November2023

PHÂN LOẠI ĐA NHÃN ĐÁNH GIÁ CỦA KHÁCH HÀNG TRÊN SÀN THƯƠNG MẠI ĐIỆN TỬ TIKI

DS102.O11.CNCL
GV: Nguyễn Văn Kiệt

Nhóm 8
21522536 - Nguyễn Phan Trúc Quỳnh
21522496 - Nguyễn Minh Quân
21522798 - Lương Triệu Hoàng Vũ
21522315 - Nguyễn Thị Cẩm Ly

Tóm lược

Trong báo cáo này, chúng tôi tập trung phân loại đa nhãn đánh giá của khách hàng trên sàn thương mại điện tử Tiki nhằm mục đích xác định những trải nghiệm của khách hàng về sản phẩm mà họ đã mua. Xây dựng bộ dữ liệu từ gần 3500 đánh giá của khách hàng về sản phẩm thuộc lĩnh vực chăm sóc da và trang điểm. Huấn luyện mô hình phân loại đa nhãn theo 3 phương pháp là Binary Relevance, Classifier Chains, Label Powerset bằng các thuật toán NB, SVC, RF, KNN, LR. Theo kết quả nghiên cứu thử nghiệm, chúng tôi nhận thấy phương pháp là **Binary Relevance** sử dụng thuật toán **RF** cho kết quả Micro F1 Score 0,85914 và Hamming Loss 0.05078 là phương pháp mang lại kết quả tốt nhất.

Nội dung

1. Mô tả bài toán
2. Nghiên cứu liên quan
3. Xây dựng bộ dữ liệu
4. Xử lý dữ liệu và trích xuất đặc trưng
5. Huấn luyện và đánh giá mô hình
6. Kết luận và hướng phát triển

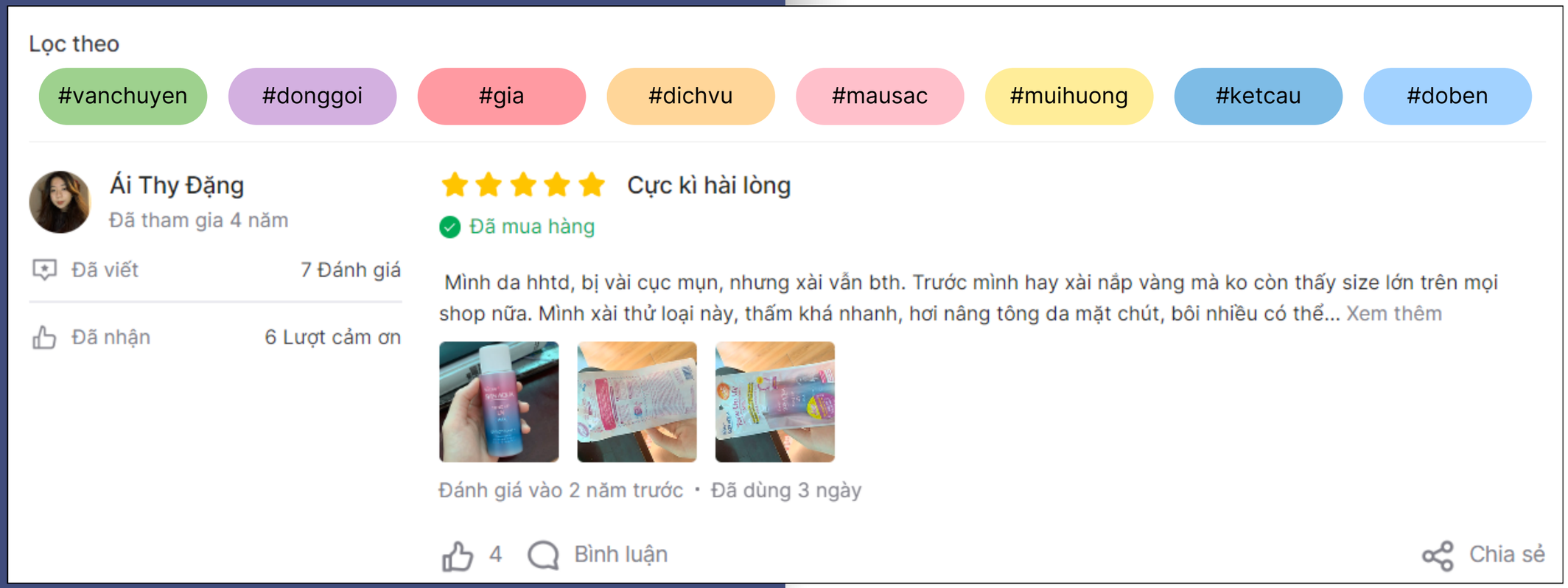
1. Mô tả bài toán

Input

Các đánh giá của khách hàng về sản phẩm chăm sóc da và trang điểm.

Output

Các đánh giá từ input kèm các nhãn liên quan.



2. Nghiên cứu liên quan

Adapting Transformers for Multi-Label Text Classification

- Phân loại văn bản đa nhãn, đề xuất kiến trúc cho các lớp phân loại.
- Sử dụng 2 dataset là AAPT (các bài báo khoa học) và Reuters-21578 (tóm tắt của các bài báo của hãng tin Reuters).
- Precision score của mô hình Random Forest đạt 97,18% với Reuter và 94,2% với AAPT

Multi-Label Classification of E-Commerce Customer Reviews via Machine Learning

- Phân loại đánh giá của khách hàng bằng nhiều nhãn.
- Tập dữ liệu đã tạo chứa hơn 50.000 đánh giá thuộc ba danh mục khác nhau và mỗi đánh giá có nhiều nhãn.
- Micro Precision 0,9157; Micro Recall 0,8837; Micro F1 Score 0,8925 và Hamming Loss 0,0278

3. Xây dựng bộ dữ liệu

Bộ dữ liệu

Thu thập dữ liệu từ API của sàn thương mại điện tử Tiki và xây dựng dữ liệu từ 3492 đánh giá của khách hàng đã mua và sử dụng các sản phẩm chăm sóc da và trang điểm.

Phân chia nhãn

Chất lượng

Dịch vụ (sản phẩm thật / giả)
Mùi hương
Kết cấu
Độ bền
Màu sắc

Thông tin

Dịch vụ (chăm sóc khách hàng)
Vận chuyển
Giá
Đóng gói

3.1. Quy tắc gán nhãn

1. Vận chuyển

Các đánh giá đề cập đến:

- Việc giao hàng: “hộp móp, méo mó”; “hàng bị ướt”.
- Thời gian giao hàng: “mới đặt hôm qua, mà nay đã tới”.
- Tốc độ giao hàng: “giao hàng nhanh”.
- Người giao hàng: “shipper thân thiện”.

2. Giá

Các đánh giá đề cập đến:

- Mức giá của sản phẩm: “Dùng cũng ổn trong tầm giá”.
- Giá cụ thể của sản phẩm: “Giá sale rẻ chỉ 9k 1 cây”.

3. Đóng gói

Các đánh giá đề cập đến:

- hình thức đóng gói: “đóng gói cẩn thận”.
- Bao bì sản phẩm: “Bao bì đẹp mắt, chắc chắn”.

4. Dịch vụ

Các đánh giá đề cập đến:

- Chăm sóc khách hàng: “giao sai hàng, đã liên hệ mà chưa thấy phản hồi”.
- Thái độ phục vụ: “shop bán hàng rất lịch sự”.
- Việc gửi hàng: “Giao sai sp cho mình rồi, vui lòng đổi hàng giúp mình”.
- Sự kiện tặng quà: “được tặng thêm 3 tuýp nhỏ rất xinh”.
- Vấn đề thật / giả, tem, nhãn mác của sản phẩm: “Sản phẩm chính hãng, có tem phụ công ty”.
- Hạn sử dụng của sản phẩm: “date xa”.

3.1. Quy tắc gán nhãn

5. Mùi hương

Các đánh giá đề cập đến:

- Hương liệu của sản phẩm: “mùi thơm, giá khá ổn”.
- Mùi hương cụ thể: “Son có mùi mật ong, mềm, dưỡng ẩm tốt”.

6. Kết cấu

Các đánh giá đề cập đến:

- độ đậm đặc, trạng thái của sản phẩm: “độ kem đặc, chống nắng tốt”.
- Trạng thái của sản phẩm khi sử dụng trên da: “chất kem thấm nhanh”.

7. Độ bền

Các đánh giá đề cập đến:

- Khả năng kiềm dầu, chống nước của sản phẩm: “Lớp nền mịn, không bóng dầu.”.
- Khả năng giữ lớp sản phẩm trên da: “Màu lì hơi khô môi”.

8. Màu sắc

Các đánh giá đề cập đến:

- Màu của sản phẩm: “Màu đẹp, không khô môi”.
- màu cụ thể của sản phẩm: “Sao tôi đặt đỏ cam lại giao màu cherry”.

3.2. Đánh giá độ đồng thuận

$$kappa(\kappa) = \frac{P_o - P_e}{1 - P_e}$$

- Po là tỷ lệ đồng ý thực tế giữa các người đánh giá.
- Pe là tỷ lệ đồng ý ngẫu nhiên.

- Đa phần các nhãn đều đạt độ đồng thuận cao.
- Cải thiện các nhãn có điểm độ đồng thuận từ 0.3 - 0.5.
- Các nhãn từ 0.8 -> tìm ra các nhãn khác nhau và thống nhất lại giữa những người gán nhãn.

Bộ 1:

0.916153957499907
0.859554726897695 9
0.904579859596499 2
0.567217398404964 7
0.95959447349766 6
0.546604754389874 4
0.790181134591521 1
0.925268008338786 9

Bộ 2:

0.964953026709007 7
0.915914560117153 6
0.937609353034578 4
0.82534253406566 84
0.98198872598331 81
0.81378789465358 7
0.30621731551423 58
0.87651551312649 17

4.1. Tiền xử lý dữ liệu

Ảnh hưởng của các phương pháp tiền xử lý đến kết quả huấn luyện mô hình

Phương pháp loại bỏ khỏi tiền xử lý	Hamming Loss	Micro F1
Xóa kí tự, emoji, khoảng trắng thừa	0.08398	0.82868
Chuẩn hóa cách gõ dấu tiếng Việt	0.08593	0.82400
Xóa các kí tự cố ý viết dài	0.08789	0.81781
Chuẩn hóa teencode	0.07813	0.84127
Chuẩn hóa unicode	0.08203	0.83200
Stopword	0.09375	0.80487
Tách từ	0.08594	0.82677

4.2. Trích xuất đặc trưng

3 phương pháp trích xuất đặc trưng

Method	Hamming Loss	Micro F1
TF-IDF	0.05025	0.86726
Word2Vec	0.11999	0.62324
Bag_of_word	0.05042	0.86847

Kết quả so sánh các phương pháp trích xuất đặc trưng

5. Huấn luyện và đánh giá mô hình

Binary Relevance

Ở cách này, mỗi nhãn sẽ được xem là một bài toán phân loại riêng biệt.

X	Y_1	Y_2	Y_3	Y_4
$x^{(1)}$	0	1	1	0
$x^{(2)}$	1	0	0	0
$x^{(3)}$	0	1	0	0
$x^{(4)}$	1	0	0	1
$x^{(5)}$	0	0	0	1

⇨

X	Y_1
$x^{(1)}$	0
$x^{(2)}$	1
$x^{(3)}$	0
$x^{(4)}$	1
$x^{(5)}$	0

X	Y_2
$x^{(1)}$	1
$x^{(2)}$	0
$x^{(3)}$	1
$x^{(4)}$	0
$x^{(5)}$	0

X	Y_3
$x^{(1)}$	1
$x^{(2)}$	0
$x^{(3)}$	0
$x^{(4)}$	0
$x^{(5)}$	0

X	Y_4
$x^{(1)}$	0
$x^{(2)}$	0
$x^{(3)}$	0
$x^{(4)}$	1
$x^{(5)}$	1

Với X là feature và Y là các labels. bài toán trên sẽ được chia làm 4 bài toán nhỏ riêng biệt (4 labels)

Label Powerset

Cách làm này sẽ xem xét các trường hợp có các nhãn là giống nhau mà gom lại thành một lớp

X	y1	y2	y3	y4
x1	0	1	1	0
x2	1	0	0	0
x3	0	1	0	0
x4	0	1	1	0
x5	1	1	1	1
x6	0	1	0	0

➡

X	y1
x1	1
x2	2
x3	3
x4	1
x5	4
x6	3

Classifier Chains

X	y1
x1	0
x2	1
x3	0

Classifier 1

X	y1	y2
x1	0	1
x2	1	0
x3	0	1

Classifier 2

X	y1	y2	y3
x1	0	1	1
x2	1	0	0
x3	0	1	0

Classifier 3

X	y1	y2	y3	y4
x1	0	1	1	0
x2	1	0	0	0
x3	0	1	0	0

Classifier 4

Nhãn đã được huấn luyện sẽ trở thành input và tiếp tục huấn luyện để dự đoán ra nhãn tiếp theo

Kết quả và nhận xét

Method	Classifier	Micro F1	Macro F1	Micro P	Macro P	Micro R	Macro R
Binary Relevance	NB	0.67998	0.62674	0.89258	0.91802	0.55636	0.51201
	RF	**0.85914**	**0.84820**	**0.94674**	**0.94678**	**0.78859**	**0.77561**
	KNN	0.21398	0.18181	0.96295	0.68376	0.12679	0.11036
	SVC	0.83023	0.81636	1.0	0.95751	0.75051	0.73722
	LR	0.71783	0.66895	0.99358	0.94588	0.57877	0.53924
Classifier Chains	NB	0.68988	0.65303	0.84250	0.85783	0.60230	0.57352
	RF	**0.85848**	**0.84678**	**0.94791**	**0.94893**	**0.78832**	**0.77955**
	KNN	0.21398	0.18181	0.96571	0.69981	0.12679	0.11036
	SVC	0.82885	0.81462	0.98210	0.95903	0.74601	0.73217
	LR	0.70397	0.65539	0.99499	0.95012	0.55969	0.52116
Label Powerset	NB	0.47738	0.35602	0.89964	0.91638	0.32686	0.26086
	RF	**0.68644**	**0.62493**	**0.92877**	**0.93085**	**0.54600**	**0.50519**
	KNN	0.21398	0.18181	0.94683	0.65222	0.12679	0.11036
	SVC	0.64809	0.59287	0.96652	0.96691	0.49823	0.45651
	LR	0.54890	0.45857	0.95809	0.95509	0.38814	0.33230

Method	Hamming Loss	Micro F1
Binary Relevance	0.05078	0.85914
Classifier Chains	0.05078	0.85848
Label Powerset	0.10300	0.68644

Kết luận

- Xây dựng được bộ dữ liệu dành cho bài toán phân phân loại đánh giá của khách hàng trên sàn thương mại điện tử gồm hơn 3500 sample với 8 nhãn liên quan và 2 nhãn là 0 và 1
- Phương pháp Binary Relevance cho kết quả tốt nhất và phương pháp Classifier Chains có kết quả gần tương đương, đạt kết quả cao nhất trên mô hình Random Forest.

Hướng phát triển

- Thực hiện kết hợp giữa bài toán này và bài toán multi - class để phân loại các tốt xấu trên từng nhãn mục tiêu và đưa ra nhận xét tổng quan cho người dùng tham khảo.
- Sẽ sử dụng PhoBert, VisoBert kết hợp cùng các thuật toán Deep Learning cho kết quả được tốt hơn.

November2023

Thank You.