

(Sesi 2)

Tutorial Dasar Analisis Sentimen Berbasis Aspek

Alfan

Fasilkom UI

Definisi Masalah

- Input:
 - Dokumen ulasan terkait restoran
 - Empat kategori aspek yang sudah ditentukan, yaitu **FOOD, AMBIENCE, SERVICE, PRICE**
- Output:
 - Identifikasi kehadiran dari empat aspek tersebut
 - Orientasi sentimen dari masing-masing aspek

Definisi Masalah

Input:

“tempatnya cozy nyaman buat lama lama disana tapi sayang banget pelayanannya buruk . Waitersnya kebanyakan ngobrol jadi tiap dipanggil mesti berkali kali , kurang senyum jadi kalo liat males banget buat manggil tapi tertolonglah sama tempat dan pemandangannya . Makanannya juga enak bangeeeeet so ga terlalu mengecewakan juga .”

Definisi Masalah

Input:

“**tempat**nya cozy nyaman buat lama lama disana tapi sayang banget **pelayanannya** buruk . **Waitersnya** kebanyakan ngobrol jadi tiap dipanggil mesti berkali kali , **kurang senyum** jadi kalo liat males banget buat manggil tapi tertolonglah sama tempat dan pemandangannya . **Makanannya juga enak bangeeeeet** so ga terlalu mengecewakan juga .”

Definisi Masalah

“**tempat**nya cozy nyaman buat lama lama disana tapi sayang banget **pelayanannya** buruk . **Waitersnya** kebanyakan ngobrol jadi tiap dipanggil mesti berkali kali , **kurang senyum** jadi kalo liat males banget buat manggil tapi tertolonglah sama tempat dan pemandangannya . **Makanannya** juga enak bangeeeeet so ga terlalu mengecewakan juga .”

Output >> Deteksi kategori aspek

FOOD

AMBIENCE

SERVICE

Definisi Masalah

“**tempat**nya cozy nyaman buat lama lama disana tapi sayang banget **pelayanannya** buruk . **Waitersnya** kebanyakan ngobrol jadi tiap dipanggil mesti berkali kali , **kurang senyum** jadi kalo liat males banget buat manggil tapi tertolonglah sama tempat dan pemandangannya . **Makanannya** juga enak bangeeeeet so ga terlalu mengecewakan juga .”

Output >> Deteksi kategori aspek >> Klasifikasi orientasi sentimen

FOOD positif

AMBIENCE positif

SERVICE negatif

Batasan & Asumsi

- Aspek sudah didefinisikan sebelumnya
- Aturan tata bahasa diasumsikan benar
- Orientasi sentimen bersifat dikotomi: **positif** atau **negatif**
 - Tidak ada yang bersifat netral
 - Tidak ada yang bersifat *conflicting*
 - “*sop ayam ini sangat enak, namun sangat berminyak dan tidak sehat.*”

Evaluasi

- Deteksi kategori aspek
 - Precision:
 - Recall
 - **F1 ???**
- Klasifikasi orientasi sentimen
 - Akurasi

Referensi: Pontiki et al., Semeval-2014 task 4:
Aspect based sentiment analysis, SemEval 2014.

Evaluasi

- *Gold Standard*
 - Dok 1: **FOOD** (pos), **SERVICE** (neg), **PRICE** (pos)
 - Dok 2: **FOOD** (neg), **AMBIENCE** (pos), **PRICE** (pos)
 - Dok 3: **AMBIENCE** (neg), **SERVICE** (pos), **PRICE** (neg)
- Program Kita
 - Dok 1: **FOOD** (pos), **AMBIENCE** (pos), **SERVICE** (neg)
 - Dok 2: **FOOD** (pos), **PRICE** (pos)
 - Dok 3: **FOOD** (neg), **SERVICE** (pos), **PRICE** (neg)

Deteksi kategori aspek: Precision = ?

Evaluasi

- *Gold Standard*
 - Dok 1: **FOOD** (pos), **SERVICE** (neg), **PRICE** (pos)
 - Dok 2: **FOOD** (neg), **AMBIENCE** (pos), **PRICE** (pos)
 - Dok 3: **AMBIENCE** (neg), **SERVICE** (pos), **PRICE** (neg)
- Program Kita
 - Dok 1: **FOOD** (pos), **AMBIENCE** (pos), **SERVICE** (neg)
 - Dok 2: **FOOD** (pos), **PRICE** (pos)
 - Dok 3: **FOOD** (neg), **SERVICE** (pos), **PRICE** (neg)

Deteksi kategori aspek: $\text{Precision} = (2 + 2 + 2) / (3 + 2 + 3) = 0.75$

Evaluasi

- *Gold Standard*
 - Dok 1: **FOOD** (pos), **SERVICE** (neg), **PRICE** (pos)
 - Dok 2: **FOOD** (neg), **AMBIENCE** (pos), **PRICE** (pos)
 - Dok 3: **AMBIENCE** (neg), **SERVICE** (pos), **PRICE** (neg)
- Program Kita
 - Dok 1: **FOOD** (pos), **AMBIENCE** (pos), **SERVICE** (neg)
 - Dok 2: **FOOD** (pos), **PRICE** (pos)
 - Dok 3: **FOOD** (neg), **SERVICE** (pos), **PRICE** (neg)

Deteksi kategori aspek: $\text{Recall} = (2 + 2 + 2) / (3 + 3 + 3) = 0.67$

Evaluasi

- *Gold Standard*
 - Dok 1: **FOOD** (pos), **SERVICE** (neg), **PRICE** (pos)
 - Dok 2: **FOOD** (neg), **AMBIENCE** (pos), **PRICE** (pos)
 - Dok 3: **AMBIENCE** (neg), **SERVICE** (pos), **PRICE** (neg)
- Program Kita
 - Dok 1: **FOOD** (pos), **AMBIENCE** (pos), **SERVICE** (neg)
 - Dok 2: **FOOD** (pos), **PRICE** (pos)
 - Dok 3: **FOOD** (neg), **SERVICE** (pos), **PRICE** (neg)

Klasifikasi orientasi sentimen: Akurasi = $(2 + 1 + 2) / (3 + 3 + 3) = 0.56$

Cross Validation

- Jangan gunakan metode *held-out*
 - Yaitu metode pisah data menjadi *training* & *testing*
- Gunakanlah metode cross validation
 - Bagi data keseluruhan menjadi **N** bagian terpisah secara **acak**
 - Gunakan **N-1** bagian untuk *training*, dan **1** bagian untuk *testing*
 - Putar peran ini sebanyak **N** kali, sehingga semua bagian pernah menjadi *testing* data

Perbandingan Antar Metode

- Gunakan uji statistik untuk menunjukkan bahwa suatu metode baru berbeda secara signifikan terhadap metode lama
 - Paired t test
 - Wilcoxon signed rank test

Perbandingan Antar Metode

- **Jangan** laporkan *kenaikan relatif* dari rata-rata aritmatika (*arithmetic mean*)
- Misal:
 - Mean precision metode A = **0.6**
 - Mean precision metode B = **0.4**
 - Laporan di buku thesis Anda: “metode A 50% lebih baik dari metode B” (X)
- Solusi:
 - laporkan saja apa adanya
 - Jika butuh peningkatan relatif, gunakan *geometric mean*; namun pastikan nilai yang dibandingkan berskala **rasio**.
 - Laporkan **effect size**: Cohen’s d, Glass delta, Hedge’s g

Data

- Lihat demo

Metode “Sederhana” Kita

- **State-of-the-art:** deep learning, representation learning, BERT, dsb.
- **Tutorial hari ini:** rule-based, machine learning sederhana

Deteksi Kategori Aspek

- Pengembangan leksikon kata aspek secara manual (*aspect term*)
 - FOOD: rasa, menu, makanan, minuman
 - PRICE: harga, ukuran
 - AMBIENCE: suasana, tempat, konsep, desain
 - SERVICE: pelayanan, staf, servis
- Leksikon ini nantinya bisa digunakan untuk pengembangan yang bersifat otomatis, atau semi-otomatis

Deteksi Kategori Aspek

- Input: dokumen ulasan restoran
- Proses:
 - $A \leftarrow$ untuk setiap kategori (*food, ambience, price, service*):
 - Kunjungi setiap kata di dokumen
 - Hitung berapa banyak kata pada dokumen yang muncul di leksikon aspek A
 - Jika kemunculannya \geq *threshold*, kita tetapkan bahwa aspek A ada pada dokumen tersebut.
 - *Threshold* adalah nilai *empiris*

Klasifikasi Orientasi Sentimen

- Kita gunakan *machine learning* → *binary classification* → 1 (**positif**), 0 (**negatif**)
- Fitur:
 - Kumpulan kata-kata yang ada di **sebelah kiri** dan **sebelah kanan** dari semua **kata aspek**.
 - Gunakan ukuran *WINDOW* tertentu

Klasifikasi Orientasi Sentimen

- “makanannya enak dan gurih, sambelnya juga pedes nampol, tetapi gw nggak suka sama pelayannya, jutek banget”

Klasifikasi Orientasi Sentimen

- “**makanannya** enak dan gurih, **sambelnya** juga pedes nampol, tetapi gw nggak suka **pelayannya**, jutek banget”
- WINDOW = 2
- FOOD = [enak, dan, gurih, juga, pedes]
- SERVICE = [nggak, suka, jutek, banget]

Klasifikasi Orientasi Sentimen

- Training data:
 - Ulasan #1
 - “restoran haji hisyam makanannya enak bikin nagih, pelayanannya bagus banget, tapi harganya mahal”
 - {**FOOD**: POSITIVE, **SERVICE**: POSITIVE, **PRICE**: NEGATIVE}
 - Ulasan #2
 - “ini restorannya bagus banget, dan minumannya bikin nagih, tapi pelayanannya jelek banget”
 - {**FOOD**: POSITIVE, **AMBIENCE**: POSITIVE, **SERVICE**: NEGATIVE}
 - Ulasan #3
 - “gw datang langsung lihat mejanya yang jelek, tapi satenya enak banget sih”
 - {**FOOD**: POSITIVE, **AMBIENCE**: NEGATIVE}

Klasifikasi Orientasi Sentimen

- Kita ekstraksi fitur-fitur pada training data beserta labelnya
- U ← untuk setiap dokumen ulasan pada data
 - A ← untuk setiap aspek yang terdaftar pada dokumen U
 - Kunjungi setiap kata pada U
 - Jika kata W ditemukan pada leksikon aspek A, kita tetapkan W adalah kata aspek berkategori A.
 - Ekstraksi kata-kata sebelah kiri dan kanan dari W dengan WINDOW yang sudah ditentukan
 - Simpan juga labelnya apakah POSITIVE atau NEGATIVE

Klasifikasi Orientasi Sentimen

- WINDOW = 2
- [haji, hisyam, enak, bikin], POSITIVE
- [bikin, nagih, bagus, banget], POSITIVE
- [banget, tapi, mahal], NEGATIVE
- [banget, dan, bikin, nagih], POSITIVE
- [ini, bagus, banget], POSITIVE
- [nagih, tapi, jelek, banget], NEGATIVE
- [jelek, tapi, enak, banget], POSITIVE
- [langsung, lihat, yang, jelek], NEGATIVE

Klasifikasi Orientasi Sentimen

- Gunakan representasi vektor untuk representasi internal dari training data yang sudah diekstrak fitur dan labelnya
 - Bag of words
 - TF / TF-IDF
- Jadikan input ke *off-the-shelf* machine learning model
 - Contoh pada python → tool scikit-learn

Klasifikasi Orientasi Sentimen

- Prediksi unseen data / testing data
- “makanannya enak dan gurih, sambelnya juga pedes nampol, tetapi gw nggak suka sama pelayannya, jutek banget”
- Gunakan program tahap 1 dahulu: ekstraksi kategori aspek

Klasifikasi Orientasi Sentimen

- “**makanannya** enak dan gurih, **sambelnya** juga pedes nampol, tetapi gw nggak suka **pelayannya**, jutek banget”
- Output dari tahap 1 (ekstraksi kata aspek):
 - WINDOW = 2
 - FOOD
 - **Kata aspek:** makanannya, sambelnya
 - [enak, dan, gurih, juga, pedes]
 - SERVICE
 - **Kata aspek:** pelayannya
 - [nggak, suka, jutek, banget]

Klasifikasi Orientasi Sentimen

- Unseen data / testing data
 - FOOD
 - [enak, dan, gurih, juga, pedes], ????
 - SERVICE
 - [nggak, suka, jutek, banget], ????
- Gunakan model yang sudah dilatih sebelumnya untuk menentukan apakah ????
- Tentunya perlu diubah kata-kata fitur tersebut ke representasi internal seperti *bag-of-words*, TF, TF-IDF, atau yang lainnya.

Ide Untuk Perbaikan

- Butuh *morphological analyser*
 - Makanannya → **makanan** + nya
- Butuh *POS Tagger*
 - Kata opini → kata sifat (ADJ)
 - Kata aspek → kata benda (NOUN)
- Butuh *coreference resolution*
 - “Jadi cuma pesen **roti dengan saus carbonara** yg gw lupa nama menunya tapi harus di coba banget . **Itu** enak banget loh .”

Ide untuk Perbaikan

- Menangani kata “tidak”, “nggak”, “gak”
 - “pelayanannya nggak **bagus**” → **NEGATIVE**
- Menggunakan atau mengembangkan leksikon opini
 - POSITIVE: bagus, mantap, enak, jos, ajib
 - NEGATIVE: jelek, buruk, hambar
- Dan yang lainnya ... bagus untuk topik penelitian Anda