Trường Đại học Công nghệ Thông tin ĐHQG TP.HCM
Tháng 8, 2020

Nhận diện cảm xúc qua các dòng trạng thái

Thành viên:
Đỗ Mạnh Quân 18521283
Huỳnh Đỗ Anh Vũ 18521665

Khoa Khoa Học Máy Tính
Lớp CS232.K21.KHCL – Thầy Mai Tiến Dũng

MỤC LỤC

1. Giới thiệu

-Nhận dạng cảm xúc có rất nhiều ứng dụng đặc biệt trong thời đại số ngày hôm nay, việc thu thập dữ liệu cũng tương đối đơn giản bao gồm các loại dữ liệu đa phương tiện. Hiện nay mọi người đều sử dụng mạng xã hội, đó là kho dữ liệu văn bản lớn để nhóm chúng tôi thực hiện đề tài này.

-Đây là bài báo cáo cuối kì của liên môn máy học và tính toán đa phương tiện. Trong môn tính toán đa phương tiện, chúng tôi xử lí dữ liệu văn bản và so sánh hai phương pháp số hoá còn ở môn máy học chúng tôi đề xuất một số giải thuật phù hợp cho bài toán song song với đó là đánh giá mô hình. Cuối cùng chúng tôi cho ra kết quả và rút ra kết luận, thông qua đồ án này chúng tôi mong muốn hiểu được cách tiếp cận một bài toán máy học và hướng giải quyết thế nào, cách sử dụng các thư viện để giải quyết quá trình từ xử lí dữ liệu, đến số hoá và thực nghiệm để đưa ra dự đoán.

2. Tập dữ liệu

-Kaggle là ngôi nhà cho những người đam mê khoa học dữ liệu, là nơi học hỏi kinh nghiệm, lưu trữ dữ liệu từ rất nhiều lĩnh vực. Tập dữ liệu chúng tôi chọn gồm 1600000 tweets đã được trích xuất bằng twitter api. Dữ liệu gồm 2 nhãn là vui và buồn. Tuy nhiên để cho mô hình xử lí nhanh thì chúng tôi chỉ lấy ra mẫu đại diện gồm 200000 tweets, nhãn của 2 lớp là phân bố đều nhau. [1]

3. Tiền xử lí dữ liệu

-Trong một đoạn văn bản thì không phải từ nào cũng quan trọng và nêu lên nội dung chủ đạo của vấn đề đang xét, do đó việc loại bỏ các từ không quan trọng là vô cùng cần thiết. Việc này giúp giảm chiều dữ liệu vì mỗi từ tương ứng với một thuộc tính khi thực hiện giai đoạn số hoá), đồng thời giúp mô hình dự đoán chính xác và tăng tốc độ huấn luyện của máy. [2]

-Các công đoạn xử lí gồm:

1. Loại bỏ các các đường dẫn url, địa chỉ mail: thông thường các địa chỉ này sẽ không đại diện cho nội dung văn bản vì là danh từ riêng.
2. Nhóm các từ có cùng ngữ nghĩa thành một nhóm: từ thì sẽ có danh, động, tính và trạng từ, khi này sẽ cần một từ đại diện là đủ nêu lên nội dung đang xét.
3. Loại bỏ các từ dừng: từ dừng là những từ xuất hiện nhiều trong các văn bản mà không thể hiện việc văn bản đang đề cập đến vấn đề nào.

4. Số hoá dữ liệu

-Việc số hoá dữ liệu chúng tôi thử nghiệm và so sánh 2 phương pháp kinh điển là bag-of-words và tf-idf.[2]

-Thuộc tính là tập hợp các từ không trùng nhau tương ứng với số cột, tổng số lượng tweets là tương ứng với số dòng. Xét phương pháp bag-of-words, với mỗi văn bản từ nào xuất hiện thì đánh số 1, tương tự cho phương pháp tf-idf song từ nào càng quan trọng thì thì giá trị càng cao dao động từ [0, 1].

-Có nhiều tham số trong 2 phương pháp số hoá kể trên, song chúng tôi đề cập đến các tham số chính ảnh hưởng đến độ chính xác. Do đó, hai tham số được lựa chọn là phạm vi n_range của từ và giá trị hiển thị(binary). Phạm vi n_range tức là ta có thể xem một nhóm từ thành một từ vì đôi khi từ gộp lại mới có ý nghĩa, giá trị hiển thị(binary) nghĩa là từ xuất hiện dưới dạng tần số(số lần) hay là chỉ xét đến sự hiển diện của từ.

5. Thuật toán

-Chúng tôi dựa trên kết quả thực nghiệm của thư viện sklearn và chọn ra các thuật toán tiêu biểu [3]:

1. Logistic Regression
2. Multinomial Naive Bayes
3. LinearSVC

-Chúng tôi sử dụng Logistic Regression cho bài toán phân lớp nhị phân bằng phương pháp One-Vs-Rest, vơi hai tham số tinh chỉnh là solver và C. solver đại diện cho phương pháp tìm ra tham số tối ưu. Thư viện sklearn gợi ý nên sử dụng trước tiên là phương pháp mặc định lbfgs, giải thuật này ước lượng đạo hàm bậc hai của ma trận thông qua việc đánh giá gradient – nó thực sự tiết kiệm bộ nhớ nhưng lại không cho ra kết quả quá nhanh với dữ liệu lớn. Do đó chúng tôi sử dụng thêm một phương pháp có tên gọi liblinear, hoạt động hiệu quả với dữ liệu nhiều chiều dành cho bài toán phân loại tuyến tính, rất phù hợp cho dữ liệu văn bản. Tham số C đại diện cho giá trị nghịch đảo của chỉnh qui hoá (regularization) tức là C càng lớn, mô hình càng cho ra kết quả tốt với tập training(overfitting) nhưng độ chính xác lại không cao với tập testing.

-Ngoài ra mô hình MultinomialNB và LinearSVC cũng được sử dụng với các tham số mặc định, chúng tôi đánh giá các mô hình bằng kiểm chứng chéo (cross validation) vì giúp tầm soát được dữ liệu đang dùng. Khi chia dữ liệu thành các k mẫu, mỗi mẫu luôn đảm bảo mật độ phân bố của các nhãn là bằng nhau, việc này sẽ giúp cho mô hình được huyến luyện với số lượng nhãn cân bằng nhau và cho ra kết quả tốt hơn.

6. Kết quả

-Chúng tôi huấn luyên dựa trên kiểm chứng chéo gồm 5 mẫu và kết quả là lấy trung bình.

-Bảng 1 chúng tôi chỉ ra kết quả giữa 2 phương pháp so sánh là bag-of-words và tf-idf sau khi tinh chỉnh các tham số với thuật toán Logistic Regression:

Bảng 1. Bag of words vs TF-IDF

	Độ chính xác (200000 mẫu)	Tham số tinh chỉnh
Bag of words	76.56%	'bow__binary': True, 'bow__ngram_range': (1, 2), 'clf__C' : 1, 'clf__solver': 'liblinear'
Tf-idf	77.2%	'clf__C': 1, 'clf__solver': 'lbfgs', 'tfidf__binary': True, 'tfidf __ngram_range': (1, 2)

-Bảng 2 cũng sử dụng kiểm chứng chéo gồm 5 mẫu và lấy kết quả trung bình, sử dụng các thuật toán Logistic Regression, MultinomialNB và LinearSVC:

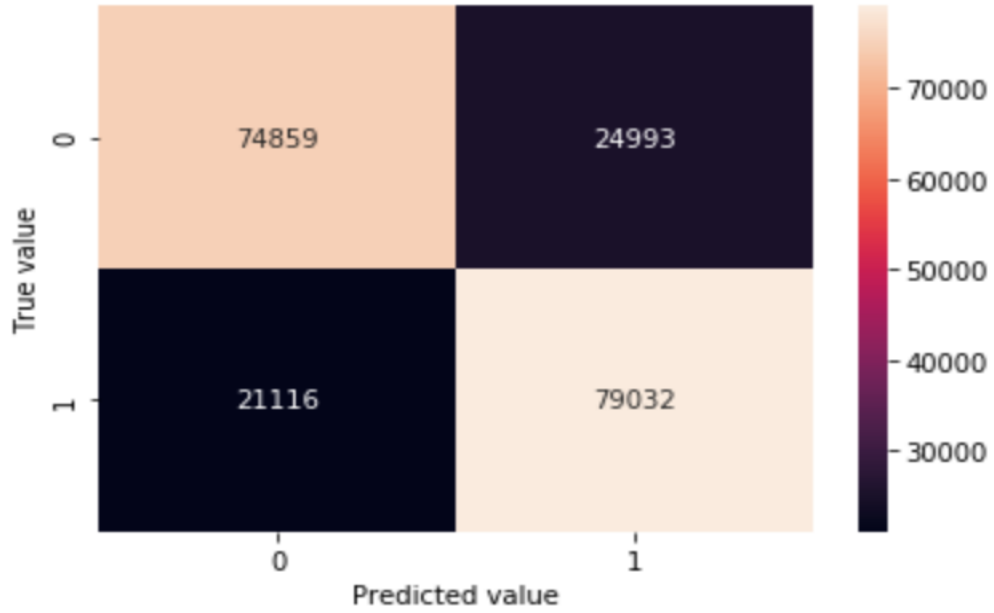
Bảng 2. TF-IDF with different algorithms

	Mẫu lớn (200000)	Mẫu nhỏ (20000)
Logistic Regression	77.2%	73.88%
Multinomial Naive Bayes	75.97%	73.04%
LinearSVC	76.93%	73.64%

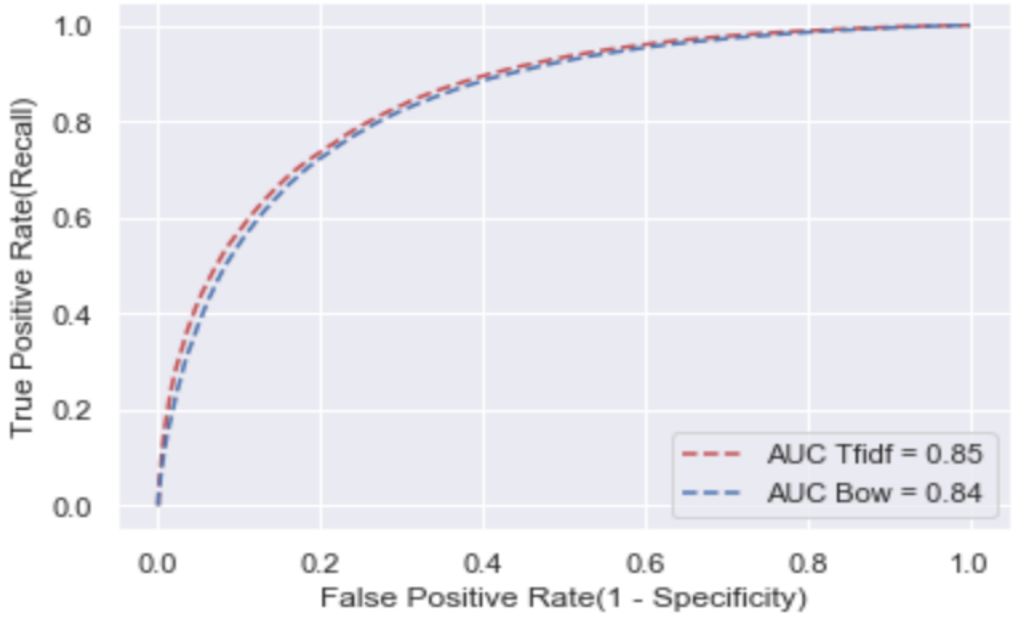
7. Đánh giá

-Chúng tôi mong muốn đánh giá các mô hình bằng việc thể hiện trực quan qua hình ảnh, đều được huyến luyện trên mẫu có kích cỡ lớn(200000):

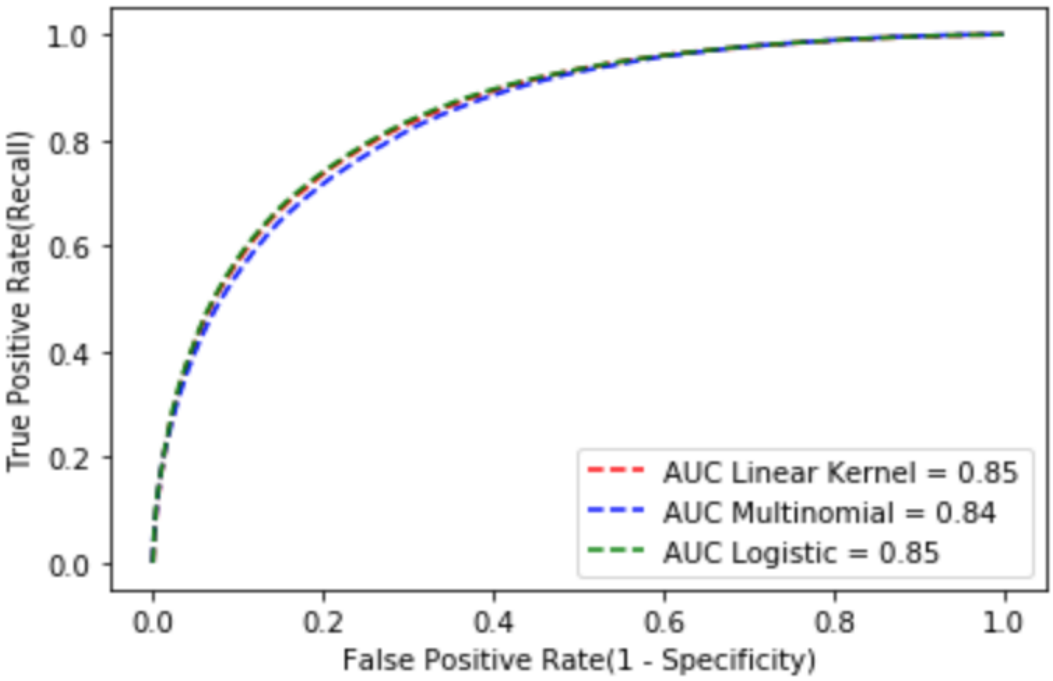
```
precision_score: 0.7597404470079308
recall_score: 0.7891520549586611
f1_score: 0.7741670054316683
```



Hình 1 thể hiện confusion matrix của mô hình tốt nhất (số hoá bằng tfidf, thuật toán logistic regression). Ngoài ra còn hiển thị thêm các thông số precision, recall và f1_score.



Hình 2 là ROC Curve của 2 phương pháp số hoá dữ liệu bằng TFIDF sử dụng thuật toán Logistic Regression



Hình 3 là ROC Curve của các thuật toán sử dụng TFIDF

8. Kết luận

-Giữa hai phương pháp số hoá dữ liệu: TFIDF cho ra kết quả tốt hơn với cùng một thuật toán, một tập dữ liệu phân bố các nhãn xấp xĩ bằng nhau và hình thức kiểm chứng chéo 5 mẫu. Điều này cũng dễ hiểu vì Bag of words chỉ nói lên sự hiển diện của từ trong văn bản, TFIDF cũng làm được như thế song lại có thể đánh giá mức độ quan trọng của từ đó nên sẽ cung cấp cho mô hình toán học nhiều thông tin hơn.

-Khi sử dụng máy học để dự đoán, dữ liệu đủ lớn một trong những điều kiện tiên quyết vì kết quả cho thấy dữ liệu đủ lớn cho ra độ chính xác cao hơn.

-Đứng trước quyết định nên sử dụng thuật toán nào cho bài toán, ta cần khoanh vùng một số thuật toán phù hợp, nên tham khảo và dựa trên các công bố khoa học, thực nghiệm của thư viện đang sử dụng.

-Việc lựa chọn tham số cho thuật toán, phương pháp số hoá tuỳ thuộc vào tập dữ liệu nên không có tham số nào là tuyệt đối. Người thực hiện cần tinh chỉnh các tham số để cho ra mô hình tốt nhất phù hợp với dữ liệu của mình.

-Với mẫu là 200000 văn bản và kết quả khoảng 77% - tương đối tốt, qua đó cho thấy các thuật toán máy học cổ điển vẫn được ứng dụng rộng rãi.

9. Hướng cải tiến

-Sử dụng tập dữ liệu lớn hơn, có thể sử dụng kĩ thuật plot learning curve để xem xét dữ liệu bao nhiêu là hợp lí (nếu có thời gian và nguồn lực).

-Mở rộng phạm vi tinh chỉnh tham số và sử dụng các thuật toán khác, kết hợp các công nghệ học sâu hiện đại.

-Tương tự, xử lí thêm các dữ liệu văn bản cũng như kết hợp các mô hình học sâu đã được huấn luyện sẵn để trích xuất các từ thực sự quan trọng.

Nguồn tham khảo

[1] Tập dữ liệu: https://www.kaggle.com/kazanova/sentiment140.
[2] Tiền xử lí, phương pháp số hoá dữ liệu: https://kavita-ganesan.com/.
[3] Thuật toán máy học: https://scikit-learn.org/stable/, hands on machine learning with scikit-learn and tensorflow.
Lựa chọn thuật toán: https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html