# BAB I Pendahuluan

Bab I berisi penjelasan mengenai latar belakang digunakannya topik pemberian peringkat pada jawaban sebagai topik penelitian ini, rumusan masalah yang dibahas pada penelitian , tujuan yang ingin dicapai dari penelitian ini serta metodologi yang digunakan dalam menyelesaikan penelitian.

## 1.1 Latar Belakang

Setiap orang pernah merasakan kebingungan dan ketidaktahuan untuk menjawab pertanyaan yang diajukan kepada mereka, baik pertanyaan di bidang akademik maupun di bidang non akademik. Terkadang dibutuhkan opini, pendapat maupun jawaban dari orang lain untuk membantu menjawab pertanyaan tersebut. Seiring berkembangnya kemajuan teknologi dan banyaknya pengguna internet dimana pengguna internet pada tahun 2015 berdasarkan statistik dari *The Statistic Portal* bahwa pengguna internet mencapai 3,17 miliar orang[1], mulailah terbentuk banyak forum diskusi online yang dengan nama lain disebut komunitas tanya-jawab (*Community Question Answering*). Setiap orang dapat bertanya maupun menjawab pada komunitas ini. Sisi positifnya banyak pilihan jawaban yang dapat diambil, namun sisi negatif dari CQA ini terkadang jawaban yang diberikan tidak sesuai atau tidak ada kaitannya dengan pertanyaan, untuk itu dibutuhkan usaha yang lebih untuk mencari jawaban yang sesuai. Banyak kasus dimana jawaban yang diberikan tidak ada kaitannya dengan pertanyaan bahkan berbeda topik pembicaraan, maka dari itu diperlukan suatu sistem yang dapat membantu memberikan peringkat pada jawaban yang paling mendekati pertanyaan yang diajukan.

Pada penelitian ini akan dibuat suatu aplikasi yang dapat memberikan peringkat pada jawaban-jawaban terhadap suatu pertanyaan. Pemberian peringkat berdasarkan tingkat kemiripan antara jawaban dan pertanyaan (Question-Comment Similarity). Untuk melihat tingkat kemiripannya, setiap jawaban dari pertanyaan akan diberikan bobot masing-masing. Bobot tersebut didapatkan dari proses ekstraksi fitur untuk setiap jawabannya. Dari bobot tersebut akan dicari nilai untuk setiap pasangan pertanyaan dan jawaban yang akan diurutkan untuk menentukan peringkat dari setiap jawaban. Dataset yang digunakan dalam penelitian ini adalah dataset pada *Qatar Living Forum* yang bersumber dari *SemEval 2016*. Pada dataset setiap jawaban dari pertanyaan sudah memiliki kelas masing-masing (*good, potentially useful* dan *bad*). Kelas tersebut diberikan berdasarkan pilihan dari pengguna *Qatar Living Website Forum*. Penelitian ini juga dilakukan untuk merangking ulang jawaban dari user tersebut.

Penelitian ini merupakan pengembangan dari penelitian pada *SemEval 2015 Task 3 Subtask A* mengenai pengklasifikasian jawaban pada *Community Question Answering*. Pada penelitian tersebut dibuat suatu sistem untuk

[1]http://www.statista.com/statistics/273018/number-of-internet-users-worldwide/. (Diakses tanggal 21 Oktober 2015)

memberikan label pada jawaban(*good, bad* dan *potentially useful*), sedangkan penelitian ini memberikan peringkat pada jawaban yang sudah memiliki label tersebut, dimana jawaban dengan kategori *good* berada diatas kategori *bad* dan *potentially useful*. Proses yang akan dilakukan pada penelitian ini meliputi proses *preprocessing* data, pembobotan jawaban menggunakan ekstraksi fitur, proses klasifikasi dan proses pemberian peringkat pada jawaban.

Ekstraksi fitur pada penelitian ini terdiri dari *lexical similarity feature, semantic similarity feature* dan *non textual feature grup* dan *Heuristic*. Namun penelitian ini lebih ditekankan pada *similarity measure feature*. *Similarity measure feature* mengukur kesamaan kata antara pertanyaan dan jawaban baik kesamaan berdasarkan struktur kata (*lexical similarity*) dan makna kata (*semantic similarity*). Pemilihan fitur-fitur yang digunakan pada penelitian ini didasari oleh pemilihan fitur yang dilakukan oleh tim JAIST [4] dan tim QCRI [5] pada penelitian SemEval 2015 task 3. *Lexical similarity, semantic similarity* dan *heuristic* merupakan fitur yang digunakan tim QCRI, dimana nilai lexical similarity memberikan kontribusi yang cukup tinggi dalam mengklasifikasikan jawaban. Sedangkan *non textual feature group* dan *semantic similarity* juga merupakan fitur-fitur yang digunakan tim jaist yang membantu mengelompokkan jawaban untuk setiap pertanyaan.

Pada proses perangkingan jawaban akan digunakan *Support Vector Machine* (SVM)dan *Logistic Regression* sebagai *classifier* dan penentuan peringkat untuk setiap pasangan pertanyaan dan jawabannya. Pemilihan SVM sebagai *classifier* dikarenakan berdasarkan penelitian sebelumnya pada *SemEval 2015* [1],tingkat akurasi dengan menggunakan SVM merupakan tingkat akurasi tertinggi yaitu 72% [1].

## 1.2 Rumusan Masalah

Adapun rumusan masalah yang dibahas adalah sebagai berikut.

1. Bagaimana pengaruh kombinasi *lexical* dan *semantic similarity feature* dalam mendeteksi kesamaan pertanyaan dan jawaban?

2. Bagaimana cara mengetahui fitur atau gabungan fitur yang sesuai untuk mengukur kesamaan kata antara pertanyaan dan jawaban?

3. Bagaimana perbandingan *classifier* SVM dan *Logistic Regression* dalam mengkategorikan jawaban?

Sedangkan pada penelitian ini terdapat beberapa batasan masalah untuk menghindari meluasnya materi pembahasan. Adapun lingkup batasan masalahnya mencakup hal-hal berikut.

1. Dataset pada penelitian ini hanya pada *Qatar Living Website Forum*.

2. Bahasa yang digunakan yaitu bahasa Inggris pada data *Qatar Living Website Forum*.

3. Jawaban sudah memiliki kategori yang terbagi kedalam tiga kategori yaitu *good, potentially useful,* dan *bad*.

4. Jawaban yang dirangking berdasarkan jawaban dari *user* pada *Qatar Living Website Forum*.

5. Proses klasifikasi dilakukan dengan bantuan *tools* klasifikasi yang sudah tersedia.

## 1.3 Tujuan

Tujuan dari penelitian ini adalah sebagai berikut:

1. Melakukan analisis terhadap keterkaitan kesamaan kata(*lexical similarity*) dan makna(*semantic similarity*) dalam pemberian peringkat pada jawaban.

2. Melakukan analisis terhadap fitur dan gabungan fitur yang menghasilkan peringkat jawaban terbaik yang sesuai dengan pertanyaan yang diajukan.

3. Melakukan analisis terhadap *classifier* yang menghasilkan pengklasifikasian dan perangkingan jawaban yang terbaik.

## 1.4 Metodologi Penyelesaian Masalah

Adapun tahapan metode yang digunakan pada penelitian ini adalah sebagai berikut:

1. Studi literatur
   Tahapan ini bertujuan untuk mempelajari konsep pembobotan antar kalimat dari suatu dokumen dan mempelajari metode-metode pembobotan kata dan kalimat. Pada tahapan ini juga mempelajari dataset yang ada dan tipe-tipe dari data tersebut.

2. *Software Requirement Analysis*
   Proses ini merupakan tahapan pengumpulan *requirement* yang dibutuhkan dalam pengembangan *software*. *Requirement* ini dapat membantu sebagai acuan dalam tahap implementasi program aplikasi.

3. *Design*
   Proses ini merupakan proses menterjemahkan *requirement* menjadi representasi dari aplikasi sebelum tahapan implementasi dibuat. Pada tahapan ini menggambarkan perancangan perilaku dan data dari aplikasi yang dibatasi oleh batasan perancangan.

4. Implementasi program aplikasi
   Merancang suatu aplikasi yang dapat mengidentifikasi bobot setiap jawaban untuk memberikan jawaban terbaik sehingga dapat memudahkan pengguna *Community Question Answering*(CQA) dalam menentukan

pilihan jawaban. Implementasi dilaksanakan berdasarkan *design* sistem yang sudah dibuat sebelumnya.

5. Pengujian aplikasi
   Dilakukan pengujian terhadap aplikasi yang telah dirancang dan memastikan kelayakan aplikasi tersebut.

6. Analisis hasil pengujian
   Proses ini merupakan proses evaluasi terhadap aplikasi yang telah dibuat dengan pengujian terhadap beberapa sampel yang hasilnya akan dianalisis tingkat akurasinya.

7. Penyusunan dan pembuatan laporan
   Tahapan ini merupakan proses penyusunan laporan hasil penelitian dimulai dari perumusan *requirement* aplikasi,*design*, implementasi, pengujian dan hasil analisis.

## 1.5 Sistematika Penulisan

Secara garis besar penulisan laporan akhir penelitian ini terdiri dari lima bab yang terdiri dari beberapa subbab. Adapun penjelasan terhadap sistematika penulisan sebagai berikut:

BAB I PENDAHULUAN

Pada bab ini menjelaskan mengenai latar belakang pemilihan topik penelitian, perumusan dan pembatasan masalah, tujuan penelitian, metodologi yang digunakan dalam menyelesaikan masalah dan sistematika penulisan laporan penelitian.

BAB II TINJAUAN PUSTAKA

Bab ini membahas mengenai *Community Question Answering*, penjelasan dataset , konsep *preprocessing* dan penjelasan ekstraksi fitur yang digunakan, serta konsep klasifikasi dan *tools* yang digunakan untuk klasifikasi. Selain itu pada bab ini juga dijelaskan proses evaluasi performansi sistem.

BAB III PERANCANGAN SISTEM

Pada bab ini dijelaskan mengenai tahapan-tahapan yang dilakukan pada saat membangun sistem pada penelitian ini. Tahap yang dilalui seperti *preprocessing* data, ekstraksi fitur, klasifikasi dan pemberian peringkat serta evaluasi performansi sistem.

BAB IV Evaluasi

Bab ini menjelaskan mengenai pengujian dan analisis yang dilakukan terhadap hasil dari sistem yang dibuat. Analisis dilakukan terhadap hasil ekstraksi fitur dan *classifier* yang digunakan.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan-kesimpulan yang didapat dari hasil penelitian dan juga saran terhadap pengembangan selanjutnya untuk sistem yang dibuat.