Text Mining and Social Media Mining

THS. LÊ NHẬT TÙNG

Phân tích Cảm xúc
Sentiment Analysis

THS. LÊ NHẬT TÙNG

Tổng quan về Phân tích Cảm xúc

- Là một lĩnh vực trong xử lý ngôn ngữ tự nhiên (NLP), còn được gọi là Opinion Mining
- Quy trình tự động để hiểu ý kiến về một chủ đề nào đó
- Rất hữu ích cho các ứng dụng thương mại:
    - Phân tích marketing
    - Quan hệ công chúng
    - Đánh giá sản phẩm
    - Phản hồi sản phẩm
    - Dịch vụ khách hàng

Phân biệt sự kiện và ý kiến

Sự kiện (Facts):

- biểu đạt khách quan về điều gì đó

Ý kiến (Opinions):

- thường là biểu đạt chủ quan, mô tả cảm xúc, đánh giá, và cảm nhận của con người về một chủ đề hoặc đối tượng

Phân loại trong Phân tích Cảm xúc

Phân tích cảm xúc có thể được xây dựng như một bài toán phân loại với hai vấn đề nhỏ:

- Phân loại câu là chủ quan hay khách quan (subjectivity classification)
- Phân loại câu thể hiện quan điểm tích cực, tiêu cực hay trung tính (polarity classification)

Các cấp độ áp dụng Phân tích Cảm xúc

Phân tích cảm xúc có thể được áp dụng ở các cấp độ khác nhau:

- Cấp độ tài liệu - xác định cảm xúc của toàn bộ tài liệu hoặc đoạn văn
- Cấp độ câu - xác định cảm xúc của một câu đơn lẻ
- Cấp độ phần tử trong câu - xác định cảm xúc của các biểu thức con trong một câu

Các loại Phân tích Cảm xúc khác

- Phân tích tập trung vào tính phân cực (tích cực, tiêu cực, [trung tính])
- Phân tích phát hiện cảm xúc và cảm giác (tức giận, vui vẻ, buồn, phấn khích...)
- Phân tích xác định ý định (quan tâm / không quan tâm)

Các phương pháp Phân tích Cảm xúc

Phương pháp dựa trên luật (Rule-based):

Thực hiện phân tích cảm xúc dựa trên tập hợp các quy tắc được tạo thủ công.

Có thể sử dụng nhiều đầu vào, như:

- Kỹ thuật NLP cổ điển như stemming, tokenization, phân tích từ loại và cú pháp
- Các tài nguyên khác, như từ điển và từ vựng (danh sách các từ và biểu thức)

Ví dụ:

- Xác định hai danh sách từ có cảm xúc (tích cực và tiêu cực)
- Với một văn bản:
  - Đếm số từ tích cực trong văn bản
  - Đếm số từ tiêu cực trong văn bản
  - Nếu số từ tích cực nhiều hơn số từ tiêu cực, trả về cảm xúc tích cực; ngược lại, trả về cảm xúc tiêu cực. Còn lại, trả về trung tính.

Các phương pháp Phân tích Cảm xúc

Phương pháp tự động (Automatic):

Nhiệm vụ phân tích cảm xúc thường được mô hình hóa như một bài toán phân loại

Một bộ phân loại được cung cấp một văn bản và trả về danh mục tương ứng (ví dụ: tích cực, tiêu cực, hoặc trung tính)

Các phương pháp Phân tích Cảm xúc

Phương pháp kết hợp (Hybrid)

Kết hợp ưu điểm của cả phương pháp dựa trên luật và phương pháp tự động

Thông thường, bằng cách kết hợp cả hai phương pháp, phương pháp có thể cải thiện độ chính xác và độ tin cậy

Thách thức trong Phân tích Cảm xúc

Phần lớn công việc trong lĩnh vực phân tích cảm xúc những năm gần đây là giải quyết các thách thức và hạn chế chính:

- Tính chủ quan và giọng điệu
- Ngữ cảnh và tính phân cực
- Sự mỉa mai và châm biếm
- So sánh Emoji
- Định nghĩa trung tính

Công cụ Phân tích Cảm xúc trong R

AFINN

- Tác giả: Finn Arup Nielsen
- Gán cho mỗi từ một điểm số từ -5 đến 5, với điểm số âm thể hiện cảm xúc tiêu cực và điểm số dương thể hiện cảm xúc tích cực
- Có thể tải từ: http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010

Công cụ Phân tích Cảm xúc trong R

BING

- Tác giả: Bing Liu và cộng sự
- Phân loại từ theo cách nhị phân vào danh mục tích cực / tiêu cực
- Có thể tải từ: https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

Công cụ Phân tích Cảm xúc trong R

NRC

- Tác giả: Saif Mohammad và Peter Turney
- Phân loại từ theo cách nhị phân (có/không) vào các danh mục cảm xúc tích cực/tiêu cực: tức giận, mong đợi, ghê tởm, sợ hãi, vui vẻ, buồn bã, ngạc nhiên, tin tưởng
- Có thể tải từ: http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

Công cụ Phân tích Cảm xúc trong R

NRC

- Tác giả: Saif Mohammad và Peter Turney
- Phân loại từ theo cách nhị phân (có/không) vào các danh mục cảm xúc tích cực/tiêu cực: tức giận, mong đợi, ghê tởm, sợ hãi, vui vẻ, buồn bã, ngạc nhiên, tin tưởng
- Có thể tải từ: http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

Phương pháp khác

Gói SentimentAnalysis

- Thực hiện phân tích cảm xúc ở cấp độ câu, không phải từng từ riêng lẻ
- Tokenize mỗi tài liệu và cuối cùng chuyển đổi đầu vào thành ma trận document-term

Công cụ Phân tích Cảm xúc trong Python

NLTK (Natural Language Toolkit)

- Hỗ trợ tương đối nhiều ngôn ngữ và công cụ
- Tài liệu phát triển tốt và cộng đồng năng động
- Cung cấp nhiều bộ phân loại cảm xúc dựa trên từ vựng
- Có tích hợp với các bộ dữ liệu chuẩn như VADER (Valence Aware Dictionary and sEntiment Reasoner)

Công cụ Phân tích Cảm xúc trong Python

```python
from nltk.sentiment.vader import SentimentIntensityAnalyzer

sia = SentimentIntensityAnalyzer()
text = "Sản phẩm này rất tuyệt vời, tôi rất hài lòng!"
sentiment_scores = sia.polarity_scores(text)
print(sentiment_scores)
# Kết quả: {'neg': 0.0, 'neu': 0.295, 'pos': 0.705, 'compound': 0.8402}
```

Công cụ Phân tích Cảm xúc trong Python

SpaCy

- Khá phù hợp cho các ứng dụng công nghiệp
- Thư viện rất nhanh và hiệu quả
- Cung cấp các pipeline xử lý ngôn ngữ toàn diện
- Dễ dàng mở rộng với các mô hình tùy chỉnh
- Tích hợp tốt với các thư viện Deep Learning như TensorFlow và PyTorch

Công cụ Phân tích Cảm xúc trong Python

```python
import spacy

nlp = spacy.load("en_core_web_sm")
doc = nlp("Sản phẩm này có chất lượng tốt nhưng giá hơi cao.")
for token in doc:
    print(token.text, token.pos_, token.dep_)
```

Công cụ Phân tích Cảm xúc trong Python

TextBlob

- Nhẹ và rất dễ tiếp cận, phù hợp cho người mới bắt đầu
- Khả năng phân tích cảm xúc tương đối phong phú (ngay từ đầu)
- API đơn giản và trực quan
- Cung cấp thang đo cảm xúc từ -1.0 (tiêu cực) đến 1.0 (tích cực)
- Hỗ trợ các tác vụ NLP khác như dịch thuật, phát hiện ngôn ngữ, và trích xuất cụm danh từ

```
1  from textblob import TextBlob
2
3  text = "Dịch vụ khách hàng tuyệt vời. Tuy nhiên, sản phẩm không được như mong đợi."
4  blob = TextBlob(text)
5  print(blob.sentiment)  # Polarity: -0.1, Subjectivity: 0.9
```

Công cụ Phân tích Cảm xúc trong Python

Gensim

- Có khả năng mở rộng và nhanh, phù hợp cho dữ liệu lớn
- Có khả năng xử lý ngữ nghĩa tiềm ẩn mạnh mẽ
- Chuyên về các mô hình chủ đề và word embeddings
- Hỗ trợ các thuật toán như Word2Vec, Doc2Vec, FastText, và LDA
- Có các phát triển thương mại và ứng dụng trong nghiên cứu học thuật

```
1  from gensim.models import Word2Vec
2  from gensim.utils import simple_preprocess
3
4  documents = ["Tôi rất thích sản phẩm này", "Dịch vụ tuyệt vời", "Không hài lòng với chất lượ
5  processed_docs = [simple_preprocess(doc) for doc in documents]
6  model = Word2Vec(processed_docs, vector_size=100, window=5, min_count=1, workers=4)
7  similar_words = model.wv.most_similar("thích")
8  print(similar_words)
```

Công cụ Phân tích Cảm xúc trong Python

Stanford CoreNLP

• Không phụ thuộc nền tảng, được viết bằng Java nhưng có giao diện Python
• Hỗ trợ đa ngôn ngữ với hiệu suất cao
• Cung cấp phân tích cảm xúc dựa trên cây cú pháp
• Phân tích sâu hơn về cấu trúc cú pháp và ngữ nghĩa
• Có demo trực tuyến và được sử dụng rộng rãi trong nghiên cứu học thuật

```python
from stanfordcorenlp import StanfordCoreNLP

# Khởi tạo CoreNLP server
nlp = StanfordCoreNLP('path/to/stanford-corenlp')

# Phân tích cảm xúc
text = "Sản phẩm này rất tuyệt vời và tôi sẽ mua lại."
sentiment = nlp.annotate(text, properties={
    'annotators': 'sentiment',
    'outputFormat': 'json'
})
print(sentiment)

# Đóng kết nối
nlp.close()
```

Công cụ Phân tích Cảm xúc trong Python

Transformers (Hugging Face)

- Thư viện hiện đại nhất cho các mô hình NLP tiên tiến
- Hỗ trợ các mô hình pre-trained như BERT, GPT, RoBERTa, XLNet
- Cung cấp khả năng fine-tuning cho nhiều tác vụ khác nhau, bao gồm phân tích cảm xúc
- Hiệu suất cao và cập nhật liên tục theo các nghiên cứu mới nhất
- Cộng đồng rộng lớn và nhiều mô hình đã huấn luyện sẵn cho tác vụ phân tích cảm xúc

```python
from transformers import pipeline

# Sử dụng pipeline phân tích cảm xúc có sẵn
sentiment_analyzer = pipeline("sentiment-analysis")
result = sentiment_analyzer("Cuốn sách này thật tuyệt vời, tôi không thể ngừng đọc được.")
print(result)  # [{'label': 'POSITIVE', 'score': 0.9998}]
```

Công cụ Phân tích Cảm xúc trong Python

Flair

•Thư viện NLP mã nguồn mở cấp tiên tiến
•Cung cấp các word embeddings tiên tiến và mô hình phân loại cảm xúc
•Hỗ trợ nhiều loại embeddings khác nhau (word, character, contextual)
•Hiệu suất cao cho nhiều tác vụ NLP khác nhau
•Dễ sử dụng và tích hợp với các thư viện khác

```python
from flair.models import TextClassifier
from flair.data import Sentence

# Tải mô hình phân tích cảm xúc
classifier = TextClassifier.load('en-sentiment')

# Tạo Sentence object
sentence = Sentence("Dịch vụ rất tốt nhưng giá cả hơi đắt.")

# Dự đoán cảm xúc
classifier.predict(sentence)

# In kết quả
print(f'Sentiment: {sentence.labels[0].value} with score {sentence.labels[0].score:.4f}')
```