Phân tích cảm xúc đánh giá phim

Võ Linh Bảo - 18520503 - CS114.K21.KHTN

Link Github:

Tóm tắt

- **Tên đề tài**: Phân tích cảm xúc đánh giá phim
- **Tóm tắt về đồ án và kết quả đạt được:** Đồ án thực hiện các bước trong việc xây dựng một ứng dụng ML để phân tích cảm xúc (tích cực hoặc tiêu cực) của các bình luận/ đánh giá (review) phim từ người xem và các chuyên gia. Mục tiêu chính là để tổng hợp và thấy được sự phản hồi một cách khách quan của phần lớn người xem, chuyên gia về bộ phim đó là như thế nào. Từ đó giúp cho khách hàng chọn được bộ phim phù hợp, tránh việc chọn những bộ phim được "đánh giá quá mức" (overated) bởi truyền thông nhưng bị đánh giá thấp bởi người xem. Có 3 thuật toán được ứng dụng: 1) Linear Regression 2) Support Vector Machine 3) Naïve Bayes với kết quả đạt được trên tập test lần lượt là 93%, 94% và 93%.

Tóm tắt

- Ảnh của các thành viên của nhóm

Bài toán

Nhu cầu giải trí bằng phim ảnh trên thế giới đang chiếm một tỉ trọng đáng kể. Về phía người dùng, họ mong muốn được xem những bộ phim hay, được nhiều người đánh giá cao; và cách đơn giản và tiết kiệm thời gian nhất là dựa trên ý kiến phản hồi từ những người đã xem bộ phim đó. Tuy nhiên, để có được một đánh giá mang tính khách quan nhất chúng ta phải tham khảo từ nhiều ý kiến phản hồi.

Để tiết kiệm thời gian cho con người, chúng ta có thể dùng các thuật toán Máy học để đánh giá mức độ cảm xúc (sentiment analysis) của một đánh giá phim bất kỳ, sau đó tổng hợp lại tất cả đánh giá về bộ phim đó, chúng ta sẽ có thể kết luận bộ phim đó có đáng xem (theo ý kiến của phần lớn người xem) hay không.

Bài toán

Đây là bài toán phân tích cảm xúc (sentiment analysis)

Input: một đánh giá (dạng văn bản) về bộ phim

Output: đánh giá đó là tích cực (positive) hay tiêu cực (negative)

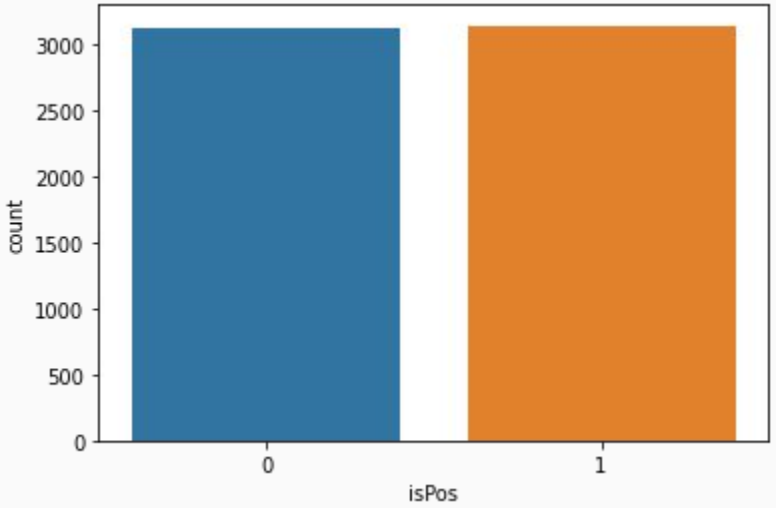
Cách giải quyết

Ta nhận thấy trong các đánh giá tích cực, tiêu cực có nhiều từ ngữ, cụm từ được sử dụng lặp đi lặp lại. Bằng cách phân loại những từ ngữ đó theo nhãn đã được gán trước (0 hoặc 1), sau đó tính toán tần số xuất hiện của chúng trong một đánh giá bằng các thuật toán Máy học, ta có thể phân loại được đánh giá đó là tích cực hay tiêu cực.

Mô tả dữ liệu

Bộ dữ liệu tự xây dựng gồm 6269 đánh giá phim từ IMDB, bao gồm 4 cột:

- Tên phim: *movie*
- Tiêu đề của đánh giá: *title*
- Nội dung của đánh giá: *review*
- Nhãn: *isPos* (0 – positive và 1 – negative)



Số lượng đánh giá ở 2 nhãn cân bằng nhau:

positive (3144) và *negative* (3125).

Mô tả dữ liệu

Cách xây dựng bộ dữ liệu:

Sử dụng Selenium để tự động khởi động trình duyệt mở đến trang đánh giá phim, sau đó tải tất cả bình luận (trigger button *Load more*) rồi tiến hành crawl các đánh giá đó về.

Gán nhãn: tiêu chí gán nhãn được dựa trên rating của bình luận đó:

- Nếu đánh giá có rating 9-10 sao: gán nhãn *positive* (*isPos*=0)
- Nếu đánh giá có rating từ 3 sao trở xuống: gán nhãn *negative* (*isPos*=1)

Các thao tác tiền xử lý dữ liệu

- Loại bỏ những ký tự KHÔNG nằm trong 26 chữ cái tiếng Anh, bao gồm: số, dấu chấm câu, ký tự đặc biệt.
- Loại bỏ những từ thừa trong quá trình crawl dữ liệu: “Was this review helpful?Sign into vote.Permalink”
- Chuyển tất cả chữ cái trong đánh giá về chữ thường (lowercase)
- Tokenize đánh giá thành nhiều từ nhỏ (sử dụng thư viện built-in của *nltk*)
- Loại bỏ stopwords
- Sử dụng kỹ thuật *Lemmatization* để biến đổi các từ về dạng nguyên mẫu.
- Kết hợp những word token vừa mới xử lý lại thành một câu. Đó là đánh giá ban đầu sau khi qua giai đoạn tiền xử lý.

Trích xuất đặc trưng và lựa chọn mô hình

Sử dụng 2 phương pháp để vector hóa các đánh giá thành các vector:

1) Count Vectorizer
2) TF-IDF Vectorizer

Có 3 thuật toán Máy học được áp dụng: 1) Linear Regression 2) Support Vector Machine 3) Naïve Bayes. Ta thử áp dụng cả 2 feature cho 3 thuật toán này để so sánh kết quả.

Đánh giá kết quả

Linear Regression (TF-IDF):

	precision	recall	f1-score	support
Positive	0.92	0.96	0.94	602
Negative	0.96	0.92	0.94	652
accuracy			0.94	1254
macro avg	0.94	0.94	0.94	1254
weighted avg	0.94	0.94	0.94	1254

Confusion Matrix:

```
[[575  27]
 [ 49 603]]
```

SVM (TF-IDF):

	precision	recall	f1-score	support
Positive	0.92	0.96	0.94	602
Negative	0.96	0.92	0.94	652
accuracy			0.94	1254
macro avg	0.94	0.94	0.94	1254
weighted avg	0.94	0.94	0.94	1254

Confusion Matrix:

```
[[575  27]
 [ 49 603]]
```

Naive Bayes (TF-IDF):

	precision	recall	f1-score	support
Positive	0.89	0.97	0.92	602
Negative	0.97	0.88	0.92	652
accuracy			0.92	1254
macro avg	0.93	0.93	0.92	1254
weighted avg	0.93	0.92	0.92	1254

Confusion Matrix:

```
[[582  20]
 [ 75 577]]
```

Linear Regression

Linear Regression (Count Vectorizer):

	precision	recall	f1-score	support
Positive	0.90	0.94	0.92	602
Negative	0.94	0.91	0.93	652
accuracy			0.92	1254
macro avg	0.92	0.92	0.92	1254
weighted avg	0.93	0.92	0.92	1254

Confusion Matrix:

```
[[567  35]
 [ 60 592]]
```

SVM

SVM (Count Vectorizer):

	precision	recall	f1-score	support
Positive	0.89	0.92	0.90	602
Negative	0.92	0.89	0.91	652
accuracy			0.91	1254
macro avg	0.90	0.91	0.91	1254
weighted avg	0.91	0.91	0.91	1254

Confusion Matrix:

```
[[552  50]
 [ 69 583]]
```

Naïve Bayes

Naive Bayes (Count Vectorizer):

	precision	recall	f1-score	support
Positive	0.90	0.97	0.93	602
Negative	0.97	0.90	0.93	652
accuracy			0.93	1254
macro avg	0.93	0.93	0.93	1254
weighted avg	0.93	0.93	0.93	1254

Confusion Matrix:

```
[[582  20]
 [ 67 585]]
```

Kết luận

- Cả ba thuật toán đều cho kết quả prediction từ 90-94%
- So sánh với bộ dữ liệu cùng lấy từ IMDB trên Kaggle với số lượng gấp khoảng 8 lần (~50k reviews), một kernel trên Kaggle cũng áp dụng những phương pháp tiền xử lý tương tự và 3 thuật toán nêu trên, kết quả prediction thu được từ 87-89%
- Nguyên nhân dẫn đến kết quả cao có thể là:
  - Số lượng nhãn chỉ là 2
  - Trong quá trình thu thập, việc gán nhãn cho dữ liệu được phân định quá rạch ròi (9-10 sao là postive, dưới 3 sao là negative), không có trường hợp nhiễu và trung tính (neutral) (e.g. các đánh giá 5-7 sao)
- Đem model thu được từ quá trình training test lại trên bộ dữ liệu crawl từ nguồn khác (Rotten Tomatoes), kết quả thu được:
- Hướng phát triển: kết hợp mở rộng bộ dữ liệu, thêm nhãn (e.g. neutral), trộn từ nhiều nguồn khác nhau