BAB I Pendahuluan

Pada bab ini akan menjelaskan tentang latar belakang, perumusan masalah, tujuan, metodologi penelitian dan sistematika penulisan yang digunakan.

1.1 Latar Belakang

Al-Quran adalah kitab suci yang menjadi sumber dan pedoman hidup bagi umat islam. Al-Quran merupakan wahyu Tuhan bagi Nabi-Nya untuk disampaikan kepada ummat islam yang terdiri dari 30 Juz, 114 Surat dan 6236 ayat, yang kebanyakan ayat merupakan kalimat pendek. Pada Al-Quran terdapat makna yang sama atau terkait namun tersebar dalam beberapa surat atau juz. Salah satu cara untuk memahami Al-Quran adalah dengan cara mencari kesamaan dan keterkaitannya agar diperoleh kandungan informasi yang lengkap. Sulit untuk sistem dalam hal mencari kesamaan dan keterkaitan ayat Al-Quran, karena tidak memiliki kemampuan intuisi seperti manusia. Dengan intuisi yang dimiliki, manusia dapat dengan tepat menentukan kesamaan dan keterkaitan dari informasi yang disediakan. Oleh karena itu, diperlukan sistem yang dapat menilai kesamaan antar ayat dengan ayat lainnya dengan *Semantic Textual Similarity*.

Semantic Textual Similarity disingkat STS merupakan suatu konsep yang dapat mengukur kesamaan makna semantik berdasarkan potongan teks yang berpasangan dengan algoritma dan model komputasi yang meniru kinerja manusia. STS telah banyak digunakan dalam penerapan disiplin ilmu *Natural Language Processing* (NLP) dan *text mining* seperti *information retrieval, machine translation, question answering, text summarization,* dan lain sebagainya [5]. Sebagai wadah untuk mendorong para peneliti bidang komputasi *linguistic* mengembangkan pendekatan baru dalam STS dilaksanakan *Semantic Evaluation* disingkat SemEval. SemEval terdiri dari berbagai *task*. Salah satu metode yang sering digunakan dalam penyelesaian *task-task* tersebut adalah *word alignment based similarity*.

Word alignment based similarity adalah metode penyejajaran kata-kata dalam kalimat yang sama. Metode ini dikembangkan oleh Sultan et al untuk penyejajaran kata-kata dalam kalimat yang sama dan diukur kesamaannya berdasarkan identifikasi urutan dan posisinya. Dalam penelitian SemEval 2015 dan 2016, fitur *alignment* dikombinasikan dengan fitur lain bertujuan untuk menghasilkan model yang lebih baik. Salah satu fitur yang dikombinasikan adalah model perhitungan kata dalam dokumen yang disebut TF-IDF. *Word alignment* pada dasarnya mengandalkan *paraphrase* dalam penyejajaran kata. Hal ini menjadi kelemahan dalam penyejajaran kata kata yang bukan *paraphrase*. Permasalahan tersebut dapat diatasi dengan merepresentasikan kata kedalam bentuk vektor [1]. Representasi ini dikenal juga dengan istilah vektor semantik.

Penelitian ini akan melakukan analisa dan implementasi perhitungan nilai semantik kemiripan kata pada ayat Al-Quran. Data yang digunakan merupakan terjemahan Al-Quran bahasa Inggris menggunakan pendekatan *word alignment* dan

vektor semantik *word2vec*. Untuk menghitung evaluasi dari penelitian ini digunakan model *Support Vector Regression* (SVR). Model SVR mencoba memprediksi tingkat kesamaan semantik antar kata atau *phrase* dengan asumsi bahwa hal tersebut dapat diwakili oleh probabilitas *annotator* (manusia) secara acak yang telah meng-*annotate* (membubuhi) keterangan pasangan *paraphrase* [6]. Hasil anotasi tersebut dijadikan *gold standard*, nilai acuan penelitian dengan skala 0 sampai 5.

Pendekatan *word alignment*, vektor semantik dan model regresi SVR ini dipilih karena merupakan salah satu metode yang diterapkan oleh Thomas Brychcin et al dari tim UWB yang meraih peringkat kedua pada kompetisi SemEval 2016 [1].

1.2 Perumusan Masalah

Berdasarkan latar belakang masalah yang telah diuraikan, maka permasalahan yang dapat dirumuskan adalah sebagai berikut:

1. Bagaimana nilai korelasi dari gabungan pendekatan *word alignment* TF-IDF dan vektor semantik menggunakan *word2vec* berdasarkan regresi *Support Vector Regression* (SVR) pada ayat Al-Quran terjemahan bahasa Inggris?

2. Bagaimana pengaruh penambahan PPDB pada fitur *word alignment* TF-IDF terhadap ayat Al-Quran terjemahan bahasa Inggris?

3. Bagaimana pengaruh fitur vektor semantik menggunakan *word2vec* terhadap ayat Al-Quran terjemahan bahasa Inggris?

4. Bagaimana hasil penggunaan metode *word alignment* TF-IDF pada ayat Al-Quran terjemahan bahasa Inggris jika dibandingkan dengan penelitian yang sudah ada?

1.3 Tujuan

Berdasarkan perumusan masalah yang telah diuraikan, maka tujuan yang diharapkan pada penelitian ini adalah:

1. Mengimplementasi dan menghitung nilai korelasi gabungan pendekatan *word alignment* TF-IDF dan vektor semantik menggunakan *word2vec* berdasarkan regresi *Support Vector Regression* (SVR) pada ayat Al-Quran terjemahan bahasa Inggris.

2. Menganalisa pengaruh penambahan PPDB pada fitur *word alignment* TF-IDF terhadap ayat Al-Quran terjemahan bahasa Inggris.

3. Menganalisa pengaruh fitur vektor semantik menggunakan *word2vec* terhadap ayat Al-Quran terjemahan bahasa Inggris.

4. Mengimplementasi dan menganalisa perbandingan hasil penggunaan metode *word alignment* TF-IDF pada ayat Al-Quran terjemahan bahasa Inggris dengan penelitian yang sudah ada.

[1] http://alt.qcri.org/semeval2016/task1/index.php?id=results

1.4 Metodologi Penelitian

Metodologi penelitian merupakan proses dalam mendapatkan data uang digunakan untuk keperluan ilmiah. Adapun metodologi yang dilakukan dalam menyelesaikan masalah adalah sebagai berikut:

1. Identifikasi Masalah
   Dilakukan metode studi literatur untuk melakukan identifikasi masalah. Studi literatur dilakukan untuk mencari informasi dan *knowledge* seputar konsep data, analisa dari topik yang diteliti melalui jurnal penelitian, buku dan referensi ilmiah.

2. Pengumpulan Data
   Pada tahapan ini dilakukan pengumpulan data pasangan ayat terjemahan bahasa Inggris untuk ditambahkan pada data pasangan ayat yang sudah dibangun pada penelitian tahun lalu. Tahapan ini dilakukan setelah selesai mengidentifikasi masalah dan penentuan metode yang tepat berdasarkan hasil dari studi literatur.

3. Permodelan dan Perancangan Sistem
   Permodelan dan perancangan sistem merupakan tahapan untuk memodelkan dan merancang sistem yang akan dibangun dengan tujuan memperoleh hasil yang terbaik.

4. Implementasi
   Tahap yang dilakukan untuk membangun sistem dengan metode *word alignment*, vektor semantik dan evaluasi menggunakan *support vector regression* berdasarkan pemodelan dan perancangan yang telah dibuat.

5. Pengujian dan Analisis Sistem
   Tahapan pengujian dilakukan setelah tahapan implementasi selesai dilaksanakan. Hal-hal yang menjadi landasan pengujian adalah parameter yang memiliki keterkaitan untuk menjawab tujuan penelitian. Pada tahapan ini juga dilakukan analisis hasil luaran dari sistem yang dibangun.

6. Kesimpulan
   Tahapan ini merupakan tahapan akhir dari penelitan. Pada tahapan ini dilakukan proses penarikan kesimpulan dan proses dokumentasi hasil penelitian kedalam bentuk laporan penelitan.

1.5 Sistematika Penulisan

Untuk mempermudah pemahaman pembaca, penulis membuat sistematika penulisan sebagai berikut:

1. Bab I Pendahuluan
   Pada bab ini akan dikemukakan latar belakang penggunaan pendekatan *word alignment* TF-IDF dan vektor semantik, tujuan penelitian, metodologi penelitian, dan sistematika penulisan.

2. Bab II Tinjauan Pustaka
   Pada bab ini menjelaskan teori-teori yang digunakan dalam penelitian dan penjelasan mengenai pendekatan yang digunakan seperti *word alignment*, TF-IDF, vektor semantik, *support vector regression* (SVR) dan teori pendukung penelitian lainnya.

3. Bab III Perancangan Sistem
   Bab ini merupakan bagian yang menjelaskan cara kerja sistem *word alignment*, vektor semantik dan SVR yang dibangun, data yang digunakan dan tahapan-tahapan yang digunakan untuk membangun sistem perhitungan nilai kesamaan semantik antara data pasangan ayat Al-Quran terjemahan bahasa Inggris.

4. Bab IV Evaluasi dan Pengujian
   Pada bab ini akan membahas dokumentasi dan analisis hasil nilai korelasi dari fitur yang digunakan.

5. Bab V Kesimpulan dan Saran
   Bab ini berisi tentang kesimpulan yang diperoleh dari penelitian yang telah dilakukan. Selain itu, bab ini juga berisikan saran untuk penyempurnaan dan pengembangan selanjutnya.