

# BAB 1

## PENDAHULUAN

### 1.1 Latar Belakang

Manusia adalah makhluk sosial yang membutuhkan orang lain. Kehadiran orang lain di dalam kehidupan manusia dapat membantu menyelesaikan permasalahan yang muncul di dalam kehidupan sehari-hari, misalnya untuk menjawab pertanyaan yang membutuhkan opini atau pendapat orang lain.

Di era dengan kemajuan teknologi yang cepat seperti sekarang ini, seorang manusia dapat dengan mudah bersosialisasi atau berkomunikasi dengan orang lain. Di Indonesia sendiri, berdasarkan survey yang dilakukan oleh Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) pada tahun 2016, pengguna internet di Indonesia mencapai angka 132.7 juta pengguna, atau setara dengan 51.7% dari total penduduk [1]. Ini menunjukkan ketergantungan manusia terhadap informasi yang tersedia di internet sangat besar.

Hal-hal di atas memicu munculnya beberapa forum diskusi *online* atau forum tanya-jawab di internet. Situs-situs seperti ini memberikan keuntungan dimana kita bisa mendapat banyak masukan terhadap suatu permasalahan sehingga tidak terpaku pada hanya satu solusi saja. Tetapi sayangnya situs-situs ini juga memiliki kelemahan dimana banyak dijumpai jawaban-jawaban yang buruk atau yang tidak relevan dengan pertanyaan yang diajukan. Selain itu, penilaian sebuah jawaban adalah baik atau buruk diberikan oleh pengguna lain sehingga jawaban yang tidak secara langsung berhubungan dengan pertanyaan bisa memiliki *rating* yang tinggi, atau jawaban bisa saja disusupi dengan tautan yang cenderung bersifat *spam*.

Pada tugas akhir ini, penulis mencoba membuat sebuah aplikasi yang dapat memberikan peringkat terhadap jawaban-jawaban dari sebuah pertanyaan. *Dataset* yang digunakan untuk pengerjaan tugas akhir ini berasal dari *SemEval2017 Task 3* dan setiap jawaban telah memiliki kelas yang terdiri dari *Good*, *Potentially Useful*, dan *Bad*. Proses yang dilakukan untuk mendapatkan ranking jawaban berupa *preprocessing data*, *feature extraction*, klasifikasi, dan pemberian peringkat.

Tahapan *preprocessing data* yang dilakukan adalah *Tag Removal, Case Folding, Tokenizing, Stopword Removal, Part of Speech (POS) Tagging, dan Lemmatization*. *Feature extraction* yang digunakan adalah *textual feature* dan *semantic similarity feature*. Sedangkan pada proses klasifikasi akan menggunakan algoritma *Support Vector Machine (SVM)* sebagai *classifier*. Untuk evaluasi performansi sistem, parameter yang digunakan sama seperti pada *SemEval2017* yaitu *Mean Average Precision (MAP)*.

Berdasarkan penelitian sebelumnya, kombinasi *Semantic Similarity* dengan *Lexical Similarity* berhasil mendapatkan skor MAP 70.96 [2] dan kombinasi *Textual Feature* dengan *Topic Modeling* berhasil mendapatkan skor MAP 71.8 [3]. Sementara itu, tim *Kelp* yang memenangkan *SemEval2016* menggunakan SVM sebagai *classifier* dan mendapatkan skor MAP 79.19 [4]

Oleh karena itu, penelitian ini akan menggunakan *Semantic Similarity* untuk mempertahankan kesamaan makna kalimat, dan juga *Textual Feature* untuk mendapatkan ciri khas dari sebuah jawaban yang baik atau buruk. Sementara pada proses klasifikasi, akan menggunakan SVM karena pada tiga penelitian yang dijadikan acuan, SVM memiliki performansi yang cukup baik.

## **1.2 Perumusan Masalah**

Beberapa masalah yang dibahas pada tugas akhir ini adalah sebagai berikut:

1. Bagaimana membuat sebuah sistem yang dapat membantu memilih mana jawaban yang baik terhadap pertanyaan dan melakukan perankingan terhadap jawaban tersebut.
2. Bagaimana pengaruh *textual feature* dan *semantic similarity feature* dalam mengetahui jawaban yang baik untuk pertanyaan yang diajukan.

Untuk menghindari agar bahasan pada tugas akhir ini tidak terlalu melebar, maka diambil beberapa batasan masalah yaitu:

1. Bahasa yang digunakan di dalam dataset adalah Bahasa Inggris.
2. Jawaban yang terdapat pada dataset hanya memiliki 3 kelas yaitu *good, potentially useful, dan bad*.
3. Proses pembangunan sistem dilakukan dengan bantuan *tools* atau *library* yang sudah tersedia.
4. Sistem dibangun dengan bahasa pemrograman *Python* dan *PHP*

5. Sistem yang dibangun hanya melakukan perangkingan ulang terhadap pasangan pertanyaan-jawaban yang terdapat di dalam dataset. Pengguna sistem tidak memiliki akses untuk memasukkan pertanyaan baru.

### 1.3 Tujuan

Adapun tujuan dari tugas akhir ini adalah sebagai berikut:

1. Membuat sebuah sistem yang dapat memberikan ranking untuk jawaban dari suatu pertanyaan agar bisa menemukan jawaban yang terbaik
2. Melakukan analisis terhadap pengaruh *textual feature* dan *semantic similarity* dalam menghasilkan jawaban terbaik dari suatu pertanyaan

### 1.4 Metodologi Penyelesaian Masalah

Tahapan-tahapan yang dilakukan untuk menyelesaikan tugas akhir ini adalah:

#### 1. Studi Literatur

Tahap ini bertujuan untuk mempelajari seluruh konsep yang berkaitan dengan sistem yang akan dibangun. Beberapa tahap yang akan dibahas diantaranya adalah mengenai ekstraksi fitur pada kalimat, memahami konsep SVM sebagai *classifier*, dan mempelajari *dataset* yang akan digunakan, Sumber-sumber referensi didapatkan dari buku, jurnal ilmiah, atau artikel di internet.

#### 2. Pengumpulan Data

Dataset yang digunakan pada tugas akhir ini telah tersedia di halaman resmi *SemEval* 2017.

#### 3. Perancangan Aplikasi

Pada tahap ini akan ditentukan *requirement* yang dibutuhkan untuk pembuatan sistem yang akan digunakan. *Requirement* yang sudah dibuat kemudian akan dijadikan sebagai acuan dalam membuat desain sistem.

#### 4. Implementasi dan Pengujian

Mengimplementasikan desain yang telah dibuat pada tahap sebelumnya ke dalam bentuk program. Sistem yang akan dibangun harus mampu melakukan perankingan jawaban untuk dapat memberikan jawaban terbaik dari suatu pertanyaan. Sistem yang sudah selesai dibangun kemudian akan diuji untuk memastikan sistem tersebut layak untuk digunakan.

#### 5. Analisis dan Pembuatan Laporan

Melakukan evaluasi terhadap *output* dari sistem yang dibangun. Evaluasi dilakukan dengan melihat nilai MAP yang dihasilkan, kemudian dilakukan proses analisis untuk mengetahui performansi sistem secara keseluruhan.

## **1.5 Sistematika Penulisan**

Laporan tugas akhir ini terdiri dari lima bab dan setiap babnya memiliki beberapa sub-bab. Adapun penjelasannya adalah sebagai berikut:

### **1. Pendahuluan**

Bab ini berisikan mengenai latar belakang permasalahan, metode atau algoritma yang digunakan, rumusan dan batasan masalah, tujuan penelitian, dan metodologi yang digunakan di dalam penelitian.

### **2. Tinjauan Pustaka**

Bab ini menjelaskan mengenai literatur studi yang berkaitan dengan penelitian yang dilakukan., seperti penjelasan mengenai *Community Question Answering*, penjelasan mengenai *dataset* yang digunakan, tahapan *preprocessing* data yang dilakukan, algoritma klasifikasi yang digunakan, dan penjelasan mengenai cara evaluasi sistem.

### **3. Perancangan Sistem**

Bab ini menjelaskan tahapan-tahapan yang dilakukan saat membangun sistem yang digunakan di dalam penelitian. Penjelasan dapat dilakukan menggunakan poin-poin atau diagram.

### **4. Evaluasi Sistem**

Bab ini menjelaskan mengenai tahapan dan hasil pengujian setelah sistem selesai dibangun. Setelah pengujian dilakukan, akan dilakukan analisis terhadap hasil pengujian.

### **5. Kesimpulan dan Saran**

Bab ini berisi kesimpulan terhadap hasil yang didapat dari pengujian sistem, serta saran untuk pengembangan sistem lebih lanjut.