

# BAB I

## Pendahuluan

Bab I megandung penjelasan latar belakang diangkatnya topik pemberian peringkat pada *community question answering*, rumusan masalah yang ingin diselesaikan, tujuan yang ingin diraih dan metodologi yang digunakan dalam penelitian.

### 1.1 Latar Belakang

Pengguna yang mencari informasi melalui forum terkadang dihadapkan dengan kondisi berupa pertanyaan yang dimiliki tidak terdapat pada forum tersebut. Untuk memperoleh jawaban, pengguna perlu mengajukan pertanyaan baru kepada forum dan menunggu komentar. Namun, terdapat kemungkinan pertanyaan yang serupa sudah pernah diajukan sebelumnya dan telah memiliki jawaban. Hal ini kerap dijumpai pada forum dan menghasilkan banyaknya duplikasi pertanyaan. Contoh pada StackOverflow pertanyaan dengan predikat duplikat<sup>1</sup> memiliki keterangan berupa “marked as duplicate by ...”. Saat pengguna lain mencari informasi serupa, ia perlu menjelajah sebagian atau semua pertanyaan duplikat untuk mencari jawaban. Jumlah waktu yang diperlukan pengguna juga akan bertambah. Oleh karena itu, dibutuhkan sistem pemeringkatan komentar secara otomatis oleh sistem *Community Question Answering*.

Pada tugas akhir ini dikembangkan sistem yang mampu memperoleh fitur-fitur tertentu dan menentukan peringkat dari setiap komentar. Dengan sistem pemberian peringkat, diharapkan waktu yang diperlukan pengguna untuk mencari solusi dapat dikurangi. Sistem akan mengolah sekumpulan pertanyaan dan komentar dalam format XML. Data yang digunakan adalah data dalam Bahasa Inggris dari forum Qatar Living yang disediakan oleh SemEval-2016 Task 3<sup>2</sup> dengan studi kasus *Question-External Comment Similarity*. *Question-External Comment Similarity* merupakan kasus dalam menemukan kemiripan antara suatu pertanyaan baru dengan komentar dari pertanyaan lain.

Penelitian yang ada sebelumnya melibatkan klasifikasi komentar. Kekurangan dari proses klasifikasi saja adalah terpencarnya komentar-komentar relevan dan pengguna perlu menjelajah seluruhnya. Untuk itu, pemberian peringkat dapat lebih membantu dengan memberikan komentar-komentar relevan pada posisi teratas pertanyaan. Peringkat diperoleh dengan menerapkan beberapa tahapan. Pertama adalah praproses, dengan tujuan mengubah dokumen ke dalam bentuk yang lebih mudah untuk diolah sistem [4]. Kedua

---

<sup>1</sup><http://stackoverflow.com/questions/1996747/add-new-value-to-an-existing-array-in-javascript>

<sup>2</sup><http://alt.qcri.org/semeval2016/task3>

adalah ekstraksi fitur. Fitur yang akan diambil berupa fitur leksikal dan semantik dengan kelompok fitur *word matching* dan *similarity*. Ketiga memanfaatkan *classifier* untuk memperoleh nilai prediksi komentar terhadap kelas tertentu. Pengurutan didasarkan kepada nilai prediksi terhadap kelas *Good* pada komentar.

## 1.2 Perumusan Masalah

Berdasarkan latar belakang, diperoleh perumusan masalah yang dipaparkan:

1. Bagaimana mengidentifikasi jawaban yang relevan?
2. Bagaimana cara menemukan fitur yang mampu mewakili keterhubungan pertanyaan dan komentar?
3. Bagaimana melakukan pemberian peringkat pada komentar?

Selain perumusan masalah, ditetapkan batasan masalah yang dipaparkan berupa:

1. Sistem menggunakan dataset *Qatar Living Forum* yang berasal dari SemEval-2016 Task 3.
2. Jenis fitur yang diambil hanya fitur leksikal dan semantik.
3. Proses klasifikasi dilakukan dengan bantuan *tools* yang telah tersedia.

## 1.3 Tujuan

Berdasarkan perumusan masalah, diperoleh tujuan yang dipaparkan:

1. Mengambil fitur yang tepat untuk mengidentifikasi jawaban yang relevan.
2. Membandingkan fitur yang memiliki kontribusi besar terhadap performa.
3. Melakukan klasifikasi komentar dan memberikan peringkat sesuai hasil klasifikasi.

## 1.4 Metodologi Penyelesaian Masalah

Metodologi berupa tahapan untuk menyelesaikan tugas akhir:

1. Studi Literatur  
Pencarian dan pemahaman mengenai materi yang berhubungan dengan pengerjaan tugas akhir. Materi yang dimaksud mengenai bidang keilmuan yang dipelajari, praproses, ekstraksi fitur dan pemberian peringkat.
2. Pengumpulan Data dan Perancangan Sistem  
Pencarian *dataset* yang berasal dari SemEval-2016 berupa kumpulan pertanyaan dan komentar dalam format XML. Secara umum sistem akan dirancang seperti yang dijelaskan pada Bab III.

### 3. Implementasi Sistem

Pada tahap ini diimplementasikannya rancangan dimana sistem mampu melakukan praproses, *feature extraction* dan penghitungan nilai akhir dengan memanfaatkan *Logistic Regression*.

### 4. Pengujian dan Analisis

Pada tahap ini dilakukan pengujian terhadap sistem dan analisis terhadap hasil pengujian. Nilai *Mean Average Precision* yang dihasilkan oleh sistem akan dihitung pada tahap ini.

### 5. Pembuatan Laporan

Pada tahap ini dilakukan pembuatan laporan berdasarkan hasil pengujian dan analisis.

## 1.5 Sistematika Penulisan

Secara umum penulisan laporan penelitian ini terdiri dari lima bab. Penjelasan mengenai setiap bab sebagai berikut:

#### 1. BAB I Pendahuluan

Menjelaskan mengenai latar belakang pemilihan topik penelitian, perumusan masalah, tujuan, metodologi yang digunakan dalam penyelesaian masalah dan sistematika penulisan laporan.

#### 2. BAB II Tinjauan Pustaka

Membahas *question answering*, jenis-jenis kasus pada *community question answering*, penelitian terkait, *dataset* yang digunakan, konsep praproses, penjelasan ekstraksi fitur, klasifikasi, pemberian peringkat, evaluasi performansi sistem dan *tools* yang digunakan untuk membantu penelitian.

#### 3. BAB III Perancangan Sistem

Menjelaskan rancangan mengenai sistem yang dibangun beserta tahapannya. Tahapan terdiri dari pengaplikasian praproses, ekstraksi fitur dan evaluasi serta pengembangan dibandingkan penelitian terdahulu.

#### 4. BAB IV Evaluasi

Menjabarkan secara singkat mengenai *dataset* yang digunakan, *baseline*, *evaluation metric* yang digunakan dan hasil pengujian. Analisis diterapkan berdasarkan hasil ekstraksi fitur dan konfigurasi *classifier* yang digunakan.

#### 5. BAB V Kesimpulan dan Saran

Berisi kesimpulan terhadap pengaruh ekstraksi fitur dan konfigurasi *classifier* terhadap nilai MAP yang dihasilkan. Saran sebagai masukan penelitian lebih lanjut juga disertakan pada bab ini.