

Proyek Akhir Kuliah
CSC4602354: Pengolahan Bahasa Manusia – Genap 2017/2018

Aspect-Based Restaurant Review Sentiment Analysis

Sub task Anotasi

Deadline: Minggu, 22 April 2018, jam 22:00

Sub task Sentiment Prediction

Progress report: Jumat, 4 Mei 2018, jam 22:00

System deadline: Selasa, 15 Mei 2018, jam 22:00

Final report deadline: (tentatif) Minggu, 21 Mei 2018, jam 22.00

A. Deskripsi Tugas

Pada proyek akhir kuliah NLP ini, Anda diminta untuk bereksperimen dalam mengembangkan sebuah model untuk menyelesaikan *Aspect-Based Restaurant Review Sentiment Analysis*. Diberikan *user review* terhadap suatu restoran, Anda diminta mengekstrak informasi terstruktur yang memuat sentimen untuk setiap aspek.

Contoh input

```
<review rid="51025">
  <text>
    Untuk harga yg cukup mahal, menurut saya rasa makanan disini kurang enak.. Kurang
    berasa sih bumbunya.. Jd ya saya cukup skali aja nyoba disini.. Ga bkin pengen balik lagi..
  </text>
</review>
```

Contoh output

```
<review rid="51025">
  <text>
    Untuk harga yg cukup mahal, menurut saya rasa makanan disini kurang enak.. Kurang
    berasa sih bumbunya.. Jd ya saya cukup skali aja nyoba disini.. Ga bkin pengen balik lagi..
  </text>
  <aspects>
    <aspect category="FOOD" polarity="NEGATIVE"/>
    <aspect category="PRICE" polarity="NEGATIVE"/>
  </aspects>
</review>
```

Bentuklah kelompok dengan anggota maksimal 3 orang. Daftarkan anggota kelompok pada forum SceLe yang telah disediakan dengan format sebagai berikut.

Nama Kelompok:

Anggota Kelompok: 1. Nama anggota – e-mail
2. Nama anggota – e-mail
3. Nama anggota – e-mail

B. Tahapan Tugas

Pengerjaan tugas ini dibagi menjadi empat tahap pengumpulan. Pengumpulan tahap 1 merupakan tugas individu. Pengumpulan tahap 2, 3, dan 4 merupakan tugas kelompok.

Tahap 1 (Anotasi Data)

- *Deadline*: Minggu, 22 April 2018, pukul 22.00
- Tugas ini dikerjakan secara individu. Setiap peserta akan menerima 200 teks *review* dan diminta menganotasi masing-masing *review* tersebut.
- Panduan dan contoh anotasi disediakan di SceLe.

Tahap 2 (Progress Report)

- *Deadline*: Jumat, 4 Mei 2018, pukul 22.00
- Setiap kelompok diminta mengumpulkan laporan kemajuan. Idealnya, di tahap ini, mahasiswa sudah melakukan studi literatur (landasan teori dan tinjauan penelitian terkait), observasi data, eksplorasi *tool*, dan rencana metode (menentukan pendekatan yang akan dikembangkan, aturan/fitur yang akan digunakan)
- Tugas cukup dikumpulkan oleh salah seorang perwakilan anggota kelompok.
- Dalam rentang waktu sepekan sejak *deadline progress report*, setiap kelompok diminta mempresentasikan kemajuan proyek di hadapan asisten. Format presentasi dapat berupa slide atau demo *progress*. Detail dan jadwal demo akan diumumkan di SceLe.

Tahap 3 (System Deadline dan Online Evaluation)

- *Deadline*: Selasa, 15 Mei 2018, pukul 22.00
- Setiap kelompok mengimplementasikan model / program yang mengekstrak aspek dan memprediksi sentimen dari teks *review*.
- Akan disediakan sebuah *online evaluation system*, di mana setiap kelompok harus *submit* minimal sekali *output* (dalam format xml) yang dihasilkan oleh model / program *sentiment prediction* yang sudah diimplementasikan. *Output* yang dikumpulkan harus merupakan hasil luaran program, bukan rekayasa (*manual tagging*). Kelompok yang melakukan kecurangan dalam *submission output* akan memperoleh nilai 0 untuk proyek akhir ini.

Online evaluation system akan mengevaluasi *output*. Peringkat kelompok akan dipublikasikan dalam *system leaderboard*. Mekanisme *submission* ke *online evaluation system* akan diinformasikan lebih lengkap via SceLe.

- Pada *deadline* tahap 3 ini, setiap kelompok juga diminta mengumpulkan *submission* ke SceLe berupa: (1) laporan akhir versi beta, (2) *source code* program.
- 2 – 3 kelompok terbaik akan dipilih oleh dosen dan tim asisten untuk mempresentasikan proyek akhirnya di sesi terakhir kelas NLP.

Tahap 4 (Final Report)

- *Deadline* (tentatif): Minggu, 21 Mei 2018, pukul 22.00
- Setiap kelompok diminta mengumpulkan *submission* ke SceLe berupa: (1) laporan akhir versi final, (2) *source code* yang dilengkapi dengan dokumentasi.

C. Data

Data yang disediakan untuk proyek akhir dibagi menjadi *training*, *testing*, dan *validation set*.

Training data terdiri dari lebih kurang 4.000 *review* yang dianotasi oleh mahasiswa dan akan tersedia di SceLe sejak tanggal 27 April 2018.

Testing data terdiri dari 2.000 *review* tanpa anotasi yang di antaranya terdapat 150 *review gold-standard* yang akan digunakan sebagai evaluasi model. Selain itu, terdapat 50 data *validation* yang telah dianotasi dan dijadikan sebagai *gold standard* pada tugas *sentiment analysis* kali ini. *Testing* dan *validation data* merupakan data yang telah dianotasi dan digunakan pada penelitian *aspect-based sentiment analysis* sebelumnya.

D. Evaluasi

Metrik evaluasi¹ yang digunakan oleh *online evaluation system* adalah sebagai berikut.

- Precision, Recall, F-1 Score: **Aspect-Sentiment Pair**
- Precision, Recall, F-1 Score: **Aspect Only**

Misalkan *gold standard* merujuk pada anotasi “id = 51025”

Sistem X memprediksi jawaban sebagai berikut

```
<aspects>
  <aspect category="FOOD" polarity="POSITIVE"/>
  <aspect category="PRICE" polarity="NEGATIVE"/>
  <aspect category="AMBIENCE" polarity="NEGATIVE"/>
</aspects>
```

¹ Peringkat kelompok akan diurutkan berdasarkan nilai metrik tersebut

Untuk penilaian **aspect-sentiment pair**, sistem X memiliki 1 True Positive, 2 False Positive, dan 1 False Negative.

Untuk penilaian **aspect-only**, sistem X memiliki 2 True Positive dan 1 False Positive

Setelah masa *online evaluation* selesai, *gold-standard label* dapat diakses oleh mahasiswa. Dalam laporan final, mahasiswa diharapkan melakukan evaluasi sistem masing-masing dengan lebih komprehensif. Misalnya, melakukan *error analysis* atau mengevaluasi sistem dengan lebih detail (*online evaluator* hanya memberikan hasil evaluasi pada level agregat).

E. Deliverable Tugas

1. Laporan Akhir

Laporan dikumpulkan dalam format .pdf

Laporan akhir (yang ideal) sekurang-kurangnya memuat konten² berikut:

- Latar belakang
Jelaskan permasalahan *aspect-based sentiment analysis* dengan bahasa Anda sendiri
- Studi literatur
Bahas penelitian-penelitian terkait yang sudah dilakukan sebelumnya
- Metodologi
Jelaskan tahapan pelaksanaan penelitian. Jelaskan juga metode / pendekatan yang digunakan. Jika menggunakan pendekatan *rule based*, aturan yang dikembangkan harus disertakan dalam laporan. Jika menggunakan pendekatan *machine learning*, jelaskan fitur yang digunakan dan proses untuk mengekstrak fitur.
- Eksperimen dan hasil
Jelaskan rancangan scenario eksperimen. Jelaskan hasil evaluasi dari eksperimen yang dilakukan.
- Kesimpulan
Simpulkan penyelesaian yang sudah dikerjakan dalam tugas ini, serta daftarkan saran implementasi / penyempurnaan solusi penyelesaian di penelitian selanjutnya.
- Daftar pustaka

2. Source Code dan/atau Model

Seluruh kode, *script*, dan model yang diimplementasikan untuk menyelesaikan tugas harus dikumpulkan (termasuk kode untuk tahap *pre* maupun *post processing*). Lengkapi kode dengan dokumentasi.

Sertakan juga *README file* yang menjelaskan cara menjalankan program.

² Judul bab / sub bab pada laporan silakan disesuaikan

F. Komponen Penilaian

Detail penilaian proyek akhir adalah sebagai berikut

- 10%: anotasi
- 15%: *online evaluation*
Semakin tinggi F-1 *score* yang diperoleh, nilai komponen ini akan semakin tinggi. Kelompok dengan peringkat lebih baik memiliki nilai yang lebih tinggi.
- 75%: *complete deliverables*,
Komponen ini dinilai dari *progress report* dan laporan final, dijabarkan sebagai berikut
 - 10%: studi literatur
 - 25%: pendekatan, metode, dan metodologi
 - 15%: eksperimen, evaluasi, dan analisis
 - 10%: kode, *script*, dan model (beserta dokumentasi)
 - 10%: penulisan laporan
 - 5%: *project management* (kerjasama tim, responsif terhadap *feedback*, dan aspek non teknis lainnya)
- (Bonus) 5 – 10%: presentasi (untuk kelompok terpilih)

Nilai proyek akhir berkontribusi 30% dari komponen penilaian akhir perkuliahan NLP.

G. Rujukan

Terdapat sejumlah pendekatan yang dapat Anda pilih untuk mengerjakan tugas, yaitu berdasarkan aturan (*rule-based*), berdasarkan pemelajaran mesin (*machine learning*), atau gabungan *rule-based* dan *machine learning* (*hybrid*). Pendekatan *machine learning* juga bermacam-macam, di antaranya *supervised*, *unsupervised*, dan *semi-supervised*.

Beberapa *tools* / *library* / *resource* yang dapat digunakan antara lain:

- Python NLTK (<https://www.nltk.org/>)
Dapat digunakan untuk melakukan tokenisasi, menghitung *n*-gram, dsb
- Stanford NLP (<https://nlp.stanford.edu/>)
- Mallet (<http://mallet.cs.umass.edu/>)
- POS Tagger
Bahasa **Indonesia** : <http://bahasa.cs.ui.ac.id/postag/corpus> ,
Bahasa **Inggris** : <http://www.corpora4learning.net/resources/corpora.html>)
- *Dependency parsing*
Bahasa **Inggris**: <http://demo.ark.cs.cmu.edu/parse>

- *Word Embedding models*
 - *Fasttext* (Bahasa Inggris: <https://fasttext.cc/>)
 - *Pretrained Word Vectors* oleh Facebook tersedia dalam Bahasa Indonesia dan Inggris
<https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md>
 - *Word2Vec* (<https://radimrehurek.com/gensim/models/word2vec.html>)
pretrained word/phrase vectors Google News (Bahasa Inggris)
<https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing>
 - *Glove* (<https://nlp.stanford.edu/projects/glove/>)
Wikipedia 2014 + Gigaword 5 (6B tokens, 400K vocab, uncased, 50d, 100d, 200d, & 300d vectors, 822 MB download) (Bahasa Inggris)
<http://nlp.stanford.edu/data/glove.6B.zip>
 - *Word2Vec Library* (<https://radimrehurek.com/gensim/models/word2vec.html>)
- *Symbolic parser, semantic analyzer, morphological analyzer*
(<http://bahasa.cs.ui.ac.id/resources.php>)