamu Alkitab (Al-Qur'an) untuk menjelaskan segala sesuatu dan petunjuk serta rahmat bagi orang-orang yang berserah diri."* (QS. An-Nahl: 89)

Petunjuk yang akan menunjukkan sesuatu (tanda/isyarat) untuk memberi tahu ketentuan atau arahan dan bimbingan bagaimana sesuatu harus dilakukan. Ibaratkan yang dilakukan pasti memiliki nilai benar atau salah sesuai yang sudah ditegaskan dalam suatu pedoman atau ajaran.

وَلَا تَلْبِسُوا الْحَقَّ بِالْبَاطِلِ وَتَكْتُمُوا الْحَقَّ وَاَنْتُمْ تَعْلَمُوْنَ

Artinya: *"Dan janganlah kamu campur adukkan kebenaran dengan kebatilan dan (janganlah) kamu sembunyikan kebenaran, sedangkan kamu mengetahuinya."* (QS. Al-Baqarah: 42)

Untuk dapat mengenal, memahami, dan menafsirkan Al-Qur'an tidak hanya berbekal pengetahuan bahasa Arab, melainkan dibutuhkan berbagai macam ilmu guna untuk mengungkapkan makna dan mampu menjadikan contoh dalam

kehidupan sehari-hari, baik yang sedang dilakukan atau dijalani. Karena itulah penulis dalam penelitian ini mengambil judul "*Part Of Speech Tagging* Al-Qur'an Menggunakan Kombinasi Dengan Metode *K-Nearest Neighbor* dan *Naive Bayes".*

1.2 Rumusan Masalah

Berdasarkan dari uraian latar belakang yang telah penulis sampaikan sebelumnya, maka rumusan masalah yang akan diteliti pada Skripsi ini adalah sebagai berikut:

1. Bagaimana proses jalannya suatu alur dalam penelitian ini dengan menggunakan *part of speech tagging*?
2. Fitur apa saja yang mempengaruhi untuk bisa menghasilkan fitur efektif dengan fitur yang tersedia?

1.3 Batasan Masalah

Untuk menjaga agar penelitian Skripsi ini dapat fokus pada rumusan masalah dan tidak menyimpang dari tujuan yang ingin diperoleh, maka penulis menentukan batasan masalah sebagai berikut:

1. Data yang digunakan berupa dokumen bahasa Arab yang diambil dari Al-Qur'an yaitu surat Al-Baqarah ayat 01- 286 dengan jumlah 6115 perkata.
2. Metode yang digunakan yaitu metode *K- Nearest Neighbor* menggunakan jarak euclidean, *Naive Bayes* menggunakan probabilistik dan strategi kombinasi menggunakan Algoritma Voting Mayoritas.
3. *Part of Speech Tagging* dibagi dengan tiga kelas kata yaitu kelas kata nominal yang berisi kata benda, kata sifat, kata ganti. Kelas kata kerja, lalu kelas kata partikel yang berisi kata depan, kata hubung dan preposisi.
4. Menggunakan 27 Set pola fitur simbol yang digunakan dalam percobaan (F1 F2 F3 F4 F5 F6 F7 F8 F9 F10 F11 F12 F13 F14 F15 F16 F17 F18 F19 F20 F21 F22 F23 F24 F25 F26 dan F27).
5. Menggunakan 27 Set pola fitur yang digunakan dalam percobaan ($P_{0,}$ $W_{-3,}$ $W_{-2,}$ $W_{-1,}$ $W_{0,}$ $W_{+1,}$ $W_{+2,}$ $W_{+3,}$ $P_{-3,}$ $P_{-2,}$ $P_{-1,}$ $P_{0,}$ $P_{+1,}$ $P_{+2,}$ $P_{+3,}$ $S_{1,}$ $S_1S_{2,}$ $S_1S_2S_{3,}$ $S_1S_2S_3S_{4,}$ $S_{0,}$ $S_0S_{-1,}$ $S_0S_{-1}S_{-2,}$ $S_0S_{-1}S_{-2}S_{-3,}$ Semua Huruf Besar, Semua Huruf Kecil, Berisi Nomor dan Panjang).

1.4 Tujuan dan Manfaat Penelitian

Berdasarkan latar belakang masalah dan rumusan masalah yang telah dijelaskan, terdapat beberapa tujuan yang ingin dicapai dalam penelitian Skripsi ini, antara lain:

1. Sebagai implementasi konsep Wahyu Memandu Ilmu, dimana objek dalam penelitian ini merupakan Teks Al-Qur'an yang diintegrasikan dengan perkembangan teknologi.
2. Untuk mengetahui proses jalannya suatu program dari awal sampai dengan hasil akhirnya, apabila menggunakan suatu *Part of Speech Tagging*.
3. Untuk mengetahui fitur apa saja yang efektif dan mendukung dalam menghasilkan nilai akurasi terbaik.

Adapun manfaat dari penelitian ini adalah sebagai berikut:

1. Hasil penelitian ini diharapkan menjadi salah satu bentuk pengembangan dan pengetahuan dalam kajian klasifikasi khususnya dalam klasifikasi data teks bahasa Arab.
2. Memberikan pengetahuan tentang pengawalan suatu program yang dimulai dari data yang terbentuk dari beberapa kalimat tersebut bisa terpisah menjadi suata kata (perkata). Di mana perkata ini akan dilakukan tahapan *Part of Speech Tagging*, lalu menggunakan metode statistika dan penggabungan metode yang digunakan.
3. Memberikan pengetahuan berupa hal yang mendukung dari jalannya suatu program dimana ada beberapa pengaruh dari suatu fitur yang digunakan, karena tidak setiap yang tersedia memiliki suatu hal yang menguntungkan.

1.5 Metode Penelitian

1. Studi Literatur

   Tahap studi literatur merupakan tahap untuk mengumpulkan data, materi dan informasi berbagai macam tentang POST, KNN, NB dan AVM (algoritma voting mayoritas) dari berbagai sumber, diantara artikel, buku, jurnal, dan lain sebagainya.

2. Analisis

Pada tahap ini, penulis mengkaji dan menganalisis hasil dari setiap tahap yang sudah dilakukan pada studi literatur sesuai dengan masalah yang ada dalam skripsi ini. Kemudian tahap ini juga dilakukan pembuatan dataset surat Al-Baqarah dari ayat 01-286.

3. Simulasi

Pada tahap ini penulis melakukan pengujian metode statistika yaitu metode *K-Nearest Neighbor* dan *Naive Bayes*, Ditambahkan dengan metode strategi kombinasi yaitu Algoritma Voting Mayoritas yang akan membuat nilai akurasi menjadi lebih baik dengan beberapa fitur yang tersedia dan menggunakan bahasa pemrograman Python. Kemudian akan menghasilkan nilai akurasi terbaik dengan beberapa fitur yang mendukung pada setiap percobaan.

1.6 Sistematika Penulisan

Sistematika penulisan pada Skripsi ini terdiri dari lima bab dan di dalam setiap bab terdiri dari beberapa subbab. Dengan sistematika penulisan sebagai berikut:

BAB I : PENDAHULUAN

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

BAB II : LANDASAN TEORI

Bab ini berisi penjelasan mengenai teori-teori yang berkaitan dengan masalah yang akan dikaji.

BAB III : METODE DAN *PART OF SPEECH TAGGING* DENGAN DATA AL-QUR'AN

Bab ini berisi pembahasan utama dari studi yang meliputi pembahasan mengenai hasil kerja dari *part of speech tagging* Teks Arab dengan Al-Qur'an yang dilabeli sesuai kelas katanya.

BAB IV : ANALISIS HASIL *PART OF SPEECH TAGGING* AL-QUR'AN MENGGUNAKAN KOMBINASI DENGAN METODE *K-NEAREST NEIGHBOR* DAN *NAIVE BAYES*

Bab ini berisi penjelasan pengujian metode KNN, NB dan AVM menggunakan data Al-Qur'an dengan fitur yang ada dan fitur akan dipilih secara random(acak).

BAB V : PENUTUP

Bab ini berisi penjelasan mengenai beberapa hal yang menjadi kesimpulan atas penelitian yang telah dilakukan serta beberapa saran yang berisi rekomendasi untuk pengembangan tulisan ini.