Xây dựng bộ dữ liệu về phân loại bình luận giả mạo trên các trang thương mại điện tử

Phan Vỹ Hào[1], Hà Minh Quân[2], and Trần Minh Quân[3]

[1] Trường Đại học Công nghệ Thông tin, Đại học Quốc gia TP.HCM
http://www.uit.edu.vn
19520524@gm.uit.edu.vn
[2] 19522076@gm.uit.edu.vn
[3] 18521288@ms.uit.edu.vn

Tóm tắt nội dung Đồ án này cung cấp một bộ dữ liệu tiếng Việt về đánh giá mua hàng, nhằm hướng đến giải quyết bài toán về phân loại bình luận giả mạo trên các sàn Thương mại Điện tử ở Việt Nam. Cụ thể, nhóm sinh viên đã thực hiện thu thập đáng kể các đánh giá của những người mua hàng trực tuyến ở Lazada. Quy trình thực thi được mô tả cụ thể trong báo cáo này. Bên cạnh đó, nhóm đã tham khảo những công trình có liên quan để nhằm đề xuất cách thức gán nhãn dữ liệu hợp lí bao gồm ba lớp dữ liệu (0 - thực tế, 1 - giả mạo, 2 - vô nghĩa). Nhóm cũng đề xuất các phương pháp xử lý và mô hình có thể áp dụng trên bộ dữ liệu này. Nhóm sinh viên tin rằng bộ dữ liệu có thể đóng góp một phần vào phát triển đẩy mạnh Học máy và Trí tuệ Nhân tạo ở Việt Nam vào không gian số, đặc biệt là giao dịch trực tuyến.

1 Giới thiệu

Xu hướng mua hàng qua mạng là không thể phủ nhận, đặc biệt là trong bối cảnh xã hội đang phải gánh chịu ảnh hưởng nặng nề bởi dịch bệnh cúm Vũ Hán. Khi đó, đánh giá (review) của một sản phẩm đóng vai trò then chốt ảnh hưởng đến quyết định mua hàng của một người tiêu dùng, bởi họ thường có xu hướng chọn đọc đánh giá về sản phẩm nhằm tìm hiểu xem sản phẩm đó có đáng tin cậy hay không và chất lượng ra sao. Về phía doanh nghiệp, các đánh giá giúp họ đưa ra quyết định và điều chỉnh phù hợp chiến lược kinh doanh.

Lợi dụng tính chất quan trọng đó, nhiều cá nhân và tổ chức cũng đã và đang thực hiện nhiều chiêu trò đánh giá sản phẩm một cách giả mạo (fake review) ngày một nhiều trên các sàn Thương Mại Điện Tử (TMĐT) với độ xác thực rất đáng quan ngại của chúng. Theo [1], trong bối cảnh nhu cầu mua sắm trực tuyến tăng hơn 57% vì đại dịch, số lượng các đánh giá giả mạo trên nền tảng này cũng tăng vọt hơn 70%. Điều này dấy lên một thực trạng rằng mặc dù ở mọi trang web TMĐT đều có hệ thống đánh giá bình luận, nhưng bây giờ hệ thống đánh giá đó ngày càng không đáng tin cậy.

Vấn đề bình luận giả mạo vốn đã được quan tâm từ năm 2007 ([3]) nhưng hiện nay vẫn đang hiện hữu, và nó có ảnh hưởng trực tiếp đến các chủ thể kinh

doanh hoạt động trên các sàn thương mại điện tử. Một nghiên cứu cho thấy có đến 82% người được hỏi cho biết họ đã đọc ít nhất một bài đánh giá giả mạo trong một năm qua, trong khi đó đọc đánh giá là một cách để họ hiểu biết thêm về chất lượng và giá trị của sản phẩm [4]. Thực trạng này có thể vô tình thúc đẩy doanh số của các sản phẩm kém chất lượng, và ngược lại, nó gây ra tổn hại cho các hoạt động kinh doanh của những người buôn bán hợp pháp.

Đứng ở vị trí một người học và làm về dữ liệu, chúng tôi hy vọng công trình này sẽ đóng góp một phần cho công cuộc tìm kiếm lời giải cho vấn đề đã nêu. Cụ thể hơn, trong nghiên cứu này, chúng tôi xây dựng bộ dữ liệu về phân loại bình luận giả mạo trên các trang thương mại điện tử bằng cách thu thập dữ liệu các bình luận trên trang thương mại điện tử Lazada bằng ngôn ngữ lập trình Python. Chúng tôi đã tham khảo phương thức gán nhãn dữ liệu ở nhiều công trình có liên quan khác, mà chúng tôi có nêu ở phần 2. Bảng 1 ở phần 3 sẽ mô tả vắn tắt bộ dữ liệu của nhóm. Quy trình thu thập dữ liệu và cách thức gán nhãn sẽ được trình bày kĩ hơn ở chương 4. Ở chương 5, chúng tôi sẽ đề xuất cài đặt những phương pháp xử lý dữ liệu và giải thuật Học máy phù hợp giải quyết bài toán phát hiện đánh giá giả mạo.

2 Công trình dữ liệu có liên quan

Trong phần này, chúng tôi sẽ trình bày những nghiên cứu liên quan đến lĩnh vực phát hiện đánh giá giả mạo.

2.1 Phát hiện đánh giá spam cho tiếng Việt

Trong bài nghiên cứu [2], tập trung chủ yếu vào nội dung phát hiện đánh giá rác. Nhưng nghiên cứu đó chưa phát hiện ra đánh giá giả mạo (fake review).

2.2 Đánh giá giả mạo

Trong [4], tác giả đã đưa ra nhận định về đánh giá giả mạo là nhằm mục đích bôi xấu thương hiệu của cá nhân hay tổ chức. Tuy nhiên, nó không đúng với mọi trường hợp, vì bình luận giả mạo cũng có thể cải thiện, gây sự chú ý, ảnh hưởng tích cực đến các cá nhân, tổ chức liên quan.

2.3 Phân tích ý kiến khách hàng

Trong [5], tác giả đề cập đến việc phân tích bình luận của khách hàng dưới mỗi sản phẩm, nhưng trong các bình luận đó có những bình luận giả mạo, không trung thực, dẫn đến kết quả bị sai lệch.

2.4 Tràn ngập đánh giá giả trên các sàn thương mại điện tử

Trong [1], tác giả đã đưa ra vấn đề, tình trạng hiện tại, hậu quả và cách nhận biết tình trạng đánh giá giả mạo trên các sàn thương mại điện tử. Nhưng vẫn chưa đưa ra cách giải quyết vấn đề triệt đó triệt để ngoài cách cấm bán hàng vĩnh viễn khi bị phát hiện .

3 Mô tả bộ dữ liệu

Bảng 1 mô tả các thông tin chi tiết về bộ dữ liệu mà nhóm đã thu thập. Trong đó, nhóm đặt tên cho bộ dữ liệu này là Lazada Products Comments.

Bảng 1. Bảng mô tả tóm tắt bộ dữ liệu.

Thông tin	Nội dung
Tên bộ dữ liệu	Lazada Products Comments
Nguồn thu thập và cách thức thu thập	Nguồn: đánh giá của những người đã mua hàng trên trang web lazada.vn Cách thức: ngôn ngữ lập trình Python, thực thi thủ công, giãn cách thời gian để tránh bị chặn
Số lượng thuộc tính	6
Thông tin tên các thuộc tính	**rating**: giá trị số từ [1-5], thể hiện số lượng sao mà người đánh giá cho điểm trên sản phẩm. **description**: mô tả của sản phẩm. **review**: nội dung đánh giá của người dùng. **titles**: tên sản phẩm được hiển thị trên trang web Lazada. **price**: giá tiền gốc của sản phẩm, chưa tính khuyến mãi. **fake**: có thể có ba giá trị: 0 (đánh giá thực sự), 1 (đánh giá giả mạo), 2 (đánh giá vô nghĩa)
Số điểm dữ liệu	1288

4 Phương pháp thực thi đề tài

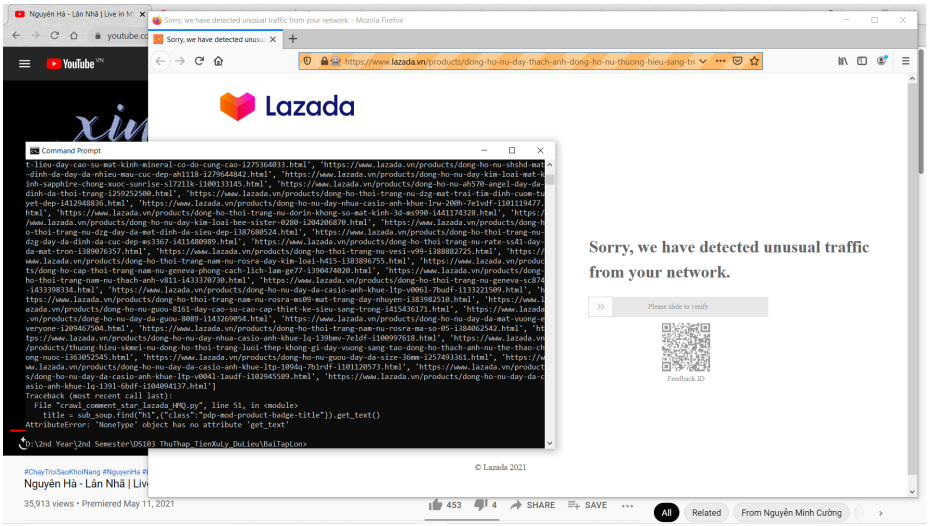
4.1 Môi trường thu thập

Chúng tôi sử dụng ngôn ngữ lập trình Python trên môi trường Python 3.7.4, và chạy trên môi trường máy tính xách tay cá nhân của mỗi thành viên trong nhóm. Chúng tôi thực thi Python bằng dòng lệnh trên Command Line như hình 1 để đạt được kết quả tốt nhất. Các thư viện sử dụng và cách thức thu thập cụ thể sẽ được trình bày ở phần 4.3.

4.2 Thách thức của công việc thu thập

Khó khăn lớn nhất mà nhóm chúng tôi gặp phải là rào cản bảo mật của hệ thống Lazada. Họ thường phát hiện truy cập và ngăn chặn khiến chúng tôi khó thực hiện việc lấy dữ liệu nhanh chóng. Các captcha thường xuyên hiển thị mỗi khi chúng tôi thực thi crawl hơn 10 sản phẩm, như trong hình 2 mô tả. Để vượt

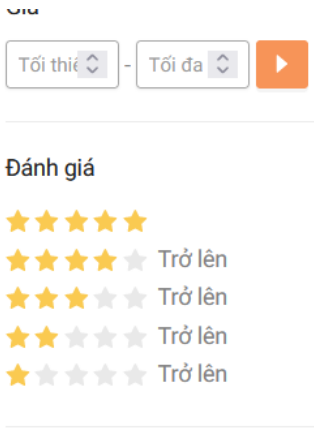
Hình 1. Thu thập dữ liệu từ ứng dụng Command Line trên Window 10.



Hình 2. Captcha bảo mật của Lazada.

qua captcha này, chúng tôi phải kéo thanh trượt sang phía phải, nhưng cũng chính khi đó lệnh Python gặp lỗi và phải ngừng lại.

Để vượt qua tình trạng này, chúng tôi thường xuyên giãn cách thời gian thực thi lệnh khoảng 5 đến 10 phút kể từ lần chạy trước. Ngoài ra, chúng tôi cũng lọc thủ công các sản phẩm đã được review (từ 1 sao trở lên, hình 3).



Hình 3. Bộ lọc số sao trên review của Lazada.

4.3 Phương pháp thu thập dữ liệu

4.3.1 Các thư viện cần dùng

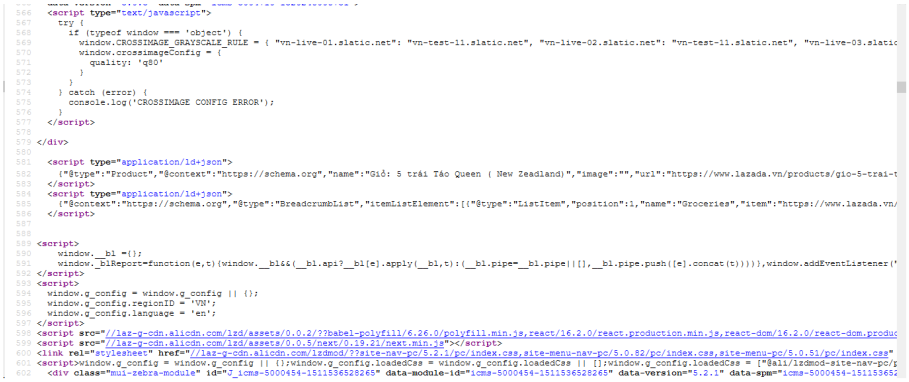
- **Selenium**: thu thập mã nguồn của lazada và nhờ vào `Driver Down` đã được tải sẵn để thao tác trực tiếp lên web khi code đang được chạy để giải quyết tình trạng Lazada xuất hiện captcha.
- **Beautiful Soup**: quét và tìm các thẻ chứa thông tin trong mã nguồn của trang web.
- **Json**: phân tích các phần tử trong thẻ được lấy về từ `Beautiful Soup` để trích xuất thông tin cần thiết cho dữ liệu.
- **Pandas**: đọc và lưu trữ dữ liệu có cấu trúc.

4.3.2 Phân tích mã nguồn (source code) Lazada

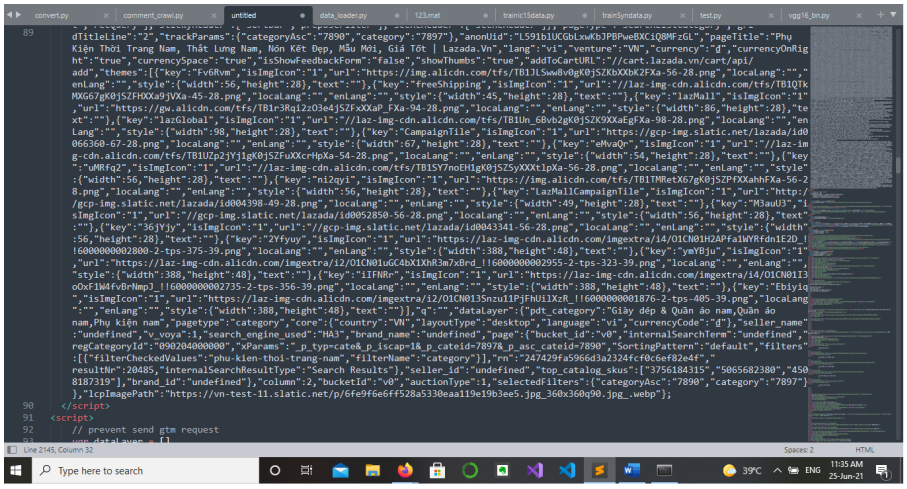
Không như các trang web khác với các thông tin được hiển thị ở trên web được sắp xếp trong mã nguồn dưới dạng các thẻ HTML tuần tự, mã nguồn của Lazada chứa nhiều thẻ `script` , và các thông tin cần thiết như : title, review, số sao, ... được gói vào các thẻ sript ở định dạng Json như hình 4.

4.3.3 Các bước thu thập dữ liệu

1. Sử dụng `selenium` để tải tất cả nội dung HTML trên Lazada bằng đường dẫn trực tiếp. Tất cả nội dung được crawl (lấy) về sẽ ở dạng dữ liệu thô như hình 5.
2. Đưa dữ liệu thô chưa xử lý về dạng dữ liệu có thể đọc được với `Beautiful Soup` về dạng `HTML` với các thẻ có thể trích xuất thông tin dễ dàng, và các thông tin cần thiết được lưu ở thẻ `<script type="application/ld+json">`.
3. Trích xuất các thông tin trong thẻ với thư viện `Json` đưa vào bảng dữ liệu và sắp xếp chúng lại thành dữ liệu có cấu trúc hoàn chỉnh.



Hình 4. Mã nguồn trên lazada.



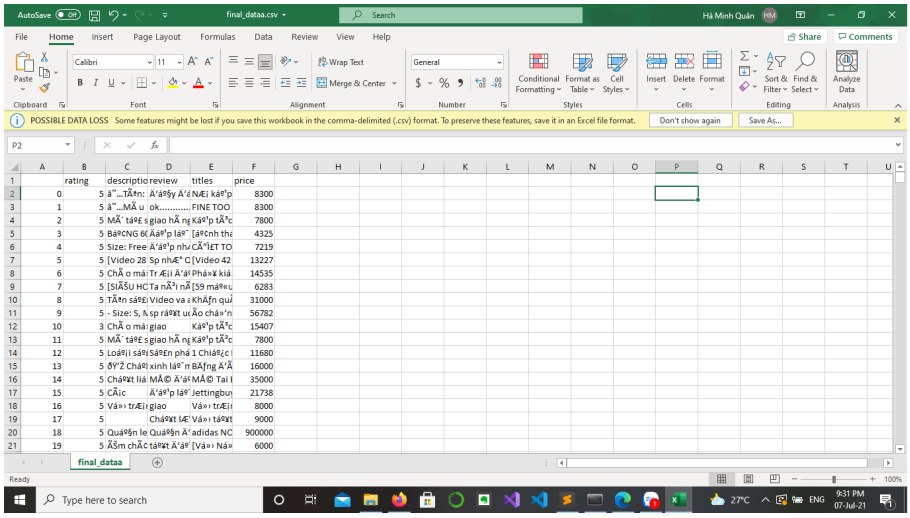
Hình 5. Raw data lazada

Ảnh mô tả dữ liệu đã được crawl như hình 6

4.4 Hướng dẫn gán nhãn

Từ việc tham khảo [2] cùng với sự chọn lọc để điều chỉnh phù hợp cho bộ dữ liệu của chúng tôi, chúng tôi quyết định ràng sẽ phân loại bộ dữ liệu thành ba nhóm:

1. **Đánh giá thực tế (0)**: Những đánh giá này mang lại ít nhất một thông tin hữu ích và đúng đắn về sản phẩm, giúp những người mua khác có thêm hiểu biết và tin tưởng khi đưa ra quyết định mua hàng.



	rating	descriptio	review	titles	price
0	5	â"...TÃ¤n:	Ã'á»§y Ã'é	NÆ¡i bá»'p	8300
1	5	â"...MÃ u	ok.....	FINE TOO	8300
2	5	MÃ¯ tá»£ s	giao hÃ ng	Ká»'p tÃ²c	7800
3	5	Bá»CNG 6(	Ã¡»'p lá»¯	[á»Cnh thá	4325
4	5	Size: Free	Ã'á»'p nhÆ	CÃ"lET TO	7219
5	5	[Video 28	Sp nhÆ° Ç	[Video 42	13227
6	5	ChÃ o má»	Tr Æ¡i Ã'á»	Phá»¥ kiá	14535
7	5	[SIÃŠU HC	Ta nÃ³i nÃ	[59 má»«u	6283
8	5	TÃ¡n sá»Ei	Video va ¿	KhÃ¡n quÃ	31000
9	5	- Size: S, M	sp rá»¥t u	Ão chá»'ng	56782
10	3	ChÃ o má»	giao	Ká»'p tÃ²c	15407
11	5	MÃ¯ tá»£ s	giao hÃ ng	Ká»'p tÃ²c	7800
12	5	Loáº¡i sáº¢n	Sáº£n phẩ	1 Chiáº¿c	11680
13	5	Ð¥Ž Chá»t	xinh lá»¯	nÐ£ng Ã'Ã	16000
14	5	Chá»¥t liá»	MÃ© Ã'á»MÃ© Tai		35000
15	5	CÃ¡c	Ã'á»'p lá»¯	Jettingbu	21738
16	5	Vá»› trÆ¡i	giao	Vá»› trÆ¡i	8000
17	5		Chá»¥t lÆ'	Vá»› tá»¥t	9000
18	5	Quá»§n le	Quá»§n Ã'adidas NC		900000
19	5	Ã§m chÃ¢	tá»¥t Ã'á»	[Vá»› Ná»	6000

Hình 6. Dataset

2. **Đánh giá giả mạo (không thực - 1)**: Những đánh giá này về mặt ngữ nghĩa thì mang nhiều thông tin nhưng sai lệch về sản phẩm. Hoặc là đánh giá dạng này sẽ phê bình quá khắt khe nhằm mục đích hạ bệ người bán hàng hoặc vì thù ghét. Hoặc là đánh giá dạng này cho nhận xét tích cực quá khích, khiến những người mua khác có thể mua phải hàng hóa kém chất lượng. Đây là lớp dữ liệu mang tính thách thức nhất vì khó xác định. Đánh giá dạng này xảy ra có thể vì một số lí do như:
   - Chính chủ doanh nghiệp hoặc người thân, bạn bè họ cố tình viết đánh giá tích cực cho sản phẩm của họ.
   - Nhân viên viết đánh giá để quảng bá cho sản phẩm của doanh nghiệp hoặc nhân viên cũ viết đánh giá tiêu cực để trả đũa công ty cũ.
3. **Đánh giá vô nghĩa (2)**: Những đánh giá này không mang lại bất cứ thông tin nào về cả sản phẩm lẫn người bán hàng, thường có rất ít kí tự hoặc kí tự vô nghĩa. Đánh giá dạng này xuất hiện thường xuyên trên các sàn thương mai bởi vì một số lí do như:
   - Người mua cố tình muốn đạt được các phần thưởng từ việc viết đánh giá nên viết vàu kì tự vô nghĩa như "OK, tốt,..." hoặc chèn các kí tự vô nghĩa để vượt qua ngưỡng mặc định về số lượng ký tự.
   - Người mua cố tình bình luận rác ở nhiều sản phẩm khác nhau với cùng nội dung bình luận hoặc không có nội dung.

Chúng tôi đã có ba ví dụ tương ứng với ba lớp dữ liệu, mô tả tại bảng 2.

Bảng 2. Minh họa cách gán nhãn dữ liệu.

rating	description	review	titles	price	class	Giải thích
5	Kiểu dáng lần này đản bảo chinh phục cả những vị khách khó tính nhất:children_crossing4màu phốiTone huyền bí: đen - đen Tone quyền lực: bạc - vàng Tone quý phái : bạc - bạc Tone tinh tế: bạc - hồngQuý khách vui lòng chọn đúng màu và size nhé !Hàng không bảo hành đổi trảCó giá sỉ sll cho cửa hàng	Rất đẹp và nam tính. Hoàn thiện ổn k bị sắc cạnh và kẹp lông tay. Haha. Chỉ lưu ý gỡ mắt cẩn thận k gãy chốt	Dây đeo Apple Watch thép RO.LEX cao cấp size 38/40 42/44mm	370 000	0	Đánh giá này mang lại thông tin cho người mua khác về diện mạo và cảm giác sử dụng (đẹp và nam tính). Ngoài ra cũng nói về thông tin sản phẩm như "hoàn thiện ổn". Ngoài ra có thêm chi tiết lưu ý (Chỉ lưu ý gỡ mắt cẩn thận k gãy chốt) chứng tỏ người mua có thực sự mua hàng và trả nghiệm sản phẩm.
5	- CAM KẾT 1 ĐỔI 1 NẾU HÀNG BỊ LỖI- CHẤT LƯỢNG ĐẢM BẢO- GIÁ CẢ HỢP LÝ- PHÙ HỢP VỚI MỌI LỨA- CÁC MẪU MỚI VỀ VÀĐANG LÀM MƯA LÀM GIÓCÁC BẠN CÓ MUỐN SỞ HỮU GU THỜI TRANG MỚI NHẤT KHÔNG???	đồng hồ bị xước mặt r	Đồng Hồ Thể Thao Mặt Vuông Nam Nữ TAIXUN FASHION SPORT 377	29 000	1	Đánh giá này cho điểm 5/5 thế nhưng nội dung lại không mang lại thông tin gì hữu ích cho người mua khác.
5	THÔNG SỐ KỸ THUẬTĐồng hồ hiệu Sun Rise của Nhật, mới 95%_ Giới tính: Nữ_ Kiểu dáng: Kiềng họa tiết hoa hồng, mặt số khảm traidây chất liệu đồi mồi, viền nạm đá Swarovski sang trọng, kim bọc vàng, cọc số học trò dễ nhìn_ Size mặt: ngang 22mm_ Size dây: <=18cm_ Độ rộng dây: 16mm_ Chất liệu vỏ: Mạ hợp kim_ Chất liệu dây: đồi mồi_ Bộ máy: Quartz Nhật bền bỉ_ Chống nước: 30 mét	tuyệt vời	Đồng hồ dạng kiềng nữ hiệu Grace của Nhật	266 000	2	Chỉ một từ "tuyệt vời" hoàn toàn không mang lại thông tin gì cho người đọc khác. Nó không giúp cho người khác hiểu biết thêm sản phẩm hoặc an tâm/ cẩn thận khi mua hàng từ người bán.

5 Đề xuất bài toán ứng dụng

5.1 Phương pháp tiền xử lý đề xuất

5.1.1 Bag of words

Bag of words là một thuật toán khá phổ biến trên thế giới, nó hoạt động bằng cách chuyển hóa các câu trong data để đưa về dạng vector các "số" trước khi đưa vào model để phân loại [6].

Thuật toán này không quan tâm tới từ ngữ có ý nghĩa như thế nào, sắp xếp trong câu ra sao,"Nó" hoạt động bằng cách thu thập tất cả các từ có trong bộ dữ liệu và chuẩn hóa từng câu về dạng vector của các "số" mà "số" đó biểu thị tần suất suất hiện của từ ngữ đó trong câu so với toàn bộ từ được thu thập.

Các bước thực hiện Bag of words như sau:

- Xây dựng vocabulary(tập hợp tất cả các từ vựng trong tập dữ liệu) từ document được sử dụng
- Vectorization(vector hóa) từ 2 câu chữ thành 2 vector với mỗi phần tử là tần suất xuất hiện trong câu

Chúng tôi lấy từ bảng 2 ra hai câu để minh họa.

- Câu thứ nhất là "**Rất đẹp và nam tính.**", câu thứ hai là "**Hoàn thiện ổn và không bị sắc cạnh.**".
- Tạo từ điển chứa các từ trong hai câu: {"rất": 1, "đẹp":1, "và": 2, "nam": 1, "tính": 1, "hoàn": 1, "thiện": 1, "ổn": 1, "không": 1, "bị": 1, "sắc": 1, "cạnh": 1 }.
- Tiến hành xây dựng véc tơ chứa số lần xuất hiện của mỗi từ cho mỗi câu. Do từ điển đang chứa 12 từ nên mỗi véc tơ sẽ có 12 phần tử:
  - $[1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0]$
  - $[0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1]$

5.1.2 TF-IDF

Phương pháp xử lý data dạng chữ phổ biến không kém là sử dụng một phương pháp thống kê có tên là TF-IDF, giá trị TF-IDF của một từ là một con số thu được qua thống kê thể hiện mức độ quan trọng của từ này trong một văn bản, mà bản thân văn bản đang xét nằm trong một tập hợp các văn bản.

Phương pháp này giải quyết các vấn đề trong văn bản có chứa nhiều từ lặp đi lặp lại nhiều lần nhưng thực chất nó không có nhiều ý nghĩa trong câu. Như trong tiếng anh có "is", "the"... tương tự tiếng việt có các từ như "là", "của", "cứ"... Chính vì vậy nếu chỉ xét theo tần số xuất hiện của từng từ thì việc phân loại văn bản rất có thể cho kết quả sai dẫn tỷ lệ chính xác sẽ thấp.

5.2 Thuật toán Máy học đề xuất

5.2.1 Hồi quy Logistic

Thuật toán hồi quy logistic là một thuật toán thống kê dựa trên hàm logistic hoặc sigmoid để phân loại đầu ra nhị phân (có/ không, 0/ 1,...). Mô hình này

nhận đầu vào là nhiều biến độc lập và đầu ra là một biến phụ thuộc (chỉ có hai giá trị cho biến phụ thuộc).

Mục tiêu của Hồi quy Logistic có thể tóm lại như công thức 1. Trong đó p là xác suất của một điểm dữ liệu thuộc lớp 0 hay 1. Các tham số h chính là các giá trị được tính toán được bởi mô hình, còn các giá trị x chính là giá trị của véc tơ đầu vào.

$$p = \frac{1}{1 + e^{-(h_0 + h_1 x_1 + \ldots + h_n x_n)}} \tag{1}$$

Đầu ra khi dự đoán sẽ bị chặn bởi một ngưỡng. Ví dụ như khi $p < 0.5$ thì lớp dữ liệu bằng 0, và bằng 1 khi $p \geq 0.5$.

5.2.2 Máy học véc tơ hỗ trợ (Support Vector Machine, SVM)

SVM là một thuật toán tìm kiếm siêu phẳng tối ưu để phân tách các điểm dữ liệu thuộc hai lớp riêng biệt. SVM đặc biệt phù hợp với bài toán này ở chỗ chúng dựa trên các giá trị của tổ hợp tuyến tính của các giá trị đặc trưng trên bộ dữ liệu huấn luyện. SVM thường phù hợp với các bài toán về văn bản, giọng nói, với lượng dữ liệu phức tạp.

Các véc tơ hỗ trợ (support vectors) là các điểm dữ liệu nằm ở rìa của các lớp dữ liệu mang nhãn khác nhau, đóng vai trò định hình cho siêu phẳng cần tìm. Ví dụ, đối với điểm dữ liệu hai chiều (chứa hai thuộc tính đầu vào), siêu phẳng cần tìm là một đường thẳng (hoặc đường cong) nằm chính giữa hai lớp dữ liệu, sao cho khoảng cách từ nó tới các véc tơ hỗ trợ của hai lớp lớn nhất.

6 Kết luận và hướng phát triển

6.1 Kết quả đạt được

Trong phạm vi đồ án môn học này, chúng tôi đã cùng nhau vận dụng những kiến thức chuyên sâu về thu thập dữ liệu và tiền xử lý vào việc thu thập một số lượng đáng kể các bình luận sản phẩm trên sàn thương mại điện tử Lazada.

Ngoài ra, chúng tôi cũng đề xuất cách gán nhãn để phục vụ cho bài toán học máy nhận diện binh luận giả mạo sau này, vốn là một bài toán mang tính chất cấp thiết trong việc tạo ra một môi trường mua sắm lành mạnh, an toàn cho người dân.

Nhóm cũng đã đề xuất hai giải thuật Học máy phù hợp là Hồi quy logistic và Máy học véc tơ hỗ trợ (SVM).

6.2 Hạn chế và thách thức

Bộ dữ liệu mà nhóm đã thu thập vẫn còn một số hạn chế nhất định về quy mô và số lượng.

Do phạm vi môn học và thời gian có hạn, nhóm vẫn chưa thể thực nghiệm mô hình học máy trên bộ dữ liệu.

6.3 Hướng phát triển của đồ án

Hướng đi chính trong tương lai của đồ án này sẽ là hướng vào việc nhận diện những đánh giá giả mạo rất khó phát hiện.

Tìm hiểu thêm các hướng thu thập dữ liệu và vượt qua hệ thống bảo mật của các sàn thương mại điện tử, nhằm tìm cách dồi dào hóa bộ dữ liệu. Có thể nói đây chính là rào cản lớn nhất cần phải vượt qua.

Thu thập ở nhiều trang thương mại điện tử lớn, nhỏ khác, thay vì chỉ tập trung vào Lazada như đồ án hiện tại.

Cài đặt và cải tiến các mô hình học máy trên đồ án.

7 Tài liệu tham khảo

[1] V. Digital. Tràn ngập đánh giá giả trên các sàn thương mại điện tử, 09 2020.

[2] T. H. H. Duong, V. D. Thang, and N. M. Vuong. Detecting Vietnamese Opinion Spam. *arXiv preprint*, 1905.06112, 2019.

[3] N. Jindal and B. Liu. Answers to the CME Questions Published in EAU-EBU Update Series Volume 4, Issue 6. *EAU-EBU Update Series*, 5(1):49–51, 2007.

[4] N. Rohr. Fake reviews: How to combat a growing online problem, 12 2020.

[5] H. T. Thành, T. T. Ánh, and H. T. Tuyền. Phân tích ý kiến khách hàng trong thương mại điện tử tiếp cận theo phương pháp học máy kết hợp kiểm định bootstrap. *Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á*, 31(11), 2021.

[6] T. thanhtt. Bag of Words (Bow) TF-IDF - Xử lý ngôn ngữ tự nhiên, 03 2021.