BAB I Pendahuluan

Pada bab ini menjelaskan latar belakang yang menjadi awal mula penelitian, rumusan masalah, tujuan dari penelitian serta metodologi penelitian yang digunakan dalam penelitian.

1.1 Latar Belakang

Semantic Textual Similarity (STS) adalah salah satu penelitian yang terkait dengan hubungan semantik antara dua konsep dalam penggunaan dan keterkaitannya [4]. Dalam konteks teks pendek merupakan salah satu penerapan teknologi penambangan teks dan *Natural Language Processing* (NLP). Identifikasi terhadap teks pendek merupakan permasalahan penelitian yang biasa diaplikasikan dalam bidang NLP (*Machine Translation, Text Summarization, Question Answering, Short Answer Scoring, Information Retrival*) [5].

Dalam bidang komputasi dan NLP terdapat kompetisi SemEval yang rutin diselenggarakan setiap tahunnya untuk mendorong peneliti agar menciptakan atau mengembangkan metode baru yang efektif dan efisien. Pada kompetisi tersebut terdapat berbagai *Task* yang dikelompokan berdasarkan bidangnya, salah satu *task* yang diperlombakan adalah STS. Pada tahun 2014 dan 2015 kompetisi SemEval bidang STS dimenangkan oleh tim Md Arafat Sultan dengan menggunakan beberapa metode. Pada tahun 2014 Sultan menggunakan metode *alignment* dan berhasil mendapatkan skor akurasi 0,859[1]. Sementara pada tahun 2015 digunakan metode tambahan yaitu vektor semantik dan memperoleh skor akurasi tertinggi 0,864 [2].

Alignment merupakan metode STS dengan mensejajarkan kata dalam kalimat. Kata dalam kalimat yang akan disejajarkan akan dilakukan identifikasi frasa dan kata yang masih mempunyai hubungan arti dan makna. Untuk mengatasi kekurangan pada *alignment* dalam mengidentifikasi kata *non*-frasa dibutuhkan vektor semantik. Vektor semantik merupakan metode yang digunakan untuk menghitung sebuah kata dari distribusi kata-kata di sekitarnya. Kata-kata ini umumnya diwakili sebagai vektor atau deret angka yang terkait dengan beberapa cara untuk diperhitungkan.

Data yang digunakan dalam penelitian STS SemEval adalah data dengan kalimat pendek sebagai contoh *headlines, answer-student, answer-forum,* dll. Kriteria data pada penelitian SemEval sama seperti data Alquran. Alquran merupakan kitab suci yang berisikan teks pedoman hidup umat Islam terdiri dari 30 Jus, 144 surat, dan 6236 ayat. Dari sekian banyak ayat memungkinkan beberapa diantaranya memiliki kesamaan arti. Untuk mempermudah mengartikan ayat dalam Alquran sudah banyak dibuat terjemahan dalam berbagai bahasa. Salah satunya adalah Alquran terjemahan bahasa Inggris.

[1] http://alt.qcri.org/semeval2014/index.php?id=evaluation-results
[2] http://alt.qcri.org/semeval2015/task2/index.php?id=results

Pada tugas akhir ini akan digunakan data terjemahan Alquran bahasa Inggris sebagai input untuk diukur nilai kesamaannya pada sistem STS dengan menggunakan pendekatan *word alignment* dan vektor semantik. Pendekatan *word alignment* dan vektor semantik dipilih karena metode tersebut merupakan salah satu metode yang cukup banyak digunakan pada salah satu kompetisi di bidang STS yaitu SemEval. Metode tersebut merupakan metode yang sederhana dan menunjukan performansi yang paling baik dalam salah satu *running* di *task* STS pada SemEval 2014 dan SemEval 2015. Untuk mengukur akurasi sistem STS yang akan dibangun pada penelitian ini dibutuhkan *gold standard* yang berisi penilaian kesamaan antar kalimat berdasarkan intuisi manusia. Hasil sistem kemudian dibandingkan dengan *gold standard* yang kemudian dihitung nilai korelasinya.

1.2 Perumusan Masalah

Terkait permasalahan yang dipaparkan diatas, dapat dirumuskan masalah yang diangkat dalam Tugas Akhir ini, yaitu:

1. Bagaimana hasil dari penggunaan penggabungan metode *word alignment* dan vektor semantik menggunakan *ridge regresi* pada data pasangan terjemahan ayat Alquran bahasa Inggris?

2. Bagaimana pengaruh metode *word alignment* serta ekstraksi fitur yang digunakan dalam mengukur nilai kesamaan semantik pada data pasangan terjemahan ayat Alquran bahasa Inggris?

3. Bagaimana pengaruh metode vektor semantik dalam mengukur nilai kesamaan semantik pada data pasangan terjemahan ayat Alquran bahasa Inggris?

1.3 Tujuan

Berdasarkan latar belakang dan rumusan masalah yang telah diuraikan diatas maka tujuan dari adanya penelitian ini adalah:

1. Mengimplementasi dan menganalisis penggabungan pendekatan *word alignment* dan vektor semantik menggunakan *ridge regresi* pada data pasangan potongan ayat Alquran.

2. Mengimplementasi dan menganalisis pengaruh metode *word alignment* dan ekstraksi fitur yang digunakan pada data pasangan potongan ayat Alquran terjemahan bahasa Inggris untuk mengetahui kesamaan semantik yang dihasilkan dan pengaruh pada nilai akurasi.

3. Mengimplementasi dan menganalisis pengaruh metode vektor semantik untuk mengetahui kesamaan semantik yang dihasilkan dan pengaruh pada nilai akurasi dengan menggunakan data pasangan potongan ayat Alquran terjemahan bahasa Inggris.

1.4 Metodologi Penelitian

Untuk menyelesaikan masalah dibutuhkan tahapan dalam pemecahan masalah tersebut. Metodologi yang dilakukan dalam menyelesaikan masalah adalah sebagai berikut:

1. Studi Literatur
   Dalam tahapan ini akan dilakukan pencarian informasi dan literatur terkait dengan data terjemahan Alquran bahasa Inggris dengan tafsir Ibnu Katsir yaitu dengan peelusuran website dan pencarian informasi mengenai metode *word alignment* serta vektor semantik yang akan diimplementasikan dalam sistem berdasarkan jurnal dan *paper* nasional dan internasional terutama pada kompetisi SemEval. Serta pencarian informasi mengenai pengukuran kesamaan semantik yang sesuai dengan menggunakan mengukur nilai korelasi. Hal ini dilakukan agar dapat menghasilkan performansi sistem yang optimal dan mencapai tujuan penelitian.

2. Pengumpulan Data
   Kemudian dilakukan pengumpulan data pasangan terjemahan Alquran untuk menambahkan data pasangan terjemahan Alquran yang sudah dibangun oleh penelitian sebelumnya sebanyak 400 pasangan ayat yang terbagi menjadi dua jenis yaitu 200 pasangan ayat tafsir Ibnu Katsir dan 200 pasangan ayat Indeks Tematik. Pada penelitian ini dilakukan penambahan 400 pasangan ayat yang terbagi menjadi dua jenis yaitu 350 pasangan ayat Ibnu Katsir dan 50 pasangan ayat Indeks Tematik. Kemudian dilakuakan penyebaran kuisioner pada beberapa anotator.

3. Perancangan Sistem
   Setelah mengumpulkan data kemudian akan dibuat suatu rancangan sistem yang sesuai untuk menyelesaikan permasalahan yang ada pada tugas akhir ini berdasarkan referensi yang sudah didapatkan dan dipelajari.

4. Pembuatan *Gold Standard*
   Data kuisioner kemudian diolah agar menghasilkan *gold standard*. *Gold Standard* yang dibuat pada penelitian ini dilakukan dengan penilaian kesamaan pasangan ayat berdasarkan intuisi manusia. Rentang nilai yang diberikan yaitu antara 0 hingga 5. Semakin besar nilai yang diberikan semakin tinggi tingkat kemiripannya. *Gold standard* berfungsi sebagai nilai yang menjadi tolak ukur pada nilai kesamaan semantik yang dihasilkan oleh sistem.

5. Implementasi
   Implementasi dilakukan untuk merealisasikan hasil perancangan sistem menggunakan penggabungan metode antara *word alignment* yaitu dengan menggunakan ekstraksi fitur *identical word*, PPDB, *sequence*, dan *name entity* dengan metode vektor semantik menggunakan *word2vec* terhadap dataset terjemahan Alquran bahasa Inggris yang dipergunakan. Metode tersebut dipilih berdasarkan sumber literatur yaitu kompetisi SemEval karena pada kom-

petisi tersebut kombinasi kedua fitur dapat menangani permasalahan dalam penelitian ini yaitu mengukur kesamaan semantik pada teks pendek.

6. Analisis Hasil
   Dari hasil implementasi yang dilakukan akan dilakukan analisis metode mana yang tepat sehingga menghasilkan performansi yang tinggi untuk data pasangan terjemahan Alquran. Untuk menghasilkan nilai performansi dilakukan pengukuran korelasi. Pengukuran korelasi pada gabungan kedua metode dilakukan dengan menggunakan regresi, hal ini dilakukan berdasarkan sumber literatur yang sudah didapatkan dan dipelajari.

7. Pembuatan Laporan
   Setelah melakukan implementasi dan analisis hasil yang didapatkan maka dilakukan kegiatan pendokumentasian hasil dari penelitian yang dilakukan dan melampirkan data pasangan terjemahan Alquran yang digunakan pada penelitian ini.

1.5 Sistematika Penulisan

Penulisan tugas akhir ini akan dibagi menjadi lima bab yang masing masing berisi penjelasan sebagai berikut:

1. Bab I Pendahuluan
   Pada bab ini dijelaskan secara singkat tentang latar belakang penggunaan metode *word alignment* dan vektor semantik pada penelitian serta data yang digunakan yaitu pasangan terjemahan ayat Alquran bahasa Inggris. Bab ini juga menjelaskan tentang perumusan masalah, tujuan dilakukannya penelitian, tahapan-tahapn yang dilakukan dalam penelitian ini, serta sistematika penulisan laporan penelitian.

2. Bab II Tinjauan Pustaka
   Bab ini menjelaskan teori-teori yang digunakan dalam penelitian dan penjelasan mengenai data yang digunakan dalam penelitian serta metode-metodenya, beberapa diantaranya yaitu *word alignment* beserta ekstraksi fitur yang digunakan, dan vektor semantik. Fitur-fitur didalam metode tersebut akan dijabarkan dalam subbab dari bab 2. Teori yang ada dalam bab ini dijadikan sebagai rujukan yang digunakan dalam penelitian ini.

3. Bab III Perancangan Sistem
   Pada bab ini akan dijelaskan bagaimana sistem dalam penelitian ini bekerja, data yang akan digunakan pada penelitian dan detail setiap tahap yang dilakukan untuk membangun sistem kesamaan semantik antar pasangan terjemahan Alquran menggunakan metode *word alignment* dan vektor semantik.

4. Bab IV Evaluasi dan Pengujian
   Bab ini akan membahas dokumentasi proses pengujian sistem dan analisis pengaruh fitur-fitur yang digunakan yaitu *word alignment* dan vektor semantik serta kombinasi antar keduanya pada penelitian ini terhadap hasil pengujian sistem menggunakan data pasangan terjemahan Alquran bahasa Inggris

5. Bab V Kesimpulan dan Saran
   Dalam bab ini akan dijelaskan kesimpulan dari hasil penelitian yang telah dilaksanakan berserta saran untuk penelitian yang akan dilanjutkan setelahnya.