Tạp chí Khoa học Đại học Cần Thơ
Phần A: Khoa học tự nhiên, Công nghệ và Môi trường
ISSN 1859-2333 | e-ISSN 2815-5599

DOI:10.22144/ctujos.2024.263

TRÍCH XUẤT VÀ PHÂN TÍCH THÔNG TIN TRÊN GOOGLE VỀ SẢN PHẨM CHĂM SÓC SẮC ĐẸP

Võ Huỳnh Quang Hiếu* và Đỗ Phúc
Trường Đại học Công nghệ thông tin, Đại học Quốc gia Thành phố Hồ Chí Minh
**Tác giả liên hệ (Corresponding author): hieuvhq.15@grad.uit.edu.vn*

Thông tin chung (Article Information)

Nhận bài (Received): 14/09/2023
Sửa bài (Revised): 09/10/2023
Duyệt đăng (Accepted): 09/10/2023

Title: *Extract and analyze information on Google about beauty care products*

Author(s): *Vo Huynh Quang Hieu* and Do Phuc*

Affiliation(s): *Vietnam National University Ho Chi Minh City*

TÓM TẮT

Tại Việt Nam, có thể nói ngành chăm sóc sắc đẹp là một trong những lĩnh vực kinh doanh có mức độ cạnh tranh cao. Việc tìm hiểu những sản phẩm nào đang được quan tâm tìm kiếm phổ biến trên Google và nắm được số liệu dự đoán tìm kiếm tương lai trên Google giúp cho các nhà đầu tư, những người phụ trách phòng kinh doanh, tiếp thị những thông tin hữu ích để có thể nghiên cứu đưa ra các chiến lược tiếp thị kinh doanh cạnh tranh với đối thủ hoặc các nhà đầu tư cân nhắc đưa ra quyết định đầu tư của mình. Bài báo này thực hiện các phương pháp thu thập, tiền xử lý dữ liệu, phân tích và trích xuất thông tin nội dung của các trang web được phổ biến trên Google. Cuối cùng là dự đoán số liệu tìm kiếm trong tương lai trên Google bằng các thuật toán học máy. Kết quả thực nghiệm đã cho biết các sản phẩm nổi bật và đề xuất mô hình phù hợp dự đoán số liệu tìm kiếm tương lai trên Google.

Từ khoá: *Google tìm kiếm, mô hình hồi quy, phân loại văn bản, xử lý ngôn ngữ tự nhiên*

ABSTRACT

In Vietnam, the beauty care industry is one of the highly competitive business fields. Knowing which products are popularly searched for on Google and understanding future search prediction data on Google helps investors and those in charge of sales and marketing departments with informed information. Researching and developing business marketing strategies to compete with competitors or for investors to consider when making investment decisions is helpful. This article implements methods for collecting, preprocessing data, analyzing, and extracting content information from websites popular on Google. Finally, predict future search figures on Google using machine learning algorithms. Experimental results have shown outstanding products and proposed suitable models to predict future search data on Google.

Keywords: *Google seach, model regression, natural languague processing, text classification*

1. GIỚI THIỆU

Ở Việt Nam, theo khảo sát Customer Barometer của Google, 73% người dùng Internet tìm hiểu thông tin trực tuyến về sản phẩm/dịch vụ trước khi mua. 93% những người này sử dụng các công cụ tìm kiếm (Google, Yahoo, Bing,…) khi nghiên cứu các thông tin sản phẩm và doanh nghiệp kinh doanh (Huyền, 2019).

Trước khi cho ra sản phẩm mới các doanh nghiệp luôn thận trọng tiến hành tìm hiểu và phân tích về sản phẩm của đối thủ. Phương pháp trong bài báo là trích xuất thông tin (Zheng et al., 2015; Singh, 2018) nổi bật trên Google tìm kiếm và dùng học máy dự đoán số liệu tìm kiếm trong tương lai. Quá trình này giúp tiết kiệm thời gian, công sức và tiền bạc trong việc tìm hiểu, lựa chọn và phân tích đối tượng, từ đó giúp doanh nghiệp tối ưu hóa chiến lược tiếp thị. Đối với nhà đầu tư, việc này cũng giúp họ hạn chế rủi ro, cân nhắc xem sản phẩm có tiềm năng đáng đầu tư hay không.

Trong bài báo này, bộ dữ liệu được thu thập trên Google Search vào tháng 8/2023 với bộ từ khóa sản phẩm chăm sóc sắc đẹp: kem trị nám. Dữ liệu của mỗi bộ từ khóa bao gồm nội dung của 50 trang web được xếp hạng từ 1 tới 50 trên Google tìm kiếm tiếng Việt. Kết quả thực nghiệm của bài báo đã liệt kê ra các sản phẩm nổi bật trong top 10 và đề xuất mô hình học máy phù hợp để dự đoán số liệu tìm kiếm tương lai trên nền tảng Google.

Sau khi trích xuất có các thông tin tên sản phẩm, bộ công cụ lập kế hoạch từ khóa trên Google được sử dụng để tải về các số liệu tìm kiếm của những năm trước. Do bài toán là dự đoán số liệu tìm kiếm của các tháng tiếp theo trong tương lai nên một số thuật toán hồi quy được đề xuất như: Simple Linear Regression (SLR), Multi Linear Regression (MLR), Decision Tree Regression (DTR), Random Forest Regression (RFR), Support Vector Regression (SVR) để dự đoán.

Linear Regression (LR) là hồi quy tuyến tính phổ biến và có lẽ là loại hồi quy cơ bản nhất trong phân tích dự đoán (Saleh & Layous, 2022). Trên thực tế, khi làm việc tập dữ liệu với một biến dự đoán thì gọi đó là hồi quy tuyến tính đơn (phương trình 1) và nhiều biến dự báo gọi là hồi quy tuyến tính bội (phương trình 2). Nói một cách đơn giản, hồi quy tuyến tính sử dụng các hàm dự đoán tuyến tính có giá trị được ước tính từ dữ liệu trong mô hình.

$$y = b_0 + b_1 x_1 \quad (1)$$

$$y = b_0 + b_1 x_1 + b_2 x_2 \quad (2)$$

Decision Tree (DT) là mô hình cây quyết định được ứng dụng khá rộng rãi và hiệu quả trong học có giám sát (Sishi & Telukdarie, 2021) của hai mô hình phân loại và dự đoán.

Random Forest (RF) là một phương pháp dự đoán tích hợp nhiều cây quyết định. Nó có thể được sử dụng cho cả phân loại, hồi quy và dự báo chuỗi thời gian (Breiman, 2001; Dogru & Subasi, 2018; Gatera et al., 2023). Từ quan điểm tính toán, RFR hấp dẫn vì chúng: Xử lý tự nhiên cả hồi quy và phân loại (đa lớp), huấn luyện và dự đoán tương đối nhanh, chỉ phụ thuộc vào một hoặc hai tham số điều chỉnh, có ước lượng sẵn về lỗi tổng quát hóa, có thể được sử dụng trực tiếp cho các bài toán nhiều chiều (Cutler et al., 2012).

Support Vector (SV) là một mô hình học máy dùng dự đoán giá trị một biến số dựa trên các biến độc lập khác. Trong khi các mô hình hồi quy tuyến tính giảm thiểu sai số giữa giá trị thực tế và dự đoán thông qua đường phù hợp nhất, thì SVR cố gắng xây dựng đường biên tốt nhất để chia các điểm dữ liệu thành hai lớp và sau đó sử dụng các điểm dữ liệu trong mỗi lớp để dự đoán giá trị cho các điểm dữ liệu mới (Gatera et al., 2023).

Kết quả được so sánh bằng cách dùng độ đo để đánh giá hiệu suất mô hình. Trong bài báo này, ba độ đo được sử dụng: R-squared (R^2) thể hiện tỷ lệ phương sai, Mean Absolute Error (MAE) thể hiện độ sai số trung bình tuyệt đối và Mean Squared Error (MSE) thể hiện độ sai số trung bình bình phương thể hiện độ sai số trung bình tuyệt đối của dữ liệu mô hình trên cả tập huấn luyện và kiểm tra.

- *R^2*

Hệ số xác định có thể được hiểu là tỷ lệ phương sai trong biến phụ thuộc có thể dự đoán được từ các biến độc lập (Chicco et al., 2021) (phương trình 3).

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(X_i - Y_i)^2}{\sum_{i=1}^{m}(\overline{Y} - Y_i)^2} \quad (3)$$

- *MAE*

MAE thể hiện độ sai số trung bình tuyệt đối có thể được sử dụng nếu các ngoại lệ đại diện cho các phần dữ liệu bị hỏng. MAE ít nhạy cảm hơn với các ngoại lệ (Chicco et al., 2021) (phương trình 4).

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|X_i - Y_i| \quad (4)$$

(Giá trị xấu nhất là: +∞, giá trị tốt nhất là: 0)

- *MSE*

MSE có thể được sử dụng nếu có các ngoại lệ cần được phát hiện. Trên thực tế, MSE rất hữu dụng khi quy các trọng số lớn hơn cho các điểm như vậy, nhờ vào chuẩn L^2: rõ ràng, nếu cuối cùng mô hình đưa ra một dự đoán rất xấu, thì phần bình phương của hàm sẽ phóng đại lỗi (Chicco et al., 2021) (phương trình 5).

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2 \quad (5)$$

(Giá trị xấu nhất là: +∞, giá trị tốt nhất là: 0)

Đầu vào là dữ liệu thu thập trên Google tìm kiếm và số liệu tìm kiếm quá khứ trên Google. Đầu ra là từ khoá cần trích xuất và số liệu dự đoán tương lai trên Google.

Các đóng góp chính của bài báo này như sau: sử dụng phương pháp học máy để dự đoán số liệu tìm kiếm tương lai dựa trên dữ liệu văn bản từ trên Google.

Phần tiếp theo của bài báo trình bày các phương pháp: thu thập dữ liệu, tiền xử lý dữ liệu, phân tích và trích xuất thông tin, thực nghiệm với các mô hình, phân tích kết quả của các thuật toán và so sánh chúng với nhau, dự đoán số liệu tìm kiếm trong tương lai và thảo luận. Cuối cùng là phần kết luận.

2. PHƯƠNG PHÁP NGHIÊN CỨU

Bài báo được thực nghiệm bằng ngôn ngữ lập trình Python trên môi trường Jupyter Notebook. Các thư viện được sử dụng bao gồm: Beautiful Soup, NLTK, Google Search để thu thập nội dung của các trang web (Zheng et al., 2015; Bisong, 2019) liên quan với những từ khóa trên mạng tìm kiếm Google. Sau đó sử dụng các thư viện: NLTK, Matplotlib để tiền xử lý, trích xuất, phân tích và trực quan hóa thông tin (Bisong, 2019; Yogish et al., 2019). Cuối cùng là dùng các thư viện: Numpy, Pandas, LinearRegression, DecisionTreeRegressor, RandomForestRegressor, SVR, Scikit-learn để dự đoán số liệu tìm kiếm trên Google trong tương lai.

Bộ từ khóa tiếng Việt được chọn liên quan tới những sản phẩm trong lĩnh vực sản phẩm chăm sóc sắc đẹp: kem trị nám.

Công việc bao gồm các bước sau:

- Bước 1: Thu thập dữ liệu văn bản.
- Bước 2: Tiền xử lý dữ liệu ban đầu.
- Bước 3: Phân tích và xử lý dữ liệu sau khi tiền xử lý.
- Bước 4: Trích xuất và trực quan hóa thông tin.
- Bước 5: Thu thập số liệu tìm kiếm thực tế trên Google và thực nghiệm với năm thuật toán: SLR, MLR, DTR, RFR, SVR để so sánh.
- Bước 6: Dự đoán số liệu tìm kiếm trên Google trong tương lai.

2.1. Thu thập dữ liệu

Dữ liệu được thực nghiệm bằng ngôn ngữ Python trên môi trường Jupyter Notebook và sử dụng các thư viện của Python: Beautiful Soup, NLTK, Google Search để thu thập nội dung từ 50 trang web được xếp hạng từ 1 tới 50 trên Google Search ngôn ngữ tiếng Việt tại thời điểm tháng 8/2023 với bộ từ khoá: *kem trị nám.*

2.2. Tiền xử lý dữ liệu

Thư viện được sử dụng trong bước tiền xử lý dữ liệu là re và unicodedata2 để xử lý các tác vụ sau:

2.2.1. Chuyển đổi về chữ thường

Tất cả dữ liệu được chuyển đổi về chữ thường (Hao et al., 2020; Thanh & Anh, 2021) thực hiện điều này sẽ đổi các chữ về cùng một loại giúp giảm thiểu nhiễu khi làm việc với ngôn ngữ, tránh việc xử lý sai hoặc không chính xác.

Đầu vào là dữ liệu thu thập được bao gồm chữ hoa và chữ thường. Đầu ra là toàn bộ dữ liệu được chuyển về chữ thường.

2.2.2. Chuyển đổi về mã Unicode

Đối với văn bản tiếng Việt thì bảng mã Unicode (Hung, 2018) có 2 loại: Unicode tổ hợp và Unicode dựng sẵn. Nếu không đổi về cùng bảng mã trước khi xử lý dữ liệu, các ký tự có thể bị nhầm lẫn hoặc bị mất đi, dẫn đến các lỗi xử lý không mong muốn.

Đầu vào là dữ liệu thu thập được soạn thảo bao gồm nhiều bảng mã Unicode. Đầu ra là toàn bộ dữ liệu được chuyển về bảng mã Unicode dựng sẵn.

2.2.3. Loại bỏ URLs

Các URL không đem lại nhiều thông tin đáng giá về nội dung của văn bản, mà chỉ đơn giản dẫn người dùng đến các trang web liên quan, do đó cần phải loại bỏ để giảm thiểu việc làm nhiễu dữ liệu.

Đầu vào là dữ liệu thu thập được có nhiều URL được chèn vào trong bài viết. Đầu ra là toàn bộ dữ liệu được xoá bỏ tất cả các liên kết.

2.2.4. Loại bỏ ký tự đặc biệt và số

Các ký tự đặc biệt và số không mang nhiều ý nghĩa (Hao et al., 2020; Thanh et al., 2021) và thường được coi là nhiễu. Bằng cách loại bỏ chúng, dữ liệu giảm đi sự phức tạp, việc phân tích và xử lý trở nên nhanh chóng, dễ dàng và chính xác hơn.

Đầu vào là dữ liệu thu thập được có nhiều số liệu như giá, số điện thoại, v.v… trong bài viết. Đầu ra là toàn bộ dữ liệu được xoá bỏ tất cả các số.

2.2.5. Loại bỏ khoảng trắng thừa

Việc loại bỏ khoảng trắng thừa trong NLP là để làm sạch dữ liệu và giảm độ phức tạp khi xử lý dữ liệu.

Đầu vào là dữ liệu thu thập được có nhiều khoảng trắng thừa. Đầu ra là toàn bộ dữ liệu được xoá bỏ tất cả các khoảng trắng thừa.

2.3. Trích xuất và trực quan hoá thông tin

Trong bước trích xuất và trực quan hoá thông tin, các thư viện được sử dụng: re và unicodedata2 để xử lý các tác vụ sau:

2.3.1. Tìm ra cụm từ liên quan

Với mong muốn trích xuất tìm ra tên sản phẩm, chính vì vậy chúng tôi thực hiện phương pháp tìm ra cụm từ liên quan đi sau với từ khóa sản phẩm theo các bước sau: tách các từ thành danh sách, tạo danh sách trống để lưu trữ cụm từ, sau đó duyệt qua danh sách tìm ra các từ đứng sau từ nám.

Đầu vào là dữ tiệu đã xử lý ở bước tiền xử lý. Đầu ra là dữ liệu chứa các cụm từ như: nám nacos cực kỳ hiệu, nám da yanhee là một,...

2.3.2. Loại bỏ các cụm từ nằm trong stopwords

Việc loại bỏ các stopwords (Hao et al., 2020; Khang & Kiet, 2020; Thanh et al., 2021) có thể giúp giảm kích thước của tài liệu văn bản và cải thiện hiệu suất phân tích và trích xuất thông tin.

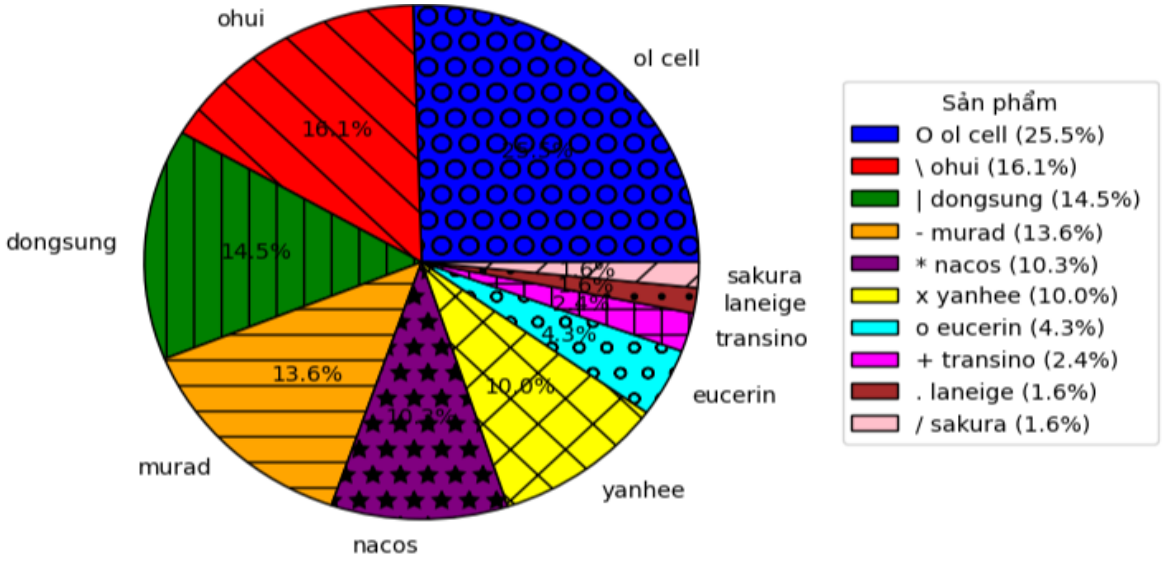
Chúng tôi dùng stopwords có sẵn (Duyet, 2017) và tạo thêm từ để hoàn thiện cho bộ stopwords tiếng Việt (số lượng 2.703 từ) dành riêng cho ngành sản phẩm chăm sóc sắc đẹp phục vụ trong thực nghiệm bài báo này.

Đầu vào là dữ liệu đã trích xuất ở mục 2.3.1 . Đầu ra là dữ liệu chứa các từ khoá như: nacos, yanhee, sakura,...

2.3.3. Trực quan hoá thông tin

Ở bước này, ngoài thư viện re và unicodedata2, thư viện matplotlib.pyplot được sử dụng để vẽ biểu đồ.

Sau khi có dữ liệu trích xuất ở mục 2.3.1 và 2.3.2, thông tin 10 tên sản phẩm được trực quan hóa và thống kê nhiều nhất bằng các phương pháp: tạo từ điển để lưu số lần xuất hiện của mỗi từ, đếm số lần xuất hiện của các từ, trích ra 10 từ xuất hiện nhiều nhất, vẽ biểu đồ. Biểu đồ trực quan hóa Hình 1 đã cho chúng tôi thấy được 10 tên sản phẩm của từ khóa "nám" được liệt kê từ cao đến thấp như sau: Ol cell (25,5%), Ohui (16,2%), Dongsung (14,5%), Murad (13,6%), Nacos (10,3%), Yanhee (10,0%), Eucerin (4,3%), Transino (2,4%), Laneige (1,6%), Sakura (1,6%).



Hình 1. Thống kê tên sản phẩm nằm trong top 10 của từ khóa: nám

2.4. Thu thập số liệu tìm kiếm quá khứ trên Google

Sau khi có tên các sản phẩm, bộ công cụ của Google được sử dụng để lập kế hoạch từ khoá nằm trong tài khoản Google Ads để lấy số liệu tìm kiếm quá khứ cho từng sản phẩm. Trong phạm vi thực nghiệm nghiên cứu bài báo này, sản phẩm Sakura được lựa chọn ngẫu nhiên. Bảng 1 mô tả số liệu (keyword volume) tìm kiếm thực tế Google với địa lý là Việt Nam và ngôn ngữ tìm kiếm là tiếng Việt của sản phẩm liên quan tới nám là: sakura.

Bảng 1. Số liệu tìm kiếm thực tế trên Google của từ khóa sakura từ tháng 01 đến tháng 12 của năm 2019 – 2022

Tháng	Năm 2019	Năm 2020	Năm 2021	Năm 2022
1	33.100	33.100	49.500	74.000
2	33.100	49.500	49.500	60.500
3	40.500	60.500	49.500	60.500
4	49.500	74.000	49.500	60.500
5	49.500	60.500	60.500	60.500
6	60.500	49.500	60.500	49.500
7	60.500	60.500	60.500	60.500
8	60.500	60.500	60.500	60.500
9	40.500	49.500	60.500	40.500
10	40.500	49.500	74.000	40.500
11	33.100	40.500	74.000	40.500
12	33.100	49.500	74.000	40.500

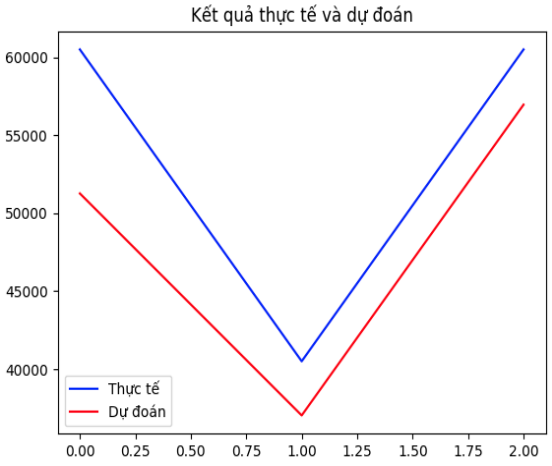
2.5. Thực nghiệm với các thuật toán

Tập dữ liệu này (Bảng 1) chia theo tỷ lệ: tập huấn luyện (80%), tập kiểm tra (20%) và áp dụng tỉ lệ này cho tất cả mô hình thuật toán thực nghiệm trong bài báo này. Các thư viện của Python: Numpy, Pandas, LinearRegression, DecisionTreeRegressor, RandomForestRegressor, SVR, Scikit-learn được dùng để thực nghiệm với các thuật toán.

Các bước thực nghiệm với các mô hình như sau: import các thư viện, đọc dữ liệu từ file số liệu tìm kiếm quá khứ Google, tách dữ liệu thành tập huấn luyện và kiểm tra, khởi tạo mô hình và huấn luyện trên tập dữ liệu huấn luyện, dự đoán giá trị cho tập huấn luyện và kiểm tra, vẽ biểu đồ so sánh giá trị dự đoán và thực tế trên tập kiểm tra, tính toán các chỉ số đánh giá mô hình: R-quared, MAE, MSE.

2.5.1. Linear Regression (LR)

Simple Linear Regression (SLR): là hồi quy tuyến tính đơn giản với mục tiêu là dự đoán số liệu tìm kiếm của biến phụ thuộc y dựa trên biến độc lập x (Saleh & Layous, 2022). Mối quan hệ giữa hai biến này được kiểm tra và sử dụng SLR để dự đoán số liệu tìm kiếm tương lai của những tháng tiếp theo trong năm 2023 dựa trên số liệu tìm kiếm trong quá khứ của từng tháng trong năm 2022.



Hình 2. Biểu đồ kết quả dự đoán (đường màu đỏ) và thực tế (đường màu xanh) sau khi áp dụng SLR của sản phẩm Sakura

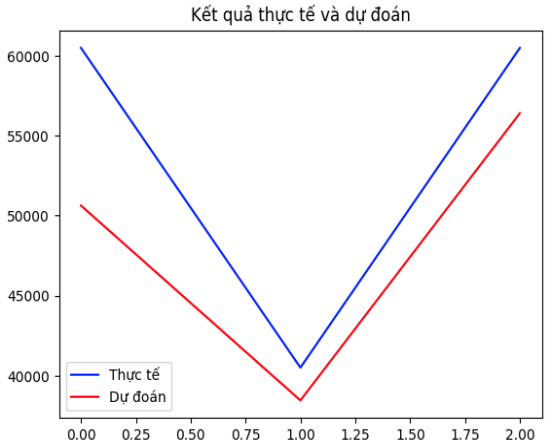
Ghi chú: Trục ox là chỉ số các điểm dữ liệu trong tập dữ liệu kiểm tra, trục oy là giá trị dữ liệu tìm kiếm Google

Bảng 2. Kết quả độ đo thuật toán SLR trên tập dữ liệu sản phẩm Sakura

	R^2 (train)	R^2 (test)	MAE (train)	MAE (test)	MSE (train)	MSE (test)
Độ đo	0,770	0,5878	4.234,5679	5.416,6666	28.929.783,9506	36.635.159,4650

Biểu đồ Hình 2 cho thấy kết quả dự đoán và thực tế của thuật toán SLR gần khớp với nhau trên tập dữ liệu sản phẩm Sakura. Bảng 2 là kết quả độ đo hiệu suất thuật toán SLR cho sản phẩm Sakura trên tập huấn luyện của có độ đo là: R^2 (0,770), MAE (4.234,5679), MSE (28.929.783,9506), trong khi tập kiểm tra có độ đo là: R-squared (0,5878), MAE (5.416,6666), MSE (36.635.159,4650).

Multi Linear Regression (MLR): theo kết quả trên cho thấy độ đo của SLR không được tốt. Chính vì vậy, MLR là hồi quy tuyến tính đa biến, tăng cường biến độc lập gồm nhiều đặc tính hơn thay vì chỉ một tính năng (Saleh & Layous, 2022) để xem kết quả có khả quan hơn hay không. Tập dữ liệu đã chỉ định cho tất cả các cột, trừ cột cuối cùng là số liệu tìm kiếm của năm 2022 vì đó là biến phụ thuộc.



Hình 3. Biểu đồ kết quả dự đoán (đường màu đỏ) và thực tế (đường màu xanh) sau khi áp dụng MLR của sản phẩm Sakura

Ghi chú: Trục ox là chỉ số các điểm dữ liệu trong tập dữ liệu kiểm tra, trục oy là giá trị dữ liệu tìm kiếm Google

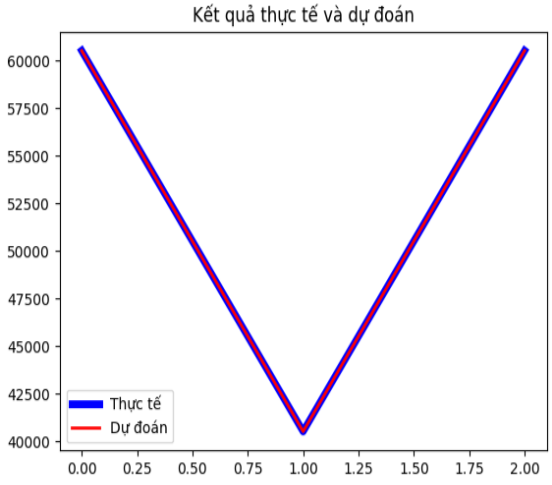
Bảng 3. Độ đo thuật toán MLR trên tập dữ liệu sản phẩm Sakura

	R^2 (train)	R^2 (test)	MAE (train)	MAE (test)	MSE (train)	MSE (test)
Độ đo	0,7770	0,5569	4.042,4241	5.332,1302	28.140.718,3639	39.379.482,6203

Biểu đồ Hình 3 là kết quả dự đoán và thực tế của thuật toán MLR gần khớp với nhau trên tập dữ liệu sản phẩm Sakura. Bảng 3 là kết quả độ đo hiệu suất của mô hình MLR cho sản phẩm Sakura trên tập huấn luyện có độ đo là: R^2 (0,7770), MAE (4.042,4241), MSE (28.140.718,3639); trong khi tập kiểm tra có độ đo là: R^2 (0,5569), MAE (5.332,1302), MSE (39.379.482,6203).

2.5.2. Decision Tree (DT) Regression:

Hình 4 cho thấy kết quả dự đoán và thực tế của thuật toán DTR hoàn toàn khớp với nhau trên tập dữ liệu sản phẩm Sakura. Bảng 4 là kết quả độ đo hiệu suất của mô hình MLR cho sản phẩm Sakura trên tập huấn luyện và tập kiểm tra: R^2 (100%), MAE (0), MSE (0). Với kết quả như trên chứng tỏ mô hình đã overfitting (quá khớp với dữ liệu huấn luyện).



Hình 4. Biểu đồ kết quả dự đoán (đường màu đỏ) và thực tế (đường màu xanh) sau khi áp dụng mô hình DTR cho sản phẩm Sakura

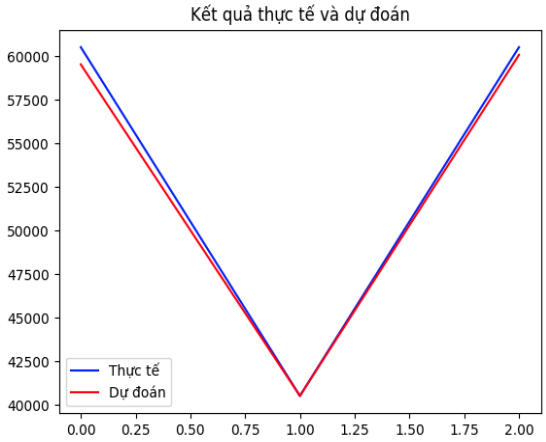
Ghi chú: Trục ox là chỉ số các điểm dữ liệu trong tập dữ liệu kiểm tra, trục oy là giá trị dữ liệu tìm kiếm Google

Bảng 4. Kết quả độ đo thuật toán DTR trên tập dữ liệu sản phẩm Sakura

	R^2 (train)	R^2 (test)	MAE (train)	MAE (test)	MSE (train)	MSE (test)
Độ đo	1,0	1,0	0,0	0,0	0,0	0,0

2.5.3. Random Forest (RF) Regression

Hình 5 cho thấy kết quả dự đoán và thực tế của thuật toán RFR gần tương đồng với nhau trên tập dữ liệu sản phẩm Sakura. Bảng 5 là kết quả độ đo hiệu suất của mô hình RFR cho sản phẩm Sakura trên tập huấn luyện là: R^2 (0,9276), MAE (2.002,7777), MSE (9.131.991,6666); trong khi tập kiểm tra có độ đo là: R^2 (0,9955), MAE (476,6666), MSE (391.233,3333). Kết quả độ đo Bảng 6 RFR cho thấy độ đo R^2, MAE và MSE hoạt động tốt trên tập dữ liệu của sản phẩm Sakura.



Hình 5. Biểu đồ kết quả dự đoán (đường màu đỏ) và thực tế (đường màu xanh) sau khi áp dụng mô hình RFR cho sản phẩm Sakura

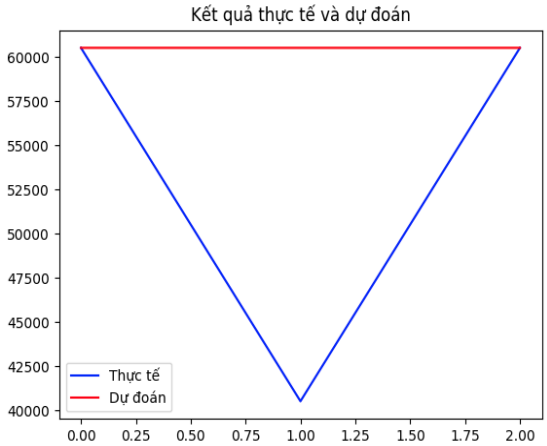
Ghi chú: Trục ox là chỉ số các điểm dữ liệu trong tập dữ liệu kiểm tra, trục oy là giá trị dữ liệu tìm kiếm Google

Bảng 5. Kết quả độ đo thuật toán RFR trên tập dữ liệu sản phẩm Sakura

	R^2 (train)	R^2 (test)	MAE (train)	MAE (test)	MSE (train)	MSE (test)
Độ đo	0,9276	0,9955	2.002,7777	476,6666	9.131.991,6666	391.233,3333

2.5.4. Support Vector (SV) Regression

Hình 6 cho thấy kết quả dự đoán và thực tế của thuật toán SVR hoàn toàn không khớp với nhau trên tập dữ liệu sản phẩm Sakura. Bảng 6 là kết quả độ đo thuật toán SVR cho sản phẩm Sakura trên tập huấn luyện là: R^2 (0,032), MAE (9.388,5427), MSE (167.012.023,6162); trong khi tập kiểm tra của chúng tôi có độ đo là: R^2 (-0,4998), MAE (6.666,4744), MSE (133.317.099,0687). Kết quả cho thấy hiệu suất mô hình RFR có độ đo R^2, MAE và MSE hoạt động không tốt trên tập dữ liệu sản phẩm Sakura.



Hình 6. Biểu đồ kết quả dự đoán (đường màu đỏ) và thực tế (đường màu xanh) sau khi áp dụng mô hình SVR của sản phẩm Sakura

Ghi chú: Trục ox là chỉ số các điểm dữ liệu trong tập dữ liệu kiểm tra, trục oy là giá trị dữ liệu tìm kiếm Google

Bảng 6. Kết quả độ đo thuật toán SVR trên tập dữ liệu sản phẩm Sakura

	R^2 (train)	R^2 (test)	MAE (train)	MAE (test)	MSE (train)	MSE (test)
Độ đo	0,032	0,4998	9.388,5427	6.666,4744	167.012.023,6162	133.317.099,0687

2.6. So sánh kết quả

Bảng 7 là kết quả độ đo của năm mô hình: SLR, MLR, DTR, RFR, SVR trên tập dữ liệu sakura. Mô hình RFR là tốt nhất trong bốn mô hình còn lại với các độ đo trên tập huấn luyện: R^2 (92%), MAE (2.003), MSE (9.131.991); tập kiểm tra có độ đo là: R^2 (99%), MAE (477), MSE (391.233).

Bảng 7. Kết quả độ đo của các thuật toán trên tập dữ liệu sản phẩm Sakura

Mô hình	R^2 (train)	R^2 (test)	MAE (train)	MAE (test)	MSE (train)	MSE (test)
SLR	0,770	0,5878	4.234,5679	5.416,6666	28.929.783,9506	36.635.159,4650
MLR	0,7770	0,5569	4.042,4241	5.332,1302	28.140.718,3639	39.379.482,6203
DTR	1,0	1,0	0,0	0,0	0,0	0,0
RFR	0,9276	0,9955	2.002,7777	476,6666	9.131.991,6666	391.233,3333
SVR	0,032	0,4998	9.388,542	6.666,4744	167.012.023,6162	133.317.099,0687

2.7. Kết quả dự đoán số liệu tìm kiếm trong tương lai

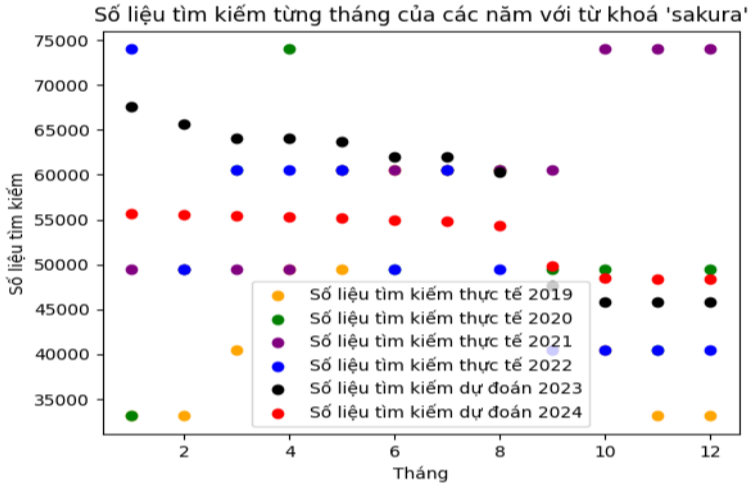
Thuật toán RFR được lựa chọn để dự đoán số liệu tìm kiếm trên Google từ tháng 01 đến 12 của năm 2023 và năm 2024.

Bảng 8 là kết quả dự đoán số liệu tìm kiếm từ tháng 01 đến tháng 12 trong năm 2023 và 2024 của sản phẩm Sakura.

Biểu đồ Hình 7 cho thấy rằng số liệu tìm kiếm thực tế cho từ khóa sakura có số liệu tìm kiếm cao và tăng giảm nhẹ giữa các tháng trong giai đoạn từ năm 2019 – 2022, ngoài ra số liệu dự đoán tìm kiếm trong năm 2023 có xu hướng tăng trưởng so với năm 2022; những tháng đầu năm 2024 có xu hướng giảm so với năm 2023 tuy nhiên những tháng cuối năm 2024 thì có xu hướng tăng.

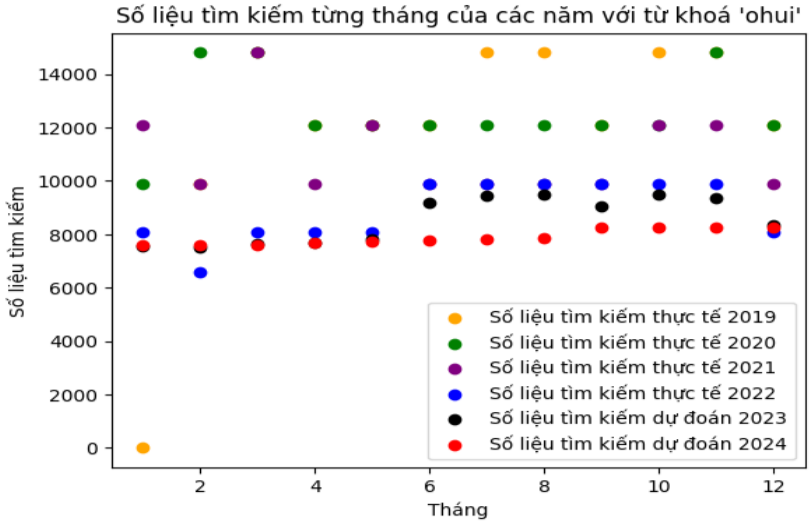
Bảng 8. Kết quả dự đoán số liệu tìm kiếm 12 tháng trong năm 2023 và 2024 của sản phẩm Sakura

Tháng	Dự đoán số liệu tìm kiếm 2023	Dự đoán số liệu tìm kiếm 2024
1	67.655	55.717
2	65.630	55.535
3	64.110	55.393
4	64.110	55.287
5	63.670	55.236
6	61.940	54.903
7	62.050	54.851
8	60.330	54.393
9	47.680	49.830
10	45.850	48.510
11	45.850	48.382
12	45.850	48.382



Hình 7. Biểu đồ số liệu tìm kiếm thực tế của sản phẩm Sakura

Tương tự, thuật toán RFR được áp dụng để dự đoán các sản phẩm còn lại (Hình 1), chẳng hạn như sản phẩm Ohui (Hình 8), sản phẩm Laneige (Hình 9), sản phẩm Ol cell (Hình 10) giúp cho doanh nghiệp có cái nhìn tổng quan từ đó so sánh đối chiếu lựa chọn sản phẩm phù hợp với chiến lược kinh doanh của họ. Đối với nhà đầu tư, việc này cũng giúp họ hạn chế rủi ro, cân nhắc xem sản phẩm là tiềm năng có đáng đầu tư hay không.

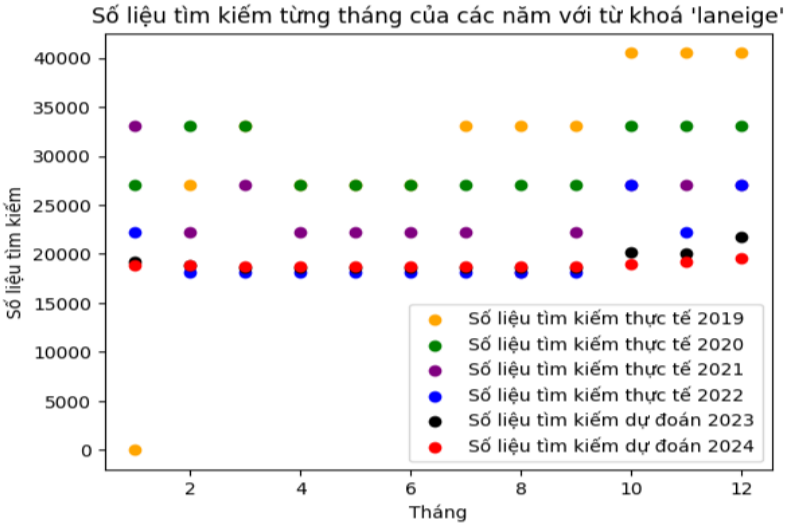


Hình 8. Biểu đồ số liệu tìm kiếm thực tế của sản phẩm Ohui

Ghi chú: màu cam năm 2019, màu xanh lá năm 2020, màu tím năm 2021, màu xanh dương năm 2022) và dự đoán (màu đen năm 2023, màu đỏ năm 2024)

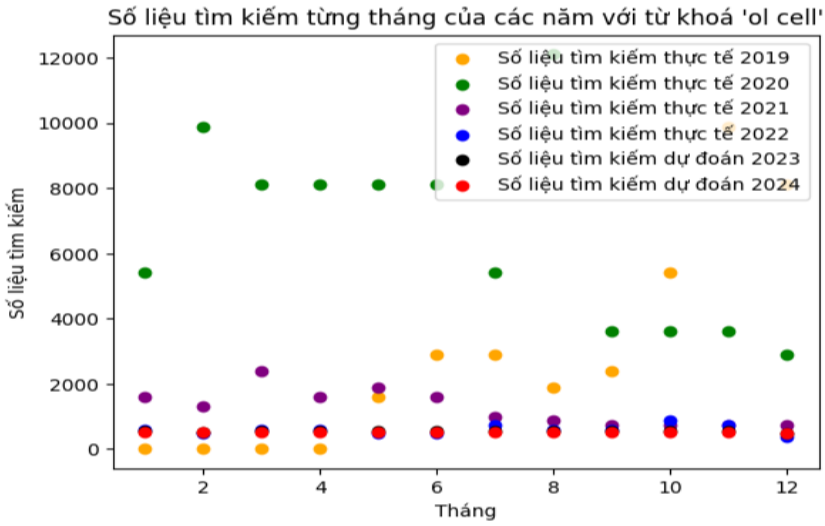
Biểu đồ Hình 8 cho thấy rằng số liệu tìm kiếm thực tế Google cho từ khóa ohui có xu hướng giảm theo từng năm từ 2019 – 2022, ngoài ra số liệu dự đoán tìm kiếm Google trong năm 2023 có xu hướng gần bằng so với năm 2022; còn trong năm 2024 thì số liệu dự đoán tìm kiếm giảm so sới năm 2023.

Biểu đồ Hình 9 cho thấy số liệu tìm kiếm thực tế Google cho từ khóa laneige có xu hướng giảm theo từng năm từ 2019 – 2022, tuy nhiên số liệu tìm kiếm dự đoán Google năm 2023 – 2024 có xu hướng tăng lên.



Hình 9. Biểu đồ số liệu tìm kiếm thực tế của sản phẩm Laneige

Ghi chú: màu cam năm 2019, màu xanh lá năm 2020, màu tím năm 2021, màu xanh dương năm 2022) và dự đoán (màu đen năm 2023, màu đỏ năm 2024)



Hình 10. Biểu đồ số liệu tìm kiếm thực tế của sản phẩm Ol cell

Ghi chú: màu cam năm 2019, màu xanh lá năm 2020, màu tím năm 2021, màu xanh dương năm 2022 và dự đoán (màu đen năm 2023, màu đỏ năm 2024)

Biểu đồ Hình 10 cho thấy rằng số liệu tìm kiếm thực tế Google cho từ khóa Ol cell có xu hướng giảm sâu theo từng năm từ 2019 – 2022, đồng thời số liệu tìm kiếm dự đoán Google năm 2023 – 2024 cho thấy xu hướng giảm.

Qua biểu đồ Hình 7, Hình 8, Hình 9 và Hình 10 với góc nhìn nhà đầu tư dựa vào số liệu tìm kiếm thực tế và dự đoán trên Google, đồng thời dựa theo xu hướng tăng trưởng thì chắc chắn doanh nghiệp sẽ ưu tiên lựa chọn sản phẩm Sakura để đưa vào kế hoạch kinh doanh của mình.

Đối với tiếng Việt, việc xử lý ngôn ngữ tự nhiên đặt ra nhiều thách thức do sự phức tạp từ vựng, ngữ pháp và cấu trúc (Hung, 2018; Hiếu, 2022). Ngoài ra, có những vấn đề khác như lỗi đánh máy, sai chính tả, các từ viết tắt hoặc ngôn ngữ tiếng lóng. Việc xử lý dữ liệu không tốt sẽ làm cho dữ liệu chứa nhiễu và sai sót, dẫn đến việc truy xuất và phân tích thông tin thiếu chính xác hoặc không chính xác.

Bên cạnh đó, các phương pháp và thuật toán khác nhau có thể gây ra sự đa dạng trong kết quả dự đoán, điều này tạo ra một thách thức khi đánh giá độ chính xác của kết quả. Ngoài ra, dữ liệu đầu vào do Google cung cấp còn nhiều hạn chế như là: số liệu tìm kiếm trong quá khứ bị giới hạn, vị trí khu vực (hiện tại chỉ cung cấp số liệu theo quốc gia, không cung cấp theo vị trí khu vực như là Hồ Chí Minh, Hà Nội), giới tính người dùng truy cập (hiện tại chỉ cung cấp tổng thể, không cung cấp theo giới tính như là nam hay nữ). Mặc khác có nhiều yếu tố tác động đến xu hướng tìm kiếm từ khóa trên Google như sau:

- Sản phẩm mới ra mắt: có thể gây ra sự hiếu kỳ và dẫn đến tăng lượng tìm kiếm đột biến.
- Các chiến dịch quảng cáo trên Google và các nền tảng khác: có thể tạo ra ảnh hưởng đến tìm kiếm từ khóa liên quan.
- Sự thay đổi thuật toán của Google: có thể làm thay đổi cách xếp hạng các từ khóa.
- Các chuyên gia, người ảnh hưởng: như các bác sỹ, dược sĩ, người nổi tiếng, … có thể tạo ra sự ảnh hưởng tìm kiếm từ khóa trên Google.
- Các chiến dịch quảng cáo trực tuyến, trực tiếp: cũng có thể tạo ra ảnh hưởng đến xu hướng tìm kiếm.
- Nhu cầu tìm kiếm của người dùng: các vấn đề họ quan tâm đến cũng làm ảnh hưởng đến tìm kiếm từ khóa.

3. KẾT LUẬN

Bài báo đã trình bày một phương pháp hiệu quả để trích xuất thông tin từ kết quả tìm kiếm Google và dự đoán xu hướng tìm kiếm trong tương lai bằng học máy. Kết quả thực nghiệm cho thấy phương pháp này có thể xác định được các sản phẩm chăm sóc sắc đẹp phổ biến và dự đoán xu hướng tìm kiếm của chúng trong tương lai trên Google.

Đặc biệt, mô hình Random Forest Regression đã cho kết quả dự đoán tốt nhất với độ chính xác cao. Điều này rất hữu ích cho các nhà đầu tư và doanh

nghiệp trong việc đánh giá tiềm năng của sản phẩm, lập kế hoạch kinh doanh và marketing hiệu quả.

Tóm lại, nghiên cứu này cung cấp một hướng tiếp cận thực tế và hiệu quả để khai thác thông tin từ dữ liệu web và dự đoán xu hướng tìm kiếm Google. Phương pháp đề xuất có thể mở rộng ứng dụng cho nhiều lĩnh vực kinh doanh khác.

TÀI LIỆU THAM KHẢO

Bisong, E. (2019). Matplotlib and Seaborn. *Building Machine Learning and Deep Learning Models on Google Cloud Platform* (pp.151-165); Apress, Berkeley, CA.. https://doi.org/10.1007/978-1-4842-4470-8_12

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1)*,* 5–32. https://doi.org/10.1023/a:1010933404324

Chicco, D.,Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science 7*(e623). https://doi.org/10.7717/peerj-cs.623

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random Forests. In C. Zhang & Y. Ma (Eds.), *Ensemble Machine Learning: Methods and Applications* (pp.157-176). Spinger. http://dx.doi.org/10.1007/978-1-4419-9326-7_5

Dogru, N., & Subasi, A. (2018). Traffic accident detection using random forest classifier. *2018 15th Learning and Technology Conference (pp.* 40–45). https://doi.org/10.1109/LT.2018.8368509

Duyet, L. V. (2017). *Vietnamese stopwords.* https://github.com/stopwords/vietnamese-stopwords/blob/master/vietnamese-stopwords.txt

Gatera, A., Kuradusenge, M., Bajpai, G., Mikeka, C., & Shrivastava, S. (2023). Comparision of random forest and support vector machine refression models for forecasting road accidents. *Scientific African, 21*(e01739)*.* https://doi.org/10.1016/j.sciaf.2023.e01739

Hao, T. H., Nghia, D. T., Dinh, Q. T., & Hiep, X. H. (2020). Vietnamese Text Classification with TextRank and Jaccard Similarity Coefficient. *Advances in Science Techonology and Engineering Systems Journal, 5*(6), 363-369. https://dx.doi.org/10.25046/aj050644

Hiếu, N. C. (2022). Khảo sát các mô hình phân loại văn bản tiếng Việt. *Tạp chí Khoa học và Công nghệ, 57*(3), 99-109. https://doi.org/10.46242/jstiuh.v57i03.4395

Hung, B. (2018). Vietnamese Diacritics Restoration Using Deep Learning Approach. *2018 10th International Conference on Knowledge and Systems Engineering* (pp. 347-351). IEEE. https://doi.org/10.1109/KSE.2018.8573427

Huyền, T. T. T. (2019). Đáp ứng nhu cầu mới về trải nghiệm mua sắm của người tiêu dùng kết nối trong thời đại 4.0. *Tạp chí Công Thương*, 18, 230-235. https://sti.vista.gov.vn/tw/Lists/TaiLieuKHCN/Attachments/278758/CVv146S182019230.pdf

Khang, P. Q. N., & Kiet, V. N. (2020). *Exploiting Vietnamese Social Media Characteristics for Textual Emotion Recognition in Vietnamese.* https://arxiv.org/pdf/2009.11005.pdf

Saleh, H., & Layous, J. A. (2022). *Machine Learning Regression*. Syrian Arab Republic Higher Institute for Applied Sciences and Technology. https://doi.org/10.13140/RG.2.2.35768.67842

Singh, S. (2018). *Natural Language Processing for Information Extraction*. https://arxiv.org/abs/1807.02383

Sishi, M., & Telukdarie, A. (2021). The Application of Decision Tree Regression to Optimize Business Processes. Proceedings of the *International Conference on Industrial Engineering and Operations Management.* https://www.ieomsociety.org/brazil2020/papers/31.pdf

Thanh, H. D., & Anh, T. T. N. (2021). A review: preprocessing techniques and data augmentation for sentiment analysis. *Computational Social Networks*, *8*(1). https://doi.org/10.1186/s40649-020-00080-x

Yogish, D. Y., Manjunath, T. N., & Hegadi, R. S. (2019). Review on Natural Language Processing Trends and Techniques Using NLTK. In K. C. Santosh & R. S. Hegadi (Eds.). *Recent Trends in Image Processing and Pattern Recognition* (pp.589-606). Springer. https://doi.org/10.1007/978-981-13-9187-3_53

Zheng, C., Hel, G., & Peng, Z. (2015). A Study of Web Information Extraction Technology Based on Beautiful Soup. *Journal of Computers, 10*(6), 381-387. https://doi.org/10.17706/jcp.10.6.381-387