Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á
Năm thứ 34, Số 5 (2023), 27–45

www.jabes.ueh.edu.vn

JABES
Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á

Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á
http://www.emeraldgrouppublishing.com/services/publishing/jabes/

Phát hiện chủ đề quan tâm của người dùng trên các phương tiện truyền thông xã hội theo độ tương quan

NGUYỄN THỊ HỘI *
Trường Đại học Thương mại, Hà Nội

THÔNG TIN

Ngày nhận: 27/11/2022
Ngày nhận lại: 21/03/2023
Duyệt đăng: 21/04/2023

Mã phân loại JEL:
C61; C63; C67.

Từ khóa:
Quan tâm;
Phương tiện truyền thông xã hội;
Wikipedia;
TF.IDF;
Độ tương quan.

Keywords:
Interest;
Social media;
Wikipedia;
TF.IDF;
Correlation.

TÓM TẮT

Phát hiện các chủ đề quan tâm của người dùng trên các phương tiện truyền thông xã hội là bài toán rất được chú ý trong thời gian gần đây, bởi khả năng ứng dụng cao trong thực tế. Bài báo giới thiệu một cách thức phát hiện chủ đề quan tâm của người dùng trên các phương tiện truyền thông xã hội dựa trên phân tích nội dung các bài đăng của người dùng, kết hợp mở rộng ngữ nghĩa theo Wikipedia, sử dụng véctơ theo TF.IDF để biểu diễn dữ liệu và ước lượng theo độ tương quan Pearson. Kết quả thực nghiệm cho thấy, cách phát hiện chủ đề này có thể áp dụng trên nhiều phương tiện truyền thông xã hội khác nhau mà không phụ thuộc vào cấu trúc, ngôn ngữ và liên kết trên các phương tiện truyền thông xã hội đó.

Abstract

Discovering user's interests on social media was an issue that has received a lot of attention recently because it has high applicability in practice. The purpose of this paper is to introduce a method to detect user interest topics on social media by analyzing the content of user posts. In this paper, a semantic expansion technique based on Wikipedia and TF.IDF weighted vector representation, then estimated based on Pearson correlation. The result of the experiment shows that the method of topic detection can be applied to many different social media sites regardless of their structure, language, and links.

* Tác giả liên hệ.
Email: hoint@tmu.edu.vn (Nguyễn Thị Hội).
Trích dẫn bài viết: Nguyễn Thị Hội. (2023). Phát hiện chủ đề quan tâm của người dùng trên các phương tiện truyền thông xã hội theo độ tương quan. *Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á, 34*(5), 27–45.

1. Giới thiệu

Các phương tiện truyền thông xã hội (Social Media) ngày càng phổ biến với đa dạng dịch vụ và được nhiều tổ chức, cá nhân lựa chọn làm kênh truyền thông chính trong các chiến lược quảng bá, marketing và bán hàng. Ảnh hưởng của các phương tiện truyền thông xã hội không chỉ trong lĩnh vực giải trí mà còn đóng vai trò quan trọng trong nhiều lĩnh vực, đặc biệt là giáo dục, chính trị, kinh doanh, cũng như các nghiên cứu về xã hội học như phát hiện lừa đảo, phát hiện tâm lý tội phạm… (Li & Sakamoto, 2014; Li và cộng sự, 2015; Samuel & Shamili, 2017; Kowsari và cộng sự, 2019). Việc xác định các chủ đề quan tâm của người dùng trên các phương tiện truyền thông xã hội ngày càng giúp các tổ chức, doanh nghiệp và người bán hàng rút ngắn thời gian phân nhóm khách hàng, xác định tốt hơn nhóm khách hàng mục tiêu, thu được nhiều kết quả khả quan hơn trong các chiến dịch marketing hoặc giới thiệu sản phẩm mới, đặc biệt trong quảng cáo cá nhân hóa người dùng (AlBanna và cộng sự, 2016; Shahzad và cộng sự, 2017; Xu & Lu, 2015; Tang và cộng sự, 2014).

Bài toán phát hiện các chủ đề quan tâm của người dùng trên phương tiện truyền thông xã hội có thể phát biểu như sau: "Cho một tập các chủ đề trên một phương tiện truyền thông xã hội và một tập hợp người dùng trên phương tiện truyền thông xã hội đó, cần đưa ra danh sách các chủ đề mà những người dùng đó quan tâm". Xét một cách tổng quát, bài toán phát hiện các chủ đề quan tâm của người dùng trên phương tiện truyền thông xã hội chính là bài toán gán nhiều nhãn cho người dùng, các nhãn chính là các chủ đề mà người dùng quan tâm.

Bài báo giới thiệu một cách thức phát hiện chủ đề quan tâm của người dùng trên các phương tiện truyền thông xã hội dựa trên phân tích nội dung các bài đăng của người dùng, kết hợp mở rộng ngữ nghĩa theo Wikipedia, dùng véctơ có trọng số theo TF.IDF[1] để biểu diễn dữ liệu và ước lượng theo độ tương quan Pearson. Kết quả có thể ứng dụng trong phân nhóm người dùng theo các chủ đề, hoặc ứng dụng chúng vào các hoạt động marketing cá nhân hóa người dùng trên các phương tiện truyền thông khác nhau. Bài báo gồm các phần: Cơ sở lý thuyết, phương pháp nghiên cứu, mô hình đề xuất, thực nghiệm và đánh giá, cuối cùng là phần kết luận.

2. Cơ sở lý thuyết

2.1. Một số khái niệm liên quan

Theo Boyd và Ellison (2007), Zafarani và cộng sự (2014), các phương tiện truyền thông xã hội là các dịch vụ dựa trên web cho phép các cá nhân có thể: Tạo lập một hồ sơ công khai hoặc bán công khai trong hệ thống có giới hạn, kết nối hoặc chia sẻ với một danh sách người dùng, và có thể cho phép xem, chia sẻ những nội dung thực hiện bởi những người dùng khác trong hệ thống. Theo Vispute và cộng sự (2014), Collin và cộng sự (2011), Samuel và Shamili (2017), phương tiện truyền thông xã hội là sự phản ánh mối quan hệ giữa các cá nhân của một xã hội trong thế giới thực vào hệ thống mạng máy tính và có thể được biểu diễn ở dạng đồ thị hoặc các mô tả, thể hiện sự liên kết và mối quan hệ của người dùng.

[1] Véc tơ trọng số TF.IDF (Viết tắt của thuật ngữ tiếng Anh: Term Frequency - Inverse Document Frequency) là trọng số của một từ/thuật ngữ trong văn bản của mẫu thử nghiệm và thể hiện mức độ quan trọng của từ/thuật ngữ này trong một văn bản.

Với sự phổ biến của mạng Internet, các phương tiện truyền thông xã hội đã và đang trở thành mảnh đất màu mỡ cho nhiều bài toán trong nhiều lĩnh vực khác nhau (Shahzad và cộng sự, 2017; Collin và cộng sự, 2011; Samuel & Shamili, 2017; Boyd & Ellison, 2007; Faizi và cộng sự, 2013), từ những bài toán ứng dụng phổ biến trong phân tích dữ liệu như: Khai phá dữ liệu, truy hồi thông tin, các hệ tư vấn… đến các lĩnh vực như: Xã hội học, các hệ thống quản lý khách hàng, các bài toán khai phá quan điểm, quản lý danh tiếng, phân tích hành vi con người…

Theo Xu và Lu (2015), Zafarari và cộng sự (2014), dữ liệu trên phương tiện truyền thông xã hội hay dữ liệu xã hội là dữ liệu nhận được từ các phương tiện truyền thông xã hội như: Các mạng xã hội, các website tìm kiếm, các trang thương mại điện tử, các trang chia sẻ hình ảnh, video… Đặc trưng cơ bản của dữ liệu trên các phương tiện truyền thông xã hội là có dung lượng lớn, có tính liên kết, chứa nhiều nhiễu, không có cấu trúc rõ ràng, và đặc biệt thường không đầy đủ, không hoàn chỉnh như các dữ liệu từ các nguồn khác.

Chủ đề (Topic) được người dùng quan tâm trên các phương tiện truyền thông xã hội (Krause và cộng sự, 2006; Lee và cộng sự, 2011; Hossen và cộng sự, 2018) thường được ẩn đằng sau những hoạt động mà người dùng thể hiện, chẳng hạn như: Có người dùng thích thể thao có thể mua các mặt hàng thể thao và xem các bài viết liên quan đến thể thao, có người dùng quan tâm đến giáo dục thì xem hoặc chia sẻ các bài viết về giáo dục, có người dùng quan tâm đến sức khỏe thì xem, đọc và viết các bài viết về sức khỏe… Như vậy, có thể phát hiện các chủ đề quan tâm của người dùng dựa trên các hoạt động của người dùng trên các phương tiện truyền thông xã hội.

2.2. Các nghiên cứu liên quan

Đã có nhiều nghiên cứu phát hiện quan tâm của người dùng trên các phương tiện truyền thông xã hội đặc biệt từ năm 2006 khi mạng xã hội Facebook ra đời, các nghiên cứu chủ yếu tập trung vào hai hướng: *Thứ nhất* là phát hiện quan tâm dựa trên các liên kết hoặc cấu trúc phương tiện truyền thông xã hội, còn gọi là hướng đặc trưng người dùng (User-Centric). *Thứ hai*, phát hiện quan tâm dựa trên dữ liệu, đối tượng do người dùng tạo ra trên các phương tiện truyền thông xã hội, còn gọi là hướng dữ liệu sinh ra bởi người dùng (Object-Centric).

Với hướng tiếp cận thứ nhất, nghiên cứu của Zafarani và Huan (2014) tập trung vào mối liên kết giữa các người dùng trên phương tiện truyền thông xã hội khác nhau để khám phá các hành vi chung và đưa ra chủ đề quan tâm của họ, sử dụng phương pháp thống kê và xây dựng dựa trên mức độ tương quan. Nori và cộng sự (2011) lại nghiên cứu và phân tích các hành động của người dùng trên phương tiện truyền thông xã hội "Delicious" và dự đoán các chủ đề quan tâm của người dùng trên mạng xã hội "Twitter". Với nghiên cứu dựa trên hành vi thích của Yin và cộng sự (2011), Lee (2016), mô hình hóa hành vi thích của người dùng rồi phân tích và dự đoán hành vi thích. Nghiên cứu cũng đưa ra cách thức thống kê lượt thích của người dùng trên các phương tiện truyền thông xã hội và mối liên quan giữa lượt thích và chủ đề quan tâm của người dùng. Với cách tiếp cận theo hướng theo đặc trưng của người dùng này, các mô hình phân tích đều phụ thuộc vào cấu trúc của phương tiện truyền thông xã hội và các kết nối của người dùng. Do đó, nếu các phương tiện truyền thông xã hội không thể hiện cấu trúc kết nối giữa các người dùng, hoặc các cấu trúc kết nối bị ẩn đi, nói cách khác, người dùng trên các phương tiện truyền thông xã hội này không thể hiện kết nối, không có mối liên kết với người dùng khác, thì rất khó để nghiên cứu theo hướng tiếp cận này.

- Với hướng tiếp cận thứ hai, tiếp cận dựa trên dữ liệu sinh ra bởi người dùng sẽ xem xét các dữ liệu hay đối tượng được sinh ra bởi người dùng trên các phương tiện truyền thông xã hội và được chia thành hai nhóm chính: Nhóm thứ nhất nghiên cứu dựa trên dữ liệu văn bản bao gồm: Nội dung bài đăng, thẻ đánh dấu, liên kết (Link) chia sẻ… Nhóm thứ hai nghiên cứu dựa trên các hành vi mà người dùng thực hiện khi tham gia trên các phương tiện truyền thông xã hội như: Các hành vi thích (Like), thể hiện cảm xúc (Emotion Icon), hành vi đăng bài (Post), hành vi gia nhập các cộng đồng (Join), hành vi chia sẻ (Share). Trong nghiên cứu thẻ đánh dấu của Bin và cộng sự (2016), Li & cộng sự (2008), đối với các bài viết không có thẻ đánh dấu sẽ bị loại bỏ, các nghiên cứu này chỉ phù hợp với các phương tiện truyền thông xã hội sử dụng thẻ đánh dấu, còn các phương tiện truyền thông xã hội khác lại không phù hợp… Nghiên cứu của Kim và cộng sự (2014) trích chọn quan tâm của người dùng dựa trên các bài viết và số lần thích của người dùng. Tuy nhiên, trong trường hợp số lượng các lần thích giống nhau trên tất cả các bài đăng hoặc nếu không có danh từ nào trong các bài đăng được xem xét thì mô hình tính toán này không đưa ra được kết quả. Nghiên cứu của Hossen và cộng sự (2018) xác định các chủ đề dựa trên các thẻ đánh dấu và nội dung của các bài viết trên Twitter, mỗi bài viết và thẻ đánh dấu có thể xác định được một chủ đề quan tâm của người dùng theo mô hình chủ đề. Nghiên cứu của Hossen và cộng sự (2018) đã sử dụng phương pháp thống kê theo thuật ngữ của bài đăng gốc, không xét đến yếu tố mở rộng hoặc các trường hợp một bài đăng có thể liên quan đến nhiều chủ đề khác nhau. Điều này có thể gây hạn chế khi ứng dụng trong quảng cáo như mục đích trong nghiên cứu của Hossen và cộng sự (2018) đã đề xuất.

Có thể thấy rằng, các mô hình nghiên cứu đã có thường tập trung nghiên cứu trên một phương tiện truyền thông xã hội cụ thể, hoặc tập trung trên một ngôn ngữ cụ thể mà chưa có mô hình có thể thực hiện trên nhiều phương tiện truyền thông xã hội khác nhau, hoặc trên nhiều ngôn ngữ khác nhau.

3. Phương pháp nghiên cứu

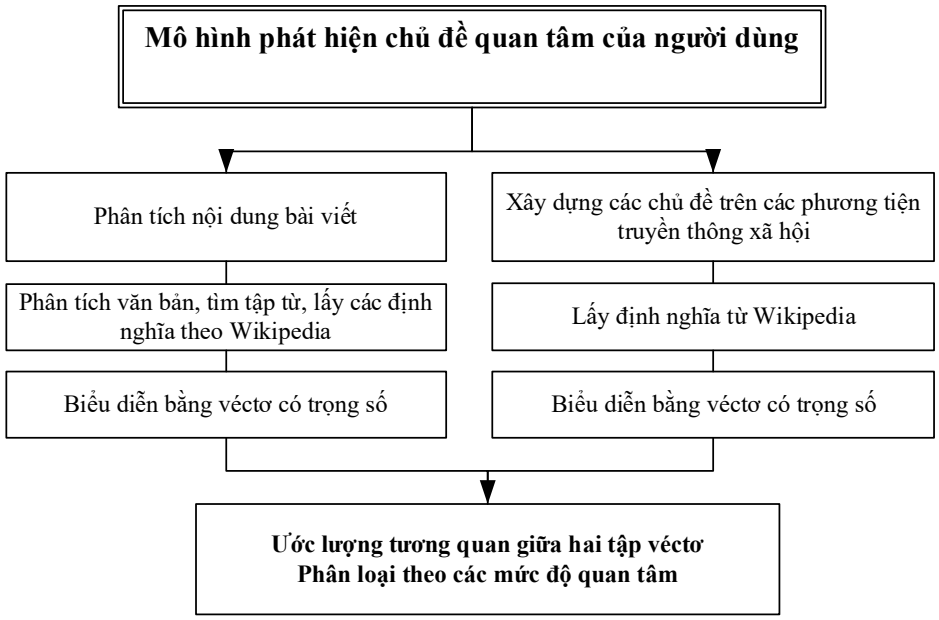
Nghiên cứu sử dụng phương pháp nghiên cứu định tính trong phân tích, so sánh, tổng hợp và đánh giá các kết quả của nghiên cứu đã có, từ đó đề xuất mô hình nghiên cứu và sử dụng phương pháp thực nghiệm trong kiểm chứng mô hình đề xuất. Nghiên cứu thực hiện đánh giá kết quả bằng độ chính xác (Precision) dựa trên ma trận nhầm lẫn (Aggarwal & Zhai, 2012; Manning và cộng sự, 2009) để tính độ chính xác của mô hình.

$$Prec = \frac{TP}{TP + FP}$$

Và sử dụng độ đo sai số bình phương trung bình (Mean Square Error) (Fouss & Saerens, 2008; Kowsari và cộng sự, 2019) để ước lượng các chủ đề quan tâm của người dùng.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2$$

Các bước thực hiện nghiên cứu cụ thể bao gồm: *Thứ nhất,* đề xuất mô hình nghiên cứu; *Thứ hai,* thực hiện thu thập và xây dựng bộ dữ liệu thực nghiệm; *Cuối cùng*, thực hiện thực nghiệm, đánh giá và so sánh với một số mô hình đã có (Hình 1).



Hình 1. Các bước thực hiện của nghiên cứu

4. Mô hình đề xuất và kết quả nghiên cứu

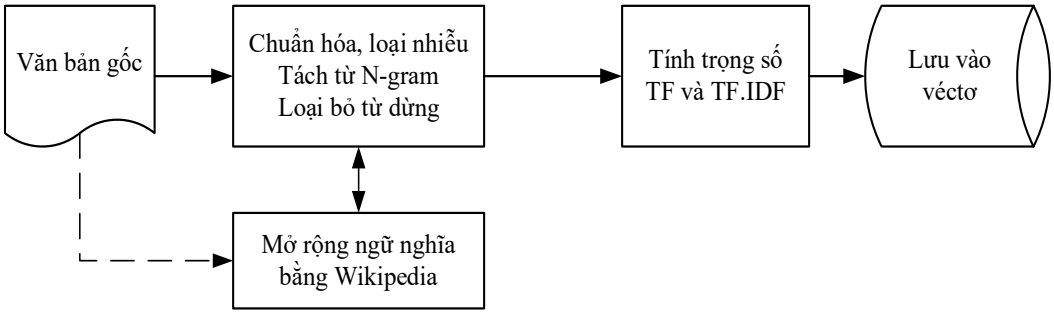
4.1. Các bước thực hiện nghiên cứu

Bài viết của người dùng trên các phương tiện truyền thông xã hội là các bài đăng mà người dùng tạo ra hoặc chia sẻ lại từ các nguồn khác trên mạng Internet, một bài viết trên một phương tiện truyền thông xã hội có thể là một video clip, một hoặc một số bức ảnh, một đoạn văn hoặc văn bản, hoặc một sự kết hợp của những thành phần này.

Như đã trình bày trong Mục 3 thì nội dung của nghiên cứu bao gồm các bước sau: Đề xuất mô hình nghiên cứu; thu thập và xây dựng bộ dữ liệu thực nghiệm; thực hiện thực nghiệm, đánh giá và so sánh với một số mô hình đã có.

4.2. Tiền xử lý dữ liệu

Các bài viết chứa văn bản của người dùng trên các phương tiện truyền thông xã hội được coi là dữ liệu xã hội, do đó, chúng thường không chuẩn về cú pháp, không chuẩn về văn phạm, thường kết hợp nhiều ngôn ngữ khác nhau trong một bài đăng… Vì vậy, nghiên cứu thực hiện tiền xử lý trước khi tính toán bằng cách: Làm sạch, tách từ và chuẩn hóa; Loại bỏ từ dừng, mở rộng tập từ theo Wikipedia; Thực hiện tách từ và tính trọng số TF.IDF rồi lưu vào các véctơ. Quy trình tiền xử lý dữ liệu của nghiên cứu được mô tả trong Hình 2.



Hình 2. Quy trình tiền xử lý văn bản trong nghiên cứu

Việc mở rộng tập từ của bài viết được nghiên cứu cải tiến từ thuật toán của Takale và Nandgaonkar (2010), trong nghiên cứu này, đề xuất mở rộng các từ đơn (Single Word) để đưa ra các từ tương tự, hoặc trong lĩnh vực nghiên cứu cải tiến thành mở rộng các từ theo các câu (Sentence) sau đó đưa vào tập từ có nghĩa. Ngoài ra, theo các nghiên cứu trước (Gattani và cộng sự, 2013; Kowsari và cộng sự, 2019; Lim & Datta, 2013) thì sử dụng từ điển Wikipedia làm cơ sở để phân tích và mở rộng từ dữ liệu văn bản xã hội cho kết quả tốt và phù hợp, bởi dữ liệu xã hội thường không theo quy chuẩn. Trong nghiên cứu sử dụng danh sách từ dừng do Wikipedia đề xuất và theo đề án nghiên cứu về tiếng Việt của (Hồ Tú Bảo & Lương Chi Mai, 2006) để loại bỏ từ dừng.

Cụ thể, việc tách từ, chuẩn hóa từ và loại bỏ từ dừng trong tiếng Việt của nghiên cứu được tiến hành trên thực nghiệm cho Bộ dữ liệu 2.000 bài viết với N = 1, N = 2, N = 3, và N = 4 tương ứng trong Bảng 1. Để tính tỷ lệ từ có nghĩa sau khi tách, nghiên cứu thực hiện việc tách từ của bài viết nguyên bản lưu vào tập tin Word_original, sau đó nghiên cứu thực hiện loại bỏ từ dừng, so sánh với từ điển thu được danh sách từ được sử dụng làm từ vựng để thực nghiệm (Word_key). Tỷ lệ thu được tính bằng phần trăm của Word_key/Word_original.

Bảng 1.

Tỷ lệ các từ có trong từ điển khi tách từ với N-gram

N-gram	Tỷ lệ Tiếng Việt 2.000 bài viết trên Facebook
N = 1	47,12%
N = 2	83,06%
N = 3	58,25%
N = 4	32,17%

Kết quả này có thể cho thấy đối với tiếng Việt, N = 2 là cho kết quả tốt nhất, N = 3 có kết quả tốt thứ 2. Vì vậy, trong thực nghiệm của nghiên cứu đã sử dụng N = 2 và N = 3.

Mở rộng ngữ nghĩa cho các bài viết bằng từ điển: Nghiên cứu sử dụng từ điển trực tuyến Wikipedia tiếng Việt[2] làm cơ sở để phân tích và mở rộng ngữ nghĩa cho bài viết của người dùng trên

[2] Truy cập từ https://vi.wikipedia.org/wiki/Wikipedia_tiếng_ Việt

mạng xã hội, ngoài ra, nghiên cứu có sử dụng thêm từ điển tiếng Việt trực truyến[3] để so sánh các kết quả. Quy trình thêm từ vựng bằng mở rộng ngữ nghĩa cho các bài viết được nghiên cứu thực hiện gồm: Tách N-gram, loại bỏ từ dừng, sau đó lưu vào véctơ từ gốc, mỗi từ xuất hiện trong danh sách từ gốc được lấy định nghĩa từ Wikipedia. Với mỗi định nghĩa thu được, nghiên cứu tiếp tục tách từ theo N-gram, loại bỏ từ dừng. Danh sách từ vựng thu được cuối cùng là danh sách từ đặc trưng của bài viết sau khi mở rộng.

4.3. Tính trọng số cho bài viết

Nghiên cứu xây dựng véctơ trọng số của tập từ thu được từ mỗi bài viết theo cách tính trọng số TF.IDF (Manning và cộng sự, 2009; Kowsari và cộng sự, 2019; Shahzad và cộng sự, 2017) và véctơ trọng số của bài viết được tính theo Nguyễn Thị Hội và Trần Đình Quế (2018). Bảng 2 minh họa tính trọng số của một bài viết theo TF.IDF.

Bảng 2.

Ví dụ về véctơ của một bài viết

Bài viết	Từ vựng và giá trị TF.IDF
Ý tưởng thu phí tác quyền âm nhạc đối với các tv trong ks của bác Phó Nhạc tuy ko mạch lạc nhưng nên áp dụng ngay với hệ thống loa phường. Và nên thu thật cao. Đặc biệt với bài Từ một ngã tư đường phố…	(âm nhạc: 0,3157); (áp dụng: 0,4024); (đối với: 0,3157); (đường phố: 0,4024); (hệ thống: 0,3157); (mạch lạc: 0,4024); (ngã tư: 0,4024)…

4.4. Biểu diễn bài viết dựa trên nội dung và các chủ đề quan tâm theo véctơ

Nghiên cứu thực hiện tính trọng số của các từ trong các không gian dữ liệu văn bản theo Định nghĩa dưới đây:

- ***Định nghĩa 1:*** Cho một tập văn bản $\boldsymbol{\mathcal{D}} = \{D_1, D_2, \dots, D_p\}$, mỗi văn bản được biểu diễn bằng một tập từ $D_i = \{d_{i1}, d_{i2}, .., d_{ip_i}\}$. Gọi $\boldsymbol{\mathcal{V}} = \{v_1, v_2, \dots, v_q\}$, là tập từ khác nhau từng đôi một. Khi đó, trọng số của từ $d \in \boldsymbol{\mathcal{V}}$ đối với D_i được tính như sau:

$$w_d = tf(d, D_i) \times idf(d, \boldsymbol{\mathcal{D}}) \quad (1)$$

Trong đó, $tf(d, D_i)$: Số lần xuất hiện của từ $d\ trong\ D_i$; và $idf(d, \boldsymbol{\mathcal{D}})$ được tính bằng:

$$idf(d, \boldsymbol{\mathcal{D}}) = log\left(\frac{\|\boldsymbol{\mathcal{D}}\|}{1+\|\{D_i | d \in D_i\}\|}\right) \quad (2)$$

Sau khi tính được trọng số của các từ thì mỗi văn bản $D_i \in \boldsymbol{\mathcal{D}}$ được biểu diễn bởi một véctơ trọng số, để tiện cho việc tính toán, mỗi véctơ được chuẩn hóa về khoảng đơn vị [0,1]. Khi đó, có thể định nghĩa văn bản $D_i \in \boldsymbol{\mathcal{D}}$ theo véctơ trọng số như sau:

- ***Định nghĩa 2:*** Cho một tập văn bản $\boldsymbol{\mathcal{D}} = \{D_1, D_2, \dots, D_p\}$, mỗi văn bản được biểu diễn bằng một tập từ $D_i = \{d_{i1}, d_{i2}, .., d_{ip_i}\}$. Gọi q là số lượng từ khác nhau từng đôi một trong không gian $\boldsymbol{\mathcal{D}}$. Khi đó, mỗi D_i được biểu diễn bởi một véctơ có q chiều:

$$\boldsymbol{w_i} = (w_{i1}, w_{i2}, .., w_{iq}) \text{ trong } \boldsymbol{\mathcal{D}}. \quad (3)$$

[3] Truy cập từ https://dict.laban.vn/

Trong đó, w_{ik} được tính theo Định nghĩa 1. Dựa trên Định nghĩa 1 và Định nghĩa 2 có thể định nghĩa biểu diễn bài viết của người dùng như sau:

- ***Định nghĩa 3:*** Cho một tập bài viết của các người dùng trên phương tiện truyền thông xã hội N là $\boldsymbol{\mathcal{E}} = \{E_1, E_2, \dots, E_q\}$, mỗi bài viết được biểu diễn bằng một tập từ $E_i = \{e_{i1}, e_{i2}, .., e_{iq_i}\}$. Gọi q là số lượng từ khác nhau từng đôi một trong không gian $\boldsymbol{\mathcal{E}}$. Khi đó, mỗi E_i được biểu diễn bởi một véctơ có q chiều: $\mathbf{w_i} = (w_{i1}, w_{i2}, .., w_{iq})$ trong không gian $\boldsymbol{\mathcal{E}}$. Trong đó, mỗi w_{ik} được tính như trong Định nghĩa 1.

Chủ đề trên các phương tiện truyền thông xã hội có thể hiểu là các vấn đề được ẩn trong các nội dung hoặc các hành vi của người dùng. Dựa trên các Định nghĩa 1, 2 và 3 có thể định nghĩa chủ đề theo hướng tiếp cận quan tâm của người dùng như sau:

- ***Định nghĩa 4:*** Cho một tập chủ đề $\boldsymbol{\mathcal{T}} = \{T_1, T_2, \dots, T_P\}$ trên phương tiện truyền thông xã hội N, khi đó, mỗi chủ đề T_i được biểu diễn bởi một tập từ: $T_i = \{t_{i1}, t_{i2}, \dots, t_{ip_i}\}$. Gọi $\boldsymbol{\mathcal{V}}_T$ là tập gồm q từ khác nhau từng đôi một trong tất cả các $T_i \in \boldsymbol{\mathcal{T}}$. Khi đó, mỗi T_i có một véctơ trọng số được ký hiệu như sau:

$$\boldsymbol{t_i} = (w_{i1}, w_{i2}, \dots, w_{iq}) \tag{4}$$

Trong đó, mỗi w_{ik} được tính như trong Định nghĩa 1.

4.5. Xây dựng và biểu diễn các chủ đề

Để xây dựng các chủ đề trong nghiên cứu, nghiên cứu thực hiện lựa chọn bằng cách thống kê các chủ đề trên một số trang tin tức điện tử phổ biến ở Việt Nam và trên thế giới, phương pháp này đã được các nghiên cứu của Bhattacharya và cộng sự (2014), Li và cộng sự (2008), Bin và cộng sự (2016) thực hiện. Các chủ đề phổ biến được thống kê từ 10 trang tin tức điện tử của Việt Nam có lượng người dùng truy cập lớn nhất (tháng 1 năm 2022)[4] cùng với 5 trang tin tức điện tử bằng tiếng Anh phổ biến trên thế giới (tháng 1 năm 2022)[5]. Danh sách các website dùng để xây dựng các chủ đề quan tâm được liệt kê trong Bảng 3.

Bảng 3.

Danh sách các trang tin tức điện tử tham khảo chủ đề

STT	Tên trang tin tức điện tử	Địa chỉ website
Các trang tin tức điện tử phổ biến ở Việt Nam		
1	Trang Vnexpress	https://www.vnexpress.net
2	Trang Việt Nam.net	https://www.vietnamnet.vn
3	Trang Tuổi trẻ	https://www.tuoitre.vn
4	Trang Thanh niên	https://www.thanhnien.vn
5	Trang Dân trí	https://www.dantri.com.vn
6	Trang Lao động	https://www.laodong.vn

[4] Theo thống kê từ https://toplist.vn/top-list/website

[5] Truy cập từ https://www.similarweb.com/top-websites/category/news-and-media

STT	Tên trang tin tức điện tử	Địa chỉ website
7	Trang Giải trí 24h	https://www.24h.com.vn
8	Trang Báo mới	https://www.baomoi.com
9	Trang Đời sống và Pháp luật	https://www.doisongphapluat.com
10	Trang Người lao động	https://www.nld.com.vn
Các trang tin tức điện tử trên thế giới		
11	Trang tin BBC	https://www.bbc.com/news
12	Trang tin Reuters	https://www.reuters.com
13	Trang tin CNN	http://edition.cnn.com
14	Trang tin Fox	https://www.foxnews.com
15	Trang tin The New York Times	https://www.nytimes.com

Sau khi thống kê các chủ đề và sắp xếp các chủ đề theo tần suất xuất hiện trên các trang tin tức, nghiên cứu thu được danh sách gồm 21 chủ đề trong Bảng 4, một số chủ đề tương tự hoặc gần nhau được nghiên cứu tích hợp lại thành một.

Bảng 4.

Danh sách các chủ đề trong thực nghiệm

STT	Tên chủ đề tiếng Việt	Tên chủ đề tiếng Anh
1	Tài chính - Kinh doanh	Business/ Finance
2	Thế giới - Quốc tế	World
3	Thời sự - Tin tức	News
4	Văn hóa - Giải trí	Entertainment
5	Khoa học - Công nghệ	Technology
6	Sức khỏe	Health
7	Thể thao	Sports
8	Đời sống - Xã hội	Life
9	Chính trị	Politics
10	Giáo dục	Education
11	Pháp luật - Nhà nước	Legal/ Law
12	Du lịch	Travel
13	Ý kiến	Opinions
14	Tâm sự	Talks
15	Con người	Humans

STT	Tên chủ đề tiếng Việt	Tên chủ đề tiếng Anh
16	Góc nhìn	Views
17	Làm đẹp - Thời trang	Fashion
18	Nhịp sống trẻ	LifeStyle
19	Môi trường	Environment
20	Khám phá	Discovery
21	Khác	Others

Mỗi chủ đề, nghiên cứu thực hiện lấy định nghĩa từ Wikipedia, sau đó xem mỗi định nghĩa là một văn bản thô và thực hiện tiền xử lý theo các bước đã minh họa trong Hình 2.

4.6. Ước lượng tương quan giữa bài viết và các chủ đề quan tâm

Dựa trên Định nghĩa 3, ký hiệu $\boldsymbol{\mathcal{U}} = \{u_1, u_2, \dots, u_n\}$ là tập hợp người dùng trên phương tiện truyền thông xã hội *N*, khi đó biểu diễn bài theo các chủ đề như sau:

- ***Định nghĩa 5:*** Giả sử $e_{ij} \in E_i$ là một bài viết của người dùng u_i trên phương tiện truyền thông xã hội N, được mô tả bởi một tập hợp các từ, khi đó, véctơ trọng số của bài viết e_{ij} đối với chủ đề T_k được định nghĩa như sau:

$$\mathbf{e_{ij}^k} = \left(e_{ij}^1, e_{ij}^2, \dots, e_{ij}^{t_{kp}}\right) \quad (5)$$

Trong đó, $e_{ij}^l = tf(t_{il}, e_{ij}) \times idf(t_{il}, E_i)$, với $t_{il} \in \mathcal{V}_T$.

Hệ số tương quan là một chỉ số thống kê đo lường mối liên hệ tương quan giữa hai biến số, hoặc hai đối tượng. Có nhiều hệ số tương quan như: Hệ số tương quan Pearson, hệ số Spearman, hệ số Kendall… (Kowsari và cộng sự, 2019; Manning và cộng sự, 2009) nhưng hệ số tương quan Pearson thông dụng nhất. Vì vậy, nghiên cứu sử dụng hệ số tương quan Pearson để ước lượng độ tương quan giữa bài viết với chủ đề và tính như sau:

$$cor(\boldsymbol{u}, \boldsymbol{v}) = \frac{\sum_i (u_i - \overline{u})(v_i - \overline{v})}{\sqrt{\sum_i (u_i - \overline{u})^2} \times \sqrt{\sum_i (v_i - \overline{v})^2}} \quad (6)$$

Tương quan giữa bài viết với các chủ đề được tính bằng mức độ tương quan của tập tất cả bài viết của người dùng đối với các chủ đề đang xét, và ký hiệu là mức độ liên quan giữa bài viết e_{ij} của người dùng u_i đối với chủ đề t_k là:

$$\alpha_{ij}^k = cor\left(\boldsymbol{e_{ij}}, \boldsymbol{t_k}\right) \quad (7)$$

Khi đó, mức độ liên quan của bài viết e_{ij} đến p chủ đề trong $\boldsymbol{\mathcal{T}}$ ký hiệu là:

$$cor\left(\boldsymbol{e_{ij}}, \mathcal{T}\right) = \left(\alpha_{ij}^1, \alpha_{ij}^2, \dots, \alpha_{ij}^p\right) \quad (8)$$

Có thể thấy rằng:

Khi số lượng các bài viết của một người dùng về cùng một chủ đề tăng lên thì mức độ quan tâm của người dùng đến chủ đề đó cũng tăng lên.

Khi số lượng các người dùng quan tâm đến một chủ đề tăng lên thì mức độ quan tâm của người dùng đến chủ đề đó cũng tăng lên.

4.7. Biểu diễn người dùng theo nội dung bài viết

Không mất tính tổng quát, giả sử rằng trên phương tiện truyền thông xã hội N có $\boldsymbol{\mathcal{E}} = \{E_1, E_2, \dots, E_m\}$ là tập bài viết tương ứng của *m* người dùng $\boldsymbol{\mathcal{U}} = \{u_1, u_2, \dots, u_m\}$. Ký hiệu mỗi bài viết là một tập từ $E_i = \{e_{i1}, e_{i2}, .., e_{im_i}\}$. Khi đó, mỗi người dùng trên phương tiện truyền thông xã hội *N* được biểu diễn bởi một véctơ gồm m_i thành phần, mỗi thành phần là một véctơ được xây dựng theo Định nghĩa 4. Ký hiệu như sau:

$u_i = \boldsymbol{u_i} = (\boldsymbol{w_{i1}}, \boldsymbol{w_{i2}}, .., \boldsymbol{w_{im_i}})$ $\forall \boldsymbol{w_{ik}} = \{w_{ik1}, w_{ik2}, .., w_{ikq}\} | \; k = 1, .., m_k$ trong không gian $\boldsymbol{\mathcal{E}}$ (9)

Cụ thể, mỗi người dùng trên phương tiện truyền thông xã hội có thể được biểu diễn như sau:

$$u_i = \begin{pmatrix} e_{i1} = \boldsymbol{w}_{i1} = (w_{i11}, w_{i12}, \dots, w_{i1q}), \\ e_{i2} = \boldsymbol{w}_{i2} = (w_{i21}, w_{i22}, \dots, w_{i2q}), \\ \dots . \\ e_{im_i} = \boldsymbol{w}_{im_i} = (w_{im1}, w_{im2}, \dots, w_{imq}) \end{pmatrix}$$

Với q là số chiều của không gian $\boldsymbol{\mathcal{E}}$ trên phương tiện truyền thông xã hội đang xem xét.

Độ quan tâm của người dùng với các chủ đề

Độ quan tâm của người dùng theo chủ đề được định nghĩa như sau:

- ***Định nghĩa 6:*** Cho một hàm số: int: $\mathcal{U} \times \mathcal{P}(\mathrm{E}) \times \mathcal{T} \to [0,1]$, gọi là quan tâm nếu nó thỏa mãn điều kiện sau: $\mathrm{int}(\mathrm{u}, \mathcal{U}, \mathrm{t}) \leq \mathrm{int}(\mathrm{v}, \mathcal{U}, \mathrm{t})$, đối với mọi $\mathrm{u}, \mathrm{v} \in \mathrm{P}(\mathrm{E_u})$ với $\mathrm{u} \leq \mathrm{v}$.

Để cho đơn giản khi tính toán và biểu diễn, trong nghiên cứu này, ký hiệu hàm quan tâm của người dùng $\mathrm{u_i}$ đến chủ đề t là $\mathrm{int}(\mathrm{u_i}, \mathrm{t})$.

- ***Bổ đề:*** Hàm số quan tâm của người dùng được tính bằng một trong ba cách sau đây:

Cách 1: $intMax(u_i, t) = \max_{\mathrm{j}} (cor(\boldsymbol{e_{ij}}, t))$ (10)

Cách 2: $intCor(u_i, t) = \frac{\sum_{\mathrm{j}} cor(\boldsymbol{e_{ij}}, t))}{\|\mathrm{E_i}\|}$ (11)

Cách 3: $intSum(u_i, t) = \frac{1}{2}\left(\frac{n_i^t}{\sum_{l \in \mathcal{T}} n_i^l} + \frac{n_i^t}{\sum_{u_k \in \mathcal{U}, l \in \mathcal{T}} n_k^l}\right)_j$ (12)

Trong đó, $cor(\boldsymbol{e_{ij}}, t)$ là mức độ liên quan của bài viết e_{ij} đến chủ đề *t,* n_i^t là số lượng bài viết liên quan đến chủ đề *t* của người dùng u_i trên phương tiện truyền thông xã hội *N.* Việc phát hiện các chủ đề quan tâm của người dùng dựa trên bài viết chính là ước lượng độ tương quan giữa các bài viết với các chủ đề trong $\boldsymbol{\mathcal{T}}$, nếu mức độ tương quan giữa bài viết và chủ đề tương ứng càng cao thì mức độ quan tâm của người dùng đối với chủ đề đó càng lớn, và ngược lại, mức độ tương quan càng gần đến giá trị 0 thì mức độ quan tâm của người dùng đến chủ đề đó càng thấp.

Ký hiệu $u_i^t = int(u_i, t)$ là độ quan tâm của người dùng u_i đến chủ đề t, khi đó, độ quan tâm của người dùng u_i đến tất cả các chủ đề trong $\boldsymbol{\mathcal{T}}$ được định nghĩa như sau:

- ***Định nghĩa 7:*** Độ quan tâm của người dùng đến tất cả các chủ đề là:

$$\mathbf{u_i^t} = (\mathbf{u_i^1}, \mathbf{u_i^2}, .., \mathbf{u_i^p}) \quad (13)$$

$\forall \mathbf{u_i^t}, t = 1, .., p$ là độ quan tâm của u_i đến chủ đề t.

Để thuận lợi cho việc phân nhóm người dùng dựa trên độ quan tâm các chủ đề trên các phương tiện truyền thông xã hội, nghiên cứu sử dụng thang đo Likert và chia độ quan tâm giữa người dùng với các chủ đề thành 5 mức tương ứng, các mức quan tâm giảm dần đến 0.

5. Thực nghiệm và đánh giá kết quả mô hình

5.1. Xây dựng bộ dữ liệu thực nghiệm

Nghiên cứu thực hiện thu thập bộ dữ liệu thực tế từ phương tiện truyền thông xã hội Facebook, sau khi loại bỏ những bài viết không chứa văn bản và loại nhiễu, nghiên cứu thu được một bộ gồm 200 người dùng, mỗi người dùng xét 10 bài viết hợp lệ, tổng cộng có 2.000 bài viết được dùng để phân tích. Thông số bộ dữ liệu thực nghiệm trong Bảng 5.

Bảng 5.

Thông số bộ dữ liệu thử nghiệm

Thuộc tính	Bộ dữ liệu tính độ tương quan
Người dùng	200
Bài viết	2.000
Chủ đề	21
Trọng số	TF.IDF
Biểu diễn	Véctơ trọng số theo không gian chủ đề

5.2. Kịch bản thực nghiệm

Xác định các chủ đề quan tâm của người dùng được thực hiện như sau:

- *Đầu vào:* Tập 200 người dùng, 2.000 bài viết và 21 chủ đề.

- *Đầu ra:* Tương quan giữa người dùng và các chủ đề.

- *Thực hiện:*

Bước 1: Xây dựng tập từ của mỗi bài viết và tập từ cho mỗi chủ đề.

Bước 2: Thực hiện tiền xử lý, loại nhiễu, tách từ, loại từ dừng.

Bước 3: Tính véctơ trọng số theo TF.IDF.

Bước 4: Tính độ tương quan của mỗi bài viết với các chủ đề theo công thức (9).

Bước 5: Phân loại mức độ quan tâm của mỗi người dùng theo 21 chủ đề dựa trên độ tương quan theo các công thức (10), (11) và (12).

- *Kết thúc.*

- *Tham số đầu ra:* Đầu ra là độ quan tâm của mỗi người dùng với 21 chủ đề trên phương tiện truyền thông xã hội. Phân loại các mức độ quan tâm theo 5 mức như sau:

Mức 0: Độ tương quan trong khoảng [0,000 – 0,025];

Mức 1: Độ tương quan trong khoảng [0,025 – 0,050];

Mức 2: Độ tương quan trong khoảng [0,050 – 0,075];

Mức 3: Độ tương quan trong khoảng [0,075 – 0,100];

Mức 4: Độ tương quan trong khoảng [0,100 – 1,000].

5.3. Kết quả các bước thực hiện thực nghiệm

Nghiên cứu thực hiện tiền xử lý dữ liệu, tính trọng số cho mỗi bài viết và các chủ đề. Sau đó tính độ tương quan giữa các bài viết với các chủ đề dựa theo công thức (9), tại bước này, với mỗi người dùng sẽ có một ma trận gồm 10 dòng (tương ứng 10 bài viết) và 21 cột (tương ứng 21 chủ đề), mỗi giá trị trong ma trận là độ tương quan của bài viết với chủ đề tương ứng (Bảng 6).

Bảng 6.

Độ tương quan của các bài viết với các chủ đề

Bài viết	Môi trường	Chính trị	Sức khỏe	Công nghệ	Thể thao	Văn hóa - Giải trí	Du lịch	Tài chính - Kinh doanh	…
e12_1	0,0272	0,0000	0,0147	0,0308	0,0465	0,0372	0,0000	0,0324	…
e12_2	0,0000	0,0000	0,0896	0,0538	0,0709	0,0072	0,0795	0,0666	…
e12_3	0,0000	0,0172	0,0000	0,0000	0,0138	0,0000	0,0000	0,0000	…
e12_4	0,0000	0,0000	0,0000	0,0000	0,0413	0,0000	0,0000	0,0000	…
e12_5	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	…
…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…
TB	0,0039	0,0025	0,0606	0,0121	0,0246	0,0093	0,0168	0,0141	…
Max	0,0272	0,0172	0,3201	0,0538	0,0709	0,0372	0,0795	0,0666	…

Sau đó, nghiên cứu thực hiện ước lượng độ tương quan của người dùng với các chủ đề theo công thức (10), (11) và (12). Với mỗi công thức sẽ thu được một ma trận gồm có 200 dòng (tương ứng 200 người dùng) và 21 cột (tương ứng 21 chủ đề), trong đó, mỗi ô là độ quan tâm của người dùng đó đến chủ đề tương ứng. Bảng 7 minh họa độ tương quan giữa người dùng với các chủ đề theo 03 công thức nghiên cứu đề xuất.

Bảng 7.

Độ tương quan của người dùng theo từng chủ đề

Chủ đề	Chính trị			Du lịch			…		
User	(12)	(11)	(10)	(12)	(11)	(10)	(12)	(11)	(10)
U001	0,0000	0,0000	0,0000	0,1378	0,0136	0,1398	…	…	…
U002	0,0000	0,0000	0,0000	0,0722	0,0035	0,0530	…	…	…

Chủ đề	Chính trị			Du lịch			…		
User	(12)	(11)	(10)	(12)	(11)	(10)	(12)	(11)	(10)
U003	0,1297	0,0118	0,0939	0,0305	0,0018	0,0379	…	…	…
U004	0,1622	0,0074	0,0527	0,0914	0,0030	0,0342	…	…	…
U005	0,0000	0,0000	0,0000	0,0451	0,0020	0,0414	…	…	…
…	…	…	…	…	…	…	…	…	…

Cuối cùng, nghiên cứu thực hiện chuyển đổi mức độ tương quan thành 05 mức như trong Bảng 8. Trong đó, mức 4 là mức cao nhất và mức 0 là mức người dùng hầu như không quan tâm đến chủ đề đó, hoặc nói cách khác, các bài viết của người dùng rất ít liên quan đến chủ đề đó.

Bảng 8.

Phân loại theo các mức quan tâm của người dùng với các chủ đề

	Chính trị			Du lịch			…		
User		(11)	(10)	(12)	(11)	(10)	(12)	(11)	(10)
U001	0	0	0	4	0	4	…	…	…
U002	0	0	0	2	0	2	…	…	…
U003	4	0	3	1	0	1	…	…	…
U004	4	0	2	3	0	1	…	…	…
U005	0	0	0	1	0	1	…	…	…
…	…	…	…	…	…	…	…	…	…

Từ kết quả phân loại trong Bảng 8, mức độ quan tâm của người dùng đến các chủ đề của thực nghiệm được minh họa trong Bảng 9.

Bảng 9.

Kết quả phân loại theo các mức quan tâm của người dùng theo từng chủ đề

STT	Chủ đề	Mức 0	Mức 1	Mức 2	Mức 3	Mức 4
1	Tài chính - Kinh doanh	18	42	31	33	76
2	Thế giới - Quốc tế	11	54	32	53	50
3	Thời sự - Tin tức	21	55	46	56	22
4	Văn hóa - Giải trí	30	17	26	26	101
5	Khoa học - Công nghệ	28	24	27	26	95
6	Sức khỏe	26	30	21	34	89
7	Thể thao	16	12	34	33	105
8	Đời sống - Xã hội	13	33	25	45	84

STT	Chủ đề	Mức 0	Mức 1	Mức 2	Mức 3	Mức 4
9	Chính trị	84	42	32	10	32
10	Giáo dục	36	49	41	28	46
11	Pháp luật - Nhà nước	55	25	37	33	50
12	Du lịch	50	50	26	28	46
13	Đánh giá sản phẩm	67	55	23	35	20
14	Gia đình	42	38	30	40	50
15	Con người	44	34	65	36	21
16	Góc nhìn	45	23	35	28	69
17	Làm đẹp - Thời trang	12	34	45	49	60
18	Nhịp sống trẻ	33	24	51	37	55
19	Môi trường	42	38	32	24	64
20	Khám phá	43	33	45	56	23
21	Khác	60	27	29	43	41
	Tổng	776	739	733	753	1.199

5.4. Thảo luận và đánh giá kết quả thực nghiệm

Để đánh giá độ chính xác của mô hình đề xuất, nghiên cứu nhờ các tình nguyện viên đánh giá bộ mẫu trước khi thực hiện thực nghiệm. Tiếp theo, nghiên cứu thực hiện so sánh kết quả thực nghiệm với kết quả của tình nguyện viên đánh giá, nếu kết quả của máy tính thực hiện trùng với kết quả của các tình nguyện viên thì độ chính xác sẽ được cộng thêm một đơn vị, ngược lại thì để nguyên. Cuối cùng, nghiên cứu sử dụng độ đo sai số bình phương trung bình (MSE) để ước lượng độ chính xác của mô hình.

Với công thức: $MSE = \frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2$ (14)

Trong đó, n: Tổng số cặp (người dùng, chủ đề); p_i: Giá trị đánh giá dự đoán của cặp (u_i, t_j); và r_i: Giá trị đánh giá thực tế từ tình nguyện viên của cặp (u_i, t_j). Với mỗi $u_i, i = 1..200$ và $t_j, j = 1..21$, như vậy, có tất cả (200 x 21) = 4.200 cặp, tức là n = 4.200. Theo công thức MSE thì giá trị của MSE càng gần đến giá trị 0 thì độ chính xác càng cao, và ngược lại. Kết quả từ thực nghiệm thu được:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2 = \frac{1}{4200}\sum_{i=1}^{4200}(p_i - r_i)^2 = 0{,}1226$$

Tương ứng với độ chính xác của thực nghiệm là:

$CR = (1 - MSE)\ x\ 100\% = (1 - 0{,}1226)\ x\ 100\% = 87{,}74\%$ (15)

5.5. So sánh với một số mô hình khác

Để so sánh kết quả của mô hình đề xuất với một số mô hình đã có, nghiên cứu thực hiện so sánh với mô hình phân tích dựa trên thẻ đánh dấu (Bin và cộng sự, 2016) và mô hình ước lượng dựa trên các Tweet (Hossen và cộng sự, 2018). Các mô hình này đều phân tích dữ liệu văn bản, biểu diễn dữ liệu bằng véctơ, tính trọng số bằng TF.IDF, tách từ bằng N-gram. Các mô hình đều sử dụng để xác định các chủ đề quan tâm của người dùng trên các phương tiện truyền thông xã hội.

Mô hình của Bin và cộng sự (2016) đã phân tích các thẻ đánh dấu trên mạng xã hội Douban.com (Một trang mạng xã hội cho phép người dùng bày tỏ quan điểm, ý kiến về đời sống, phim ảnh, âm nhạc, sự kiện…). Mỗi sản phẩm giải trí trên Douban thường là một bộ phim, một bài hát, một đĩa nhạc… được mô tả bằng một số từ khóa, các từ khóa này cho biết nội dung và thể loại của sản phẩm giải trí đó. Nghiên cứu đã biểu diễn mỗi sản phẩm giải trí bằng hai véctơ: Một véctơ chứa từ khóa của các thẻ đánh dấu, và một véctơ chứa từ khóa về sản phẩm giải trí. Trọng số của một từ *i* trong tài liệu *j* được tính theo TF dựa trên công thức:

$$a_{ij}^{tf} = \frac{f_{ij}}{\sqrt{\sum_{k=1}^{t} f_{ij}^2}} \quad (16)$$

Và TF.IDF của một từ I trong tài liệu j được tính bằng:

$$a_{ij}^{tf \times idf} = \frac{b_{ij}}{\sqrt{\sum_{k=1}^{t} b_{ij}^2}} \quad (17)$$

Với $b_{ij} = f_{ij} \log(\frac{d}{D_i})$; và D_i là số các tài liệu có chứa từ *i.*

Mô hình Hossen và cộng sự (2018) được thực hiện bằng cách xây dựng một mô hình gọi là Giá trị chủ đề giao của các thực thể (Entity Intersect Categorized Value – EICV), trong đó, W_n: Tập hợp các từ; R_n: Các từ dư thừa; K_n: Tập hợp các từ trích chọn từ tri thức của các thực thể; T_n: Tập hợp các chủ đề; $T_{tag[n]}$: Tập hợp các thẻ đánh dấu của mỗi chủ đề trong T_n.

Khi đó, nếu mỗi W_n có liên quan đến chủ đề t_k thì:

$K_n = F_{twitter}(W_n - R_n)$ (18)

Trong đó, $F_{twitter}$ là một hàm lấy các thực thể từ mạng Twitter, hay là các phương thức, giao thức kết nối với các thư viện và ứng dụng khác (Application Programming Interface – API) dùng để lấy dữ liệu từ mạng Twitter. Khi đó, giá trị giao được tính bằng:

$V_i = number\ of\ (K_n \cap T_{tag[n]})$ (19)

Với mỗi chủ đề $t_k \in T_n$ nó tạo ra một giá trị V_i. Hossen và cộng sự (2018) đã đặt một giá trị ngưỡng T_v và lấy các chủ đề lớn hơn ngưỡng này để xem xét các chủ đề quan tâm của người dùng.

Nghiên cứu thực hiện cài đặt nhằm so sánh giữa các mô hình trên Python Ver 3.8, có thể tóm tắt các kỹ thuật để so sánh mô hình đề xuất trong nghiên cứu và hai mô hình đã trình bày trong Bảng 10 gồm: Đối tượng, kỹ thuật, và độ đo.

Bảng 10.

Tóm tắt các mô hình so sánh

Tên mô hình	Mô hình Bin và cộng sự (2016)	Mô hình Hossen và cộng sự (2018)	Mô hình đề xuất trong nghiên cứu
Đối tượng	Các thẻ đánh dấu	Nội dung bài viết	Nội dung bài viết
Kỹ thuật tách từ	N-gram Loại từ dừng	Dựa trên dấu cách Loại từ dừng	N-gram Loại từ dừng
Trọng số	TF.IDF	Words	TF.IDF
Độ tương quan	Cosine	Giao của hai tập hợp	Pearson

Kết quả thực nghiệm thu được trong Bảng 11.

Bảng 11.

Độ chính xác của các mô hình

Tên mô hình	Mô hình Bin và cộng sự (2016)	Mô hình Hossen và cộng sự (2018)	Mô hình đề xuất
Độ chính xác	62,40%	85,26%	87,74%

Từ kết quả trên có thể thấy rằng, mô hình của Bin và cộng sự (2016) có độ chính xác thấp nhất là 62,40%, nghĩa là có 1.248/2.000 mẫu đúng so với đánh giá thực tế, điều này có thể giải thích do mô hình này dựa trên thẻ đánh dấu, mà Facebook có định dạng thẻ đánh dấu ít từ, vì vậy, khi tính TF.IDF với các chủ đề thì giá trị sẽ nhỏ. Mô hình của Hossen và cộng sự (2018) có độ chính xác cao hơn, đạt 85,26%, tương ứng có 1.705/2.000 mẫu đúng, do xét đến nội dung bài viết và tính theo từ, không mở rộng không gian các tập từ, mặt khác, trong mô hình này so sánh dựa trên hai tập hợp và lấy giao của tập từ trong bài viết và tập từ các của chủ đề, nên độ chính xác của mô hình cao hơn. Mô hình đề xuất của nghiên cứu cho kết quả tốt nhất với độ chính xác là 87,74%, tương ứng có 1.754/2.000 mẫu đúng, có thể hiểu được do nghiên cứu xét các bài viết đã đăng và tích hợp thêm việc mở rộng tập từ theo từ điển Wikipedia nên tăng độ chính xác của mô hình.

6. Kết luận

Việc phát hiện chủ đề quan tâm của người dùng giúp các tổ chức, doanh nghiệp tăng hiệu quả các chiến lược kinh doanh, thu được kết quả tốt hơn trong các chiến dịch quảng bá thương hiệu, giới thiệu sản phẩm. Với mô hình đề xuất trong nghiên cứu có thể ứng dụng trong phát hiện các chủ đề quan tâm của người dùng, phân nhóm người dùng theo chủ đề hoặc cá nhân hóa quảng cáo trên các phương tiện truyền thông xã hội mà không phụ thuộc vào cấu trúc và ngôn ngữ. Tuy nhiên, nghiên cứu chỉ mới phân tích dựa trên nội dung bài viết, việc mở rộng đối tượng nghiên cứu sang dữ liệu ảnh hoặc tích hợp thêm các đối tượng khác là hướng nghiên cứu tiếp theo của nhóm nghiên cứu.

Lời cảm ơn

Tác giả trân trọng cám ơn các bạn sinh viên Ngành HTTT Quản lý K52S, K53S Trường Đại học Thương mại đã hỗ trợ thu thập dữ liệu và đóng vai trò đánh giá thực tế cho bộ mẫu dữ liệu thực nghiệm.

Tài liệu tham khảo

Aggarwal, C. C., & Zhai, C. X. (2012). A survey of text clustering algorithms. In Aggarwal, C., Zhai, C. (eds), *Mining Text Data*. Boston, MA: Springer. doi: 10.1007/978-1-4614-3223-4_4

AlBanna, B., Sakr, M., Moussa, S., & Moawad, I. (2016). Interest aware location – Based recommender system using geo-tagged social media. *ISPRS International Journal of Geo-Information, 5*(12), 245–264.

Bhattacharya, P., Zafar, M. B., Ganguly, N., Ghosh, S., & Gummadi, K. P. (2014). *Inferring user interests in the Twitter social network.* Proceedings of the 8th ACM Conference on Recommender Systems (pp. 357–360), 6–10 October, 2014, Foster City, Silicon Valley, California, USA.

Bin, S., Sun, G., Zhang, P., & Zhou, Y. (2016). Tag-based interest-matching users discovery approach in online social network. *International Journal of Hybrid Information Technology, 9*(5), 61–70.

Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication, 13*(1), 210–230.

Collin, P., Rahilly, K., Richardson, I., & Third, A. (2011). *The benefits of social networking services: A literature review.* Melbourne: Cooperative Research Centre for Young People, Technology and Wellbeing.

Faizi, R., El Afia, A., & Chiheb, R. (2013). Exploring the potential benefits of using social media in education. *International Journal of Engineering Pedagogy (iJEP), 3*(4), 50–53. doi: 10.3991/ijep.v3i4.2836

Fouss, F., & Saerens, M. (2008). *Evaluating performance of recommender systems: An experimental comparison.* In 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (pp. 735–738), 09–12 December 2008, Sydney, NSW, Australia.

Gattani, A., Lamba, D. S., Garera, N., Tiwari, M., Chai, X., Das, S., & Rajaraman, A. (2013). Entity extraction, linking, classification, and tagging for social media: A Wikipedia-based approach. *Proceedings of the VLDB Endowment, 6*(1), 1126–1137.

Hossen, M. K., Faiad, A., Chowdhury, S. A., & Islam, S. (2018). Discovering user's topic of interest from Tweet. *International Journal of Computer Science and Information Technology, 10*(1), 95–105. doi: 10.5121/ijcsit.2018.10108

Hồ Tú Bảo, & Lương Chi Mai. (2006). *Về xử lý tiếng Việt trong công nghệ thông tin.* Truy cập từ http://www.jaist.ac.jp/~bao/Writings/VLSPwhitepaper%20-%20Final.pdf

Kim, J., Choi, D., Ko, B., Lee, E., & Kim, P. (2014). Extracting user interests on Facebook. *International Journal of Distributed Sensor Networks, 10*(6), 1–5.

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information, 10*(4), 150. doi: 10.3390/info10040150

Krause, A., Leskovec, J., & Guestrin, C. (2006). *Data association for topic intensity tracking.* Proceedings of the 23rd International Conference on Machine Learning (pp. 497–504). Pittsburgh, PA, USA: Carnegie Mellon University.

Lee, D. (2016). *Likeology: Modeling, predicting, and aggregating likes in social media.* Proceedings of the 8th ACM Conference on Web Science, Hannover, Germany.

Lee, K., Palsetia, D., Narayanan, R., Patwary, M. M. A., Agrawal, A., & Choudhary, A. (2011). *Twitter trending topic classification.* In 2011 IEEE 11th International Conference on Data Mining Workshops (pp. 251–258), 11–11 December 2011, Vancouver, BC, Canada. doi: 10.1109/ICDMW.2011.171

Li, H., & Sakamoto, Y. (2014). Social impacts in social media: An examination of perceived truthfulness and sharing of information. *Computer in Human Behavior, 41,* 278–287. doi: 10.1016/j.chb.2014.08.009

Li, L., Peng, W., Kataria, S., Sun, T., & Li, T. (2015). Recommending users and communities in social media. *ACM Transactions on Knowledge Discovery from Data, 10*(2), 1–27.

Li, X., Guo, L., & Zhao, Y. E. (2008). *Tag-based social interest discovery.* Proceedings of the 17th International Conference on World Wide Web (pp. 675–684), 21–25 April, 2008, Beijing, China.

Lim, K. H., & Datta, A. (2013). *Interest classification of Twitter users using Wikipedia.* Proceedings of the 9th International Symposium on Open Collaboration (pp. 1-2), Hong Kong, China.

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *Introduction to Information Retrieval.* New York, USA: Cambridge University Press.

Nori, N., Bollegala, D., & Ishizuka, M. (2011). *Exploiting user interest on social media for aggregating diverse data and predicting interest.* Proceedings of the 5th International Conference on Weblogs and Social Media, 17–21 July, 2011, Barcelona, Catalonia, Spain.

Nguyễn Thị Hội, & Trần Đình Quế. (2018). Ước lượng quan tâm người dùng trên mạng xã hội dựa trên tương tự bài viết. *Tạp chí Khoa học và Công nghệ - Đại học Đà Nẵng, 7*(128), 28–32.

Samuel, C. J., & Shamili, S. (2017). A study on impact of social media on education, business and society. *International Journal of Research in Management & Business Studies, 4*(3), 52–56.

Shahzad, B., Lali, I., Nawaz, M. S., Aslam, W., Mustafa, R., & Mashkoor, A. (2017). Discovery and classification of user interests on social media. *Information Discovery and Delivery, 45*(3), 130–138. doi: 10.1108/IDD-03-2017-0023

Si, H., Zhou, J., Chen, Z., Wan, J., Xiong, N. N., Zhang, W., & Vasilakos, A. V. (2019). Association rules mining among interests and applications for users on social networks. *IEEE Access, 7*, 116014–116026.

Takale, S. A., & Nandgaonkar, S. S. (2010). Measuring semantic similarity between words using web documents. *International Journal of Advanced Computer Science and Applications, 1*(4), 78–85.

Tang, J., Chang, Y., & Liu, H. (2014). Mining social media with social theories: A survey. *ACM SIGKDD Explorations Newsletter, 15*(2), 20–29. doi: 10.1145/2641190.2641195

Vispute, A., Jadhav, P., & Kharat, P. V. (2014). Collective behavior of social networking sites. *Journal of Computer Engineering (IOSR-JCE), 16*(2), 75–79.

Xu, J., & Lu, T.-C. (2015). *Toward precise user-topic alignment in online social media.* International Conference on Big Data (pp. 767–775), 29 October, 2015–01 November, 2015, Santa Clara, CA, USA.

Yin, D., Hong, L., & Davison, B. D. (2011). *Exploiting session-like behaviors in tag prediction.* Proceedings of the 20th International Conference on World Wide Web (pp. 167–168), March 28–April 1, 2011, India.

Zafarani, R., & Huan, L. (2014). Behavior analysis in social media. *IEEE Intelligent Systems, 29*(4), 69–71.

Zafarani, R., Abbasi, M. A., & Liu, H. (2014). *Social Media Mining: An Introduction.* New York, USA: Cambridge University Press.