Mô hình khai phá ý kiến và phân tích cảm xúc khách hàng trực tuyến trong ngành thực phẩm

A text-based model for opinion mining and sentiment analysis from online customer reviews in food industry

Nguyễn Đăng Lập Bằng¹, Nguyễn Văn Hồ², Hồ Trung Thành^{1*}

¹Trường Đại học Kinh tế - Luật, ĐHQG-HCM, Việt Nam

²Trường Đại học Kinh tế Thành phố Hồ Chí Minh, Việt Nam

*Tác giả liên hệ, Email: thanhht@uel.edu.vn

THÔNG TIN

DOI:10.46223/HCMCOUJS.
econ.vi.16.1.1388.2021

Ngày nhận: 18/05/2020

Ngày nhận lại: 22/06/2020

Duyệt đăng: 22/06/2020

Từ khóa:

học máy, khai phá ý kiến,
phân tích cảm xúc, thương
mại điện tử, ý kiến khách
hàng

Keywords:

customer reviews, e-
commerce, machine learning,
opinion mining, sentiment
analysis

TÓM TẮT

Với sự phát triển mạnh mẽ của công nghệ thông tin và Internet, các website Thương mại điện tử ra đời như một phương tiện hữu ích giúp khách hàng thực hiện mua hàng, đặt thực phẩm trực tuyến cũng như chia sẻ những trải nghiệm, bình luận và đánh giá sau giao dịch. Chính vì vậy để có thể thấu hiểu hành vi khách hàng thông qua ý kiến tích cực hay tiêu cực về sản phẩm và dịch vụ được trải nghiệm là một trong những vấn đề quan trọng. Giải pháp cho vấn đề này, nghiên cứu đề xuất phương pháp khai thác ý kiến và phân tích cảm xúc khách hàng thông qua việc thu thập tập dữ liệu là ý kiến bình luận của khách hàng trên website Foody.vn - một trang Thương mại điện tử hàng đầu trong lĩnh vực dịch vụ đặt hàng trực tuyến. Sau đó, tiến hành thực nghiệm bằng phương pháp học máy để khai phá ý kiến từ bình luận dạng văn bản của khách hàng và trực quan hóa kết quả hỗ trợ ra quyết định. Kết quả thực nghiệm cho thấy độ chính xác 90% của phương pháp đề xuất và kết quả khai thác được tập thông tin, tri thức tiềm ẩn có giá trị từ tập ngữ liệu nhằm giúp các cửa hàng, nhà quản trị hiểu được các ưu nhược điểm về sản phẩm, dịch vụ để cải thiện chiến lược kinh doanh tốt hơn.

ABSTRACT

In the rapid growth of technology and the Internet over recent years, e-commerce websites have been developed as a useful online media channel for users to easily make transactions such as online shopping and ordering food and drinks online, then share experience and feedbacks. Therefore, to be able to understand customer behaviors through positive or negative reviews about the products and services is an important desideratum. To offer a solution for this problem, the research proposes a method for customers opinion mining and sentiment analysis based on collecting data sets as customer reviews from the website Foody.vn - a top ranking website in the field of online ordering services. Machine learning models were conducted and evaluated to choose

the best model and then dashboards were created as visualizing results. The experimental results show that 90% accuracy of the proposed method; and valuable information and latent knowledge discovered from the corpus can support businessmen to capture the advantages and disadvantages of products and services and improve business with better strategies.

1. Giới thiệu

Những năm gần đây, chúng ta chứng kiến sự trỗi dậy của thị trường giao đồ ăn trực tuyến khi mà các ứng dụng giao đồ ăn ngày càng hoàn thiện hơn, thanh toán tiện dụng hơn. Mặt khác các mạng xã hội chuyên nhận xét về đồ ăn rất được nhiều người dùng truy cập như Foody, Now có rất nhiều dữ liệu các bình luận, đánh giá về đồ ăn của người tiêu dùng. Các thương hiệu đồ uống như trà sữa TocoToco, Bobabop rất được người dùng chú ý. Ý kiến khách hàng là những phản hồi mà khách hàng cảm nhận được sau khi sử dụng dịch vụ, sản phẩm của doanh nghiệp (Kumar, Desai, & Majumdar, 2016). Những ý kiến của khách hàng có thể tiêu cực hoặc tích cực. Dựa theo những nhận xét tích cực của khách hàng, doanh nghiệp sẽ biết được những ưu điểm của sản phẩm hay dịch vụ. Những ý kiến đó của khách hàng có thể dùng để quảng bá hay truyền thông. Bởi vậy các doanh nghiệp luôn luôn cải thiện chất lượng dịch vụ để có thể dẫn đầu.

Cạnh tranh giữa các doanh nghiệp ngày càng tăng. Theo Sharma, Agarwal, Dhir, và Sikka (2016), để chinh phục khách hàng thì không thể không tìm hiểu về nhu cầu của họ. Một trong bước để biết khách hàng có phù hợp là thu hút khách hàng trải nghiệm sản phẩm. Sau đó đánh giá sự thỏa mãn của khách hàng với sản phẩm hay dịch vụ. Tuy nhiên, vấn đề làm sao doanh nghiệp có thể biết được khách hàng đang hài lòng và không hài lòng về vấn đề này hay thương hiệu đang được người dùng sử dụng nhiều. Để giải quyết bài toán này nghiên cứu đề xuất giải pháp khai thác các bình luận của khách hàng về sản phẩm của các cửa hàng để lại trên trang web Foody. Tuy nhiên dữ liệu chỉ ở mức độ sơ cấp, và lượng dữ liệu rất lớn các doanh nghiệp không thể dựa vào dữ liệu thô này để ra quyết định được, họ cần biết được các tri thức được phân tích từ tập dữ liệu này. Do đó, chúng tôi đã áp dụng các phương pháp học máy để phân loại dữ liệu, xem bình luận nào là tích cực, bình luận nào là tiêu cực và dùng các phương pháp phân tích và dự đoán. Cuối cùng, nghiên cứu khai thác công cụ để trực quan hóa dữ liệu trên các báo cáo thông minh (dashboards). Kết quả nghiên cứu sẽ giúp các cửa hàng, nhà quản lý doanh nghiệp nắm bắt thông tin một cách dễ dàng và nhanh chóng, từ đó việc phát triển kinh doanh được cải thiện và nâng cao, chẳng hạn việc nâng cao sự hài lòng của khách hàng và giữ chân khách hàng tốt hơn.

Tiếp theo, là Mục 2 của bài báo, trình bày các cơ sở lý thuyết và các nghiên cứu liên quan. Mô hình nghiên cứu sẽ được trình bày chi tiết ở Mục 3. Mục 4 là kết quả thực nghiệm, đánh giá mô hình và trực quan hóa kết quả. Cuối cùng chúng tôi kết luận và đề xuất hướng phát triển ở Mục 5.

2. Cơ sở lý thuyết và các nghiên cứu liên quan

2.1. Phân tích cảm xúc tiếp cận theo xử lý ngôn ngữ tự nhiên

Các ý kiến, bình luận của khách hàng là dạng ngôn ngữ tự nhiên được viết ra (Eisenstein, 2019; Popescu & Etzioni, 2007). Trong một số nghiên cứu của Buche, Chandak, và Zadgaonkar (2013), Sun, Luo, và Chen (2017), Thanh và Phuc (2015) đã đưa ra một số phương pháp và kỹ thuật xử lý ngôn ngữ tự nhiên trong việc phân tích ý kiến và cảm xúc khách hàng thông qua bình luận trực tuyến. Như vậy, việc chuẩn bị tập dữ liệu để phân tích, ở đây là dữ liệu văn bản là các nội dung bình luận của khách hàng để lại sau khi trải nghiệm những sản phẩm và dịch của các cửa

hàng, có thể trên website, trên các trang mạng xã hội. Tiếp theo là tiền xử lý, ta tiến hành làm sạch dữ liệu, loại bỏ các kí tự đặc biệt, các dữ liệu rác, các dữ liệu không chuẩn hóa, chuẩn hóa dữ liệu về ngữ pháp ngữ nghĩa. Khảo sát phân tích dữ liệu, xem dữ liệu đã đầy đủ chưa, phân bố độ dài của nội dung. Giai đoạn này nghiên cứu sẽ phát họa khái quát tính chất, nội dung, số lượng của tập dữ liệu mình thu được. Lựa chọn các yếu tố đầu vào để phân tích, và dữ liệu ban đầu sẽ có rất nhiều chiều. Lựa chọn chiều nào thích hợp nhất để phân tích là việc rất quan trọng. Các chiều đầu vào càng chính xác thì kết quả phân tích sẽ có độ chính xác càng cao. Bước cuối cùng là đánh giá kết quả và triển khai dự án.

2.2. Phân tích cảm xúc tiếp cận theo phương pháp Học máy

Phân tích cảm xúc đã được định nghĩa là tính toán nghiên cứu ý kiến, tình cảm và cảm xúc thể hiện trong văn bản (Liu, 2012). Nói cách khác, khai thác ý kiến là một phương pháp trích xuất ý kiến của người đã tạo ra một tài liệu cụ thể gần đây đã trở thành mối quan tâm nghiên cứu lớn nhất trong mạng xã hội (Pang & Lee, 2008). Tầm quan trọng ngày càng tăng của phân tích tình cảm tăng dần cùng với sự phát triển của phương tiện truyền thông xã hội như đánh giá, thảo luận diễn đàn, và mạng xã hội. Đặc biệt, trong thời đại phát triển kỹ thuật số, chúng ta hiện có một khối lượng dữ liệu lớn được ghi lại dưới dạng văn bản để phân tích.

Học máy là một ứng dụng của Trí tuệ nhân tạo, là lĩnh vực giúp hệ thống tự động hiểu dữ liệu từ dữ liệu được đào tạo mà không cần lập trình cụ thể. Học máy tập trung vào vấn đề cung cấp hệ thống tự động hiểu dữ liệu và thực hiện các phép dự đoán. Học máy chia làm 4 phần (Das, Dey, Pal, & Roy, 2015): học có giám sát, học bán giám sát, học không giám sát và học củng cố.

Máy học có giám sát là thuật toán dự đoán dữ liệu đầu ra dựa vào các tập dữ liệu (dữ liệu đầu vào, kết quả đầu ra) đã biết từ trước. Có hai loại máy học có giám sát đó là phân loại và hồi quy. Phân loại thì dự đoán kết quả phân chia thành các nhóm dữ liệu có cùng tính chất, hồi quy thì cho ra kết quả dự đoán là một số thực cụ thể thay vì chỉ phân nhóm như học máy phân loại.

Máy học không giám sát là thuật toán dự đoán dữ liệu đầu ra dựa vào duy nhất tập dữ liệu đầu vào, dữ liệu đầu vào sẽ không được dán nhãn hoặc kết quả đầu ra. Thuật toán sẽ dựa vào cấu trúc dữ liệu để thực hiện lưu trữ và tính toán. Máy học không giám sát bao gồm phân nhóm và tích hợp. Thuật toán phân nhóm dựa sẽ phân nhóm toàn bộ dữ liệu thành các nhóm nhỏ dựa trên dự liên quan của các dữ liệu trong nhóm. Thuật toán tích hợp sẽ khai phá một số quy luật dựa trên nhiều dữ liệu cho trước.

Học bán giám sát là thuật toán kết hợp cả hai thuật toán có giám sát và không giám sát. Áp dụng với một phần tập dữ liệu đã được dán nhãn, phần còn lại thì không được dán nhãn.

Học củng cố là thuật toán giúp hệ thống tự động xác định các hành vi để đạt hiệu quả tối ưu nhất.

Trong nghiên cứu này, chúng tôi chọn phương pháp học có giám sát để áp dụng cho bài toán phân loại cảm xúc khách hàng dựa trên bình luận.

2.3. Thuật toán Hồi quy Logistic

Thuật toán Hồi quy Logistic (Hieu, 2018) thuộc học máy có giám sát để phân loại dữ liệu. Mô hình hồi quy Logistic áp dụng cho biến phụ thuộc là biến định tính hoặc định lượng chỉ có hai giá trị (có hoặc không) hay nhị phân là 0 hoặc 1. Điều này phù hợp với bài toán phân loại bình luận người dùng. Đầu ra của bài toán đó là xác định bình luận đó là tích cực hay tiêu cực. Phương trình tổng quát (hàm **Sigmoid**) hoặc hàm **Logistic**:

$$y = f(s) = \frac{1}{1 + e^{-s}} \quad (1)$$

Trong đó, $f(s)$ là xác suất xảy ra giá trị $y = 1$ hoặc $y = 0$, s là phương trình tuyến tính phụ thuộc vào các biến đầu vào. Phương trình mô hình đơn biến: $s = \alpha_0 + \alpha_1 x_1$, phương trình tuyến tính phụ thuộc vào duy nhất biến x_1 . Phương trình mô hình đa biến: $s = \alpha_0 + \alpha_1 x_1 + \dots + \alpha_n x_n$, phương trình tuyến tính phụ thuộc vào các biến x . Dạng ma trận khi $\alpha_0 = 0$ là

$$[\alpha_1 \alpha_2 \dots \alpha_n] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Hình 1. Đồ thị hàm Sigmoid (Hieu, 2018)

Đồ thị hàm số thể hiện:

$$s > 0 \Leftrightarrow e^{-s} < 1 \Leftrightarrow f(s) > 0.5$$

Chia làm hai lớp:

$$y = 0 \text{ nếu } s < 0$$

$$y = 1 \text{ nếu } s \geq 0$$

Các tính chất hàm **Logistic**:

- Miền xác định: Tất cả các số thực;
- Miền giá trị: $(0,1)$;
- Hàm liên tục;
- Hàm tăng trên miền xác định;
- Hàm đối xứng qua điểm $(0, \frac{1}{2})$, không phải hàm chẵn cũng không phải hàm lẻ;
- Bị giới hạn trên và dưới;
- Không có cực trị địa phương;

- Tiệm cận ngang: $y = 0$ và $y = 1$;
- Không có tiệm cận đứng;
- Mượt (smooth) nên có đạo hàm mọi nơi, có thể được lợi trong việc tối ưu hàm Sigmoid.

Giải thích:

- Giới hạn

$$\lim_{s \rightarrow -\infty} f(s) = \lim_{s \rightarrow -\infty} \left(\frac{1}{1 + e^{-s}} \right) = 0 \quad (2)$$

$$\lim_{s \rightarrow +\infty} f(s) = \lim_{s \rightarrow +\infty} \left(\frac{1}{1 + e^{-s}} \right) = 1 \quad (3)$$

Hàm mất mát (Jurafsky & Martin, 2008): hàm mất mát là hàm số xác định sự chênh lệch giữa đầu ra y dự đoán so với kết quả đầu ra y đã đúng (y dùng trong huấn luyện). Việc tối ưu hàm mất mát sẽ cho ra kết quả bài toán chính xác hơn

$$j(\theta) = \frac{1}{m} \sum_{i=1}^m (\text{cost}(h_{\theta}(x^{(i)}), y^{(i)})) = \frac{1}{m} \sum_{i=1}^m [-y^{(i)} \log(h_{\theta}(x^{(i)})) - (1 - y^{(i)}) \log(1 - h_{\theta}(x^{(i)}))] \quad (4)$$

$$\text{cost}(h_{\theta}(x), y) = \begin{cases} -\log(h_{\theta}(x)) & \text{if } y=1 \\ -\log(1 - h_{\theta}(x)) & \text{if } y=0 \end{cases}$$

3. Nghiên cứu thực nghiệm

3.1. Mô hình nghiên cứu tổng quan

Trong nghiên cứu này, trước tiên chúng tôi tiến hành thu thập dữ liệu thô từ trang web Foody. Sau đó dữ liệu thô được tiền xử lý và lấy mẫu, và gán nhãn trước khi tiến hành học máy. Dữ liệu lấy mẫu được chia thành ba nhóm: tập dữ liệu huấn luyện (training data), tập dữ liệu xác nhận (validation data) và tập dữ liệu kiểm tra (test data). Tập dữ liệu huấn luyện được sử dụng để thiết lập các mô hình học máy, bộ dữ liệu xác nhận được sử dụng để lặp lại và tinh chỉnh các mô hình được chọn, chúng tôi dựa trên kết quả phân loại chính xác trên dữ liệu tập kiểm tra để tìm ra mô hình học máy phù hợp nhất. Các bộ dữ liệu kiểm tra chỉ được sử dụng một lần là bước cuối cùng để báo cáo tỷ lệ lỗi ước tính cho dự đoán trong tương lai (Shmueli & Koppius, 2011). Hình 2 là tổng quan mô hình nghiên cứu chúng tôi đã thực hiện.

Hình 2. Mô hình nghiên cứu tổng quan

3.2. Thu thập dữ liệu

Các thư viện Beautiful Soup và Selenium trên ngôn ngữ Python được dùng để thu thập dữ liệu trên website. Việc thu thập dữ liệu dựa vào cấu trúc Hypertext Markup Language (HTML)

của trang Foody.vn. Muốn thu thập dữ liệu của thông tin nào ta tiến hành truy xuất dữ liệu tương ứng với thẻ HTML chứa thông tin đó. Tập dữ liệu thu thập được có 32,000 dòng, gồm các thông tin như tên cửa hàng, địa chỉ, tên khách hàng bình luận, thời gian bình luận, nội dung bình luận, tỷ lệ đánh giá của khách hàng đối với cửa hàng đó.

3.3. Tiền xử lý dữ liệu

Dữ liệu thu thập về sẽ có dạng thô, do chưa qua xử lý nên có thể dữ liệu bị rỗng, dữ liệu sai chính tả, dữ liệu quá ngắn, quá dài hoặc chứa các biểu tượng icon. Điều này sẽ gây ảnh hưởng đến kết quả của việc phân tích, vì vậy ta cần làm sạch dữ liệu.

- **Xóa các icon, kí tự đặc biệt:** các kí tự đặc biệt không mang ý nghĩa phân loại, mặt khác sẽ gây nhiễu trong quá trình phân tích. Chuyển tất cả về chữ thường: mỗi số, ký tự đặc biệt, ký tự là đại diện cho một dãy nhị phân trong bộ nhớ máy tính. Chữ in hoa sẽ có mã Unicode khác chữ in thường, về mặt ngữ nghĩa là giống nhau tuy nhiên máy tính sẽ không thể phân biệt dữ liệu đầu vào, dẫn đến có thể kết quả dự đoán bị ảnh hưởng. Vì vậy việc chuyển toàn bộ chữ về chữ thường là hợp lý cho hệ thống phân tích và dự đoán.

- **Chuyển dạng từ rõ nghĩa:** việc chuyển dạng từ rõ nghĩa là cần thiết cho bước tiền xử lý dữ liệu. Các bình luận trên Foody do người dùng bình luận tiếng Việt nên việc viết tắt hoặc sai chính tả. Chẳng hạn từ ko ngon (không ngon), vs (với), 15k (15000) ... hay dữ liệu không đồng bộ, không chuẩn hóa. Việc này sẽ ảnh hưởng gây nhiễu kết quả phân tích. Trong quá trình huấn luyện của học máy, dữ liệu đưa vào là “không ngon”, nhưng khi dự đoán dự liệu đầu ra, cụm từ “ko ngon” không xuất hiện trong quá trình huấn luyện, vì vậy sẽ khó thể nhận diện cảm xúc và dự đoán kết quả được.

- **Xóa dòng dữ liệu:** tập dữ liệu thu về sẽ có nhiều dữ liệu bị trống, dữ liệu trống không có ý nghĩa trong quá trình phân tích, gây tốn bộ nhớ lưu trữ.

3.4. Gán nhãn dữ liệu

Để thực hiện quá trình gán nhãn dữ liệu trước khi đưa vào huấn luyện, nghiên cứu áp dụng phương pháp phân loại cảm xúc theo điểm số đánh giá (Rating) của khách hàng (Liu, 2017) để phân chia tập dữ liệu đã thu thập được thành 2 bộ dữ liệu được gán nhãn theo quy tắc sau:

Rate ≤ 5 : bình luận nào đánh giá dưới 5 sao sẽ được dán nhãn là tiêu cực (negative).

Rate > 5 : bình luận nào đánh giá trên 5 sao sẽ được dán nhãn là tích cực (positive).

Kết quả gán nhãn cho thấy, chiếm đa số dữ liệu là các bình luận tích cực 70% so với tổng bình luận, bình luận tiêu cực chiếm 30% tổng bình luận.

3.5. Phương pháp biểu diễn văn bản

Trong học máy, máy tính không thể hiểu trực tiếp ngôn ngữ tự nhiên mà chỉ hiểu được ngôn ngữ khi chúng được biểu diễn dưới dạng không gian vector. Các chiều thuộc tính đầu vào sẽ được biểu diễn dưới dạng ma trận vector, có nhiều phương pháp để biểu diễn văn bản sang dạng ma trận vector chẳng hạn: cách truyền thống như mô hình Bag of N-grams, mô hình TF-IDF, mô hình chủ đề hay các cách cải tiến như các mô hình Word2Vec, GloVe, FastText (Sarkar, 2019). Trong nghiên cứu này, chúng tôi áp dụng hai phương pháp là Bag of N-grams và TF-IDF để thử nghiệm mô hình và biểu diễn dữ liệu.

Phương pháp Bag of word (BoW): mô hình BoW chỉ tập hợp tất cả các từ dạng một từ duy nhất, không chứa các cụm từ gồm nhiều từ ghép lại. Mô hình Bag of N-Grams sẽ giải quyết vấn đề này. Bag of N-grams sẽ thành lập một tập hợp các cụm từ gồm n-từ ghép lại với nhau tùy

thuộc vào nhu cầu. Ví dụ, xem xét ba bình luận sau đây “đồ ăn ngon quá”, “phục vụ quá tệ”, “đồ ăn không ngon”.

Hình 3. Sơ đồ cấu thành các n-grams

Bình luận 1: có tập từ

đồ ăn ăn ngon ngon quá

Bình luận 2: có tập từ

phục vụ vụ quá quá tệ

Bình luận 3: có tập từ

đồ ăn ăn không không ngon

Các từ nào trùng lặp thì mình chỉ lấy một từ, sau đó sắp xếp theo thứ tự bảng chữ cái, ta có ma trận:

		m thuộc tính						
		ăn ngon	ăn không	đồ ăn	không ngon	ngon quá	phục vụ	quá tệ
n bình luận	1	1	0	1	0	1	0	1
	2	0	0	0	0	0	1	1
	3	0	1	1	1	0	0	0

Hình 4. Ma trận từ theo mô hình Bag-of-word

Phương pháp TF-IDF: mô hình Bag of word n-grams gặp một vài vấn đề đối với tập dữ liệu lớn, đó là các từ có tần suất xuất hiện nhiều ở đa số các đoạn văn bản, nhưng không có ý nghĩa phân loại, ví dụ như các từ “này”, “đó”, “rất”, “cửa hàng”, ... Khi đó chỉ số TF-IDF sẽ được dùng để tính toán và phát hiện các từ có trọng số cao và thấp.

Bước 1: Tính TF theo công thức

$$TF(t, d) = \frac{df(d, t)}{|D|} \quad (5)$$

Bước 2: Tính IDF theo công thức

$$IDF(t, D) = \log \frac{|D|}{df(d, t)} \quad (6)$$

Bước 3: Tính TF-IDF theo công thức

$$TF-IDF(t, d, D) = tf(t, d) \frac{|D|}{df(d, t)} \quad (7)$$

Ở đây:

- $|D|$ là số lượng các văn bản trong tập ngữ liệu;
- $df(d, t)$ là số lượng các văn bản mà từ t xuất hiện;
- $tf(t, d)$ là tần suất các từ xuất hiện trong một văn bản.

Phân bố dữ liệu theo thuật toán BoW

Hình 5. Phân bố dữ liệu theo BoW

Trước tiên, thuật toán BoW biểu diễn giá trị của các thuộc tính bằng giá trị 1 và 0. Từ không gian nhiều chiều ta chuyển đổi thành không gian 2 chiều thì các giá trị sẽ phân bố dọc theo trục y, giá trị 0 của trục x cố định và giá trị của trục y sẽ biến thiên. Dữ liệu phân bố theo một đường thẳng như Hình 5 sẽ khó cho việc phân loại dữ liệu vì vậy nghiên cứu tiếp tục thực hiện phân bố dữ liệu theo *TF-IDF*.

Phân bố dữ liệu theo Phương pháp TF-IDF

Hình 6. Phân bố dữ liệu theo TF-IDF

Thuật toán TF-IDF không biểu diễn giá trị của các thuộc tính bằng giá trị 0 và 1 mà sẽ biểu diễn với giá trị trọng số TF-IDF đã tính. Chính vì vậy khi biểu diễn trên đồ thị giảm từ nhiều chiều sang 2 chiều, các giá trị của dữ liệu phân bố phụ thuộc cả hai chiều, khi trục x tăng thay đổi thì cũng kéo theo giá trị trục y thay đổi. Do vậy dữ liệu phân bố rời rạc và tách biệt hơn, việc này giúp quá trình phân loại sẽ dễ dàng hơn.

3.6. Dùng Google API để xác định tọa độ các cửa hàng

Dùng Google Map API để lấy tọa độ các cửa hàng, ứng dụng hiển thị vị trí các cửa hàng lên bản đồ để nhìn thấy tổng quan về sự phân bố dữ liệu ở các vị trí quận phường. Đăng ký tài khoản google map API để lấy mã truy cập, mã này đại diện cho duy nhất một tài khoản.

Kết quả thu được là tọa độ (kinh độ, vĩ độ) của từng vị trí các cửa hàng. Hình 7 là kết quả xác định tọa độ các cửa hàng. Kết quả này sẽ được áp dụng để hiển thị vị trí các cửa hàng trên bản đồ, báo cáo để thể hiện phân bố các cửa hàng theo cảm xúc.

Store	Address	latitude	longitude
["Three O'clock", "Coffee & Tea", "Nguyễn Thị Minh Khai"]	["462 Nguyễn Thị Minh Khai", "P. 2", "Quận 3", "TP. HCM"]	10.76791	106.6836
["Jollibee", "Hoa Sứ"]	["Tầng 2", "Coopmart Rạch Miễu", "48 Hoa Sứ", "P. 7", "Quận Phú Nhuận", "TP. HCM"]	10.79885	106.6893
["Hebe Tea & Coffee", "Sư Vạn Hạnh"]	["345 Sư Vạn Hạnh", "Quận 10", "TP. HCM"]	10.77004	106.6704
["Hot & Cold", "Trà Sữa & Xiên Que", "Vincom Thủ Đức"]	["Tầng 5", "Quầy 3", "Tầng L3", "Vincom Thủ Đức", "216 Võ Văn Ngân", "P. Bình Thọ", "Quận Thủ Đức", "TP. HCM"]	10.8501	106.7651
["Hot & Cold", "Trà Sữa & Xiên Que", "Hậu Giang"]	["64 Hậu Giang", "P. 6", "Quận 6", "TP. HCM"]	10.74962	106.647
["Hot & Cold", "Trà Sữa & Xiên Que", "Vincom Thủ Đức"]	["Tầng 5", "Quầy 3", "Tầng L3", "Vincom Thủ Đức", "216 Võ Văn Ngân", "P. Bình Thọ", "Quận Thủ Đức", "TP. HCM"]	10.8501	106.7651
["Trà Sữa GemTea House", "323 Lê Văn Khương"]	["323 Lê Văn Khương", "P. Hiệp Thành", "Quận 12", "TP. HCM"]	10.88917	106.6472
["Pizza Hut", "Võ Văn Ngân"]	["24 Võ Văn Ngân", "Quận Thủ Đức", "TP. HCM"]	10.85108	106.7564
["Trà Sữa Quyên"]	["48 Lê Lợi", "P. 4", "Quận Gò Vấp", "TP. HCM"]	10.82067	106.6871
["Chewy Chewy", "Cafe & Bánh", "Trần Hưng Đạo"]	["34 Trần Hưng Đạo", "P. Phạm Ngũ Lão", "Quận 1", "TP. HCM"]	10.76716	106.6944
["Baoz Dimsum Restaurant", "The Garden Mall"]	["Tầng Trệt", "Tầng Trệt The Garden Mall", "190 Hồng Bàng", "P. 12", "Quận 5", "TP. HCM"]	10.75458	106.6571
["Trà Sữa & Xiên Que Bo Mập"]	["29 Huỳnh Tịnh Của", "P. 8", "Quận 3", "TP. HCM"]	10.78878	106.6879
["Tuyệt Nhiệt Đới", "Milk Tea & Coffee", "Xóm Chiếu"]	["2A Xóm Chiếu", "P. 13", "Quận 4", "TP. HCM"]	10.7598	106.7116
["Pizza Hut", "Trương Định"]	["61 Trương Định", "Quận 3", "TP. HCM"]	10.7769	106.6904
["Trà Sữa A Little"]	["98F Lê Lai", "P. Bến Thành", "Quận 1", "TP. HCM"]	10.76993	106.6932
["Youth House Coffee", "Nguyễn Văn Nghi"]	["29 Nguyễn Văn Nghi", "P. 4", "Quận Gò Vấp", "TP. HCM"]	10.76791	106.6836
["Trà Sữa Cuộc Sống Dễ Dàng", "Nguyễn Tri Phương"]	["323 Nguyễn Tri Phương", "P. 5", "Quận 10", "TP. HCM"]	10.79885	106.6893
["Trà Sữa Hoa Hướng Dương", "Nguyễn Tri Phương"]	["339 Nguyễn Tri Phương", "Quận 10", "TP. HCM"]	10.77004	106.6704
["Trà Sữa Cuộc Sống Dễ Dàng", "Bầu Cát"]	["154 Bầu Cát", "P. 14", "Quận Tân Bình", "TP. HCM"]	10.8501	106.7651
["Trà Sữa Trần Châu Hong Kong", "Trần Phú"]	["354 Trần Phú (Ngã 4 Nguyễn Tri Phương)", "P. 7", "Quận 5", "TP. HCM"]	10.74962	106.647

Hình 7. Kết quả xác định tọa độ của hàng

4. Kết quả thực nghiệm

4.1. Đánh giá mô hình

Tập dữ liệu đã được huấn luyện bằng mô hình học máy, sử dụng các thuật toán gồm: Decision Tree, Naïve Bayes, hồi quy Logistic. Với kết quả đánh giá mô hình, xác suất dự đoán như Bảng 1, nhận xét cụ thể như sau:

$$F_score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (8)$$

- True Positive (TP): tổng số lượng tích cực dự đoán ĐÚNG với số lượng tích cực thực tế;
- False Positive (FP): tổng số lượng tích cực dự đoán SAI với số lượng tích cực thực tế;
- True Negative (TN): tổng số lượng tiêu cực dự đoán ĐÚNG với số lượng tiêu cực thực tế;
- False Negative (FN): tổng số lượng tiêu cực dự đoán SAI với số lượng tiêu cực thực tế.

Accuracy: độ chính xác trung bình các thuật toán, là tỷ lệ giữa kết quả dự đoán với dữ liệu thực tế. Cây quyết định và Hồi quy Logistic dự đoán 90%, nghĩa là trong 100 dữ liệu dự đoán thì hai mô hình này dự đoán đúng 90 dữ liệu so với kết quả thực tế.

Precision: được định nghĩa là số lượng dự đoán được thực hiện chính xác hoặc có liên quan trong số tất cả các dự đoán dựa trên lớp tích cực. Thuật toán Cây quyết định có độ chính xác là 90,075 % đối với dự đoán tích cực, có nghĩa là trong 100 dữ liệu tích cực thực tế thì mô hình dự đoán đúng 90,075 dữ liệu tích cực.

Recall: chỉ số thể hiện trong tất cả các trường hợp Positive, bao nhiêu trường hợp đã được dự đoán chính xác. Recall của Cây quyết định là 94.996% nghĩa là trong 100 dự đoán tích cực thì có khoảng 94.996 dự đoán là đúng.

F_score: có một số trường hợp chúng tôi muốn tối ưu hóa cân bằng cả độ chính xác và thu hồi. Điểm F1 là giá trị trung bình hài hòa của độ chính xác và thu hồi và giúp chúng tôi tối ưu hóa một bộ phân loại cho độ chính xác cân bằng và hiệu suất thu hồi.

Thời gian huấn luyện và dự đoán lâu nhất là thuật toán cây quyết định (huấn luyện 48.3s và dự đoán 328 ms), thời gian dự đoán nhanh nhất là hồi quy Logistic, thời gian huấn luyện nhanh nhất là Naïve Bayes bởi vì thuật toán này chạy dựa trên lý thuyết các biến dữ liệu độc lập với nhau. Độ chính xác cao nhất là 90% của thuật toán hồi quy Logistic, thấp nhất là Naïve Bayes với 78%. Như vậy có thể thấy Hồi quy Logistic là thuật toán tốt hơn so với các thuật toán còn lại khi xét tổng thể về tốc độ thực thi và độ chính xác.

Bảng 1.

Kết quả đánh giá mô hình

Thuật toán	Decision Tree		Naïve Bayes		Hồi quy Logistic	
	Positive	Negative	Positive	Negative	Positive	Negative
Precision	90.075	84.312	77.125	98.868	88.635	92.587
Recall	94.996	71.982	99.911	20.681	98.085	67.392
F_score	92.471	77.661	87.502	34.207	93.085	78.006
Accuracy	89%		78%		90%	
Thời gian huấn luyện	48.3 s		96.3 ms		1.79 s	
Thời gian dự đoán	328 ms		24.4 ms		11 ms	

4.2. Trực quan hóa kết quả khai phá ý kiến

Sau khi thực thi mô hình, nhằm dễ dàng theo dõi và ra quyết định, kết quả thực nghiệm được nghiên cứu tiếp tục trực quan hóa trên các báo cáo thông minh. Hình 7, 8 trình bày kết quả được phân tích và trực quan.

Hình 8. Báo cáo thống kê theo quận

Báo cáo thống kê theo quận thể hiện các số liệu tổng quan ở mỗi quận trong tất cả các năm từ 2012 đến nay. Báo cáo bao gồm số lượng cửa hàng mỗi quận, tổng các bình luận của khách hàng ở từng quận, tỷ lệ đánh giá trung bình theo các tiêu chí (địa điểm, dịch vụ, chất lượng, giá cả và không gian), cuối cùng là chỉ số đánh giá chung của người dùng theo quận.

Biểu đồ Trung bình đánh giá: thể hiện điểm đánh giá trung bình của tất cả các cửa hàng từng quận.

Biểu đồ Số lượng cửa hàng: thể hiện tổng số lượng quán của từng quận, hình tròn càng lớn chứng tỏ số lượng càng nhiều.

Biểu đồ Tổng số bình luận: tính tổng lượt bình luận của khách hàng trên một quận.

Biểu đồ Trung bình đánh giá theo tiêu chí: thể hiện điểm đánh giá theo địa điểm, dịch vụ, chất lượng, giá cả và không gian của tất cả các cửa hàng từng quận

Phân tích: Nhìn vào biểu đồ Trung bình đánh giá ta dễ dàng nhận thấy tỷ lệ đánh giá của các quận. Biểu đồ Số lượng cửa hàng kết hợp với các số liệu đánh giá điểm về dịch vụ của biểu đồ Trung bình đánh giá theo tiêu chí giúp người quản lý có thể suy luận các điểm mạnh yếu về vị trí, giá cả, chất lượng dịch vụ, phục vụ cũng như không gian quán. Như vậy các doanh nghiệp có thể xem xét kết quả này nhằm cải thiện dịch vụ để nhằm nâng cao lợi thế cạnh tranh.

Hình 9. Báo cáo Thống kê theo cửa hàng

Báo cáo thống kê theo cửa hàng thể hiện các số liệu chi tiết đến từng quán, báo cáo này sẽ giúp doanh nghiệp sẽ xác định được cụ thể các quán đối thủ, các đánh giá của khách hàng đối với quán đó, xác định vị trí và các quán xung quanh, ngoài ra báo cáo còn thể hiện thứ tự các cửa hàng được yêu thích nhất.

Biểu đồ Thống kê từng cửa hàng: thể hiện thông tin mỗi quận, mỗi phường sẽ có các quán nào, tên gì và điểm đánh giá là bao nhiêu.

Biểu đồ Định vị cửa hàng: biểu diễn vị trí các cửa hàng lên bản đồ, phân loại theo kết quả tích cực hay tiêu cực.

Biểu đồ Số người bình luận: tổng số người bình luận của phường, xác định được phường đó có bao nhiêu bình luận tốt, bao nhiêu bình luận xấu.

Biểu đồ Top cửa hàng nhiều bình luận tốt: hiển thị tên các quán có lượt bình luận tốt cao nhất.

Phân tích: Biểu đồ Top cửa hàng nhận nhiều bình luận tốt thể hiện rằng ở quận Tân Bình các cửa hàng như Royaltea, TocoToco, Bobapop, ... được nhiều khách hàng bình luận tích cực hơn so với các cửa hàng khác. Kết hợp các số liệu của biểu đồ Số người bình luận và Thống kê từng cửa hàng thì doanh nghiệp sẽ có góc nhìn tổng quan về đánh giá từng cửa hàng, cụ thể doanh nghiệp sẽ biết được ở mỗi quận thì phường nào có nhiều quán được đánh giá cao. Ngoài ra biểu đồ Định vị cửa hàng thể hiện phân bố đánh giá tiêu cực hay tích cực trực quan hóa trên bản đồ, giúp doanh nghiệp dễ dàng xác định các khu vực nhiều tích cực hoặc tiêu cực.

5. Kết luận và hướng phát triển

Trong nghiên cứu này, chúng tôi đã hoàn thành một giải pháp ứng dụng trong phân tích ngôn ngữ tự nhiên, cụ thể là phân tích cảm xúc khách hàng dựa trên bình luận được đăng tải trên

trang web Foody.vn. Giải pháp được thực hiện trên nhiều thuật toán để so sánh được các ưu và khuyết điểm của mô hình, và chọn ra mô hình tốt nhất thông qua độ đo F-Score. Đặc biệt tạo ra các báo cáo trực quan phục vụ nhu cầu ra quyết định của các doanh nghiệp. Giải quyết được bài toán trong thời kỳ bùng nổ dữ liệu đó là cung cấp các thông tin trải nghiệm khách hàng ở các địa điểm. Từ đó doanh nghiệp trong lĩnh vực đồ ăn thức uống sẽ có chiến lược để phát triển dịch vụ sản phẩm tốt hơn nhằm thu hút và giữ chân khách hàng tốt hơn. Ngoài ra nghiên cứu sẽ là tiền đề cho các ứng dụng phân tích dữ liệu, sử dụng giải pháp này để tích hợp vào các ứng dụng với mục đích khảo sát cảm xúc trải nghiệm khách hàng đối với tất cả sản phẩm dịch vụ khác nhau.

Trong các nghiên cứu tiếp theo, chúng tôi sẽ mở rộng bằng cách cài đặt hệ thống để tự động cập nhật dữ liệu. Dữ liệu sẽ tự trích xuất dữ liệu từ trên website và loại bỏ dữ liệu trùng lặp trước khi lưu vào cơ sở dữ liệu. Thu thập dữ liệu nhiều hơn nữa từ nhiều nguồn, và phát triển nghiên cứu theo hướng phân tích dữ liệu lớn (Bigdata). Ứng dụng triển khai các báo cáo phân tích ý kiến người dùng trên website, đặc biệt là trên thiết bị di động, giúp doanh nghiệp tiện lợi hơn trong việc xem báo cáo và ra quyết định tốt hơn.

Tài liệu tham khảo

- Agarwal, B., & Mittal, N. (2016). Machine learning approach for sentiment analysis. In *Prominent feature extraction for sentiment analysis* (pp. 21-45). Berlin, Germany: Springer.
- Boiy, E., & Moens, M. F. (2009). A machine learning approach to sentiment analysis in multilingual Web texts. *Information Retrieval*, 12(5), 526-558.
- Buche, A., Chandak, D., & Zadgaonkar, A. (2013). Opinion mining and analysis: A survey. *International Journal on Natural Language Computing (IJNLC)*, 2(3), 39-48.
- Das, S., Dey, A., Pal, A., & Roy, N. (2015). Applications of artificial intelligence in machine learning: Review and prospect. *International Journal of Computer Applications*, 115(9), 31-41.
- Eisenstein, J. (2019). *Introduction to natural language processing*. Cambridge, MA: The MIT Press.
- Gautam, G., & Yadav, D. (2014). Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In *2014 Seventh international conference on contemporary computing (IC3)* (pp. 437-442). Noida, India: IEEE.
- Hao, M., Rohrdantz, C., Janetzko, H., Dayal, U., Keim, D. A., Haug, L. E., & Hsu, M. C. (2011). Visual sentiment analysis on twitter data streams. In *2011 IEEE conference on visual analytics science and technology (VAST)* (pp. 277-278). Providence, MA: IEEE.
- Hieu, N. T. (2018). Dạy học hàm số logistic ở Mỹ [Teaching logistics functions in America]. *Tạp chí Khoa học*, 15(1), 60-67.
- Hosseini, B. (2004). *The internet encyclopedia*. Hoboken, NJ: John Wiley & Sons Inc.
- Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing: An introduction to speech recognition, computational linguistics and natural language processing*. Upper Saddle River, NJ: Prentice Hall.
- Kumar, K. S., Desai, J., & Majumdar, J. (2016). Opinion mining and sentiment analysis on online customer review. In *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)* (pp. 1-4). Chennai, India: IEEE.

- Liu, B. (2012). *Sentiment analysis and opinion mining*. New York, NY: Morgan & Claypool Publishers.
- Liu, B. (2017). Many facets of sentiment analysis. In *A practical guide to sentiment analysis* (pp. 11-39). Cham, Switzerland: Springer.
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1/2), 1-135.
- Popescu, A. M., & Etzioni, O. (2007). Extracting product features and opinions from reviews. In *Natural language processing and text mining* (pp. 9-28). London, UK: Springer.
- Sarkar, D. (2019). *Text analytics with Python: A practitioner's guide to natural language processing*. New York, NY: Apress.
- Sharma, V., Agarwal, A., Dhir, R., & Sikka, G. (2016). Sentiments mining and classification of music lyrics using SentiWordNet. In *2016 Symposium on Colossal Data Analysis and Networking (CDAN)* (pp. 1-6). Noida, India: IEEE.
- Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly*, 35(3), 553-572.
- Srinivasa-Desikan, B. (2018). *Natural language processing and computational linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Birmingham, UK: Packt Publishing Ltd.
- Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information Fusion*, 36(2017), 10-25.
- Thanh, H., & Phuc, D (2015). Analyzing users' interests with the temporal factor based on topic modeling. In *Asian conference on intelligent information and database systems* (pp. 105-115). Pa Tong, Thailand: Springer.
- Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... Zhou, Z. H. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1), 1-37.
- Ye, Z., Tafti, A. P., He, K. Y., Wang, K., & He, M. M. (2016). Sparktext: Biomedical text mining on big data framework. *PloS One*, 11(9), 1-15.