ĐÁNH GIÁ GIẢ MẠO: PHÂN TÍCH TỔNG QUAN VÀ PHƯƠNG PHÁP PHÁT HIỆN
FAKE REVIEWS: AN OVERVIEW AND DETECTION METHODS

Huỳnh Thị Khánh Vy[1], Phạm Thị Kim Anh[2], Vũ Thị Khánh Huyền[3], Trần Kim Thoa[4], Phạm Trần Gia Huyên[5], Trần Thành Công[*]

Trường Đại học Kinh tế-Tài chính Thành Phố Hồ Chí Minh, TP. Hồ Chí Minh, Việt Nam
[1]vyhtk20@uef.edu.vn
[2]anhptk220@uef.edu.vn
[3]huyenvtk20@uef.edu.vn
[4]thoatk20@uef.edu.vn
[5]huyenptg20@uef.edu.vn

***Tóm tắt**: Thương mại điện tử phát triển nhanh chóng trong những năm gần đây, cùng với đó thì các đánh giá giả mạo trong thương mại điện tử cũng xuất hiện ngày càng nhiều và chúng có ảnh hưởng tương đối lớn đến quyết định mua hàng của người tiêu dùng. Mặc dù hiện tại đã có rất nhiều giải pháp được đưa ra để giảm thiểu và ngăn chặn sự xuất hiện của đánh giá giả mạo nhưng các đánh giá mạo vẫn còn tồn tại trên các nền tảng thương mại điện tử. Mục tiêu chính của nghiên cứu là đưa ra cái nhìn tổng quan về đánh giá giả mạo hiện nay cùng với một số biện pháp để phát hiện chúng. Nghiên cứu cũng đề ra một số phương pháp để phát hiện các bài đánh giá giả mạo nhằm mang đến cho người tiêu dùng, doanh nghiệp một môi trường thương mại điện tử có tính cạnh tranh lành mạnh.*

***Từ khóa**: Đánh giá giả mạo, phương pháp phát hiện, NLP, máy học, thương mại điện tử.*

***Abstract**: With the fast expansion of e-commerce in recent years, phony evaluations in e-commerce have become increasingly common, and they have an attractive substantial effect on customer decision-making. Although the reality of many initiatives has already been adopted to avoid the emergence of fake reviews, there are still fraudulent reviews on e-commerce platforms. The paper's primary goal was to present an overview of existing fake reviews as well as some methods for detecting them. In order to offer consumers and companies a healthy competitive e-commerce environment, the research also suggests a variety of strategies for detecting fake reviews.*

***Keywords**: Fake reviews, detection methods, NLP, machine learning, e-commerce.*

1. Giới thiệu vấn đề

Trong những năm gần đây, thương mại điện tử ngày càng phát triển vượt bậc có thể thấy rằng những đánh giá giả mạo cũng đang ngày càng gia tăng. Tỷ lệ đánh giá giả mạo dao động từ 16% [1], 20% [2] và 25% [3] đến 33,3% [4]. Ngay từ năm 2012, khoảng 10,3% sản phẩm trực tuyến đã bị thao túng để xem xét [5]. Các đánh giá giả ngày càng tinh vi, chúng không chỉ được tạo ra bằng máy mà có thể do chính con người tạo ra. Điều này khiến cho việc phát hiện và ngăn chặn các đánh giá giả mạo ngày càng khó thực hiện. Nhưng hiện nay đã có một số giải pháp được đưa ra và thực hiện giúp giảm thiểu được tác hại của các đánh giá giả mạo mang lại.

2. Tổng quan về đánh giá giả mạo trong môi trường thương mại điện tử:

2.1. Định nghĩa:

Đánh giá giả mạo hay còn được gọi là đánh giá ảo theo Jana Valant-Thành viên của EPRS (10/2015) được định nghĩa là đánh giá tích cực, trung lập hoặc tiêu cực không phải là ý kiến thực của người tiêu dùng thực tế hoặc không phản ánh trải nghiệm chân thực của người tiêu dùng về sản phẩm, dịch vụ hoặc doanh nghiệp.

2.2. Các loại đánh giá giả mạo

- Đánh giá giả mạo tích cực: Là những đánh giá được viết theo hướng nói tốt cho sản phẩm, dịch vụ, nhãn hàng nào đó. Chúng

nhằm mục đích làm tăng ý định mua của người tiêu dùng chưa được trải nghiệm sản phẩm hay dịch vụ đó.

- Đánh giá giả mạo tiêu cực: Là những đánh giá được viết theo hướng nói không tốt cho sản phẩm, dịch vụ, nhãn hàng nào đó. Các đánh giá này nhằm mục đích làm giảm ý định mua của người tiêu dùng chưa được trải nghiệm sản phẩm hay dịch vụ đó. Loại đánh giá giả mạo này thường được tạo ra bởi những đối thủ cạnh tranh.

- Không phải đánh giá (non-reviews): các đánh giá không liên quan khác không chứa ý kiến (ví dụ: câu hỏi, câu trả lời và văn bản ngẫu nhiên không có nghĩa, không liên quan đến sản phẩm). Đây là loại đánh giá thường được dùng để nhận thưởng từ nền tảng. Ví dụ điển hình là ở Sàn thương mại điện tử Shopee loại đánh giá này xuất hiện rất nhiều.

Hình 1. *Một đánh giá không liên quan đến sản phẩm trên Shopee*

2.3. Phương pháp tạo ra đánh giá giả mạo

Đánh giá giả có thể được tạo theo hai cách chính sau:

- Đầu tiên, theo cách *do con người tạo ra* người bán sẽ trả tiền cho những người tạo nội dung để viết các bài đánh giá về sản phẩm - trong trường hợp này, tác giả bài đánh giá chưa bao giờ nhìn thấy sản phẩm nhưng vẫn viết về chúng. Bên cạnh đó người bán còn có thể thực hiện hoàn lại tiền/tặng quà khi người mua đăng các bài đánh giá khen ngợi để lôi kéo họ đăng bài đánh giá giả mạo. Có thể nói các bài đánh giá giả mạo do con người tạo ra được giao dịch như hàng hóa trong "thị trường hàng giả".

- Thứ hai, theo cách *do máy tính tạo ra* bằng cách sử dụng thuật toán tạo văn bản để tự động hóa việc tạo đánh giá giả mạo.

2.4. Hậu quả của đánh giá giả mạo đối với người tiêu dùng và doanh nghiệp

2.4.1. Ảnh hưởng đến các bên liên quan

Đánh giá giả mạo ảnh hưởng đáng kể đến các bên liên quan khác nhau, chẳng hạn như người tiêu dùng, người bán và nền tảng. Hơn nữa, chúng còn làm tăng sự không chắc chắn và đánh lừa người tiêu dùng.

Sự mất lòng tin của người tiêu dùng đối với các đánh giá sản phẩm trực tuyến vì các đánh giá giả mạo thường dẫn đến tâm lý khó chịu cho họ. Điều này đã làm tổn hại đến danh tiếng của doanh nghiệp, làm giảm ý định mua hàng của người tiêu dùng và giảm độ chính xác của hệ thống đưa ra sản phẩm khuyến nghị.

Đánh giá giả mạo có thể ảnh hưởng tích cực đến xếp hạng của sản phẩm (khi đánh giá giả mạo là tích cực) hoặc tiêu cực (khi đánh giá giả mạo là tiêu cực). Điều này là do các thuật toán của thị trường trực tuyến sử dụng các bài đánh giá làm tiêu chí để xác định thứ hạng của sản phẩm so với các sản phẩm khác trong cùng danh mục [6]. Thứ hạng của sản phẩm có thể sẽ bị thổi phồng hoặc giảm đi một cách quá mức so với thực tế do tác động của đánh giá giả mạo.

2.4.2. Ảnh hưởng đến chất lượng của các đánh giá sản phẩm trong môi trường thương mại điện tử nói chung

Vì đánh giá giả mạo là một dạng đánh giá sản phẩm trực tuyến bị bóp méo, nên các đánh giá này sẽ làm trầm trọng thêm sự phân tán của xếp hạng đánh giá [1].

Các bài đánh giá giả mạo làm giảm tính thông tin, chất lượng thông tin và tác dụng của bài đánh giá mang lại. Đánh giá giả mạo cũng làm giảm độ tin cậy của đánh giá sản phẩm và ảnh hưởng tiêu cực đến tính hữu ích của các bài đánh giá.

2.4.3. Ảnh hưởng đến toàn thị trường và xã hội

Các đánh giá giả mạo ảnh hưởng đến sự phát triển của các bài đánh giá sản phẩm trực

tuyến và các bên liên quan khác nhau, đồng thời ảnh hưởng đến toàn bộ thị trường hoặc xã hội. Đặc biệt, các đánh giá giả mạo làm suy yếu tính cạnh tranh của thị trường và có tác động tiêu cực đến phúc lợi xã hội.

3. Phân tích một số giải pháp hiện có

3.1. Phương pháp thủ công

Cách tiếp cận này dựa trên tiền đề rằng con người có thể phát hiện khi những người khác cư xử theo cách thức lừa đảo - tức là kiến thức về "tâm lý nói dối" (theo Bella M.DePaulo và cs., 1996). Sử dụng phương pháp thủ công trong giải quyết vấn đề đánh giá giả mạo tuy có mang lại hiệu quả nhưng không quá cao và phương pháp này có rất nhiều những nhược điểm. Theo một số nghiên cứu hiệu quả của phương pháp thủ công có kết quả như sau:

- Ott và cộng sự (2011) đã tuyển dụng một nhóm người để đánh giá xem một bài đánh giá có phải là giả mạo hay không, và nhận thấy rằng độ chính xác cao nhất đối với con người (65%) thấp hơn đáng kể so với mô hình máy học-Machine learning (86%).

- Trong một nghiên cứu khác của Sun và cs. (2013), độ chính xác do con người thực hiện là 52%, cho thấy con người rất khó phân biệt các đánh giá giả và đánh giá thực.

- Plotkina và cộng sự (2020) nhận thấy rằng con người có khả năng phát hiện chính xác là 57%, ngay cả khi bị ảnh hưởng bởi các tín hiệu thông tin về các bài đánh giá giả mạo.

3.1.1. Quy trình thực hiện

Sau khi tiến hành nghiên cứu từ nhiều nguồn, nhóm nghiên cứu chúng tôi đã thống nhất và đưa ra một quy trình phát hiện đánh giá giả mạo bằng phương pháp thủ công gồm các bước sau:

- *Thu thập dữ liệu đánh giá*

Tìm kiếm và tổng hợp lại các đánh giá của khách hàng về sản phẩm hoặc dịch vụ xuất hiện trên trang bán hàng cần kiểm tra bao gồm: nội dung đánh giá, thông tin khách hàng, thời gian đánh giá, thông tin sản phẩm.

- *Sàng lọc đánh giá:*

- Xem xét chi tiết nội dung của đánh giá: Ta cần xem xét các chi tiết trong nội dung của đánh giá bao gồm: độ dài, ngữ pháp, ý nghĩa và cách sử dụng từ ngữ. Đánh giá giả mạo thường có ngôn ngữ khá rập khuôn, sử dụng nhiều từ ngữ quá ca ngợi sản phẩm hoặc chỉ trích quá tiêu cực.

- Xem xét thời gian đăng đánh giá: Nếu có quá nhiều đánh giá được đăng trong một khoảng thời gian ngắn, đặc biệt là các đánh giá tương tự nhau, ta có thể nghi ngờ về tính chân thực của các đánh giá đó.

- Kiểm tra hồ sơ của người đánh giá: Thông qua hồ sơ có thể kiểm tra lịch sử đánh giá của họ. Nếu người dùng đó chỉ có một số ít đánh giá, các đánh giá tương tự nhau hoặc các đánh giá đều là tích cực/tiêu cực thì ta nên gắn cờ cho hồ sơ người dùng đó vì rất có thể họ là người chuyên viết các đánh giá giả mạo.

- Kiểm tra đối chiếu với các nguồn đánh giá khác: Ta nên đối chiếu thông tin trong đánh giá xếp hạng sản phẩm với các nguồn đánh giá xếp hạng khác để có cái nhìn tổng quát hơn về sản phẩm. Nếu có sự khác biệt lớn giữa các nguồn thì khả năng có đánh giá giả mạo khá cao.

- *Đưa ra nhận định*

Dựa trên kết quả của các bước trên, đưa ra kết luận về tính chính xác của các đánh giá được phân loại và xác minh.

3.1.2. Ưu điểm và nhược điểm của phương pháp thủ công

- Ưu điểm

- Có khả năng đánh giá review dựa trên ngữ cảnh thực tế.

- Khá linh hoạt.

- Có khả năng nhận biết tinh tế.

- Nhược điểm

- Tốn nhiều thời gian và công sức.

- Tốn nhiều chi phí thuê nhân sự.

- Không thể xử lý số lượng cực kỳ lớn của các bài đánh giá.

- Một khi các quy tắc của quy trình phát hiện đánh giá giả trở thành kiến thức phổ biến, những kẻ gửi đánh giá giả mạo sẽ thích nghi với chúng và thay đổi hành vi của chúng, khiến các quy tắc trở nên vô hiệu (Mattson và cộng sự, 2021).

3.2. Phương pháp tự động ứng dụng kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) và máy học (Machine Learning)

Theo một nghiên cứu của S.Feng, R.Banerjee, & Y.Choi (2012) cho thấy mức độ hiệu quả của một số mô hình máy học sử dụng NLP là khá cao. Nhóm nghiên cứu đã sử dụng dữ liệu từ bài luận của Mihalcea và Strapparava (2009), kết quả đạt độ chính xác lên tới 85,0%.

3.2.1. Quy trình thực hiện

Ứng dụng kỹ thuật xử lý ngôn ngữ tự nhiên và máy học có thể lọc ra đánh giá giả mạo thông qua việc huấn luyện một mô hình phân loại để xem một đánh giá có phải là đánh giá giả mạo hay không. Sau khi nghiên cứu một số mô hình phát hiện đánh giá giả mạo, nhóm nghiên cứu đưa ra một số bước cơ bản để xây dựng mô hình như sau:

1. Thu thập dữ liệu: Thu thập một lượng lớn dữ liệu đánh giá, bao gồm cả đánh giá thật và giả mạo. Đảm bảo dữ liệu đa dạng về nội dung, ngôn ngữ và nguồn gốc.

2. Phân chia dữ liệu: Mô hình được đào tạo và đánh giá bằng cách sử dụng phân chia 80/20 thường được áp dụng, có nghĩa là 80% tập dữ liệu được sử dụng để đào tạo mô hình (training) và 20% còn lại được giữ lại cho mục đích đánh giá (testing). Nói cách khác, bộ kiểm tra chứa các mẫu mà mô hình không được tiếp xúc trong quá trình đào tạo mô hình. [7]

3. Tiền xử lý dữ liệu: Sau khi đưa dữ liệu các đánh giá giả và thật vào thuật toán thì ta tiến hành quá trình tiền xử lý dữ liệu bao gồm việc chuẩn hóa chữ hoa/chữ thường, tách từ, loại bỏ từ dừng (stop words).

- Chuẩn hoá chữ hoa chữ thường là quá trình đưa tất cả các chữ cái trong một chuỗi văn bản về cùng một dạng, thông thường là chữ thường hoặc chữ hoa. Điều này giúp việc xử lý và phân tích văn bản trở nên dễ dàng hơn rất nhiều.

- Tách từ: Một trong những vấn đề nền tảng của việc phân tích văn bản là chia văn bản thành các đơn vị cơ sở nhỏ nhất đó chính là từ. Trong tiếng Anh việc tách từ sẽ khá đơn giản ta có thể dễ dàng tách các từ tiếng Anh dựa vào khoảng trắng vì về mặt ngữ nghĩa mỗi từ đều mang trong nó ngữ nghĩa riêng. Nhưng trong tiếng Việt đó là một thách thức vô cùng to lớn vì sự phức tạp tồn tại trong nó. Tiếng Việt có những nét khá giống với các ngôn ngữ ngữ âm (phonographic language) ở chỗ ranh giới từ khó xác định, có sự khác biệt về ngữ âm, ngữ pháp và nghĩa khi so sánh với các ngôn ngữ Ấn-Âu như tiếng Anh, tiếng Pháp…

Một số phương pháp sử dụng để tách từ phải kể đến là: Chuyển đổi trạng thái trọng số hữu hạn (Weighted Finite State Transducer), Độ hỗn loạn cực đại (Maximum Entropy – ME), So khớp từ dài nhất (Longest Matching), So khớp cực đại (Maximum Matching).

Trong nghiên cứu này chúng tôi xin giới thiệu công cụ để tách từ tiếng Việt. Đó là vnTokenizer của nhóm tác giả Lê Hồng Phượng, Nguyễn Thị Minh Huyền, Vũ Xuân Lương. Công cụ nêu trên phát triển dựa trên phương pháp So khớp cực đại (Maximum Matching). Độ chính xác khoảng 95%.

- Loại bỏ từ dừng: Từ dừng là những từ trong bất kỳ ngôn ngữ nào không bổ sung nhiều ý nghĩa cho một câu. Chúng có thể được bỏ qua một cách an toàn mà không làm mất đi ý nghĩa của câu. Các từ dừng thường được xóa khỏi văn bản trước khi huấn luyện mô hình máy học (machine learning model). Trong quá trình loại bỏ từ dừng các ký tự đặc biệt như <, >, *, @,…cũng sẽ bị loại bỏ. Ví dụ từ dừng: Trong tiếng Anh là the, is, at, which, on… Trong tiếng Việt gồm các từ nếu, thì, nhưng, là, các,…

Ta có thể loại bỏ từ dừng bằng cách so sánh với một từ điển từ dừng đã được xây dựng và loại bỏ chúng ra khỏi văn bản đầu vào bằng một thuật toán phát hiện từ dừng. Danh sách từ sẽ được tạo ra trong bước này.

3. Trích chọn đặc trưng (Feature Engineering): Chuyển đổi dữ liệu văn bản thành đặc trưng có thể dễ dàng sử dụng trong các mô hình máy học. Các phương pháp phổ biến gồm: Bag of Words, TF-IDF.

Trong một số các thuật toán máy học, giả sử các điểm dữ liệu được biểu diễn bằng các vector, được gọi là feature vector hay vector đặc trưng. Khi làm việc với các bài toán Machine Learning thực tế, nhìn chung chúng ta chỉ có được dữ liệu thô (raw) chưa qua chỉnh sửa, chọn lọc. Chúng ta cần phải tìm một phép biến đổi để loại ra những dữ liệu nhiễu (noise), và để đưa dữ liệu thô với số

chiều khác nhau về cùng một chuẩn (cùng là các vector hoặc ma trận). Dữ liệu chuẩn mới này phải đảm bảo giữ được những thông tin đặc trưng (features) cho dữ liệu thô ban đầu. Không những thế, tùy vào từng bài toán, ta cần thiết kế những phép biến đổi để có những features phù hợp. Quá trình quan trọng này được gọi là Feature Extraction, hoặc Feature Engineering, một số tài liệu tiếng Việt gọi nó là trích chọn đặc trưng. [8]

Một trong những phương pháp trích chọn đặc trưng có tên là Bag of Words (BoW) (Túi đựng Từ). Dưới đây là ví dụ để mọi người hiểu về phương pháp BoW.

Giả sử có bài toán phân loại tin rác. Ta nhận thấy rằng nếu một tin nào đó có chứa các từ như: khuyến mại, giảm giá, miễn phí, quà tặng… thì khả năng đó là một tin nhắn rác rất cao. Vậy phương pháp đơn giản nhất là đếm xem trong tin đó có bao nhiêu từ thuộc vào các từ trên, nếu nhiều hơn 1 ngưỡng nào đó thì ta quyết định đó là tin rác. Tuy nhiên với bài toán thực tế thì độ phức tạp cao hơn rất nhiều khi các từ có thể được viết dạng không dấu, bị cố tình viết sai chính tả, hoặc dùng ngôn ngữ teen.

Trong bước này mỗi câu trong văn bản đều sẽ được so sánh với danh sách từ đã được tạo ra trong bước 3. Từ đó có thể dựa vào số lượng các từ trong từng loại để làm các vector đặc trưng cho từng văn bản. Các vectơ này có thể được sử dụng trong các thuật toán ML để phân loại và dự đoán dữ liệu.

5. Lựa chọn và huấn luyện mô hình: Chọn một hoặc nhiều mô hình phù hợp để phân loại đánh giá giả và thật, ví dụ: Logistic Regression, Naive Bayes, Support Vector Machines (SVM), Random Forest,...

6. Đánh giá và tinh chỉnh mô hình: Sử dụng tập kiểm tra để đánh giá hiệu suất của mô hình. Đưa ra các chỉ số như độ chính xác (precision), độ nhạy (recall), độ đặc hiệu (support) và F1-score để đánh giá mô hình. Nếu cần thiết, tiến hành tinh chỉnh tham số mô hình để cải thiện hiệu suất.

7. Triển khai mô hình: Sau khi có mô hình phù hợp và hiệu suất chấp nhận được, triển khai mô hình vào hệ thống thực tế để tự động phát hiện và lọc các đánh giá.

3.2.2. Ưu điểm và nhược điểm của phương pháp tự động

- Ưu điểm

- Độ chính xác cao, nhất quán và có thể lặp lại.
- Xử lý được các tệp dữ liệu lớn.
- Tiết kiệm chi phí nhân công.
- Tiết kiệm thời gian: có thể chạy thử nghiệm nhiều lần trong thời gian ngắn.

- Nhược điểm

- Cần kỹ thuật cao và có kiến thức chuyên sâu về máy học.
- Độ phức tạp của ngôn ngữ có thể gây khó khăn cho các thuật toán NLP.
- Không thể xử lý linh hoạt với các ngữ cảnh mới.
- Có thể bị đánh lừa khi đánh giá có các kí tự đặc biệt như @, dấu chấm…

3.3. Phương pháp kết hợp tự động và thủ công

Phương pháp kết hợp tự động và thủ công là kết hợp sự hiểu biết giữa kỹ thuật xử lý ngôn ngữ tự nhiên và máy học cùng với sự hiểu biết của con người để kiểm tra các đánh giá trên các trang web và nền tảng thương mại điện tử.

Phương pháp tự động kết hợp thủ công trong giải quyết vấn đề đánh giá giả mạo đã mang lại nhiều hiệu quả cho các trang web bán hàng trực tuyến, kết quả này được chứng minh qua một số nghiên cứu nổi bật như:

- "Combining human and machine intelligence to identify and verify social media rumors"- nghiên cứu được thực hiện bởi Zubiaga và cộng sự vào năm 2016. Các tác giả sử dụng các thuật toán học máy để phát hiện các tin giả trên Twitter. Cụ thể, độ chính xác của phương pháp kết hợp là 90,9%, trong đó độ chính xác của phương pháp chỉ sử dụng thuật toán học máy là 86,7%.

- "Combining Machine Learning and Human Intelligence for Accurate Evaluation of Online Reviews"- nghiên cứu được thực hiện bởi Trabelsi và cộng sự vào năm 2018. Nghiên cứu này sử dụng một bộ dữ liệu gồm các đánh giá về khách sạn được thu thập từ trang web TripAdvisor. Cụ thể, độ chính xác của phương pháp kết hợp là 84,9%, trong đó

độ chính xác của phương pháp chỉ sử dụng thuật toán học máy là 77,6%.

Kết quả của các nghiên cứu này cho thấy rằng sự can thiệp của con người có thể giúp cải thiện độ chính xác của phương pháp phát hiện đánh giá giả mạo trên các trang web đánh giá.

3.3.1. Quy trình thực hiện

- *Giai đoạn 1: Giai đoạn làm việc tự động*

Sau khi thu thập dữ liệu đánh giá từ các trang web khác nhau. Dữ liệu bao gồm các thông tin về sản phẩm hoặc dịch vụ được đánh giá, thông tin về người đánh giá, nội dung đánh giá và thời gian đánh giá. Các thuật toán xử lý ngôn ngữ tự nhiên và máy học sẽ được sử dụng để phân tích về dữ liệu các đánh giá của người dùng, kiểm tra độ tin cậy của người đánh giá, tính khớp của nội dung đánh giá với sản phẩm hoặc dịch vụ được đánh giá, và tần suất của các đánh giá như cách thực hiện của phương pháp tự động được đề cập ở mục 3.2. Sau đó, các đánh giá có dấu hiệu giả mạo sẽ được đánh dấu.

Ở giai đoạn này nguồn lực làm việc chính để xử lý các đánh giá giả mạo là các thuật toán máy học làm việc tự động mà không có sự can thiệp của con người. Thời gian để giai đoạn 1 hoàn thành khá nhanh và nó giống như 1 chiếc lưới lọc hiệu quả tiết kiệm công sức của con người. Hiệu quả làm việc của nó có thể chiếm từ 70% đến 80% toàn bộ quy trình phát hiện đánh giá giả mạo.

- *Giai đoạn 2: Giai đoạn thủ công*

Sau khi bộ phận máy học hoàn thành giai đoạn 1 và lọc ra những đánh giá bị gắn cờ giả mạo, bộ phận chuyên gia sẽ tiến hành kiểm tra thủ công các đánh giá đó thêm lần nữa. Việc này giúp ta có thể xác nhận kết quả của mô hình máy học và đưa ra quyết định chính xác hơn trong việc loại bỏ hay giữ lại đánh giá đó.

Nếu có sự khác biệt giữa kết quả của mô hình và kiểm tra thủ công, ta cần cập nhật cho mô hình thông tin mới từ kiểm tra thủ công để có thể cải thiện được hiệu suất của mô hình trong tương lai.

Tuy sự can thiệp của con người trong phương pháp này là rất nhỏ ở những bước hoàn thiện cuối cùng, có thể nói những kinh nghiệm của các chuyên gia đã cải thiện phần lớn nhược điểm của phương pháp tự động.

3.3.2. Ưu điểm và nhược điểm của phương pháp kết hợp

- Ưu điểm:

Phương pháp kết hợp tự động và thủ công phát hiện đánh giá giả mạo là một phương pháp hiệu quả cao và đáng tin cậy:

- Tiết kiệm thời gian và nhân lực.
- Hiệu quả chi phí.
- Độ linh hoạt (Phù hợp với nhiều loại sản phẩm hoặc dịch vụ).
- Độ chính xác cao.

- Nhược điểm:

Bên cạnh những ưu điểm nổi bật của phương pháp kết hợp tự động và thủ công, trong quá trình hoạt động vẫn có những nhược điểm nhất định:

- Chi phí và thời gian có thể tăng lên do việc kết hợp phương pháp thủ công vào quy trình.
- Tốc độ xử lý sẽ bị giảm đi so với việc chỉ sử dụng phương pháp tự động.

4. Nghiên cứu quy trình giải quyết đánh giá giả mạo trên ReviewMeta

4.1. Giới thiệu về ReviewMeta

ReviewMeta là một công cụ phân tích sản phẩm của Amazon và lọc ra các bài đánh giá có vẻ đáng ngờ hoặc không tự nhiên để đưa ra đánh giá thực sự về sản phẩm mà bạn có thể tin tưởng.

4.2. Cách thức hoạt động

- Đầu tiên, hệ thống thu thập tất cả dữ liệu đánh giá có sẵn cho một sản phẩm cùng với thông tin của những người đánh giá đã để lại đánh giá.
- Tiếp theo, hệ thống chạy dữ liệu thông qua bộ phân tích ReviewMeta. Tất cả các bài đánh giá đều phải trải qua 12 bài kiểm tra khác nhau.
- Nó chỉ định điểm Đạt (Pass), Cảnh báo (Warn) hoặc Không đạt (Fail) cho các bài đánh giá. Ngoài điểm, báo cáo cũng hiển thị xếp hạng được điều chỉnh so với xếp hạng ban đầu.

- Báo cáo cũng hiển thị 10 bài đánh giá đáng tin cậy nhất và kém tin cậy nhất (hoặc ít hơn).

- Cuối cùng, báo cáo hiển thị chi tiết chuyên sâu về phân tích đánh giá để bạn có thể biết chính xác lý do tại sao một đánh giá được coi là trung thực hoặc đáng ngờ.

4.3. Ưu điểm của ReviewMeta

- Hệ thống sao chép và dán trên trang web rất dễ sử dụng.

- Việc chấm điểm khá đơn giản và bất kỳ người mua sắm nào cũng có thể nhanh chóng hiểu được.

- Hiển thị xếp hạng trung bình có nhiều khả năng hơn nếu các bài đánh giá đáng ngờ không tồn tại.

- Bạn có thể xem tất cả các số liệu được sử dụng để chấm điểm đánh giá.

4.4. Nhược điểm của RreviewMeta

- Chỉ hoạt động được với các sản phẩm trên Amazon.

- Độ chính xác chưa cao. Có một vài trường hợp người đánh giá thật lại bị cho là đánh giá giả.

5. Thực hiện quy trình phát hiện đánh giá giả mạo bằng phương pháp thủ công

5.1. Thu thập dữ liệu đánh giá

Bước đầu tiên trong quá trình nghiên cứu chúng tôi đã tổng hợp lại một số đánh giá của khách hàng về sản phẩm của shop Hanes trên Amazon. Những mục tiêu mà chúng tôi đề ra cho quá trình xem xét đánh giá bao gồm: nội dung đánh giá, thông tin khách hàng, thời gian đánh giá và thông tin của sản phẩm.

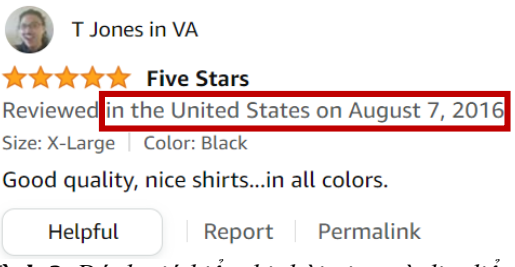
5.2. Sàng lọc đánh giá

- Xem xét chi tiết nội dung bài đánh giá:

Ở bước này chúng tôi bắt đầu với việc xem xét chi tiết của bài đánh giá qua các tiêu chí: độ dài, ngữ pháp, ý nghĩa và cách sử dụng từ ngữ.

Trong bài đánh giá được lấy từ một gian hàng trên Amazon tên là Hanes Store này chúng tôi nhận thấy có một số bất thường về nội dung như: Bài đánh giá quá ngắn không cung cấp được thông tin tham khảo cho khách hàng đến sau. Các từ trong bài đánh giá cũng khá phổ biến, trùng lặp thường được dùng trong các bài đánh giá giả mạo.

- Xem xét thời gian đăng đánh giá



Hình 2. *Đánh giá hiển thị thời gian và địa điểm đăng bài.*

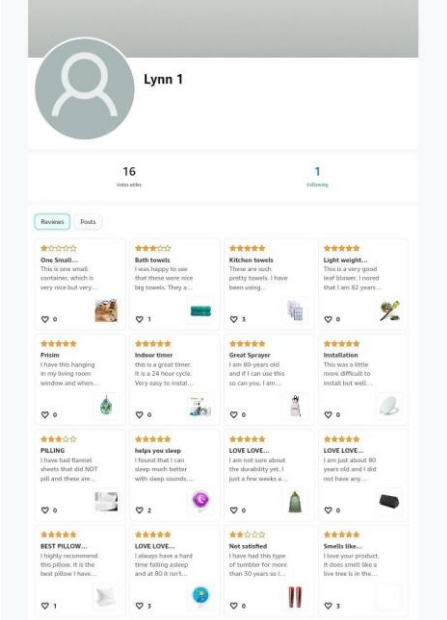
Mỗi đánh giá trên trang web sẽ có hiển thị kèm theo thời gian và địa điểm đánh giá. Dựa vào chi tiết này, các chuyên gia của chúng tôi sẽ kiểm tra đánh giá được đăng vào thời gian nào. Nếu nó được đăng vào cùng một ngày hoặc trong một khoảng thời gian ngắn, đó có thể là một dấu hiệu của đánh giá giả mạo.

Trên Amazon có rất nhiều đánh giá xuất hiện trong một thời gian ngắn, những đánh giá này do máy tạo ra và chúng xuất hiện hàng loạt với nội dung tương tự như nhau. Vì vậy khi xem xét đánh giá chúng tôi thường xem xét các đánh giá trong những khoảng thời gian gần nhau để tìm ra điểm bất thường và đưa ra kết luận.

- Kiểm tra hồ sơ của người đánh giá

Trong quá trình xem xét hồ sơ chúng tôi cũng quan tâm đến cả lịch sử đánh giá của của họ.

Qua hồ sơ của người dùng Amazon chúng tôi còn nhận thấy được những phiếu bầu hữu ích mà những khách hàng khác đã để lại, điều này cũng một phần nào thể hiện được độ tin cậy của người đánh giá. Đối với người đánh giá tên Lynn 1 lượt phiếu bầu hữu ích cho tác giả này cũng khá thấp chỉ ở mức 16 phiếu. Đến với phần lịch sử đánh giá, chúng tôi nhận thấy có rất nhiều đánh giá 5 sao với những từ ngữ lặp lại, rập khuôn xuất hiện nhiều lần ở nhiều bài đánh giá khác nhau.



Hình 3. *Hồ sơ của người dùng Lynn 1*

- Kiểm tra đối chiếu với nhiều nguồn khác

Sau quá trình xem xét và đưa ra kết luận với một số đánh giá bị nghi ngờ là giả, chúng tôi sử dụng một số phương pháp khác để chắc chắn hơn về kết luận này.

Trong nghiên cứu này chúng tôi dùng giải pháp đánh giá trên trang web ReviewMeta để đánh giá lại những đánh giá mà chúng tôi nhận thấy có dấu hiệu là giả.

Để biết được độ chính xác và sự khác biệt trong phương pháp thủ công và web đánh giá ReviewMeta chúng tôi tiến hành làm và so sánh một vài đánh giá như sau:

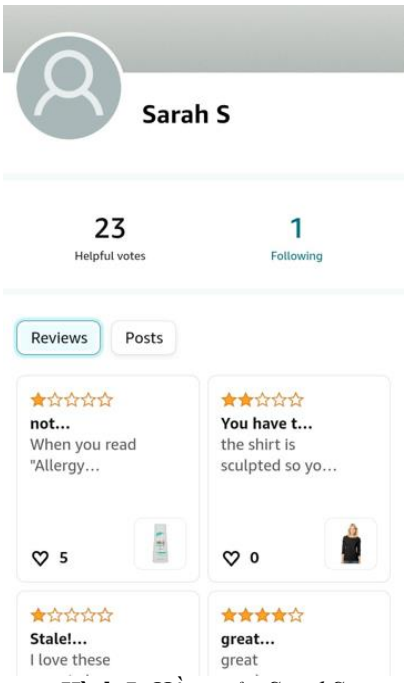
Bài đánh giá thứ nhất:

Hình 4. *Bài đánh giá của SarahS*

Đối với phương pháp thủ công:

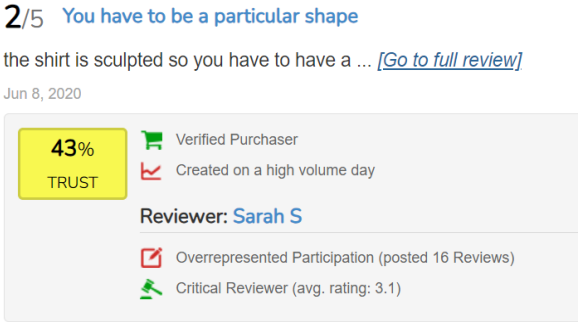
Xét về mặt nội dung và thời gian đánh giá: Chúng tôi cho rằng đây là một bài đánh giá giả mạo bởi khi nhìn vào nội dung chúng tôi nhận thấy có quá nhiều từ ngữ mang lại sự tiêu cực của bài đánh giá. Về mặt thời gian theo quan sát chúng tôi không nhận thấy bất kỳ bất thường nào về mặt thời gian.

Xét về phần hồ sơ của tác giả chúng tôi nhận thấy lịch sử đánh giả của người này cũng đều là những đánh giá không tốt. Người dùng này cũng chỉ nhận được tổng cộng 23 phiếu hữu ích mà người dùng khác bầu chọn cho các bài đánh giá.



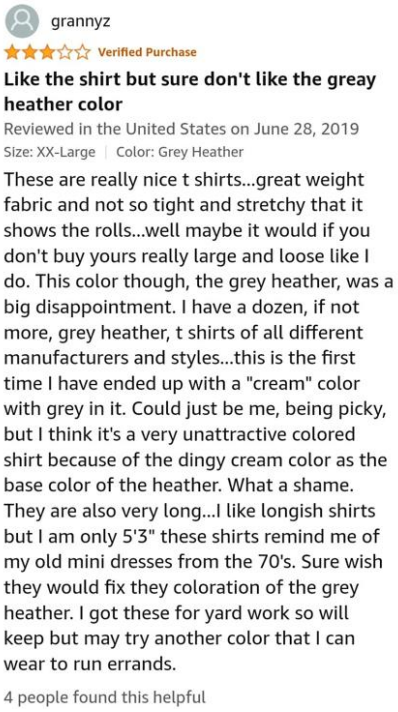
Hình 5. *Hồ sơ của SarahS*

Đối với web ReviewMeta: Bài đánh giá này được đánh giá là màu vàng mang ý nghĩa cảnh báo bởi đây có thể là đánh giá giả. Bài đánh giá ít được tin cậy với chỉ số tin cậy là 43%. Hồ sơ của người đánh giá được đánh giá ở mức trung bình 3.1



Hình 6. *Bài đánh giá của SarahS được xem xét bởi ReviewMeta*

Bài đánh giá thứ hai:



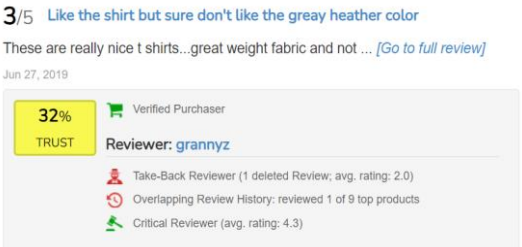
Hình 7. *Bài đánh giá của grannyz*

Đối với phương pháp thủ công:

Xét về mặt nội dung và thời gian đánh giá: Tuy bài đánh giá cho thấy được mặt lợi, mặt hại của sản phẩm và mang lại lợi ích cho các khách hàng khác đọc để có cái nhìn khách quan về sản phẩm nhưng đánh giá có nội dung quá dài. Về mặt thời gian theo quan sát chúng tôi không nhận thấy bất kỳ bất thường nào về mặt thời gian.

Xét về hồ sơ: Tổng lượt bầu chọn các bài đánh giá hữu ích là 199 phiếu và người dùng có 25 người theo dõi. Trong phần lịch sử các bài đánh giá, tác giả cũng để lại các đánh giá khá trực quan về sản phẩm chứ không có dấu hiệu quảng bá hay hướng tới điều tiêu cực với bất kỳ cửa hàng nào.

Đối với web đánh giá ReviewMeta:



Hình 8. *Bài đánh giá của grannyz được xem xét bởi ReviewMeta*

Bài đánh giá này được ReviewMeta đưa vào mức cảnh báo màu vàng là ít tin cậy. Theo thông tin mà ReviewMeta cho thấy người này đã có 1 bài đánh giá vi phạm và bị xóa, tuy nhiên người này có xếp hạng trung bình khá cao con số được đánh giá là 4.3. Bài đánh giá này rơi vào mức cảnh bảo bởi trong bài có một số từ trùng lặp và xuất hiện nhiều trong các bài đánh giá của sản phẩm này.

5.3. Đưa ra nhận định

Sau khi các chuyên gia hoàn thành và nhận kết quả từ hai bước nêu trên, bước 3 được thực hiện cuối cùng nhằm tổng hợp kết quả bao gồm: số lượng đánh giá bị đánh dấu, danh sách các tài khoản vi phạm, tên sản phẩm/ dịch vụ. Đồng thời để kết thúc vấn đề bộ phận các chuyên gia cũng cần đưa ra phương án xử lý và đưa ra hình phạt cho các hành vi đánh giá giả mạo bị phát hiện. Tuy phương pháp này khá tốn thời gian và công sức, nhưng kết quả mà nó mang lại cũng đóng góp một phần nhỏ cho các trang web trong vấn đề phát hiện và ngăn chặn "Đánh giá giả mạo".

6. Kết luận

Theo "Sách trắng Thương mại điện tử năm 2020", tại Việt Nam, 56% người tiêu dùng coi các bình luận, đánh giá trên mạng là lý do lựa chọn để mua hàng qua mạng. Điều này cho thấy các đánh giá thực sự rất quan trọng. Những nhận xét tích cực mang đến niềm tin cho người tiêu dùng, ngược lại những nhận xét tiêu cực mang đến sự lưỡng lự cho khách hàng. Các đánh giá trực tuyến mang lại hiệu quả rất to lớn nên những năm gần đây xuất hiện các đánh giá giả mạo gây xáo trộn trong quá trình lựa chọn và đưa ra quyết định mua sắm trên các nền tảng thương mại điện tử.

Hiện nay có một số phương pháp được đưa ra như: phương pháp thủ công, phương pháp phát hiện tự động áp dụng kỹ thuật NLP và máy học, hay phương pháp kết hợp tự động và thủ công. Các phương pháp này đã giúp giảm thiểu được phần nào các đánh giá giả mạo trên thị trường nhưng vẫn còn một số hạn chế cần được cải thiện.

Một số khuyến cáo mà chúng tôi đưa ra giúp giảm thiểu ngăn chặn đánh giá giả mạo.

- Đối với người tiêu dùng:

- Nắm bắt được những dấu hiệu cơ bản của các đánh giá giả để phòng tránh, ngăn chặn những tác động xấu của các đánh giá giả gây ra.

- Cảnh giác những đánh giá có dấu hiệu là giả và báo cáo với người quản trị trang bán hàng.

- Người tiêu dùng nên chọn mua ở những sàn thương mại điện tử chỉ cho phép để lại đánh giá, bình luận khi đã mua sản phẩm dịch vụ tại đó.

- Đối với doanh nghiệp:

- Tăng tương tác với khách hàng từ đó phát hiện sớm các bất thường trong hoạt động.

- Đưa ra những điều khoản, chính sách cho các bài đánh giá đảm bảo tính minh bạch cho các bài đánh giá.

- Sử dụng kết hợp nhiều giải pháp trong quá trình làm giảm và ngăn chặn các đánh giá giả để mang lại kết quả khách quan.

Tài liệu tham khảo

[1] M. Luca, G. Zervas (2016), "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud," *Management Science,* vol. 62, no. 12, pp. 3412 -3427.

[2] M. Schuckert, X. Liu, R. Law (2016), "Insights into suspicious online ratings: direct evidence from TripAdvisor," *Asia Pacific Journal of Tourism Research,* vol. 21, no. 3, pp. 259 -272.

[3] A. Munzel (2016), "Assisting consumers in detecting fake reviews: The Role of Identity Information Disclosure and Consensus," *Journal of Retailing and Consumer Services,* vol. 32, pp. 96 -108.

[4] S. Salehi -Esfahani, A.B. Ozturk (2018), "Negative reviews: Formation, spread, and halt of opportunistic behavior," *International Journal of Hospitality Management,* vol. 74, pp. 138 -146.

[5] T. Hennig-Thurau, K.P. Gwinner, G. Walsh and D.D. Gremler (2001), "Electronic Word- Of-Mouth Via Consumer-Opinion Platforms: What Motivates Consumers to Articulate Themselves on the Internet?," *Journal of Interactive Marketin,* vol. 18, no. 1, p. 38–52.

[6] N. Gobi & A. Rathinavelu (2019), "Analyzing cloud based reviews for product ranking using feature based clustering algorithm," *Cluster Computing,* vol. 22, pp. 6977-6984.

[7] Joni Salminen, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon-gyo Jung, Bernard J. Jansen (2022), "Creating and detecting fake reviews of online products," *Journal of Retailing and Consumer Services,* vol. 64.

[8] V. H. Tiệp, "Machine Learning cơ bản," 6 Feb 2017. [Online]. Available: https://machinelearningcoban.com/general/2017/02/06/featureengineering/#-gioi-thieu. [Accessed 18 Jun 2023].

[9] M. Ott, Y. Choi, C. Cardie, & J. T. Hancock (2011), "Finding deceptive opinion spam by any stretch of the imagination," *arXiv preprint arXiv,* p. 1107.4557.

[10] D.Plotkina, A.Munzel, & J.Pallud (2020), "llusions of truth—Experimental insights into human and algorithmic detections of fake online reviews," *Journal of Business Research,* vol. 109, pp. 511-523.

[11] H. Sun, A. Morales, & X. Yan (2013), "Synthetic review spamming and defense," *In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining,* pp. 1088-1096.

[12] C.Mattson, R.L.Bushardt & A.R.Artino Jr. (2021), "When a measure becomes a target, it ceases to be a good measure," *Journal of Graduate Medical Education,* vol. 13, no. 1, pp. 2-5.

[13] SOMAIT, "Ứng dụng của AI trong thương mại điện tử là gì?," 26 February 2023. [Online]. Available:

https://som.edu.vn/ung-dung-cua-ai-trong-thuong-mai-dien-tu-la-gi/.

[14] P. Sudhakaran, S. Hariharan, J. Lu (2016), "A framework investigating the online user reviews to measure the biasness for sentiment analysis," *Asian Journal of Information Technology,* vol. 15, no. 12, pp. 1890 -1898.

[15] A. Agnihotri, S. Bhattacharya (2016), "Online review helpfulness: Role of qualitative factors," *Psychology & Marketing,* vol. 33, no. 11, pp. 1006 - 1017.