ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH

Báo cáo bài tập 2

Xử lý ngôn ngữ tự nhiên

Logistic Regression

GV hướng dẫn: TS. Nguyễn Thị Quý

Họ và tên	*MSSV*	*Mã lớp*
Trần Hoàng Bảo Ly	21521109	CS231.N21.KHTN
Lê Thanh Minh	21520063	CS231.N21.KHTN
Nguyễn Quốc Trường	21521604	CS231.N21.KHTN
Lê Thu Hà	21520800	CS231.N21.KHTN
Trần Xuân Minh	21520352	CS231.N21.KHTN

Hồ Chí Minh, tháng 4 năm 2023

Mục lục

1 Giới thiệu bài toán

1.1 Phân tích quan điểm (Sentiment Analysis)

Phân tích quan điểm là một ứng dụng của trí tuệ nhân tạo, nó sử dụng các thuật toán phức tạp để xử lý ngôn ngữ tự nhiên của con người (NLP) và xác định các đặc điểm cảm xúc tiêu cực/tích cực/trung lập, tại một thời điểm thông qua văn bản hoặc lời nói. Trong khuôn khổ môn học chúng ta chỉ làm việc với văn bản (text)

Hình 1: Mô tả về phân tích quan điểm (Nguồn: capitalhub)

Dựa trên phân tích ngôn ngữ tự nhiên dạng văn bản (text), để xác định xem liệu quan điểm của người viết về một vấn đề gì đó là gì.

1.2 Phân tích quan điểm dựa trên khía cạnh (aspect-based sentiment analysis - ABSA)

Phân tích quan điểm dựa trên khía cạnh (ABSA) là một kỹ thuật phân tích văn bản giúp phân loại dữ liệu theo khía cạnh và xác định quan điểm được gán cho từng khía cạnh đó. ABSA có thể được sử dụng để phân tích phản hồi của khách hàng bằng cách liên kết quan điểm cụ thể với các khía cạnh khác nhau của sản phẩm hoặc dịch vụ.

Ở đây, khía cạnh đề cập đến các thuộc tính hoặc thành phần của sản phẩm hoặc dịch vụ, ví dụ: giá cả, chất lượng, tốc độ sản phẩm/dịch vụ, ngoại hình/giao diện sản phẩm, ...

Như vậy ABSA có thể trích xuất:

- Quan điểm: ý kiến tích cực hoặc tiêu cực về một khía cạnh cụ thể.
- Các khía cạnh: danh mục, tính năng hoặc chủ đề đang được nói đến

1.3 Logistic Regression

Thuật toán logistic regression là thuật toán mà chúng ta tìm hiểu trong chương 5, logistic regression là thuật toán học máy(machine learning) được dùng cho bài toán phân loại, ví dụ email spam hay không spam, cảm xúc tích cực hay tiêu cực, thậm chí là phân loại với nhiều lớp như phân loại chó, mèo, gà, vit,... Đầu vào của bài toán có thể là hình ảnh, ngôn ngữ,... Tuy nhiên nó phải được trích xuất các đặc trưng thành các ma trận số. Trong bài tập lần này đặc trưng được sử dụng được trích xuất bằng phương pháp BOW (bag of words), Lý do sử dụng đặc trưng này cho mô hình logistic regression (giống với naive Bayes ở bài tập trước sẽ được giải thích ở phần giới thiệu bài toán).

2 Giải quyết

2.1 Giới thiệu

Trong bài toán về nhà hàng và khách sạn, các khía cạnh được quan tâm gồm: Thực thể (ENTITY), thuộc tính (ATTRIBUTE), và quan điểm (SENTIMENT) dựa trên khía cạnh đó. Nghĩa là mô hình của chúng ta phải phân tích được cả 3 khía cạnh kể trên, việc trích xuất đặc trưng cho từng mô hình là cực kỳ tốn kém. Vậy nên, một cách tổng quát cho việc giải quyết vấn đề ở trên là sử dụng phương pháp truyền thống BOW.

Vì bài toán chúng ta cần giải quyết tương tự với bài tập 1, nên phần dưới đây tương tự với báo cáo ở bài tập 1 cho đền mục 3.3

2.2 Hướng giải quyết

Với bài toán phân tích quan điểm dựa trên khía cạnh, mỗi lời nhận xét có thể bao gồm nhiều khía cạnh. Giả sử mỗi một câu là một quan điểm về một khía cạnh nào đó. Vậy nên với mỗi một lời nhận xét chúng ta cố gắng tách thành các câu có nghĩa đầy đủ, mỗi một câu là một khía cạnh.

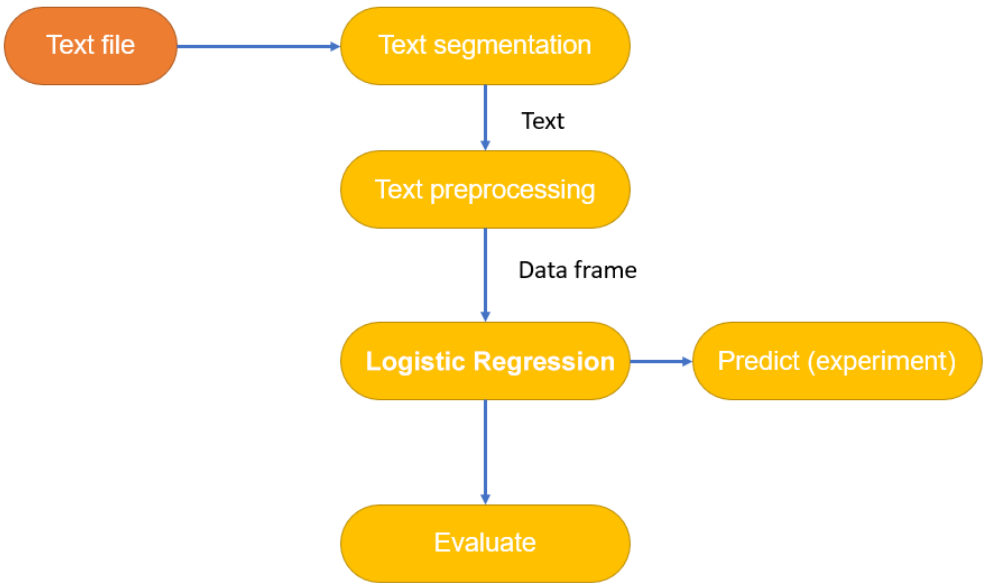
Mỗi một câu đi kèm với nó là ENTITY, ATTRIBUTE và SENTIMENT, việc của chúng ta là cố gắng dự đoán câu trên nói về cái gì, nói về vấn đề gì, và quan điểm của câu nói đó là gì.

Như vậy chúng ta sẽ cần 3 logistic model để giải quyết bài toán trên.

- Một model để xác định thực thể (Entity)
- Một model để xác định vấn đề (Attribute)
- Một model để xác định quan điểm (sentiment)

3 Quy trình thực hiện

3.1 Sơ đồ thực hiện



Hình 2: Quy trình thực hiện

3.2 Giải thích các bước

.

- Text_segmentation: Sau khi đọc file text, text gốc được tách thành các câu khác nhau, dựa vào các nhãn là các aspect cho trước.

- Text_preprocessing: Xử lý text đã được tách bằng cách loại bỏ các dấu câu, thực hiện ghép các từ đơn thành các từ ghép có nghĩa, chuẩn hóa. Sau đó sắp xếp thành dataframe với các aspect và sentiment tương ứng.
- Model: model được sử dụng là Logistic regression (từ thư viện sci-kit learn). Tiến hành fit model với các features được chọn theo phương pháp bag of words.
- Predict: Tiến hành thử nghiệm model với các text cụ thể, đánh giá kết quả sơ bộ, đặc biệt là với bài toán phân tích quan điểm dựa trên khía cạnh.
- Evaluate: đánh giá hiệu suất tổng thể của model trên cả tập train và tập test.

3.3 Hướng tiếp cận 1

Sử dụng StandardScaler để scale lại các feature trước thực hiện fit model.

Kết quả thu được với tập dữ liệu Restaurant (train).

precision	**recall**	**f1-score**	**support**	
negative	1.00	0.99	0.99	466
neutral	0.99	0.98	0.98	1321
positive	0.99	1.00	1.00	5455
accuracy	0.99	7242		
macro avg	0.99	0.99	0.99	7242
weighted avg	0.99	0.99	0.99	7242
precision	**recall**	**f1-score**	**support**	
AMBIENCE	0.99	0.99	0.99	607
DRINKS	1.00	0.99	0.99	272
FOOD	0.99	1.00	0.99	4562
LOCATION	1.00	0.99	1.00	358
RESTAURANT	1.00	0.98	0.99	952
SERVICE	0.99	0.97	0.98	491
accuracy	0.99	7242		
macro avg	0.99	0.99	0.99	7242
weighted avg	0.99	0.99	0.99	7242
precision	**recall**	**f1-score**	**support**	
GENERAL	0.99	0.98	0.98	2070
MISCELLANEOUS	0.99	1.00	1.00	155
PRICES	0.99	0.98	0.98	1234
QUALITY	0.97	0.99	0.98	2264
STYLE&OPTIONS	0.99	0.98	0.99	1519
accuracy	0.98	7242		
macro avg	0.99	0.99	0.99	7242
weighted avg	0.98	0.98	0.98	7242

Kết quả đánh giá với tập dữ liệu Restaurant (test)

precision	**recall**	**f1-score**	**support**	
negative	0.21	0.21	0.21	123
neutral	0.21	0.25	0.23	333
positive	0.77	0.74	0.76	1355
accuracy	0.61	1811		
macro avg	0.40	0.40	0.40	1811
weighted avg	0.63	0.61	0.62	1811
precision	**recall**	**f1-score**	**support**	
AMBIENCE	0.12	0.15	0.13	154
DRINKS	0.07	0.07	0.07	60
FOOD	0.69	0.65	0.67	1157
LOCATION	0.31	0.20	0.24	84
RESTAURANT	0.20	0.24	0.22	235
SERVICE	0.12	0.13	0.12	121
accuracy	0.48	1811		
macro avg	0.25	0.24	0.24	1811
weighted avg	0.50	0.48	0.49	1811
precision	**recall**	**f1-score**	**support**	
GENERAL	0.36	0.35	0.36	510
MISCELLANEOUS	0.06	0.03	0.04	38
PRICES	0.23	0.25	0.24	326
QUALITY	0.37	0.39	0.38	551
STYLE&OPTIONS	0.26	0.26	0.26	386
accuracy	0.32	1811		
macro avg	0.26	0.25	0.26	1811
weighted avg	0.32	0.32	0.32	1811

Có thể thấy, với hướng tiếp cận trên. Chúng ta đã có kết quả siêu siêu tốt đối với tập train, phân tích quan điểm dựa và các khía cạnh gần như đặt điểm tuyệt đối. Tuy nhiên, kết quả lại quá tệ đối với tập test, mô hình logistic regression của chúng ta đã bị overfitting quá nặng, làm mất tính tổng quát hóa. Vậy nên, chúng ta sẽ đến với hướng giải quyết khác có tình tổng quát hóa hơn (tuy nhiên kết quả đánh giá với tập train thấp hơn rất nhiều so với hướng tiếp cận này, đây là điều phải chấp nhận).

3.4 Hướng tiếp cận 2

Chỉ sử dụng TfidfTransfomer và RobustScaler để xử lý các features trước khi thực hiện (được trình bày trong file notebook), kết quả đánh giá cũng bao gồm trong file ở trên.

4 Đánh giá hiệu suất thực hiện

Phần đánh giá hiệu suất thực hiện đã được bao gồm trong notebook nhằm tăng tính thực tiễn.

Link đến github của assignment tại đây