HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN THỊ HỘI

MÔ HÌNH HÀNH VI VÀ QUAN TÂM CỦA NGƯỜI DÙNG TRÊN CÁC MẠNG XÃ HỘI

LUẬN ÁN TIẾN SĨ KỸ THUẬT

HÀ NỘI - 2021

BỘ THÔNG TIN VÀ TRUYỀN THÔNG
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN THỊ HỘI

MÔ HÌNH HÀNH VI VÀ QUAN TÂM CỦA NGƯỜI DÙNG TRÊN CÁC MẠNG XÃ HỘI

CHUYÊN NGÀNH: **HỆ THỐNG THÔNG TIN**

MÃ SỐ: **9.48.01.048**

LUẬN ÁN TIẾN SĨ KỸ THUẬT

NGƯỜI HƯỚNG DẪN KHOA HỌC:

1. **PGS.TS TRẦN ĐÌNH QUẾ**
2. **PGS.TS ĐÀM GIA MẠNH**

HÀ NỘI - 2021

LỜI CAM ĐOAN

Tôi xin cam đoan tất cả các nội dung trong luận án: ***"Mô hình hành vi và quan tâm của người dùng trên các mạng xã hội"*** là công trình nghiên cứu của riêng tôi, dưới sự hướng dẫn khoa học của PGS.TS.Trần Đình Quế và PGS.TS.Đàm Gia Mạnh. Tất cả các tài liệu tham khảo sử dụng trong luận án đều được nêu rõ nguồn gốc trong danh mục các tài liệu tham khảo. Tất cả các kết quả, số liệu sử dụng trong luận án là trung thực và chưa được người khác công bố trong bất kỳ công trình khoa học nào.

Hà Nội, ngày tháng năm 2021

TM. TẬP THỂ HƯỚNG DẪN KHOA HỌC	TÁC GIẢ LUẬN ÁN
PGS.TS. Trần Đình Quế	Nguyễn Thị Hội

LỜI CẢM ƠN

Trong quá trình hoàn thành luận án này, tôi đã được các thầy hướng dẫn tận tình chỉ bảo. Tôi xin kính gửi lòng biết ơn sâu sắc nhất đến thầy PGS.TS Trần Đình Quế, thầy đã tận tình hướng dẫn trong quá trình định hướng nghiên cứu, đặt vấn đề nghiên cứu, phương pháp nghiên cứu khoa học, cho đến những công việc cụ thể trong trình bày các bài báo khoa học, các báo cáo và luận án. Tôi cũng bày tỏ lòng biết ơn sâu sắc đến thầy PGS.TS Đàm Gia Mạnh, thầy đã tận tình giúp đỡ và thường xuyên động viên khích lệ tôi, hướng dẫn tôi cách viết tỉ mỉ, hướng tiếp cận cũng như hoàn thiện các báo cáo và luận án.

Tôi xin chân thành cảm ơn Ban lãnh đạo Học viện Công nghệ Bưu chính Viễn thông, các thầy cô Khoa Đào tạo Sau đại học đã động viên, giúp đỡ và tạo điều kiện thuận lợi cho tôi trong suốt quá trình thực hiện luận án. Tôi cũng xin cảm ơn các thầy cô Khoa Công nghệ thông tin đã có nhiều đóng góp quý báu giúp tôi hoàn thiện luận án, sự tận tình hướng dẫn, động viên của các thầy cô đã giúp tôi tự tin hơn trong con đường nghiên cứu khoa học. Tôi cảm thấy mình thật sự đã học hỏi được rất nhiều kỹ năng trong nghiên cứu, thu nhận được nhiều kiến thức hơn sau những năm tháng học tập và nghiên cứu tại cơ sở đào tạo của Học viện.

Luận án này không thể hoàn thành tốt nếu như không có sự hỗ trợ và tạo điều kiện thuận lợi từ Ban giám hiệu Trường Đại học Thương mại và các thầy cô ở Khoa Hệ thống thông tin kinh tế và Thương mại điện tử cũng như các thầy cô ở Bộ môn Công nghệ thông tin. Đặc biệt tôi rất cám ơn các bạn sinh viên K50S, K51S và K52S đã hỗ trợ tôi trong việc thu thập dữ liệu phục vụ cho quá trình thực nghiệm.

Con xin cảm ơn mẹ, chồng và hai con cùng các anh chị trong gia đình, đặc biệt em trai PGS.TS Nguyễn Mạnh Hùng, đã luôn động viên, giúp đỡ, khích lệ và góp ý cho luận án được hoàn thành.

Tác giả luận án

Nguyễn Thị Hội

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

TỪ VIẾT TẮT	DIỄN GIẢI	
	TIẾNG ANH	TIẾNG VIỆT
IDF	Inverse Document Frequency	Tần số nghịch đảo của một từ, cụm từ trong văn bản
IRS	Information Retrieval Similarity	Độ tương tự trích xuất thông tin
LSA	Latent Semantic Analysis	Phân tích ngữ nghĩa tiềm ẩn
Sim	Similarity	Độ tương tự
TF	Term Frequency	Tần suất của một từ, một cụm từ xuất hiện trong văn bản
TCAM	Temporal Context-Aware Mixture Model	Mô hình hỗn hợp thống kê lớp tiềm ẩn
UIW	User Interest Weight	Trọng số quan tâm của người dùng
WFST	Weighted Finite State Transducer	Chuyển đổi trạng thái trọng số hữu hạn
TBTĐ		Trung bình độ lệch tuyệt đối
TBTgĐ		Trung bình độ lệch tương đối

CÁC KÝ HIỆU

Ký hiệu	DIỄN GIẢI	
	TIẾNG ANH	TIẾNG VIỆT
B	Behavior	Hành vi
C	Comment	Bình luận
c_i		Bình luận thứ i
E	Entry	Bài viết
e_j		Bài viết thứ j
$\boldsymbol{e_j}$		Véctơ của bài viết j
G	Group	Nhóm/ Cộng đồng
g_k		Nhóm thứ k
$\boldsymbol{g_k}$		Véctơ của nhóm thứ k
J	Join a group	Gia nhập một nhóm
L	Like an entry	Thích một bài viết
N	Network	Mạng
P	Post an entry	Đăng một bài viết
T	Topic	Chủ đề
t_x		Chủ đề x
$\boldsymbol{t_x}$		Véctơ của chủ đề x
U	User	Người dùng
u_y		Người dùng y
$\boldsymbol{u_y}$		Véctơ của người dùng y
$cont$	Content	Nội dung
cat	Category	Thể loại
des	Description	Mô tả
emo	Emotion	Cảm xúc
$name$	Name	Tên
tag	Tag	Đánh dấu
$sent$	Sentiment	Quan điểm
sty	Style	Kiểu/ Loại
$cor(e_x,t)$	Corellation between e_x and t	Mức độ liên quan của bài viết e_x với chủ đề t
$sim(x,y)$	Similar between x and y	Độ tương tự giữa x và y
$int(x,y)$	Interest of x to y	Quan tâm của x đến y

DANH MỤC BẢNG BIỂU

DANH MỤC HÌNH

MỞ ĐẦU

Tính cấp thiết của luận án

Mạng xã hội (social network) xuất hiện vào những năm cuối thế kỷ 20 đã tạo điều kiện thuận lợi cho hàng triệu người trên thế giới kết nối, thiết lập và duy trì các mối quan hệ cũng như tiếp cận và chia sẻ thông tin với nhau. Ảnh hưởng của mạng xã hội đến mọi mặt trong đời sống xã hội đang ngày càng khẳng định rõ vai trò của chúng trong nhiều lĩnh vực từ giáo dục, kinh doanh, sức khỏe, du lịch… đến các vấn đề xã hội như phát hiện gian lận hoặc lừa đảo, phát hiện tâm lý tội phạm, bạo lực xã hội, phát hiện tin tức giả (fake news) được thể hiện trong nhiều công trình nghiên cứu như [30] [37] [38] [73] [81] [93] [137] [146].

Mạng xã hội đã được người dùng cá nhân, các doanh nghiệp, các nhà quản lý sử dụng như một kênh truyền thông quảng bá mới, với nhiều ưu thế như chi phí tiết kiệm, có hiệu quả lan truyền cao, có thể tiếp cận với nhiều nhóm đối tượng khác nhau trong các hoạt động sản xuất kinh doanh của các tổ chức, doanh nghiệp. Nhiều công trình nghiên cứu [1] [7] [12] [38] [44] [69] [73] [85] đã xem xét đến hiệu quả và sự phổ biến của mạng xã hội trong các hoạt động sản xuất kinh doanh của các tổ chức, doanh nghiệp.

Các nghiên cứu về khai phá quan tâm của người dùng (*user interest*) có vai trò quan trọng đối với các tổ chức, doanh nghiệp trong các chiến dịch quảng bá thương hiệu, giới thiệu sản phẩm, gợi ý dịch vụ, đặc biệt có nhiều ứng dụng trong thực tế như [1] [2] [9] [12] [16] [18] [22] [25]: xây dựng hệ thống khuyến nghị người dùng (*user recommendation system*); các ứng dụng của các chương trình hay chiến lược quảng cáo (*advertising campaign*); ứng dụng hệ thống giới thiệu sản phẩm (*product introduction systems*)…

Bên cạnh đó, việc xác định được xu hướng quan tâm (*interest trend*) của người dùng trên các trang mạng xã hội, các trang web, hay các phương tiện truyền thông xã hội (*social media*) ngày càng được chú ý và đóng vai trò quan trọng trong các ứng

dụng thực tiễn đối các tổ chức, doanh nghiệp và người bán hàng. Chúng giúp người dùng rút ngắn thời gian phân nhóm khách hàng, xác định tốt hơn nhóm khách hàng mục tiêu cho trong hoạt động sản xuất, kinh doanh và điều phối các chiến lược cũng như xây dựng được các chiến lược quảng cáo cá nhân hóa người dùng hiệu quả hơn [25] [28] [32] [37] [43] [47] [49] [50] [60] [72] [77] [108] [111] [114] [118] [143] [148] [158].

Khi sử dụng các phương tiện truyền thông xã hội và các mạng xã hội, các chiến dịch quảng cáo, các chiến lược bán hàng của các tổ chức, doanh nghiệp đã chuyển dần sang phương thức tương tác, trao đổi giữa người bán và người mua hơn là các chương trình chạy quảng bá, không tập trung vào các mục tiêu cụ thể như trước đây. Hành vi và xu hướng quan tâm của người dùng trên các mạng xã hội thường được thể hiện thông qua các bài đăng (*tweets, status, posts* ...), các câu lệnh tìm kiếm (*search queries*), các bài đánh giá (*reviews*), các bài chia sẻ từ phương tiện truyền thông xã hội khác, các hành vi thích (*like*), theo dõi (f*ollow*) ...

Theo khảo sát của luận án, có một số cách phát hiện mối quan tâm của người dùng phổ biến trên các trang mạng xã hội bao gồm:

- Phát hiện quan tâm của người dùng dựa trên trích xuất thông tin cá nhân (*profile*) [14] [31] [103] [166];
- Phát hiện quan tâm của người dùng dựa trên phân tích các liên kết của người dùng (*follows, link*) [4] [25] [28] [43] [48] [107];
- Phát hiện quan tâm của người dùng dựa trên phân tích hành vi thích, đánh dấu hoặc đăng bài (*like, tags, post*) [50] [63] [76] [77] [87] [108] [121] [144].

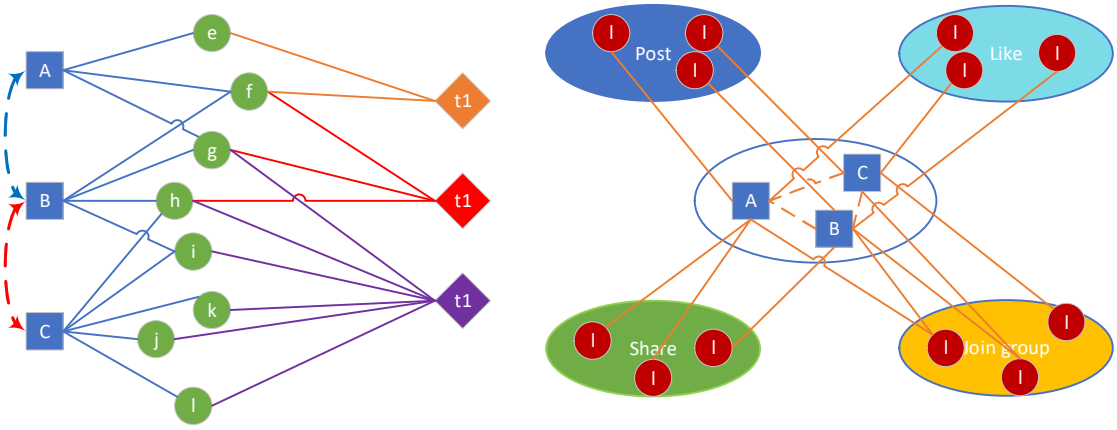
Tuy nhiên, hiện nay các thông tin cá nhân của người dùng trên các mạng xã hội rất khó thu thập bởi yêu cầu bảo mật người dùng của các hệ thống, hoặc người dùng thường xuyên không cung cấp, cập nhật đầy đủ các thông tin, hoặc các thông tin của người dùng thường quá rời rạc cũng gây trở ngại trong các nghiên cứu. Vì vậy, các nghiên cứu về phát hiện quan tâm của người dùng trên các mạng xã hội gần đây thường đi theo hai hướng tiếp cận chính:

- Tập trung phân tích về các liên kết, cấu trúc của mạng xã hội, các kết nối quan hệ bạn bè, danh sách những người được theo dõi... của người dùng trên các mạng xã hội như trong [4] [21] [23] [28] [43] [60] [105] [108] [111]
- Tập trung phân tích các bài đăng, các thẻ đánh dấu, các bài chia sẻ, các bình luận và các đối tượng được tạo ra trong quá trình hoạt động của người dùng trên các mạng xã hội [107], [114] [118] [124] [125] [143] [145] [157] [159], hướng tiếp cận này sẽ loại bỏ được vấn đề về cấu trúc mạng, sự khó khăn trong tiếp cận thông tin cá nhân người dùng cũng như trong thu thập các liên kết bạn bè của người dùng. Đa số các công trình nghiên cứu hiện nay đều theo cách tiếp cận này và luận án cũng tập trung xem xét các đối tượng được sinh ra trong quá trình tương tác của người dùng trên các mạng xã hội bao gồm các bài viết, thẻ đánh dấu, các nhóm tham gia, các bài chia sẻ…

Từ khảo sát các kết quả nghiên cứu có được tác giả cho rằng các nghiên cứu phát hiện quan tâm của người dùng cho đến nay tập trung chủ yếu vào việc xác định hoặc khám phá quan tâm của từng cá nhân người dùng dựa trên từng đối tượng nghiên cứu được tiếp cận. Có rất ít nghiên cứu xem xét sự liên quan hay mối tương quan giữa những người dùng có cùng quan tâm với nhau. Ví dụ như: có hai người dùng ***a*** và ***b***, cùng quan tâm đến các trận đấu bóng đá ngoại hạng. Họ thường xuyên đăng, thích, bình luận các bài viết về các trận đấu, về một số cầu thủ, về lịch trình thi đấu của một số câu lạc bộ... Khi đó có thể nói rằng hai người dùng ***a*** và ***b*** có cùng quan tâm đến nội dung bóng đá hoặc rộng hơn là chủ đề thể thao.

Câu hỏi đặt ra là: Khi có một bài viết về một trận đấu bóng đá mà người dùng ***a*** thích và chia sẻ lại thì liệu người dùng ***b*** có thích và chia sẻ lại bài viết đó hay không? Hoặc liệu hai người dùng này có thể cùng tham gia một nhóm có các chủ đề về bóng đá hay không? Hoặc khi có một sự kiện thể thao nào đó xảy ra trên mạng xã hội, nếu người dùng ***b*** chú ý đến và theo dõi sự kiện đó thì liệu người dùng ***a*** có quan tâm và theo dõi sự kiện đó hay không?

Để trả lời các câu hỏi này, ngoài việc xác định được chủ đề quan tâm của từng cá nhân người dùng thì còn cần phải làm rõ ràng hơn *mối tương quan giữa các chủ đề quan tâm của người dùng đó với những người dùng khác trên mạng xã hội*.



Hình 0.1: Bài toán phát hiện quan tâm của người dùng

Do đó, luận án nghiên cứu và phân tích các bài đăng của người dùng như trạng thái trên mạng Facebook.com, các nội dung đăng trên mạng Twitter.com ... Các hành vi của người dùng như đăng bài viết, chia sẻ bài viết, thích bài viết, hành vi gia nhập nhóm ... Từ đó, mô hình hóa người dùng dựa trên các đối tượng này và xây dựng một độ đo tương tự để xác định mối tương quan giữa chủ đề quan tâm của người dùng trên các mạng xã hội.

Mục tiêu của luận án và nội dung nghiên cứu

Mục tiêu của luận án

Mục tiêu của luận án là giải quyết ba bài toán sau:

- Thứ nhất, mô hình hóa bài viết của người dùng trên các mạng xã hội dựa trên nhiều đặc trưng và phân loại các bài viết đó theo các chủ đề. Các bài viết được luận án đề xuất biểu diễn dựa trên năm đặc trưng gồm: nội dung, thể loại, thẻ đánh dấu, quan điểm và cảm xúc. Dựa trên cách biểu diễn này luận án ước lượng độ tương quan của các bài viết với các chủ đề nhằm phát hiện các quan tâm của người dùng theo các chủ đề đó.

- Thứ hai, mô hình hóa người dùng trên các mạng xã hội theo các hành vi và phân loại họ dựa trên các chủ đề mà họ quan tâm. Luận án đề xuất biểu diễn người dùng trên các mạng xã hội dựa trên các hành vi đăng bài viết, chia sẻ bài viết, thích bài viết, tham gia nhóm trên các mạng xã hội. Dựa trên cách biểu diễn người dùng này, luận án ước lượng độ tương quan giữa các người dùng theo các chủ đề để tìm ra các quan tâm của họ.
- Cuối cùng, ước lượng độ tương tự giữa hai người dùng theo các chủ đề và xem xét mối tương quan giữa những người dùng đó dựa trên các hành vi họ đã thực hiện.

Nội dung nghiên cứu của luận án

Dựa trên mục tiêu đã trình bày luận án tập trung giải quyết các bài toán sau đây:

Mô hình hóa bài viết của người dùng trên các mạng xã hội và phân loại các bài viết theo các chủ đề. Để giải quyết bài toán này, luận án nghiên cứu và phân tích các đặc trưng liên quan đến bài viết của người dùng trên các mạng xã hội. Do các bài viết trên mạng xã hội là các văn bản ngắn (*short-text*) nên cần xem xét các kỹ thuật để bổ sung ngữ nghĩa cho bài viết rồi biểu diễn theo véctơ bài viết của người dùng dựa trên các đặc trưng này.

Mô hình hóa các chủ đề dựa trên danh sách từ đặc trưng và biểu diễn dưới dạng véctơ đặc trưng. Dựa trên mô hình bài viết và mô hình biểu diễn chủ đề, luận án xây dựng một độ đo tương tự giữa các bài viết và các chủ đề để phân loại các bài viết theo các chủ đề dựa trên độ đo tương tự này.

Mô hình hóa người dùng trên các mạng xã hội và phân loại các người dùng theo các chủ đề. Nghiên cứu và phân tích các hành vi đặc trưng liên quan đến các hành động phổ biến của người dùng trên các mạng xã hội, sau đó biểu diễn người dùng dựa trên các hành vi đã nghiên cứu. Để làm được điều này, luận án sẽ biểu diễn các hành vi của người dùng thành các véctơ theo không gian của các bài viết và không gian

các chủ đề. Xây dựng một độ đo tương tự giữa người dùng và các chủ đề dựa trên các hành vi, từ đó, phân loại người dùng theo các chủ đề dựa trên độ đo tương tự này.

Xác định mối tương quan giữa quan tâm của người dùng trên các mạng xã hội với các hành vi của họ. Luận án thực hiện so sánh và ước lượng giữa độ tương tự theo người dùng dựa trên các hành vi và độ tương tự của người dùng dựa trên các chủ đề quan tâm của họ. Mục tiêu là chỉ rõ được mối tương quan giữa các chủ đề quan tâm và hành vi mà người dùng thực hiện trên các mạng xã hội.

Các vấn đề nghiên cứu của luận án được minh họa như trong **Hình 0.2,** luận án phân tích, nghiên cứu các hành vi phổ biến của người dùng bao gồm: hành vi đăng bài (*post*), hành vi thích (*like*) bài viết, thích các bình luận, hoặc bày tỏ cảm xúc qua các biểu tượng cảm xúc (*emotion icon*), hành vi bình luận (*comment*) trong các bài viết, hành vi chia sẻ (*share*) các bài viết, hành vi tham gia các nhóm (*join group*) trên mạng xã hội. Luận án nghiên cứu và phân tích các đặc trưng của bài viết gồm: nội dung (*content*) bài viết, các đánh dấu (*tags*), các biểu tượng cảm xúc (*emotion*), các phân loại của bài viết (*category*) và quan điểm của bài viết (*sentiment*).

Đối tượng nghiên cứu và phạm vi nghiên cứu

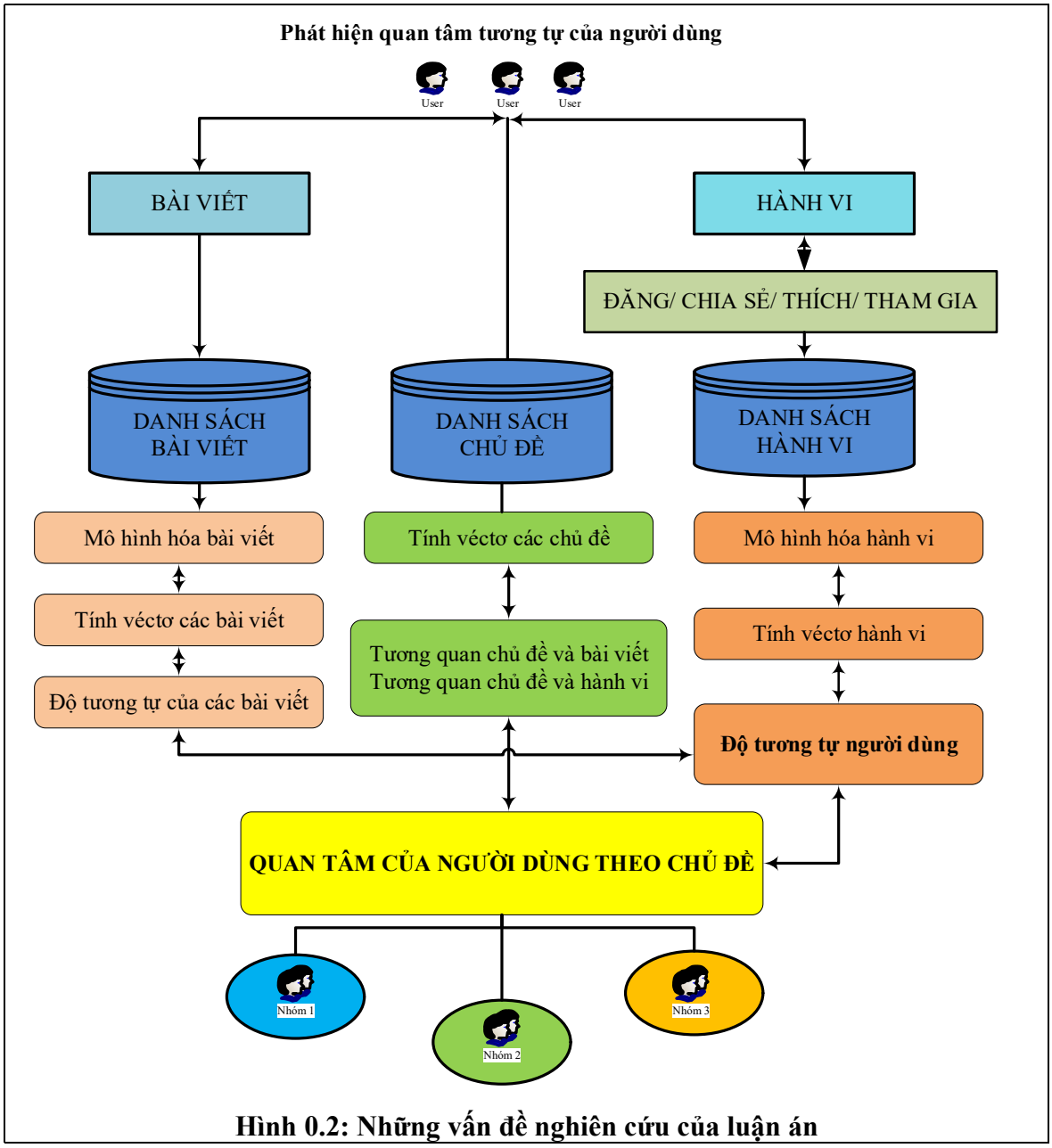
Đối tượng nghiên cứu

Với mục tiêu đã đề ra của luận án, đối tượng nghiên cứu của luận án bao gồm: Các kỹ thuật và phương thức tiền xử lý cho các văn bản ngắn; Các mô hình và phương pháp ước lượng độ tương tự giữa hai đối tượng có nhiều đặc trưng .

Phạm vi nghiên cứu

- Nghiên cứu và phân tích các đối tượng chứa văn bản sinh ra dựa trên hoạt động của người dùng cùng các hành vi của người dùng trên mạng xã hội.
- Nghiên cứu và phân tích các chủ đề trên mạng xã hội cùng các độ đo tương tự giữa các đối tượng trên mạng xã hội.
- Tổng hợp nghiên cứu, phân tích các đặc trưng chứa văn bản của bài viết và một số hành vi phổ biến của của người dùng trên các mạng xã hội cùng với các độ

đo tương tự để trả lời cho câu hỏi: Nếu có hai người dùng tương tự nhau theo các hành vi trên mạng xã hội thì họ có quan tâm các chủ đề tương tự nhau hay không? Và nếu hai người dùng thường xuyên quan tâm các chủ đề giống nhau liệu họ có nhiều điểm tương đồng nhau theo các hành vi hay không?



Hình 0.2: Những vấn đề nghiên cứu của luận án

Hiện nay, dữ liệu trên các mạng xã hội rất phong phú, đa dạng với nhiều loại dữ liệu khác nhau như dữ liệu văn bản (*text*), dữ liệu hình ảnh (*image*), dữ liệu phim (*video*), dữ liệu là các ký hiệu (*symbol*) ... Tuy nhiên, trong luận án này chỉ nghiên cứu và phân tích dữ liệu văn bản cùng các biểu tượng thể hiện cảm xúc và một số

hành vi phổ biến mà một số trang mạng xã hội cung cấp như hành vi đăng bài viết, hành vi thích và hành vi gia nhập một nhóm trên mạng xã hội. Còn các loại dữ liệu xã hội khác không phải là đối tượng nghiên cứu của luận án này.

Phương pháp nghiên cứu

Các phương pháp nghiên cứu:

- *Phương pháp luận:* Phân tích, so sánh, tổng hợp, đánh giá trên các kết quả nghiên cứu đã có, từ đó đề xuất hướng giải quyết và cách tiếp cận của luận án
- *Phương pháp đánh giá dựa trên cơ sở toán học*: Kiểm nghiệm các mô hình đề xuất bằng các thực nghiệm và đánh giá
- *Phương pháp đánh giá bằng thực nghiệm*: Thu thập dữ liệu, cài đặt các mô hình đề xuất, xây dựng các bộ dữ liệu mẫu, thực hiện thử nghiệm trên các bộ dữ liệu mẫu và phân tích, đánh giá kết quả thử nghiệm.

Thu thập dữ liệu thực nghiệm và đánh giá

Để đánh giá và kiểm nghiệm các mô hình đề xuất trong luận án, luận án thực hiện thu thập dữ liệu từ 03 nguồn dữ liệu chính là Facebook.com, Twitter.com và YouTube.com

Facebook.com là một dịch vụ mạng xã hội do công ty Facebook Inc. điều hành, có trụ sở tại Menlo Park, California, USA. Tính đến tháng 9 năm 2020, Facebook hiện có hơn 2.8 tỷ người sử dụng hằng tháng, hiện nay, Facebook là mạng xã hội phổ biến và có lượng người dùng lớn nhất trên thế giới.

Twitter.com là một dịch vụ mạng xã hội cho phép người dùng có thể cập nhật các mẩu tin nhỏ lên tường của mình, mỗi mẩu tin nhỏ đó gọi là tweet. Twitter được sở hữu bởi Twitter Inc. Hiện có hơn 35 công ty khắp thế giới và số lượng người dùng đang ngày càng tăng lên.

YouTube.com là trang dịch vụ chia sẻ video, YouTube do ba nhân viên cũ của PayPal là Chad Hurley, Steve Chen và Jawed Karim thành lập vào năm 2005. Sử dụng công nghệ HTML5 để hiển thị nhiều nội dung và đặc trưng của các loại video

khác nhau.

Ngoài ra, luận án có sử dụng thêm hai bộ dữ liệu chuẩn để so sánh khi thực nghiệm là 20 NewsGroups [41] [106] và SemEval-2017 [74][106].

- Bộ dữ liệu 20 NewsGroups có với 20 danh mục hay nhãn, có 11.293 tài liệu trong tập huấn luyện, có 7.528 trong tập kiểm thử. Bộ dữ liệu được lưu trong tập tin list.csv gồm 2 cột: Số thứ tự tài liệu (document_id number) và tên nhãn. Ngoài ra kèm theo 20 tập tin, mỗi tập tin chứa các tài liệu của 1 nhóm tương ứng.
- Bộ dữ liệu SemEval-2017, chứa dữ liệu để phân loại cảm xúc được thu thập trên mạng xã hội Twitter. Bộ dữ liệu SemEval-2017 được xây dựng dựa trên 5 bước chính xử lý trên tiếng Ả Rập và tiếng Anh: A là đưa ra một tweet, xác định xem cảm xúc của nó là *tích cực, tiêu cực hay trung lập*. B là đưa ra một tweet và một chủ đề, phân loại cảm xúc được truyền đạt về chủ đề đó trên hai thang điểm *tích cực và tiêu cực*. C là đưa ra một tweet và một chủ đề, phân loại tình cảm được truyền tải trong tweet về chủ đề đó theo năm thang điểm: *rất tích cực, khá tích cực, trung lập, khá tiêu cực và rất tiêu cực.* D là đưa ra một tập hợp các tweet về một chủ đề, ước lượng sự phân bố các tweet trên các lớp *tích cực và tiêu cực.* E là đưa ra một tập hợp các tweet về một chủ đề, hãy ước lượng sự phân bố của các tweet trong năm lớp: *rất tích cực, tích cực, trung lập, tiêu cực và rất tiêu cực.*
- Bộ dữ liệu thực được luận án sử dụng trong các thực nghiệm được thu thập qua ba giai đoạn, dựa trên thu thập tự động (API module thu thập tự động [136]) và dựa trên 4 nhóm sinh viên tình nguyện thu thập thủ công. Mỗi nhóm từ 8 - 16 sinh viên thu thập vào thời gian thống kê trong Bảng 0.1. Bộ dữ liệu được chia dùng để xác định thể loại, quan điểm và cảm xúc lưu trong tập tin scv gồm có 03 cột gồm số thứ tự bài viết (id_number, bài viết, giá trị).

  Chi tiết các bộ nhỏ được lưu trong Bảng 0.2

Bảng 0.1: Chi tiết thu thập dữ liệu thực nghiệm

Nhóm	Số lượng sinh viên	Thời gian thu thập
1	12	12/2015 – 02/2016, 03/2016 – 05/2016
2	10	03/2017 – 05/2017
3	16	03/2018 – 05/2018, 09/2018 – 10/2018

Cấu trúc bộ dữ liệu: Do phạm vi luận án chỉ tập trung nghiên cứu các bài viết, các đặc trưng của bài viết và các hành vi thể hiện trên các bài viết chứa văn bản nên sau khi đã loại bỏ những bài viết không chứa văn bản, những hành vi không được đưa vào nghiên cứu, những người dùng không tham gia bất kỳ nhóm cộng đồng nào, hoặc chưa từng đăng một bài viết nào, ... luận án thu được bộ dữ liệu như trong Bảng 0.2.

Bảng 0.2: Cấu trúc tập dữ liệu thu thập của luận án

	Người dùng	Bài viết		Hành vi	
		Số lượng	Đặc trưng	Số lượng	Hành động
Facebook	200	2000	Nội dung Thể loại Thẻ đánh dấu Quan điểm Cảm xúc	6000	Đăng/ Chia sẻ Thích Bình luận Tham gia nhóm
Twitter	200	2000	Nội dung Thể loại Thẻ đánh dấu Quan điểm Cảm xúc	6000	Đăng/ Chia sẻ Thích Bình luận Theo dõi nhóm
YouTube	200	2000	Tiêu đề Thể loại Thẻ đánh dấu Quan điểm Cảm xúc	6000	Xem Thích Theo dõi Bình luận

Kịch bản các thực nghiệm

Kịch bản thực nghiệm các mô hình ước lượng được đề xuất trong luận án được thực hiện theo 03 bước:

- Xây dựng bộ dữ liệu mẫu thử nghiệm;
- Xây dựng kịch bản và thực hiện chạy mô hình trên bộ dữ liệu mẫu và lưu kết quả các tham số đầu ra;

- Thảo luận, so sánh, đánh giá kết quả thực hiện, có thể so sánh với các mô hình khác và tính độ chính xác của các mô hình đề xuất.

Phương pháp đánh giá

Bảng 0.3: Các độ đo được sử dụng để đánh giá trong luận án

Tên độ đo	**Công thức tính**	**Nghiên cứu liên quan**	**Luận án sử dụng**
Độ chính xác (Precision)	Dựa trên ma trận nhầm lẫn $Prec = \frac{TP}{TP + FP}$	Aggarwal C.C et al. [11], D.Manning et al. [41], …	Đánh giá độ tương tự giữa hai bài viết Đánh giá độ tương tự giữa hai người dùng
Độ nhạy Recall	Dựa trên ma trận nhầm lẫn $Recall = \frac{TP}{TP + FN}$	Aggarwal C.C et al. [11], D.Manning et al. [41], …	Đánh giá độ tương tự giữa hai bài viết Đánh giá độ tương tự giữa hai người dùng
F1- measure hay F1- score	Dựa trên ma trận nhầm lẫn $F_1 = \frac{2 * P * R}{P + R}$	Aggarwal C.C et al. [11], D.Manning et al. [41], …	Đánh giá các thuật toán gán nhãn để tính thể loại, quản điểm và cảm xúc
Độ chính xác (Accuracy)	Dựa trên ma trận nhầm lẫn $Accu = \frac{TP + TN}{TP + FP + FN + TN}$	Aggarwal C.C et al. [11], D.Manning et al. [41], …	Đánh giá các thuật toán gán nhãn để tính thể loại, quản điểm và cảm xúc
*Sai số bình phương trung bình (*Mean Square Error)	$MSE = \frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2$	Fouss F. et al. [55], Kowsari et al. [80]	Độ tương tự giữa các cặp bài viết với các chủ đề Độ tương tự giữa các cặp người dùng với chủ đề
*Sai số tuyệt đối trung bình (*Mean Absolute Error)	$MAE = \frac{1}{n}\sum_{i=1}^{n}(\lvert p_i - r_i \rvert)$	Fouss F. et al. [55], Kowsari et al. [80]	Độ tương tự giữa các cặp bài viết với các chủ đề Độ tương tự giữa các cặp người dùng với chủ đề

Có rất nhiều độ đo được dùng để đánh giá hiệu suất hoặc độ chính xác của các mô hình khi kiểm nghiệm trong các nghiên cứu khoa học, trong luận án này, việc thực hiện đánh giá hiệu suất hoặc độ chính xác của các mô hình đề xuất được tính toán dựa theo một số phương pháp như sau: Đánh giá dựa trên độ chính xác (*accuracy*), độ nhạy (*recall*) và đánh giá dựa trên độ lệch trung bình như các nghiên cứu [13] [15] [42] [56] [80] [106] [156].

Những đóng góp chính của luận án

- Thứ nhất đề xuất biểu diễn bài viết và các chủ đề bằng véctơ; xây dựng độ đo tương tự giữa hai bài viết và độ tương quan giữa bài viết với các chủ đề. Mô hình này đã được công bố trên Tạp chí *International Journal of Advanced Computer Science and Applications (IJACSA)* (Vol. 6, No. 2, 2015) và công bố trên Tạp chí *Southeast Asian Journal of Sciences, Vol 7 No 2 (2019), ISSN 2286 – 7724*
- Thứ hai đề xuất mô hình biểu diễn bài viết mở rộng dựa trên năm đặc trưng là nội dung, thể loại, thẻ đánh dấu, quan điểm và cảm xúc; xây dựng độ đo tương tự giữa hai bài viết mở rộng và độ tương quan giữa bài viết với các chủ đề. Kết quả được công bố trên Kỷ yếu *Hội nghị quốc gia lần 9 về Nghiên cứu Cơ bản và Ứng dụng (9th National Symposium on Fundamental and Applied IT Research – FAIR'9),* 2016, trên Kỷ yếu của Hội nghị khoa học quốc tế *Advances in Information and Communication Technology, ICTA 12 – Vietnam,* 2016*,* Springer International Publishing và trên Kỷ yếu Hội nghị quốc gia lần 10 về Nghiên cứu Cơ bản và Ứng dụng (*10th National Symposium on Fundamental and Applied IT Research – FAIR'10),* 2017
- Thứ ba đề xuất mô hình biểu diễn người dùng dựa trên các hành vi đăng/chia sẻ bài viết, thích bài viết, bình luận trong bài viết và tham gia các nhóm trên mạng xã hội; xây dựng độ đo tương tự giữa hai người dùng theo các hành vi và độ tương quan giữa hành vi của người dùng với các chủ đề. Kết quả được công bố trên Kỷ yếu khoa học quốc tế *Conferences EAI International*

Conference on Industrial Networks and Intelligent Systems, INISCOM 2017, Vietnam, Springer International Publishing, và trên Tạp chí *Vietnam Journal of Computer Science*, (2018) 5:165–175, Springer Open. Tạp chí *Khoa học Công nghệ Đại học Đà Nẵng*, No.07 (128), 2018. Kỷ yếu *Hội nghị quốc gia lần 11 về Nghiên cứu Cơ bản và Ứng dụng* (11th National Symposium on Fundamental and Applied IT Research – FAIR'11, 08-2018) và Tạp chí *Journal of Science and Technology on Information and Communications*, 2018

Bố cục luận án

Ngoài phần mở đầu, kết luận và hướng phát triển cùng tài liệu tham khảo, luận án được chia thành 4 chương như sau:

Chương 1: Tổng quan về hành vi, quan tâm và mô hình người dùng trên các mạng xã hội. Trình bày một số khái niệm liên quan đến đề tài luận án như mạng xã hội, dữ liệu trên mạng xã hội, người dùng và cộng đồng người dùng trên mạng xã hội, chủ đề và quan tâm của người dùng trên mạng xã hội. Ngoài ra, các nghiên cứu liên quan đến đề tài luận án trong nước và quốc tế cũng như một số kiến thức nền tảng được sử dụng trong quá trình phân tích, đánh giá và ước lượng đối với kiểu dữ liệu văn bản ngắn trên mạng xã hội cũng được trình bày trong chương này.

Chương 2: Mô hình và quan tâm của người dùng theo nội dung bài viết. Luận án trình bày cách thức biểu diễn nội dung bài viết theo véctơ trọng số dựa trên không gian bài viết và không gian các chủ đề. Dựa trên các định nghĩa và cách thức biểu diễn này, luận án tính độ tương quan giữa người dùng với các chủ đề theo véctơ trọng số và ước lượng độ tương tự giữa hai người dùng dựa trên các véctơ trọng số này. Cuối chương hai là các thực nghiệm kiểm nghiệm lại các định nghĩa đã đề xuất và các công thức đã xây dựng dựa trên nội dung của bài viết.

Chương 3: Mô hình và quan tâm của người dùng dựa trên bài viết mở rộng nhiều đặc trưng. Luận án phân tích những nhược điểm của cách thức tính toán dựa

trên nội dung bài viết. Sau đó, luận án đề xuất mô hình biểu diễn bài viết dựa trên năm đặc trưng là nội dung, thẻ đánh dấu, thể loại, quan điểm và cảm xúc. Với mô hình bài viết này, luận án đề xuất cách ước lượng độ tương quan của người dùng theo các chủ đề và độ tương tự của các cặp người dùng. Cuối chương ba, luận án thực hiện các thực nghiệm để kiểm nghiệm và so sánh các kết quả của mô hình đề xuất trên bộ dữ liệu thực với kết quả của chương hai.

Chương 4: Hành vi và quan tâm của người dùng theo các hành vi. Luận án đề xuất mô hình biểu diễn người dùng dựa trên các hành vi đăng bài, thích bài viết, chia sẻ bài viết, tham gia nhóm trên mạng xã hội trong chương bốn. Dựa trên mô hình đề xuất này, luận án đề xuất cách tính độ tương quan giữa người dùng theo chủ đề và ước lượng độ tương tự giữa hai người dùng. Cuối chương bốn là thực nghiệm kiểm nghiệm và so sánh kết quả với các thưc nghiệm ở chương ba và chương hai, ngoài ra, cuối chương bốn luận án còn tính toán sự tương quan giữa tương tự người dùng và các chủ đề quan tâm của họ trên mạng xã hội cùng các thực nghiệm kiểm chứng và so sánh kết quả với một số mô hình khác.

CHƯƠNG 1: TỔNG QUAN VỀ HÀNH VI, QUAN TÂM VÀ MÔ HÌNH NGƯỜI DÙNG TRÊN CÁC MẠNG XÃ HỘI

Chương một luận án trình bày một số khái niệm được sử dụng phổ biến trên các mạng xã hội theo hướng tiếp cận của luận án trong mục 1.1, bao gồm: Mạng xã hội, dữ liệu trên mạng xã hội, người dùng và cộng đồng người dùng trên mạng xã hội, chủ đề, quan tâm mô hình người dùng trên mạng xã hội. Trong mục 1.2, luận án trình bày bài toán và một số công trình nghiên cứu liên quan đến bài toán phát hiện quan tâm của người dùng trên các mạng xã hội cũng như các vấn đề liên quan đến bài toán. Cuối chương một trình bày ngắn gọn về cách thức biểu diễn dữ liệu văn bản, phân tích dữ liệu văn bản ngắn, tính véctơ trọng số cho các dữ liệu văn bản được nghiên cứu trong luận án.

1.1. Mạng xã hội và hành vi của người dùng trên mạng xã hội

Trong mục 1.1 luận án trình bày các khái niệm cơ bản về mạng xã hội, dữ liệu trên mạng xã hội và những đối tượng liên quan sẽ được nghiên cứu trong luận án bao gồm: người dùng, cộng đồng người dùng, chủ đề và quan tâm của người dùng trên mạng xã hội.

1.1.1. Mạng xã hội

Mạng xã hội hay còn gọi là mạng xã hội ảo (social network) là một cấu trúc xã hội được tạo ra bởi cá nhân hoặc các tổ chức (gọi là các "node - nút"). Các nút được liên kết bởi một hoặc nhiều mối liên kết như: tình bạn, quan hệ đồng nghiệp, trao đổi lợi ích chung, trao đổi tài chính, mua bán hàng, các quan tâm chung hoặc những mối quan hệ khác [10] [35] [36] [38] [41] [64] [77] [84] [126] [131] [139] [156]. Theo nghiên cứu [41] và [156] thì các mạng xã hội là các dịch vụ dựa trên web cho phép các cá nhân có thể: (1) tạo lập một hồ sơ công khai hoặc bán công khai trong hệ thống có giới hạn, (2) kết nối hoặc chia sẻ với một danh sách người dùng, và (3) cho phép xem, chia sẻ những nội dung thực hiện bởi những người dùng khác trong hệ thống. Còn theo nghiên cứu [7] [139] thì mạng 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 được biểu diễn ở dạng đồ thị hoặc các mô tả. Theo nghiên cứu [10] thì mạng xã hội là một thuật ngữ dùng để mô tả các dịch vụ dựa trên web 2.0 cho phép các cá nhân có thể tạo lập một hồ sơ công khai hoặc bán công khai trong một miền ảo mà họ có thể kết nối và trao đổi với người dùng khác.

Các cấu trúc và dịch vụ cung cấp của mỗi mạng xã hội có thể không giống nhau, nhưng mục đích của các mạng xã hội đều dùng để kết nối người dùng trong một mạng mà trên đó cung cấp sẵn một số dịch vụ để người dùng có thể tương tác với nhau. Ví dụ như mạng xã hội Facebook (*www.facebook.com*) có cấu trúc kết nối giữa những người dùng chính là mối quan hệ bạn bè, sự theo dõi giữa những người dùng cá nhân đến các người dùng cá nhân, trang thông tin của các cá nhân hoặc tổ chức khác. Nếu người dùng cá nhân chú ý đến một trang (*page*) hoặc tham gia vào một cộng đồng thì có hành vi thích hoặc theo dõi trên trang đó. Trên mạng xã hội Twitter (*www.twitter.com*) thì cấu trúc của mạng xã hội được hình thành dựa trên hành vi theo dõi. Một người dùng có thể theo dõi một danh sách người dùng khác hoặc cùng theo dõi đến một nhóm, một tài khoản cá nhân hoặc một trang khác. Trên trang YouTube (*www.YouTube.com*) thì kết nối giữa các người dùng và các tài khoản cá nhân là theo dõi, thích, chia sẻ các video của các tài khoản khác. Các dịch vụ mà các mạng xã hội cung cấp sẽ quyết định cách thức người dùng trên mạng xã hội đó có thể tương tác với nhau như thế nào, các đối tượng nào được sinh ra trên mạng xã hội đó, các kiểu dữ liệu nào có thể được tạo ra, việc chia sẻ và phân phối trên mạng xã hội đó ...

Theo thống kê của trang *http://statistic.com* (tháng 9/2020) thì hiện nay có 4.6 tỷ người dùng Internet, trong đó có đến 3.617 tỷ người dùng các trang mạng xã hội, chiếm khoảng 44% dân số trên thế giới. Theo thống kê của *http://www.emarsys.com* và *http://www.BusinessWire.com* thì lượng người dùng khổng lồ trên các mạng xã hội có ảnh hưởng tích cực đến các hoạt động của các tổ chức, doanh nghiệp cũng như người dùng cá nhân, đặc biệt trong các hoạt động marketing, hoạt động bán hàng,

hoạt động quảng bá ... [1] [3] [35] [37] [44] [49] [69] [73]. Hơn thế nữa, các mạng xã hội đã và đang trở thành mảnh đất màu mỡ cho các bài toán ứng dụng của nhiều lĩnh vực khác nhau, từ những bài toán ứng dụng phổ biến trong phân tích dữ liệu như khai phá dữ liệu [8] [10] [14] [20] [23] [43] [48] [53], truy hồi thông tin (*information retrieval*) [10] [50] [54] [64] [72], các hệ tư vấn (*recommender systems*), khoa học web (*web science*) [4] [8] [10] [23] [42] [48]... đến nhiều ngành khoa học xã hội khác như y tế và chăm sóc sức khỏe, giáo dục, điều tra các tổ chức xã hội, đặc biệt trong các nghiên cứu về xã hội học, tâm lý học tội phạm, phân tích tin giả (fake news), y tế và sức khỏe ... hay các bài toán khai phá quan điểm (*opinion mining*), quản lý danh tiếng (*reputation management*), phân tích hành vi con người (*human behavior analysis*) [39] [45] [47] [51] [54] [57] [61].

1.1.2. Dữ liệu trên mạng xã hội

Theo nghiên cứu [132] [156] thì dữ liệu trên mạng xã hội hay dữ liệu xã hội (social data) là dữ liệu nhận được từ các phương tiện truyền thông xã hội như các trang mạng xã hội, các trang web tìm kiếm, các trang thương mại điện tử, các trang chia sẻ hình ảnh, video ...

Theo nghiên cứu [8] [9] thì dữ liệu trên các mạng xã hội biểu diễn mối quan hệ hoặc các tương tác của người dùng trên các mạng xã hội đó. Các mối quan hệ có thể hai chiều và có sự tương tác qua lại giữa những người dùng hoặc được kết nối một chiều, chẳng hạn như các kết nối theo dõi/được theo dõi trong Twitter, Facebook ...

Dữ liệu trên các mạng xã hội có thể là văn bản, hình ảnh, các video hoặc kết hợp nhiều loại dữ liệu đó với nhau. Đặc trưng cơ bản của dữ liệu trên các mạng xã hội là có dung lượng lớn, có tính liên kết, chứa nhiều nhiễu, không có cấu trúc hoặc ngữ pháp chuẩn và đặc biệt thường không đầy đủ, không hoàn chỉnh như các dữ liệu từ các nguồn sinh dữ liệu khác.

- Lớn (Big): Theo thống kê của *Facebook.com* thì mỗi ngày có khoảng 2.5 tỉ nội dung được tạo ra, có hơn 500 TB dữ liệu được lưu trữ, có 2.7 tỉ hành vi

thích và 300 triệu bức ảnh được đăng lên Facebook. Theo thống kê của www.statistic.com thì năm 2019 người dùng trên các phương tiện truyền thông xã hội đã tăng hơn 9% so với năm 2017. Các số liệu thống kê cho thấy rằng, dữ liệu trên mạng xã hội càng ngày càng khổng lồ và vẫn tiếp tục tăng thêm hàng phút, hàng giây.

- Liên kết (Linked): Bản chất mạng xã hội là sự liên kết giữa những người sử dụng trên mạng, vì vậy, dữ liệu trên mạng xã hội đều có sự liên kết hay kết nối với nhau. Các mối liên kết trên mạng xã hội có thể khác nhau, nhưng chủ yếu là dựa trên các mối quan hệ như: quan hệ bạn bè, quan hệ gia đình, quan hệ trường lớp, quan tâm chung, sở thích chung, các nhóm chia sẻ nội dung, nhóm người hâm mộ ...
- Nhiều nhiễu (Noisy): Một đặc điểm quan trọng của dữ liệu trên các mạng xã hội là nhiều nhiễu, bởi mỗi người dùng bất kỳ có thể là người mua hàng, có thể là người bán hàng, có thể là người tạo ra thông tin và cũng có thể là người thu thập thông tin. Nhiễu của dữ liệu trên các mạng xã hội thường đến từ hai nguồn chính: nhiễu từ các spammer hay những người dùng chuyên gửi các nội dung rác, truyền mã độc và nhiễu sinh ra từ các mối quan hệ của người dùng trên các trang mạng xã hội.
- Không có cấu trúc (Unstructured): Các dữ liệu do người dùng tạo ra trên các mạng xã hội thường không có cấu trúc, do nhiều người dùng sử dụng thiết bị di động để xuất bản nội dung lên các mạng xã hội như cập nhật trạng thái, gửi bài viết... kết quả là (1) văn bản thường rất ngắn, có những văn bản chỉ có một từ, một dấu hỏi (?), một dấu chấm than (!) hoặc một biểu tượng (icon) và (2) có nhiều lỗi chính tả, lỗi ngữ pháp và sự pha trộn nhiều ngôn ngữ trong một đoạn văn bản. Ví dụ sử dụng mã ASCII như :) và :(, sử dụng các từ viết tắt (ví dụ: h r u? g9!).
- Chưa hoàn chỉnh (Incomplete): Nhiều người dùng tạo ra hoặc cập nhật các thông tin trên các mạng xã hội không đầy đủ, hoặc không cho phép người khác có thể đọc được, vì vậy các thông tin về người dùng thường rời rạc,

không đầy đủ, hoặc chưa hoàn chỉnh. Ngoài ra, các dữ liệu khác được sinh ra từ người dùng trên các mạng xã hội cũng chỉ thể hiện một khía cạnh nào đó của người dùng, chúng không đầy đủ và không được thể hiện rõ ràng trên các trang cá nhân.

Dựa trên các đặc điểm của dữ liệu trên các trang mạng xã hội, có thể thấy rằng, các dữ liệu trên các mạng xã hội thường không theo quy chuẩn, không hoàn chỉnh và có nhiều nhiễu.

1.1.3. Người dùng và cộng đồng người dùng trên các mạng xã hội

Người sử dụng hay người dùng (user) trên các mạng xã hội là những người tham gia vào các mạng xã hội đó, họ thiết lập các kết nối với người dùng khác và có thể trao đổi với nhau, đọc tin tức, chơi trò chơi, tham gia vào các nhóm, tạo ra các thông tin, chia sẻ thông tin, chia sẻ dữ liệu trên các mạng xã hội [8] [9] [23] [35] [41] [51]. Những người dùng trên các mạng xã hội chính là những mắt xích kết nối trên các trang mạng xã hội, giúp các mạng xã hội tạo nên sự kết nối và duy trì sự phát triển theo thời gian. Số lượng người dùng trên các mạng xã hội ngày một phát triển nhanh chóng chính là mục tiêu nghiên cứu của các ứng dụng nhằm khai thác những ưu thế và lợi ích mà mạng xã hội mang lại.

Theo khảo sát và thống kê của LLNRR (Leibniz Library Network for Research Information) công bố trên website *www.goportis.de* vào năm 2017 thì hiện nay, trên các mạng xã hội tồn tại bốn nhóm người dùng phổ biến là người dùng sáng tạo (Maker), người dùng công nghệ (Technological), người dùng cổ điển (Classical) và người dùng kiểu mọt sách (Nerd).

- Người dùng sáng tạo thường làm việc trong các trường Đại học, trung tâm nghiên cứu ... họ sử dụng các kênh truyền thông xã hội khác nhau một cách thường xuyên, họ có thể áp dụng các kỹ năng của mình trong quá trình sử dụng và khai thác dữ liệu trên các trang mạng xã hội.

- Người dùng công nghệ là các học giả, các chuyên gia nghiên cứu, đây là nhóm người dùng phổ biến nhất trong học tập cũng như các cộng đồng nghiên cứu, những người dùng này có mối liên hệ chặt chẽ đối với công nghệ và các phương tiện truyền thông xã hội, rất thích dùng các công cụ mới, sử dụng thường xuyên là: Wikipedia, mạng xã hội như Facebook và Google+ ...
- Người dùng cổ điển là những người sử dụng ít các phương tiện truyền thông xã hội khác trong cộng đồng học tập và nghiên cứu, không cần các công cụ mới, chủ yếu sử dụng dịch vụ Web 2.0 mỗi tháng một lần hoặc ít hơn, vì lý do thực dụng hoặc vì đó là một phần bắt buộc trong công việc của mình.
- Cuối cùng, người dùng mọt sách là những người dùng tích cực tham gia vào lĩnh vực truyền thông xã hội, rất dễ tiếp thu các phương tiện truyền thông xã hội mới, tương đối không quan tâm đến các vấn đề bảo mật, quyền riêng tư và rất thích thử thách trong sử dụng các công cụ và tính năng mới của Web 2.0. Người dùng mọt sách là nhóm người dùng sử dụng nhiều công cụ truyền thông xã hội hơn bất kỳ nhóm người dùng nào khác trên các trang mạng xã hội, đặc biệt các công cụ mới, hoặc các tiện ích mà các nhà cung cấp dịch vụ vừa giới thiệu.

Cộng đồng người dùng theo [4] [9] [35] [41] [54] [64] [111] là một tập hợp người dùng trên một mạng xã hội cùng chia sẻ các sở thích, quan tâm chung về một sự kiện, đối tượng hay chủ đề nào đó. Họ có mối liên kết chặt chẽ với nhau theo cùng một mối quan tâm chung hơn so với những người dùng khác. Trong một mạng xã hội bất kỳ, có nhiều người dùng cùng quan tâm đến một chủ đề, một đối tượng hoặc một sự kiện thì họ có xu hướng kết nối với nhau để cùng chia sẻ các mối quan tâm chung đó. Các kết nối của người dùng thường theo các kiểu quan hệ gần với các quan hệ thực tế ngoài xã hội, chẳng hạn như quan hệ bạn bè, quan hệ gia đình, quan hệ đồng nghiệp ...

Nhóm hay cộng đồng người dùng trên mạng xã hội thường phụ thuộc vào tính năng được cung cấp bởi các mạng xã hội mà họ tham gia. Chẳng hạn như mạng xã

hội Facebook.com có tính năng nhóm (group), mạng xã hội Twitter.com có tính năng nhóm (list), mạng xã hội Weibo có tính năng vòng bạn bè ... Từ nghiên cứu [41] [54] [64] [111] thì đặc điểm của các nhóm hay cộng đồng người dùng trên các mạng xã hội chính là mục đích hoạt động của nhóm, có nhóm dùng làm nơi chia sẻ thông tin như các nhóm học tập, nghiên cứu, định hướng, có nhóm dùng để chia sẻ quan tâm, sở thích như nhóm hâm mộ các thần tượng, các bộ phim, người nổi tiếng, có nhóm dùng để mua bán các mặt hàng.

Mỗi nhóm thường có người đứng đầu (*admin*) và những người dùng trong nhóm là các thành viên (*member*), các thành viên trong nhóm có thể là bạn bè của nhau hoặc không. Mỗi nhóm đều có những quy định và các thành viên tham gia đều phải tuân thủ như: quy định về nội dung đăng bài, quy định về cách ứng xử, quy định về điều kiện tham gia nhóm ... Mỗi nhóm thường có đặc trưng là tên nhóm, kiểu nhóm và các mô tả hay quy định chung về hoạt động của nhóm. Nhóm người dùng cũng tạo ra rất nhiều thông tin và nhóm cũng là nơi các thành viên chia sẻ mối quan tâm chung của mình với những người dùng khác.

1.1.4. Mô hình người dùng trên các mạng xã hội

Mô hình người dùng (*user moderling*) là cách thức biểu diễn thông tin cá nhân của người dùng thông qua các đặc trưng mà người dùng thể hiện trên các mạng xã hội. Mô hình người dùng theo các nghiên cứu [8] [9] [135] [18] thường được xây dựng dựa trên các đặc trưng sau của người dùng:

- Đặc điểm cá nhân hoặc nhân khẩu học (*personal characteristics or demographics*): bao gồm từ thông tin cơ bản như giới tính hoặc tuổi tác đến các thông tin xã hội như tình trạng hôn nhân, nghề nghiệp ...
- Quan tâm và sở thích (*interests and preferences*): trong một hệ thống có thể đáp ứng mô tả sự quan tâm của người dùng đối với các đối tượng nhất định các sản phẩm, tin tức hoặc tài liệu ...

- Nhu cầu và mục tiêu (*needs and goals*): Các nhu cầu và mục tiêu mà người dùng muốn đạt được khi sử dụng mạng xã hội như mua hàng nhanh hơn, tìm thấy sản phẩm tốt hơn, hoặc có nhiều thông tin liên quan hơn ...
- Trạng thái tinh thần và thể chất (*mental and physical state*): mô tả các đặc điểm cá nhân của người dùng như giới hạn về thể chất (khả năng nhìn, khả năng đi lại, nhịp tim, huyết áp...) hoặc trạng thái tinh thần (chịu áp lực, tải trọng nhận thức) ...
- Nền tảng tri thức (*knowledge and background*): mô tả tri thức của người dùng về một chủ đề hoặc hệ thống, tri thức này sẽ thay đổi theo thời gian.
- Hành vi của người dùng (*user behavior*): Việc quan sát và phân tích hành vi của người dùng thường là giai đoạn sơ bộ để suy ra thông tin cho một trong nhiều vấn đề về người dùng trên mạng xã hội. Thường khai thác dựa trên lịch sử tương tác hoặc lịch sử đã để lại của người dùng trên các mạng xã hội.
- Ngữ cảnh (*context*) là những thông tin mô tả đặc trưng của tình huống mà sự việc xảy ra, trên mạng xã hội thường đề cập đến môi trường của người dùng như vị trí địa lý, thời gian, thiết bị sử dụng trong các tương tác của người dùng.
- Đặc điểm tính cách cá nhân (*individual traits*): Đề cập đến các đặc điểm tính cách cá nhân của người dùng như hướng nội hoặc hướng ngoại hoặc kiểu nhận thức và phong cách học tập ...

Cũng theo các nghiên cứu [8] [9] [135] [18] thì sau khi xây dựng xong mô hình của người dùng, mỗi người dùng sẽ được biểu diễn bằng một tập thông tin cá nhân gọi là hồ sơ người dùng (*user's profile*) về vấn đề được nghiên cứu. Khi đó, mô hình người dùng sẽ tương ứng với hồ sơ chứa thông tin cá nhân tương ứng như mô hình quan tâm của người dùng (*user's interest profile*), mô hình di chuyển của người dùng (*user's mobility profile*), mô hình đặc điểm cá nhân của người dùng (*user's personality profile*) ...

1.1.5. Quan tâm của người dùng trên mạng xã hội

Quan tâm là thái độ hay sự chú ý của người dùng đối với một chủ đề, sự kiện, hiện tượng nào đó xảy ra trên các trang mạng xã hội. Có nhiều khái niệm về thái độ hay sự chú ý, theo nghiên cứu [156] và [159] thì thái độ có thể là một kết quả của quá trình tâm lý, thái độ không thể quan sát trực tiếp được, nhưng *thái độ hay quan tâm có thể suy ra được từ những lời nói và các hành vi của con người*. Từ đó có thể thấy rằng, các hành vi của người dùng thể hiện trên các mạng xã hội chính là thái độ hay quan tâm của người dùng đó đến các đối tượng, sự kiện hay hiện tượng trên các trang mạng xã hội.

Theo nghiên cứu [47] thì quan tâm của người dùng trên các mạng xã hội được chia thành hai nhóm: tường minh và không tường minh. Quan tâm tường minh là những quan tâm mà các mạng xã hội như Twitter hiển thị để hỏi người dùng trực tiếp hoặc phân tích các thẻ đánh dấu có sẵn chủ đề quan tâm. Quan tâm không tường minh phải phân tích mới phát hiện được như phân tích hành vi, phân tích liên kết mạng, phân tích cấu trúc mạng ...

Khi người dùng thể hiện quan tâm hay tỏ thái độ về một đối tượng, sự kiện, hiện tượng nào đó đều dựa trên nhận thức và niềm tin. Bởi vậy, hành vi hay thái độ của người dùng đối với các đối tượng, sự kiện, hiện tượng trên các mạng xã hội là cách thức người dùng thể hiện mối quan tâm của họ đối với các vấn đề xảy đến đối với các đối tượng đó. Chẳng hạn như một người dùng thường xuyên chia sẻ các bài viết về lịch trình các trận tennis, tham gia vào nhóm cổ động viên của Man U, thường xuyên thích hoặc bình luận hình ảnh các vận động viên thì có thể xem người dùng đó quan tâm đến chủ đề thể thao, hoặc một người dùng thường xuyên chú ý đến các bộ phim bom tấn, tham gia nhóm hâm mộ một ca sĩ, diễn viên nào đó, thường xuyên theo dõi lịch chiếu và các sự kiện bên lề của các liên hoan phim thì có thể xem người dùng đó quan tâm đến chủ đề phim ảnh, giải trí. Người dùng có quan tâm hay thể hiện sự chú ý đến các đối tượng, sự kiện hoặc hiện tượng thì họ có thể tạo nên các bài viết, chia sẻ lại các bài viết từ người dùng khác hoặc nguồn khác. Ngoài ra, người dùng có thể

bày tỏ thái độ thông qua hành vi thích, hoặc rõ ràng hơn thì để lại các bình luận dưới các bài viết đó, hoặc họ tham gia các nhóm, các cộng đồng có cùng mối quan tâm với họ.

Như vậy, có thể nói rằng: *Quan tâm của người dùng trên các mạng xã hội là sự để tâm và chú ý thường xuyên đến một hoặc một số đối tượng, sự kiện, hiện tượng nào đó. Việc nghiên cứu phát hiện quan tâm của người dùng trên các mạng xã hội có thể thực hiện được bằng cách phân tích các hành vi của người dùng thực hiện trên các đối tượng, sự kiện, hiện tượng trên các mạng xã hội đó.*

1.1.6. Chủ đề trên các trang mạng xã hội

Theo từ điển Tiếng Việt thì chủ đề là vấn đề chính được đặt ra trong một tác phẩm nghệ thuật hoặc là đề tài được chọn làm nội dung chủ yếu trong học tập, sáng tác. Theo từ điển Cambridge (*https://dictionary.cambridge.org*), chủ đề là một vấn đề được thảo luận, viết hay nghiên cứu, còn theo từ điển Oxford (*https://en.oxforddictionaries.com*), chủ đề là một vấn đề được trình bày trong văn bản, bài luận hay trong cuộc hội thoại. Chủ đề của các nội dung mà người dùng đăng trên các trang mạng thường ẩn trong nội dung, người dùng thông qua các nội dung trong các bài đăng của mình để gián tiếp trình bày các chủ đề mà họ quan tâm đến, hay lĩnh vực cụ thể được trao đổi thông qua các hành vi của mình.

Có thể nói rằng, chủ đề là một mô tả vấn đề, lĩnh vực mà người dùng quan tâm đến trên các phương tiện truyền thông xã hội, chủ đề được biểu diễn là một đoạn văn bản nêu lên đặc trưng để thấy sự khác biệt của vấn đề, lĩnh vực đó với vấn đề và lĩnh vực khác.

Biểu diễn chủ đề trên các mạng xã hội được biểu diễn bằng các từ khóa, tập hợp các từ khóa, các khái niệm, các liên kết giữa các khái niệm dựa trên các tập hợp từ.

1.1.7. Hành vi của người dùng trên các mạng xã hội

Hành vi trên mạng xã hội được xem là toàn bộ các hành động có chủ đích hoặc không có chủ đích mà người dùng thực hiện trên các đối tượng được các nhà cung

cấp dịch vụ mạng xã hội sinh ra. Hành vi là toàn bộ những phản ứng, cách cư xử ra bên ngoài của một người trong một hoàn cảnh cụ thể. Hành vi là một giá trị có thể thay đổi theo thời gian, hành vi của một người có thể bị ảnh hưởng và chi phối bởi nhiều yếu tố, ví dụ như trong nghiên cứu [57] hành vi mua hàng của người dùng bị chi phối bởi 4 yếu tố là: *yếu tố trình độ văn hoá, những yếu tố mang tính chất xã hội, những yếu tố mang tính chất cá nhân* và *những yếu tố mang tính chất tâm lý*.

Theo [35] [37] [104] thì hành vi của con người mô tả những cách mà con người hành động và tương tác với thế giới bên ngoài, hay hành động và phản ứng của con người đối với người dùng khác, các thực thể khác, các tình huống hoặc sự kiện trong môi trường ảo hoặc môi trường vật lý. Phân tích hành vi là một khái niệm quan trọng trong việc nhận thức các động lực và nguyên nhân trong nhiều lĩnh vực như khoa học hành vi, khoa học máy tính, mạng xã hội, hệ thống tư vấn, quản lý quan hệ khách hàng, tổ chức đa phương tiện, phát hiện cộng đồng ...

Trong nghiên cứu về mạng xã hội, người ta đồng ý rộng rãi rằng phân tích hành vi của người dùng là chủ đề quan trọng để hiểu sâu sắc về người dùng. Các hành vi được chia thành các hành vi định tính và định lượng dựa trên loại hành động có liên quan. Hành vi định tính được đặc trưng bởi hành động của các tác nhân và hành vi định lượng được định lượng bằng đặc trưng của các thực thể. Theo [35] thì hành vi của người dùng trên mạng xã hội có ảnh hưởng trực tiếp đến các mối quan hệ, công việc, giải trí, học tập ... chúng có thể hàm chứa nhiều ý nghĩa khác nhau và có vai trò quan trọng trong nhiều ứng dụng như: khai thác thông tin và cộng đồng người dùng, các dịch vụ chăm sóc sức khỏe, học tập và nghiên cứu. Vì vậy, luận án lựa chọn phân tích hành vi làm đối tượng nghiên cứu.

1.2. Phát hiện các chủ đề quan tâm của người dùng trên các mạng xã hội

1.2.1. Phát biểu bài toán và câu hỏi nghiên cứu

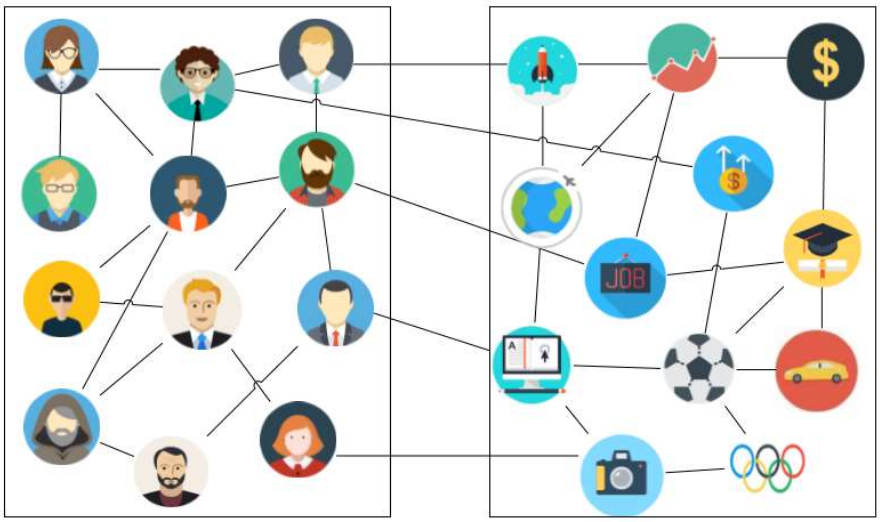
Từ các khái niệm về người dùng và chủ đề trong Mục 1.1 có thể thấy rằng, mỗi người dùng trên mạng xã hội được đặc trưng bởi rất nhiều tính chất khác nhau, có

những đặc trưng cố định như tên người dùng, vị trí địa lí, nghề nghiệp ... đến những đặc trưng không cố định như danh sách các bài viết của người dùng, danh sách bạn bè của người dùng, các liên kết mà người dùng theo dõi, các nhóm mà người dùng đang tham gia ... vì vậy, để phát hiện các chủ đề quan tâm của người dùng trên các mạng xã hội cần đi tìm quan tâm tường minh và quan tâm không tường minh thông qua các hành vi của người dùng.

Bài toán phát hiện các chủ đề quan tâm của người dùng dựa trên hành vi có thể phát biểu như sau: *Cho một tập các chủ đề trên một mạng xã hội và một tập hợp người dùng cùng các đặc trưng của họ trên mạng xã hội đó, cần đưa ra danh sách các chủ đề mà những người dùng quan tâm, chú ý đến dựa trên việc phân tích các hành vi đặc trưng của những người dùng đó.*

Xét một cách tổng quát thì bài toán phát hiện các chủ đề quan tâm của người dùng trên các mạng xã hội chính là một bài toán gán nhiều nhãn cho người dùng, các nhãn ở đây chính là các chủ đề mà người dùng quan tâm, các đặc trưng của người dùng chính là các đối tượng dùng để phân tích để gán nhãn như trong Hình 1.1. Những câu hỏi cần giải quyết của bài toán bao gồm:

- *Đối tượng nghiên cứu được lựa chọn của bài toán là gì*? chính xác hơn là những đặc trưng nào của người dùng trên mạng xã hội sẽ được lựa chọn làm đối tượng nghiên cứu và phân tích nhằm phát hiện chủ đề quan tâm của họ?
- *Những người dùng trên các mạng xã hội được biểu diễn như thế nào để phân tích và ước lượng nhằm phát hiện các quan tâm của họ?*
- *Các phương pháp hay các kỹ thuật nào sẽ được sử dụng*? trong nghiên cứu, phân tích và ước lượng để phát hiện được các chủ đề quan tâm của người dùng trên các mạng xã hội.
- *Các chủ đề quan tâm được xây dựng và biểu diễn như thế nào?*



Hình 1.1. Minh họa bài toán phát hiện chủ đề quan tâm của người dùng

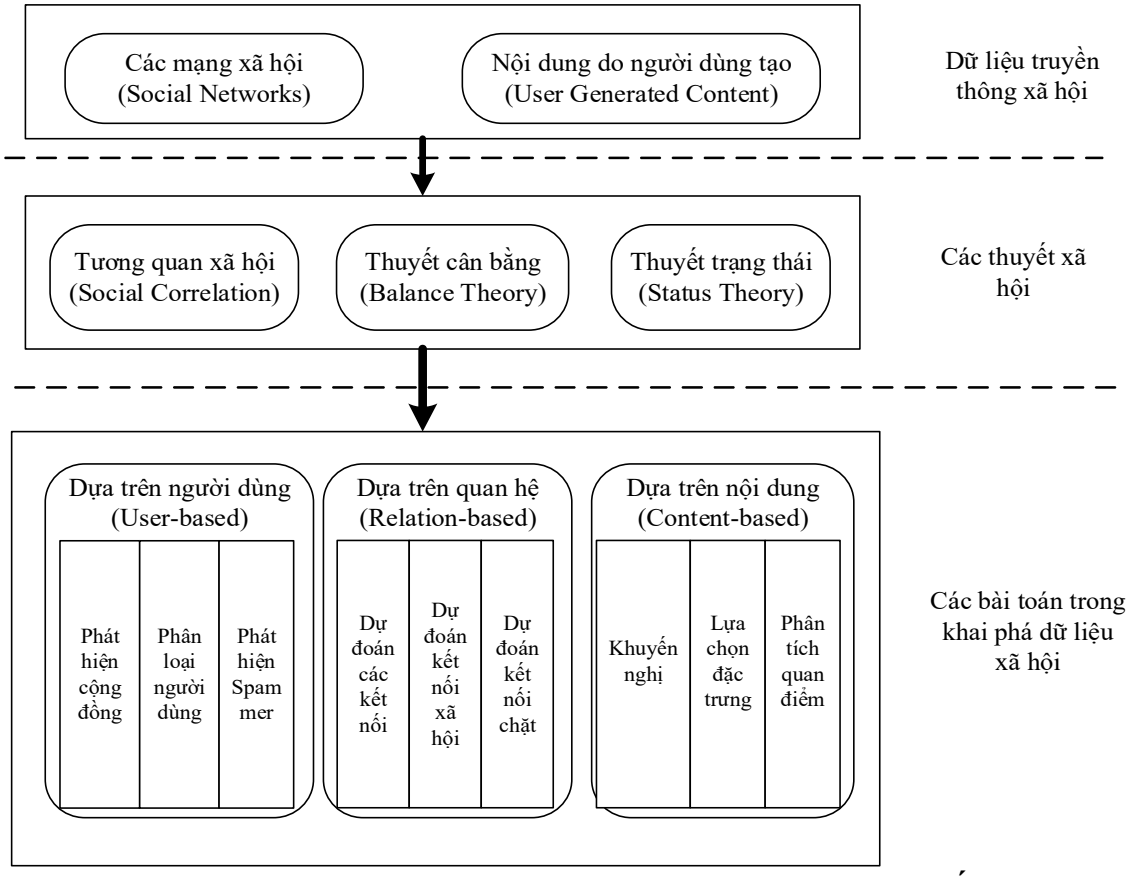
(Nguồn: Dhelm S.N. et al. [47])

1.2.2. Ứng dụng của phát hiện quan tâm của người dùng trên mạng xã hội

Theo thống kê trong [132] thì các bài toán khai phá dựa trên dữ liệu xã hội (social data) có thể tóm tắt như trong Hình 1.2, các dữ liệu xã hội là các dữ liệu có được từ các phương tiện truyền thông xã hội như các mạng xã hội và từ các phương tiện truyền thông xã hội khác. Các nghiên cứu dữ liệu xã hội chủ yếu dựa trên ba học thuyết: *thuyết tương quan xã hội, thuyết cân bằng và thuyết trạng thái*. Dựa trên các thuyết này, các nghiên cứu tập trung vào ba nhóm chính:

- Các nghiên cứu dựa trên các ứng dụng cho người dùng như phát hiện cộng đồng, phân loại các nhóm người dùng và phát hiện người dùng xấu. Các ứng dụng chủ yếu trong các lĩnh vực như giáo dục, y tế và sức khỏe, tội phạm xã hội tiềm ẩn, các tin tức giả ...
- Các nghiên cứu dựa trên các mối quan hệ của các người dùng như dự đoán các kết nối của người dùng, dự đoán các kết nối xã hội chặt chẽ và dự đoán các mối quan hệ lâu dài của các nhóm người dùng. Các ứng dụng chủ yếu trong các lĩnh vực kinh doanh, xã hội học, các hệ thống tư vấn, các hệ thống dự báo ...

- Các nghiên cứu dựa trên nội dung của các đối tượng được sinh ra bởi người dùng như các bài toán khuyến nghị người dùng, các bài toán trích chọn đặc trưng và các bài toán phân tích quan điểm. Các ứng dụng chủ yếu trong các lĩnh vực như quảng cáo, giới thiệu sản phẩm, phân tích quan điểm và so sánh, nhận xét về các sản phẩm ...



Hình 1.2: Các bài toán khai phá dữ liệu xã hội dựa trên các thuyết xã hội

(*Nguồn Tang Jiliang et al.[132]*)

1.3. Các nghiên cứu liên quan đến bài toán

1.3.1. Các hướng tiếp cận của bài toán

Theo [10] [54] và [60] thì bài toán phát hiện quan tâm của người dùng trên các mạng xã hội thường được xem xét dựa trên *nguồn thông tin được phân tích, cách thức biểu diễn các chủ đề được so sánh, các kỹ thuật được sử dụng để khai thác các mô hình và các phương pháp để đánh giá*:

- Các nguồn thông tin (*information sources*): Nguồn thông tin là các nguồn được sử dụng để trích chọn thông tin nhằm tìm kiếm chủ đề quan tâm của người dùng. Chẳng hạn như nội dung văn bản (*posts text, comments, tags*), cấu trúc mạng xã hội (*social network structures*), ảnh và video (*images and video*)... Các nguồn thông tin thường được chia thành nguồn thông tin bên trong (internal) và nguồn thông tin bên ngoài (external). Nguồn thông tin bên trong thường là các mạng xã hội đơn (single – OSN) như LinkedIn, Facebook, Twitter... và các mạng xã hội liên kết chéo (cross-system) như các trang Google+, các trang có thông tin kết nối liên quan... Nguồn thông tin bên ngoài thường là các bài đăng trên các website, các mạng tri thức, các từ điển tri thức...
- Cách thức biểu diễn các chủ đề quan tâm của người dùng (*user interest representation units*): Là cách thức dùng để biểu diễn các chủ đề quan tâm của người dùng được làm cơ sở để phát hiện, ước lượng hoặc so sánh. Thường các chủ đề quan tâm của người dùng có thể biểu diễn dựa trên từ khóa (*keywords*), hoặc biểu diễn dựa trên nhóm từ khóa (*group of keywords*), có thể biểu diễn dựa trên khái niệm (*concepts*) hoặc biểu diễn dựa trên nhóm các khái niệm (*group of concepts*) dựa trên mạng tri thức như thông qua thực thể hoặc các thể loại.
- Các kỹ thuật khai thác và phân tích cơ sở dữ liệu liên quan đến người dùng (*underlying techniques*): Hiện nay có khá nhiều kỹ thuật được sử dụng trong khai thác và phân tích để phát hiện các chủ đề quan tâm của người dùng. Điển hình như các mạng nơ ron nhúng (*embeddings neuron*); các hệ thống lọc cộng tác (*collaborative filtering*); mô hình chủ đề (*topic modelling*); dự đoán liên kết (*link prediction*); hồi quy tuyến tính (*regression*); các phương pháp dựa trên đồ thị *(graph methods*); các kỹ thuật khai thác web ngữ nghĩa. Bên cạnh đó, hiện nay các cơ sở dữ liệu chuẩn để phát hiện và phân tích về chủ đề quan tâm của người dùng vẫn chưa có một bộ dữ liệu chuẩn, các cơ

sở dữ liệu phân tích hầu hết được các nghiên cứu tự thu thập dựa trên các API hoặc dựa trên phương pháp thủ công.

- Các phương thức để đánh giá (*evaluation methodology*): Đánh giá bên trong và đánh giá bên ngoài (*intrinsic và extrinsic*) là các phương pháp được sử dụng trong đánh giá các kỹ thuật và phương pháp phát hiện chủ đề quan tâm của người dùng. Đánh giá bên trong là phương pháp đánh giá về chất lượng của cấu trúc các thông tin quan tâm của người dùng dựa trên nghiên cứu người dùng hay dựa trên đánh giá của chính người dùng. Đánh giá bên ngoài là phương pháp đánh giá cấu trúc thông tin quan tâm của người dùng dựa trên xem xét ảnh hưởng của các chủ đề quan tâm đến các ứng dụng như các hệ thống khuyến nghị người dùng, các hệ thống dự báo ...

Theo khảo sát của luận án, bài toán phát hiện quan tâm của người dùng trên mạng xã hội có hai hướng tiếp cận chính:

Một là, tập trung vào các mối liên kết của người dùng bao gồm theo cấu trúc của mạng xã hội, theo các kết nối, các liên kết của người dùng còn gọi là tập trung vào người dùng (*user-centric*)

Hai là tập trung vào các đối tượng được sinh ra bởi người dùng trên mạng xã hội như các bài viết, các bài chia sẻ, các hành vi như thích, bình luận còn gọi là hướng đối tượng (*object-centric*).

Với hướng tiếp cận *user-centric*, các nghiên cứu tập trung phát hiện quan tâm của người dùng bằng cách phân tích các mối liên kết, các cấu trúc mạng, các đặc trưng ít thay đổi của người dùng trên các mạng xã hội như thông tin cá nhân của người dùng (*profile's user*), vị trí địa lý (*locality*), các tương tác và các hành vi di chuyển (*mobility and social interactions*), các kết nối bạn bè, hàng xóm (*friends, neighbors*), các liên kết được theo dõi (*follows experts, follow famous people*), các kết nối và tương tác thông qua các ứng dụng (*internet connections*), các kết nối của người dùng *(node in social network*), các cộng đồng (*community extraction*) ... Một số nghiên cứu theo hướng này được tóm tắt trong Bảng 1.1.

Bảng 1.1: Tóm tắt về các nghiên cứu theo hướng tiếp cận user-centric

STT	Nghiên cứu	Nguồn thông tin	Biểu diễn chủ đề quan tâm	Mô hình và phương pháp phân tích	Cách đánh giá
1	A.Basma et al., ISPRS' 2016 [2]	Chú thích trên các phương tiện được gắn vị trí (geo-tagged)	Từ khóa về các địa danh nổi tiếng và được yêu thích	- Khuyến nghị quan tâm theo cá nhân và cho điểm - Sử dụng Location-Based SN (LBSN) score	Độ chính xác, Độ nhạy và F1
2	B. Jiang & Ying Sha, PCSE' 2015 [18]	Tri thức về các chủ đề quan tâm của người dùng	Khái niệm theo cây phân cấp dựa trên mạng tri thức	- Nắm bắt sự thay đổi chủ đề quan tâm của người dùng theo thời gian - Sử dụng cho điểm và xếp loại (score và ranked)	Cây phân cấp theo chủ đề Các chủ đề nóng (Hot topic)
3	Budak. C et al., MSR-TR' 2014 [28]	Các phát ngôn của người dùng trên mạng và thông tin trên mạng	Các tài khoản liên quan đến mạng xã hội, người dùng và hàng xóm lân cận	- Nắm bắt các tương tác phức tạp giữa các mối quan tâm khác nhau của người dùng - Sử dụng mô hình xác suất	Độ chính xác, Độ nhạy, Độ tin cậy
4	Parantapa Bhattacharya et al., ACM 2014 [21]	Các theo dõi chuyên gia, lựa chọn đặc trưng qua danh sách	Các khái niệm xã hội	- Mô hình chủ đề - Labeled LDA trên các tweets	- Xếp hạng và đánh giá của người
5	Itai Himelboim et al., 2017 [62]	Các cấu trúc mức mạng của người dùng	Các khái niệm và các luồng thông tin mẫu	Dán nhãn với 6 nhãn (Polarized Crowds, In-Group, Brand, Communities, Broadcast, Support	Xếp hạng và so sánh
6	Min Jiming et al., ENIR'12 [103]	Các tập tin thông tin cá nhân (profiles)	Hệ tri thức của từ điển Wikipedia	- Xây dựng danh mục thể loại và đánh chỉ mục ID từ danh mục lịch sử tìm kiếm của người dùng - Phân loại dùng k-means	- Xếp hạng
7	Noulas, A. et al., 2009 [108]	Thông tin di chuyển và vị trí của người dùng	Sử dụng từ khóa chỉ vị trí	- Khai thác tập dữ liệu khác nhau chứa thông tin về tương tác giữa những người trong hội nghị - Phân loại bằng giám sát và bán giám sát	Độ chính xác, Độ nhạy và F1

8	Palesta D. et al., KDD'2012 [111]	Thông tin về cộng đồng	Sử dụng khái niệm cấu trúc mạng động	- Tính các thành phần của đồ thị bằng thông tin người dùng và tính trọng số bằng độ đo Jaccard	- So sánh giữa CNM và INC
9	Tingting Wang et al., PAKDD'2013 [143]	Kết nối xã hội giữa các người dùng trong mạng	Sử dụng khái niệm kết nối xã hội dựa trên xu hướng (homophily)	- Sử dụng kỹ thuật random-walk dựa trên mô hình học củng cố kết hợp giữa văn bản và các thông tin liên kết	Độ chính xác
10	Zhiheng Xu et al., IEEE'2011 [148]	Các tweet đơn và mối quan hệ với các tác giả	Các khái niệm liên quan đến thông tin có liên quan đến tác giả	- Sử dụng mô hình chủ đề - Tác giả với tên là twitter-user	Cho điểm (score)

Với hướng tiếp cận *object-centric*, các nghiên cứu tập trung phát hiện quan tâm của người dùng bằng cách phân tích các đối tượng được tạo ra trong quá trình tương tác của người dùng trên các mạng xã hội như nội dung các bài đăng (*post, tweets…*), các thẻ đánh dấu (*tags*), các bình luận (*comment*), các thông tin trong các bài chia sẻ (*content of sharing*), các hành vi thích (*like*), thả cảm xúc (*emotions)* ... Hướng tiếp cận này được phân tích dựa trên nhiều đối tượng khác nhau như: phát hiện quan tâm của người dùng trên mạng xã hội thông qua việc khám phá các chủ đề tiềm ẩn dựa trên các thẻ đánh dấu như [145] phát hiện những quan tâm chung của người dùng được chia sẻ bởi các người dùng khác trong nhóm trên mạng xã hội bằng cách phân tích thẻ đánh dấu của người dùng trên mạng del.icio.us. Trong các hệ thống này, người dùng sử dụng thẻ đánh dấu như một nhãn mô tả để chú thích các nội dung mà họ quan tâm và chia sẻ các thẻ đánh dấu này với người dùng khác trong mạng, có thể hiểu rằng các thẻ đánh dấu này chính là các quan tâm của một người dùng trong mạng xã hội. Cải tiến từ thẻ đánh dấu, trong [125] phân tích và nghiên cứu quan tâm của người dùng dựa trên các thẻ đánh dấu để khám phá các cộng đồng có quan tâm chung trên mạng xã hội; phát hiện quan tâm dựa trên hành vi thích của [43] và [40], đã mô hình hóa hành vi thích của người dùng và dự đoán hành vi thích của người dùng trên mạng xã hội, bên cạnh đó nghiên cứu cũng đưa ra cách thức thống kê lượt thích của người dùng trên các mạng xã hội và mối liên quan giữa lượt thích và quan tâm của người dùng. Tuy nhiên, nghiên cứu dựa trên độ đo Jacard để ước lượng độ tương tự của hành vi thích với quan tâm của người dùng, không đưa ra mối quan tâm của người

dùng trong tổng thể hành vi của họ, đặc biệt trên thực tế, các hành vi thích có thể chưa thể hiện thực sự người dùng quan tâm đến bài viết hoặc chủ đề của bài viết mà chỉ là có mối quan hệ bạn bè, hoặc thực hiện nút thích kiểu "xã giao" thì việc nghiên cứu chỉ riêng hành vi thích có thể đưa ra kết quả chưa chính xác. Một số nghiên cứu theo hướng thứ hai được tóm tắt trong Bảng 1.2.

Bảng 1.2: Tóm tắt về các nghiên cứu theo hướng tiếp cận object-centric

TT	Nghiên cứu	Nguồn thông tin	Biểu diễn chủ đề quan tâm	Mô hình phân tích	Phương pháp đánh giá
1	Guy Ido et al., IBM Lab, WWW'13, 2013 [60]	Sử dụng các thông tin trên các ứng dụng của các mạng xã hội như các dữ liệu chuyên gia để phân tích	Sử dụng mạng tri thức với biểu đồ mối quan hệ giữa người dùng, các tài liệu và các thuật ngữ	Kỹ thuật đánh chỉ mục (indexed) trên tất cả các tài liệu được xây dựng theo siêu dữ liệu (metadata)	Khảo sát đánh giá cho điểm và phản hồi của người dùng
2	Al Kouz Akram, 2013 [14]	Các dữ liệu văn bản do người dùng tạo nên	Mạng ngữ nghĩa sử dụng từ điển Wikipedia	Dựa trên phân loại theo mô hình Bayesian Network	Độ chính xác, Độ nhạy và F1
3	Basit Shahzad et al., IDD' 2017 [23]	Thông tin của các sự kiện nổi bật trên mạng (famous events)	Sử dụng khái niệm của 30 thể loại theo Twitter lingo	Sử dung học máy với SVM với biểu diễn dữ liệu dựa trên Data Visualization	Độ chính xác, Độ nhạy và F1
4	Dhelim Sahraooui et al., KBS' 2020 [47]	Sử dụng 5 nhóm tính cách của người dùng (Big Five personality traits)	Biểu diễn dựa trên mô hình đồ thị, và túi từ BOW (bag-of-words)	- Sử dụng đối sánh để phát hiện meta-path - Độ đo tương tự	Độ chính xác, Độ nhạy và F1
5	Fattane Zarinkalam et al., '2019 [54]	Thông tin văn bản các tweets	Tập hợp từ khóa biểu diễn chủ đề	- Phương pháp phát hiện chủ đề	Độ chính xác, Độ nhạy, F1
6	Muhammad H. et al., IJCSIT' 2018 [63]	Thông tin văn bản các tweets	Tập hợp từ khóa biểu diễn chủ đề	- Đối sánh với tập các thể loại so sánh trên mạng xã hội Tweetter	Độ chính xác

7	Jain, Arti et al., ICIoTC'2018 [68]	Các văn bản đính kèm với ảnh được đăng lên mạng	Tập hợp từ khóa biểu diễn chủ đề	Sử dụng học máy với cấu trúc mảng dữ liệu và mạng nơron nhân tạo (Artificial Neural Network - ANN)	So sánh và đối chiếu
8	Kapanipathi P. et al., ESWC'2014 [72]	Thông tin văn bản các tweets	- Sử dụng từ điển Wikipedia	- Biểu diễn thông qua mô hình thực thể và cây phân cấp - Sử dụng học máy	Cho điểm và xếp loại
9	Kim J. et al., IJSE IA'2013 [77]	Các bài đăng của người dùng	- Tập hợp các từ khóa	- Trích chọn các danh từ và tính trọng số bằng cosine	So sánh
10	Kwan Hui Lim et al., WikiSym'13 [82]	Các bài đăng tweets của người dùng	- Các từ khóa là tên chủ đề	- Sử dụng Wikipedia để phân loại chủ động theo các chủ đề	- Đếm số người lựa chọn các chủ đề khi khảo sát
11	M. Shafik et al., ICICIS'2019 [97]	Các bài đăng tweets của người dùng	- Danh sách các từ khóa biểu diễn chủ đề	- Sử dụng mô hình chủ đề lai giữa các giải thuật học máy và các đặc trưng của người dùng	Độ chính xác, Độ nhạy và F1
12	Nori Nozomi et al., ICWSM' 2011 [107]	Thông tin chứa trong hành vi đánh dấu và yêu thích	Biểu diễn bằng mô hình đồ thị với các yếu tố phụ thuộc thời gian	- Sử dụng lý thuyết đồ thị hành động Action Graph và LDA	R-Precision
13	Pengtao Xie et al., AAAI'15 [114]	Các ảnh và văn bản đính kèm ảnh cá nhân	Tự định nghĩa dựa trên nội dung ảnh	Sử dụng mô hình không gian ảnh của người dùng (User Image Latent Space Model) với bốn mức: themes, semantic regions, visual words, pixels	Sử dụng các kết quả định tính để đánh giá
14	Fang-Yu Chao et al., APSIPA'2016 [50]	Các bài đăng trên mạng Pinterest (Pin of Pinterest)	Nhóm các từ khóa để biểu diễn chủ đề	- Sử dụng ma trận thuật ngữ của các tài liệu, các túi từ trực quan - Sử dụng (DLDA) rời rạc	Độ chính xác

15	Rytsarev Igor et al., ITNT'2016 [119]	Các bài đăng trên mạng Twitter	Nhóm các từ khóa để biểu diễn chủ đề	- Sử dụng Latent Dirichlet Allocation (LDA) và giải thuật k-means	Độ chính xác
16	Shuiqiao Yang et al., ' IEEE'2019 [121]	Các nội dung bài đăng	Thuật ngữ của văn bản ngắn (Term of short text)	Khám phá các thuật ngữ theo cách biểu diễn các chủ đề (TRTD) với trọng số của các nút và trọng số các cạnh theo đồ thị các từ (node-weighted, edge-weighted word graph - NEWG)	Độ chính xác
17	Sengkey C.H et al., ICNCC'16 [122]	Các thông tin được chia sẻ	Mô hình chủ đề (LDA) dựa trên Wikipedia	Sử dụng các hàm ContentInterest, SharingInterest, và InteractionInterest để phát hiện các chủ đề	Độ chính xác, Độ chính xác trung bình
18	Shangsong Liang et al., ACM'2017 [124]	Luồng các văn bản ngắn	Các thuật ngữ trong các văn bản ngắn (Term of short text)	- Sử dụng mô hình kết hợp của dynamic multinomial Dirichlet và mô hình UCT	Độ chính xác, Độ chuẩn
19	Sheng Bin et al., IJHIT'2016 [125]	Các thẻ đánh dấu	Các từ khóa	- Sử dụng không gian véc tơ và TF dựa trên mô hình TAUT (taking advantage of user tags)	Xếp hạng
20	Xin Li et al., [145]	Các thẻ đánh dấu	Các từ khóa và các khái niệm liên quan đến nội dung	- Sử dụng mô hình TAUT và (ISID - Internet Social Interest Discovery system,	Độ tương tự nội tại và độ tương tự của các chủ đề
21	V. Ranjith,V. Rajaram, IJRAR'2018 [138]	Thông tin quan tâm cá nhân, độ tương tự quan tâm giữa	Biểu diễn dựa trên quan điểm của tài liệu, các	- Dự đoạn dựa trên các xếp loại của các dịch vụ người dùng	Độ chính xác

		các cá nhân và ảnh hưởng của bản thân họ	câu, các thực thể.	- Phân tích bằng ngôn ngữ tự nhiên	
22	Yoad Lewenberg et al., IEEE'2015 [152]	Các cảm xúc (emotions)	Các từ khóa gồm sports, movies, technology & computing, politics, news, economics, science, arts, health & religion	- Sử dụng mô hình học máy theo sáu mức cảm xúc của Ekman's	Độ chính xác, và đánh giá mô hình dự đoán
23	Fattane Zarrinkalam et al., IR'2019 [54]	Thông tin các bài đăng trên Twitter	Các khái niệm biểu diễn ngữ nghĩa dựa trên Wikipedia	- Sử dụng mô hình phân cấp theo các thể loại của Wikipedia - Sử dụng phân tích dựa trên yếu tố thời gian	Độ chính xác, Độ nhạy và F1
24	Lin Gong et al., KDD'2018 [89]	Các bài đăng trên Twitter	Từ khóa tên chủ đề	- Sử dụng hệ thống nhúng và so sánh	Độ chính xác
25	Singh Arabzadeh et al., CIKM'2019 [129]	Thông tin các bài đăng trên Facebook	Nhóm các từ biểu diễn chủ đề	- Sử dụng phương pháp lan truyền	Độ chính xác

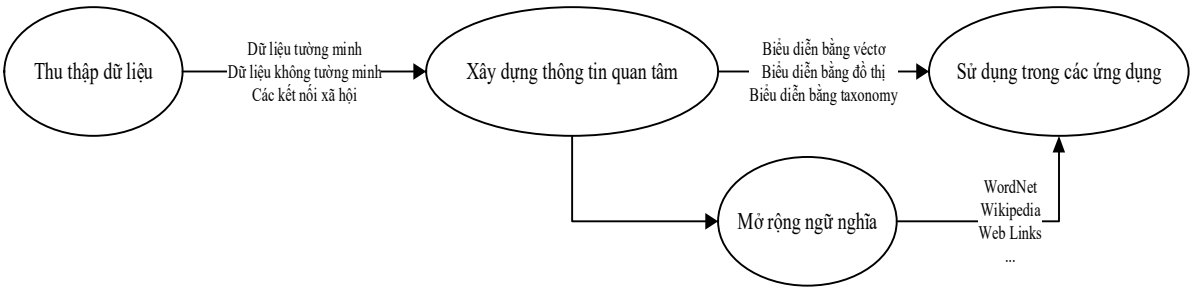
Dựa trên Bảng 1.2 có thể thấy rằng, các nghiên cứu về quan tâm của người dùng hiện nay tập trung nhiều vào việc phân tích thông tin trong các bài đăng, cảm xúc, thẻ đánh dấu, hoặc các thông tin chia sẻ được một cách riêng lẻ. Các kỹ thuật và phương pháp phân tích thường dùng là từ điển, xây dựng nhóm từ và các hệ thống từ vựng dựa trên các phép toán so sánh, các hệ thống học máy.

1.3.2. Các bước xây dựng hồ sơ quan tâm của người dùng

Theo [9] và [54] thì quá trình xây dựng hồ sơ quan tâm của người dùng (*user interest profile*) là quá trình thu thập, trích xuất và biểu diễn cho các chủ đề quan tâm

của người dùng. Quá trình này thường có ba giai đoạn: *Thu thập dữ liệu, xây dựng đặc trưng và đưa vào các ứng dụng* như Hình 1.3.

- Thu thập dữ liệu (*Collect data*): Là bước thu thập dữ liệu liên quan đến người dùng để lựa chọn và xây dựng các đặc trưng của người dùng đó. Các dữ liệu thu thập có thể là các bài đăng, các kết nối, các liên kết, các thông tin trong hồ sơ; có thể khai thác không tường minh như phân tích quan điểm, phân tích cảm xúc…
- Xây dựng thông tin quan tâm (*Interest profile construction*): Là bước phân tích, trích chọn và biểu diễn thông tin quan tâm của người dùng. Sử dụng các phương thức khác nhau để đưa ra mức độ quan tâm của người dùng đến các chủ đề. Kết quả là thông tin quan tâm của mỗi người dùng có thể được biểu diễn bằng véctơ, bằng đồ thị hoặc bằng cây phân cấp, ...
- Sử dụng trong các ứng dụng: Bước này là mô tả các ứng dụng có thể sử dụng kết quả của thông tin quan tâm của người dùng như các hệ thống khuyến nghị người dùng, các quảng cáo cá nhân hóa hoặc các hệ thống thương mại điện tử…



Hình 1.3: Quy trình xây dựng thông tin quan tâm của người dùng

Nguồn (Abdel. H Ahmad và Xu Yue (2013) [9] và Fattane Z et al. (2019) [54])

Quy trình xây dựng thông tin quan tâm của cá nhân người dùng thường được gọi tên dựa trên mô hình dùng để khám phá chủ đề quan tâm của người dùng, có thể hiểu là các kỹ thuật hay các phương thức sẽ được sử dụng trong phát hiện các chủ đề

quan tâm của người dùng, thường chia thành các mô hình dựa trên các đối tượng, các mô hình dựa trên các khai phá ẩn và các mô hình liên quan đến yếu tố thời gian.

- Mô hình dựa trên các đối tượng là các mô hình được nhìn thấy dễ dàng như các nội dung văn bản thể hiện rõ quan tâm của người dùng *(textual contents*) hoặc các mối kết nối rõ ràng liên quan đến các chủ đề quan tâm của người dùng (*user relationships*). Các phương pháp sử dụng để phân tích thường là phát hiện chủ đề (*topic detection*) hoặc xác định các thực thể có tên (*entity recognition identification*).
- Mô hình dựa trên các khai phá ẩn là các mô hình sử dụng các kỹ thuật khác nhau để phát hiện hay khám phá quan tâm của người dùng như các mối quan hệ bên trong của người dùng (*inter-user relations*) như các theo dõi, các bài đăng, các chia sẻ, các trích chọn được đề cập đến, hoặc các chủ đề bên trong các nội dung văn bản như phân cấp thể loại (*category hierarchical*), dựa trên đồ thị (*graph-based*). Các phương pháp sử dụng để phân tích và phát hiện thường là lọc cộng tác (*collaborative filtering*) hoặc khai phá mẫu tuần tự (*frequent pattern mining*) ...
- Mô hình có liên quan đến yếu tố thời gian là các mô hình dùng để phân tích trong quan tâm của người dùng trong các khoảng thời gian khác nhau (*user's interest change over time*). Thường các mô hình này có yếu tố ràng buộc về mặt thời gian như theo tuần, theo tháng, theo quý, theo năm hoặc theo một khoảng thời gian cố định trước.

1.3.3. Những nội dung đang nghiên cứu về mạng xã hội

Với lượng người dùng khổng lồ và số lượng thông tin cập nhật hàng ngày rất lớn, các mạng xã hội hiện được xem là thị trường mới đầy tiềm năng cho các tổ chức, doanh nghiệp, các nhà sản xuất cũng như các nhà nghiên cứu trong nước và trên thế giới. Tuy nhiên, việc thu thập dữ liệu xã hội từ các mạng xã hội càng ngày càng khó khăn do hạn chế của các nhà cung cấp dịch vụ và yêu cầu bảo mật thông tin cá nhân của người dùng. Thêm vào đó, dù có chung nhiều đặc điểm về mặt phương pháp luận

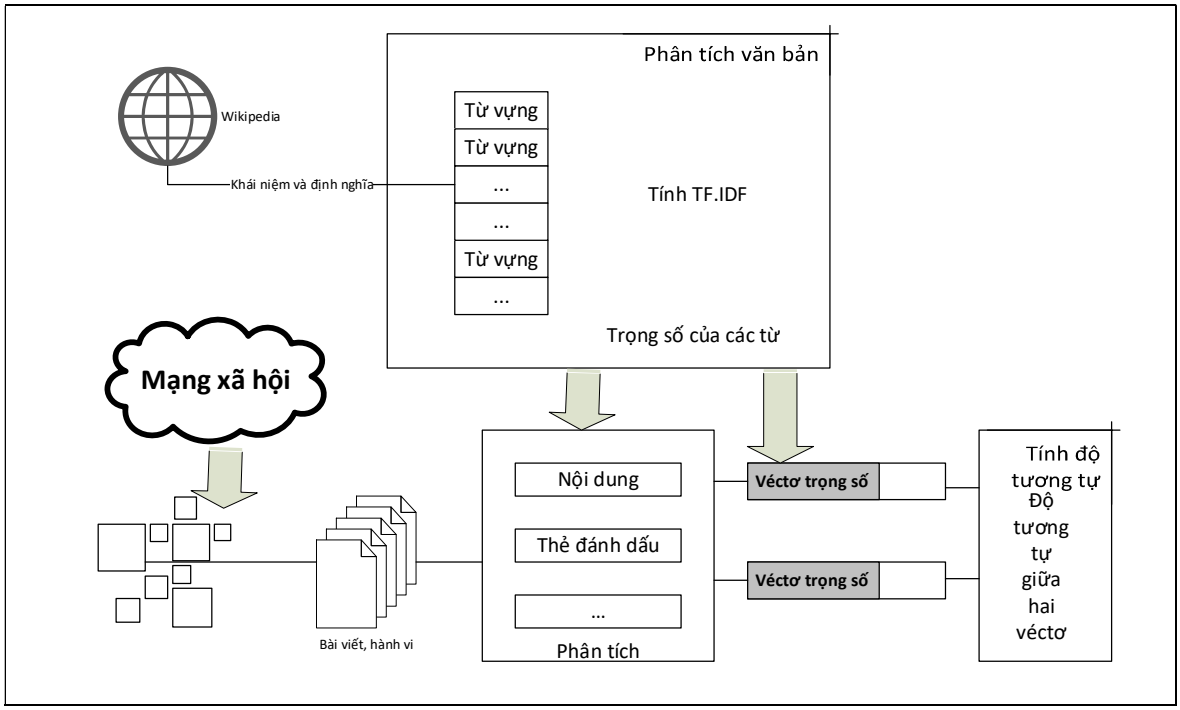
với các lĩnh vực nghiên cứu khác như xử lý ngôn ngữ tự nhiên, phân tích cấu trúc mạng, xử lý văn bản… Nhưng các mạng xã hội và các hành vi của người dùng trên các mạng xã hội là một lĩnh vực vẫn còn nhiều vấn đề chưa được nghiên cứu đầy đủ.

Việc phát hiện quan tâm của người dùng và phân nhóm người dùng trên các mạng xã hội là một bài toán không mới nhưng luôn có nhiều thách thức đối với các nhà nghiên cứu, do tính chất của mạng xã hội ngày càng phức tạp, mỗi nhà cung cấp dịch vụ lại muốn có những tính chất và đặc trưng riêng để thu hút người dùng tham gia vào các mạng xã hội cũng như sử dụng các dịch vụ mà họ cung cấp. Xây dựng một mô hình chung để phát hiện quan tâm nhằm phân nhóm người dùng hoặc xác định những người dùng tương tự nhau trên các mạng xã hội hiện nay đang là một bài toán mở, chưa có lời giải tốt cho các ứng dụng trên các mạng xã hội hiện nay. Do đó, việc xác định các quan tâm của người dùng trong các cộng đồng trực tuyến luôn luôn thu hút các nhà nghiên cứu quan tâm, mặc dầu có nhiều nghiên cứu đưa ra các kết quả khả quan nhưng việc ước lượng và phát hiện quan tâm tương tự của người dùng trên các mạng xã hội vẫn là một bài toán mở, còn nhiều thách thức với các nhà nghiên cứu.

1.3.4. Hướng nghiên cứu của luận án

Dựa trên kết quả phân tích các nghiên cứu đã có, luận án tiếp cận bài toán phát hiện quan tâm của người dùng dựa trên việc phân tích các hành vi của người dùng trên các mạng xã hội. Đây là hướng tiếp cận dựa trên phân tích dữ liệu do người dùng tạo ra trên các mạng xã hội hay các đối tượng được sinh ra trong quá trình hoạt động của người dùng trên các trang mạng xã hội. Hướng tiếp cận này sẽ loại bỏ được khó khăn mà hướng tiếp cận tập trung vào người dùng gặp phải như vấn đề phân tích dựa trên cấu trúc mạng, hoặc phân tích các liên kết của người dùng, phân tích dựa trên việc chia sẻ thông tin cá nhân của người dùng trên các trang mạng xã hội… Hình 1.4 mô tả hướng nghiên cứu của luận án với bài toán xây dựng hồ sơ thông tin quan tâm của người dùng gồm hai giai đoạn chính:

- ***Giai đoạn thu thập dữ liệu phân tích***: Thứ nhất là thu thập dữ liệu về người dùng bao gồm: nội dung bài đăng, các thẻ đánh dấu, các cảm xúc, các nhóm tham gia, các hành vi đăng bài, thích bài viết, bình luận trong bài viết, tham gia nhóm trên mạng xã hội. Thứ hai là thu thập và xây dựng các chủ đề để so sánh
- ***Giai đoạn xây dựng hồ sơ quan tâm của người dùng***: Giai đoạn này chia thành ba giai đoạn nhỏ, thứ nhất là tiền xử lý dữ liệu gồm tách từ, loại bỏ từ dừng, mở rộng ngữ nghĩa, tính trọng số. Giai đoạn thứ hai là biểu diễn các chủ đề, nội dung bài viết, mô hình bài viết và mô hình người dùng bằng véctơ trọng số. Giai đoạn cuối cùng là tính toán độ tương tự và độ tương quan để xác định mức độ quan tâm của người dùng đến các chủ đề và mối liên quan khác.

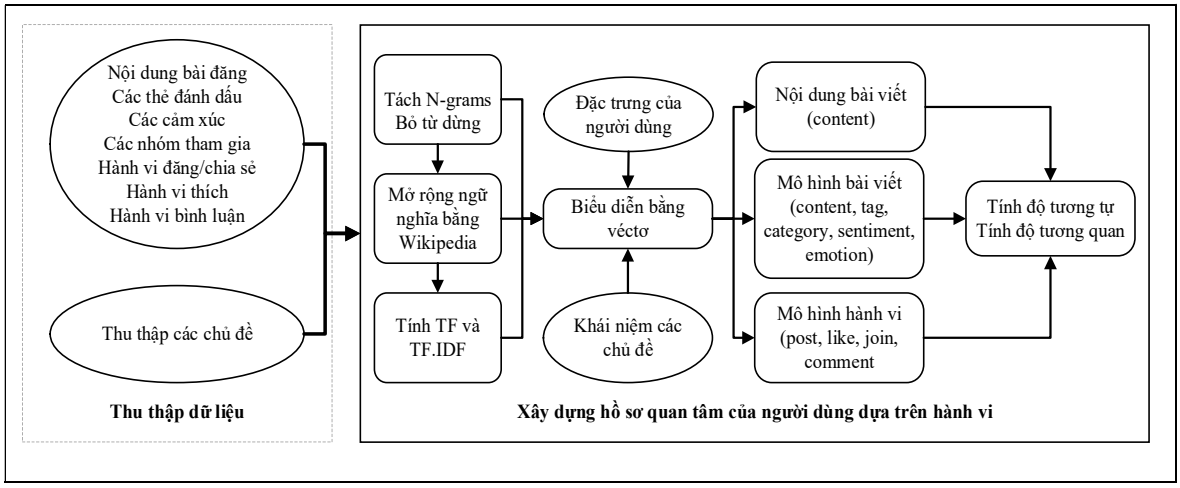


Hình 1.4: Hướng tiếp cận của luận án

Như vậy, để giải quyết bài toán phát hiện quan tâm của người dùng dựa trên các hành vi theo hướng tiếp cận *vào đối tượng*, nhiệm vụ của luận án là giải quyết ba các bài toán cụ thể như sau: Mô hình hóa bài viết dựa trên các đặc trưng để biểu diễn

người dùng trên các mạng xã hội, mục đích để mô hình hóa đối tượng nghiên cứu để có thể áp dụng trên nhiều mạng xã hội khác nhau; Thứ hai là mô hình hóa các hành vi để biểu diễn người dùng dựa trên các hành vi; Cuối cùng là xây dựng cách thức biểu diễn các bài viết và hành vi của người dùng để có thể ước lượng được dựa trên độ đo tương tự và tính tương quan của các đối tượng đó với các chủ đề trên các trang mạng xã hội.

Hướng tiếp cận của luận án chi tiết được minh họa trong Hình 1.5



Hình 1.5: Hướng tiếp cận của luận án chi tiết

Về dữ liệu nghiên cứu, luận án tập trung nghiên cứu, phân tích kiểu dữ liệu văn bản (text) và các biểu tượng (icon) trong các đặc trưng của bài viết. Các hành vi thể hiện trên các bài viết và các đặc trưng của nhóm hay cộng đồng trên mạng xã hội mà người dùng gia nhập cũng được nghiên cứu trong luận án. Các kiểu dữ liệu khác sẽ không xem xét trong luận án này;

Về công cụ lưu trữ dữ liệu, luận án sử dụng một số công cụ lưu trữ dữ liệu văn bản phổ biến bao gồm cấu trúc dữ liệu kiểu mảng, kiểu chuỗi, kiểu véctơ trọng số và ma trận trọng số để lưu trữ dữ liệu sơ cấp và thứ cấp trong xử lý dữ liệu;

Về kỹ thuật phân tích và xử lý dữ liệu, luận án sử dụng kỹ thuật N-gram để phân tách văn bản thành các từ khóa, sử dụng bộ từ dừng của Wikipedia kết hợp bộ từ dừng để loại bỏ từ dừng. Cuối cùng luận án xây dựng ma trận trọng số bằng TF-IDF để tính

toán và ước lượng. Ngoài ra, luận án có sử dụng một số thuật toán học có giám sát trong tính toán các giá trị cho một số đặc trưng của bài viết trước khi ước lượng trong mô hình bao gồm các thuật toán: Naive Bayes (NB), Multinomial Naive Bayes (MNB), Support Véctơ Machine (SVM), K-Nearest Neighbors (KNN or IBK), C4.5 (J48), Rotation Forest (RF); Các chủ đề sử dụng trong luận án được xây dựng dựa trên việc thống kê và trích chọn từ các trang báo điện tử có lượng truy cập lớn nhất ở Việt nam và một số trang nổi tiếng có thứ hạng cao trên thế giới, cách thức xác định các chủ đề này đã được nhiều nghiên cứu về mạng xã hội sử dụng [25] [26] [63] [86] [89] [100].

Về thực nghiệm và đánh giá, luận án tiến hành thực nghiệm và kiểm nghiệm các mô hình đề xuất trên ba bộ dữ liệu tự thu thập trên các mạng xã hội Facebook.com, Twitter.com và YouTube.com. Ngoài ra, để lựa chọn thuật toán phù hợp nhất với bộ dữ liệu văn bản trên mạng xã hội mà luận án thu thập được, hai bộ dữ liệu mẫu chuẩn là 20 NewsGroups [99] và SemEval-2017 [99] được lựa chọn để so sánh kết quả với bộ dữ liệu thực của luận án.

1.4. Xử lý dữ liệu văn bản ngắn trên mạng xã hội

Đối tượng nghiên cứu của luận án là các bài viết và hành vi của người dùng trên các mạng xã hội có các đặc trưng là văn bản nên các bài toán cần giải quyết trong luận án đều liên quan đến vấn đề xử lý dữ liệu văn bản. Do đó, trong mục này, luận án trình bày một số kiến thức nền tảng về xử lý văn bản cũng như các phương pháp ước lượng và tính toán được sử dụng trên kiểu dữ liệu văn bản trong luận án. Nội dung của mục này bao gồm các kỹ thuật dùng để biểu diễn và phân tích văn bản, các phương pháp tính toán và ước lượng độ tương tự giữa các văn bản, cuối cùng là các phương pháp đánh giá kết quả thực nghiệm khi phân tích và xử lý dữ liệu văn bản trên các trang mạng xã hội.

1.4.1. Biểu diễn và tiền xử lý văn bản

Trong các nghiên cứu liên quan đến dữ liệu văn bản, cách thức biểu diễn văn bản là một trong các bước quan trọng làm tiền đề cho việc ứng dụng các phương pháp và kỹ thuật để khai thác, phân tích và xử lý dữ liệu văn bản. Tùy thuộc vào từng bài toán, từng thuật toán ứng dụng khác nhau mà các nghiên cứu lựa chọn mô hình biểu diễn dữ liệu văn bản phù hợp, các mô hình biểu diễn văn bản phổ biến hiện nay gồm mô hình logic, mô hình véctơ, mô hình đồ thị ...

Mô hình logic: Theo các nghiên cứu [15] [40] [45] [48] [71] thì dữ liệu văn bản biểu diễn theo mô hình logic thông qua các danh sách các từ và chỉ số của chúng trong văn bản. Chỉ số của các từ loại trong văn bản thường là tần suất xuất hiện của từ loại hoặc trọng số của từ loại, ... Các từ loại được đánh chỉ số trong văn bản gốc đều là các từ có nghĩa. Mỗi văn bản được đánh chỉ số theo qui tắc liệt kê các từ có nghĩa trong văn bản theo thứ tự xuất hiện trong văn bản hoặc được sắp xếp theo thứ tự từ điển. Nói cách khác, cách biểu diễn văn bản theo mô hình logic là đánh chỉ số index cho các từ xuất hiện trong văn bản cần xem xét.

Mô hình véctơ: Theo các nghiên cứu [15] [40] [45] [48] [71] thì mô hình véctơ là một trong những mô hình đơn giản và thường được sử dụng trong phần lớn các bài toán xử lý dữ liệu văn bản. Theo mô hình này, mỗi văn bản được biểu diễn thành một véctơ, mỗi thành phần của véctơ là một từ khóa trong tập văn bản gốc và được gán một giá trị trọng số, trọng số có thể được xác định bằng tần suất xuất hiện của từ trong văn bản TF.

Phát biểu của mô hình như sau: Mỗi văn bản *d* được biểu diễn dưới dạng một véctơ: $\overrightarrow{\boldsymbol{v}} = (v_1, v_2, \dots, v_n)$ (gọi là véctơ đặc trưng cho văn bản *d*), với n là số lượng các từ khóa đặc trưng hay số chiều của véctơ (thường là tổng số từ khóa), v_i là trọng số của từ khóa thứ i hay đặc trưng thứ i (với $1 \leq i \leq n$). Trọng số của đặc trưng v_i có thể được tính dựa theo tần số xuất hiện của từ khóa trong văn bản *d.* Nếu xét một tập hợp các văn bản D thì sẽ có một tập các véctơ, khi đó được gọi là ma trận trọng số $w = \{w_{ij}\}$ được xác định dựa trên tần số xuất hiện của từ tập hợp các văn bản D.

Mô hình đồ thị: Mô hình được John F. Sowa đưa ra lần đầu tiên vào năm 1976, còn được gọi là mô hình đồ thị khái niệm (Conceptual Graphs CGs) [39] [48] [98], mô hình đồ thị hiện nay được khá nhiều ứng dụng sử dụng vào các bài toán xử lý văn bản. Trong mô hình đồ thị, mỗi đồ thị là một văn bản, đỉnh của đồ thị có thể là câu, hoặc từ, hoặc kết hợp câu và từ, cạnh nối giữa các đỉnh là vô hướng hoặc có hướng, thể hiện mối quan hệ trong đồ thị. Ngoài ra, có thể sử dụng đồ thị chứa trọng số để biểu diễn văn bản, khi đó, nhãn của đỉnh thường là tần số xuất hiện của từ xuất hiện trong câu, trong văn bản, còn nhãn của cạnh là tên liên kết khái niệm giữa hai đỉnh, hay tần số xuất hiện chung của hai đỉnh trong một phạm vi nào đó, hay tên vùng mà đỉnh xuất hiện trong văn bản.

Trong luận án sử dụng mô hình véctơ kết hợp trọng số để biểu diễn dữ liệu khi phân tích và xây dựng các bộ dữ liệu mẫu để kiểm nghiệm các mô hình

1.4.2. Véctơ hóa dựa trên TF.IDF

Cách tính trọng số đã được nhiều nghiên cứu đề cập đến và phân tích như nghiên cứu [48] [63] và [83] đều giới thiệu chi tiết về cách tính trọng số cho các từ, thuật ngữ trong xử lý dữ liệu văn bản. Trọng số của đặc trưng v_i trong véctơ có thể được tính dựa theo tần số xuất hiện của thuật ngữ trong văn bản *d.* Nếu xét một tập hợp các văn bản D thì sẽ có một tập các véctơ, khi đó được gọi là ma trận trọng số $w = \{w_{ij}\}$ được xác định dựa trên tần số xuất hiện của từ tập hợp các văn bản D. Có một số cách tính trọng số cho các thuật ngữ trong văn bản *d* như sau:

***Tính trọng số dựa trên tần số từ khóa (Term Frequency*):** Với văn bản d_i trong D, sau khi được phân tách d_i thành các từ khóa, được biểu diễn dưới dạng véctơ $\vec{v} = (w_1, w_2, .. w_m)$. Gọi f_{ij} là tần suất xuất hiện của từ khóa w_i trong văn bản d_j, khi đó giá trị của các thành phần trong ma trận trọng số w_{ij} được tính bằng một trong ba cách thức đề xuất ở công thức (1.1).

$$w_{ij} = \begin{cases} f_{ij} \quad \text{hoặc,} \\ 1 + log(f_{ij}) \text{ hoặc,} \\ \sqrt{f_{ij}} \end{cases}$$

(1.1)

Nếu số lần xuất hiện từ khóa w_i trong văn bản d_j càng nhiều thì văn bản d_j càng phụ thuộc vào từ khóa w_i, hay nói cách khác từ khóa w_i mang nhiều thông tin trong văn bản d_j.

Tính trọng số dựa trên tần suất nghịch đảo của văn bản (Inverse Document Frequency): Với văn bản d_i trong D, sau khi được phân tách d_i thành các từ khóa, được biểu diễn dưới dạng véctơ $\vec{v} = (w_1, w_2, .. w_m)$. Gọi tf_{ij} là tần suất xuất hiện của từ khóa w_i trong văn bản d_j tính theo công thức (1.1). Khi đó, ma trận trọng số w_{ij} của văn bản d_j được tính theo tần suất nghịch đảo của từ khóa w_i trong văn bản d_j theo công thức (1.2) như sau:

$$w_{ij} = \begin{cases} log(\frac{N}{1 + df_i}) \, nếu \, tf_{ij} \geq 1 \\ 0 \, nếu \, tf_{ij} = 0 \end{cases}$$

(1.2)

Trong đó *N* là số lượng văn bản có trong *D* và df_i là số lượng các văn bản có chứa từ khóa w_i. Trong công thức này, trọng số w_{ij} được tính dựa trên độ quan trọng của từ khóa w_i trong văn bản d_j. Nếu w_i xuất hiện trong càng ít văn bản của *D*, thì khi nó xuất hiện trong d_j, trọng số của nó đối với d_j càng lớn (do tính nghịch đảo của hàm log), tức là hàm lượng thông tin trong nó càng lớn. Nói cách khác w_i là từ quan trọng để phân biệt d_j với các văn bản khác trong *D*.

Tính bằng TF × IDF. Đây là phương pháp kết hợp của hai phương pháp TF và IDF đã tính theo (1.1) và (1.2). Gọi tf_{ij} là tần suất xuất hiện của w_i trong văn bản d_j

được tính theo công thức (1.2). Trọng số w_{ij} được tính bằng tần số xuất hiện của từ khóa w_i trong văn bản d_j và độ hiếm của từ khóa w_i trong tập văn bản *D* như sau:

$$w_{ij} = \begin{cases} TF * IDF = (1 + log(f_{ij})) * log\left(\frac{N}{1 + df_i}\right) nếu\ tf_{ij} \geq 1 \\ 0\ nếu\ tf_{ij} = 0 \end{cases} \quad (1.3)$$

Trong đó w_{ij} là trọng số của w_i trong văn bản d_j, f_{ij} (term frequency) là tần suất xuất hiện của của w_i trong văn bản d_j, f_{ij} càng cao thì từ đó càng miêu tả tốt nội dung văn bản. df_i (document frequency) là số lượng các văn bản trong *D* có chứa w_i.

1.5. Kết luận

Trong những năm gần đây, các phương tiện truyền thông xã hội đặc biệt là các mạng xã hội ngày càng hiện diện sâu rộng trong nhiều lĩnh vực sản xuất kinh doanh của các tổ chức, doanh nghiệp làm thúc đẩy sự quan tâm nghiên cứu nhiều ứng dụng khác nhau như ứng dụng phân tích dữ liệu người dùng trong Marketing, ứng dụng các hệ tư vấn sản phẩm, ứng dụng trong chăm sóc khách hàng, ứng dụng trong tìm kiếm khách hàng tiềm năng… trong đó bài toán phát hiện các chủ đề quan tâm của người dùng đóng vai trò quan trọng trong phân tích dữ liệu của người dùng trên các mạng xã hội. Trong chương một, luận án đã trình bày sơ lược một số khái niệm liên quan đến mạng xã hội và những vấn đề nghiên cứu liên quan đến bài toán phát hiện chủ đề quan tâm của người dùng trên các mạng xã hội. Ba vấn đề chính được quan tâm trong bài toán phát hiện chủ đề quan tâm của người dùng trên mạng xã hội bao gồm: đối tượng dùng để phân tích, nghiên cứu; công cụ dùng để biểu diễn đối tượng và phương thức dùng để tính toán, ước lượng. Đây cũng chính là ba nội dung chính luận án sẽ trình bày và đề xuất trong các chương tiếp theo.

CHƯƠNG 2: MÔ HÌNH VÀ QUAN TÂM CỦA NGƯỜI DÙNG THEO NỘI DUNG BÀI VIẾT

Trong chương này, luận án trình bày khái niệm bài viết và nội dung bài viết, cách thức xử lý, biểu diễn bài viết dựa trên nội dung và ước lượng với các chủ đề trên mạng xã hội. Các chủ đề được luận án xây dựng từ danh sách các trang tin tức điện tử phổ biến ở Việt Nam và trên thế giới. Nội dung bài viết và các chủ đề được biểu diễn dưới dạng các véctơ trọng số. Cuối cùng, luận án thực hiện ước lượng độ tương tự giữa véctơ của nội dung bài viết và véctơ của chủ đề để xác định mức độ quan tâm của người dùng.

2.1. MÔ HÌNH NGƯỜI DÙNG THEO NỘI DUNG BÀI VIẾT

Trong mục này, luận án trình bày nội dung bài viết của người dùng trên mạng xã hội và cách biểu diễn bài viết của người dùng bằng véctơ trọng số tính bằng TF.IDF. Dựa trên cách biểu diễn bài viết này, luận án đưa ra độ đo tương tự của hai người dùng trên mạng xã hội theo nội dung bài viết dựa trên độ đo tương tự giữa hai véctơ biểu diễn người dùng theo nội dung bài viết.

2.1.1. Biểu diễn véctơ bài viết bằng TF.IDF

a. Bài viết trên mạng xã hội

Bài viết của người dùng trên các mạng 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 mạng xã hội có thể là một video clip, một hoặc một số bức ảnh, một văn bản, hoặc một sự kết hợp những thành phần này.

Với bài viết trên mạng xã hội Twitter.com trong Hình 2.1, bài viết này không có tiêu đề, nội dung là đoạn văn bản *"When I was growing up, my mom taught me that I could be anything I wanted and supported me along every step of my path. I can still count on her for help, whether I'm wrestling with a tough issue or just need somebody to talk to. Happy #MotherDay Mom"*, trong đoạn văn bản của bài viết thì ngoài nội dung còn có

đánh dấu *@melindagates*, *#MotherDay*, cảm xúc thể hiện qua hình bông hoa (happy – hạnh phúc), quan điểm và thể loại bị ẩn trong nội dung của bài viết. Bên cạnh đó, còn có ngày tháng đăng bài, số lượng người thích là 1.8 nghìn người, có 169 chia sẻ và có 69 bình luận.

Hình 2.1: Bài viết trên mạng xã hội Twitter.com và Facebook.com

Trong Hình 2.2 là hai bài viết được người dùng chia sẻ lại, bài viết thứ nhất được chia sẻ từ người dùng khác nhưng cùng trên mạng xã hội Facebook nên có đầy đủ các dữ liệu của một bài viết trên trạng mạng Facebook, trong khi bài viết thứ hai được người dùng chia sẻ lại từ một trang tin tức điện tử (www.vnexpress.net). Các bài viết được chia sẻ lại từ các nguồn khác như bài viết thứ hai thường sẽ có hai phần: Phần thứ nhất là dữ liệu sinh ra bởi người dùng và phần thứ hai là toàn bộ dữ liệu của bài viết từ nguồn được chia sẻ. Phần thứ nhất có thể có hoặc không, nếu có thường là quan điểm hay thái độ của người dùng đối với bài viết gốc, phần thứ hai thường được giữ nguyên tình trạng như nó được tạo ra, trong Hình 2.2 thì bài viết này có phần thứ nhất là đoạn văn bản: "*Công an quận 4 đang thụ lý giải quyết theo quy trình xử lý tin tố giác hành vi ...*" và phần thứ hai là bài viết trên trang tin tức vnexpress.net có tiêu đề là "*Lời khai của bé gái bị xâm hại trong thang máy ở Sài Gòn*" kèm theo toàn bộ nội dung của bài viết.

Luận án xem các bài viết trong mục này là phần nội dung văn bản mà người dùng đã đăng lên hoặc đã tạo ra kèm với nội dung văn bản được chia sẻ lại từ người khác hoặc phương tiện truyền thông xã hội khác.

Hình 2.2: Bài viết chia sẻ lại từ nguồn khác và người dùng khác

(Nguồn: Facebook.com)

Các bài viết chứa nội dung là văn bản của người dùng trên các mạng xã hội được coi là dữ liệu xã hội, do đó, chúng thường không theo quy chuẩn và có độ dài khác nhau. Vì vậy, việc xử lý dữ liệu bài viết chứa văn bản trên các mạng xã hội được xếp vào nhóm dữ liệu văn bản ngắn (short-text) và sử dụng các kỹ thuật áp dụng trên dữ liệu văn bản ngắn để thực hiện.

Bảng 2.1: Ví dụ về văn bản ngắn trên mạng xã hội

Bài đăng (Post)	Em rất thích ăn bánh xèo và Hủ tiếu miền nam nhưng chưa tìm ra được ở Hà Nội có quán nào ăn ưng ý. Bánh xèo mấy quán em ăn thấy toàn nhiều mỡ, bánh dày quá ko giòn tan. Hủ tiếu thì chưa thấy có quán nào cả....chắc tại e chưa tìm ra. Mọi người ai có biết chỗ nào ngon giới thiệu em với ạ chứ thèm quá mà ko biết ăn ở đâu 😂 😂 😂
UserA	Bánh xèo 2 cô chú thân thiện số nhà 57 ngõ 68 triều khúc. Đây là 3 suất 60k. Bánh đầy đặn, nước chấm đậm đà. Rau, nộm và dưa chuột thì ăn xả láng. <3
UserB	bánh xèo 6 phước ở 74 cầu đất (nếu mình nhớ k nhầm mua chắc chắn ở cầu đất) ngon cực kì phục vụ nhiệt tình nhé
UserC	E hay ăn ở đội cấn rất ngon. Mà e k nhớ là quán nào. View trong ngõ, khá ok

Bài viết của người dùng được luận án xử lý như là dữ liệu văn bản ngắn, vì vậy, trong mục này, luận án trình bày phương pháp biểu diễn dữ liệu văn bản ngắn bằng

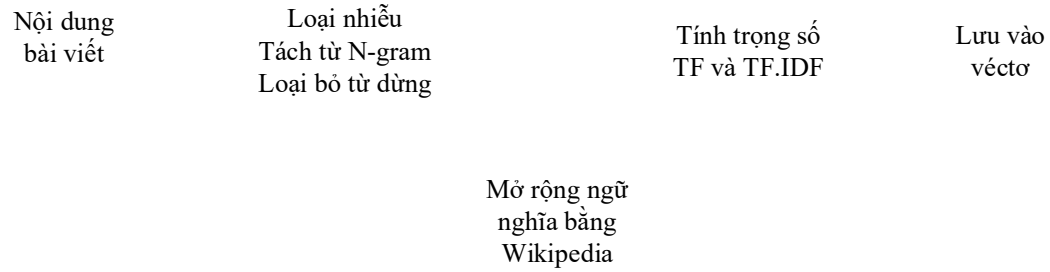
véctơ trọng số để phục vụ cho các tính toán của các mô hình sẽ đề xuất ở các phần tiếp theo.

b. Xử lý văn bản ngắn

Trong mục này, luận án trình bày một số kỹ thuật xử lý dữ liệu với văn bản ngắn được sử dụng trong quá trình tiền xử lý cho các bộ mẫu dữ liệu thực nghiệm. Ví dụ về một số bài viết trên mạng xã hội trong Bảng 2.1 có thể thấy rõ sự không chuẩn về cú pháp và văn phạm của ngôn ngữ, trong bài viết này, văn bản được viết theo kiểu của văn nói trong đời sống với nhiều từ viết tắt *"không"* viết thành *"ko"* hoặc *"k"; "60 nghìn"* viết thành *"60k"*; *"em"* viết thành *"e"*; kết hợp các biểu tượng "😂", các từ viết tắt "*6 phước*", các từ lóng "*xả láng"*, tiếng nước ngoài như *"view"*...

Theo [33] [53] [80] [119] [130] thì phương pháp xử lý cho dữ liệu văn bản ngắn gồm hai bước chính: Thứ nhất, làm sạch và tách từ theo N-gram; Thứ hai, mở rộng ngữ nghĩa (nếu cần), loại bỏ từ dừng và tính trọng số của từ. Đối với các văn bản quá ngắn, các nghiên cứu thường mở rộng ngữ nghĩa bằng cách sử dụng từ điển hoặc mạng từ, mục đích nhằm mở rộng thêm số lượng từ, thuật ngữ liên quan đến các từ, thuật ngữ gốc để thu được các kết quả có độ chính xác cao hơn.

Trong mục 1.1.2 của chương một đã giới thiệu các nội dung văn bản của các bài viết là dữ liệu xã hội và được coi là các văn bản ngắn [53] [80] [121] [130]. Vì vậy, luận án thực hiện xử lý nội dung các bài viết theo hướng xử lý văn bản ngắn, bao gồm bỏ nhiễu, tách từ, loại bỏ từ dừng, mở rộng ngữ nghĩa, tính trọng số minh họa trong Hình 2.3



Hình 2.3: Quy trình xử lý nội dung bài viết của luận án

Các bước tiền xử lý dữ liệu văn bản của bài viết được luận án thực hiện qua các bước sau*: làm sạch dữ liệu, tách bài viết thành các từ và thuật ngữ, chuẩn hóa danh sách từ, loại bỏ từ dừng, mở rộng danh sách từ theo Wikipedia* đối với các bài viết quá ngắn, cuối cùng là xây dựng véctơ cho bài viết.

Làm sạch dữ liệu: Do các bài viết trên mạng xã hội thông thường không chỉ chứa dữ liệu văn bản mà còn chứa các ký pháp khác như các biểu tượng, các dấu ngoặc, các dấu chấm than… ví dụ như trong Bảng 2.2 sau đây:

Bảng 2.2: Danh sách các biểu tượng, dấu câu, ký tự đặc biệt được loại bỏ

Kiểu	Ví dụ
Các biểu tượng	☺, ☹,
Các dấu câu	. ?, :, `, …
Các ký tự đặc biệt	=, >=, $, !, #, ?
…	…

Tách từ sử dụng N-gram: Sau khi loại bỏ các biểu tượng, các ký pháp không chuẩn của bước làm sạch dữ liệu văn bản của các bài viết, luận án thực hiện việc tách từ sử dụng kỹ thuật N-gram [11] [33] [53].

Chuẩn hóa từ vựng: Do các bài viết có chứa nhiều từ vựng không chuẩn cú pháp, ví dụ như: *"aaaaa", "ko", "kkkk"*... Vì vậy, luận án thực hiện xây dựng một bộ gồm các từ vựng và danh sách các từ vựng thay thế, ví dụ: *"aaaaa"* được thay bằng *"a", "ko*" hoặc "*k*" được thay thế bằng "*không*"… Một số từ thường dùng không chuẩn mực trên mạng xã hội được thay thế như trong Bảng 2.3.

Bảng 2.3: Ví dụ làm sạch dữ liệu với văn bản thay thế

Từ trong bài	Từ thay thế
a, aa, aaaaa, a!!!	a
k, ko, o, koooo, kko, koooo	không
và, vàooo, vvaooooooo	vào
hiii, hii, hiiiiiiiiii	hi
…	…

Loại bỏ từ dừng: Từ dừng là những từ không giúp ích nhiều trong việc phân biệt nội dung của các dữ liệu văn bản, danh sách từ dừng này được luận án tham khảo và kết hợp trên các website:

https://xltiengviet.fandom.com/wiki/Danh_s%C3%A1ch_stop_word và danh sách từ dừng Tiếng Việt do Wikipedia đề xuất. Danh sách các từ dừng được sử dụng trong luận án dược liệt kê trong Phụ lục C

Việc tách từ, chuẩn hóa từ và loại bỏ từ dừng trong tiếng Việt và tiếng Anh có sự khác nhau khi lựa chọn N, trong quá trình luận án tiến hành thực nghiệm với cùng giá trị N nhưng tỉ lệ các từ Tiếng Việt và các từ Tiếng Anh sau khi tách có sự khác nhau và được liệt kê trong Bảng 2.3. Tỉ lệ này được luận án thực hiện bằng cách tách toàn bộ 2000 bài viết với N=1, 2, 3 và 4 tương ứng, được tổng số từ sau khi tách nguyên bản *Word_original*, sau đó luận án 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*. Tỉ lệ thu được tính bằng phần trăm của *Word_key /Word_original.*

Bảng 2.4: Bảng so sánh tỉ lệ các từ có trong từ điển khi tách từ

N-gram	Tỉ lệ Tiếng Việt 2000 bài viết trên Facebook	Tỉ lệ Tiếng Anh 2000 bài viết trên Twitter
N=1	47.12%	86.25%
N=2	83.06%	63.25%
N=3	58.25%	38.17%
N=4	32.17%	18.25%

Kết quả này có thể cho thấy đối với tiếng Việt N=2 là cho kết quả tốt nhất nhưng với tiếng Anh thì N=1 là cho kết quả tốt nhất. Vì vậy, trong thực nghiệm đối với tiếng Việt luận án thực hiện tách từ với N=2 và N=3.

Mở rộng ngữ nghĩa cho các bài viết bằng từ điển: Luận án sử dụng từ điển trực tuyến Wikipedia Tiếng Việt theo *(https://vi.wikipedia.org/wiki/Wikipedia_tiếng_ Việt)* làm cơ sở để phân tích và mở rộng ngữ nghĩa cho bài viết của người dùng trên mạng xã hội, ngoài ra luận án có sử

dụng thêm từ điển Tiếng Việt trực truyến (*https://dict.laban.vn/*) để so sánh các kết quả. Quy trình thêm từ vựng bằng mở rộng ngữ nghĩa cho các bài viết được luận án 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, luận án 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. Trong các phân tích và xử lý dữ liệu văn bản ngắn trên mạng xã hội, từ điển Wikipedia được khá nhiều nghiên cứu sử dụng trong quá trình làm giàu thêm từ vựng cho các khái niệm. Cụ thể được thực hiện theo Thuật toán 2.1 trong Bảng 2.5

Bảng 2.5: Thuật toán 2.1 (Mở rộng ngữ nghĩa theo Wikipedia)

```
                Thuật toán mở rộng từ vựng theo Wikipedia, openWordWiki(x,y)
Input:          Danh sách từ, thuật ngữ của bài viết ngắn x
Output:         Danh sách từ, thuật ngữ đã mở rộng của bài viết
Thực hiện:      W← ∅ // Khởi tạo
                For i=1 to all(x)
                      Begin
                      W[i]← W[i] ∪ getDefineWiki(x[i]) ;//Lấy định nghĩa
                       For j←2 to 4 do //Tách từ cho định nghĩa
                            y ← separateNgram(W[i],j);
                       End For
                         y ← y ∪ removeStopWord(y);
                EndFor
                Return
```

Trong thuật toán 2.1 có dùng đến hàm *getDefineWiki()* được sử dụng dựa trên API với lệnh:

```
$jsoncontent =
file_get_contents("https://vi.wikipedia.org/w/api.php?format=json&action=
query&titles=".urlencode($keyword)."&prop=extracts&exintro=&explaintext=
&callback=?&redirects=1");
```

Sau khi lấy nội dung định nghĩa của các từ xong được lưu vào tập tin định nghĩa gốc theo mảng, mỗi giá trị trong mảng là một định nghĩa của từ trong các bài viết.

Cách để lọc nhiễu và chuẩn hóa từ vựng của các bài viết được so sánh dựa trên các từ chuẩn lấy trong từ điển trực tuyến *http://dict.laban.vn* theo hàm fetchUrl().

fetchUrl("http://dict.laban.vn/find?query=".urlencode($keyword), true);

Hàm này trả về các từ có trong từ điển và luận án sử dụng các từ này để thay thế các từ không chuẩn mực trong bài viết.

Các nghiên cứu [4] [10] [72] [82] [103] đã chỉ ra rằng từ điển Wikipedia rất phù hợp với việc mở rộng từ vựng cho các bài viết ngắn. Wikipedia Tiếng Việt hiện có hơn 1.225.000 định nghĩa và khái niệm, là một trong 10 ngôn ngữ có nhiều hơn 1 triệu định nghĩa và khái niệm trên mạng Internet. Vì vậy, việc sử dụng định nghĩa từ Wikipedia có thể đáp ứng được đầy đủ các từ vựng có nghĩa của Tiếng Việt.

Phương pháp mở rộng ngữ nghĩa của luận án được dựa trên mở rộng từ trong thuật toán mà [120] đề xuất trên các từ Tiếng Anh được luận án mở rộng theo các từ và thuật ngữ và thực hiện trên từ Tiếng Việt.

Các bước mở rộng ngữ nghĩa được thực hiện bằng việc tìm định nghĩa của các từ, tải xuống đánh chỉ mục theo các từ và lưu vào cơ sở dữ liệu. Các khái niệm sau đó được sử dụng lại cho các bước sau bao gồm cả mở rộng cho các chủ đề, các thẻ đánh dấu, thể loại, tên nhóm, mô tả nhóm cho các chương sau của luận án. Cách thực hiện được trình bày trong Bảng 2.5, một bài viết ví dụ được minh họa trong Bảng 2.6

Bảng 2.6: Ví dụ về mở rộng ngữ nghĩa cho bài viết

Bài viết	Danh sách từ gốc	Danh sách từ sau khi mở rộng
Lí do tôi chọn ngành thực phẩm sạch: được nhà cung cấp đích thân tới nhà thể hiện tài chế biến!!! Wow! Tuyệt vời! Hihi :D	Lí do, chọn, ngành thực phẩm, cung cấp, nhà cung cấp, chế biến, thể hiện, tuyệt vời	Lí do, giải thích, điều để giải thích, đem lại, lựa chọn, ngành thực phẩm, thức ăn, cung cấp, mang lại, đưa đến, chế biến, biến đổi, thay thế, thể hiện, biểu hiện, tuyệt vời, rất hay, rất đẹp, …

Xây dựng véctơ cho bài viết: Luận án xây dựng véctơ cho bài viết dựa trên danh sách từ và thuật ngữ thu được của mỗi bài viết. Trọng số của các từ và thuật ngữ được tính toán theo TF.IDF. Sau đó luận án thực hiện các bước sau: Với mỗi bài viết $e_i \in$

$E|E \in N$ sau khi phân tích và xử lý, mỗi bài viết thu được hai danh sách: một danh sách từ, thuật ngữ của bài viết và một danh sách trọng số của các từ, thuật ngữ trong không gian các bài viết.

Véctơ trọng số của bài viết được tính theo Định nghĩa 2.1 và biểu diễn như Định nghĩa 2.5. Giả sử có bài viết: "*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*", trên mạng xã hội N, khi đó, bài viết trên được tiền xử lý và phân tích theo năm bước đã trình bày, thực hiện tính TF.IDF thì có kết quả như trong Bảng 2.7 và véctơ của biểu diễn của bài viết đó được biểu diễn như sau: $\mathbf{e}_{\mathrm{i}} = \big((\text{âm nhạc}; 0.3157), \dots, (\text{ý tưởng}; 0.4024)\big)$

Bảng 2.7: 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); ...

c. Biểu diễn văn bản bằng véctơ trọng số

Trong luận án, trọng số của các từ, thuật ngữ trong các không gian dữ liệu văn bản xem xét được tính theo Định nghĩa 2.1 dưới đây.

Định nghĩa 2.1:

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

$$w_d = tf(d, D_i) \times idf(d, \boldsymbol{\mathcal{D}}) \tag{2.1}$$

Trong đó, $tf(d, D_i)$ *là số lần xuất hiện của thuật ngữ* $d\ trong\ D_i$ *và* $idf(d, \boldsymbol{\mathcal{D}})$ *được tính bằng*

$$idf(d, \boldsymbol{\mathcal{D}}) = log\ (\frac{\|\boldsymbol{\mathcal{D}}\|}{1 + \|\{\ D_i | d \in\ D_i\}\|}) \tag{2.2}$$

Như vậy, sau khi tính được trọng số của các thuật ngữ thì mỗi văn bản $D_i \in \boldsymbol{\mathcal{D}}$ được biểu diễn bởi một véctơ trọng số, và 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 2.2:

Cho một tập các văn bản $\boldsymbol{\mathcal{D}} = \{D_1, D_2, \dots, D_p\}$*, mỗi một văn bản được biểu diễn bằng một tập các thuật ngữ* $D_i = \{d_{i1}, d_{i2}, .. d_{ip_i}\}$*. Gọi q là số các thuật ngữ 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:* $\boldsymbol{w_i} = (w_{i1}, w_{i2}, .. w_{iq})$ *trong không gian* $\boldsymbol{\mathcal{D}}$*. Trong đó,* w_{ik} *được tính theo Định nghĩa 2.1.*

d. Biểu diễn nội dung bài viết bằng véctơ trọng số

Trước khi đưa ra khái niệm bài viết được dùng trong mục này, luận án đề xuất Định nghĩa về mạng xã hội theo hướng tiếp cận của luận án như sau:

Định nghĩa 2.3:

Một mạng xã hội $\boldsymbol{\mathcal{N}}$ *là một bộ bốn:* $\boldsymbol{\mathcal{N}} =< U, E, G, B >$*. Trong đó:*

- $U = \{u_i\}$ *là tập những người dùng (user) trên mạng xã hội* $\boldsymbol{\mathcal{N}}$*,* u_i *là kí hiệu người dùng thứ i trong tập U.*
- $E = \{e_i\}$ *là tập các bài đã đăng/đã chia sẻ (entry) trên mạng xã hội* $\boldsymbol{\mathcal{N}}$*,* e_i *là kí hiệu bài đăng thứ i trong tập E.*
- $G = \{g_i\}$ *là tập các nhóm/ cộng đồng người dùng đã tham gia trên mạng xã hội* $\boldsymbol{\mathcal{N}}$*,* g_i *là kí hiệu nhóm thứ i trong tập G.*

- *B là tập các hành vi của người dùng trên mạng xã hội $\mathcal{N}$, các hành vi được luận án xem xét và phân tích trong chương 4 của luận án*

Theo khái niệm về bài viết trong mục 2.1, luận án đề xuất định nghĩa bài viết trên mạng xã hội theo hướng tiếp cận dữ liệu văn bản như sau:

Bài viết e trên mạng xã hội $\mathcal{N}$ là một văn bản ngắn được biểu diễn bởi một tập các từ, ký hiệu: $e = \{w_i\}, i = 1,2, \dots i_q, e \in E$, với E là tập các bài viết trên mạng xã hội $\mathcal{N}$.

Dựa trên định nghĩa 2.1 và định nghĩa 2.2 có thể biểu diễn một bài viết theo không gian các bài viết của người dùng như sau:

Định nghĩa 2.4:

Cho một tập các bài viết của người dùng $E = \{e_1, e_2, \dots, e_q\}$, mỗi bài viết được biểu diễn bằng một tập thuật ngữ $e_i = \{e_{i1}, e_{i2}, .. e_{iq_i}\}$. Gọi q là số thuật ngữ khác nhau từng đôi một trong không gian E. Khi đó, mỗi E_i được biểu diễn bởi một véctơ có q chiều: $\boldsymbol{w_i} = (w_{i1}, w_{i2}, .. w_{iq})$ trong không gian E. Trong đó, mỗi w_{ik} được tính như trong định nghĩa 2.1.

Ví dụ một bài viết trên mạng xã hội Facebook.com như sau: *"Lí do tôi chọn ngành thực phẩm sạch: được nhà cung cấp đích thân tới nhà thể hiện tài chế biến!!! Wow! Tuyệt vời! Hihi:D".* Như vậy, bài viết e trong trường hợp này được xác định bằng đoạn văn bản: "*Lí do tôi chọn ngành thực phẩm sạch: được nhà cung cấp đích thân tới nhà thể hiện tài chế biến!!! Wow! Tuyệt vời! Hihi*"

Sau khi tiến hành xử lý thì dữ liệu văn bản là nội dung của bài viết thì luận án thu được e = {cung cấp; đích thân; hiện tài; thể hiện; thực phẩm} đây là biểu diễn bài viết theo danh sách từ vựng.

d. Các thuật toán tiền xử lý dữ liệu văn bản

Để xử lý dữ liệu văn bản của các bài viết trên các mạng xã hội, luận án xây dựng hai thuật toán để tính toán và xử lý. Thuật toán thứ nhất dùng để phân tách văn bản

theo N-gram và loại bỏ từ dừng Thuật toán 2.2: *Thuật toán phân tách văn bản và xác định từ, thuật ngữ* và thuật toán thứ hai sau khi xác định được danh sách từ vựng của bài viết thì thực hiện tính trọng số cho mỗi từ xuất hiện trong bài viết đối với toàn bộ các bài viết của người dùng Thuật toán 2.2: *Xây dựng véctơ trọng số* cho nội dung các bài viết.

Ví dụ với bài viết: "*Từ năm 1950 đến nay, con người đã sản xuất ra khoảng 8 tỷ tấn nylon, và dưới 10% được tái chế. Hiện có 4.9 tỷ tấn rác thải nylon như thế nằm vương vãi ngoài môi trường. Đến năm 2050, con số này sẽ tăng lên gấp đôi. Bao bì nhựa thì chỉ có 14% được tái chế…… Tôi trân trọng những người sống với giá trị của mình, làm những việc dù rất nhỏ vì gía trị đó của mình. Giá trị của bạn là gì? Và bạn sẽ làm việc nhỏ gì để lan truyền cảm hứng đến một cộng đồng lớn hơn?.* ☹ ☹ *#Environment#*"

Bảng 2.8: Thuật toán 2.2 (Phân tích văn bản và xác định từ, thuật ngữ)

Thuật toán 2.2: Phân tích bài viết và xây dựng từ, thuật ngữ getTerm(x,y)
Input: Một bài viết trên mạng xã hội **Ouput:** Danh sách các từ của văn bản, Term

1: x←Text; y←∅; T1← ∅; T2← ∅; W← ∅; T3← ∅; *//Khởi tạo*
2: x← cleanText(x); *// Làm sạch văn bản x*
3: x← formatText(x); *//Chuẩn hóa các từ vựng trong x*
4: For i←2 to 4 do *//Tách từ cho x*
T1← T1 ∪ separateNgram(x,i) ; // N=2,3,4
End For
5: T2 ← removeStopWord(T1); //Loại bỏ từ dừng
6: If count(T2)← 10 then //Mở rộng từ vựng nếu cần
Open_word(T2,T3)
Else T3← T2;
End If
7: Return T3

Phân tích bằng Thuật toán 2.1 sẽ thu được danh sách từ khóa sau khi thực hiện sắp xếp theo thứ tự từ điển là: *{bao bì; cảm hứng; cảm thấy; chỉ có; con người; con số; cộng đồng; cửa hàng; đâu đó; đến nay; gấp đôi; giá trị; giải quyết; hàng không; khả năng; khác nhau; làm việc; lan truyền; môi trường; mục tiêu; người làm; nhìn thấy; như thế; sản*

phẩm; sản xuất; thấy có; thuỷ tinh; trân trọng; truyền cảm; vấn đề; vương vãi, …}

Sau khi thực hiện việc phân tách văn bản và thu được danh sách các từ, thuật ngữ cho mỗi bài viết trên mạng xã hội dựa trên Thuật toán 2.2, luận án thực hiện tính trọng số của các từ, thuật ngữ để xây dựng véctơ trọng số của mỗi bài viết bằng Thuật toán 2.3: *Xây dựng véctơ trọng số cho bài viết*. Thuật toán 2.3 được thực hiện như sau: Mỗi từ, thuật ngữ của bài viết, luận án thực hiện tính trọng số theo TF.IDF của chúng như đã trình bày trong định nghĩa 2.1.

Ví dụ với bài viết và danh sách từ khóa thu được trong mục 2.2.1 thì véctơ trọng số của bài viết theo tần suất xuất hiện TF.IDF là: *{(bao bì, 0.073); (cảm hứng, 0.049); (cảm thấy, 0.024); (chỉ có, 0.024); (con người, 0.024); (con số, 0.024); (cộng đồng, 0.024); (cửa hàng, 0.073); (đâu đó, 0.024); (đến nay, 0.024); (gấp đôi, 0.024); (giá trị, 0.049); (gía trị, 0.024); (giải quyết, 0.024); (hàng không, 0.024); (khả năng, 0.024); (khác nhau, 0.024); (làm việc, 0.049); (lan truyền, 0.049); (môi trường, 0.049); (mục tiêu, 0.024); (người làm, 0.024); (nhìn thấy, 0.024); (như thế, 0.024); (sản phẩm, 0.024); (sản xuất, 0.024); (thấy có, 0.024); (thuỷ tinh, 0.024); (trân trọng, 0.024); (truyền cảm, 0.024); (vấn đề, 0.024); (vương vãi, 0.024}*

Bảng 2.9: Thuật toán 2.3 (Xây dựng các véctơ trọng số cho bài viết)

Thuật toán 2.3: Tính các véctơ trọng số getWeightWord(x)
Input: Danh sách từ, thuật ngữ của bài viết e trên mạng xã hội N **Ouput:** Véctơ trọng số TF-IDF của bài viết e

1: **w**←∅; **wtfidf**←∅; *//Khởi tạo*
2: For i← to count(x) do *//Đếm tần suất của các từ khóa trong x*
w[i]← count(x[i]) ; N ← tổng số lượng các tài liệu
df_i ← số lượng các tài liệu mà từ w_i xuất hiện.
If **w**[i]>=1 then
wtfidf[i]←$(1+\log(f_{ij}))\log\left(\frac{N}{1+df_i}\right)$
else **wtfidf**[i]←0; *//Tính TF.IDF*
End For
3: Return **wtfidf;**

2.1.2. Biểu diễn người dùng bằng véctơ

Trên mỗi mạng xã hội, mỗi người dùng có một tập các bài viết, vì vậy, mỗi người dùng có thể biểu diễn dưới dạng một tập các bài viết. Không mất tính tổng quát, giả sử rằng trên $E = \{e_1, e_2, \dots, e_q\}$ là tập các bài viết tương ứng của *i* người dùng $\boldsymbol{U} = \{u_1, u_2, \dots, u_q\}$. Ký hiệu $e_i = \{e_{i1}, e_{i2}, .. e_{ik_i}\}$ là tập các bài viết của người dùng thứ *i.* Khi đó mỗi người dùng được biểu diễn bởi một véctơ gồm i_{k_i} thành phần, mỗi thành phần là một véctơ được xây dựng theo định nghĩa 2.4. Ký hiệu như sau:

$$u_i = \boldsymbol{u_i} = \left(\boldsymbol{w_{i1}}, \boldsymbol{w_{i2}}, .. \boldsymbol{w_{ik_i}}\right), \boldsymbol{w_{ik}} = (w_{ik1}, w_{ik2}, .. w_{ikq}) |\ k = 1, \dots, i_{kq}$$

trong không gian E. (2.3)

Cụ thể mỗi người dùng trên mạng xã hội có thể được biểu diễn như sau:

$$u_i = \begin{pmatrix} e_{i1} = \boldsymbol{w}_{i1} = (w_{i1}\ , w_{i1}\ , \dots, w_{i1q}), \\ e_{i2} = \boldsymbol{w}_{i2} = \left(w_{i21}, w_{i22}, \dots, w_{i2q}\right), \\ \dots. \\ e_{ik_i} = \boldsymbol{w}_{ik_i} = \left(w_{ik1}, w_{ik2}, \dots, w_{ikq}\right) \end{pmatrix} \tag{2.4}$$

Với *q* là số chiều của không gian E trên mạng xã hội đang xem xét.

2.1.3. Độ đo tương tự và độ tương quan giữa hai đối tượng

Luận án sử dụng độ đo Cosine để tính độ tương tự giữa hai đối tượng theo các véc tơ biểu diễn của hai đối tượng tương ứng như sau:

Giả sử có hai véc tơ $\boldsymbol{u} = (u_1, u_2, .. u_n)$ và $\boldsymbol{v} = (v_1, v_2, .. v_n)$ khi đó độ tương tự của u và v được tính bằng:

$$sim(u, v) = \frac{< \boldsymbol{u}, \boldsymbol{v} >}{\|\boldsymbol{u}\| * \|\boldsymbol{v}\|} \tag{2.5}$$

Trong đó, $< \boldsymbol{u}, \boldsymbol{v} >$ là tích vô hướng của 2 véctơ $\boldsymbol{u}$ và $\boldsymbol{v}$, còn $\|\boldsymbol{x}\|$ là độ dài Euclidean của véctơ $\boldsymbol{x}$

Để tính độ tương quan giữa hai đối tượng, luận án sử dụng độ tương quan Pearson theo công thức như sau:

$$cor(\boldsymbol{u}, \boldsymbol{v}) = \frac{\sum_i (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_i (u_i - \bar{u})^2} * \sqrt{\sum_i (v_i - \bar{v})^2}} \quad (2.6)$$

Trong đó, $\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i$ và $\bar{v} = \frac{1}{n}\sum_{i=1}^{n} v_i$ khi đó, $cor(\boldsymbol{u}, \boldsymbol{v})$ là độ tương quan giữa $\boldsymbol{u}$ và $\boldsymbol{v}$.

Luận án sử dụng hai độ đo này trong tính toán và ước lượng độ tương tự và độ tương quan giữa hai đối tượng như độ tương tự giữa hai bài viết, giữa hai người dùng, độ tương quan giữa bài viết và chủ đề, người dùng và chủ đề trong chương này và các chương sau.

2.1.4. Độ tương tự giữa hai người dùng theo nội dung bài viết

Trong mục này, luận án trình bày cách thức ước lượng độ tương tự giữa hai người dùng trên mạng xã hội dựa trên nội dung bài viết. Mục đích là để xem xét mối tương quan giữa hai người dùng tương tự nhau thì các chủ đề quan tâm của họ có tương tự nhau hay không và ngược lại. Để ước lượng độ tương tự giữa hai người dùng theo bài viết, luận án ước lượng độ tương tự của hai tập các bài viết tương ứng của hai người dùng trên mạng xã hội.

a. Độ tương tự giữa hai bài viết

Giả sử có hai bài viết e_{il} và e_{jk} của hai người dùng u_i và u_j tương ứng trên mạng xã hội $\boldsymbol{\mathcal{N}}$. Khi đó, độ tương tự giữa hai bài viết e_{il} và e_{jk} được tính bằng độ tương tự giữa hai véctơ trọng số tương ứng của e_{il} và e_{jk} như sau:

$$sim(\boldsymbol{e_{il}}, \boldsymbol{e_{jk}}) = \frac{< \boldsymbol{e_{il}}, \boldsymbol{e_{jk}} >}{\|\boldsymbol{e_{il}}\| \times \|\boldsymbol{e_{jk}}\|} \quad (2.7)$$

Trong đó, $\boldsymbol{e_{pq}}$ được tính theo định nghĩa 2.1.

Giả sử có hai người dùng u_i và u_j với hai tập bài viết E_i và E_j tương ứng trên mạng xã hội $\mathcal{N}$. Khi đó, độ tương tự giữa hai tập bài viết E_i và E_j được tính bằng độ tương tự giữa hai tập các véctơ trọng số tương ứng của u_i và u_j được ký hiệu là:

$$sim(\boldsymbol{E_i}, \boldsymbol{E_j}) = \max_{ik,il}(sim(\boldsymbol{e_{il}}, \boldsymbol{e_{jk}})) \quad (2.8)$$

Trong đó, $sim(\boldsymbol{e_{il}}, \boldsymbol{e_{jk}})$ được tính theo công thức (2.7)

b. Độ tương tự giữa hai người dùng theo nội dung bài viết

Để tính độ tương tự giữa hai người dùng u_i và u_j trên mạng xã hội $\mathcal{N}$ theo nội dung bài viết, luận án thực hiện tính độ tương tự giữa hai véctơ biểu diễn hai người dùng đó theo nội dung bài viết như công thức (2.3). Như vậy, độ tương tự giữa hai người dùng được xác định dựa trên độ tương tự giữa hai véctơ $\boldsymbol{u_i}$ và $\boldsymbol{u_j}$ và được định nghĩa một cách hình thức như sau:

Định nghĩa 2.5:

Cho hai người dùng u_i và u_j với hai tập bài viết E_i và E_j tương ứng trên mạng xã hội $\mathcal{N}$. Độ tương tự của hai người dùng được tính bằng:

$$sim(u_i, u_j) = sim(\boldsymbol{u_i}, \boldsymbol{u_j}) = sim(\boldsymbol{E_i}, \boldsymbol{E_j}) \quad (2.9)$$

Trong đó, $sim(\boldsymbol{E_i}, \boldsymbol{E_j})$ được tính theo công thức (2.8). Như vậy, độ tương tự của hai người dùng theo nội dung bài viết chính là độ tương tự giữa hai tập các bài viết của hai người dùng đó. Có thể thấy rằng độ tương tự của hai người dùng nằm trong khoảng đơn vị [0,1].

Để tiện lợi cho việc phân nhóm người dùng dựa trên độ tương tư, luận án sử dụng thang đo Likert và chia độ tương tự giữa hai người dùng thành 5 mức theo lý thuyết trong [88], chi tiết các mức được luận án sử dụng như trong bảng Bảng 2.10.

Bảng 2.10: Mức độ tương tự giữa hai đối tượng

Mức	Giả thiết	Khoảng	Ghi chú
0	Hai người dùng không tương tự nhau hoặc khác nhau	[0.000 – 0.025]	
1	Hai người dùng ít tương tự nhau	[0.025 – 0.050]	
2	Hai người dùng khá tương tự nhau	[0.050 – 0.075]	
3	Hai người dùng tương tự nhau	[0.075 – 0.100]	
4	Hai người dùng tương tự nhau nhiều hoặc rất nhiều	[0.100 – 1.000]	

2.2. MÔ HÌNH QUAN TÂM CỦA NGƯỜI DÙNG THEO CHỦ ĐỀ

2.2.1. Biểu diễn véctơ trọng số của chủ đề

Trong mục này luận án trình bày một định nghĩa về chủ đề theo hướng tiếp cận quan tâm của người dùng và cách thức xây dựng chủ đề cũng như cách biểu diễn chúng để tính toán trong luận án. Chủ đề được coi là chủ thể hay vấn đề cốt yếu của các bài viết mà người dùng thể hiện trên trang chủ của mình.

Như vậy có thể đưa ra khái niệm về chủ đề như sau: *Cho một tập các chủ đề về các lĩnh vực trên mạng xã hội. Khi đó, mỗi một chủ đề sẽ được biểu diễn bởi một tập hợp từ, thuật ngữ đặc trưng để mô tả và diễn giải về chủ đề đó.*

Giả sử rằng $\boldsymbol{\mathcal{T}} = \{T_1, T_2, \dots, T_p\}$ là tập các chủ đề trên mạng xã hội $\boldsymbol{\mathcal{N}}$, trong đó mỗi chủ đề được biểu diễn bằng một tập các từ $T_i = \{t_{i1}, t_{i2}, \dots, t_{ip_i}\}$. Khi đó, theo định nghĩa 2.2 ta có thể định nghĩa chủ đề một cách hình thức như sau:

Định nghĩa 2.6:

Cho một tập các chủ đề $\boldsymbol{\mathcal{T}} = \{T_1, T_2, \dots, T_P\}$ trên mạng xã hội $\boldsymbol{\mathcal{N}}$, khi đó, mỗi chủ đề T_i được biểu diễn bởi một tập các thuật ngữ hoặc các 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 tương ứng một véctơ trọng số được ký hiệu như sau:

$$\boldsymbol{t_i} = (w_{i1}, w_{i2}, \dots, w_{iq}) \tag{2.10}$$

Trong đó, mỗi w_{ik} được tính như trong Định nghĩa 2.1

Luận án không đi sâu vào bài toán phát hiện các chủ đề trên mạng xã hội, mà chỉ kế thừa các chủ đề phổ biến mà người dùng thường quan tâm làm đến để phân loại các bài viết của người dùng trong các nghiên cứu [14] [21] [28] [60] [62] [96]. Mỗi chủ đề trên mạng xã hội được luận án biểu diễn bằng một tên gọi hay thuật ngữ và được biểu diễn bằng một danh sách từ. Danh sách từ biểu diễn chủ đề được xây dựng dựa trên từ điển Wikipedia như đã trình bày trong Hình 2.3, các bước xây dựng các chủ đề và danh sách từ vựng của các chủ đề được trình bày trong các mục tiếp theo.

2.2.2. Xây dựng các chủ đề trên mạng xã hội

Do số lượng các chủ đề trên các mạng xã hội rất đa dạng, phong phú và khó thống kê một cách chính xác, nên trong luận án chỉ lựa chọn một số chủ đề phổ biến trên các phương tiện truyền thông xã hội để phục vụ cho nghiên cứu và thực nghiệm.

Bảng 2.11: Danh sách các trang tin tức điện tử tham khảo chủ đề

STT	Tên trang báo điện tử	Địa chỉ trang 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
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

Luận án thực hiện lựa chọn các chủ đề 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 [25] [145] [125]. 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 theo thống kê của *https://toplist.vn/top-list/website* cùng với 5 trang tin tức điện tử bằng Tiếng Anh phổ biến trên thế giới của *https://www.similarweb.com/top-websites/category/news-and-media.* Luận án thu được danh sách gồm 21 chủ đề có tần suất xuất hiện nhiều nhất trên 15 trang tin tức như trong Bảng 2.11 và Bảng 2.12

Bảng 2.12: Danh sách các chủ đề trên mạng xã hội

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
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

Khi xây dựng, các chủ đề tương tự nhau hoặc gần nhau được luận án tích hợp lại thành một, chẳng hạn như chủ đề "*Tài chính*" và "*Kinh doanh*" được tổ hợp lại

thành "*Tài chính – Kinh doanh*"; "*Sự kiện văn hóa*" và "*Tin tức giải trí*" được tích hợp thành "*Văn hóa – Giải trí*"; ... Để thuận tiện cho tính toán và ước lượng, luận án thực hiện tính danh sách từ vựng và trọng số của mỗi chủ đề. Mỗi chủ đề được lấy định nghĩa theo Wikipedia, sau đó thực hiện tách từ bằng N-gram, loại bỏ từ dừng, cuối cùng tính TF.IDF cho mỗi từ, thuật ngữ trong không gian 21 chủ đề đã xây dựng.

Các bước tính toán và ước lượng từ vựng và trọng số của mỗi chủ đề được thực hiện bằng hai thuật toán: Thuật toán 2.4: *Xây dựng danh sách từ vựng cho chủ đề* và Thuật toán 2.5: *Xây dựng véctơ trọng số cho mỗi chủ đề.*

Bảng 2.13: Thuật toán 2.4 (Xây dựng danh sách từ vựng cho các chủ đề)

Thuật toán 2.4: Xây dựng từ vựng cho các chủ đề, topicWord()
Input: Chủ đề t trên mạng xã hội N **Ouput:** Danh sách các từ vựng của chủ đề t

1: x ←∅; **tW** ← ∅; *//Khởi tạo*
2: x ← getDefineWiki(t); *// Lấy Định nghĩa từ Wikipedia cho t*
3: For i←2 to 4 do *//Tách từ cho x*
tW ← **tW** ∪ separateNgram(x,i) ; // N=2,3,4
End For
4: tW ← removeStopWord(tW); //Loại bỏ từ dừng
5: Return **tW;**

Thuật toán 2.4: *Xây dựng danh sách từ vựng cho chủ đề* được luận án thực hiện dựa trên thuật toán mở rộng từ vựng theo từ điển Wikipedia. Danh sách từ vựng thu được chính là danh sách các từ vựng của chủ đề *t* đang xem xét.

Bảng 2.14: Danh sách từ vựng của chủ đề

Chủ đề	**Danh sách từ vựng**
Giáo dục	Giáo dục, tiếng Anh, học tập, kiến thức, thói quen, thế hệ, giảng dạy, đào tạo, nghiên cứu, trải nghiệm, giáo dục, tiểu học, trung học, từ nguyên, từ đồng, tiếng Việt, toàn cầu, Quốc tế, Kinh tế, Xã hội, Văn hóa, Quốc công, cha mẹ, trực tuyến, Liên Hiệp Quốc, học trực tuyến, giáo dục tiểu học, …
Môi trường	Môi trường, tổ hợp, tự nhiên, xã hội, hệ thống, tập hợp, tương tác, định nghĩa, con người, không khí, độ ẩm, sinh vật, loài người, môi trường, vật chất, đối tượng, tập hợp con, …

Ví dụ với chủ đề *"Giáo dục"* và chủ đề *"Môi trường*" sau khi thực hiện Thuật toán 2.4 sẽ thu được danh sách từ vựng như trong Bảng 2.14. Sau khi thu được danh sách từ vựng của mỗi chủ đề, luận án thực hiện Thuật toán 2.5: *Xây dựng véctơ trọng số cho mỗi chủ đề* để tính trọng số cho mỗi từ vựng của các chủ đề. Trọng số của mỗi từ vựng trong mỗi chủ đề được bằng tần suất xuất hiện của các từ vựng đó trong không gian các chủ đề TF.IDF.

Bảng 2.15: Thuật toán 2.5 (Xây dựng véctơ trọng số cho mỗi chủ đề)

Thuật toán 2.5: Xây dựng véctơ trọng số getWeightTopic()
Input: Một danh sách từ vựng của chủ đề t **Ouput:** Véctơ trọng số TF-IDF của chủ đề t

1: **w**←∅; **wtfidftp**←∅; *//Khởi tạo*
2: For i← to count(t) do *//Đếm tần suất của các từ khóa trong t*
w[i]← count(tW[i]) ; N ← số lượng các chủ đề trong T
df_i ← số lượng các chủ đề mà từ khóa w_i xuất hiện.
If **w**[i]>=1 then
wtfidftp[i]←$(1+\log(f_{ij}))\log\left(\frac{N}{df_i}\right)$
else **wtfidftp**[i]←0; *//Tính TF.IDF*
End For
3: Return **w, wtfidftp;**

Ví dụ với chủ đề *"Môi trường"* sau khi thực hiện với thuật toán 2.4, luận án thực hiện Thuật toán 2.5 sẽ thu được danh sách các trọng số tương ứng như trong Bảng 2.16, trong đó danh sách từ vựng của chủ đề *"Môi trường"* đã được sắp xếp tăng dần theo thứ tự từ điển. Sau khi tính toán xong, luận án thu được một tập gồm 21 véctơ tương ứng với 21 chủ đề chứa danh sách từ và véctơ trọng số tương ứng như công thức (2.7).

$$\mathcal{T} = \begin{pmatrix} t_1 = \boldsymbol{t_1} = (w_{i1}, w_{i2}, \dots, w_{iq}), \\ t_2 = \boldsymbol{t_2} = (w_{i1}, w_{i2}, \dots, w_{iq}), \\ \dots. \\ t_{21} = \boldsymbol{t_{21}} = (w_{i1}, w_{i2}, \dots, w_{iq}) \end{pmatrix} \tag{2.11}$$

Trong đó, mỗi w_{ik} được tính như trong Định nghĩa 2.1

Bảng 2.16: Minh họa chủ đề và các trọng số của từ vựng tương ứng

Chủ đề	Termtp	TF-IDF	Termtp	TF-IDF
Môi trường	ảnh hưởng	0.0239	nào đó	0.1959
	bao gồm	0.0141	ngữ cảnh	0.1015
	bao quanh	0.1522	rõ ràng	0.0507
	bên ngoài	0.0392	sinh vật	0.0507
	có tính	0.0324	sử dụng	0.0324
	con người	0.0483	tác động	0.0124
	điều kiện	0.0418	tập hợp	0.0649
	định nghĩa	0.0276	tình trạng	0.0973
	độ ẩm	0.0507	tổ hợp	0.0392
	đối tượng	0.0507	tồn tại	0.0507
	hệ thống	0.1382	tự nhiên	0.0324
	hoàn cảnh	0.0507	tương tác	0.0478
	hoạt động	0.0247	vật chất	0.0392
	khác nhau	0.0324	xã hội	0.0276
	khách thể	0.1522	xác định	0.0323
	không khí	0.0507	xem xét	0.0276
	loài người	0.0324	xu hướng	0.0784
	môi trường	0.0239	ý nghĩa	0.0507
	…	…	…	…

2.2.3. Biểu diễn véctơ nội dung bài viết theo chủ đề

Khái niệm quan tâm đã được luận án trình bày trong chương một, trong mục này, luận án trình bày phương thức phát hiện quan tâm dựa trên các nội dung bài viết. Theo định nghĩa 2.3, ký hiệu $\boldsymbol{\mathcal{U}} = \{u_1, u_2, \dots, u_q\}$ là tập hợp những người dùng trên mạng xã hội $\boldsymbol{\mathcal{N}}$, trong một khoảng thời gian xác định, mỗi người dùng u_i sẽ đăng hoặc chia sẻ một tập hợp các bài viết $e_i = \{e_{i1}, e_{i2}, \dots, e_{in_i}\}$, ký hiệu $E = \{e_1, e_2, \dots, e_n\}$. Khi đó, véctơ biểu diễn nội dung bài viết theo không gian các chủ đề được định nghĩa một cách hình thức như sau:

Định nghĩa 2.7:

Giả sử $e_{ij} \in e_i$ là một bài viết của người dùng u_i trên mạng xã hội $\mathcal{N}$, 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:

$$\boldsymbol{e_{ij}^{k}} = \left(e_{ij}^{1}, e_{ij}^{2}, \dots, e_{ij}^{t_{kp}}\right) \tag{2.12}$$

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

2.2.4. Độ quan tâm của người dùng theo các chủ đề trên mạng xã hội

Độ quan tâm của người dùng theo các chủ đề thể hiện mối quan tâm của người dùng đó đến các chủ đề và được tính bằng mức độ liên quan của tập tất cả các bài viết của người dùng đó đối với các chủ đề. Ký hiệu mức độ liên quan giữa bài viết e_{ij} của người dùng u_i đối với chủ đề t_k là:

$$\alpha_{ij}^{k} = cor\left(e_{ij}, t_k\right) \tag{2.13}$$

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

$$cor\left(e_{ij}, p\right) = (\alpha_{ij}^{1}, \alpha_{ij}^{2}, \dots, \alpha_{ij}^{p}) \tag{2.14}$$

Có thể thấy rằng:

(1) 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.

(2) 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.

Như vậy, có thể định nghĩa mức độ quan tâm của người dùng như sau:

Định nghĩa 2.8:

Hàm số: $int: \mathcal{U} \times \mathcal{P}(E) \times \mathcal{T} \rightarrow [0,1]$ *được gọi là độ đo quan tâm nếu nó thỏa mãn điều kiện sau:* $int(u, U, t) \leq int(v, V, t)$*, đối với mọi* $U, V \in P(E_u)$ *với* $U \subseteq V$

Để cho đơn giản khi tính toán và biểu diễn, trong luận án này ký hiệu hàm quan tâm của người dùng u_i đến chủ đề t là $int(u_i, t)$. Dễ dàng chứng minh rằng:

Mệnh đề 2.8.1: *Các hàm số sau:*

$$(i) \quad intMax(u_i, t) = max_j\ (cor(e_{ij}, t)) \tag{2.15}$$

$$(ii) \quad intCor(u_i, t) = \frac{\sum_j cor(e_{ij}, t))}{\|E_i\|} \tag{2.16}$$

$$(iii) \quad intSum(u_i, t) = \frac{1}{2}\left(\frac{n_i^t}{\sum_{l \in \mathcal{T}} n_i^l} + \frac{n_i^t}{\sum_{u_k \in \mathcal{U}, l \in \mathcal{T}} n_k^l}\right)_j \tag{2.17}$$

là các độ đo quan tâm của người dùng đối với các chủ đề.

Trong đó, $cor(e_{ij}, t)$ là mức độ liên quan của bài viết e_{ij} đến chủ đề *t,* n_i^t là số lượng các bài viết liên quan đến chủ đề *t* của người dùng u_i trên mạng xã hội $\boldsymbol{\mathcal{N}}$.

Việc phát hiện các chủ đề quan tâm của người dùng dựa trên nội dung 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 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 không thì mức độ quan tâm của người dùng đến chủ đề đó càng thấp.

Luận án cũng sử dụng thang đo Likert và chia độ quan tâm giữa người dùng với các chủ đề thành 5 mức như trong Bảng 2.10 để phân loại mức độ quan tâm của người dùng theo các chủ đề.

2.2.5. Tương tự quan tâm theo chủ đề của người dùng

Độ tương tự giữa hai người dùng theo chủ đề mà họ quan tâm được tính dựa trên mức độ quan tâm của hai người dùng đối với các chủ đề trên mạng xã hội. Ký hiệu mức độ quan tâm của người dùng u_i đến chủ đề t trên mạng xã hội *N* là: $u_i^t =$

$int(u_i, t)$. Mức 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 2.9:

Độ quan tâm của người dùng u_i đến p chủ đề trong $\boldsymbol{\mathcal{T}}$ là một véctơ quan tâm, được biểu diễn như sau:

$$\boldsymbol{u_i^t} = (u_i^1, u_i^2, \dots, u_i^p) \tag{2.18}$$

Trong đó, mỗi u_i^k là độ quan tâm của u_i đến chủ đề thứ k, k=1, 2, ..., p, các u_i^k được tính theo một trong ba công thức của mệnh đề 2.9.1.

Khi đó, độ tương tự giữa hai người dùng theo chủ đề được định nghĩa như sau:

Định nghĩa 2.10:

Độ tương tự theo các chủ đề quan tâm của hai người dùng u_i, u_j được tính bằng độ tương tự cosine giữa hai véctơ quan tâm đến tất cả các chủ đề theo công thức:

$$sim_{int}(u_i, u_j) = sim(\boldsymbol{u_i^t}, \boldsymbol{u_j^t}) = \frac{< \boldsymbol{u_i^t}, \boldsymbol{u_j^t} >}{\| \boldsymbol{u_i^t}\| \times \|\boldsymbol{u_j^t}\|} \tag{2.19}$$

Trong đó, $< \boldsymbol{u_i^t}, \boldsymbol{u_j^t} >$ là tích vô hướng của hai véctơ, $\|\boldsymbol{X}\|$ là độ dài của véctơ. Dễ dàng thấy rằng, $sim_{int}(u_i, u_j)$ nằm trong khoảng [0,1].

2.3. TƯƠNG QUAN GIỮA TƯƠNG TỰ NGƯỜI DÙNG VÀ QUAN TÂM

2.3.1. Mối tương quan giữa tương tự và quan tâm của người dùng

Một câu hỏi cần đặt ra là liệu có mối quan hệ nào giữa những người dùng tương tự nhau theo không gian bài viết và những người dùng tương tự nhau theo không gian các chủ đề hay không? Nói cách khác là nếu hai người dùng có độ tương tự với nhau theo không gian các bài viết liệu họ có tương tự nhau theo không gian các chủ đề hay không và ngược lại?

Xét bài toán sau:

Cho một tập người dùng U trên mạng xã hội $\mathcal{N}$. Gọi SimU là tập những người dùng tương tự nhau dựa trên nội dung các bài viết và CorrU tập người dùng tương tự nhau theo chủ đề.

Để xem xét mối tương quan giữa hai tập *SimU và CorrU,* luận án tiến hành hai thực nghiệm trình bày trong mục 2.3.2.

- Thực nghiệm thứ nhất là tính độ tương tự giữa hai người dùng theo không gian bài viết dựa trên độ đo tương tự giữa hai người dùng đã trình bày trong mục 2.1.3 để tìm tập SimU
- Thực nghiệm thứ hai là xác định độ tương quan theo các chủ đề quan tâm của người dùng theo không gian các chủ đề như đã trình bày trong mục 2.2.5 để tìm tập CorrU

Sau đó, luận án so sánh và thống kê để tìm giao của hai nhóm SimU và CorrU. Nếu có người dùng xuất hiện trong cả hai nhóm thì có thể kết luận rằng SimU và CorrU có sự tương quan với nhau.

Khi tính toán và phân loại theo độ tương tự dựa trên không gian bài viết thu được 5 nhóm với 5 mức độ tương tự nhau là: $\{S_{sim0,}S_{sim1,}S_{sim2,}S_{sim3,}S_{sim4,}\}$ và theo độ tương quan dựa trên không gian chủ đề thu được 5 nhóm với 5 mức độ tương quan là: $\{S_{int0,}S_{int1,}S_{int2,}S_{int3,}S_{in}\ ,\}$ trong đó, mỗi $S_k = \{u_{k1}, u_{k1}, .. u_{km}\}$ là tập hợp những người tương tự nhau theo mức k trên một trong hai không gian đã trình bày.

Để xét độ tương quan giữa hai độ đo này, luận án thực hiện tính các tập người dùng giao giữa các tập theo mức độ tương ứng $\{S_{cor0,}S_{cor1,}S_{cor2,}S_{cor3,}S_{cor4,}\}$ trong đó $S_{cork=}S_{simk} \cap S_{int}, với\ k = 0, .. 4$. Sau đó, so sánh lực lượng của các tập hợp theo các mức độ tương ứng.

Như vậy, nếu các S_{cork} càng gần với S_{simk} hoặc S_{intk} thì hai độ đo này càng có mối liên quan đến nhau. Hay nói chính xác là khi đó, những người dùng tương tự nhau theo không gian bài viết sẽ có các chủ đề quan tâm tương tự nhau và ngược lại.

Nếu các S_{cork} càng xa với S_{simk} hoặc S_{intk} thì có thể kết luận rằng những người dùng tương tự nhau theo không gian bài viết sẽ không tương tự nhau theo không gian chủ đề và ngược lại.

Để khẳng định cho các nhận định và giả thiết trên, luận án tiến hành các thực nghiệm chi tiết trong mục 2.3.2.

2.3.2. Xác định độ quan tâm và vấn đề tương quan

Như đã trình bày trong mục 2.3.1, mục đích của các thực nghiệm ngoài việc phân nhóm người dùng theo các mức độ tương tự theo không gian nội dung các bài viết và không gian các chủ đề trên mạng xã hội. Các kết quả thực nghiệm còn giúp luận án tìm ra mối tương quan giữa những người dùng theo hai không gian này dựa trên nội dung của bài viết.

a. Xây dựng các bộ dữ liệu thực nghiệm

Để tiến hành thực nghiệm, luận án thực hiện thu thập một bộ dữ liệu thực từ mạng xã hội Facebook (http://www.facebook.com) như đã trình bày trong phần mở đầu. Sau khi loại bỏ những bài viết không chứa văn bản, các bài viết chỉ chứa các biểu tượng và các bài viết quá ngắn chỉ có một đến hai từ hoặc các ký tự đặc biệt, luận án thu được một bộ gồm 200 người dùng với mỗi người có 10 bài viết hợp lệ, tổng cộng có 2000 nội dung bài viết, một số người dùng có số lượng các bài viết quá ít dưới 5 bài thỏa mãn thì luận án loại bỏ ra khỏi bộ dữ liệu thử nghiệm.

Bảng 2.17: Thông số bộ dữ liệu thử nghiệm

	Bộ dữ liệu tính độ tương tự	Bộ dữ liệu tính độ tương quan
User	200	200
Entry	2000	2000
Topic	0	21
Trọng số	TF.IDF	TF.IDF
Biểu diễn	Theo không gian bài viết	Theo không gian chủ đề

b. Ước lượng độ tương tự giữa hai người dùng

- **Kịch bản thực nghiệm:** Kịch bản thực nghiệm xác định độ tương tự của hai người dùng như sau:

  Đầu vào: Danh sách 2000 bài viết của 200 người dùng

  Đầu ra: Độ tương tự giữa các cặp người dùng theo nội dung bài viết

  Thực hiện

  - Xây dựng danh sách từ, thuật ngữ cho các bài viết

  - Tính véctơ trọng số theo TF.IDF cho mỗi bài viết theo không gian bài viết

  - Tính độ tương tự giữa các cặp bài viết

  - Phân loại các cặp người dùng theo độ tương tự của hai tập bài viết tương ứng

  Kết thúc

- **Tham số đầu ra:** Đầu ra là độ tương tự của mỗi cặp người dùng, tham số đầu ra được tính dựa trên các công thức (2.7) và (2.8) theo định nghĩa 2.6

Bảng 2.18: Độ tương tự giữa các cặp bài viết

	e1	**e2**	**e3**	**e4**	**e5**	**e6**	**e7**	**e8**	**e9**	**e10**
e1	1.000	0.327	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
e2	0.327	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
e3	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
e4	0.000	0.000	0.000	1.000	0.000	0.316	0.000	0.000	0.000	0.000
e5	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
e6	0.000	0.000	0.000	0.316	0.000	1.000	0.000	0.000	0.000	0.000
e7	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
e8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000
e9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000
e10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
e11	0.000	0.000	0.000	0.000	0.000	0.000	0.089	0.000	0.000	0.000
e12	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
e13	0.000	0.000	0.000	0.000	0.000	0.000	0.114	0.000	0.000	0.000
e14	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
e15	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

- **Kết quả thực nghiệm:** Bước 1 và Bước 2 được thực hiện như trong thực nghiệm trong mục 2.7.2, chỉ khác là các véctơ được tính trong không gian các bài viết, không tính theo chủ đề. Bước 3 tính độ tương tự giữa các cặp bài viết dựa trên Định nghĩa 2.6 theo công thức (2.7), minh họa trong Bảng 2.18. Bước 4 tính độ tương tự giữa các cặp người dùng dựa trên định nghĩa 2.6 theo công thức 2.8 sau khi phân loại theo mức độ tương tự được minh họa như trong Bảng 2.19. Các cặp người dùng được lưu thành một ma trận và có độ tương tự được tính theo độ tương của hai tập bài viết tương ứng.

Sau khi tính toán, luận án tính được số các cặp người dùng tương tự nhau theo không gian bài viết như sau: Tổng số các cặp người dùng tính toán: (200 * 200) /2 = 20.000 cặp do các cặp (u_i, u_j); (u_j, u_i) là có độ tương tự bằng nhau, nên luận án chỉ tính một lần số các cặp người dùng trong bộ dữ liệu.

Bảng 2.19: Độ tương tự giữa các cặp người dùng theo không gian bài viết

	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8	U_9	U_{10}	U_{11}	U_{12}	U_{13}	U_{14}	U_{15}
U_1	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
U_2	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0
U_3	1	1	4	0	0	0	0	0	0	0	0	0	0	0	0
U_4	1	2	3	4	0	0	0	0	0	0	0	0	0	0	0
U_5	2	1	1	1	4	0	0	0	0	0	0	0	0	0	0
U_6	2	3	2	2	2	4	0	0	0	0	0	0	0	0	0
U_7	2	2	2	2	1	2	4	0	0	0	0	0	0	0	0
U_8	2	0	2	2	1	0	3	4	0	0	0	0	0	0	0
U_9	0	1	1	0	0	0	1	0	4	0	0	0	0	0	0
U_{10}	1	1	3	3	1	1	2	2	2	4	0	0	0	0	0
U_{11}	2	4	3	4	3	3	4	3	2	2	4	0	0	0	0
U_{12}	1	1	1	1	1	0	2	1	1	0	2	4	0	0	0
U_{13}	0	2	0	2	2	1	2	1	0	0	2	2	4	0	0
U_{14}	0	1	1	1	1	0	1	1	0	0	1	1	1	4	0
U_{15}	1	0	1	1	1	1	2	3	1	0	2	4	3	1	4

Theo mức độ tương tự đã xây dựng trong Bảng 2.10, luận án thu được kết quả thực nghiệm theo các mức tương tự ước lượng trong không gian các bài viết của người dùng và được trình bày trong Bảng 2.20

Bảng 2.20: Nhóm các cặp người dùng tương tự theo không gian bài viết

Mức 0	Mức 1	Mức 2	Mức 3	Mức 4
5094	8961	4122	1355	468

Kết quả trong Bảng 2.20 có thể thấy rằng độ tương tự theo không gian bài viết của các cặp người dùng ở mức 1 là nhiều nhất, và ở mức 4 là ít nhất.

c. Thực nghiệm xác định các mức độ quan tâm của người dùng theo các chủ đề

- **Kịch bản thực nghiệm:** Kịch bản xác định mức độ quan tâm của người dùng theo các chủ đề thực hiện như sau:

Đầu vào: Danh sách 2000 bài viết và 21 chủ đề dùng làm nhãn để phân loại

Đầu ra: Mức độ tương quan giữa mỗi người dùng và các chủ đề

Thực hiện

- Xây dựng danh sách từ, thuật ngữ cho các bài viết và các chủ đề

- Tính véctơ trọng số theo TF.IDF cho mỗi bài viết và mỗi chủ đề

- Tính độ tương quan giữa mỗi bài viết và các chủ đề

- Phân loại người dùng theo mức độ quan tâm với 21 chủ đề

Kết thúc

- **Tham số đầu ra**

  Đầu ra của kịch bản là độ tương quan của mỗi người dùng với 21 chủ đề trên mạng xã hội, tham số đầu ra được tính dựa trên các công thức (2.15), (2.16) và (2.17) theo Định nghĩa 2.7. Phân loại các mức độ quan tâm và các độ đo tương theo 5 mức như trình bày trong Bảng 2.10

- **Kết quả thực nghiệm như sau**

  Bước 1 và bước 2 xây dựng danh sách từ, thuật ngữ cho bài viết và tính véctơ trọng số theo TF.IDF, sử dụng các thuật toán 2.1, 2.2, 2.3, 2.4 và 2.5. Bước 3 tính độ tương quan giữa các bài viết với các chủ đề, kết quả được độ tương quan dựa trên Định nghĩa 2.7 theo công thức (2.14)

Kết quả thực nghiệm về độ tương quan của các bài viết với chủ đề được trình bày trong Bảng 2.21, mỗi ô trong bảng là độ tương quan giữa bài viết ở cột BV với các chủ đề tương ứng, độ tương quan được xác định dựa trên độ tương tự giữa véctơ bài viết $\boldsymbol{e_{ij\_k}}$ với chủ đề $\boldsymbol{t_l}$.

Độ tương quan sau đó được phân loại vào 5 nhóm từ mức 0 đến mức 4 như trong Bảng 2.10. Mức 0 là hầu như không tương quan và mức 4 là có sự tương quan lớn nhất. Nghĩa là bài viết đó gần với chủ đề đang xét nhất. Các ô có chữ đậm là có độ tương quan lớn nhất của chủ đề đang xét.

Hàng TB là độ tương quan trung bình của các bài viết của người dùng ij với các chủ đề trong thực nghiệm.

Bảng 2.21: Độ tương quan của các bài viết với các chủ đề

BV	Môi trường	Chính trị	Sức khỏe	Công nghệ	Thể thao	Văn hóa - Giải trí	Du lịch	Tài chính - Kinh doanh	Giáo dục	Khoa học kỹ thuật
e12_1	**0.0272**	0.0000	0.0147	0.0308	0.0465	**0.0372**	0.0000	0.0324	0.0105	**0.0112**
e12_2	0.0000	0.0000	0.0896	**0.0538**	**0.0709**	0.0072	**0.0795**	**0.0666**	0.0000	0.0098
e12_3	0.0000	**0.0172**	0.0000	0.0000	0.0138	0.0000	0.0000	0.0000	0.0100	0.0000
e12_4	0.0000	0.0000	0.0000	0.0000	0.0413	0.0000	0.0000	0.0000	0.0000	0.0000
e12_5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
e12_6	0.0000	0.0000	**0.3201**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
e12_7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0209	0.0382	0.0000	**0.0265**	0.0000
TB	**0.0039**	**0.0025**	**0.0606**	**0.0121**	**0.0246**	**0.0093**	**0.0168**	**0.0141**	**0.0067**	**0.0030**
Max	**0.0272**	**0.0172**	**0.3201**	**0.0538**	**0.0709**	**0.0372**	**0.0795**	**0.0666**	**0.0265**	**0.0112**

Bước 4 tính và phân loại độ quan tâm của người dùng theo các chủ đề dựa trên Định nghĩa 2.7 và các công thức (2.15), (2.16) và (2.17). Các kết quả của độ tương tự được minh họa trong các Bảng 2.22

Bảng 2.22: Độ tương quan của người dùng theo chủ đề theo công thức (2.15)

User	Chính trị	Du lịch	Gia đình	KH - CN	Môi trường	NT-GT	Sức khỏe	TC - KD	Thể thao
U001	0.0000	0.1398	0.0566	0.1183	0.0420	0.0204	0.0213	0.0369	0.1147
U002	0.0000	0.0530	0.0801	0.0524	0.0476	0.0867	0.0311	0.0253	0.0474
U003	0.0939	0.0379	0.0397	0.0444	0.0800	0.1325	0.0725	0.1239	0.0869
U004	0.0527	0.0342	**0.1664**	0.1109	0.0849	0.0714	**0.1280**	0.0836	0.0615
U005	0.0000	0.0414	0.1448	0.0766	0.0185	0.0000	0.0000	0.0081	0.0158
U006	0.0791	0.0000	0.1114	0.1132	0.0890	0.0642	0.0640	0.0261	0.0271
U007	0.0895	0.1303	0.1093	0.1034	0.0625	0.0735	0.0435	**0.2256**	0.0659
U008	0.0598	**0.2272**	0.0431	0.0144	0.0609	0.0613	0.0878	0.1970	0.0766
U009	**0.3873**	0.0000	0.0000	0.0000	0.0000	0.0000	0.1045	0.0112	0.0164
U010	0.1581	0.0468	0.1448	0.0588	0.1483	0.0000	0.0640	0.0166	**0.2059**
U011	0.0994	0.1543	0.1000	**0.2322**	**0.1861**	0.1513	0.1143	0.2173	0.0413
U012	0.0289	0.0000	0.0000	0.0179	0.0108	0.0613	0.0234	0.0095	0.1432
U013	0.0000	0.0367	0.0930	0.0000	0.0000	0.0419	0.0242	0.0394	0.0287
U014	0.0456	0.0000	0.0283	0.1157	0.0514	0.0000	0.0370	0.0000	0.0164
U015	0.0674	0.0000	0.1288	0.0000	0.0839	**0.1529**	0.0302	0.1778	0.0648

Bảng 2.22 minh họa độ tương quan của từng người dùng với các chủ đề dựa trên nội dung các bài viết. Độ tương quan của của người dùng được tính dựa trên một trong ba công thức trong mệnh đề 2.8.1. Trong Bảng 2.22 là độ tương quan tính theo công thức 2.15 tức là dựa trên bài viết có độ tương quan lớn nhất trong 10 bài viết đang xét của người dùng trong cột User. Tính theo ba công thức trong mệnh đề 2.8.1 thì độ tương quan được minh họa trong Bảng 2.23.

Phân loại theo 5 mức độ quan tâm của người dùng theo Bảng 2.10 thì Bảng 2.23 sẽ thu được kết quả như Bảng 2.24, các mức độ quan tâm được chia thành 5 mức để phân nhóm người dùng.

Dựa vào cách phân loại này có thể thấy, tính theo công thức 2.14 thì mức độ tương quan đạt cao nhất, tiếp theo là tính theo công thức 2.13, công thức 2.12 cho kết quả phân loại thấp nhất. Điều này có thể thấy, nếu tính trung bình dựa trên nhiều bài viết sẽ ra mức độ thấp hơn là dựa trên mức độ tương quan cao nhất hoặc dựa trên số lượng các bài viết cùng liên quan đến một chủ đề. Nghĩa là khi người dùng có nhiều

bài viết liên quan đến một chủ đề nhiều thì mức độ quan tâm của người dùng đối với chủ đề đó sẽ cao hơn là chỉ có một số ít các bài liên quan. Điều này cũng đúng với thực tế, khi một người dùng quan tâm đến chủ đề hoặc sự kiện nào đó thì họ thường đăng hoặc chia sẻ nhiều bài viết về chủ đề đó hơn là các chủ đề khác.

Bảng 2.23: Độ tương quan của người dùng theo (2.15), (2.16) và (2.17)

Chủ đề	Chính trị			Du lịch			Gia đình		
User	2.17	2.16	2.15	2.17	2.16	2.15	2.17	2.16	2.15
U001	0.0000	0.0000	0.0000	0.1378	0.0136	0.1398	0.1023	0.0061	0.0566
U002	0.0000	0.0000	0.0000	0.0722	0.0035	0.0530	0.1072	0.0086	0.0801
U003	0.1297	0.0118	0.0939	0.0305	0.0018	0.0379	0.1507	0.0070	0.0397
U004	0.1622	0.0074	0.0527	0.0914	0.0030	0.0342	0.1808	0.0178	0.1664
U005	0.0000	0.0000	0.0000	0.0451	0.0020	0.0414	0.0896	0.0120	0.1448
U006	0.0672	0.0044	0.0791	0.0000	0.0000	0.0000	0.1253	0.0123	0.1114
U007	0.1297	0.0107	0.0895	0.1219	0.0104	0.1303	0.1206	0.0075	0.1093
U008	0.0336	0.0028	0.0598	0.0633	0.0132	0.2272	0.0627	0.0034	0.0431
U009	0.0642	0.0184	0.3873	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
U010	0.1048	0.0111	0.1581	0.0660	0.0037	0.0468	0.1632	0.0163	0.1448
U011	0.6486	0.0428	0.0994	0.3962	0.0272	0.1543	0.3918	0.0266	0.1000
U012	0.0380	0.0014	0.0289	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
U013	0.0000	0.0000	0.0000	0.0379	0.0017	0.0367	0.0752	0.0059	0.0930
U014	0.0399	0.0022	0.0456	0.0000	0.0000	0.0000	0.0376	0.0013	0.0283
U015	0.0380	0.0032	0.0674	0.0000	0.0000	0.0000	0.0715	0.0086	0.1288

Bảng 2.24: Phân loại theo các mức quan tâm của người dùng với các chủ đề

Topic	Chính trị			Du lịch			Gia đình			Giáo dục		
User	2.17	2.16	2.15	2.17	2.16	2.15	2.17	2.16	2.15	2.17	2.16	2.15
U001	0	0	0	4	0	4	4	0	2	1	0	3
U002	0	0	0	2	0	2	4	0	3	2	0	0
U003	4	0	3	1	0	1	4	0	1	4	0	4
U004	4	0	2	3	0	1	4	0	4	4	0	3
U005	0	0	0	1	0	1	3	0	4	3	0	0
U006	2	0	3	0	0	0	4	0	4	1	0	0
U007	4	0	3	4	0	4	4	0	4	2	0	4
U008	1	0	2	2	0	4	2	0	1	3	0	4
U009	2	0	4	0	0	0	0	0	0	0	0	0
U010	4	0	4	2	0	1	4	0	4	4	1	4
U011	4	1	3	4	1	4	4	1	4	4	1	4
U012	1	0	1	0	0	0	0	0	0	2	0	0
U013	0	0	0	1	0	1	3	0	3	1	0	0
U014	1	0	1	0	0	0	1	0	1	1	0	1
U015	1	0	2	0	0	0	2	0	4	4	0	1

Kết quả thực nghiệm thu được dựa trên các mức tương tự theo không gian chủ đề các nhóm như minh họa trong Bảng 2.25. Trong Bảng 2.25 các nhóm người dùng được phân tích từ các tập $\{S_{int0,}S_{int1,}S_{int2,}S_{int3,}S_{int4,}\}$. Mỗi tập này luận án đếm số các người dùng khác nhau từng đôi một để suy ra số lượng người dùng quan tâm đến các chủ đề.

Bảng 2.25: Phân loại theo các mức theo chủ đề quan tâm

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
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**	**1199**

Ví dụ với chủ đề "*Tài chính – Kinh doanh*", mức 0 có 18 người nghĩa là có 18/200 người có mức độ tương quan nằm trong khoảng [0.000 – 0.025] tương ứng với hầu như không có bài viết nào liên quan đến chủ đề này hay không quan tâm đến

chủ đề này, mức 1 có 42 người có mức độ tương quan nằm trong khoảng [0.025 – 0.050] nghĩa là có liên quan ít đến chủ đề này hay hơi quan tâm đến chủ đề này, mức 2 có 31 người có mức độ liên quan nằm trong khoảng [0.050 – 0.075] nghĩa là liên quan đến chủ đề này hay là quan tâm đến chủ đề này, mức 3 có 33 người nằm trong khoảng [0.075 – 0.100] nghĩa là có 33 người khá liên quan đến chủ đề hay 33 người khá quan tâm đến chủ đề này và có 76 người có mức độ liên quan nằm trong khoảng [0.100 – 1.000] nghĩa là có 76 người rất quan tâm đến chủ đề này.

2.3.3. Thảo luận về kết quả

Để đánh giá độ chính xác của việc xác định các chủ đề quan tâm của người dùng theo độ tương quan giữa các bài viết và chủ đề và tính độ tương tự giữa hai người theo các bài viết, luận án thực hiện đánh giá của các tình nguyện viên cho các bộ mẫu. Sau đó, luận án thực hiện so sánh kết quả đó với kết quả thu được từ thực nghiệm các mô hình đề xuất, 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. Độ chính xác của thực nghiệm xác định chủ đề quan tâm của người dùng được tính dựa trên *Sai số bình phương trung bình (*MSE - Mean Square Error)*:*

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2 \tag{2.19}$$

Trong đó, *n* tổng số cặp người dùng với các chủ đề, trong đó p_i là giá trị đánh giá dự đoán của cặp (u_i, t_j) và r_i là giá trị đánh giá thực tế của cặp (u_i, t_j). Với $u_i, i = 1..200$ và $t_j, j = 1..21$. Như vậy có tất cả là (200 * 21) = 4200 cặp, tức là n=4200. Với cách tính này thì MSE càng gần đến giá trị 0 thì độ chính xác càng cao và ngược lại. Kết quả :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2 = \frac{1}{4200}\sum_{i=1}^{4200}(p_i - r_i)^2 = 0.352$$

Tương ứng với độ chính xác của thực nghiệm là:

$$CR = (1\text{-}MSE) * 100\% = (1 - 0.315) * 100\% = 68.5\% \quad (2.20)$$

Độ chính xác của thực nghiệm xác định độ tương tự của các cặp người dùng cũng được tính dựa trên *Sai số bình phương trung bình (*MSE - Mean Square Error), khi đó, *n* tổng số cặp người dùng thực tế, p_i là giá trị đánh giá dự đoán của cặp (u_i, u_j) và r_i là giá trị đánh giá thực tế của cặp (u_i, u_j).

Với $u_i, i = 1..200$ và $u_j, j = 1..200$. Tuy nhiên trên thực tế là độ tương tự của $(u_i, u_j) = (u_j, u_i)$, vì vậy, luận án thực hiện trên một nữa số cặp hay (200 * 200)/2= 20.000 cặp, tức là n=20000. Kết quả :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2 = \frac{1}{20000}\sum_{i=1}^{20000}(p_i - r_i)^2 = 0.296$$

Tương ứng với độ chính xác bằng

$$CR = (1\text{-}MSE) * 100\% = (1 - 0.296) * 100\% = 70.8\% \quad (2.21)$$

Như vậy, kết quả thực hiện trên nội dung của bài viết khá khả quan với độ chính xác không quá tốt chỉ được 68.5% và 70.8%, nguyên nhân là các bài viết trên mạng xã hội ngắn, nhiều bài viết khi so sánh không đưa ra được chính xác độ tương tự với nhau hoặc độ tương quan với các chủ đề.

Để xem xét mức độ tương quan giữa các độ đo trong các nhóm người dùng, luận án sử dụng cách thức thống kê số lượng những người có cùng mức độ tương tự theo không gian bài viết sau và những người có cùng độ tương tự theo không gian các chủ đề để tìm sự tương quan giữa hai độ đo này. Cách tính các chỉ số trong Bảng 2.25 như sau:

Tổng các cặp người dùng được xét theo độ tương tự là 20000 cặp;

Tổng số cặp xét theo tương quan với chủ đề là 4200 cặp, sau đó luận án đếm các người dùng khác nhau đôi một trong mỗi mức, rồi lấy giao của các tập người dùng này.

Bảng 2.26: Nhóm các cặp người dùng tương tự theo không gian bài viết

	Mức 0	**Mức 1**	**Mức 2**	**Mức 3**	**Mức 4**
Phân loại theo độ tương tự	5094	8961	4122	1355	468
Tập người dùng	67	77	126	75	87
Phân loại theo chủ đề	776	739	733	753	1199
Tập người dùng theo chủ đề	95	76	74	110	143
Tập giao của hai không gian	43	45	57	59	67
Tỷ lệ trùng nhau	**53.08%**	**58.82%**	**57.00%**	**63.78%**	**58.26%**

Nhìn vào kết quả của Bảng 2.26 có thể thấy rằng, tỷ lệ các nhóm người dùng cùng mức độ tương tự giao nhau theo hai không gian đều trên 50%. Điều đó cho thấy rằng có sự tương quan giữa độ đo tương tự khi thực hiện tính theo nội dung của bài viết. Nói cách khác nếu hai người dùng có độ tương tự nhau theo nội dung bài viết thì sẽ có khả năng trên 50% là họ sẽ có các chủ đề quan tâm tương tự nhau.

Tuy nhiên tỷ lệ khá thấp, cao nhất là 64% ở mức số 3. Vậy nếu xem xét thêm các đặc trưng khác của người dùng thì tỷ lệ này có tăng hay không? Chẳng hạn nếu xem xét các bài viết không chỉ chứa chỉ nội dung văn bản mà còn kèm theo các đặc tính khác như các thẻ đánh dấu, các nhóm, các biểu tượng thể hiện cảm xúc hay quan điểm ẩn chứa trong các bài viết... thì có thể cải thiện được độ chính xác khi phân loại các mức quan tâm của người dùng thông qua bài viết của họ đến các chủ đề hay không? Độ tương tự của các mức có tăng lên và tỷ lệ tương quan giữa các độ đo này có mang lại kết quả tốt hơn? Vấn đề này sẽ được luận án đề cập và đưa ra hướng giải quyết cũng như trình bày chi tiết trong Chương 3.

2.4. KẾT LUẬN

Chương 2 luận án đã trình bày cách thức biểu diễn nội dung bài viết trên mạng xã hội bằng véctơ trọng số trong không gian các bài viết và trong không gian các chủ đề. Các nội dung bài viết được tính toán dựa trên trọng số của các từ và thuật ngữ, trọng số của nội dung bài viết theo hai không gian này được tính theo TF.IDF. Dựa trên cách biểu diễn bằng véctơ trọng số, luận án đã đưa ra cách tính độ tương tự giữa hai người dùng theo không gian các nội dung bài viết và theo không gian các chủ đề trên mạng xã hội. Mối tương quan giữa hai độ đo này cũng được xem xét và chứng minh bằng thực nghiệm, từ kết quả thực nghiệm có thể thấy rằng, nếu hai người dùng tương tự nhau theo không gian nội dung các bài viết thì cũng có thể tương tự nhau các chủ đề quan tâm. Nhưng nếu chỉ xét dựa trên nội dung các bài viết của người dùng thì tỷ lệ này là không cao, vì vậy, luận án đề xuất sẽ xem xét thêm các đặc trưng khác của bài viết để có thể làm kết quả của cách thức tính toán được tốt hơn và sẽ được trình bày trong chương 3. Kết quả nghiên cứu trong chương này đã được công bố trên Tạp chí *Southeast Asian Journal of Sciences, Vol. 09, No 1 (2019), pp. 01–10. ISSN 2286 – 7724 và* Tạp chí *International Journal of Advanced Computer Science and Applications (IJACSA)* (Vol. 6, No. 2, 2015)

CHƯƠNG 3: MÔ HÌNH VÀ QUAN TÂM CỦA NGƯỜI DÙNG DỰA TRÊN BÀI VIẾT MỞ RỘNG

Trong chương này luận án phân tích một số hạn chế đã đề cập ở cuối Chương hai, để cải tiến mô hình tính toán và xử lý, luận án đề xuất biểu diễn người dùng qua mô hình bài viết có nhiều đặc trưng bao gồm nội dung, các thẻ đánh dấu, thể loại, quan điểm và cảm xúc. Để thực hiện đề xuất này, luận án đưa ra cách thức để ước lượng giá trị cho các đặc trưng và biểu diễn chúng dưới dạng các véctơ trọng số trong mục 3.2. Dựa trên mô hình bài viết với nhiều đặc trưng, luận án biểu diễn người dùng và các chủ đề cùng phương thức tính mức độ quan tâm của người dùng theo bài viết có nhiều đặc trưng trong mục 3.3, 3.4 và 3.5.

3.1. XÁC ĐỊNH QUAN TÂM CỦA NGƯỜI DÙNG THEO BÀI VIẾT

Bài toán phát hiện quan tâm của người dùng đã được nghiên cứu và phân tích theo nhiều hướng tiếp cận khác nhau. Trong đó hướng tiếp cận theo bài viết hay bài đăng là một trong những hướng nghiên cứu được sử dụng phổ biến bởi việc thu thập dữ liệu không phụ thuộc quá nhiều vào cấu trúc và các dịch vụ mà mạng xã hội cung cấp. Theo hướng tiếp cận phân tích bài viết, các nghiên cứu đã có thường nghiên cứu nội dung bài viết [21] [23] [50], theo thẻ đánh dấu như nghiên cứu [145] [125], theo cảm xúc [152] [163] và [58].

Tuy nhiên, khi phân loại bài viết theo nội dung, có thể gặp một số vấn đề có thể dẫn đến hiệu quả của mô hình không đạt kết quả cao như mong đợi, chẳng hạn như có những bài viết quá ngắn hoặc quá ít từ loại, việc bài viết quá ngắn hay quá ít từ sẽ bị loại bỏ khi xây dựng bộ dữ liệu thực nghiệm như vậy sẽ ảnh hưởng trực tiếp tỷ lệ được phân loại cũng như tác động đến số lượng các bài viết bị loại bỏ. Ví dụ với bài viết: *"Một ngày mùa hè ☺#Sam Son beach#, #my family#"*, có nội dung của bài viết chỉ có 4 từ *"Một ngày mùa hè"* nó có thể xếp vào chủ đề thời tiết, hoặc chủ đề xã hội, nhưng nếu phân tích thêm thẻ đánh dấu *#Sam Son beach#* thì bài viết này có liên quan đến chủ đề "*Du lịch*", phân tích thêm thẻ đánh dấu *#my family#* thì bài viết này liên

quan đến chủ đề "*Gia đình*", ... Qua đó, có thể thấy rằng, nếu chỉ phân tích phần nội dung của bài viết sẽ không phát hiện được hết các chủ đề quan tâm của người dùng thể hiện qua bài viết này. Hoặc như phân loại các bài viết theo biểu hiện cảm xúc (emotion) mà [152] nghiên cứu thì không thể xác định được các chủ đề mà người dùng thực sự quan tâm, chỉ xác định được mức độ quan tâm dựa trên cảm xúc thông qua 6 biểu tượng cảm xúc đã nghiên cứu dựa trên khung cảm xúc của Paul Ekman đề xuất. Hoặc như các nghiên cứu trên thẻ đánh dấu trong [125] và [145] thì đối với các bài viết không có thẻ đánh dấu sẽ bị loại bỏ, hoặc các nghiên cứu đó chỉ phù hợp với các mạng xã hội sử dụng thẻ đánh dấu, còn các mạng xã hội khác lại không phù hợp. Hoặc như nghiên cứu [77] trích chọn quan tâm của người dùng dựa trên các nội dung bài viết và số lần thích của người dùng. Nếu bài đăng không có nội dung hoặc số lần thích của các bài đăng là giống nhau thì các kết quả phân tích không phân biệt được mức độ quan tâm cũng như không đưa ra được các chủ đề quan tâm của người dùng. Hoặc nghiên cứu của [63] xác định các chủ đề dựa trên các thẻ đánh dấu và nội dung của các tweet trên mạng xã hội Twitter.com, 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. Tuy nhiên, cách phân tích này chỉ xếp mỗi bài đăng của người dùng vào một chủ đề mà không xét đến trường hợp, mỗi bài đăng của người dùng có thể liên quan đến nhiều chủ đề khác nhau. Điều này có thể gây hạn chế khi ứng dụng trong quảng cáo, khuyến nghị sản phẩm hay đưa ra các chủ đề quan tâm của người dùng.

Qua đó có thể thấy rằng, việc phân tích chỉ có nội dung bài viết, chỉ có thẻ đánh dấu, hoặc cảm xúc, hoặc các hành vi đơn lẻ như thích, theo dõi, ... có thể dẫn đến những thiếu sót khi phân tích tổng quát, hoặc dữ liệu thu thập được sẽ không đáp ứng được các yêu cầu trong các ứng dụng, hoặc không áp dụng được cho nhiều mạng xã hội khác nhau, hoặc mô hình không xác định được chính xác toàn bộ các chủ đề mà người dùng thực sự quan tâm. Vì vậy, với mục tiêu đưa ra được một đối tượng nghiên cứu nhằm cải thiện được các hạn chế đó, luận án đề xuất mô hình bài viết với nhiều đặc trưng có thể áp dụng cho nhiều nghiên cứu khác nhau trên các mạng xã hội khác nhau đặc biệt trong bài toán phát hiện quan tâm của người dùng trên các mạng xã hội.

Mô hình biểu diễn bài viết được luận án đề xuất bao gồm năm đặc trưng: nội dung, thể loại, thẻ đánh dấu, cảm xúc và quan điểm để phân loại bài viết theo các mức độ quan tâm đến các chủ đề. Các đặc trưng nội dung, thẻ đánh dấu có thể được thu thập trực tiếp từ các bài viết công khai của người dùng, cảm xúc có thể phân tích trực tiếp hoặc gián tiếp từ nội dung bài viết, quan điểm và thể loại có thể phát hiện từ nội dung bài viết và thẻ đánh dấu. Với lựa chọn này, luận án có thể sử dụng các thuật toán phân tích dữ liệu văn bản để ước lượng và hạn chế được những vấn đề như dữ liệu không đầy đủ, dữ liệu không hoàn chỉnh, hoặc dữ liệu bị thiếu hoặc rời rạc.

3.2. MÔ HÌNH BÀI VIẾT MỞ RỘNG

3.2.1. Mô hình bài viết

Trong định nghĩa 2.4 luận án đề cập đến bài viết được biểu diễn dựa trên nội dung, tuy nhiên với những hạn chế và các lý do đã trình bày trong mục 3.1, luận án mở rộng cách thức biểu diễn bài viết của người dùng trên mạng xã hội dựa trên năm đặc trưng gồm nội dung, thể loại, thẻ đánh dấu, quan điểm và cảm xúc. Như trong chương 2 đã định nghĩa, bài viết của người dùng trên các mạng 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 mạng xã hội có thể là một video clip, một hoặc một số bức ảnh, một văn bản, hoặc một sự kết hợp những thành phần này. Khi đó, một bài viết mở rộng có thể định nghĩa:

Định nghĩa 3.1:

Một bài viết $e_i \in E$ *trên mạng xã hội* $\mathcal{N}$ *được biểu diễn bởi năm đặc trưng:* $e_i = \{cont_i, cat_i, tag_i, sent_i, emo_i\}$. *Trong đó:*

- $cont_i$ *là nội dung (content) của bài viết* $e_i \in E$,
- cat_i *là thể loại (category) của bài viết* $e_i \in E$,
- tag_i *là thẻ đánh dấu (tag) của bài viết* $e_i \in E$,
- $sent_i$ *là quan điểm (sentiment) của bài viết* $e_i \in E$,
- emo_i *là cảm xúc (emotion) trong bài viết* $e_i \in E$.

Như vậy, mỗi bài viết $e_i \in E$ trên mạng xã hội $\mathcal{N}$, được biểu diễn bởi năm đặc trưng là nội dung, thể loại, thẻ đánh dấu, quan điểm và cảm xúc. Các đặc trưng của bài viết được mô tả chi tiết như sau:

- *Nội dung (Content) của bài viết e_i ký hiệu là: $cont_i$.* Phần nội dung của bài viết trên thực tế có thể là một video clip, một hoặc một số bức ảnh, một văn bản hoặc một sự kết hợp giữa chúng. Trong phạm vi của luận án, đặc trưng nội dung được xác định là toàn bộ văn bản chứa trong bài viết của người dùng, nội dung là đặc trưng tường minh của bài viết. Vì vậy, nội dung của bài viết trong luận án có thể là một văn bản, một đoạn văn ngắn, một câu hoặc một thuật ngữ. Nếu trong trường hợp đặc trưng nội dung không chứa văn bản, luận án sẽ coi đặc trưng này không có hoặc không tồn tại trong bài viết đó, và giá trị của đặc trưng này được tính là rỗng.
- *Thể loại (Category) của bài viết e_i ký hiệu là: cat_i.* Thể loại hay nhóm của các bài viết có thể hiểu là các vấn đề được ẩn chứa trong các nội dung hoặc các thẻ đánh dấu. Trên mỗi mạng xã hội N, mỗi bài viết có thể liên quan đến một hoặc nhiều thể loại, tùy theo nội dung của bài viết hoặc sự phân loại của người dùng.
- *Thẻ đánh dấu (Tag) của bài viết e_i ký hiệu là: tag_i*. Mỗi bài viết $e_i \in E$ trên mạng xã hội *N*, có thể được gắn vào một hoặc một tập các thẻ đánh dấu, cũng có thể không chứa bất kỳ thẻ đánh dấu nào, đặc trưng thẻ đánh dấu của bài viết được xác định là phần văn bản nằm giữa các ký hiệu đặc biệt như dấu # #, @, ...
- *Quan điểm (Sentiment) của bài viết e_i ký hiệu là: $sent_i$.* Quan điểm chính là góc nhìn hay khía cạnh của vấn đề mà người dùng suy nghĩ đến, hoặc là cách xem xét và hiểu các sự vật, hiện tượng, sự kiện, các vấn đề của người dùng trên mạng xã hội. Quan điểm của bài viết trên mạng xã hội có thể là thể hiện sự đồng ý hay tích cực, sự không đồng ý hoặc tiêu cực, không ý kiến hay trung lập đối với các đối tượng, sự kiện, hiện tượng. Trong luận

án, giá trị của đặc trưng quan điểm của các bài viết được xem xét như trình bày trong Bảng 3.1, mỗi bài viết có thể có giá trị của đặc trưng quan điểm là tích cực, tiêu cực hoặc trung lập.

Bảng 3.1: Giá trị của đặc trưng quan điểm

STT	Giá trị	Diễn giải
1	Positive	Tích cực
2	Neutral	Trung lập
3	Negative	Tiêu cực

Bảng 3.2: Giá trị của đặc trưng cảm xúc

STT	Biểu tượng	Giá trị	Diễn giải	Nhóm
1		Enjoy	Vui vẻ	Tích cực
2		Happyfor	Hạnh phúc	Tích cực
3		Love	Yêu thương	Tích cực
4		Gratitude	Biết ơn	Tích cực
5		Admiration	Ngưỡng mộ	Tích cực
6		Pride	Tự hào	Tích cực
7		Hope	Mong chờ	Tích cực
8		Sad	Buồn	Tiêu cực
9		Sorry	Tiếc nuối	Tiêu cực
10		Fear	Sợ hãi	Tiêu cực
11		Regret	Hối tiếc	Tiêu cực
12		Disappointed	Thất vọng	Tiêu cực
13		Disgust	Ghê tởm	Tiêu cực
14		Angry	Tức giận	Tiêu cực
15		Confused	Bối rối	Trung lập
16		No Emotion	Không cảm xúc	Trung lập

• *Cảm xúc (Emotion) của bài viết* e_i *ký hiệu là:* emo_i. Cảm xúc của bài viết trên mạng xã hội là một hình thức thể hiện thái độ của người dùng đối với chủ đề trình bày trong bài viết, hoặc thái độ đối với sự vật, hiện tượng trên các mạng xã hội. Cảm xúc có nhiều loại: cảm xúc đạo đức,

cảm xúc thẩm mỹ, cảm xúc trí tuệ... Đặc điểm của của cảm xúc là có tính đối lập: yêu và ghét, ưa thích và không ưa thích, xúc động và dửng dưng... Các cảm xúc được xem xét trên các mạng xã hội hiện nay có rất nhiều trạng thái, tuy nhiên trong luận án chỉ xem xét 16 giá trị trong Bảng 3.2 được dùng chung trên các trang mạng xã hội phổ biến như facebook.com, twitter.com, instagram.com ...

Theo định nghĩa 3.1 và dựa trên các đặc trưng đã xem xét thì mỗi bài viết $e_i \in E$ có thể biểu diễn một cách hình thức như công thức (3.1):

$$e_i = (cont_i, cat_i, tag_i, sent_i, emo_i),\ i = 1,..n, \forall e \in E | \boldsymbol{\mathcal{N}} \qquad (3.1)$$

Để thực hiện các ước lượng và tính toán đối với các bài viết theo mô hình đã đề xuất, luận án thực hiện tính giá trị của các đặc trưng của bài viết như sau:

- Đặc trưng nội dung được xác định là phần nội dung văn bản trong mỗi bài viết, đặc trưng nội dung là tường minh, được xác định trực tiếp.
- Đặc trưng thẻ đánh dấu là phần văn bản có thể xác định trực tiếp từ bài viết thông qua các ký hiệu đặc trưng như ##, @ ...
- Các đặc trưng thể loại, quan điểm và cảm xúc không thể xác định trực tiếp từ bài viết hay nói cách khác ba đặc trưng này là các giá trị không tường minh. Vì vậy, luận án lựa chọn phương thức kế thừa một thuật toán học có giám sát đã có để xác định giá trị cho các đặc trưng không tường minh này. Các nhãn dùng để gán giá trị cho đặc trưng thể loại của bài viết được tính toán vào phương pháp thống kê đã trình bày trong Chương 2, các nhãn dùng để gán giá trị cho đặc trưng quan điểm của bài viết được trình bày trong Bảng 3.1, còn các nhãn dùng để gán giá trị cho đặc trưng cảm xúc của bài viết được luận án chuyển đổi dựa trên các biểu tượng cảm xúc và trình bày chi tiết trong Bảng 3.3 của luận án.

Hiện nay có rất nhiều thuật toán gãn nhãn văn bản theo hướng học có giám sát được giới thiệu và sử dụng trong các nghiên cứu liên quan đến dữ liệu văn bản, tuy

nhiên, với đặc trưng dữ liệu trên mạng xã hội có nhiều khác biệt với các bộ dữ liệu chuẩn như sự đa dạng trong ngôn ngữ, sự sai sót trong biểu diễn văn bản, nội dung văn bản thường ngắn... Luận án lựa chọn đã một số thuật toán sử dụng phương pháp thống kê, bởi vì một số lí do sau đây:

- Thứ nhất, nếu dùng phương pháp thống kê, luận án có thể dễ dàng thực hiện trên nhiều ngôn ngữ khác nhau cho các bộ dữ liệu thực khi thu thập dữ liệu từ các trang mạng xã hội khác nhau. Điều này giúp mô hình nghiên cứu gần như không phải thay đổi hay cập nhật lại trong quá trình thực nghiệm.
- Thứ hai, các thuật toán sử dụng phương pháp ngữ nghĩa thì các mô hình đề xuất khi thực thi đều phụ thuộc vào ngôn ngữ trong mô hình đề xuất, hoặc phải dựa vào các bản thể học (ontology) để thực hiện, trong khi đó, bản thể học cho Tiếng Việt thì chưa có nhiều và chưa có chuẩn chung.
- Cuối cùng, dữ liệu văn bản từ các bài đăng, các bình luận, các thẻ đánh dấu trên các mạng xã hội thường không đúng chuẩn ngữ pháp mà thường viết tắt, dùng từ lóng theo giới trẻ, thậm chí nhiều ngôn ngữ pha trộn trong cùng một đoạn văn bản. Do đó, việc áp dụng các phương pháp ngữ nghĩa sẽ gặp khó khăn hơn so với việc sử dụng các phương pháp thống kê. Vì vậy, trong phạm vi nghiên cứu của luận án này, các thuật toán theo phương pháp thống kê sẽ được tập trung xem xét để lựa chọn tính toán giá trị cho đặc trưng thể loại, quan điểm và cảm xúc của bài viết. Tuy nhiên, với mô hình đề xuất trong luận án, hoàn toàn có thể sử dụng một thuật toán phân lớp văn bản dựa theo tiếp cận ngữ nghĩa để ứng dụng.

Các thuật toán phân loại hay gán nhãn cho dữ liệu văn bản theo phương pháp học có giám sát với hướng tiếp cận thống kê có thể kể đến như thuật toán CNN, thuật toán MNB, thuật toán NB...

- *Thuật toán học sâu CNN* [11] [80]: Thuật toán này dựa trên mạng nơ-ron tích chập trong học sâu còn gọi Convolutional Neuron Network (CNN). Thuật toán CNN hiện đang được coi là xu hướng mới của lĩnh vực học máy,

thuật toán CNN đã được chứng minh khá hiệu quả trong bài toán phân loại, gán nhãn văn bản, đặc biệt văn bản ngắn. Các lớp cơ bản trong một mạng CNN bao gồm: Lớp tích chập (Convolutional), Lớp kích hoạt phi tuyến ReLU (Rectified Linear Unit), Lớp lấy mẫu (Pooling) và Lớp kết nối đầy đủ (Fully-connected), được thay đổi về số lượng và cách sắp xếp để tạo ra các mô hình huấn luyện phù hợp cho từng bài toán khác nhau;

- *Thuật toán dựa trên Word2Vec* [11] [80]: Thuật toán này tính điểm mỗi từ theo xác suất của từ đó xuất hiện trong các văn bản có nhãn hay không;
- *Thuật toán MNB* [5] [11] [80]: Đây là thuật toán Multinomial Naive Bayes được công bố năm 2014, dựa trên thuật toán Naive Bayes. Thuật toán MNB dựa trên đặc trưng là véctơ TF-IDF của văn bản để phân lớp.
- Ngoài ra, luận án cũng thử nghiệm với một số thuật toán phân lớp phổ biến như: Naive Bayes [80], Support Véctơ Machine [80], K-Nearest Neighbors [80], C4.5 ...

Việc lựa chọn thuật toán phù hợp cho dữ liệu là văn bản ngắn trên mạng xã hội được luận án thực hiện dựa trên thực nghiệm và trình bày chi tiết trong Phụ lục B của luận án. Các thuật toán học có giám sát được luận án so sánh với nhau dựa trên kết quả gán nhãn các bộ dữ liệu mẫu và bộ dữ liệu thực, thuật toán cho kết quả phù hợp nhất sẽ được dùng để lựa chọn cho mô hình tính toán của luận án. Sau đó, luận án tiến hành gán nhãn và gán các giá trị vào cho đặc trưng của tất cả các bài viết trong bộ dữ liệu mẫu thử nghiệm.

3.2.2. Biểu diễn bài viết bằng véctơ

Với mỗi bài viết $e_i = (cont_i, cat_i, tag_i, sent_i, emo_i) \in E$ trên mạng xã hội $\boldsymbol{\mathcal{N}}$ như Định nghĩa 3.1 và các đặc trưng của bài viết đã trình bày chi tiết trong mục 3.2.1, luận án thực hiện tính toán giá trị cho năm đặc trưng: nội dung, thể loại, thẻ đánh dấu, quan điểm và cảm xúc. Để biểu diễn bài viết dựa trên các đặc trưng, luận án sử dụng một véctơ gồm năm thành phần của năm đặc trưng. Các thành phần được phân tích như Định nghĩa 2.2.

Ký hiệu $\boldsymbol{E} = \{e_1, e_2, \dots, e_n\}$ là tập tất các các bài viết đang xét trên mạng xã hội $\boldsymbol{\mathcal{N}}$, khi đó theo Định nghĩa 2.2 ở Chương 2, luận án ký hiệu lần lượt:

- E_{cont} là tập tất cả các từ vựng khác nhau từng đôi một của đặc trưng nội dung của tất cả các bài viết trong E
- E_{cat} là tập tất cả các từ vựng khác nhau từng đôi một của đặc trưng thể loại của tất cả các bài viết trong E
- E_{tag} là tập tất cả các từ vựng khác nhau từng đôi một của đặc trưng thẻ đánh dấu của tất cả các bài viết trong E
- E_{sent} là tập tất cả các từ vựng khác nhau từng đôi một của đặc trưng quan điểm của tất cả các bài viết trong E
- E_{emo} là tập tất cả các từ vựng khác nhau từng đôi một của đặc trưng cảm xúc của tất cả các bài viết trong E

Khi đó, đặc trưng nội dung được xem là đoạn văn bản ngắn nên luận án sử dụng Định nghĩa 2.2. trong không gian các nội dung của bài viết ta có:

$$cont_i = \boldsymbol{v}_{cont} = (w_{i1}, w_{i2}, .. w_{iq}) \quad (3.2)$$

Trong đó, q là tổng số từ vựng khác nhau từng đôi một của đặc trưng nội dung của tất cả các bài viết đang xét sau khi thực hiện tiền xử lý $\boldsymbol{E_{cont}}$, $(w_{ik}), k = 1, .., iq$ tương ứng được tính theo Định nghĩa 2.1 ở Chương 2.

Đặc trưng thẻ đánh dấu được xác định là phần văn bản hoặc thuật ngữ sau dấu @ hoặc giữa dấu ## của bài viết $e_i \in E$. Giá trị của đặc trưng thẻ đánh dấu thường là chuỗi văn bản có dấu hoặc không dấu được viết liền nhau nên giá trị của chúng bằng một véctơ chứa tập hợp các ký tự như trong công thức (3.3):

$$tag_i = \boldsymbol{v}_{tag} = (w_{i1}, w_{i2}, .. w_{ip}) \quad (3.3)$$

Trong đó p là số từ của không gian thẻ đánh dấu $\boldsymbol{E_{tag}}$, $(w_{ik}), k = 1, .., ip$ tương ứng được tính theo Định nghĩa 2.1 ở chương 2.

Đặc trưng thể loại, quan điểm và cảm xúc là không tường minh nên luận án thực hiện việc xác định dựa trên việc gán nhãn theo nội dung hoặc các chuỗi văn bản theo biểu tượng cảm xúc đính kèm theo nội dung của bài viết. Những bài viết đã có giá trị của đặc trưng thể loại thì giá trị của chúng sẽ là thuật ngữ được xác định trực tiếp, tương tự một số bài viết đã có cảm xúc thì được xác định trực tiếp, còn những bài viết chưa xác định được giá trị của đặc trưng thể loại, hoặc cảm xúc sẽ được xác định gián tiếp bằng một thuật toán phân loại văn bản. Khi đó giá trị của đặc trưng thể loại được tính bằng:

$$cat_i = \boldsymbol{v}_{cat} = (w_{i1}, w_{i2}, .. w_{il}) \quad (3.4)$$

Trong đó l là số từ của không gian thể loại $\boldsymbol{E_{cat}}$, $(w_{ik}), k = 1, .., il$ tương ứng được tính theo Định nghĩa 2.1 ở chương 2.

Giá trị của đặc trưng cảm xúc là:

$$emo_i = \boldsymbol{v}_{emo} = (w_{i1}, w_{i2}, .. w_{ir}) \quad (3.5)$$

Trong đó r là số từ của không gian thể loại $\boldsymbol{E_{emo}}$, $(w_{ik}), k = 1, .., ir$ tương ứng được tính theo Định nghĩa 2.1 ở chương 2.

Giá trị của đặc trưng quan điểm là:

$$sent_i = \boldsymbol{v}_{sent} = (w_{i1}, w_{i2}, .. w_{it}) \quad (3.6)$$

Trong đó t là số từ của không gian thể loại $\boldsymbol{E_{sent}}$, $(w_{ik}), k = 1, .., it$ tương ứng được tính theo Định nghĩa 2.1 ở chương 2.

Như vậy, mỗi bài viết $e_i \in E$ trên mạng xã hội $\boldsymbol{\mathcal{N}}$, được mô hình hóa bởi năm đặc trưng nội dung, thể loại, thẻ đánh dấu, quan điểm và cảm xúc, được biểu diễn bởi một véctơ có năm thành phần như trong công thức (3.7).

$$e_i = \begin{cases} cont_i = \boldsymbol{v}_{cont} = (w_{i1}, w_{i2}, .. w_{iq}), \\ cat_i = \boldsymbol{v}_{cat} = (w_{i1}, w_{i2}, .. w_{ip}), \\ tag_i = \boldsymbol{v}_{tag} = (w_{i1}, w_{i2}, .. w_{il}), \\ sent_i = \boldsymbol{v}_{emo} = (w_{i1}, w_{i2}, .. w_{ir}), \\ emo_i = \boldsymbol{v}_{sent} = (w_{i1}, w_{i2}, .. w_{it}) \end{cases} \quad (3.7)$$

Ví dụ với bài viết, e ="*Khu vực nhà tớ sẽ làm nơi đầu tiên được xem Nhật thực toàn phần vào 21/8. Dân Khoa học và du lịch khắp nơi trên thế giới đến rất đông. Ra đường hôm nay toàn thấy biển báo chấp nhận tắc đường do nhật thực. Lần đầu tiên được ngắm nhật thực là năm lớp 12. Lúc ấy vừa ra khỏi trường thì trời tối sầm lại. ☺ #NhatThuc2018#*"

Khi đó, giá trị các đặc trưng của bài viết được tính như sau:

- Giá trị của đặc trưng "*Nội dung*" của bài viết là: "*Khu vực nhà tớ sẽ làm nơi đầu tiên được xem Nhật thực toàn phần vào 21/8. Dân Khoa học và du lịch khắp nơi trên thế giới đến rất đông. Ra đường hôm nay toàn thấy biển báo chấp nhận tắc đường do nhật thực. Lần đầu tiên được ngắm nhật thực là năm lớp 12. Lúc ấy vừa ra khỏi trường thì trời tối sầm lại*"

=> Sau khi thực hiện tiền xử lý, danh sách từ vựng và trọng số của bài viết e (đã sắp xếp theo thứ tự chữ cái) tương ứng là:*{chấp nhận; du lịch; đầu tiên; hôm nay; khắp nơi; khoa học; khu vực; nhật thực; thế giới; toàn phần; tối sầm}*

- Giá trị của đặc trưng "*Thể loại*" của bài viết được xác định là: *{Khoa; học; khoa học; công nghệ}*
- Giá trị của đặc trưng *"Thẻ đánh dấu*" là: *{Nhat; thuc; 2018}*
- Giá trị của đặc trưng "*Quan điểm*" là :{*tích; cực; tích cực}*
- Giá trị của đặc trưng "*Cảm xúc*": *{Vui; vui vẻ}*

Và véctơ biểu diễn của bài viết e khi đó tính theo Định nghĩa 2.1 có giá trị là:

$$\boldsymbol{e_i} = \begin{cases} cont_i = \boldsymbol{v}_{cont} = (0.071, 0.071, 0.143, .. 0), \\ cat_i = \boldsymbol{v}_{cat} = (2.146, 1.450, 2.146, .. 0), \\ tag_i = \boldsymbol{v}_{tag} = (1.670, 0.812, 2.430, 0, .. 0), \\ sent_i = \boldsymbol{v}_{emo} = (0.3974, 0.0703, 0.3974, .. 0), \\ emo_i = \boldsymbol{v}_{sent} = (2.665, 3.216, .. 0) \end{cases}$$

3.2.3. Độ tương tự giữa hai bài viết mở rộng

Trong mục này, luận án đề xuất mô hình ước lượng độ tương tự giữa hai bài viết trên mạng xã hội dựa vào việc tích hợp độ tương tự giữa các đặc trưng của bài viết.

Mục đích đề xuất độ tương tự giữa hai bài viết mở rộng để phân nhóm các bài viết theo độ đo tương tự, sau đó có thể tích hợp để phân nhóm người dùng theo các bài viết đã đăng trên mạng xã hội.

a. Mô hình ước lượng tổng quát

Giả sử có hai bài viết $e_i, e_j \in E$ trên mạng xã hội $\mathcal{N}$, khi đó, mỗi bài viết được biểu diễn bởi năm đặc trưng tương ứng với năm thành phần của hai véctơ $\boldsymbol{e_i}, \boldsymbol{e_j}$ tương ứng như trong công thức (3.7), khi đó độ tương tự của hai bài viết $e_i, e_j \in E$ chính là độ tương tự giữa hai véctơ biểu diễn chúng, và độ tương tự của hai bài viết được tính dựa trên độ tương tự của mỗi thành phần tương ứng của hai véctơ theo trọng số.

Như vậy, độ tương tự giữa hai bài viết $e_i, e_j \in E$ trên mạng xã hội $\mathcal{N}$ theo định nghĩa 3.1 được tính như sau:

$$s_{entry}(e_i, e_j) = w_{cont} * s_{cont}(cont_i, cont_j) + w_{cat} * s_{cat}(cat_i, cat_j) + w_{tag} * s_{tag}(tag_i, tag_j) + w_{sent} * s_{sent}(sent_i, sent_j) + w_{emo} * s_{emo}(emo_i, emo_j) \quad (3.8)$$

Trong đó,

- $w_{cont}, w_{cat}, w_{tag}, w_{sent}, w_{emo}$ lần lượt là trọng số trên các đặc trưng nội dung, thể loại, thẻ đánh dấu, quan điểm, và cảm xúc của bài viết, thỏa mãn điều kiện: $w_{cont} + w_{cat} + w_{tag} + w_{sent} + w_{emo} = 1$
- $s_{cont}(cont_i, cont_j)$ là độ tương tự theo đặc trưng nội dung của hai bài viết $e_i, e_j \in E$ trên mạng xã hội $\mathcal{N}$
- $s_{cat}(cat_i, cat_j)$ là độ tương tự trên đặc trưng thể loại của hai bài viết $e_i, e_j \in E$ trên mạng xã hội $\mathcal{N}$
- $s_{tag}(tag_i, tag_j)$ là độ tương tự trên đặc trưng thẻ đánh dấu của hai bài viết $e_i, e_j \in E$ trên mạng xã hội $\mathcal{N}$

- $s_{sent}(sent_i, sent_j)$ là độ tương tự trên đặc trưng quan điểm của hai bài viết $e_i, e_j \in E$ trên mạng xã hội $\boldsymbol{\mathcal{N}}$
- $s_{emo}(emo_i, emo_j)$ là độ tương tự trên đặc trưng cảm xúc của hai bài viết $e_i, e_j \in E$ trên mạng xã hội $\boldsymbol{\mathcal{N}}$

Các mục tiếp theo trong phần này, luận án sẽ trình bày chi tiết cách tính độ tương tự trên từng đặc trưng của các bài viết.

a. Ước lượng độ tương tự trên từng đặc trưng của bài viết

- *Độ tương tự trên đặc trưng nội dung:* Đặc trưng nội dung của bài viết là một văn bản ngắn, vì vậy, độ tương tự giữa hai đặc trưng nội dung được tính theo công thức (2.4):

$$s_{cont}(cont_i, cont_j) = sim_{cont}\left(\boldsymbol{v}_{cont_i}, \boldsymbol{v}_{cont_j}\right) = \frac{< \boldsymbol{v}_{cont_i}, \boldsymbol{v}_{cont_j} >}{\left\|\boldsymbol{v}_{cont_i}\right\| \times \left\|\boldsymbol{v}_{cont_j}\right\|} \tag{3.9}$$

- *Độ tương tự trên đặc trưng thể loại:* Đặc trưng thể loại của bài viết sau khi gán nhãn có giá trị là một văn bản ngắn, vì vậy, độ tượng tự giữa hai đặc trưng thể loại được tính theo công thức (2.4) như sau:

$$s_{cat}(cat_i, cat_j) = sim_{cat}\left(\boldsymbol{v}_{cat_i}, \boldsymbol{v}_{cat_j}\right) = \frac{< \boldsymbol{v}_{cat_i}, \boldsymbol{v}_{cat_j} >}{\left\|\boldsymbol{v}_{cat_i}\right\| \times \left\|\boldsymbol{v}_{cat_j}\right\|} \tag{3.10}$$

- *Độ tương tự trên đặc trưng thẻ đánh dấu:* Đặc trưng thẻ đánh dấu của bài viết là một văn bản ngắn, vì vậy, độ tượng tự giữa hai đặc trưng thẻ đánh dấu được tính theo công thức (2.4) như sau:

$$s_{tag}(tag_i, tag_j) = sim_{tag}\left(\boldsymbol{v}_{tag_i}, \boldsymbol{v}_{tag_j}\right) = \frac{< \boldsymbol{v}_{tag_i}, \boldsymbol{v}_{tag_j} >}{\left\|\boldsymbol{v}_{tag_i}\right\| \times \left\|\boldsymbol{v}_{tag_j}\right\|} \tag{3.11}$$

- *Độ tương tự trên đặc trưng quan điểm:* Đặc trưng quan điểm của bài viết sau khi gán nhãn là một văn bản ngắn, vì vậy, độ tượng tự giữa hai đặc trưng quan điểm được tính theo công thức (2.4) như sau:

$$s_{sent}(sent_i, sent_j) = sim_{sent}\left(\boldsymbol{v}_{sent_i}, \boldsymbol{v}_{sent_j}\right) = \frac{< \boldsymbol{v}_{sent_i}, \boldsymbol{v}_{sent_j} >}{\left\|\boldsymbol{v}_{sent_i}\right\| \times \left\|\boldsymbol{v}_{sent_j}\right\|} \quad (3.12)$$

- *Độ tương tự trên đặc trưng cảm xúc:* Đặc trưng cảm xúc của bài viết sau khi gán nhãn là một văn bản ngắn, vì vậy, độ tượng tự giữa hai đặc trưng cảm xúc c được tính theo công thức (2.4) như sau:

$$s_{emo}(emo_i, emo_j) = sim_{emo}\left(\boldsymbol{v}_{emo_i}, \boldsymbol{v}_{emo_j}\right) = \frac{< \boldsymbol{v}_{emo_i}, \boldsymbol{v}_{emo_j} >}{\left\|\boldsymbol{v}_{emo_i}\right\| \times \left\|\boldsymbol{v}_{emo_j}\right\|} \quad (3.13)$$

Trong các công thức (3.9), (3.10), (3.11), (3.12) và (3.13) các $\boldsymbol{v_{ki}}$ được tính theo công thức (3.7), dựa trên các công thức này, luận án thực hiện một thực nghiệm nhằm tìm kiếm bộ trọng số tốt nhất cho năm đặc trưng của bài viết khi ước lượng. Thực nghiệm lựa chọn bộ trọng số cho năm đặc trưng của bài viết được trình bày chi tiết trong mục 3.5.

3.3. MÔ HÌNH NGƯỜI DÙNG THEO BÀI VIẾT MỞ RỘNG

3.3.1. Biểu diễn người dùng theo bài viết mở rộng

Dựa trên mô hình bài viết có năm đặc trưng, mỗi người dùng trên mạng xã hội được biểu diễn bằng một tập các bài viết. Không mất tính tổng quát, giả sử rằng trên $E = \{e_1, e_2, \dots, e_p\}$ là tập các bài viết tương ứng của p người dùng $\boldsymbol{U} = \{u_1, u_2, \dots, u_p\}$. Ký hiệu mỗi một bài viết là một véctơ theo công thức (3.7), $e_i = \{\boldsymbol{e_{i1}}, \boldsymbol{e_{i2}}, .. \boldsymbol{e_{im_i}}\}$ là tập các bài viết của người dùng thứ *i*. Khi đó mỗi người dùng trên mạng xã hội $\boldsymbol{\mathcal{N}}$ được biểu diễn bởi một véctơ gồm m_i thành phần, mỗi thành phần là một véctơ được xây dựng theo công thức 3.7. Ký hiệu như sau:

$$u_i = \boldsymbol{u_i} = (\boldsymbol{e_{i1}}, \boldsymbol{e_{i2}}, .. \boldsymbol{e_{im_i}}) \tag{3.14}$$

Cụ thể mỗi người dùng trên mạng xã hội có thể được biểu diễn như sau:

$$u_i = \begin{pmatrix} \boldsymbol{e_{i1}} = \begin{cases} cont_{i1} = \boldsymbol{v}_{cont} = (w_{i1}, w_{i2}, .. w_{iq}), \\ cat_{i1} = \boldsymbol{v}_{cat} = (w_{i1}, w_{i2}, .. w_{ip}), \\ tag_{i1} = \boldsymbol{v}_{tag} = (w_{i1}, w_{i2}, .. w_{il}), \\ sent_{i1} = \boldsymbol{v}_{emo} = (w_{i1}, w_{i2}, .. w_{ir}) \ , \\ emo_{i1} = \boldsymbol{v}_{sent} = (w_{i1}, w_{i2}, .. w_{it}) \end{cases} \\ \dots \dots \\ \boldsymbol{e_{im_i}} = \begin{cases} cont_{i\boldsymbol{m_i}} = \boldsymbol{v}_{cont} = (w_{i1}, w_{i2}, .. w_{iq}), \\ cat_{i\boldsymbol{m_i}} = \boldsymbol{v}_{cat} = (w_{i1}, w_{i2}, .. w_{ip}), \\ tag_{i\boldsymbol{m_i}} = \boldsymbol{v}_{tag} = (w_{i1}, w_{i2}, .. w_{il}), \\ sent_{i\boldsymbol{m_i}} = \boldsymbol{v}_{emo} = (w_{i1}, w_{i2}, .. w_{ir}) \ , \\ emo_{i\boldsymbol{m_i}} = \boldsymbol{v}_{sent} = (w_{i1}, w_{i2}, .. w_{it}) \end{cases} \end{pmatrix}$$

Với *q, p, l, r, t* là số chiều của các không gian $E_{cont}, E_{cat}, E_{tag}, E_{sent}, E_{emo}$ trên mạng xã hội đang xem xét.

3.3.2. Độ tương tự giữa hai người dùng theo mô hình bài viết mở rộng

Để ước lượng độ tương tự giữa hai người dùng theo bài viết có năm đặc trưng , luận án thực hiện ước lượng độ tương tự giữa hai tập bài viết đã đăng của hai người dùng tương ứng. Giả sử có hai người dùng u_i và u_j với hai tập bài viết E_i và E_j tương ứng trên mạng xã hội *N*. Khi đó, độ tương tự giữa hai tập bài viết E_i và E_j được tính bằng độ tương tự giữa hai tập các véctơ trọng số tương ứng của u_i và u_j được tính như sau:

$$sim(\boldsymbol{E_i}, \boldsymbol{E_j}) = \max_{\text{ik,il}}(sim(\boldsymbol{e_{il}}, \boldsymbol{e_{jk}}))$$

Trong đó các $sim(\boldsymbol{e_{il}}, \boldsymbol{e_{jk}})$ được tính theo công thức (3.8). Khi đó độ tương tự của hai người dùng được tính bằng:

$$sim(u_i, u_j) = sim(\boldsymbol{u_i}, \boldsymbol{u_j}) = sim(\boldsymbol{E_i}, \boldsymbol{E_j}) \tag{3.15}$$

3.4. QUAN TÂM CỦA NGƯỜI DÙNG THEO MÔ HÌNH BÀI VIẾT MỞ RỘNG

3.4.1. Biểu diễn bài viết theo chủ đề

Tương tự như biểu diễn theo không gian chủ đề đã trình bày trong Chương 2, luận án sử dụng tập hợp gồm từ vựng của 21 chủ đề đã xây dựng để xây dựng trọng số cho năm đặc trưng của bài viết. Khi đó, giả sử rằng $\boldsymbol{\mathcal{T}} = \{T_1, T_2, \dots, T_q\}$ là một tập các chủ đề trên mạng xã hội N, trong đó mỗi chủ đề được biểu diễn bằng một tập các thuật ngữ hoặc các từ $T_i = \{t_{i1}, t_{i2}, \dots, t_{iq_i}\}$. Véctơ trọng số của các bài viết được tính dựa trên công thức (2.9) như sau:

Gọi $e_{ij} \in E_i$ là một bài viết của người dùng u_i trên mạng xã hội $\boldsymbol{\mathcal{N}}$, được mô tả bởi năm đặc trưng, mỗi đặc trưng là 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:

$$\boldsymbol{e_{ij}^{k}} = \left(e_{ij}^{1}, e_{ij}^{2}, \dots, e_{ij}^{t_{kp}}\right) \tag{3.16}$$

Trong đó, $e_{ij}^{l} = w_k * tf(t_{il}, e_{ij}) \times idf(t_{il}, E_i)$ với $t_{il} \in \mathcal{V}_T$, $w_k, k = 1,..5$ là trọng số của các đặc trưng tương ứng của bài viết.

3.4.2. Xác định mối tương quan giữa người dùng và các chủ đề

Dựa trên công thức (2.10) và (2.11), luận án biểu diễn mức độ quan tâm của người dùng theo các chủ đề trên mạng xã hội chính bằng mức độ liên quan của tập tất cả các bài viết của người dùng đó đối với các chủ đề, 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:

$$\alpha_{ij}^{k} = cor\left(e_{ij}, t_k\right) \tag{3.17}$$

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

$$cor\left(e_{ij}, \boldsymbol{\mathcal{T}}\right) = (\alpha_{ij}^{1}, \alpha_{ij}^{2}, \dots, \alpha_{ij}^{q}) \tag{3.18}$$

Khi đó mức độ quan tâm của người dùng được xác định theo Định nghĩa 2.9, và hàm $int(u_i, t)$ được tính dựa theo một trong ba công thức (2.12), (2.13) hoặc (2.14) chỉ khác là các e_{ij} được tính như công thức (3.16)

3.4.3. Độ tương tự quan tâm của người dùng theo chủ đề

Dựa trên mức độ quan tâm của người dùng theo các chủ đề, mỗi người dùng được biểu diễn bởi một véctơ quan tâm như sau:

$$\boldsymbol{u_i^t} = (u_i^1, u_i^2, \dots, u_i^p) \tag{3.19}$$

Trong đó, mỗi u_i^k là mức độ quan tâm của người dùng u_i đến chủ đề thứ k trong tập $\boldsymbol{\mathcal{T}}$**.** Khi đó, độ tương tự của hai người dùng theo các chủ đề cũng được tính dựa trên công thức (2.16) như sau:

$$sim_{int}(u_i, u_j) = sim(\boldsymbol{u_i^t}, \boldsymbol{u_j^t}) = \frac{< \boldsymbol{u_i^t}, \boldsymbol{u_j^t} >}{\| \boldsymbol{u_i^t}\| \times \|\boldsymbol{u_j^t}\|} \tag{3.20}$$

3.5. TƯƠNG QUAN GIỮA TƯƠNG TỰ NGƯỜI DÙNG VÀ QUAN TÂM

3.5.1. Bài toán xác định tương quan giữa tương tự người dùng và chủ đề

Xét bài toán sau:

Cho một tập người dùng U trên mạng xã hội $\boldsymbol{\mathcal{N}}$*. Gọi SimU là tập những người dùng tương tự nhau dựa trên nội dung các bài viết và CorrU tập người dùng tương tự nhau theo chủ đề.*

Để xem xét mối tương quan giữa hai tập *SimU và CorrU,* luận án tiến hành hai thực nghiệm trình bày trong mục 3.5.2. Luận án so sánh và thống kê để tìm giao của hai nhóm SimU và CorrU. Nếu có người dùng xuất hiện trong cả hai nhóm thì có thể kết luận rằng SimU và CorrU có sự tương quan với nhau. Trong chương 2 luận án đã tìm thấy sự tương quan khi chỉ xét trên đặc trưng nội dung. Dựa trên những phân tích ở mục 3.1, luận án thực hiện thực nghiệm dựa trên mô hình xem xét bài viết với năm

đặc trưng và các phương thức tính toán đã trình bày trong các mục 3.2, 3.3 và 3.4 để so sánh liệu rằng với mô hình biểu diễn bài viết bằng năm đặc trưng nội dung, thể loại, thẻ đánh dấu, quan điểm và cảm xúc, liệu có thu được kết quả tốt hơn? Ngoài ra, do mô hình bài viết được đề xuất có năm đặc trưng, trên thực tế các đặc trưng này có các trọng số khác nhau khi tính toán, vì vậy, luận án thực hiện thêm một thực nghiệm dựa trên độ tương tự giữa hai bài viết để xác định bộ trọng số tối ưu cho các đặc trưng của bài viết

- Thực nghiệm ước lượng độ tương tự giữa hai người dùng theo mô hình bài viết mở rộng để tìm tập SimU
- Thực nghiệm xác định độ tương quan theo các chủ đề quan tâm của người dùng dựa trên bài viết mở rộng với các chủ đề để tìm tập CorrU
- Thực nghiệ xác định bộ trọng số tối ưu cho năm đặc trưng cho mô hình bài viết mở rộng

3.5.2. Thực nghiệm và đánh giá

a. Xây dựng bộ dữ liệu thử nghiệm

Luận án xây dựng bộ dữ liệu thử nghiệm bằng dữ liệu thực tế thu được từ mạng xã hội Facebook.

Bảng 3.3: Mô tả bộ dữ liệu thực nghiệm

		Bộ dữ liệu tính độ tương tự (Theo Entry)	Bộ dữ liệu tính độ tương quan (Theo Topic)
Người dùng (User)		200	200
Bài viết (Entry)	Nội dung (Cont)	2000	2000
	Thể loại (Cat)	2000	2000
	Thẻ đánh dấu (Tag)	2000	2000
	Quan điểm (Sent)	2000	2000
	Cảm xúc (Emo)	2000	2000
Chủ đề (Topic)		0	21
Trọng số (Weight)		TF.IDF	TF.IDF
Biểu diễn		Theo không gian năm đặc trưng của bài viết	Theo không gian chủ đề

Bao gồm 200 người dùng và 2000 bài viết, mỗi bài viết ngoài đặc trưng nội dung đã tính toán trong chương 2, luận án thực hiện thu thập và phân tích thêm các đặc trưng thẻ đánh dấu, còn ba đặc trưng thể loại, quan điểm và cảm xúc được tính toán vào bước tiền xử lý.

b. Lựa chọn bộ trọng số của các đặc trưng của bài viết

Để tính toán và đưa ra bộ trọng số phù hợp cho các đặc trưng của bài viết trong mô hình đề xuất, luận án thử nghiệm ước lượng độ tương tự giữa hai bài viết dựa trên một bộ mẫu gồm 500 mẫu xây dựng theo không gian bài viết mở rộng.

Mỗi mẫu trong bộ dữ liệu thực nghiệm có chứa 3 thành phần: Các id của mẫu; Giá trị của mẫu, giá trị có thể bằng 1 hoặc 2; Mỗi mẫu chứa ba bài viết thu thập từ Twitter.com, các bài viết này được gán nhãn là A, B, và C.

Bảng 3.4: Một mẫu minh họa trong bộ mẫu thực nghiệm

ID	354
Value	2
A	Một hành động đẹp tôi yêu iOS 5 @apple #iPhone
B	@AsimRang @apple @umber các ứng dụng cho máy bàn đang bị bỏ qua
C	Cám ơn @apple trong việc tìm lại chiếc Mac của tôi – chỉ việc định vị và lướt tìm cái máy bị mất

Giá trị của mẫu được xác định như sau:

Nếu bài viết A tương tự với bài viết B nhiều hơn bài viết C, thì giá trị của mẫu này bằng 1. Ngược lại, nếu bài viết A tương tự với bài viết C nhiều hơn bài viết B thì giá trị của mẫu này bằng 2.

Trong quá trình thực nghiệm, luân án đã xây dựng và sử dụng 500 mẫu, ví dụ, một mẫu được trình bày trong Bảng 3.4, trong mẫu này, bài viết A tương tự với bài viết C nhiều hơn bài viết B, vì vậy giá trị của mẫu bằng 2.

Chi tiết được thực hiện lần lượt theo tổ hợp các đặc trưng được thống kê trong Bảng 3.5 gồm có: 1 đặc trưng có 5 tổ hợp; 2 đặc trưng có 10 tổ hợp; 3 đặc trưng có 10 tổ hợp; 4 đặc trưng có 5 tổ hợp và cuối cùng 5 đặc trưng có 1 tổ hợp.

Bảng 3.5: Các tổ hợp khảo sát chọn bộ trọng số

Số đặc trưng	Số tổ hợp	Tổ hợp các đặc tính
1/5	5	Nội dung Chủ đề Thẻ đánh dấu Quan điểm Cảm xúc
2/5	10	Nội dung - Chủ đề Nội dung - Thẻ đánh dấu Nội dung - Quan điểm Nội dung - Cảm xúc Chủ đề - Thẻ đánh dấu Chủ đề - Quan điểm Chủ đề - Cảm xúc Thẻ đánh dấu - Quan điểm Thẻ đánh dấu - Cảm xúc Quan điểm - Cảm xúc
3/5	10	Nội dung - Chủ đề - Thẻ đánh dấu Nội dung - Chủ đề - Quan điểm Nội dung - Chủ đề - Cảm xúc Chủ đề - Thẻ đánh dấu - Quan điểm Chủ đề - Thẻ đánh dấu - Cảm xúc Chủ đề - Quan điểm - Cảm xúc Thẻ đánh dấu - Quan điểm - Cảm xúc Nội dung, Thẻ đánh dấu, Quan điểm Nội dung, Thẻ đánh dấu, Cảm xúc
4/5	5	Nội dung - Chủ đề - Thẻ đánh dấu - Quan điểm Nội dung - Chủ đề - Thẻ đánh dấu - Cảm xúc Nội dung - Thẻ đánh dấu - Quan điểm - Cảm xúc Nội dung – Chủ đề - Quan điểm - Thẻ đánh dấu Chủ đề - Thẻ đánh dấu – Quan điểm - Cảm xúc
5/5	1	Nội dung – Chủ đề - Thẻ đánh dấu – Quan điểm - Cảm xúc

Luận án thực nghiệm chọn trọng số phù hợp như sau:

Bước 1: Thực hiện việc chạy trên toàn bộ mẫu thử nghiệm nhiều lần, mỗi lần chỉ với một tổ hợp đặc trưng của bài viết mở rộng được kết hợp trong Bảng 3.5.

Bước 2: Lưu toàn bộ các giá trị của các lần thử nghiệm, tính độ chính xác CR, với mỗi bộ kết hợp các đặc tính này, luận án chọn bộ trọng số phù hợp nhất cho các tổ hợp minh họa trong Bảng 3.6

Bảng 3.6: Khảo sát và lựa chọn bộ trọng số ước lượng

Chiến lược	w_1 Nội dung	w_2 Thể loại	w_3 Thẻ đánh dấu	w_4 Quan điểm	w_5 Cảm xúc	**#mẫu đúng** (Tổng 500 mẫu)	**Tỉ lệ (%)**
Chỉ dùng 1/5 đặc trưng					1.00	140	28.00
		1.00				148	29.60
				1.00		216	43.20
			1.00			283	56.60
	1.00					***345***	***69.00***
Sử dụng 2/5 đặc trưng			0.50		0.50	239	47.80
				0.30	0.70	254	50.80
			0.75	0.25		303	60.60
		0.50	0.50			318	63.60
	0.50		0.50			358	71.60
		0.60			0.40	366	73.20
	0.90			0.10		373	74.60
	0.75				0.25	378	75.60
		0.60		0.40		392	78.40
	0.70		***0.30***			***416***	***83.20***
Sử dụng 3/5 đặc trưng			0.60	0.10	0.30	333	66.60
	0.50			0.05	0.45	386	77.20
		0.40	0.30		0.30	387	77.40
	0.40		0.40	0.20		390	78.00
	0.35		0.35		0.30	392	78.40
		0.40	0.40	0.20		410	82.00
		0.40		0.35	0.25	414	82.80
	0.45	0.15	0.40			421	84.20
	0.60	0.30		0.10		446	89.20
	0.60	***0.10***			***0.30***	***450***	***90.00***
Sử dụng 4/5 đặc trưng	0.35		0.30	0.05	0.30	406	81.20
		0.30	0.35	0.10	0.25	432	86.40
	0.50	0.25	0.15	0.10		450	90.00
	0.35	0.20	0.25		0.20	454	90.80
	0.35	***0.25***		***0.10***	***0.30***	***460***	***92.00***
5/5 đặc trưng	***0.35***	***0.15***	***0.25***	***0.05***	***0.20***	***465***	***93.00***
Trọng số phù hợp nhất	**0.35**	**0.15**	**0.25**	**0.05**	**0.20**	**465**	**93.00**

Trong đó các đặc trưng tương ứng với các cột thứ 2 đến thứ 6 của Bảng 3.6. Cột thứ 7 là số mẫu đúng thu được so với kết quả của các cộng tác viên đánh giá. Cột cuối cùng là tỉ lệ được tính bằng số mẫu đúng trên tổng số mẫu của thực nghiệm.

Để đánh giá các kết quả, luận án sử dụng tỷ lệ chính xác trên bộ mẫu đã cho, được tính như sau: $\mathrm{CR} = \frac{\text{Số lượng các mẫu đúng}}{\text{Tổng số các mẫu}} * 100\%$

Như vậy, nếu dùng 5 đặc trưng như mô hình đề xuất thì chỉ có 1 bộ kết hợp duy nhất, bộ này cho kết quả phù hợp nhất với bộ trọng số tương ứng các đặc trưng <*nội dung, thể loại, thẻ đánh dấu, quan điểm, cảm xúc*> ~ < ***0.35, 0.15, 0.25, 0.05, 0.20***>, sự kết hợp này cũng cho kết quả phù hợp nhất trên bộ mẫu dữ liệu thực nghiệm với độ chính xác đến 93.00%.

c. Xác định độ tương tự của hai người dùng theo bài viết mở rộng

- **Kịch bản thực nghiệm:** Kịch bản xác định độ tương tự của hai người dùng dựa trên mô hình bài viết mở rộng được thực hiện như sau:

Đầu vào: Danh sách 2000 bài viết của 200 người dùng

Đầu ra: Độ tương tự giữa các cặp người dùng theo các bài viết có năm đặc trưng

Thực hiện

- Tính giá trị và xây dựng bộ thuật ngữ cho các đặc trưng của các bài viết

- Tính véctơ trọng số theo TF.IDF cho mỗi đặc trưng của bài viết theo các đặc trưng của bài viết

- Tính độ tương tự dựa trên tích hợp có trọng số giữa các cặp bài viết mở rộng

- Phân loại mức độ tương tự của các cặp người dùng theo tập bài viết mở rộng

Kết thúc

- **Tham số đầu ra:** Đầu ra là độ tương tự của mỗi cặp người dùng dựa trên tập các bài viết mở rộng.
- **Các bước thực hiện chi tiết như sau:**

Bước 1 và Bước 2 được thực hiện như trong thực nghiệm trong mục 3.3.1, các véctơ tương ứng được tính trong không gian các đặc trưng của bài viết.
Bước 3 tính độ tương tự giữa các cặp bài viết dựa trên công thức (3.8) với bộ trọng số lấy từ thực nghiệm ở mục 3.5.2.b.
Bước 4 tính độ tương tự giữa các cặp người dùng dựa trên Định nghĩa 2.7 theo công thức (3.18), sau khi phân loại theo mức độ tương tự được minh họa như trong **Bảng 3.7**

Đây là kết quả thực nghiệm thu được dựa trên các mức tương tự theo không gian bài viết mở rộng, so sánh với số cặp người dùng tương tự nhau khi chỉ xét trên thuộc tính nội dung.

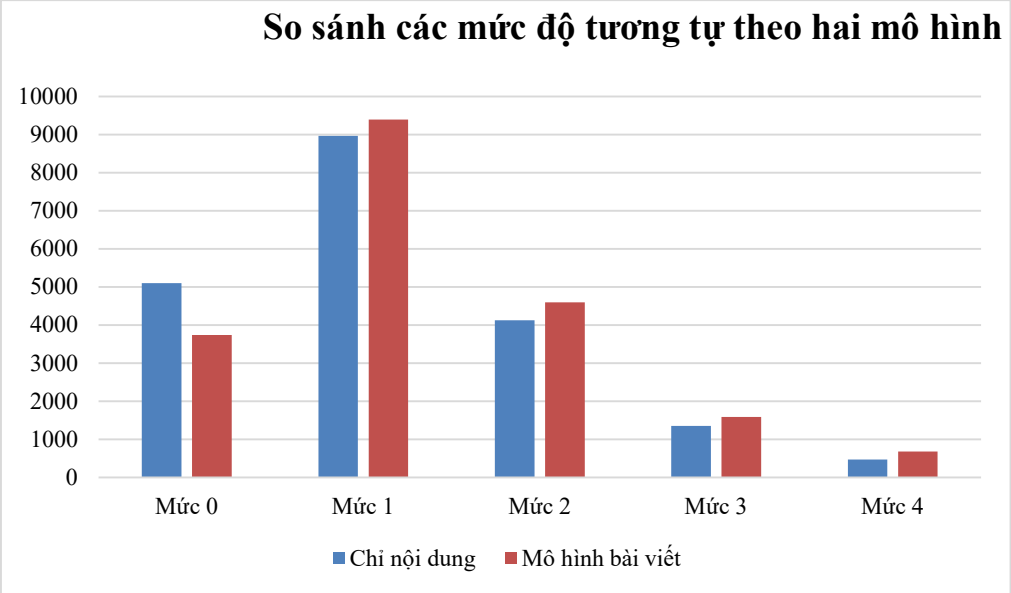
Bảng 3.7: Nhóm các cặp người dùng tương tự theo không gian bài viết

	Mức 0	**Mức 1**	**Mức 2**	**Mức 3**	**Mức 4**
Bài viết chỉ có nội dung	5094	8961	4122	1355	468
Bài viết có 5 đặc trưng	4123	9350	4437	1469	621

Kết quả so sánh trong **Bảng 3.7** có thể thấy rằng độ tương tự theo không gian bài viết của các cặp người dùng theo mô hình bài viết mở rộng có năm đặc trưng.

Có thể thấy rõ hơn trong **Hình 3.1** giảm ở mức 0 và tăng các mức còn lại, đặc biệt số lượng các cặp có mức tương tự nhau ở mức 4 đã tăng lên rất đáng kể. Nghĩa là nếu số lượng các đặc trưng tăng lên thì số lượng các bài viết tương tự nhau tăng lên và các cặp người dùng tương tự nhau cũng tăng lên.

Trong quá trình thực nghiệm luận án có so sánh mô hình biểu diễn người dùng dựa trên bài viết mở rộng với mô hình của Buscaldi et al. [41] chỉ tính đến đặc tính nội dung của các bài viết, và mô hình đề xuất của luận án trong hai trường hợp, chỉ có 3 đặc trưng (nội dung, thể loại và đánh dấu), có đủ 5 đặc trưng (nội dung, thể loại, đánh dấu, quan điểm và cảm xúc).



Hình 3.1: So sánh độ tương tự giữa hai người dùng

Mô hình của Buscaldi et al. được dùng để so sánh hai người dùng dựa trên nội dung các câu hỏi và các đoạn văn trong các biểu mẫu khi hỏi, phương pháp của Buscaldi et al. có ưu điểm là không phụ thuộc vào ngôn ngữ. Phương pháp được thực hiện bằng cách tách từ, tính TF.IDF sau đó xếp loại các từ trong câu và so sánh chúng để tìm độ tương tự giữa các câu hỏi hoặc đoạn văn điền trong biểu mẫu của bảng hỏi.

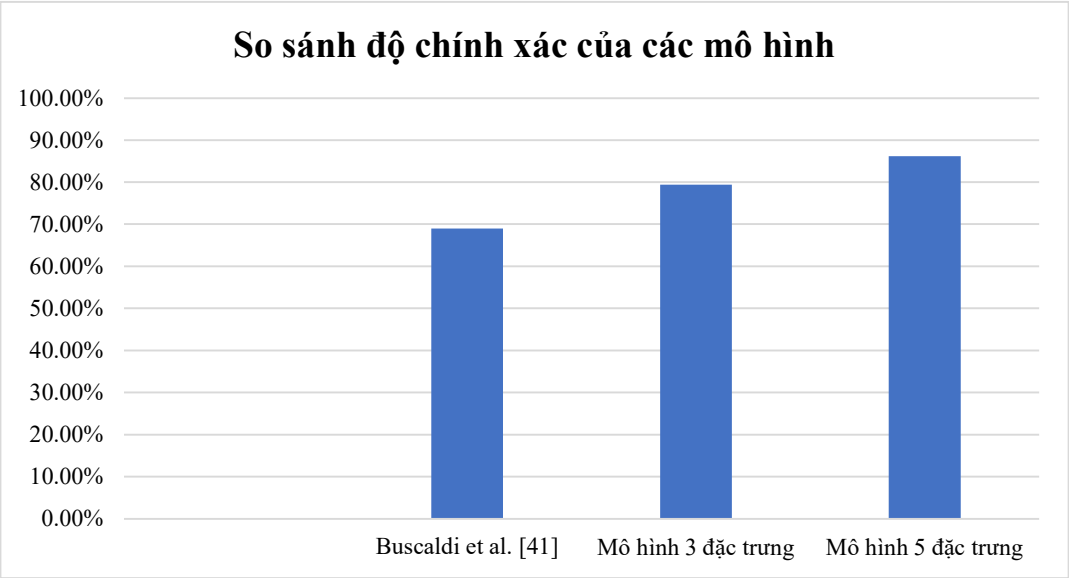
Cùng bộ dữ liệu, luận án sử dụng kết quả phân tách nội dung bài viết và so sánh với mô hình đề xuất kết quả như trong Bảng 3.8 như sau:

Bảng 3.8: Kết quả thực nghiệm so sánh với mô hình khác

Mô hình	**CR**	**Các đặc trưng đưa vào tính toán**				
		w_1 Nội dung	w_2 Thể loại	w_3 Thẻ đánh dấu	w_4 Quan điểm	w_5 Cảm xúc
Buscaldi et al. [41]	69.00%	w_1				
Mô hình 3 đặc trưng	79.40%	w_1	w_2	w_3		
Mô hình 5 đặc trưng	86.20%	w_1	w_2	w_3	w_4	w_5

Dựa vào bảng 3.7 có thể thấy rằng nếu chỉ sử dụng một đặc trưng thì kết quả xác định độ tương tự giữa hai người dùng có độ chính xác thấp nhất, thêm đặc trưng thể loại và đặc trưng thẻ đánh dấu thì độ chính xác tăng hơn 10%. Cuối cùng nếu xét

thêm đặc trưng quan điểm và cảm xúc thì độ chính xác tăng lên thành 86.20%. Các kết quả so sánh được minh họa trong Hình 3.2.



Hình 3.2: So sánh độ chính xác của các mô hình

d. Xác định mức độ quan tâm của người dùng theo chủ đề

- **Kịch bản thực nghiệm:** Kịch bản thực hiện như sau:

Đầu vào: Danh sách 200 người dùng với 2000 bài viết và 21 chủ đề dùng làm nhãn để phân loại

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

Thực hiện

- Xây dựng danh sách từ, thuật ngữ cho các bài viết và các chủ đề

- Tính véctơ trọng số theo TF.IDF cho mỗi bài viết và mỗi chủ đề

- Tính độ tương quan giữa mỗi bài viết và các chủ đề

- Phân loại các bài viết theo 21 chủ đề dựa trên độ đo tương quan

Kết thúc

- **Tham số đầu ra:** Đầu ra của kịch bản là độ tương quan của mỗi người dùng với 21 chủ đề được xem xét

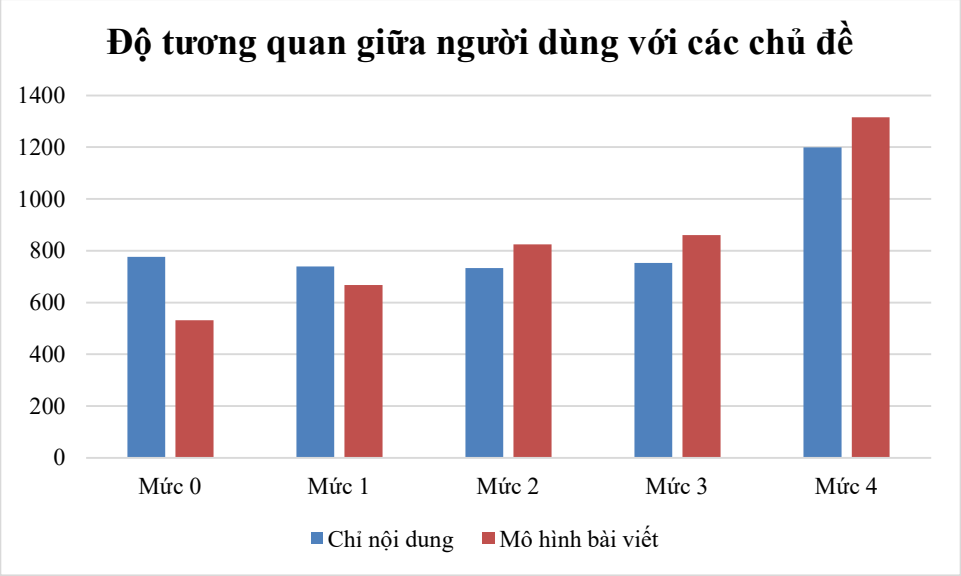
Bảng 3.9: Phân loại theo các mức quan tâm của người dùng với các chủ đề

STT	Chủ đề	Mức 0	Mức 1	Mức 2	Mức 3	Mức 4
1	Tài chính - Kinh doanh	16	26	38	40	80
2	Thế giới - Quốc tế	8	39	37	58	58
3	Thời sự - Tin tức	14	45	47	59	35
4	Văn hóa - Giải trí	17	21	28	29	105
5	Khoa học - Công nghệ	22	25	26	27	100
6	Sức khỏe	22	27	19	39	93
7	Thể thao	9	9	37	38	107
8	Đời sống – xã hội	8	27	31	46	88
9	Chính trị	55	42	44	20	39
10	Giáo dục	25	42	45	36	52
11	Pháp luật – Nhà nước	31	27	39	43	60
12	Du lịch	34	42	37	34	53
13	Đánh giá sản phẩm	47	47	31	45	30
14	Gia đình	33	31	37	46	53
15	Con người	26	44	65	38	27
16	Góc nhìn	32	26	39	33	70
17	Làm đẹp – Thời trang	7	26	49	52	66
18	Nhịp sống trẻ	21	27	55	38	59
19	Môi trường	32	25	44	32	67
20	Khám phá	37	37	45	56	25
21	Khác	36	33	32	51	48
Tổng		**532**	**668**	**825**	**860**	**1315**

- **Các bước thực hiện chi tiết như sau:** Bước 1 và bước 2 xây dựng danh sách từ, thuật ngữ cho bài viết và tính véctơ trọng số theo TF.IDF theo không gian của chủ đề có tính trọng số của các thuật ngữ theo các đặc trưng. Bước 3, tính độ tương quan giữa các bài viết với các chủ đề, bước 4 tính và phân loại mức độ quan tâm của người dùng theo các chủ đề dựa trên Định nghĩa 2.6 và các công thức (2.11), (2.12) và (2.13). Tại bước này, luận án tính mỗi công thức sẽ thu được một bảng gồm có 200 dòng và 21 cột, trong đó mỗi ô là mức độ quan tâm của người dùng đó đến các

chủ đề tương ứng. Kết quả thực nghiệm thu được dựa trên các mức tương quan theo không gian chủ đề các nhóm như minh họa trong **Bảng 3.9**

So sánh giữa **Bảng 2.24** với **Bảng 3.9** có thể thấy rằng các mức 0 và mức 1 đều giảm, và các mức liên quan ở mức 2, mức 3 đều tăng nhẹ, mức độ quan tâm ở mức 4 tăng lên rất đáng kể. Có thể nhìn thấy rõ ràng hơn trong **Hình 3.3**



Hình 3.3: So sánh mức độ tương quan giữa người dùng và chủ đề

3.5.3. Thảo luận về kết quả thực nghiệm

Độ chính xác của thực nghiệm xác định chủ đề quan tâm của người dùng được tính dựa trên *Sai số bình phương trung bình (*MSE - Mean Square Error) như công thức (2.19). Với cách tính này thì MSE càng gần đến giá trị 0 thì độ chính xác càng cao và ngược lại. Kết quả:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2 = \frac{1}{4200}\sum_{i=1}^{4200}(p_i - r_i)^2 = 0.1820$$

Tương ứng với độ chính xác của thực nghiệm là:

$$CR = (1\text{-}MSE) * 100\% = (1 - 0.1820) * 100\% = 81.8\% \qquad (3.21)$$

Độ chính xác của thực nghiệm xác định độ tương tự của các cặp người dùng

cũng được tính dựa trên *Sai số bình phương trung bình (*MSE - Mean Square Error). Kết quả:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2 = \frac{1}{20000}\sum_{i=1}^{20000}(p_i - r_i)^2 = 0.1311$$

Tương ứng với độ chính xác bằng

$$CR = (1\text{-}MSE) * 100\% = (1 - 0.1311) * 100\% = 86.9\% \quad (3.22)$$

Như vậy, kết quả thực hiện trên nội dung của bài viết đã tăng lên đáng kể với kết quả đạt được là 81.8% và 86.9% so với độ chính xác khi chỉ phân tích nội dung là 68.5% và 70.8%. Điều này chứng tỏ, mô hình bài viết với năm đặc trưng mang lại kết quả phân tích tốt hơn rất nhiều so với chỉ xét nội dung của bài viết.

Ngoài ra, khi xem xét độ tương quan giữa các độ đo trong các nhóm người dùng, luận án sử dụng cách thức thống kê số lượng những người có cùng mức độ tương tự theo không gian bài viết sau và những người có cùng độ tương tự theo không gian các chủ đề để tìm sự tương quan giữa hai độ đo này.

Bảng 3.10: Nhóm các cặp người dùng tương tự theo không gian bài viết

	Mức 0	**Mức 1**	**Mức 2**	**Mức 3**	**Mức 4**
Phân loại theo độ tương tự	4123	9350	4437	1469	621
Tập người dùng	54	108	132	89	103
Phân loại theo chủ đề	532	668	825	860	1315
Tập người dùng theo chủ đề	75	97	93	122	157
Tập giao của hai không gian	41	73	85	88	93
Tỷ lệ trùng nhau	**63.56%**	**70.53%**	**75.56%**	**83.41%**	**71.54%**

Kết quả được như **Bảng 3.10** cho thấy rằng tỷ lệ trùng nhau giữa các tập hợp đã tăng lên rất nhiều. Các mức khác không đều đạt trên 70%. Có nghĩa là nếu hai người dùng có độ tương tự theo bài viết thì trên 70% khả năng các chủ đề quan tâm của họ là tương tự nhau.

Bảng 3.11 so sánh kết quả thu được khi phân tích và so sánh độ tương tự giữa hai người dùng theo bài viết nhiều đặc trưng với người dùng chỉ xét đặc trưng nội dung của bài viết ở Chương 2.

Có thể thấy rằng tỷ lệ các mức độ tương tự ở mức 2, mức 3 và mức 4 tăng lên, nghĩa là khi phân tích người dùng dựa trên bài viết nhiều đặc trưng sẽ tìm được nhiều người có độ tương tự với nhau hơn so với chỉ xét nội dung của bài viết.

Bảng 3.11: So sánh với chỉ có nội dung bài viết

	Mức 0	**Mức 1**	**Mức 2**	**Mức 3**	**Mức 4**
Tỷ lệ trùng nhau khi chỉ xét nội dung	53.08%	58.82%	57.00%	63.78%	58.26%
Tỷ lệ trùng nhau khi xét bài viết nhiều đặc trưng	63.56%	70.53%	75.56%	83.41%	71.54%
Số cặp người dùng tương tự nhau khi chỉ xét đặc trưng nội dung	5094	8961	4122	1355	468
Số cặp người dùng tương tự nhau khi xét bài viết nhiều đặc trưng	4123	9350	4437	1469	621

Bên cạnh đó, mức độ tương quan của người dùng với các chủ đề cũng tăng lên ở các mức 2 mức 3 và mức 4, nghĩa là sự liên quan giữa các bài viết mở rộng với các chủ đề sẽ tăng lên. Có thể nhìn thấy sự chênh lệch giữa các mức như trong **Bảng 3.11** khi so sánh các tỷ lệ trùng nhau khi chỉ xét nội dung của mức 2 là 57.00% nhưng nếu theo mô hình bài viết mở rộng là 75.56%, mức 3 tăng từ 63.78% lên 83.41% và ở mức 4 là tăng từ 58.26% lên 71.54%

3.6. KẾT LUẬN

Chương ba luận án đã trình bày một số vấn đề giới hạn khi phân tích các đặc trưng riêng rẽ của bài viết trên các mạng xã hội, từ đó làm cơ sở đề xuất mô hình biểu diễn bài viết dựa trên năm đặc trưng là nội dung, thể loại, thẻ đánh dấu, quan điểm và cảm xúc. Dựa trên mô hình bài viết này, luận án đưa ra cách biểu diễn mô hình người dùng dựa trên bài viết nhiều đặc trưng bằng véctơ trọng số và cách thức ước lượng hai độ đo: Độ đo tương tự giữa hai người dùng theo không gian các bài viết dựa trên

nhiều đặc trưng và mức độ quan tâm của người dùng theo các chủ đề dựa trên không gian các chủ đề. Dựa trên hai độ đo này, luận án xem xét đến sự tương quan giữa độ tương tự của hai người dùng theo không gian các bài viết và độ quan tâm các chủ đề của người dùng. Các kết quả thực nghiệm đã chỉ ra rằng, nếu hai người dùng có độ tương tự nhau theo bài viết thì các chủ đề quan tâm trên mạng xã hội cũng tương tự nhau và ngược lại, nếu hai người dùng có các chủ đề quan tâm tương tự nhau thì họ cũng có nhiều bài viết tương tự nhau trên mạng xã hội. Các kết quả nghiên cứu liên quan đến chương ba đã được công bố trong một số Tạp chí Khoa học chuyên ngành và Kỷ yếu Hội nghị Khoa học uy tín, bao gồm: Kỷ yếu Hội nghị quốc gia lần thứ 9 về *Nghiên cứu Cơ bản và Ứng dụng (FAIR'9), 2016*. Kỷ yếu của Hội nghị khoa học quốc tế *Advances in Information and Communication Technology, ICTA 12 – Vietnam, 2016, Springer International Publishing*. Kỷ yếu Hội nghị quốc gia lần 10 về *Nghiên cứu Cơ bản và Ứng dụng (FAIR'10), 2017.* Kỷ yếu khoa học quốc tế *Conferences EAI International Conference on Industrial Networks and Intelligent Systems, INISCOM 2017, Vietnam,* Springer International Publishing và Tạp chí *Khoa học và Công nghệ, Trường Đại học Đà Nẵng*, ISSN 1859-1531 – Số 7(128). 2018.

Tuy nhiên, trong mô hình xem xét và phân tích mức độ quan tâm của người dùng đến các chủ đề chưa xem xét đến các hành vi như thích, bình luận, chia sẻ, ... Có thể thấy rằng mô hình người dùng dựa trên bài viết gồm năm đặc trưng vẫn chưa phải bao quát hết các trường hợp khác nhau của người dùng trên các mạng xã hội. Do vậy, luận án đề xuất một mô hình người dùng tổng quát hơn trong Chương 4 để có thể sử dụng trên nhiều mạng xã hội khác nhau và các ngôn ngữ khác nhau.

CHƯƠNG 4: HÀNH VI VÀ QUAN TÂM CỦA NGƯỜI DÙNG THEO HÀNH VI TRÊN MẠNG XÃ HỘI

Chương bốn luận án trình bày mô hình người dùng theo bài viết mở rộng có năm đặc trưng và mô hình người dùng theo các hành vi như đăng bài, thích bài viết, bình luận, chia sẻ và tham gia nhóm. Dựa trên mô hình người dùng này, luận án đề xuất cách thức ước lượng độ tương tự giữa hai người dùng theo mô hình hành vi và cách thức xác định mức độ quan tâm đến các chủ đề của người dùng dựa trên mô hình hành vi. Cuối Chương bốn là hai thực nghiệm dựa trên bộ dữ liệu thực để so sánh với mô hình biểu diễn người dùng dựa trên bài viết ở Chương ba và xem xét mối tương quan giữa tương tự người dùng và tương quan với chủ đề quan tâm của người dùng.

4.1. HÀNH VI CỦA NGƯỜI DÙNG TRÊN MẠNG XÃ HỘI

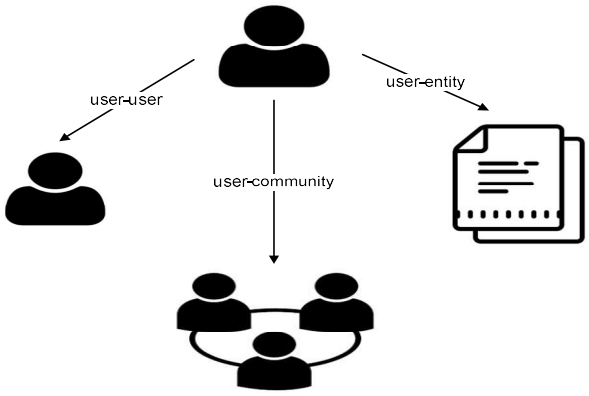
4.1.1. Hành vi và phân loại các hành vi của người dùng trên mạng xã hội

Theo [65] [91] [147] [154] và [104] thì hành vi của người dùng trên các trang mạng xã hội là các cách thức người dùng hoạt động và tương tác với các sự kiện, hiện tượng trên mạng xã hội. Hành vi của người dùng bao gồm các hành động được người dùng thể hiện trên các trang mạng xã hội như chia sẻ bài viết, đăng bài viết, thích bài viết, bình luận trong bài viết, đánh dấu, kết bạn, theo dõi, tạo và gia nhập nhóm/cộng đồng, ... Các hành vi này được phân loại theo hành vi cá nhân (*individual behavior*) và hành vi tập thể (*collective behavior*) minh họa trong **Hình 4.1**.

Hành vi cá nhân được thể hiện qua hành động của cá nhân người dùng đối với các đối tượng, sự kiện trên các trang mạng xã hội, còn hành vi tập thể có thể quan sát được khi một nhóm người dùng có cùng biểu hiện. Hành vi cá nhân được [147] [154] và [104] đề xuất theo ba nhóm là hành vi giữa người dùng với người dùng (*user-user behavior),* hành vi giữa người dùng với các thực thể (*user-entity behavior)* và hành vi giữa người dùng với các cộng đồng (*user-community behavior*):

- Hành vi thể hiện giữa người dùng với người dùng (*user-user*): Đây là hành vi thể hiện sự tương tác qua lại, hoặc mối liên quan giữa hai người dùng trên một mạng xã hội, chẳng hạn như trở thành bạn của nhau, cùng theo dõi lẫn nhau, đánh dấu tên nhau trong các bài viết;
- Hành vi của người dùng với các đối tượng (thực thể) trên một mạng xã hội (*user-entity*): Đây là dạng hành vi thể hiện hành động, thái độ của người dùng trên mạng xã hội chẳng hạn như tạo ra một nội dung trên trang mạng xã hội, thích một nội dung nào đó, như là bài viết, hình ảnh, bình luận về nội dung một bài viết hoặc thể hiện cảm xúc với một bài viết trên trang mạng xã hội;
- Hành vi của người dùng với cộng đồng (*user-community*): Đây là dạng hành vi người dùng thể hiện với cộng đồng như tham gia, rời bỏ một nhóm trên mạng xã hội, theo dõi một nhóm, bình luận nội dung trong nhóm, ...

Hành vi tập thể được xem xét bao gồm việc phân tích các cá nhân thể hiện hành vi một cách độc lập, hoặc bằng cách phân tích các cá nhân thể hiện hành vi tập thể như một nhóm. Hành vi tập thể thường được nghiên cứu trong các nhóm, các cộng đồng trên các mạng xã hội.

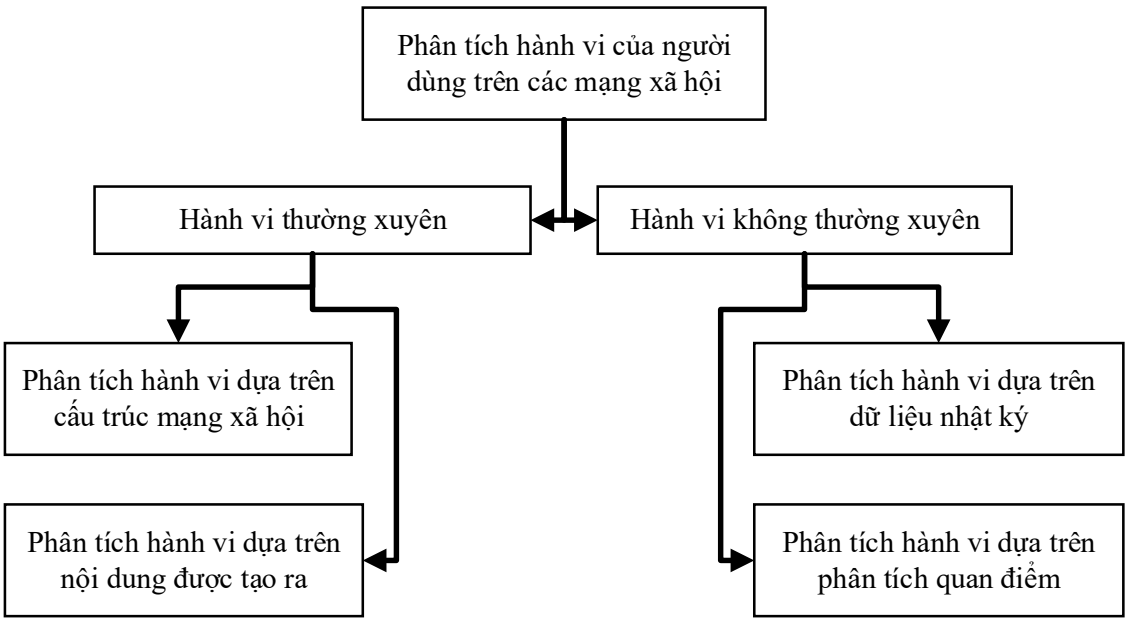


Hình 4.1: Các loại hành vi cá nhân trên mạng xã hội

Nguồn (Huan Liu và Zafarani Reza, (2013) [65])

Các nghiên cứu về hành vi của người dùng trên các mạng xã hội có thể chia thành hai nhóm được tóm tắt như trong **Hình 4.2** bao gồm:

- Các nghiên cứu về các hành vi có tính chất bản năng, cố định do các mạng xã hội cung cấp (*persistent behavior*). Nhóm này thường nghiên cứu và phân tích hành vi người dùng dựa trên cấu trúc thông tin người dùng của mạng xã hội như dựa trên cấu trúc mạng theo đồ thị hoặc mô tả, các bài toán nghiên cứu là so sánh, tính toán các mô-đun, tính chất các kết nối trong mạng, sự gắn kết các thực thể, ... hoặc nghiên cứu, phân tích dựa trên nội dung do người dùng tạo ra trên mạng xã hội như các bài đăng, các thông báo, các bình luận, các hành vi khác mà người dùng thể hiện.
- Các nghiên cứu về các hành vi không cố định, thay đổi thường xuyên, không thường xuyên (*non- persistent behavior*. Nhóm này thường nghiên cứu, phân tích dựa trên các hành vi tiềm ẩn dựa trên dữ liệu lịch sử của người dùng như phân tích các mẫu, các nhận dạng, truy cập các website, hành vi đăng nhập, đăng xuất các hệ thống, ... hoặc là nghiên cứu, phân tích theo hành vi xã hội để xác định các quan điểm (*opinions mining*).



Hình 4.2: Phân loại các nghiên cứu về hành vi của người dùng trên mạng xã hội
Theo Mojtahedi et al. [104]

Hành vi cá nhân thể hiện hành động, thái độ của người dùng đối với các đối tượng, sự kiện trên mạng xã hội và hành vi của người dùng với cộng đồng là những hành vi được xem là tự nhiên và gần với hành vi trên thực tế của người dùng hơn cả [6] [7] [31] [35] [83] [91] [104] [123]. Vì vậy, với mục tiêu phát hiện các chủ đề quan tâm của người dùng theo hướng tiếp cận object-centric như đã trình bày trong các chương 1, 2 và 3, luận án lựa chọn nghiên cứu nhóm hành vi cá nhân theo hướng phân tích các hành vi thường xuyên dựa trên phân tích các nội dung được tạo ra của người dùng.

Theo hướng tiếp cận này của luận án, các hành vi cá nhân được nghiên cứu gồm các hành vi giữa người dùng với các thực thể và các hành vi giữa người dùng với cộng đồng. Theo thống kê từ [65] [91] [147] [104], và [132] thì trên một mạng xã hội, các hành vi của một người dùng bất kỳ thường bao gồm:

- *Đăng bài viết (Post) trên trang cá nhân* – có thể đăng bất cứ một suy nghĩ, cảm xúc, bức ảnh, thể hiện, các đoạn phim, các trích dẫn, chia sẻ từ các phương tiện truyền thông xã hội khác, ...
- *Thích (Like)* thường được gắn với một nội dung cụ thể, người dùng có thể nhấn nút "Thích" khi xem một nội dung bài viết hoặc đọc một bình luận trong một bài viết
- *Bình luận (Comment)* được đưa ra khi người dùng muốn đưa ra quan điểm, ý kiến, hay nhận xét về một nội dung bài viết hay một bình luận khác trong cùng bài viết mà người dùng quan tâm.
- *Tham gia hay gia nhập nhóm (Join group)* là khi người dùng thể hiện sự quan tâm đến nội dung, hoạt động hoặc đơn giản vì họ muốn tìm thấy các thông tin về chủ đề hoặc nội dung mà nhóm tạo ra.
- *Kết bạn (Add friend)* với một người nào đó mà người dùng mong muốn biết thông tin hoặc trao đổi, hoặc giao lưu.
- *Theo dõi (Follow)* với một người nào đó mà người dùng mong muốn, thường người dùng hay thể hiện việc theo dõi để quan tâm đến những cá nhân đặc biệt trên các mạng xã hội như nhà văn nổi tiếng, diễn viên, các nhân vật chính trị, các nhân vật nổi tiếng trên mạng xã hội, các cá nhân kiệt xuất như Mark Zuckerberg, Bill Gates, ...

- *Tạo/tham gia các sự kiện (Event)*, việc tạo và tham gia các sự kiện giúp cho người dùng thu hút sự chú ý của những người dùng khác trên mạng xã hội, việc tham gia các sự kiện thể hiện quan tâm của người dùng đến nội dung của các sự kiện đó
- *Đánh dấu (Tag)* là đánh dấu một hoặc một số người dùng, một hoặc một số nội dung, một hoặc một số địa điểm, ... liên quan đến bài đăng của người dùng, hoặc nội dung cần quan tâm trên các mạng xã hội
- *Chia sẻ (Share)* một bài viết trên trang cá nhân là thể hiện quan tâm của người dùng đó đến nội dung của bài viết có thể trên trang mạng xã hội đó, hoặc trên các phương tiện truyền thông xã hội khác.

4.1.2. Phát hiện quan tâm của người dùng dựa trên hành vi

Bài toán phát hiện quan tâm của người dùng dựa trên phân tích các hành vi của người dùng thường lựa chọn nghiên cứu các hành vi như: hành vi đăng bài [3] [5] [35] [56] [57] [63] [87] [167], hành vi đánh dấu [2] [4] [7] [125] [145] [151], hành vi chia sẻ lại [143] [154] [122] [24] [35] hành vi thích và thể hiện cảm xúc [40] [126] [129] [152], và các hành vi tham gia nhóm [124] [150], ...

Các nghiên cứu phát hiện quan tâm người dùng dựa trên các hành vi có thể tóm tắt ngắn gọn như trong Bảng 4.1.

Bảng 4.1. Tóm tắt các nghiên cứu phát hiện quan tâm từ hành vi người dùng

STT	Nghiên cứu	Hành vi phân tích	Phương pháp và cách tiếp cận	Phương thức đánh giá
1	R.Chinnaiyan et al. IJASCSE'15 [116]	Hành vi truy cập vào log files	- Khai thác nội dung các bài đăng dựa trên lịch sử truy cập của người dùng để phát hiện các chủ đề quan tâm của người dùng - Phương pháp lọc mẫu dựa trên nội dung thông tin	Thống kê các chủ đề
2	A. Gattani et al. VLDB'2013 [4]	Hành vi đánh dấu	- Phân tích các từ khóa trong thẻ đánh dấú, sử dụng mạng tri thức để phân tích	Đánh giá dựa trên độ chính

			- Dựa trên 23 từ khóa là tên của các chủ đề để đối sánh, sử dụng đối sánh theo cặp (string, KB node) sau đó cho điểm	xác độ nhạy và F1
3	Basit Shahzad et al. ‘ IDD’2017 [23]	Hành vi đăng bài	- Phân tích nội dung các tweets trên mạng Tweetter, sau đó phân tách thành hai nhóm là tích cực và tiêu cực. Bỏ hết các tweets tiêu cực và sử dụng SVM để phân loại	Đánh giá dựa trên thống kê
4	Carine Mukamakuza et al. WIM’2018 [32]	Hành vi cho điểm các sản phẩm	- Phân tích dựa trên quan sát định lượng - Sử dụng phương pháp lọc cộng tác kết hợp phân loại	Đánh giá dựa trên P-value
5	F.Chao et al. APSIPA’2016 [50]	Hành vi đăng	- Phân tích nội dung trong bài đăng và ngữ cảnh đăng bài - Sử dụng mô hình Drichle rời rạc (DLDA) cho các đối tượng với 64 đặc trưng cho các đối tượng đồ họa và 6 màu cho hình ảnh. Văn bản dùng túi từ	Độ chính xác
6	Hossen et al. IJCSIT’2018 [63]	Hành vi đăng	- Phân tích nội dung các tweets khi người dùng đăng trên Twitter - So sánh giữa danh sách chủ đề có sẵn với nội dung các tweets	Đánh giá bằng phân loại
7	Jeongin Kim et la. IJSEIA’2013 [77]	Hành vi thích	- Phân tích dựa trên thống kê số lượng các hành vi thích và thống kê các danh từ tìm được trong bài đăng đã được người dùng thích - Dùng phương pháp thống kê và tính trọng số bằng TF.IDF	Đánh giá bằng thống kê
8	Manel Mezghani et al. ICEIS’2014 [99]	Hành vi đánh dấu	- Phân tích các thẻ đánh dấu và trích chọn nội dung của các thẻ đánh dấu - So sánh độ tương tự dựa trên WordNet của các thẻ	Sử dụng độ chính xác trung bình

9	Diana Palsetia et al. KDD'2012 [111]	Hành vi gia nhập cộng đồng	- Phân tích hành vi bình luận trong các cộng đồng để phát hiện quan tâm - Sử dụng độ đo tương tự theo khoảng cách	Đánh giá theo độ chính xác
10	Sengkey C.H et al. ICNCC'16 [122]	Hành vi chia sẻ	- Phân tích nội dung bài đăng được chia sẻ hoặc theo dõi trên Twitter - Phân loại và xử lý theo LDA và cho điểm dựa trên độ tương tự	Đánh giá dựa trên độ chính xác
11	ShengBin et al. IJHIT'2016 [125]	Hành vi đánh dấu	- Phân tích và thống kê các đánh dấu - So sánh và ước lượng dựa trên TF.IDF	Đánh giá dựa trên độ chính xác
12	Tingting Wang et al. PAKDD'2013 [143]	Hành vi chia sẻ	- Phân tích nội dung các bài chia sẻ - Sử dụng độ đo tương tự	Đánh giá bằng độ chính xác
13	Xin Li et al. WWW'08 [145]	Hành vi đánh dấu	- Phân tích và thống kê các đánh dấu - So sánh và ước lượng dựa trên TF.IDF	Đánh giá dựa trên độ chính xác
14	Yin Dawei et al. WSDM'2013 [151]	Hành vi bình luận và đánh dấu	- Phân tích các bình luận và các đánh dấu - Sử dụng mô hình phân tích nhân tố và mạng Bayes	Đánh giá dựa trên độ chính xác
15	Yoad Lewenberg et al. IEEE'2015 [152]	Hành vi thể hiện cảm xúc	- Phân tích các bài viết mà người dùng đã thể hiện cảm xúc, xác định mối quan hệ giữa nội dung bài viết và cảm xúc - Đánh giá độ quan tâm bằng thực nghiệm	Đánh giá dựa trên độ chính xác

Từ các nghiên cứu trong Bảng 4.1 cho thấy rằng, hầu hết các phân tích về hành vi của người dùng đều đi vào phân tích các nội dung văn bản của các bài viết, các thẻ đánh dấu, các bình luận, các nội dung bài viết đã thích. Trong luận án tập trung đề xuất một cách biểu diễn người dùng dựa trên các hành vi có thể phân tích được nhằm phát hiện các chủ đề quan tâm của người dùng dựa trên phân tích dữ liệu văn bản.

4.1.3. Nhóm hay cộng đồng người dùng trên mạng xã hội

Trên các mạng xã hội, nhóm hay cộng đồng thông thường có một tên gọi, một đoạn văn bản để mô tả và được phân chia thành nhiều kiểu như học tập, giải trí, buôn bán, ... Vì vậy, luận án định nghĩa một nhóm hay cộng đồng trên mạng xã hội như sau:

Định nghĩa 4.1:

Một nhóm hay một cộng đồng $g_i \in G$ trên mạng xã hội N, được đặc trưng bởi ba đặc trưng : $g_i = \{name_i, sty_i, des_i\}$. Trong đó:

- $name_i$ *là tên (name) của nhóm* g_i,
- sty_i *là kiểu (style) của nhóm* g_i
- des_i *là mô tả (description) về nhóm* g_i.

Bảng 4.2. Một nhóm trên mạng xã hội Facebook.com

Tên nhóm	Freecycle Vietnam - Nơi cho tặng đồ hoàn toàn miễn phí https://www.facebook.com/groups/freecyclevn/?ref=group_header
Kiểu	Tổng quát, chung
Mô tả	Freecycle Vietnam là nơi cho, tặng và tìm kiếm các loại đồ cũ, mới hoàn toàn miễn phí. Nếu bạn có món đồ nào bạn không cần nhưng lại có giá trị sử dụng đối với người khác, hãy cho tặng món đồ của bạn tại Freecycle Vietnam để một ai đó có thể dùng nó, thay vì bạn vứt nó đi hay cứ để lưu cữu vô ích rồi sẽ hao mòn hỏng hóc theo thời gian. Như vậy là bạn đã giúp đỡ một được một ai đó đang rất cần món đồ như của bạn. Bạn cũng sẽ giúp xã hội tránh lãng phí tiền bạc và tài nguyên thiên nhiên. Bạn đã bảo vệ môi trường một cách thiết thực nhất. Và cuối cùng nhưng không kém phần quan trọng, bạn đã tặng chính bạn một niềm vui nho nhỏ khi cho đi một cách hữu ích. Ngược lại, nếu bạn đang cần món đồ nào đó thì thay vì ngay lập tức mua nó, bạn có thể tìm kiếm trên Freecycle Vietnam biết đâu có món đồ bạn cần, vừa tiết kiệm chi phí, vừa tránh lãng phí chung.

Trong Bảng 4.2 minh họa một nhóm hay một cộng đồng trên mạng xã hội Facebook. Nhóm có tên là:*"Freecycle Vietnam - Nơi cho tặng đồ hoàn toàn miễn phí"*, kiểu nhóm là: "*Tổng quát*" hay *"Chung",* và mô tả nhóm là đoạn văn bản: "*Freecycle Vietnam là nơi cho, tặng và tìm kiếm các loại đồ cũ, mới hoàn toàn miễn phí. Nếu bạn có món đồ nào bạn không cần nhưng lại có giá trị sử dụng đối với người khác, hãy cho tặng món đồ của bạn tại Freecycle Vietnam để một ai đó có thể dùng nó, thay vì bạn vứt nó đi hay cứ để lưu cữu vô ích rồi sẽ hao mòn hỏng hóc theo thời gian. Như vậy là bạn đã giúp đỡ một được một ai đó đang rất cần món đồ như của bạn. Bạn cũng sẽ giúp xã hội tránh lãng phí tiền bạc và tài nguyên thiên nhiên ... vừa tránh lãng phí chung*."

Khi một người dùng tham gia một nhóm hay một cộng đồng thì hành vi này được coi là hành động của cá nhân người dùng với cộng đồng, có thể dựa trên hành vi này để đưa ra các phân tích về hành vi của người dùng trên mạng xã hội.

4.2. MÔ HÌNH NGƯỜI DÙNG THEO HÀNH VI

Như đã đề cập vào cuối Chương 3, mô hình biểu diễn người dùng dựa trên bài viết mở rộng có năm đặc trưng đã cải thiện được hiệu quả của bài toán phát hiện người dùng trên các mạng xã hội nhưng vẫn chưa bao gồm được khá nhiều hành vi như đăng bài viết, thích bài viết, chia sẻ bài viết... Mục tiêu của luận án là đề xuất một cách biểu diễn người dùng trên các mạng xã hội một cách tổng quát để có thể phân tích và ứng dụng trong nhiều bài toán khác nhau, trong đó đặc biệt nhất là bài toán phát hiện quan tâm của người dùng. Vì vậy, trong mục này, luận án đề xuất mô hình người dùng dựa trên các hành vi.

4.2.1. Mô hình biểu diễn người dùng

Trong mục này, luận án đề xuất một mô hình biểu diễn người dùng qua hành vi mà họ thể hiện trên mạng xã hội. Mỗi người dùng biểu diễn thông qua các hành vi là đăng và chia sẻ bài viết, hành vi bình luận trong bài viết, hành vi thích bài viết và hành vi gia nhập nhóm. Hành vi chia sẻ bài viết được luận án xem như hành vi đăng

bài, vì bài viết được chia sẻ được luận án phân tích và tính toán như bài viết trong hành vi đăng. Tương tự, hành vi bình luận trong bài viết được tích hợp vào nội dung của hành vi đăng bài đó hoặc chia sẻ bài viết. Như vậy, mỗi người dùng có thể biểu diễn tổng quát dựa trên ba hành vi chính là đăng bài viết, thích bài viết và gia nhập nhóm trên các mạng xã hội. Trước khi mô tả và đề xuất mô hình biểu diễn người dùng dựa trên các hành vi, luận án định nghĩa các hành vi như sau:

Định nghĩa 4.2:

Trong mạng xã hội $\mathcal{N} =< U, E, G, B >$, tập các hành vi của người dùng B trên mạng xã hội đang xem xét bao gồm:

- $P = \{post_i\}$ *tập hành vi đăng/chia sẻ (post) bài viết trên mạng xã hội N của người dùng,* p_i *là kí hiệu hành vi đăng bài* i *trong tập P.*
- $L = \{like_i\}$ *tập hành vi thích (like) bài viết trên mạng xã hội N,* l_i *là kí hiệu hành vi thích bài viết* i *trong tập L.*
- $C = \{comt_i\}$ *tập các bình luận của người dùng trong bài viết trên mạng xã hội đó,* c_i *là kí hiệu bình luận thứ i trong tập C*
- $J = \{join_i\}$ *tập các hành vi gia nhập nhóm hay cộng đồng người dùng trên mạng xã hội đó,* j_i *là kí hiệu hành vi gia nhập nhóm thứ i trong tập J*

  Mỗi người dùng u_i *khi biểu diễn theo các hành vi sẽ là một bộ bốn như sau:*

  $u_i =< P_i, L_i, C_i, J_i >$

Định nghĩa 4.3:

P là hành vi đăng bài viết (Post an entry). Theo đó, người dùng $u_i \in U$ *đăng bài viết* $e_j \in E$ *trên mạng xã hội* $\mathcal{N}$ *được xác định bởi một ánh xạ:*

$f_{post}: U \times E \rightarrow \{0,1\}$*, xác định như sau:*

$$\begin{cases} f_{post}(u_i, e_j) = 1 \text{ nếu } u_i \text{ đăng bài viết } e_j \in E \\ f_{post}(u_i, e_j) = 0 \text{ nếu } u_i \text{ không đăng bài viết } e_j \in E \end{cases}$$

Định nghĩa 4.4:

L là hành vi thích bài viết (Like an entry). Theo đó, người dùng $u_i \in U$ *thích bài viết* $e_j \in E$ *trên mạng xã hội* $\mathcal{N}$ *được xác định bởi một ánh xạ:*

$f_{like}: U \times E \rightarrow \{0,1\}$, *xác định như sau:*

$$\begin{cases} f_{like}(u_i, e_j) = 1 \text{ nếu } u_i \text{ thích bài viết } e_j \in E \\ f_{like}(u_i, e_j) = 0 \text{ nếu } u_i \text{ không thích bài viết } e_j \in E \end{cases}$$

Định nghĩa 4.5:

Tập các bài viết của người dùng $u_i \in U$ *đã đăng/chia sẻ trên mạng xã hội* $\mathcal{N}$ *được định nghĩa như sau:* $E_i^{post} = \{e_j \in E \mid \forall j, f_{post}(u_i, e_j) = 1\}$

Tập các bài viết $e_j \in E$ *mà người dùng* $u_i \in U$ *đã thích trên mạng xã hội* $\mathcal{N}$ *được định nghĩa như sau:* $E_i^{like} = \{e_j \in E \mid \forall j, f_{like}(u_i, e_j) = 1\}$

Định nghĩa 4.6:

C là hành vi bình luận trong bài viết (Comment in an entry). Theo đó, người dùng $u_i \in U$ *bình luận trong bài viết* $e_j \in E$ *trên mạng xã hội* $\mathcal{N}$ *được xác định bởi một ánh xạ:*

$f_{comt}: U \times E \rightarrow \{0,1\}$, *xác định như sau:*

$$\begin{cases} f_{comt}(u_i, e_j) = 1 \text{ nếu } u_i \text{ bình luận trong bài viết } e_j \in E \\ f_{comt}(u_i, e_j) = 0 \text{ nếu } u_i \text{ không bình luận trong bài viết } e_j \in E \end{cases}$$

Định nghĩa 4.7:

J là hành vi tham gia nhóm/cộng đồng (Join a group/page). Theo đó, người dùng u_i *tham gia vào nhóm* g_j *được xác định bởi một ánh xạ:* $f_{join}: U \times G \rightarrow \{0,1\}$, *xác định như sau:*

$$\begin{cases} f_{join}(u_i, g_j) = 1 \text{ nếu } u_i \text{ có tham gia vào nhóm } g_j \in G \\ f_{join}(u_i, g_j) = 0 \text{ nếu } u_i \text{ không tham gia vào nhóm } g_j \in G \end{cases}$$

Định nghĩa 4.8:

Tập các nhóm/cộng đồng mà người dùng $u_i \in U$ đã tham gia trên mạng xã hội N được Định nghĩa như sau: $G_i^{join} = \{g_k \in G \mid \forall k, f_{join}(u_i, g_k) = 1\}$

Theo Định nghĩa 4.2, mỗi người dùng được biểu diễn bởi ba hành vi là đăng bài viết, thích bài viết, và tham gia vào nhóm hoặc cộng đồng trên mạng xã hội.

- *Hành vi đăng (post)* bài viết $e_i \in E$ của một người dùng $u_i \in U$ trên mạng xã hội $\mathcal{N}$, ký hiệu là: $post_i$, là hành vi đăng một suy nghĩ, cảm xúc, bức ảnh, các đoạn phim, các trích dẫn của chính người dùng tạo ra lên trang cá nhân của người dùng đó.
- *Hành vi chia sẻ một bài viết* cũng được được xếp vào hành vi đăng bài viết bởi vì việc chia sẻ chính là hành vi đăng lại một bài viết, một nội dung nào đó từ chính mạng xã hội đó hoặc các phương tiện truyền thông xã hội khác lên trên trang cá nhân của người dùng, hoặc sao chép có nguồn gốc bài viết của người dùng khác đăng lên trang của họ.
- *Hành vi thích (like)* bài viết $e_i \in E$ của một người dùng $u_i \in U$ trên mạng xã hội $\mathcal{N}$, ký hiệu là: $like_i$, là hành vi kích chuột vào nút "thích" của một bài viết trên mạng xã hội $\mathcal{N}$. Người dùng có thể nhấn nút "thích" khi xem một bài viết, đọc một bình luận trong bài viết, hoặc đọc một nội dung được chia sẻ lại từ tài khoản của người dùng khác. Thông thường, khi người dùng nhấn nút "thích" đã gián tiếp chỉ ra rằng, người dùng đã thể hiện một phản hồi hay một quan tâm tích cực đến nội dung hoặc chủ đề thể hiện trong bài viết mà họ đang quan tâm, hoặc đối tượng mà họ quan tâm.
- *Hành vi bình luận trong bài viết (comment):* Nếu người dùng bình luận trong bài viết đã đăng hoặc chia sẻ của người dùng $e_i \in E$ của một người dùng $u_i \in U$ trên mạng xã hội $\mathcal{N}$, ký hiệu là: $comt_i$, là hành vi để lại một bình luận là văn bản trong một bài viết trên mạng xã hội $\mathcal{N}$. Bình luận của người dùng được tính là một đoạn văn bản và thể hiện thái độ đồng ý (tích cực), hoặc không đồng ý (không tích cực hay tiêu cực), hoặc trung lập

(không đồng ý cũng không phản đối) của người dùng đối với bài viết đã được đăng hoặc đã được chia sẻ.

- *Hành vi tham gia hay gia nhập nhóm (join group)* $g_i \in G$ của một người dùng $u_i \in U$ trên mạng xã hội $\mathcal{N}$, ký hiệu là: $join_i$, là hành vi mà người dùng tham gia vào một nhóm hay cộng đồng trên mạng xã hội. Việc tham gia vào nhóm hay cộng đồng trên mạng xã hội $\mathcal{N}$ cho thấy người dùng thể hiện sự quan tâm đến nội dung, hoạt động hoặc đơn giản vì người dùng muốn tìm thấy các thông tin về chủ đề hoặc nội dung mà các nhóm hay cộng đồng đó mang lại. Việc tham gia một nhóm, thích một nhóm hoặc theo dõi một nhóm trên các trang mạng xã hội thể hiện xu hướng và quan tâm của người dùng trên mạng xã hội của người dùng.

Khi đó mỗi người dùng u_i được biểu diễn dựa trên các hành vi:

$$u_i =< P_i, L_i, C_i, J_i > = \{post_i, like_i, comt_i, join_i\} | u_i \in U \quad (4.1)$$

4.2.2. Biểu diễn mô hình người dùng bằng véc tơ trọng số

Dựa trên công thức (4.1), luận án thực hiện tính toán và biểu diễn các giá trị cho các hành vi của người dùng dựa trên các véctơ có trọng số theo không gian bài viết và không gian các nhóm tham gia.

a. Tính giá trị cho các hành vi

Các hành vi được xem xét tính toán là hành vi đăng bài viết, hành vi thích bài viết, hành vi bình luận trong bài viết và hành vi gia nhập nhóm trên mạng xã hội.

- Giá trị của hành vi đăng bài viết

Giá trị của hành vi đăng bài viết của người dùng $u_i \in U$ trên mạng xã hội $\mathcal{N}$ được xác định bằng tập các bài viết đã đăng và đã chia sẻ của người dùng, ký hiệu là $E_i^{post} \in E$ trên mạng xã hội $\mathcal{N}$ theo Định nghĩa 4.3.

Mỗi bài viết được biểu diễn dựa trên năm đặc trưng theo Định nghĩa 3.1 và công thức (3.7) ở Chương 3.

Giả sử $u_i \in U$ có n bài viết đã đăng và chia sẻ $E_i^{post} = \{e_{i1}, e_{i2}, .. e_{in}\} \in E$ trên mạng xã hội $\boldsymbol{\mathcal{N}}$, khi đó giá trị của hành vi đăng bài viết của người dùng u_i được tính véctơ $\boldsymbol{p_i}$ có n thành phần, mỗi thành phần chính là véctơ trọng số của bài viết đã đăng, đã chia sẻ tương ứng trong E_i^{post} được tính theo công thức (3.7)

$$u_{ipost} = post_i = \boldsymbol{p_i} = (\boldsymbol{e_{i1}}, \boldsymbol{e_{i2}}, .. \boldsymbol{e_{in}}) \quad (4.2)$$

- Giá trị của hành vi thích bài viết

Tương tự với hành vi đăng, giá trị của hành vi thích bài viết của người dùng $u_i \in U$ trên mạng xã hội $\boldsymbol{\mathcal{N}}$ được xác định bằng tập các bài viết đã thích của người dùng $E_i^{like} \in E$ trên mạng xã hội $\boldsymbol{\mathcal{N}}$ theo Định nghĩa 4.4.

Mỗi bài viết được biểu diễn dựa trên năm đặc trưng theo Định nghĩa 3.1 và công thức (3.7) ở Chương 3.

Giả sử $u_i \in U$ có m bài viết đã thích $E_i^{like} = \{e_{i1}, e_{i2}, .. e_{im}\} \in E$ trên mạng xã hội $\boldsymbol{\mathcal{N}}$, khi đó giá trị của hành vi thích các bài viết của người dùng u_i được được tính véctơ $\boldsymbol{l_i}$ có m thành phần, mỗi thành phần chính là véctơ trọng số của các bài viết đã thích tương ứng trong E_i^{like} được tính theo công thức (3.7)

$$u_{ilike} = like_i = \boldsymbol{l_i} = (\boldsymbol{e_{i1}}, \boldsymbol{e_{i2}}, .. \boldsymbol{e_{im}}) \quad (4.3)$$

- Giá trị của hành vi bình luận trong bài viết

Một số người dùng có thể bình luận hoặc thích một vài bình luận mà các người dùng khác đã đăng hoặc chia sẻ. Trong trường hợp này, luận án ước lượng dựa trên các nguyên tắc dưới đây:

- Nếu người dùng u_i không bình luận trong bài viết $e_k \in E$ thì bài viết đó không được xét trong luận án.
- Nếu người dùng u_i có bình luận trong bài viết $e_k \in E$ thì giá trị của bình luận sẽ được tính với giá trị là tích cực, tiêu cực, hoặc trung lập. Nếu trong một bài viết mà người dùng u_i thực hiện nhiều bình luận thì luận án chỉ xét các bình luận có quan điểm tích cực, hoặc quan điểm tiêu cực còn quan điểm trung lập

sẽ bị loại bỏ hay không được tính. Khi đó bài viết có bình luận của người dùng sẽ được đưa vào danh sách các bài viết có bình luận để xét.

Tương tự với hành vi đăng, giá trị của hành vi bình luận trong bài viết của người dùng $u_i \in U$ trên mạng xã hội $\boldsymbol{\mathcal{N}}$ được xác định bằng tập các bài viết đã có bình luận của người dùng $E_i^{comt} \in E$ trên mạng xã hội $\boldsymbol{\mathcal{N}}$ theo Định nghĩa 4.4.

Giả sử $u_i \in U$ có p bài viết đã thực hiện bình luận $E_i^{comt} = \{e_{i1}, e_{i2}, .. e_{ip}\} \in E$ trên mạng xã hội $\boldsymbol{\mathcal{N}}$, khi đó giá trị của hành vi bình luận trong các bài viết của người dùng u_i được được tính véctơ $\boldsymbol{c_i}$ có p thành phần, mỗi thành phần chính là véctơ trọng số của các bài viết đã bình luận tương ứng trong E_i^{comt} được tính theo công thức (3.7)

$$u_{icomt} = comt_i = \boldsymbol{c_i} = \boldsymbol{(e_{i1}, e_{i2}, .. e_{ip})} \quad (4.4)$$

- Giá trị của hành vi gia nhập một nhóm trên mạng xã hội

Hành vi gia nhập một nhóm $g_i \in G_i^{join}$ trên mạng xã hội $\boldsymbol{\mathcal{N}}$của người dùng được xác định dựa trên Định nghĩa 4.7 và Định nghĩa 4.8, để xét hành vi gia nhập nhóm thì xét và phân tích tập các nhóm đã tham gia của người dùng.

Ký hiệu là: $u_{ijoin} = join_i = G_i^{join}$ trong đó, $G_i^{join} = \{g_i | \forall i\}$ mỗi nhóm được biểu diễn bởi ba đặc trưng $g_i = \{name_i, sty_i, des_i\} \in G$.

Cách xác định giá trị cho các g_i dựa trên Định nghĩa 2.1, toàn bộ ba đặc trưng của nhóm gồm tên, mô tả và thể loại được đưa vào không gian các nhóm $\boldsymbol{\mathcal{G}}$. Từ đó, mỗi nhóm được biểu diễn bởi một véc tơ trọng số như sau.

Cho một tập các nhóm $\boldsymbol{\mathcal{G}} = \{g_1, g_2, \dots, g_p\}$, mỗi một nhóm được biểu diễn bằng một tập các thuật ngữ $g_i = \{d_{i1}, d_{i2}, .. d_{ip_i}\}$. Gọi q là các từ vựng khác nhau từng đôi một trong không gian $\boldsymbol{\mathcal{G}}$. Khi đó, mỗi g_i được biểu diễn bởi một véctơ có q chiều: $\boldsymbol{g_i} = (w_{i1}, w_{i2}, .. w_{iq})$ trong không gian $\boldsymbol{\mathcal{G}}$. Như vậy, hành vi gia nhập một nhóm:

$$u_{ijoin} = join_i = \boldsymbol{j_i} = (\boldsymbol{g_1}, \boldsymbol{g_2}, \dots, \boldsymbol{g_p}) \quad (4.5)$$

Như vậy, mỗi người dùng u_i trên mạng xã hội được biểu diễn bằng một véctơ dựa trên các hành vi có các thành phần như sau:

$$u_i = (\boldsymbol{p_i}, \boldsymbol{l_i}, \boldsymbol{c_{i,}}\, \boldsymbol{j_i}) \tag{4.6}$$

Trong đó, $\boldsymbol{p_i}, \boldsymbol{l_i}, \boldsymbol{c_{i,}}\, \boldsymbol{j_i}$ lần lượt được tính dựa trên các công thức (4.2), (4.3), (4.4) và (4.5) chiều của $\boldsymbol{p_i}$ bằng tập tất cả các từ vựng của các bài viết mà người dùng đã đăng hoặc chia sẻ, chiều của $\boldsymbol{l_i}$ bằng tập tất cả các từ vựng của các bài viết mà người dùng đã thích, chiều của $\boldsymbol{c_i}$ bằng tập tất cả các từ vựng của các bài viết đã được người dùng bình luận có giá trị không trung lập, chiều của $\boldsymbol{j_i}$ bằng tập tất cả các từ vựng xuất hiện trong các nhóm mà người dùng đã tham gia.

Nói cách khác có thể biểu diễn người dùng dựa trên các hành vi như sau:

$$u_i = (post_i, like_i, comt_i, join_i) = \begin{cases} E_i^{post} = \boldsymbol{p_i} = (\boldsymbol{e_{i1}}, \boldsymbol{e_{i2}}, .. \boldsymbol{e_{in}}), \\ E_i^{like} = \boldsymbol{l_i} = (\boldsymbol{e_{i1}}, \boldsymbol{e_{i2}}, .. \boldsymbol{e_{im}}), \\ E_i^{comt} = \boldsymbol{c_i} = (\boldsymbol{e_{i1}}, \boldsymbol{e_{i2}}, .. \boldsymbol{e_{ik}}), \\ G_i^{join} = \boldsymbol{j_i} = (\boldsymbol{g_{i1}}, \boldsymbol{g_{i2}}, .. \boldsymbol{g_{ip}}) \end{cases} \tag{4.7}$$

4.2.3. Độ tương tự giữa hai người dùng theo hành vi

Trong mục này, luận án đề xuất một cách thức phân nhóm người dùng theo độ đo tương tự dựa trên mô hình biểu diễn người dùng theo hành vi. Mô hình ước lượng được tính toán dựa trên độ tương tự trên từng hành vi, sau đó tích hợp có trọng số để thu được độ tương tự của các cặp người dùng theo hành vi. Luận án tiến hành phân loại người dùng theo các mức độ tương tự để chia thành các nhóm. Kết quả này được dùng để so sánh với việc phân loại người dùng theo các chủ đề quan tâm dựa trên hành vi, để xem xét mối tương quan giữa các người dùng và các chủ đề mà họ quan tâm có liên quan đến nhau hay không?

a. Mô hình ước lượng tổng quát

Giả sử có hai người dùng u_i và u_k trên mạng xã hội N, độ đo tương tự của hai người dùng theo hành vi được luận án tính bằng tích hợp có trọng số độ đo tương tự trên các hành vi của người dùng trên mạng xã hội N, được tính theo công thức (4.8) như sau:

$$s_{beha}(u_i, u_k) = w_{post} * s_{post}(u_i, u_k) + w_{like} * s_{like}(u_i, u_k) + w_{comt} * s_{comt}(u_i, u_k) + w_{join} * s_{join}(u_i, u_k) \quad (4.8)$$

Trong đó:

- $w_{post}, w_{like}, w_{comt}, w_{join}$, lần lượt là trọng số của hành vi đăng/ chia sẻ bài viết, hành vi thích bài viết, hành vi bình luận trong bài viết và hành vi tham gia một nhóm trên mạng xã hội, và chúng thỏa mãn điều kiện: $w_{post} + w_{like} + w_{comt} + w_{join} = 1$
- $s_x(u_i, u_k)$ là độ tương tự trên từng hành vi của hai người dùng u_i, u_k.

b. Độ tương tự trên từng hành vi

- *Độ tương tự dựa trên hành vi đăng bài viết:* Độ tương tự trên hành vi đăng/chia sẻ bài viết của hai người dùng được tính bằng độ tương tự giữa hai tập bài đã đăng của hai người dùng tương ứng trên mạng xã hội N.

  Giả sử có hai người dùng u_i và u_k trên mạng xã hội $\mathcal{N}$, với $u_i, u_k \in U$ và E_i^{post}, $E_k^{post} \in E$ tương ứng là hai tập các bài viết đã đăng của hai người dùng u_i và u_k. Khi đó, độ tương tự của hai người dùng u_i và u_k theo hành vi đăng/chia sẻ bài viết trên mạng xã hội được tính dựa trên công thức theo công thức (4.9) như sau:

  $$s_{post}(u_i, u_k) = sim(E_i^{post}, E_k^{post}) = sim(\boldsymbol{p_i}, \boldsymbol{p_k}) \quad (4.9)$$

  Trong đó, $\boldsymbol{p_i}, \boldsymbol{p_k}$ là hai véctơ tương ứng của hai tập bài viết E_i^{post}, E_k^{post}, được tính theo công thức 4.2, $sim(\boldsymbol{p_i}, \boldsymbol{p_k})$ được tính theo công thức (3.15)

- *Độ tương tự dựa trên hành vi thích bài viết:* Giả sử có hai người dùng u_i và u_k trên mạng xã hội $\mathcal{N}$, với $u_i, u_k \in U$ và E_i^{like}, $E_k^{like} \in E$ tương ứng là hai tập các bài viết đã thích của hai người dùng u_i và u_k. Khi đó, độ tương tự của hai người dùng u_i và u_k theo hành vi thích bài viết trên mạng xã hội được tính dựa trên công thức (4.10) như sau:

$$s_{like}(u_i, u_k) = sim\left(E_i^{like}, E_k^{like}\right) = sim(\boldsymbol{l_i}, \boldsymbol{l_k}) \quad (4.10)$$

Trong đó, $\boldsymbol{l_i}, \boldsymbol{l_k}$ là hai véctơ tương ứng của hai tập bài viết E_i^{like}, E_k^{like}, được tính theo công thức 4.2, $sim(\boldsymbol{l_i}, \boldsymbol{l_k})$ được tính theo công thức (3.15)

- *Độ tương tự dựa trên hành vi bình luận trong bài viết:* Gọi E_{posi}^i và E_{nega}^i lần lượt là tập hợp các bài viết có giá trị tích cực và bài viết có giá trị tiêu cực của người dùng u^i. Gọi E_{posi}^k và E_{nega}^k lần lượt là tập hợp các bài viết có giá trị tích cực và bài viết có giá trị tiêu cực của người dùng u^k. Để xác định độ tương tự hành vi thích bình luận của người dùng u^i và u^k luận án sử dụng các giả thiết sau đây:
    - Nếu có càng nhiều hai bộ E_{posi}^i và E_{posi}^k giống nhau thì độ tương tự hành vi bình luận của người dùng u^i và u^k càng cao.
    - Nếu có càng nhiều hai bộ E_{nega}^ivà E_{nega}^k giống nhau thì độ tương tự hành vi bình luận của người dùng u^i và u^k càng cao.
    - Nếu có càng ít hai bộ E_{posi}^ivà E_{nega}^k giống nhau thì độ tương tự hành vi bình luận của người dùng u^i và u^k càng cao.
    - Nếu có càng ít hai bộ E_{posi}^k và E_{nega}^i giống nhau thì độ tương tự hành vi bình luận của người dùng u^i và u^k càng cao.

Đặt $\begin{cases} S1 = sim\left(E_{posi}^i, E_{posi}^k\right) + sim\left(E_{nega}^i, E_{nega}^k\right) \\ S2 = sim\left(E_{posi}^i, E_{nega}^k\right) + sim\left(E_{nega}^i, E_{posi}^k\right) \end{cases}$

Khi đó, độ tương tự về hành vi bình luận của người dùng u^i và u^k được định nghĩa bằng công thức sau:

$$sim_{comt}(u_i, u_k) = min\,(1, max(0, |S1 - S2|)) \quad (4.11)$$

- *Độ tương tự dựa trên hành vi gia nhập nhóm:* Giả sử có hai người dùng u_i và u_k trên mạng xã hội $\mathcal{N}$, với $u_i, u_k \in U$ và G_i^{join}, $G_k^{join} \in G$ tương ứng là hai tập các nhóm đã gia nhập của hai người dùng u_i và u_k. Độ tương tự của hai người dùng u_i và $u_k \in U$ trên mạng xã hội $\mathcal{N}$ theo hành vi gia nhập nhóm trên mạng xã hội được tính dựa trên công thức (4.12) như sau:

$$sim_{join}(u_i, u_k) = sim\left(G_i^{join}, G_k^{join}\right) = sim(\boldsymbol{j_i}, \boldsymbol{j_k}) \quad (4.12)$$

Trong đó, $\boldsymbol{j_i}, \boldsymbol{j_k}$ là hai véctơ tương ứng của hai tập nhóm G_i^{join}, G_k^{join}, được tính theo công thức 4.4, $sim(\boldsymbol{j_i}, \boldsymbol{j_k})$ được tính theo công thức (3.15)

Như vậy dựa trên công thức (4.8) kết hợp các công thức (4.9), (4.10) và (4.11) cùng trọng số của các hành vi, có thể thấy rằng độ tương tự giữa hai người dùng u_i và $u_k \in U$ theo các hành vi trên mạng xã hội nằm trong khoảng đơn vị [0,1]. Để xác định bộ trọng số cho từng hành vi trên mạng xã hội áp dụng trong tính toán và xử lý, luận án thực hiện một thực nghiệm để thực hiện tìm bộ trọng số trong mục 4.4.

4.3. QUAN TÂM CỦA NGƯỜI DÙNG THEO MÔ HÌNH HÀNH VI

4.3.1. Biểu diễn mô hình hành vi người dùng theo không gian chủ đề

Trong mục này, luận án trình bày cách thức xác định độ quan tâm các chủ đề của người dùng dựa trên các hành vi. Để ước lượng mức độ quan tâm của người dùng theo hành vi, luận án thực hiện mô hình hóa các hành vi của người dùng theo không gian các chủ đề trên mạng xã hội, sau đó ước lượng mức độ liên quan giữa các hành vi với các chủ đề để xác định mức độ quan tâm của người dùng.

Khi đó, giả sử rằng $\boldsymbol{\mathcal{T}} = \{T_1, T_2, \dots, T_p\}$ là một tập các chủ đề trên mạng xã hội N, trong đó mỗi chủ đề được biểu diễn bằng một tập các thuật ngữ hoặc các từ $T_i = \{t_{i1}, t_{i2}, \dots, t_{ip_i}\}$. Như vậy véctơ trọng số của các bài viết được tính dựa trên công thức (2.9) , kết hợp công thức (3.16) thì có thể biểu diễn mô hình người dùng theo hành vi dựa theo không gian chủ đề như sau:

Mỗi bài viết được xét trong hành vi đăng, hành vi thích, hành vi bình luận và mỗi nhóm người dùng đã tham gia được biểu diễn theo không gian các chủ đề theo công thức (3.16) như vậy mỗi người dùng sẽ được biểu diễn bằng:

$$u_i^t = \begin{cases} E_i^{post} = \boldsymbol{p_i}^{\boldsymbol{t}} = (\boldsymbol{e_{i1}}, \boldsymbol{e_{i2}}, .. \boldsymbol{e_{in}}), \\ E_i^{like} = \boldsymbol{l_i}^{\boldsymbol{t}} = (\boldsymbol{e_{i1}}, \boldsymbol{e_{i2}}, .. \boldsymbol{e_{im}}), \\ E_i^{comt} = \boldsymbol{c_i}^{\boldsymbol{t}} = (\boldsymbol{c_{i1}}, \boldsymbol{c_{i2}}, .. \boldsymbol{c_{ik}}) \\ G_i^{join} = \boldsymbol{j_i}^{\boldsymbol{t}} = (\boldsymbol{g_{i1}}, \boldsymbol{g_{i2}}, .. \boldsymbol{g_{ip}}) \end{cases} \quad (4.13)$$

Trong đó, $\boldsymbol{e_{ij}^k} = \left(e_{ij}^1, e_{ij}^2, \dots, e_{ij}^{t_{kp}}\right), e_{ij}^k = tf(t_{il}, e_{ij}) \times idf(t_{il}, E_i)$ *với* $t_{il} \in \boldsymbol{\mathcal{T}}$

4.3.2. Xác định chủ đề quan tâm theo hành vi

Theo mô hình người dùng theo hành vi, mỗi người dùng u_i trên mạng xã hội N được biểu diễn thông qua ba hành vi tương ứng và đăng bài viết, thích bài viết và gia nhập nhóm cộng đồng. Để biểu diễn người dùng theo các hành vi theo không gian các chủ đề, luận án dựa trên mô hình biểu diễn người dùng theo bài viết dựa trên chủ đề đã trình bày ở Chương 3, kết hợp công thứ (3.18) và (3.19) thì mô hình hành vi người dùng theo không gian chủ đề được tính như sau:

Giả sử rằng $\boldsymbol{\mathcal{T}} = \{T_1, T_2, \dots, T_p\}$ là một tập các chủ đề trên mạng xã hội N, khi đó, mức độ liên quan của các hành vi đăng bài viết, thích bài viết và gia nhập của người dùng u_i với các chủ đề trong $\boldsymbol{\mathcal{T}}$ được tính bằng mức độ liên quan của các tập bài viết $E_i^{post}, E_i^{like}, G_i^{join}$ với các chủ đề đang xem xét. Ký hiệu tương ứng là:

$$\boldsymbol{u_{ipost}^t} = (u_i^1, u_i^2, \dots, u_i^p) \quad (4.12)$$

Trong đó, mỗi u_i^k là mức độ quan tâm của người dùng u_i đến chủ đề thứ k trong tập $\boldsymbol{\mathcal{T}}$ theo các bài viết đã đăng và được tính theo công thức (3.18)

$$\boldsymbol{u_{ilike}^t} = (u_i^1, u_i^2, \dots, u_i^p) \quad (4.13)$$

Trong đó, mỗi u_i^k là mức độ quan tâm của người dùng u_i đến chủ đề thứ k trong tập $\boldsymbol{\mathcal{T}}$ theo các bài viết đã thích và được tính theo công thức (3.18)

$$\boldsymbol{u_{icomt}^t} = (u_i^1, u_i^2, \dots, u_i^p) \quad (4.13)$$

Trong đó, mỗi u_i^k là mức độ quan tâm của người dùng u_i đến chủ đề thứ k trong tập $\boldsymbol{\mathcal{T}}$ theo các bài viết đã bình luận và được tính theo công thức (3.18)

$$\boldsymbol{u_{ijoin}^t} = (u_i^1, u_i^2, \dots, u_i^p) \quad (4.14)$$

Trong đó, mỗi u_i^k là mức độ quan tâm của người dùng u_i đến chủ đề thứ k trong tập $\boldsymbol{\mathcal{T}}$ theo các nhóm đã gia nhập và được tính bằng mức độ tương tương của nhóm gia nhập với chủ đề k theo công thức (3.17)

Khi đó, mức độ quan tâm của người dùng u_i với các chủ đề trong $\boldsymbol{\mathcal{T}}$ được tính theo công thức:

$$\boldsymbol{u_i^t} = w_p * \boldsymbol{u_{ipost}^t} + w_l * \boldsymbol{u_{ilike}^t} + w_j * \boldsymbol{u_{ijoin}^t} \quad (4.15)$$

Trong đó, w_p, w_l, w_j là trọng số của các hành vi thỏa mãn $w_p + w_l + w_j$ =1 và các $\boldsymbol{u_k^t}$ là các độ đo mức quan tâm của người dùng đến các chủ đề trong tập $\boldsymbol{\mathcal{T}}$.

4.3.3. Độ tương tự quan tâm của người dùng theo chủ đề

Để ước lượng độ tương tự quan tâm các chủ đề giữa hai người dùng $u_i, u_j \in U$ trên mạng xã hội $\boldsymbol{\mathcal{N}}$ theo các chủ đề *t*, luận án dựa trên độ tương tự giữa hai tập các hành vi tương ứng với mức độ liên quan của các chủ đề để tính toán.

Giả sử có hai người dùng $u_i, u_j \in U$ có hai giá trị quan tâm $\boldsymbol{u_i^t}, \boldsymbol{u_j^t}$ tương ứng với các chủ đề $\boldsymbol{\mathcal{T}}$.

Khi đó độ quan tâm tương tự của hai người dùng theo hành vi dựa trên chủ đề được tình bằng

$$sim_{int}(u_i, u_j) = sim(\boldsymbol{u_i^t}, \boldsymbol{u_j^t}) \quad (4.16)$$

Trong đó các $\boldsymbol{u_i^t}$, $\boldsymbol{u_j^t}$ được tính theo công thức (4.15), và $sim(\boldsymbol{u_i^t}, \boldsymbol{u_j^t})$ được tính như công thức (2.16). Từ đó có thể thấy rằng $sim_{int}(u_i, u_j)$ nằm trong khoảng [0,1].

4.4. TƯƠNG QUAN GIỮA TƯƠNG TỰ NGƯỜI DÙNG VÀ QUAN TÂM

4.4.1. Bài toán xác định tương quan giữa tương tự người dùng và chủ đề

Quay trở lại bài toán

Cho một tập người dùng U trên mạng xã hội $\mathcal{N}$. Gọi SimU là tập những người dùng tương tự nhau dựa trên các hành vi và CorrU tập người dùng tương tự nhau theo chủ đề. Với $SimU \subseteq U$ và $CorrU \subseteq U$.

Chứng minh rằng: $SimU \cap CorrU > 0$

Trong Chương 3 luận án đã tìm thấy sự tương quan đã tăng lên khi xét trên bai viết mở rộng với năm đặc trưng. Trong mục này luận thực hiện 03 thực nghiệm chính để xem xét và so sánh mức độ tương quan khi biểu diễn người dùng theo các hành vi bao gồm:

- Thực nghiệm ước lượng độ tương tự giữa hai người dùng theo mô hình hành vi để tìm tập SimU
- Thực nghiệm xác định độ tương quan theo các chủ đề quan tâm của người dùng dựa trên hành vi với các chủ đề để tìm tập CorrU
- Thực nghiệm xác định bộ trọng số tối ưu cho ba hành vi

4.4.2. Thực nghiệm đánh giá

a. Xây dựng bộ dữ liệu thử nghiệm

Luận án xây dựng bộ dữ liệu thử nghiệm bằng dữ liệu thực thu được từ mạng xã hội Facebook. Bao gồm 200 người dùng và bộ các bài viết đã đăng, đã thích và các nhóm đã tham gia tương ứng. Hành vi đăng bài của một người dùng được xét trên 10 bài viết, hành vi thích của người dùng cũng được thực hiện trên 10 bài viết, riêng

hành vi gia nhập nhóm thì có người tham gia rất nhiều nhóm, có người không tham gia nhóm nào, nên để bình quân, luận án chỉ xét mỗi người dùng là 5 nhóm tham gia.

Hai bộ dữ liệu chính để thực hiện thực nghiệm được mô tả như sau:

Bảng 4.3. Mô tả bộ dữ liệu thực nghiệm

	Bộ dữ liệu tính độ tương tự	Bộ dữ liệu tính độ tương quan
Người dùng (User)	200	200
Bài đã đăng (Entry post)	2000	2000
Bài đã thích (Entry like)	2000	2000
Nhóm đã tham gia (Group join)	800	800
Chủ đề (Topic)	0	21
Trọng số	TF.IDF	TF.IDF

b. Thực nghiệm xác định bộ trọng số cho các hành vi

Để tính toán và đưa ra bộ trọng số cho các hành vi của người dùng trong mô hình đề xuất, luận án thử nghiệm ước lượng độ tương tự giữa hai người dùng dựa trên một bộ mẫu gồm 300 mẫu xây dựng theo không gian bài viết và nhóm tham gia. Sau đó tiến hành chạy trên bộ dữ liệu theo ba bước trình bày tiếp sau đây, để khảo sát bộ trọng số cho các hành vi trong mô hình đề xuất như sau:

Bước 1: Thực hiện việc chạy trên toàn bộ mẫu thử nghiệm nhiều lần, mỗi lần chỉ với 1 hành vi của người dùng. Sau đó tiếp tục thực hiện chạy nhiều lần với tổ hợp 2 hành vi, sau đó là các tổ hợp 3 hành vi, cuối cùng là tổ hợp gồm 4 hành vi của người dùng.

Bước đầu cho các trọng số có giá trị bằng nhau, sau đó thay đổi dựa trên kết quả thu được, các tổ hợp được liệt kê trong Bảng 4.5 là toàn bộ các tổ hợp được xem xét để thực nghiệm.

Bước 2: Lưu toàn bộ các giá trị của các lần thử nghiệm, tính độ chính xác CR và so sánh giữa các kết quả với nhau, với mỗi bộ kết hợp các hành vi này, luận án cho chạy mô hình với các trọng số bắt đầu với các giá trị bằng nhau, sau đó thay đổi các giá trị trọng số để xem xét các kết quả.

Bảng 4.6 trình bày các bộ trọng số có kết quả phù hợp nhất từ thực nghiệm cho các hành vi.

Bước 3: Ước lượng và tính toán các lần chạy như trong Bảng 4.4 là các tổ hợp khảo sát để chọn bộ trọng số và Bảng 4.6 là kết quả ước lượng để chọn bộ trọng số phù hợp nhất cho thực nghiệm.

Bảng 4.4: Các tổ hợp khảo sát chọn bộ trọng số

Số đặc trưng	Số tổ hợp	Tổ hợp các đặc tính
1/4	4	Đăng bài viết Thích bài viết Bình luận trong bài viết Tham gia nhóm
2/4	6	Đăng bài viết – Thích bài viết Đăng bài viết - Bình luận trong bài viết Đăng bài viết - Tham gia nhóm Thích bài viết – Bình luận trong bài viết Thích bài viết – Tham gia nhóm Bình luận trong bài viết – Tham gia nhóm
3/4	4	Đăng bài viết - Thích bài viết – Bình luận trong bài viết Thích bài viết – Bình luận trong bài viết –Tham gia nhóm Đăng bài viết - Bình luận trong bài viết - Tham gia nhóm Đăng bài viết - Thích bài viết – Tham gia nhóm
4/4	1	Đăng bài viết – Thích bài viết – Bình luận trong bài viết – Tham gia nhóm

Kết quả được trình bày trong ***Bảng 4.5***, thứ tự các nhóm kết quả được trình bày lần lượt theo mức số lượng các hành vi của người dùng được kết hợp, đến các bộ hành vi được kết hợp (sắp xếp theo chiều tăng dần – tốt dần lên) trong mỗi mức.

Giá trị trong ô tương ứng với mỗi hành vi là trọng số tối ưu của hành vi đó, hai cột cuối cùng biểu diễn số lượng mẫu đúng và tỉ lệ phần trăm mẫu đúng.

Bảng 4.5: Khảo sát và lựa chọn bộ trọng số ước lượng

Số hành vi	w_1 **Đăng bài viết**	w_2 **Thích bài viết**	w_3 **Bình luận trong bài viết**	w_4 **Gia nhập nhóm**	**#mẫu đúng (Tổng 300 mẫu)**	**Tỉ lệ (%)**
Dùng ¼ hành vi				1.00	113	37.67
		1.00			120	40.00
			1.00		143	47.67
	1.00				***196***	***65.33***
Dùng 2/4 hành vi		0.65		0.35	187	62.33
		0.45	0.55		192	64.00
			0.65	0.35	201	67.00
	0.75			0.25	236	78.67
	0.70	0.30			241	80.33
	0.75		***0.25***		***247***	***82.33***
Dùng ¾ hành vi		0.3	0.45	0.25	242	80.67
	0.60	0.25		0.15	257	85.67
	0.60		0.30	0.10	258	86.00
	0.60	***0.25***	***0.35***		***263***	***87.33***
Dùng 4/4 hành vi	***0.35***	***0.25***	***0.30***	***0.10***	***269***	***89.67***
Bộ trọng số phù hợp nhất	***0.35***	***0.25***	***0.30***	***0.10***	***269***	***89.67***

Như vậy, nếu dùng 4 hành vi như mô hình đề xuất thì chỉ có 1 bộ kết hợp duy nhất, bộ này cho kết quả phù hợp nhất với bộ trọng số là < 0.35, 0.25, 0.30, 0.10> lần lượt tương ứng với các hành vi <*đăng bài viết, thích bài viết, bình luận trong bài viết, tham gia nhóm*>.

Sự kết hợp này cũng cho kết quả tốt nhất trên bộ mẫu được xây dựng, với độ chính xác đến 89.67%. Nhìn vào Bảng 4.6 cũng thấy rằng, hành vi đăng bài viết có trọng số cao nhất trong tất cả các bộ tổ hợp.

c. Ước lượng độ tương tự giữa hai người dùng theo hành vi

- **Kịch bản thực nghiệm:** *Kịch bản thực nghiệm xác định độ tương tự của hai người dùng dựa trên mô hình hành vi được thực hiện như sau:*

Đầu vào: Gồm 200 người dùng với 2000 bài viết của hành vi đăng bài, 2000 bài của hành vi thích bài viết và 800 nhóm của hành vi gia nhập nhóm.

Đầu ra: Độ tương tự giữa các cặp người dùng theo các hành vi dựa trên bộ trọng số (0.6, 0.25, 0.15)

Thực hiện

- Tính giá trị và xây dựng bộ thuật ngữ cho các đặc trưng của các bài viết và nhóm mà người dùng đã tham gia

- Tính véctơ trọng số theo TF.IDF cho mỗi đặc trưng của bài viết và các nhóm theo không gian của các hành vi

- Tính độ tương tự theo trên từng hành vi và tích hợp có trọng số

- Phân loại mức độ tương tự của các cặp người dùng theo hành vi

Kết thúc

- **Tham số đầu ra:** Là độ tương tự của mỗi cặp người dùng
- **Các bước thực hiện chi tiết như sau:** Bước 1 và Bước 2 được thực hiện như trong thực nghiệm trong mục 3.3.1 và mục 4.2.2 các véctơ tương ứng được tính dựa trên không gian bài viết và nhóm. Bước 3 Tính độ tương tự giữa các cặp bài viết dựa trên công thức (3.8) với bộ trọng số lấy từ thực nghiệm ở mục 3.5.2.b. Tính độ tương tự giữa các nhóm theo công thức (4.4) và (4.10). Bước 4 Tính độ tương tự giữa các cặp người dùng dựa trên Định nghĩa 2.7 theo công thức (4.7) với bộ trọng

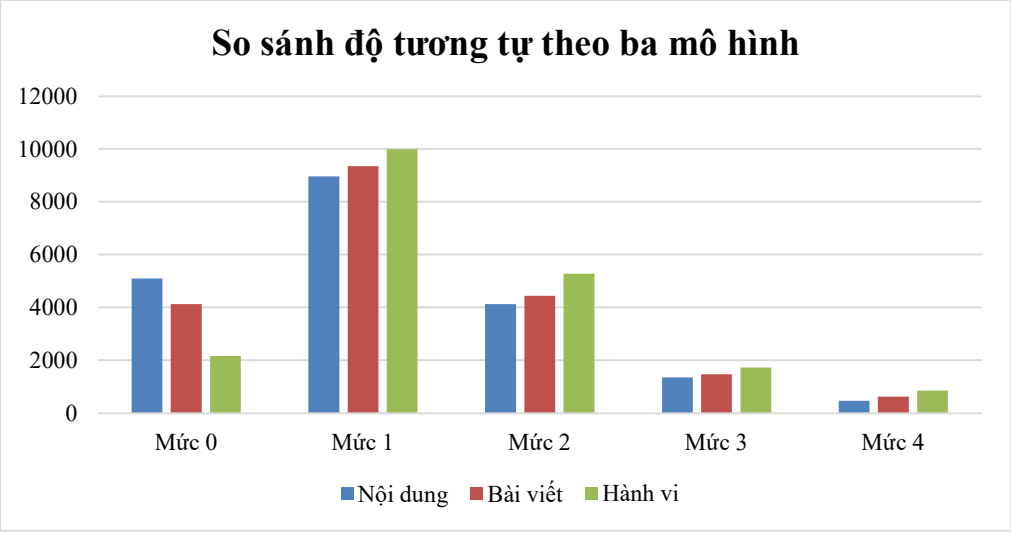
số của các hành vi, sau khi phân loại theo mức độ tương tự được minh họa như trong Bảng 4.6.

Kết quả thực nghiệm thu được dựa trên các mức tương tự theo không gian bài viết các mức như sau:

Bảng 4.6: Nhóm các cặp người dùng theo độ tương tự

	Mức 0	Mức 1	Mức 2	Mức 3	Mức 4
Chỉ có nội dung bài viết	5094	8961	4122	1355	468
Bài viết có 5 đặc trưng	4123	9350	4437	1469	621
Theo mô hình hành vi	2758	9955	5037	1521	729

Nhìn vào kết quả Bảng 4.6 có thể thấy rằng độ tương tự theo của người dùng theo mô hình hành vi xác định được nhiều cặp tương tự nhau hơn, số lượng các cặp người dùng ở mức 0 là thấp nhất chỉ bằng khoảng ½ so với mô hình người dùng chỉ có nội dung bài viết. Điều này được mình họa rõ nét trong Hình 4.3.



Hình 4.3: So sánh độ tương tự giữa hai người dùng

d. Xác định các chủ đề quan tâm của người dùng theo hành vi

- **Kịch bản thực nghiệm:** Kịch bản thực nghiệm xác định các chủ đề quan tâm của người dùng được thực hiện như sau:

Đầu vào: Danh sách 200 người dùng với 2000 bài viết theo hành vi đăng bài, 2000 bài viết theo hành vi thích, 800 nhóm theo hành vi gia nhập nhóm và 21 chủ đề dùng làm nhãn để phân loại

Đầu ra: Tương quan giữa người dùng và 21 chủ đề

Thực hiện

- Xây dựng danh sách từ, thuật ngữ cho các bài viết, các nhóm theo các chủ đề, các chủ đề đã xây dựng trong chương hai

- Tính véctơ trọng số theo TF.IDF cho mỗi bài viết, nhóm và mỗi chủ đề

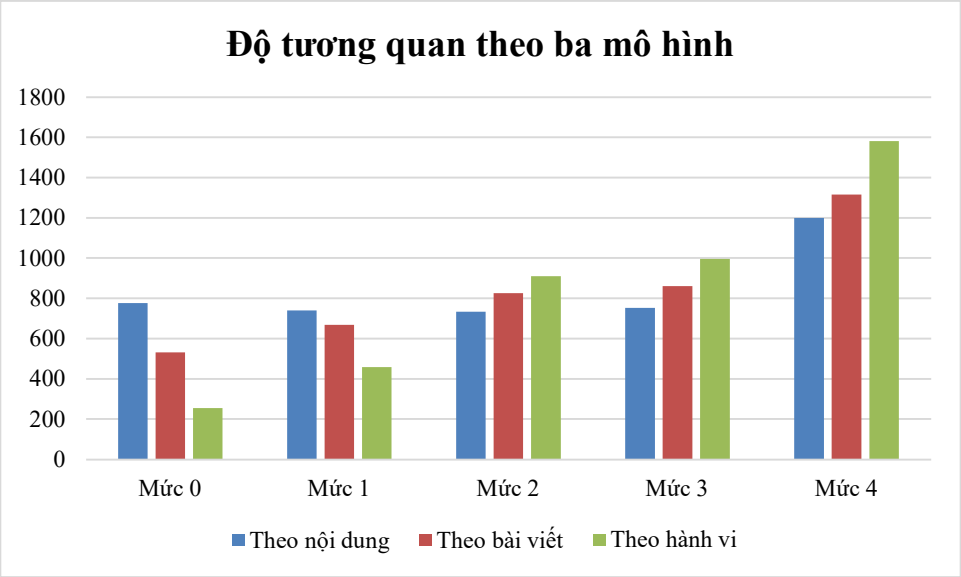
- Tính độ tương quan giữa mỗi bài viết, nhóm và các chủ đề, tích hợp theo trọng số cho mỗi người dùng

- Phân loại các theo 21 chủ đề dựa trên độ đo tương quan

Kết thúc

- **Tham số đầu ra:** là độ tương quan của mỗi người dùng với 21 chủ đề
- **Các bước thực hiện chi tiết như sau:** Bước 1 và Bước 2: Xây dựng danh sách từ, thuật ngữ cho bài viết và tính véctơ trọng số theo TF.IDF theo không gian của chủ đề. Bước 3: Tính độ tương quan giữa các bài viết, nhóm với các chủ đề. Sau đó tích hợp tính theo hành vi dựa trên bộ trọng số. Bước 4 Tính và phân loại độ quan tâm của người dùng theo các chủ đề dựa trên Định nghĩa 2.6 và các công thức (4.12), (4.13), (4.14) và (4.15). Tại bước này, luận án tính mỗi công thức sẽ thu được một bảng gồm có 200 dòng và 21 cột, trong đó mỗi ô là độ quan tâm của

người dùng của hàng đó đến chủ đề tương ứng. Kết quả thực nghiệm thu được dựa trên các mức tương tự theo không gian chủ đề theo các mức.



Hình 4.4: So sánh mức độ tương quan giữa người dùng và chủ đề

Theo kết quả thu được dựa trên tổng số các người dùng quan tâm theo các chủ đề cũng thấy rõ, các mức 0 và mức 1 đều giảm, và các mức liên quan ở mức 2, mức 3 và mức 4 đều tăng mạnh, đặc biệt mức 4 tăng nhanh. Điều đó cho thấy rằng, biểu diễn người dùng theo hành vi thì việc xác định các chủ đề quan tâm của người dùng cho kết quả tốt hơn.

4.4.3. Thảo luận về kết quả thực nghiệm

Độ chính xác của thực nghiệm xác định chủ đề quan tâm của người dùng được tính dựa trên *Sai số bình phương trung bình (*MSE - Mean Square Error) như công thức (2.19). Với cách tính này thì MSE càng gần đến giá trị 0 thì độ chính xác càng cao và ngược lại. Kết quả:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2 = \frac{1}{4200}\sum_{i=1}^{4200}(p_i - r_i)^2 = 0.0913$$

Tương ứng với độ chính xác của thực nghiệm là:

$$CR = (1\text{-}MSE) * 100\% = (1 - 0.0913) * 100\% = 90{,}87\% \qquad (4.21)$$

Độ chính xác của thực nghiệm xác định độ tương tự của các cặp người dùng cũng được tính dựa trên *Sai số bình phương trung bình (*MSE - Mean Square Error). Kết quả :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2 = \frac{1}{20000}\sum_{i=1}^{20000}(p_i - r_i)^2 = 0.0733$$

$$CR = (1\text{-}MSE) * 100\% = (1 - 0.0733) * 100\% = 92.67\% \qquad (4.22)$$

Như vậy, kết quả thực hiện trên nội dung của bài viết đã tăng lên đáng kể với kết quả đạt được là 90.87% và 92.67% so với mô hình bài viết là 81.8% và 86.9%, và so với độ chính xác khi chỉ phân tích nội dung là 68.5% và 70.8%. Điều này chứng tỏ, mô hình người dùng hành vi mang lại kết quả phân tích tốt hơn rất nhiều so với chỉ xét nội dung của bài viết, kết quả trình bày trong Bảng 4.7

Bảng 4.7: Độ chính xác của các mô hình

Độ chính xác	**Tương quan với chủ đề**	**Tương tự người dùng**
Chỉ có nội dung bài viết	68.50%	70.80%
Bài viết có 5 đặc trưng	81.81%	86.91%
Theo mô hình hành vi	90.87%	92.67%

Ngoài ra, khi xem xét độ tương quan giữa các độ đo trong các nhóm người dùng, luận án sử dụng cách thức thống kê số lượng những người có cùng mức độ tương tự theo không gian bài viết sau và những người có cùng độ tương tự theo không gian các chủ đề để tìm sự tương quan giữa hai độ đo này.

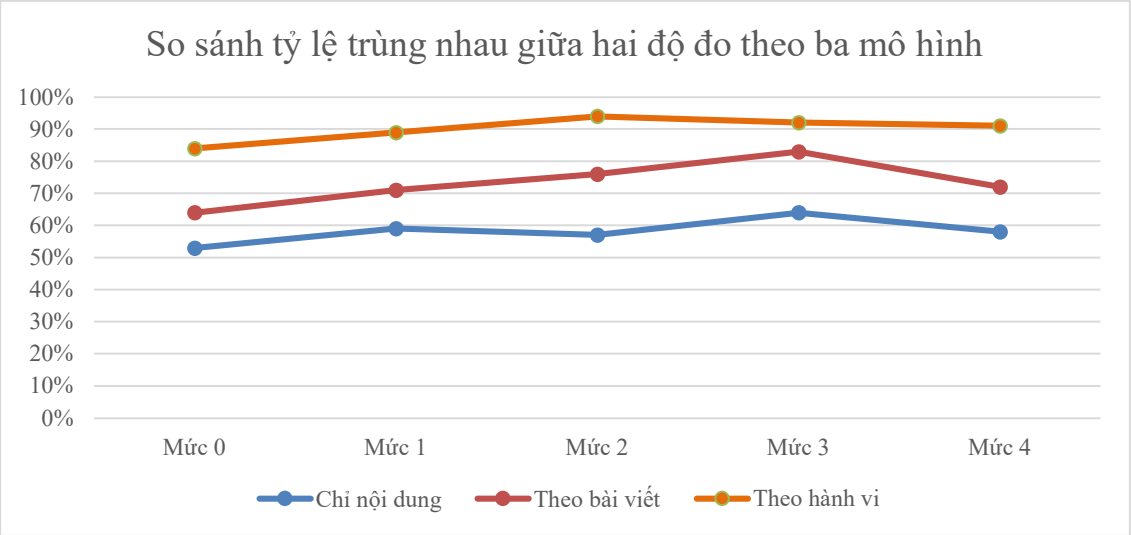
Kết quả được như Bảng 4.8 cho thấy rằng tỷ lệ trùng nhau giữa các tập hợp đã tăng lên rất nhiều. Các mức khác không đều đạt trên 90%.

Có nghĩa là nếu hai người dùng có độ tương tự theo hành vi thì trên 90% khả năng có các chủ đề quan tâm tương tự nhau.

Bảng 4.8: Tỷ lệ trùng nhau theo các mô hình

Mô hình	Mức 0	Mức 1	Mức 2	Mức 3	Mức 4
Chỉ có nội dung bài viết	**53.08%**	**58.82%**	**57.00%**	**63.78%**	**58.26%**
Bài viết có 5 đặc trưng	**63.56%**	**70.53%**	**75.56%**	**83.41%**	**71.54%**
Theo mô hình hành vi	**83.06%**	**89.21%**	**94.41%**	**92.52%**	**91.45%**

Minh họa trong Hình 4.5. có thể thấy rõ sự vượt trội của mô hình hành vi so với mô hình bài viết mở rộng và mô hình bài viết chỉ có nội dung



Hình 4.5: So sánh tỷ lệ trùng nhau giữa hai độ đo theo ba mô hình

Trong Hình 4.5 có thể thấy rằng, tỷ lệ trùng nhau giữa các mức đều tăng theo tỷ lệ thuận, nghĩa là các mức tương tự nhau đều tăng lên khi xem xét các thêm đối tượng dữ liệu và mở rộng thêm các hành vi.

4.5. SO SÁNH VỚI MỘT SỐ MÔ HÌNH KHÁC

4.5.1. Các mô hình so sánh

Luận án thực hiện việc so sánh kết quả thực hiện mô hình với 03 mô hình tính toán dựa trên TF.IDF và dữ liệu là văn bản ngắn gồm: Mô hình ước lượng độ quan tâm dựa trên thẻ đánh dấu của Sheng Bin et al. [125]; Mô hình ước lượng phát hiện các chủ đề quan tâm của người dùng dựa trên các Tweet của Hossen M. F. et al. [63]

và mô hình ước lượng chủ đề quan tâm dựa trên hành vi đăng bài (post) và hành vi thích (like) của Kim J. Ko et al. [77].

Các mô hình này được lựa chọn để so sánh với mô hình đề xuất trong luận án do chúng đều phân tích dữ liệu văn bản ngắn trên các mạng xã hội, biểu diễn bằng giá trị bằng các véctơ hoặc trọng số tính bằng TF.IDF, có phân tách từ và phân tách bằng N-gram. Các mô hình đều được sử dụng để xác định các chủ đề quan tâm của người dùng trên mạng xã hội.

Mô hình của Sheng Bin et al. [125]: Trong nghiên cứu này, Sheng Bin et al. phân tích các thẻ đánh dấu (tag) thay vì nội dung trên Douban (một mạng xã hội phổ biến ở Trung Quốc, Trung Quốc không dùng Facebook). Mỗi sản phẩm giải trí (media item) trên Douban thường là một bộ phim, một bài hát, một album 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. Mỗi một người dùng có thể tự tạo ra một thẻ đánh dấu cho sản phẩm giải trí mà họ muốn chia sẻ trên mạng xã hội này.

Trong nghiên cứu này, thay vì phân tích nội dung và mô tả của sản phẩm giải trí thì nhóm nghiên cứu đi thu thập và phân tích các thẻ đánh dấu mà người dùng tạo ra cho sản phẩm giải trí đó.

Nhóm nghiên cứu sử dụng không gian véctơ SVM để biểu diễn mỗi sản phẩm giải trí, mỗi sản phẩm giải trí này được biểu diễn bằng 02 véctơ: Một véctơ chứa từ khóa của các thẻ đánh dấu và một véc tơ 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:

$$a_{ij}^{tf} = \frac{f_{ij}}{\sqrt{\sum_{k=1}^{t} f_{ij}^2}} \tag{4.23}$$

và *TF.IDF* của một từ *i* trong tài liệu *j* được tính bằng:

$$a_{ij}^{tf\times idf} = \frac{b_{ij}}{\sqrt{\sum_{k=1}^{t} b_{ij}^2}} \tag{4.24}$$

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

Sau đó nhóm nghiên cứu biểu diễn các chủ đề của các sản phẩm giải trí thành một tập các thẻ đánh dấu, sau đó thực hiện việc so sánh và phân loại các véctơ theo hai nhóm: một nhóm là các sản phẩm giải trí có chứa tất cả các thẻ đánh dấu về chủ đề và một nhóm là những người dùng mà họ đã sử dụng các thẻ đánh dấu của chủ đề. Như vậy, sẽ tìm được sự tương quan giữa các người dùng và các chủ đề dựa trên các thẻ đánh dấu.

Mô hình tính toán của Hossen M. F. et al. [63]: Mô hình của nhóm nghiên cứu thực hiện xây dựng một mô hình gọi là EICV (Entity Intersect Categorized Value) trong đó W_n là tập hợp các từ, R_n là các từ dư thừa, K_n là tập hợp các từ trích chọn từ tri thức của các thực thể, T_n là 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 tính:

$$K_n = F_{twitter}(W_n - R_n) \tag{4.25}$$

Trong đó, $F_{twitter}$ là một hàm lấy các thực thể từ mạng Twitter, hay là API dùng để lấy dữ liệu từ mạng Twitter. Khi đó giá trị giao được tính bằng:

$$V_i = number\ of\ (K_n \cap T_{tag[n]}) \tag{4.26}$$

Với mỗi chủ đề $t_k \in T_n$ nó tạo ra một giá trị V_i. Nhóm nghiên cứu đặt một giá trị ngưỡng T_v và lấy các chủ đề lớn hơn ngưỡng này để xem là các chủ đề quan tâm của người dùng.

Mô hình tính toán của Kim J. Ko et al. [77]: Nhóm nghiên cứu xây dựng một mô hình để phân tích hành vi đăng bài và hành vi thích của người dùng dựa trên tính

trọng số TF.IDF để xác định các chủ đề quan tâm của người dùng. Với 715 bài đăng tương ứng với 715 hành vi đăng bài và 1477 lượt thích tương ứng.

Mô hình phân tích bài đăng tương ứng với hành vi đăng bài và tính thêm số lượt thích kèm theo bài đăng đó để tính trọng số của véctơ hành vi đăng bài. Trọng số của hành vi đăng bài được tính bằng tần suất của các danh từ xuất hiện trong bài đăng và số lần thích của bài đăng đó. Khi đó trọng số quan tâm của người dùng được Kim J. Ko et al. [77] tính bằng:

$$UIW = log(TF + 1) \times (log(like + 1) + 1) \tag{4.27}$$

Trong đó, *TF* là tần số xuất hiện của danh từ trong bài viết và *like* là số lượt thích của bài viết đang xét. Các danh từ sau đó được so sánh với tên các chủ đề trên mạng xã hội và được coi là chủ đề quan tâm của người dùng đó.

4.5.2. Các bước thực hiện

Các mô hình này và mô hình của luận án đề xuất được cài đặt trên Python version 3.8 trên hệ điều hành Windows 10. Môi trường soạn thảo và thực thi mã nguồn IDE Python là Wing 101 version 7.2.50. Trong quá trình thực hiện luận án có sử dụng một số thư viện và một số mã nguồn trên hệ thống Python online như chuẩn hóa từ vựng từ Python's Natural Language Toolkit (http://www.nltk.org)

Để so sánh kết quả của các mô hình, luận án tiến hành xây dựng bộ dữ liệu thử nghiệm trên nhóm người dùng với 21 chủ đề đã giới thiệu trong chương hai và tập mẫu dữ liệu thu thập trên mạng xã hội Facebook.

Cách thức thực hiện thực nghiệm nhằm mục đích so sánh độ chính xác của các mô hình được tính bằng số lượng các chủ đề mô hình xác định được so với số lượng các chủ đề đã được gán nhãn trước cho các người dùng.

Mô hình của ShengBin et al [125] được luận án ước lượng dựa trên trọng số của các bài viết có chứa thẻ đánh dấu, sau đó phân tách theo N-Gram và tính TF.IDF dựa

trên các bài viết theo các chủ đề trong phần đăng bài của người dùng. Sau đó xác định các người dùng tương quan với các chủ đề theo véctơ trọng số này.

Mô hình của Hossen M. F. et al. [63] được luận án ước lượng từ bài đăng của người dùng, sau đó tách theo N-gram và lấy giao của tập từ thu được với tập từ các chủ đề. Nếu giữa bài viết và chủ đề không có bất kỳ từ nào trùng nhau thì bài viết đó không liên quan đến chủ đề đang xét, và nếu một người dùng không có bất kỳ bài viết nào liên quan đến chủ đề đang xét thì được coi là không liên quan đến chủ đề đó.

Mô hình của Kim J. Ko et al. [77] được luận án ước lượng dựa trên các bài đăng (hành vi đăng bài) cùng với số lượng lượt thích của bài đăng đó và tính theo công thức (4.27) để đưa ra véctơ trọng số và sắp xếp các danh từ theo bài đăng dựa trên trọng số này, sau đó sắp xếp các từ theo trọng số giảm dần rồi xét các từ trong các chủ đề để xem xét và ước lượng các chủ đề quan tâm của người dùng.

Mẫu thử nghiệm được xây dựng như sau: Với 200 người và 2000 bài viết tương ứng (mỗi người có 10 bài viết đã đăng có thẻ đánh dấu, riêng mô hình đề xuất có tính thêm mỗi người dùng 05 bài viết đã thích và 05 bài viết đã tham gia bình luận và 05 nhóm đã tham gia).

Trong đó, bài viết _ei có tương quan với chủ đề tj thì giá trị được gán là 1, nếu không có tương quan được gán bằng 0. Mẫu được các tình nguyện viên đánh giá trước với tổng bộ mẫu là 200 user với mỗi user 10 bài viết.

Bảng 4.9: Giá trị một mẫu của mô hình

User01						
	_t1	_t2	…	_t19	_t20	_t21
_e01	1	0	…	1	0	0
_e02	0	0	…	1	1	0
…	…	…	…	…	…	…
_e09	0	0	…	0	1	1
_e10	1	1	…	0	0	1

Mỗi bài viết được tách thành các từ theo các kỹ thuật được mô tả ngắn gọn trong Bảng 4.10. Sau đó thực hiện tiền xử lý để tính toán và ước lượng theo các bước gồm tách từ, loại bỏ từ dừng, tính trọng số và lưu vào véctơ.

Các chủ đề được tách từ, loại bỏ từ dừng và thu được tập các từ của các chủ đề, sau đó tính trọng số trong không gian các chủ đề. Mỗi mẫu được biểu diễn thành một bảng gồm một user, 10 bài viết, 21 chủ đề như Bảng 4.9

Do chỉ so sánh về các chủ đề mà người dùng quan tâm nên luận án không so sánh các mức độ chênh lệch giữa các độ quan tâm của người dùng theo các chủ đề.

Có thể tóm tắt các kỹ thuật để so sánh mô hình đề xuất trong luận án và 03 mô hình đã trình bày ở trên Bảng 4.10

Bảng 4.10: Kỹ thuật tính toán của các mô hình

Tên mô hình	**Mô hình của ShengBin et al [125]**	**Mô hình của Hossen M. F. et al. [63]**	**Mô hình của Kim J. Ko et al. [77]**	**Mô hình đề xuất**
Đối tượng phân tích	Bài viết chứa thẻ đánh dấu	Bài viết được đăng	Bài viết được đăng và số lượt thích	Bài viết đã đăng, đã thích, đã bình luận và nhóm đã tham gia
Kỹ thuật tách từ	N-gram Loại từ dừng	Dựa trên dấu cách Loại từ dừng	N-gram Loại từ dừng	N-gram Loại từ dừng
Giá trị ước lượng	TF.IDF	Words	UIW	TF.IDF
Giá trị so sánh	Độ tương tự Cosine	Dựa trên phép giao của hai tập hợp	Xếp hạng và lấy giao của hai tập hợp	Độ tương tự Cosine

Kịch bản thực nghiệm được sử dụng cho mỗi mô hình như sau: Với mỗi bài viết được các tình nguyện viên thực hiện gán các nhãn theo các chủ đề, mỗi bài viết có thể được gán nhiều nhãn từ danh sách các định nghĩa của các chủ đề xem xét trong luận án. Tổng hợp các chủ đề quan tâm của mỗi người dùng được tính theo giá trị so sánh và dựa trên ngưỡng, nếu một chủ đề được xác định theo

Độ chính xác được tính bằng:

$$\text{Precision} = \frac{\text{Tổng số mẫu đúng}}{\text{Tổng số số mẫu}} * 100\% \quad (4.28)$$

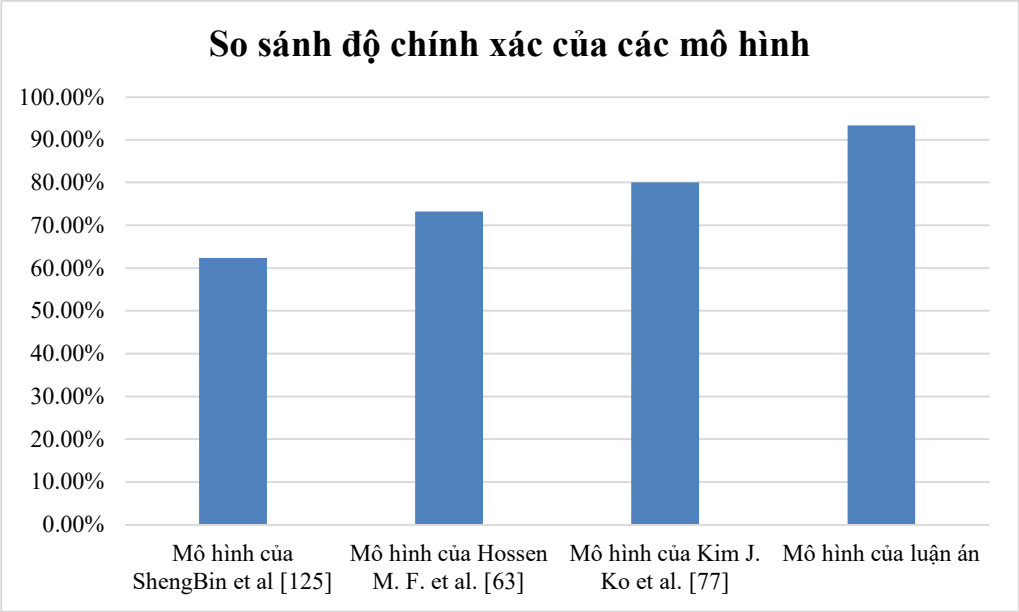
4.5.3. Kết quả so sánh các mô hình và thảo luận

Sau khi thực hiện kết quả thu được như trong Bảng 4.11

Bảng 4.11: Độ chính xác so sánh giữa các mô hình

Tên mô hình	Mô hình của ShengBin et al [125]	Mô hình của Hossen M. F. et al. [63]	Mô hình của Kim J. Ko et al. [77]	Mô hình đề xuất
Độ chính xác	62.40%	73.25%	80.06%	93.35%

Minh họa các kết quả so sánh giữa các mô hình trong Hình 4.6.



Hình 4.6: Kết quả so sánh các mô hình

Qua kết quả trên có thể thấy rằng, mô hình của ShengBin et al [125] có độ chính xác thấp nhất là 62.40% nghĩa là có (1248/2000 mẫu đúng so với giá trị đã xác định), điều này có thể giải thích vì mô hình của ShengBin et al. dựa trên thẻ đánh dấu, trên

mạng xã hội Facebook thẻ đánh dấu thường ngắn và có ít từ, vì vậy, khi tính TF.IDF với các chủ đề thì giá trị sẽ nhỏ.

Mô hình của Hossen M. F. et al. [63] có độ chính xác cao hơn là đạt 73.25% (1465/2000 mẫu) vì xét đến nội dung bài viết, các bài viết trên FB sẽ có số lượng từ cao hơn, mặt khác, trong mô hình này Hossen M. F. et al so sánh dựa trên hai tập hợp và lấy giao của tập các từ trong nội dung bài viết và tập các từ trong định nghĩa của chủ đề, nên độ chính xác của mô hình cao hơn.

Mô hình của Kim J. Ko et al. [77] xét đến cả nội dung và số lượt thích để xem xét và xếp hạng dựa trên các danh từ tìm được trong bài viết nên có độ chính xác cao hơn với độ chính xác là 80.06% (1612/2000 mẫu). Tuy nhiên, thời gian phân tích lâu hơn vì so sánh tất cả các danh từ với tất cả các từ trong tập từ của chủ đề, sau đó cần thêm thao tác xếp hạng từ và đếm số lượt thích.

Mô hình đề xuất của luận án cho kết quả tốt nhất với độ chính xác là 93.35% (1867/2000 mẫu) có thể giải thích do xét các bài viết đã đăng, đã thích, đã bình luận và thêm các nhóm đã tham gia nên độ chính xác khi so sánh được tăng hơn rất nhiều và thời gian phân tích là lâu hơn. Nếu đã thực hiện tiền xử lý trước khi thực hiện so sánh thì thời gian thực hiện không chênh quá nhiều, nhưng nếu thực hiện trực tuyến hoặc trực tiếp trên hệ thống cho từng người dùng thì thời gian thực hiện mô hình đề xuất của luận án là lâu nhất.

4.6. KẾT LUẬN

Chương bốn của luận án đã trình bày hành vi của người dùng trên các mạng xã hội, phân loại các kiểu hành vi cũng như các nghiên cứu liên quan đến bài toán phát hiện quan tâm của người dùng dựa trên các hành vi. Sau đó, luận án đề xuất mô hình biểu diễn người dùng dựa trên bốn hành vi là: hành vi đăng bài viết, hành vi thích bài viết, hành vi bình luận trong bài viết và hành vi gia nhập nhóm trên mạng xã hội. Dựa trên mô hình hành vi này, luận án đưa ra cách biểu diễn mô hình người dùng dựa trên các hành vi với cách thức biểu diễn véctơ dựa trên trọng số. Từ mô hình biểu diễn

người dùng đã đề xuất, luận án đưa ra cách thức ước lượng hai độ đo: Độ đo tương tự giữa hai người dùng và mức độ quan tâm của người dùng theo các chủ đề. Dựa trên hai độ đo này, luận án xem xét đến sự tương quan giữa độ tương tự của hai người dùng theo không gian các hành vi và độ quan tâm các chủ đề của người dùng. Luận án đã tiến hành thực nghiệm để tìm được bộ trọng số tối ưu cho bốn hành vi trong mô hình người dùng, từ đó, tiến hành hai thực nghiệm là xác định độ tương tự người dùng theo hành vi và xác định độ quan tâm của người dùng theo chủ đề dựa theo các hành vi. Các kết quả thực nghiệm đã chỉ ra rằng, nếu hai người dùng có độ tương tự nhau theo hành vi thì các chủ đề quan tâm trên mạng xã hội cũng tương tự nhau và ngược lại, nếu hai người dùng có các chủ đề quan tâm tương tự nhau thì họ cũng có nhiều hành vi tương tự nhau trên mạng xã hội. Ngoài ra, luận án cũng thực hiện thực nghiệm so sánh mô hình đề xuất với 03 mô hình về xác định chủ đề quan tâm của người dùng dựa trên phân tích hành vi và trạng thái đăng bài. Kết quả cho thấy mô hình luận án đề xuất cho kết quả tốt hơn 03 mô hình đã so sánh. Các kết quả nghiên cứu liên quan đến chương bốn đã được công bố trong một số Tạp chí Khoa học chuyên ngành và Kỷ yếu Hội nghị Khoa học uy tín, bao gồm: Kết quả của các nghiên cứu này đã được công bố trên Tạp chí *Vietnam Journal of Computer Science*, (2018)5:165–175, Springer Open. Tạp chí *Journal of Science and Technology on Information and Communications (JSTIC), No. 3-4, 2018* và trên Kỷ yếu Hội nghị quốc gia lần 11 về Nghiên cứu Cơ bản và Ứng dụng *(11th National Symposium on Fundamental and Applied IT Research – FAIR'11, 08-2018).*

KẾT LUẬN

Những kết quả nghiên cứu của luận án

Mục tiêu của luận án là nghiên cứu bài toán phát hiện quan tâm của người dùng dựa trên dữ liệu từ các bài viết và hành vi của người dùng trên mạng xã hội. Bài viết và hành vi của người dùng trên mạng xã hội thể hiện cách ứng xử của người dùng đối với các đối tượng, các sự kiện và các chủ đề mà họ quan tâm. Mỗi bài viết và hành vi của người dùng được luận án mở rộng ngữ nghĩa theo từ điển Wikipedia, sau đó được véctơ hóa theo trọng số TF.IDF của các bài viết và các chủ đề. Các độ đo tương tự giữa các véctơ bài viết và giữa các véctơ hành vi của người dùng trên mạng xã hội với các véctơ chủ đề được luận án sử dụng để xác định quan tâm của người dùng và phân loại người dùng dựa trên những chủ đề quan tâm này. Các đóng góp chính của luận án bao gồm:

- Đề xuất mô hình biểu diễn bài viết của người dùng trên mạng xã hội dựa trên năm đặc trưng là nội dung, thể loại, thẻ đánh dấu, quan điểm và cảm xúc. Mỗi bài viết được tính toán, mở rộng ngữ nghĩa theo Wikipedia và biểu diễn dưới dạng một véctơ có trọng số theo TF.IDF theo các đặc trưng của chúng.
- Đề xuất mô hình biểu diễn hành vi của người dùng dựa trên các hành vi đăng/chia sẻ bài viết, hành vi thích bài viết, bình luận trong bài viết và hành vi gia nhập nhóm/cộng đồng trên mạng xã hội.
- Đề xuất cách xác định các chủ đề quan tâm của người dùng dựa trên ước lượng độ tương quan giữa các bài viết của người dùng với các chủ đề. Độ tương quan giữa tập hợp các bài viết của người dùng với các chủ đề là mức độ quan tâm của người dùng đến các chủ đề đó trên mạng xã hội.
- Đề xuất cách thức ước lượng độ tương tự hai người dùng theo mô hình bài viết và mô hình hành vi. Độ tương tự giữa hai người dùng theo mô hình bài viết được tính dựa trên việc tích hợp có trọng số độ tương tự các đặc trưng của bài viết và giữa hai tập bài viết của người dùng. Độ tương tự giữa hai

người dùng theo hành vi cũng được tính dựa trên tích hợp có trọng số độ tương tự giữa các hành vi của người dùng.

- Luận án cũng thực hiện thu thập các bộ dữ liệu thực từ ba mạng xã hội là Facebook.com, Twitter.com và YouTube.com để thực hiện kiểm nghiệm các mô hình và cách thức ước lượng đã đề xuất trong chương 2, chương 3 và chương 4 của luận án. Luận án đã thực hiện bốn nhóm thực nghiệm chính, bao gồm:
    - Thực nghiệm nhằm lựa chọn thuật toán gán nhãn phân loại văn bản phù hợp nhất để xác định các giá trị của các đặc trưng thể loại, cảm xúc và quan điểm của bài viết trên mạng xã hội. Từ thực nghiệm đã chỉ ra rằng thuật toán gán nhãn MNB là phù hợp nhất cho việc gán nhãn dữ liệu văn bản trên các trang mạng xã hội.
    - Nhóm thực nghiệm ước lượng độ tương tự giữa hai người dùng theo mô hình bài viết và theo mô hình hành vi. Mục đích của các thực nghiệm dùng để phân nhóm người dùng trên các trang mạng xã hội dựa trên độ tương tự theo các bài viết và các hành vi của người dùng
    - Nhóm thực nghiệm xác định các chủ đề quan tâm của người dùng theo mô hình bài viết và mô hình hành vi. Mục đích của các thực nghiệm nhằm xác định các chủ đề quan tâm của người dùng và phân loại các người dùng theo các mức độ quan tâm đến các chủ đề trên mạng xã hội. Từ đó, xem xét mối tương quan giữa hai độ đo tương tự theo người dùng và tương tự các chủ đề quan tâm trên mạng xã hội.
    - Nhóm thực nghiệm so sánh mô hình đề xuất với 03 mô hình xác định chủ đề quan tâm của người dùng trên mạng xã hội
    - Nhóm thực nghiệm cuối cùng dùng để xác định hai bộ trọng số: Bộ trọng số thứ nhất là dùng để tính toán trong mô hình người dùng dựa trên bài viết có năm đặc trưng; Bộ trọng số thứ hai dùng

để xác định bộ trọng số cho bốn hành vi trong mô hình biểu diễn người dùng dựa trên các hành vi.

Ý nghĩa và khả năng ứng dụng vào thực tiễn

Phát hiện quan tâm của người dùng trên các trang mạng xã hội là một bài toán có vai trò quan trọng trong nghiên cứu cũng như ứng dụng thực tiễn đối với các hoạt động của các tổ chức, doanh nghiệp, đặc biệt các tổ chức, doanh nghiệp kinh doanh trực tuyến.

Phát hiện quan tâm của người dùng dựa trên các hành vi của họ là một bài toán có tính ứng dụng cao, bởi các hành vi của người dùng trên mạng xã hội có thể gián tiếp thể hiện quan tâm của người dùng trên mạng xã hội. Việc phát hiện 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. Khả năng lan truyền và chia sẻ nội dung trên mạng xã hội cũng giúp cho các tổ chức, doanh nghiệp thu được kết quả tốt hơn trong các chiến dịch quảng cáo và giới thiệu sản phẩm. Việc xác định được quan tâm của người dùng trên mạng xã hội 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 trong các hoạt động quảng bá, giới thiệu sản phẩm.

Với các mô hình bài viết và mô hình hành vi của người dùng trên mạng xã hội đã được đề xuất trong luận án, những mô hình này có thể làm cơ sở cho các nghiên cứu như phân loại người dùng, phát hiện quan tâm của người dùng, phát hiện quan điểm, cảm xúc của người dùng, phân nhóm khách hàng, ...

Trên thực tế, bài toán phân nhóm khách hàng là một trong các bài toán trọng điểm trong hoạt động quản trị quan hệ khách hàng của tổ chức, doanh nghiệp, đặc biệt đối với các hoạt động kinh doanh trực tuyến, do đó các tổ chức, doanh nghiệp xác định được quan tâm của khách hàng sẽ rút ngắn được thời gian phân nhóm khách hàng, tiến gần hơn đến các hoạt động chăm sóc cá nhân hóa khách hàng trong tổ chức.

Ngoài ra, khi phân nhóm được khách hàng, các tổ chức, doanh nghiệp sẽ tiết kiệm được nhiều tài nguyên trong các chiến dịch quảng cáo và tiếp thị trực tuyến, tăng khả năng phản hồi của khách hàng.

Những vấn đề còn hạn chế của luận án

Ngoài những đóng góp chính của luận án, do thời gian và mục tiêu của luận án nên luận án vẫn còn những vấn đề cần tiếp tục nghiên cứu và cải thiện bao gồm:

Thứ nhất, luận án tập trung nghiên cứu phân tích dữ liệu văn bản mà chưa đề cập đến việc phân tích các kiểu dữ liệu khác như hình ảnh, video, … trên mạng xã hội.

Thứ hai, trong luận án chưa đề cập đến yếu tố thời gian, quan tâm của người dùng có thể thay đổi theo thời gian, vì vậy, hướng nghiên cứu tiếp theo có thể mở rộng thêm yếu tố thời gian vào các đặc trưng để phân tích và ước lượng.

Mặc dù mạng xã hội là phổ biến nhưng những thông tin của người dùng trên mạng xã hội là tài sản và thuộc sở hữu riêng của người dùng, việc tính toán, nghiên cứu và ước lượng về quan tâm của người dùng chỉ có thể thực hiện trên các dữ liệu được công bố của người dùng, do đó không tránh khỏi những kết quả mang tính chất dựa trên số liệu.

Hướng nghiên cứu tiếp theo

Từ những kết quả nghiên cứu đã thực hiện luận án đề xuất một số hướng nghiên cứu mở rộng từ kết quả nghiên cứu của luận án như sau:

Thứ nhất, luận án có thể bổ sung và mở rộng thêm các kiểu dữ liệu cho nghiên cứu như dữ liệu ảnh, dữ liệu video hoặc các dữ liệu khác trong các bài viết của người dùng trên mạng xã hội;

Thứ hai, luận án có thể bổ sung hoặc thêm yếu tố thời gian trong các bài viết hoặc hành vi nghiên cứu và phân tích. Yếu tố thời gian có thể đưa vào thành một thuộc tính hoặc nghiên cứu dựa trên thực nghiệm bằng cách thu thập dữ liệu vào các

khoảng thời gian khác nhau của cùng nhóm người dùng để thực nghiệm. Vì vậy, nghiên cứu phân tích để phát hiện quan tâm của người dùng theo thời gian cũng là một hướng tiếp cận thú vị;

DANH MỤC CÁC CÔNG TRÌNH NGHIÊN CỦA TÁC GIẢ LIÊN QUAN ĐẾN LUẬN ÁN

TẠP CHÍ KHOA HỌC

[1]. Manh Hung Nguyen, Thi Hoi Nguyen. *A general model for similarity measurement between objects*. International Journal of Advanced Computer Science and Applications (IJACSA), 6(2):235 - 239, 2015.

[2]. Thi Hoi Nguyen, Dinh Que Tran, Gia Manh Dam, Manh Hung Nguyen, *Estimating the similarity of social network users based on behaviors*, Vietnam Journal of Computer Science (2018) 5:165–175, Springer Opens .

[3]. Nguyễn Thị Hội, Trần Đình Quế, *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 (JST-UD), Trường Đại học Đà Nẵng, ISSN 1859-1531 – Số 7(128). 2018

[4]. Nguyen Thi Hoi, Tran Dinh Que, *Estimating user's interest on social networks based on behaviors*, Journal of Science and Technology on Information and Communications, Vol 3, CS.01 (2018), 9-15, ISSN 2525 – 2224

[5]. Dinh Que Tran, Thi Hoi Nguyen, Phuong Thanh Pham, *Modeling user's interests, similarity and trustworthiness based on vectors of entries in social networks*, Southeast Asian Journal of Sciences, Vol. 09, No 1 (2020), pp. 01–10

HỘI THẢO KHOA HỌC

[6]. Thi Hoi Nguyen, Dinh Que Tran, Gia Manh Dam, and Manh Hung Nguyen. *Multi-feature Based Similarity Among Entries on Media Portals*, Advances in Information and Communication Technology, Proceedings of the International Conference, ICTA 12 - 2016, Advances in Intelligent Systems and Computing, ISBN 978-3-319-49072-4, Springer International Publishing. Advances in Intelligent Systems and Computing, 538 AISC, pp. 373-382, (2017).

[7]. Nguyen, Thi Hoi; Tran, Dinh Que; Dam, Gia Manh; Nguyen, Manh Hung. *Integrated Sentiment and Emotion into Estimating the Similarity among Entries on Social Network*, 3rd EAI International Conference on Industrial Networks and Intelligent Systems. Sep 4, 2017, Springer International Publishing. Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST, 221, pp. 242-253, (2018).

[8]. Nguyễn Thị Hội, Đàm Gia Mạnh, Trần Đình Quế. *Độ tương đồng ngữ nghĩa các bài viết trên mạng xã hội dựa trên Wikipedia*, Kỷ yếu Hội thảo Fundamental and Applied IT Research - FAIR'10, Đà Nẵng 08/2017, NXB Khoa học Tự nhiên và Công nghệ.

[9]. Nguyễn Thị Hội, Trần Đình Quế. *Ước lượng tương tự quan tâm người dùng trên mạng xã hội dựa vào các nhóm tham gia*, Kỷ yếu Hội thảo Fundamental and Applied IT Research - FAIR'11, Hà Nội 08/2018, NXB KHTN và CN

TÀI LIỆU THAM KHẢO

[1] A. Abdul-Rahim, et al. (2014),"Determinants of Online Buying Behavior of Social Media Users in Saudi Arabia: An Exploratory Study", SSRN Electronic Journal, DOI:10.2139/ssrn.2519254, India.

[2] A. Basma, et al. (2016), "Interest Aware Location-Based Recommender System Using Geo-Tagged Social Media", ISPRS International Journal of Geo-Information, vol. 5, no. 12, p. 245.

[3] A. Ezgi and S. Mardikyan (2014), "Analyzing factors affecting users' behavior intention to use social media: Twitter case", International Journal of Business and Social Science, p. 5-11.

[4] A. Gattani, et al. (2013), "Entity Extraction, Linking, Classification, and Tagging for Social Media: A Wikipedia-based Aproach", Proc. VLDB, vol. 6, no. 1, p. 1126 - 1137, VLDB Endowment.

[5] A. M. Kibriya, et al. (2004), "Multinomial Naive Bayes for Text Categorization Revisited", in Proceedings of the 17th Australian Joint Conference on Advances in Artificial Intelligence, AI'04, Springer-Verlag, Berlin, Heidelberg.

[6] A. Vispute, et al. (2014), "Collective Behavior of social Networking Sites", India IOSR IOSR Journal of Computer Engineering (IOSR-JCE), vol. 16, no. 2, p. 75-79.

[7] Abbasi, Mohammad-Ali and Chai, Sun-Ki and Liu, Huan and Sagoo, Kiran, (2012), "Real-World Behavior Analysis through a Social Media Lens", 18-26. 10.1007/978-3-642-29047-3_3.

[8] Abdallah, Wasan & almougy, Samir & Sarhan, Shahenda. (2018). Social networks user modeling. AL-Qadisiyah Journal of pure Science, Vol.23 No.1 Year 2018

[9] Abdel-Hafez, Ahmad & Xu, Yue (2013) A survey of user modelling in social media websites. Computer and Information Science, 6(4), pp. 59-71

[10] Adedoyin-Olowe, Mariam & Gaber, Mohamed & Stahl, Frederic. (2013). A Survey of Data Mining Techniques for Social Media Analysis. Journal of Data Mining and Digital Humanities.

[11] Aggarwal C.C. Zhai C. (2012) A Survey of Text Clustering Algorithms. In: Aggarwal C. Zhai C. (eds) Mining Text Data. Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-3223-4_4

[12] Akram, Waseem. (2018). A Study on Positive and Negative Effects of Social Media on Society. International Journal of Computer Sciences and Engineering. 5. 10.26438/ijcse/v5i10.351354.

[13] Alex Smola and S.V.N. Vishwanathan, (2008), "Introduction to Machine Learning", Cambridge University Press The Edinburgh Building, Cambridge CB2 2RU, UK

[14] Al-Kouz, Akram. (2013). Interests Discovery in Social Networks Based on a Semantically Enriched Bayesian Network Model. 10.14279/depositonce-3720.

[15] Allahyari, Mehdi, et al. (2017), "A brief survey of text mining: Classification, clustering and extraction techniques." arXiv preprint arXiv:1707.02919.

[16] Araujo, C. Mourão, F. & Wagner Meira, J. (2014). Someato : characterizing and exploiting behavior and interests of users in social media , Corpus ID: 18512928

[17] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2012. Effects of user similarity in social media. In Proceedings of the fifth ACM international conference on Web search and data mining (WSDM'12). Association for Computing Machinery, New York, NY, USA, 703–712. DOI:https://doi.org/10.1145/2124295.2124378

[18] B. Jiang and Ying Sha, (2015), "Modeling Temporal Dynamics of User Interests in Online Social Networks", Proceedings Computer Science Elsevier, vol. 51, no. 1, p. 503-512.

[19] B. Jie, et al. (2016), "Geo-social Media Data Analytic for User Modeling and Location-based Services", Journal SIGSPATIAL Special, vol. 7, no. 3, p. 11-18, 2016.

[20] B. Ohana and B.Tierney, (2009), "Sentiment classification of reviews using SentiWordNet", in School of Computing, 9th. IT and T Conference, Dublin Institute of Technology, Dublin.

[21] B. Parantapa, et al. (2014), "Inferring User Interests in the Twitter Social Network", Proceedings of the 8th ACM Conference on Recommender Systems, RecSys'14, 2014, Silicon Valley, California, USA, p 357-360, ACM, New York.

[22] Bagadia, Sameep & Jindal, Pranav & Mundra, Rohit. (2014). Influence of Topical Interests on Users' Social Networks. 10.13140/2.1.2838.3526.

[23] Basit Shahzad, Ikramullah Lali, M. Saqib Nawaz, Waqar Aslam, Raza Mustafa, Atif Mashkoor, (2017), "Discovery and classification of user interests on social media", Information Discovery and Delivery, https://doi.org/10.1108/IDD-03-2017-0023

[24] Berndt, Jan Ole and Rodermund, Stephanie and Lorig, Fabian and Timm, Ingo. (2017), "Modeling User Behavior in Social Media with Complex Agents", Conference: Third International Conference on Human and Social Analytics (HUSO 2017), Nice, France

[25] Bhattacharya, P. Zafar, M. B. Ganguly, N. Goush, S. and Gummadi. K. P (2013), "Inferring user interests in the Twitter social network", In Proceedings of the 8th ACM Conference on Recommender Systems, p. 357

[26] Bhattacharyya, P. Garg, A. & Wu, S.F. Analysis of user keyword similarity in online social networks. Soc. Netw. Anal. Min. 1, 143–158 (2011). https://doi.org/10.1007/s13278-010-0006-4

[27] Bollegala, D. (2009). A Study on Attributional and Relational Similarity between Word Pairs on the Web.

[28] Budak, C. Kannan, A. Agrawal, R. and Pedersen, J. (2014), "Inferring user interests from microblogs", In Technical Report, MSR-TR-2014-68.

[29] C. Hung-Hsuan, (2013), "Identifying Similar Objects in Social Networks and Digital Libraries.", Pennsylvania State University, Pennsylvania, USA.

[30] C. John Samuel, IIS. Shamili, (2017), “A study on Impact of Social Media on Education, Business and Society”, International Journal of Research in Management and Business Studies (IJRMBS 2017), Vol. 4 Issue 3 (SPL 2) Jul. - Sept. 2017 ISSN : 2348-6503 (Online) ISSN : 2348-893X (Print)

[31] Can Wang, Tao Bo, Yun Wei Zhao, et al. (2018), “Behavior-Interior-Aware User Preference Analysis Based on Social Networks,” Complexity, vol. 2018, Article ID 7371209, 18 pages. https://doi.org/10.1155/2018/7371209.

[32] Carine Mukamakuza, Dimitris Sacharidis, and Hannes Werthner, (2018), "Mining User Behavior in Social Recommender Systems". In Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics (WIMS '18), Association for Computing Machinery, New York, NY, USA, Article 37, 1–6. DOI:https://doi.org/10.1145/3227609.32276

[33] Cavnar, W. & Trenkle, J. (1994). N-gram-based text categorization. Ann Arbor MI, 48113, 161--175.

[34] Cazzanti, Luca & Gupta, M.R.. (2006). Information-theoretic and Set-theoretic Similarity. 1836 - 1840. 10.1109/ISIT.2006.261752.

[35] Cheok, Adrian & Edwards, Bosede & Muniru, Idris. (2017). Human Behavior and Social Networks. 10.1007/978-1-4614-7163-9_235-1.

[36] Chinthakayala, K.C. Zhao, C. Kong, J. et al. A comparative study of three social networking websites. World Wide Web 17, 1233–1259 (2014). https://doi.org/10.1007/s11280-013-0222-8

[37] Christo Wilson, Bryce Boe, Alessandra Sala, Krishna P.N. Puttaswamy, and Ben Y. Zhao. 2009. User interactions in social networks and their implications. In Proceedings of the 4th ACM European conference on Computer systems (EuroSys '09). Association for Computing Machinery, New York, NY, USA, 205–218. DOI:https://doi.org/10.1145/1519065.1519089

[38] Collin, P. Rahilly, K. Richardson, (2011), “The Benefits of Social Networking Services: A literature review", Cooperative Research Centre for Young People, Technology and Wellbeing. Melbourne.

[39] Courtney Corley and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment (EMSEE '05). Association for Computational Linguistics, USA, 13–18.

[40] D. Lee, (2016), "Likeology: Modeling, Predicting, and Aggregating Likes in Social Media", in Proceedings of the 8th ACM Conference on Web Science, WebSci '16, Hannover, Germany.

D. Buscaldi, P. Rosso, J. M. Gomez-Soriano, and E. Sanchis, (2010) “Answering questions with an n-gram based passage retrieval engine," *Journal of Intelligent Information Systems,* vol. 34, no. 2, pp. 113{134, 2010

[41] D. M. Boyd, et al. (2007), "Social Network Sites: Definition, History, and Scholarship", Journal of Computer-Mediated Communication, vol. 13, no. 1, p. 210-230.

[42] D. Manning, et al. (2008), “Introduction to Information Retrieval”, New York, USA: Cambridge University Press, ISBN: 0521865719, 9780521865715.

[43] D.Yin, et al. (2011), "Exploiting session-like behaviors in tag prediction", In Proceedings of the 20th International Conference Companion on World wide web (WWW '11), Hyderabad, India.

[44] D‘Silva, B. Bhuptani, R. & Menon, S. (2011). Influence of Social Media Marketing on Brand Choice Behaviour among Youth in India: An Empirical Study. International Conference on Technology and Business Management, 756-763

[45] David Crandall, Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, and Siddharth Suri. 2008. Feedback effects between similarity and social influence in online communities. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '08). Association for Computing Machinery, New York, NY, USA, 160–168. DOI:https://doi.org/10.1145/1401890.1401914

[46] De Meo, Pasquale and Ferrara, Emilio and Fiumara, Giacomo (2011) Finding Similar Users in Facebook. [Book Chapter]

[47] Dhelim, Sahraoui and Aung, Nyothiri and Ning, Huansheng. (2020), "Mining user interest based on personality-aware hybrid filtering in social networks", Knowledge-Based Systems. 106227. 10.1016/j.knosys.2020.106227.

[48] Editor, Ijcsis. (2016), “A Method for Mining Social Media to Discovering Influential Users”, 10.6084/M9.FIGSHARE.3362416.V1.

[49] Elisabeta, Ioanas. (2014). Social Media and its Impact on Consumers Behavior. International Journal of Economic Practices and Theories,. 4.

[50] F. Chao, J. Xu and C. Lin, (2016), "Mining user interests from social media by fusing textual and visual features", 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, 2016, p. 1-8, doi: 10.1109/APSIPA.2016.7820713.

[51] Fabrício Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgílio Almeida, (2009), "Characterizing user behavior in online social networks". In Proceedings of the 9th ACM SIGCOMM (IMC '09), Association for Computing Machinery, New York, NY, USA, 49–62. DOI:https://doi.org/10.1145/1644893.1644900

[52] Falomir, Zoe & Gonzalez-Abril, Luis & Museros, Lledó & Ortega, Juan. (2012). Measures of Similarity Between Objects Based on Qualitative Shape Descriptions. Spatial Cognition and Computation - SPAT COGN COMPUT. 13. 10.1080/13875868.2012.700463.

[53] Faris Kateb and Jugal Kalita, (2015), “Article: Classifying Short Text in Social Media: Twitter as Case Study", International Journal of Computer Aplications 111(9):1-12, February 2015.

[54] Fattane Zarrinkalam, Hossein Fani, and Ebrahim Bagheri. 2019. Social User Interest Mining: Methods and Applications. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19). Association for Computing Machinery, New York, NY, USA, 3235–3236. DOI:https://doi.org/10.1145/3292500.3332279

[55] Fishbein, M. & Ajzen, Icek. (1975). Belief, attitude, intention and behaviour: An introduction to theory and research.

[56] Fouss F. and M. Saerens, (2008), "Evaluating performance of recommender systems: an experimental comparison.", Web Intelligence, IEEE, p. 735–738.

[57] G. Sharad and G. Daniel, (2014), "Predicting Individual Behavior with Social Networks", Marketing Science, INFORMS, Linthicum, Maryland, USA, vol. 33, no. 1, p. 8293.

[58] Gaind, Bharat & Syal, Varun & Padgalwar, Sneha. (2019). Emotion Detection and Analysis on Social Media.Global Journal of Engineering Science and Researches (ICRTCET-18) (2019) 78-89

[59] Godoy, Daniela and Amandi, Analía. (2006), "Modeling user interests by conceptual clustering", Information Systems. 247-265. 10.1016/j.is.2005.02.008.

[60] Guy Ido, et al. (2013), "Mining Expertise and Interests from Social Media", in Proceedings of the 22nd International Conference on World Wide Web, WWW '13, Rio de Janeiro, Brazil.

[61] Hall, M. Frank, et al. (2009), "The WEKA data mining software: An update", ACM SIGKKD Exploration Newsletter, 11(1), 10-18.

[62] Himelboim, I. Smith, M. A. Rainie, L. Shneiderman, B. & Espina, C. (2017). Classifying Twitter Topic-Networks Using Social Network Analysis. Social Media + Society. https://doi.org/10.1177/2056305117691545

[63] Hossen. Muhammad, Faiad. Md, Chowdhury. Md and Islam. Md. (2018), "Discovering Users Topic of Interest from Tweet", International Journal of Computer Science and Information Technology. 10. 10.5121/ijcsit.2018.10108.

[64] Hsinchun Chen, et al. (2010), "Data Mining for Social Network Data, Annalys of Information Systems", volume 12, Springer.

[65] Huan Liu and Zafarani Reza, (2013), "Connecting Users Across Social Media Sites: A Behavioral-modeling Aproach", in Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13, Chicago, Illinois, USA.

[66] Ido Guy, Michal Jacovi, Adam Perer, Inbal Ronen, and Erel Uziel. 2010. Same places, same things, same people? mining user similarity on social media. In Proceedings of the 2010 ACM conference on Computer supported cooperative work (CSCW '10). Association for Computing Machinery, New York, NY, USA, 41–50. DOI:https://doi.org/10.1145/1718918.1718928

[67] J. Tang, Y. Wang, K. Zheng, and Q. Mei, (2017), "End-to-end learning for short text expansion", in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2017, p. 1105–1113, can, August 2017.

[68] Jain, Arti and Gupta, Ashutosh and Sharma, Nikhil and Joshi, Shubham and Yadav, Divakar, (2018), "Mining Aplication on Analyzing Users' Interests from Twitter ", Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT), 2018 held at Malaviya National Institute of Technology,

Jaipur (India) on March 26-27, 2018, Available at SSRN: https://ssrn.com/abstract=3166015

[69] Johnston, Patrick & Kristoff, Nicholas & McGinness, Heather & Vu, Phuong & Wong, Nathaniel & Wright, Jason & Scherer, Bill & Burkett, Matt. (2006). Strategic Online Advertising: Modeling Internet User Behavior with Advertising.com. 10.1109/SIEDS.2006.278732.

[70] Jung, R. & Adolf, M. (2018). Extracting Realistic User Behavior Models. Software Engineering.

[71] K. Deng, L. Xing, L. Zheng, H. Wu, P. Xie and F. Gao, "A User Identification Algorithm Based on User Behavior Analysis in Social Networks," in IEEE Access, vol. 7, pp. 47114-47123, 2019, doi: 10.1109/ACCESS.2019.2909089.

[72] Kapanipathi P. Jain P. Venkataramani C. Sheth A. (2014) User Interests Identification on Twitter Using a Hierarchical Knowledge Base. In: Presutti V. d'Amato C. Gandon F. d'Aquin M. Staab S. Tordai A. (eds) The Semantic Web: Trends and Challenges. ESWC 2014. Lecture Notes in Computer Science, vol 8465. Springer, Cham. https://doi.org/10.1007/978-3-319-07443-6_8

[73] Kapoor, K.K. Tamilmani, K. Rana, N.P. et al. (2018), "Advances in Social Media Research: Past, Present and Future", Inf Syst Front 20, 531–558 (2018), https://doi.org/10.1007/s10796-017-9810-y

[74] Kateb, Faris & Kalita, Jugal. (2015). Classifying Short Text in Social Media: Twitter as Case Study. International Journal of Computer Applications. 111. 1-12. 10.5120/19563-1321.

[75] Kim Soo-Min and Hovy Eduard, (2004), "Determining the Sentiment of Opinions", in Proceedings of the 20th International Conference on Computational Linguistics COLING '04, Geneva, Switzerland, Association for Computational Linguistics, Stroudsburg, PA, USA.

[76] Kim, Cheonsoo & Yang, Sung-Un. (2017). Like, comment, and share on Facebook: How each behavior differs from the other. Public Relations Review. 10.1016/j.pubrev.2017.02.006.

[77] Kim, J. Ko, B. Jeong, H. and Kim. P. (2013), "Extracting user interests on Facebook", International Journal of Software Engineering and Its Aplications, 7(2).

[78] Kinsella S. Passant A. Breslin J.G. (2011) Topic Classification in Social Media Using Metadata from Hyperlinked Objects. In: Clough P. et al. (eds) Advances in Information Retrieval. ECIR 2011. Lecture Notes in Computer Science, vol 6611. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-20161-5_20

[79] Konsolakis, K. Hermens, H. Villalonga, C. Vollenbroek-Hutten, M. and Banos, O. (2018), "Human Behaviour Analysis through Smartphones", In Multidisciplinary Digital Publishing Institute Proceedings (Vol. 2, No. 19, p. 1243)

[80] Kowsari, K. Meimandi, K. J. Heidarysafa, M. Mendu, S. Barnes, L. E. and Brown, D. E. (2019), "Text Classification Algorithms: A Survey", ACM Journal.

[81] Kumari, Archana and Verma, Jyotsna. (2015), “Impact of social networking sites on social interaction – a study of college students”, International Journal of Humanities and Social Sciences (IJHSS), 4. 55-61.

[82] Kwan Hui Lim and Amitava Datta. 2013. Interest classification of Twitter users using Wikipedia. In Proceedings of the 9th International Symposium on Open Collaboration (WikiSym '13). Association for Computing Machinery, New York, NY, USA, Article 22, 1–2. DOI:https://doi.org/10.1145/2491055.2491078

[83] L. Jin, Y. Chen, T. Wang, P. Hui and A. V. Vasilakos, "Understanding user behavior in online social networks: a survey," in IEEE Communications Magazine, vol. 51, no. 9, pp. 144-150, September 2013, doi: 10.1109/MCOM.2013.6588663.

[84] Lei Li, (2017), "Behavior Analysis in Social Networks", In: Alhajj R. Rokne J. (eds), "Encyclopedia of Social Network Analysis and Mining". Springer, New York, NY. https://doi.org/10.1007/978-1-4614-7163-9_110198-1

[85] LI, Wei & Darban, Ayda. (2012). The impact of online social networks on consumers' purchasing decision : The study of food retailers.

[86] Lim, B.H. Lu, D. Chen, T. & Kan, M. (2015). #mytweet via Instagram: Exploring user behaviour across multiple social networks. 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 113-120.

[87] Lim, K. H. and Datta, A. (2013), “Interest classification of Twitter users using Wikipedia”, In Proceedings of the 9th International Symposium on Open Collaboration, p. 22.

[88] Lin Dekang, (1998), "An information-theoretic definition of similarity", in Proc. 15th International Conf. on Machine Learning, San Francisco, CA.

[89] Lin Gong and Hongning Wang. 2018. When Sentiment Analysis Meets Social Network: A Holistic User Behavior Modeling in Opinionated Data. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18). Association for Computing Machinery, New York, NY, USA, 1455–1464. DOI:https://doi.org/10.1145/3219819.3220120

[90] Lirong Qiu and Jia Yu, (2018), “CLDA: An Effective Topic Model for Mining User Interest Preference under Big Data Background”, Complexity, vol. 2018, Article ID 2503816, 10 pages, 2018. https://doi.org/10.1155/2018/2503816.

[91] Liu Huan and Reza Zafarani, (2014), "Behavior Analysis in Social Media", Arizona State University, Arizona, USA.

[92] Liu, Zhenyan & Qiu, Yueshi & Zhang, Zhe & Mao, Limin & Zheng, Xiaohan. (2019). Research on Hot Topics Discovery Based on Short Texts of Online Reviews. Journal of Physics: Conference Series. 1288. 012046. 10.1088/1742-6596/1288/1/012046.

[93] Looijenga, M.S. (2018) The Detection of Fake Messages using Machine Learning.EEMCS: Electrical Engineering, Mathematics and Computer Science

[94] M. A. Mouriño García, R. P. Rodríguez, M. V. Ferro and L. A. Rifón, "Wikipedia-Based Hybrid Document Representation for Textual News Classification," 2016 3rd International Conference on Soft Computing & Machine Intelligence (ISCMI), Dubai, 2016, pp. 148-153, doi: 10.1109/ISCMI.2016.31.

[95] M. H. Nguyen, (2018), "On the Distinction of Subjectivity and Objectivity of Emotions in Texts", International Journal of Advanced Computer Science and Aplications (IJACSA), 9(9), p.584-589, 2018

[96] M. Michelson and S. A. Macskassy, (2010), "Discovering customers' topics of interest on Twitter: a first look", In ACM Workshop on Analytics for Noisy Unstructured Text Data.

[97] M. Shafik, R. Elgohary, I. Moawad and M. Roushdy, (2019), "Hybrid Method for Modeling User Interests based on Social Network", 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS), Cairo, Egypt, 2019, p. 68-73, doi: 10.1109/ICICIS46948.2019.9014724.

[98] Maia, M. Almeida, J. and Almeida, V. (2008), "Identifying user behavior in online social networks", In Proceedings of the 1st workshop on Social network systems (p. 1-6), ACM.

[99] Manel Mezghani, et al. (2014), "Analyzing Tagged Resources for Social Interests Detection", in 16th International Conference on Enterprise Information Systems (ICEIS 2014), Lisbonne, Portugal.

[100] Meenakshi, & Geetika (2014). Survey on Classification Methods using WEKA. International Journal of Computer Applications, 86, 16-19.

[101] Metzler D. Dumais S. Meek C. (2007) Similarity Measures for Short Segments of Text. In: Amati G. Carpineto C. Romano G. (eds) Advances in Information Retrieval. ECIR 2007. Lecture Notes in Computer Science, vol 4425. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-71496-5_5

[102] Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. (2013), "Efficient estimation of word representations in vector space", arXiv 2013, arXiv:1301-3781.

[103] Min, Jinming and Jones, Gareth J.F. (2011) Building user interest profiles from wikipedia clusters. In: The Workshop on Enriching Information Retrieval (ENIR 2011) at SIGIR 2011, 28 July 2011, Beijing, China.

[104] Mojtahedi, Rahebeh and Rahmani, Amir and Alizadeh, Sasan, (2019), "User behavior mining on social media: a systematic literature review", Multimedia Tools and Aplications. 78. 10.1007/s11042-019-08046-6.

[105] Motoyama, Marti & Varghese, George. (2009). I seek you: searching and matching individuals in social networks. 67-75. 10.1145/1651587.1651604.

[106] Nguyen, Manh Hung. (2020), A Label-Oriented Aproach For Text Classification. International journal of innovative computing, information and control: IJICIC. 16. 1593-1609. 10.24507/ijicic.16.05.1593.

[107] Nori, Nozomi, Bollegala, Danushka and Ishizuka, Mitsuru, (2011), "Exploiting User Interest on Social Media for Aggregating Diverse Data and Predicting

Interest", Proceedings of the fifth International Conference on Weblogs and Social Media, ICWSM. 11. 2011

[108] Noulas, A. Musolesi, M. Pontil, M. & Mascolo, C. (2009). Inferring Interests from Mobility and Social Interactions.

[109] Opuszko and J. Ruhland, "Classification Analysis in Complex Online Social Networks Using Semantic Web Technologies," 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Istanbul, 2012, pp. 1032-1039, doi: 10.1109/ASONAM.2012.179.

[110] P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, and H. Hao, (2016), "Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification," Neurocomputing, vol. 174, p. 806–814, 2016.

[111] Palsetia, D. Patwary, M. Zhang, K. Lee, K. Moran, C. Xie, Y. Honbo, D. Agrawal, A. Liao, W. & Choudhary, A. (2012). User-Interest based Community Extraction in Social Networks. KDD 2012.

[112] Panagiotou, N. Katakis, I. & Gunopulos, D. (2016). Detecting Events in Online Social Networks: Definitions, Trends and Challenges. Solving Large Scale Learning Tasks.

[113] Parag Singla and Matthew Richardson. 2008. Yes, there is a correlation: - from social networks to personal behavior on the web. In Proceedings of the 17th international conference on World Wide Web (WWW '08). Association for Computing Machinery, New York, NY, USA, 655–664. DOI:https://doi.org/10.1145/1367497.1367586

[114] Pengtao Xie, Yulong Pei, Yuan Xie, and Eric Xing, (2015), "Mining user interests from personal photos", In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15), AAAI Press, 1896–1902.

[115] Pennacchiotti, M. & Popescu, A. (2011). A Machine Learning Approach to Twitter User Classification. ICWSM.

[116] R.Chinnaiyan and V.Ilango, Analyzing the User Behaviours by Mining Web Access Log Files, (2015), International Journal of advanced studies in Computer Science and Engineering IJASCSE Volume 4, Issue 11, 2015

[117] Ristoski, P. Petrovski, P. Mika, P. and Paulheim, H. (2018), "A machine learning aproach for product matching and categorization", Semantic web, (Preprint), p.1-22.

[118] Ryen W. White, Peter Bailey, and Liwei Chen (2009), "Predicting user interests from contextual information", In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR '09), Association for Computing Machinery, New York, NY, USA, 363–370. DOI:https://doi.org/10.1145/1571941.1572005

[119] Rytsarev, Igor & Blagov, Aleksandr. (2016). Classification of text data from the social network Twitter. 851-856. 10.18287/1613-0073-2016-1638-851-856.

[120] S. Sheetal A Takale, (2010), "Measuring semantic similarity between words using web documents", International Journal of Advanced Computer Science and Aplications (IJACSA), Volume 1, Issue 4. 2010.

[121] S. Yang, G. Huang and B. Cai, "Discovering Topic Representative Terms for Short Text Clustering," in IEEE Access, vol. 7, pp. 92037-92047, 2019, doi: 10.1109/ACCESS.2019.2927345.

[122] Sengkey, C.H. & Chang, H. (2016). A Topic of Interest-based Approach on Online Social Network Services for Information Sharing. ICNCC '16.

[123] Servizi, Valentino & Pereira, Francisco & Anderson, Marie & Nielsen, Otto. (2019). Mining User Behaviour from Smartphone data, a literature review.

[124] Shangsong Liang, et al. (2017), "Inferring Dynamic User Interests in Streams of Short Texts for User Clustering", ACM Transactions on Information Systems, vol. 36, no. 1, p. 1-37.

[125] Sheng Bin, et al. (2016), "Tag-Based Interest-Matching Users Discovery Aproach in Online Social Network", International Journal of Hybrid Information Technology, vol. 9, no. 5, p. 61-70.

[126] Shuang-Hong Yang, Bo Long, Alex Smola, Narayanan Sadagopan, Zhaohui Zheng, and Hongyuan Zha. (2011), “Like like alike: joint friendship and interest propagation in social networks”, In Proceedings of the 20th International Conference on World wide web (WWW ’11), Association for Computing Machinery, New York, USA.

[127] Si, H. Zhou, J et al. (2019), "Association Rules Mining Among Interests and Aplications for Users on Social Networks", IEEE Access, 7, p.116014-116026, 2019.

[128] Siddiqui, Tamanna & Aalam, Parvej. (2015). Short Text Clustering; Challenges & Solutions: A Literature Review.

[129] Singh, A. Halgamuge, M. N. and Moses, B. (2019), "An Analysis of Demographic and Behavior Trends Using Social Media: Facebook, Twitter, and Instagram", Social Network Analytics, 87–108. https://doi.org/10.1016/B978-0-12-815458-8.00005-0

[130] Somnath Banerjee, Krishnan Ramanathan, and Ajay Gupta. (2007), “Clustering short texts using wikipedia”, In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR ’07), Association for Computing Machinery, New York, USA, 787–788. DOI:https://doi.org/10.1145/1277741.1277909

[131] Sonam, Surjeet Kumar, (2019), Mining Social Networking Sites: A Specific area of Facebook, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-4, November 2019 1943

[132] Tang Jiliang, et al. (2013), "Mining Social Media with Social Theories: A Survey", SIGKDD Explor. Newsl. vol. 15, no. 2, p. 20-29, 2013.

[133] Tang, G. Xia, Y. Wang, W. Lau, R. & Zheng, T. F. (2014). Clustering tweets using Wikipedia concepts. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14) (pp. 2262-2267). European Language Resources Association (ELRA).

[134] Tsapatsoulis N and Djouvas C (2019) Opinion Mining From Social Media Short Texts: Does Collective Intelligence Beat Deep Learning? Front. Robot. AI 5:138. doi: 10.3389/frobt.2018.00138

[135] Tuna, T. Akbas, E. Aksoy, A. et al. User characterization for online social networks. Soc. Netw. Anal. Min. 6, 104 (2016). https://doi.org/10.1007/s13278-016-0412-3

[136] Twitter API. (2016), Retrieved from website: https://dev.Twitter.com/rest/public/search, November 12

[137] Udayanan, H.K. (2019). Factors Influencing Online Shopping Intention: A study among online shoppers in Oman. The International Journal of Academic Research in Business and Social Sciences, 9.

[138] V. Ranjith,V. Rajaram,(2018), Mining user's interest by mining interpersonal interest similarity and polarity identification, 2018 IJRAR September 2018, Volume 5, Issue 3, www.ijrar.org (E-ISSN 2348-1269, P- ISSN 2349-5138)

[139] W. M. Campbell, et al. (2003), "Social Network Analysis with Content and Graphs", Lincoln Laboratory Journal, vol. 20, no. 1, p. 63, 2013.

[140] W. Meng, L. Lin, W. Jing, P. Yu, J. Liu, and X. Fei, (2013), "Improving Short Text Classification Using Public Search Engines", Uncertainty in Knowledge Modelling and Decision Making, Springer, Heidelberg, Germany, 2013.

[141] Waheed H, Anjum M, Rehman M, Khawaja A, (2017), "Investigation of user behavior on social networking sites", PLoS ONE 12(2): e0169693. https://doi.org/10.1371/journal.pone.0169693

[142] Wang H. (2013) Understanding Short Texts. In: Ishikawa Y. Li J. Wang W. Zhang R. Zhang W. (eds) Web Technologies and Applications. APWeb 2013. Lecture Notes in Computer Science, vol 7808. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-37401-2_1

[143] Wang T. Liu H. He J. Du X. (2013) Mining User Interests from Information Sharing Behaviors in Social Media. In: Pei J. Tseng V.S. Cao L. Motoda H. Xu G. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science, vol 7819. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_8

[144] Xiaohua Hu, Xiaodan Zhang, Caimei Lu, E. K. Park, and Xiaohua Zhou. 2009. Exploiting Wikipedia as external knowledge for document clustering. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09). Association for Computing Machinery, New York, NY, USA, 389–396. DOI:https://doi.org/10.1145/1557019.1557066

[145] Xin Li, et al. (2008), "Tag-based Social Interest Discovery", in WWW 2008 April 21–25, 2008, International World Wide Web Conference Committee, Beijing, China.

[146] Xinyi Zhou and Reza Zafarani. 2020. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. ACM Comput. Surv. 53, 5, Article 109 (October 2020), 40 pages. DOI:https://doi.org/10.1145/3395046

[147] Xu, Xuhai and Hassan, Ahmed and Dumais, Susan and Omar, Farheen and Pop, Bogdan and Rounthwaite, Robert and Jahanbakhsh, Farnaz. (2020), "Understanding User Behavior For Document Recommendation". 3012-3018. 10.1145/3366423.3380071.

[148] Xu, Z. Lu, R. Xiang, L. and Yang, Q. (2011), "Discovering user interest on Twitter with a modified Author-Topic model", In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, p. 422-429.

[149] Y. Kim. (2014), "Convolutional Neural Networks for Sentence Classification", in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP-2014), Doha, Qatar, p.1746-1751, 2014

[150] Y. Teng, Y. Shi, J. Tsai, H. Shuai, C. Tai and D. Yang,(2019), "Optimizing Social-Topic Engagement on Social Network and Knowledge Graph", 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, USA, 2019, p. 1-6, doi: 10.1109/GLOBECOM38437.2019.9013546.

[151] Yin Dawei, et al. (2013), "Connecting Comments and Tags: Improved Modeling of Social Tagging Systems", in Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM '13, Rome, Italy.

[152] Yoad Lewenberg, et al, (2015), "Using Emotions to Predict User Interest Areas in Online Social Networks", 978-1-4673-8273-1/15,IEEE, Paris, France.

[153] Younus, F.R. (2015). Identifying the Factors Affecting Customer Purchase Intention. Global Journal of Management and Business Research, 15.

[154] Yousukkee, Sawita. (2016), "Survey of analysis of user behavior in online social network". MIT-128. 10.1109/MITICON.2016.8025232.

[155] Z. Yongzheng and P. Marco, (2013), "Predicting Purchase Behaviors from Social Media", in Proceedings of the 22nd International Conference on World Wide Web, WWW '13, Rio de Janeiro, Brazil.

[156] Zafarani Reza, et al. (2014), "Social Media Mining: An Introduction", 1107018854, 9781107018853, New York, USA: Cambridge University Press.

[157] Zarrinkalam, F. Fani, H. Bagheri, E. Kahani, M. and Du, W. (2015), "Semantics-Enabled User Interest Detection from Twitter", In Proceedings of International Conference of Web Intelligence and Intelligent Agent Technology, p. 469-476.

[158] Zarrinkalam, F. Kahani, M. and Bagheri, E. (2019), "User interest prediction over future unobserved topics on social networks", Inf Retrieval J 22, 93–128 (2019), https://doi.org/10.1007/s10791-018-9337-y

[159] Zarrinkalam, Fattane and Fani, Hossein and Bagheri, Ebrahim. (2019), "Extracting, Mining and Predicting Users' Interests from Social Networks", 1407-1408. 10.1145/3331184.3331383.

[160] Zarrinkalam, Fattane and Kahani, Mohsen and Bagheri, Ebrahim. (2018), "Mining user interests over active topics on social networks", Information Processing and Management. 54. 339-357. 10.1016/j.ipm.2017.12.003.

[161] Zeinab Abbassi, Aditya Bhaskara, and Vishal Misra. 2015. Optimizing Display Advertising in Online Social Networks. In Proceedings of the 24th International Conference on World Wide Web (WWW '15). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1–11. DOI:https://doi.org/10.1145/2736277.2741648

[162] Zhang, W. Yoshida, T. and Tang, X. (2011), "A comparative study of TF∗IDF, LSI and multi-words for text classification", Expert Systems with Aplications, 38(3), 2758–2765.

[163] Zhang, Xiuzhen & Wang, Shuliang & Cong, Gao & Cuzzocrea, Alfredo. (2019). Social Big Data: Mining, Applications, and Beyond. Complexity. 2019. 1-2. 10.1155/2019/2059075.

[164] Zhang, Z. & Lan, M. (2014). Estimating Semantic Similarity between Expanded Query and Tweet Content for Microblog Retrieval. TREC.

[165] Zhao, Y. Wang, J. and Shen, Q. (2012), "Feature analysis based on Weibo user interest detection algorithm", Telecom Engineering Technics and Standardization, 25(11), 79-83.

[166] Zhe Zhao, Zhiyuan Cheng, Lichan Hong, and Ed H. Chi. 2015. Improving User Topic Interest Profiles by Behavior Factorization. In Proceedings of the 24th International Conference on World Wide Web (WWW '15). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1406–1416. DOI:https://doi.org/10.1145/2736277.2741656

[167] Zhiheng Xu, Yang Zhang, Yao Wu, and Qing Yang. 2012. Modeling user posting behavior on social media. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '12). Association for Computing Machinery, New York, NY, USA, 545–554. DOI:https://doi.org/10.1145/2348283.2348358

PHỤ LỤC

PHỤ LỤC A: MỘT SỐ THUẬT NGỮ SỬ DỤNG TRÊN MẠNG XÃ HỘI

Một số thuật ngữ được sử dụng trên mạng xã hội Facebook.com

***Status**: Việc đăng hay viết lên trang cá nhân trên Facebook được gọi là status. **Like:** Thể hiện cảm xúc của mình với một bài viết hay bình luận nào đó. **Share**: Chia sẻ lại status của chính mình hoặc người dùng khác hoặc từ mạng xã hội khác. **Follow:** Theo dõi một người dùng trên mạng, là bạn bè hoặc không là bạn bè. **Hashtag:** hashtag là một dạng như từ khóa liên quan đến sự kiện nào đó. Hashtag thường được nhóm các status lại với nhau ví dụ #Monan. Hashtag thường được sử dụng trong các mạng lưới thảo luận, trò chuyện ngang hàng, tức là trực tiếp giữa người dùng này với người dùng khác.*

Một số thuật ngữ sử dụng trên mạng Twitter.com

***Tweet**: Việc đăng hay viết lên trang cá nhân trên Twitter được gọi là **Tweet**. Tweet này cũng tương đương với status trên Facebook. **Retweet**: Retweet thì giống như share (chia sẻ) của Facebook. **Likes:** Thích tweet. **Follow:** Follow là theo dõi, do Twitter không có dịch vụ kết bạn giống như Facebook nên Follow là chức năng theo dõi. Nếu hai người cùng theo dõi lẫn nhau thì cũng giống kết bạn trên Facebook*

***Hashtag**: hashtag là một dạng như từ khóa liên quan đến sự kiện nào đó. Hashtag thường được sử dụng trong các mạng lưới thảo luận, trò chuyện ngang hàng, tức là trực tiếp giữa người dùng này với người dùng khác. Hashtag được sử dụng nhiều nhất trên Twitter sau đó Facebook mới cập nhật chức năng này.*

Một số thuật ngữ sử dụng trên mạng YouTube.com

***Channel:** Kênh YouTube của người dùng. **Video**: Việc đăng một video lên tài khoản cá nhân trên YouTube, tương đương với status bên Facebook hay Tweet trên Twitter*

***Views:** Số lượt xem video thể hiện video có được quan tâm hay không. **Share**: Chia sẻ lại video của người dùng khác trên kênh của mình. **I like this:** Thích video - **I dislike this:** Không thích video*

***Subcribe:** Đăng ký nhận tin, do YouTube không có dịch vụ kết bạn giống như Facebook nên Subscribe là chức năng đăng ký nhận tin từ tài khoản người dùng, tương đương với theo dõi.*

***Hashtag**: hashtag là một dạng như từ khóa liên quan đến sự kiện nào đó. Hashtag thường được sử dụng trong các mạng lưới thảo luận, trò chuyện ngang hàng, tức là trực tiếp giữa người dùng này với người dùng khác, kênh này với kênh khác*

PHỤ LỤC B: THỰC NGHIỆM LỰA CHỌN THUẬT TOÁN TÍNH GIÁ TRỊ CHO THỂ LOẠI, QUAN ĐIỂM VÀ CẢM XÚC

Bài toán tính giá trị cho các đặc trưng thể loại, quan điểm, cảm xúc của bài viết thực chất là bài toán gán nhãn cho dữ liệu văn bản ngắn trên mạng xã hội, vì vậy, có thể sử dụng bất kỳ một thuật toán gãn nhãn nào đó đã được giới thiệu để thực hiện như các thuật toán CNN, Word2vec, MNB, NB ... Tuy nhiên, đặc trưng của dữ liệu văn bản ngắn trên mạng xã hội là không đầy đủ, không có cấu trúc thống nhất, ngữ pháp và văn phạm của văn nói không giống các dữ liệu văn bản chuẩn mực. Do vậy, luận án cần lựa chọn thuật toán phù hợp nhất với bộ dữ liệu thực được xây dựng trong các thực nghiệm. Do đó, mục đích của thực nghiệm này nhằm so sánh và lựa chọn được thuật toán phù hợp nhất cho việc gán nhãn cho dữ liệu văn bản ngắn trên mạng xã hội được xử lý và thực nghiệm trong luận án.

PL2.1. Một số thuật toán gán nhãn dữ liệu văn bản trong thực nghiệm

Bộ ngữ liệu để thử nghiệm:

Luận án tiến hành thực nghiệm trên hai bộ ngữ liệu, bộ ngữ liệu thực tự thu thập và xây dựng gồm 2000 bài viết trên mạng xã hội Facebook và bộ ngữ liệu chuẩn 20 NewsGroups và bộ ngữ liệu cảm xúc SemEval-2017 dùng để so sánh độ chính xác.

PL2.2. Kịch bản thực nghiệm và tham số đầu ra

Luận án tiến hành thực nghiệm theo kịch bản dựa theo nhãn. Với mỗi thực nghiệm (tương ứng với một bộ ngữ liệu), các bước tiến hành tương tự phương pháp One-vs-All [11] [15] như sau:

- Với mỗi bộ ngữ liệu, luận án xây dựng bộ mẫu như sau:
  - Với mỗi nhãn có trong bộ dữ liệu lấy N văn bản để thử nghiệm, trong N văn bản này có N/2 mẫu có nhãn đang xét (chọn ngẫu nhiên) và N/2 mẫu còn lại là tập các văn bản của các nhãn còn lại (chọn ngẫu nhiên).
  - Cụ thể với hai bộ ngữ liệu 20 NewsGroups và SemEval-2017 có 1000 mẫu.

Bộ ngữ liệu phân loại chủ đề lấy bằng N = 200 và bộ ngữ liệu phân loại cảm xúc cũng lấy N=200.

- Sử dụng phương pháp cross-validation: Chia N mẫu văn bản thành 10 nhóm con (10-folds cross-validation), mỗi lần chạy sẽ lấy một nhóm làm bộ kiểm tra (bộ test), 9 nhóm còn lại làm bộ dữ liệu học (bộ training).
- Sau đó thực hiện việc lặp 10 lần cho mỗi nhóm con như sau:
  - Coi mẫu có nhãn tương ứng nhãn đang xét là YES, các mẫu có nhãn khác đều gán thành nhãn NO. Bài toán trở thành phân loại văn bản theo hai nhãn YES và NO.
  - Dùng nhóm đang xét làm bộ kiểm thử (test), dùng 9 nhóm còn lại làm bộ dữ liệu huấn luyện (training)
  - Lần lượt áp dụng các thuật toán được xem xét để huấn luyện và kiểm thử trên hai tập trên.
  - Quan sát các tham số đầu ra.
  - Tính trung bình giá trị trên từng tham số đầu ra cho từng nhãn.
- So sánh kết quả đầu ra sau 10 lần lặp, tham số đầu ra là F1- score (tính từ Precision và Recall) hay F1-measure đã trình bày trong chương một luận án.

Bảng PL2.1: Danh sách các thuật toán đưa vào thực nghiệm

Thuật toán	Đặc trưng đầu vào	Gán đơn nhãn	Gán đa nhãn
CNN	Văn bản gốc	x	x
W2V	Văn bản gốc	x	x
MNB	TF-IDF	x	
NB	TF-IDF	x	
SVM	TF-IDF	x	
K-NN	TF-IDF	x	
C4.5	TF-IDF	x	

PL2.3. Kết quả thực nghiệm

Kết quả độ chính xác (Accuracy) trên bộ ngữ liệu 20 NewsGroups được trình bày trong Bảng PL2.2, đây là kết quả trung bình của 10 lần chạy cho mỗi nhãn tương ứng trong bộ mẫu thử nghiệm

Từ kết quả cho thấy thuật toán C45 đạt giá trị Accuracy cao nhất trong 19/20 nhãn, thuật toán SVM cho giá trị Accuracy cao nhất trên nhãn "*talk.politics.mideast*". Xét kết quả trung bình trên tất cả các nhãn thì thuật toán C45 cho giá trị Accuracy cao nhất, tiếp theo lần lượt là các thuật toán RF, SVM, và MNB.

Bảng PL2.2: Độ chính xác Accuracy trên bộ ngữ liệu 20 NewsGroups

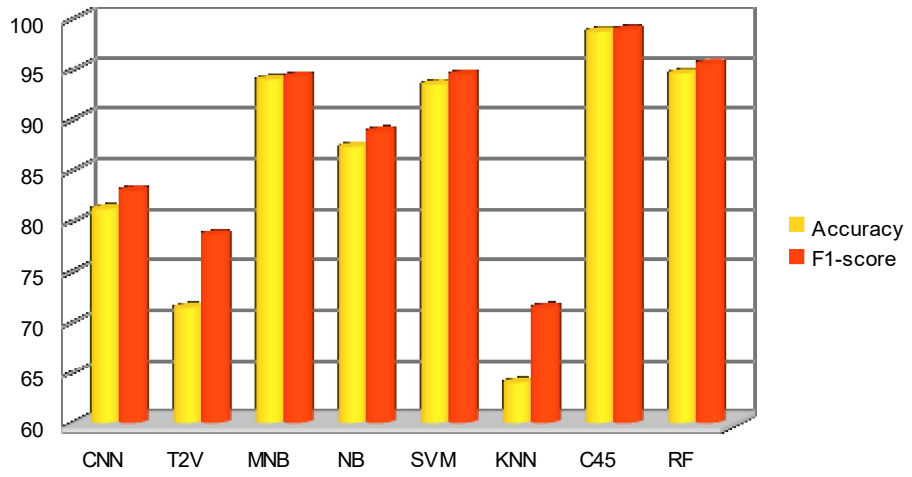
Nhãn	**CNN**	**W2V**	**MNB**	**NB**	**SVM**	**K-NN**	**C4.5**	**RF**
alt.atheism	80.89	77.14	95.45	90.34	93.75	61.48	97.39	95.80
comp.graphics	81.48	67.33	90.00	83.86	87.73	56.93	97.73	86.70
comp.os.ms-windows.misc	81.14	65.91	87.16	87.95	94.20	58.07	99.89	92.27
comp.sys.ibm.pc.hardware	78.62	71.25	87.73	86.14	89.43	65.45	99.43	94.32
comp.sys.mac.hardware	73.52	72.37	90.57	82.05	91.82	63.07	98.98	94.66
comp.windows.x	80.97	73.25	92.73	81.25	90.68	58.30	97.05	91.70
misc.forsale	83.36	76.25	91.14	79.66	92.61	61.14	99.32	92.39
rec.autos	79.28	75.91	93.86	84.77	89.77	59.66	99.66	93.64
rec.motorcycles	84.32	80.42	95.45	91.93	93.18	62.16	99.77	97.16
rec.sport.baseball	82.81	70.57	96.82	93.41	93.75	63.18	99.43	96.02
rec.sport.hockey	87.27	70.84	97.95	94.66	97.27	66.14	99.66	97.39
sci.crypt	84.66	65.11	94.43	91.14	95.11	61.59	99.32	97.50
sci.electronics	78.72	75.91	91.36	85.68	91.93	57.84	98.30	90.23
sci.med	82.27	63.64	93.30	84.89	93.30	61.82	98.41	96.25
sci.space	81.93	72.27	95.91	87.05	94.09	66.48	99.20	97.95
soc.religion.christian	85.80	62.00	98.07	96.93	99.89	72.95	99.43	98.98
talk.politics.guns	79.98	71.02	94.43	86.02	93.98	76.14	98.30	94.43
talk.politics.mideast	80.57	69.08	96.82	90.68	97.84	65.23	97.05	97.05
talk.politics.misc	75.64	72.16	87.61	83.18	94.20	69.66	96.25	93.30
talk.religion.misc	79.25	75.10	93.07	83.75	93.86	70.57	98.41	92.95
Trung bình các nhãn	81.12	71.38	93.19	87.27	93.42	63.89	**98.65**	94.53

Từ kết quả cho thấy thuật toán C45 đạt giá trị Accuracy cao nhất trong 19/20 nhãn, thuật toán SVM cho giá trị Accuracy cao nhất trên nhãn "*talk.politics.mideast*". Xét kết quả trung bình trên tất cả các nhãn thì thuật toán C45 cho giá trị Accuracy cao nhất, tiếp theo lần lượt là các thuật toán RF, SVM, và MNB.

Bảng PL2.3: Độ chính xác F1- score trên bộ ngữ liệu 20 NewsGroups

Nhãn	**CNN**	**W2V**	**MNB**	**NB**	**SVM**	**K-NN**	**C4.5**	**RF**
alt.atheism	84.22	83.03	96.09	91.60	94.51	73.15	97.70	96.32
comp.graphics	82.73	76.60	91.69	86.33	89.96	67.35	97.99	89.50
comp.os.ms-windows.misc	84.15	55.35	87.44	89.68	95.08	72.65	99.90	93.68
comp.sys.ibm.pc.hardware	79.10	79.99	90.16	87.90	91.29	72.62	99.50	95.23
comp.sys.mac.hardware	71.50	80.77	92.20	84.48	93.08	71.84	99.10	95.52
comp.windows.x	81.55	80.65	93.76	81.55	92.14	62.03	97.43	95.23
misc.forsale	83.26	83.12	92.59	81.14	93.87	72.94	99.40	93.81
rec.autos	82.49	78.30	94.78	87.02	91.49	71.63	99.69	94.70
rec.motorcycles	86.26	84.77	96.12	92.89	94.20	70.20	99.80	97.58
rec.sport.baseball	82.76	79.76	97.28	94.27	94.74	74.78	99.50	96.64
rec.sport.hockey	88.68	79.66	98.24	95.32	97.66	70.14	99.70	97.76
sci.crypt	86.38	76.57	95.30	92.34	95.65	72.56	99.40	97.80
sci.electronics	82.82	83.03	92.74	87.65	93.29	65.59	98.51	92.13
sci.med	84.49	75.68	94.34	86.18	94.19	67.31	98.61	96.80
sci.space	83.50	80.26	96.46	88.46	94.73	70.97	99.30	98.23
soc.religion.christian	88.18	74.95	98.33	97.36	99.90	80.80	99.50	99.11
talk.politics.guns	83.50	78.88	95.24	87.89	94.71	77.73	98.50	95.25
talk.politics.mideast	81.36	77.85	97.26	91.91	98.08	75.64	97.44	97.46
talk.politics.misc	78.96	80.25	90.12	86.08	95.03	68.31	96.77	94.28
talk.religion.misc	82.91	82.07	94.15	85.68	94.64	68.83	98.63	94.07
Trung bình các nhãn	82.94	78.58	94.21	88.79	94.41	71.35	**98.82**	95.45

Kết quả giá trị F1- score thu được từ bộ ngữ liệu 20 NewsGroups được trình bày trong Bảng PL2.3. Từ kết quả cho thấy, thuật toán C45 đạt giá trị F1- score cao nhất trên 18/20 nhãn, thuật toán SVM đạt giá trị F1- score cao nhất trên hai nhãn còn lại: *"soc.religion.christian"* và *"talk.politics.mideast"*.



Hình PL2.1: So sánh Accuracy và F1- score trên bộ 20 NewsGroups

Tổng hợp kết quả thực nghiệm từ bộ dữ liệu 20 NewsGroups cho thấy rằng thuật toán C45 cho kết quả tốt nhất, tiếp theo lần lượt là các thuật toán RF, SVM, và MNB. So sánh kết quả thực nghiệm giữa độ chính xác Accuracy và F1-score của các thuật toán trên bộ ngữ liệu 20 NewsGroups được minh họa trong Hình PL2.1

Kết quả độ chính xác Accuracy từ bộ ngữ liệu cảm xúc SemEval-2017 được trình bày trong Bảng PL2.4.

Bảng PL2.4: Độ chính xác của các thuật toán trên bộ ngữ liệu SemEval-2017

Nhãn	**CNN**	**W2V**	**MNB**	**NB**	**SVM**	**K-NN**	**C4.5**	**RF**
anger	64.04	66.18	78.67	70.20	83.67	53.47	53.16	74.69
fear	59.69	66.36	76.12	67.55	80.31	56.22	52.35	79.39
joy	65.18	72.81	78.47	66.43	86.22	60.41	57.55	88.27
sadness	62.08	65.65	78.67	70.51	85.10	55.61	54.90	81.43
Trung bình các nhãn	62.75	67.75	77.98	68.67	**83.83**	56.43	54.49	80.94

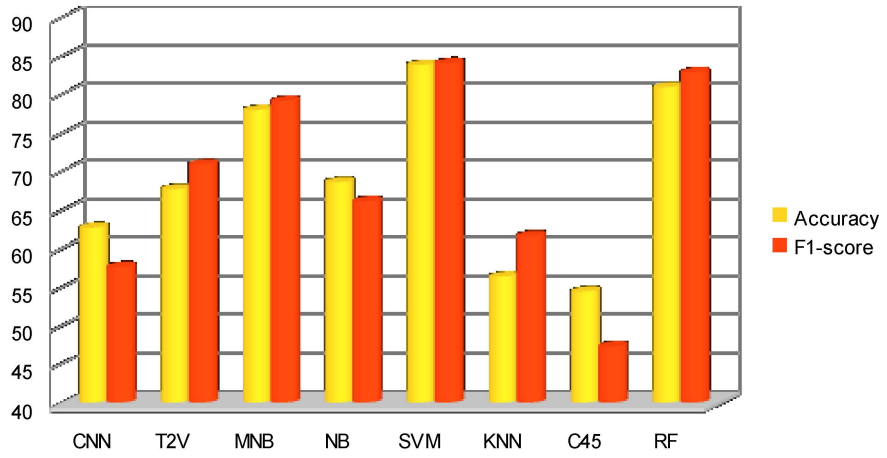
Từ kết quả cho thấy thuật toán SVM đạt giá trị Accuracy cao nhất trong 3/4 nhãn, thuật toán RF cho giá trị Accuracy cao nhất trên nhãn còn lại là *"joy"*.

Kết quả trung bình trên tất cả các nhãn, thuật toán SVM cho giá trị Accuracy cao nhất, tiếp theo lần lượt là các thuật toán RF và MNB.

Bảng PL2.5: F1 - score của các thuật toán trên bộ ngữ liệu SemEval-2017

Nhãn	**CNN**		**W2V**	**MNB**	**NB**	**SVM**	**K-NN**	**C4.5**	**RF**
anger	59.69		69.58	79.71	72.87	83.48	67.94	60.83	79.03
fear	54.05		66.99	77.27	63.45	81.57	69.56	52.27	81.95
joy	55.39		75.74	79.45	60.41	86.22	40.70	41.16	88.00
sadness	61.54		71.56	80.26	67.94	85.00	68.59	35.23	82.57
Trung bình các nhãn	57.66		70.97	79.17	66.17	**84.07**	61.70	47.37	82.89

Tổng hợp kết quả F1- score từ bộ ngữ liệu cảm xúc SemEval-2017 trình bày trong Bảng PL2.5. Từ kết quả đó cho thấy rằng, thuật toán SVM đạt giá trị F1- score cao nhất trên hai nhãn là *"joy"* và *"sadness"*. Kết quả trung bình trên tất cả các nhãn của bộ ngữ liệu thì thuật toán SVM cho kết quả F1-score cao nhất, tiếp theo lần lượt là các thuật toán RF và MNB.



Hình PL2.2: So sánh Accuracy và F1- score trên bộ SemEval-2017

Tổng hợp kết quả từ bộ ngữ liệu cảm xúc SemEval-2017 thì thuật toán SVM cho kết quả tốt nhất, tiếp theo lần lượt là thuật toán RF và MNB. Riêng trường hợp thuật toán C45 cho kết quả tốt với bộ dữ liệu 20 NewsGroups, nhưng với bộ dữ liệu cảm xúc được gọi là văn bản ngắn SemEval-2017, thì thuật toán C45 cho kết quả không cao, thậm chí là thấp nhất trong các thuật toán được xem xét. So sánh kết quả thực nghiệm giữa độ chính xác Accuracy và F1-score của các thuật toán trên bộ ngữ liệu SemEval-2017 được minh họa trong Hình PL2.2

Bảng PL2.6: Độ chính xác các thuật toán trên bộ ngữ liệu bài viết của luận án

Nhãn	**CNN**	**W2V**	**MNB**	**NB**	**SVM**	**K-NN**	**C4.5**	**RF**
Chính trị	71.91	66.38	***76.17***	76.17	68.51	58.72	73.62	62.13
Đời sống – Xã hội	63.91	62.17	***70.87***	70.00	63.91	58.70	70.00	60.43
Giáo dục	72.77	60.85	***78.72***	68.94	68.94	54.89	74.47	64.26
Khoa học – Công nghệ	62.76	68.80	71.91	62.55	69.36	42.55	62.55	***72.77***
Kinh doanh	***71.91***	68.09	66.38	71.06	66.81	58.30	65.53	69.79
Thời sự	56.52	56.35	57.39	49.13	56.09	57.83	56.96	***59.13***
Văn hóa – Giải trí	69.36	60.85	***77.02***	61.70	65.53	58.72	71.06	59.15
Pháp luật	73.62	77.02	***87.66***	65.96	84.26	45.96	70.64	74.04
Thể thao	67.83	76.65	***86.09***	70.00	68.70	40.87	80.43	69.13
Sức khỏe	76.49	78.30	***83.40***	68.09	73.19	56.60	72.34	73.19
Trung bình các nhãn	68.71	67.55	**75.56**	66.36	68.53	53.31	69.76	66.40

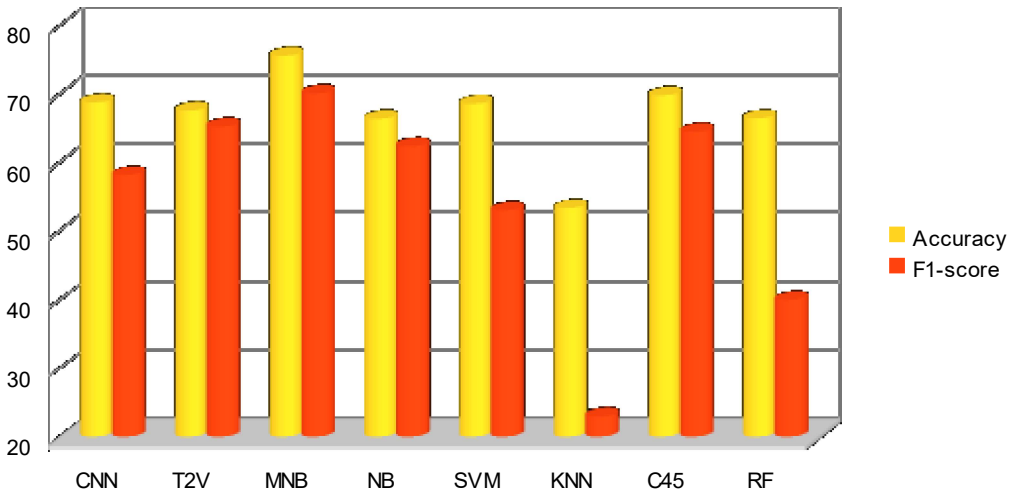
Với hai bộ ngữ liệu chủ đề và cảm xúc được xây dựng từ dữ liệu thu thập thực tế của luận án, kết quả độ chính xác Accuracy từ bộ ngữ liệu chủ đề của luận án trình bày trong Bảng PL2.6.

Từ kết quả cho thấy thuật toán MNB đạt giá trị Accuracy cao nhất trong 7/10 nhãn, thuật toán RF cho giá trị Accuracy cao nhất trên hai nhãn "*Khoa học – công nghệ"* và "*Thời sự"*, còn thuật toán CNN cho kết quả cao nhất trên nhãn "*Kinh doanh"*. Xét kết quả trung bình trên tất cả các nhãn, thuật toán MNB cho giá trị Accuracy cao nhất, tiếp theo lần lượt là các thuật toán C45, CNN và SVM.

Kết quả F1- score thu được từ bộ ngữ liệu chủ đề của luận án trình bày trong Bảng PL2.7. Từ kết quả cho thấy thuật toán MNB đạt giá trị F1- score cao nhất trong 8/10 nhãn. Thuật toán W2V đạt giá trị F1- score cao nhất trên hai nhãn là *"Thời sự"* và *"Pháp luật"*. Thuật toán NB cho giá trị F1- score cao nhất trên nhãn *"Kinh doanh"*. Xét kết quả trung bình trên tất cả các nhãn của bộ ngữ liệu chủ đề thì thuật toán MNB cho giá trị F1- score cao nhất, tiếp theo là các thuật toán W2V và C45.

Bảng PL2.7: Kết quả F1- score trên bộ ngữ liệu bài viết của luận án

Nhãn	**CNN**	**W2V**	**MNB**	**NB**	**SVM**	**K-NN**	**C4.5**	**RF**
Chính trị	60.66	57.47	***75.51***	74.34	47.66	6.90	67.81	27.23
Đời sống – Xã hội	65.49	67.17	***71.05***	65.21	36.86	0.00	63.10	9.64
Giáo dục	66.91	68.16	***79.04***	61.39	50.07	22.09	69.03	30.79
Khoa học – Công nghệ	34.14	58.84	***67.50***	48.05	48.28	59.70	55.69	52.44
Kinh doanh	65.62	59.53	48.05	***70.46***	69.04	8.87	60.22	69.38
Thời sự	52.57	***54.58***	43.65	45.71	16.86	0.00	48.88	2.00
Văn hóa – Giải trí	47.07	63.73	***72.12***	58.18	36.10	5.71	68.75	7.45
Pháp luật	57.09	77.05	***84.86***	63.75	80.53	49.42	66.25	63.74
Thể thao	69.89	68.77	***79.82***	70.17	71.03	58.02	76.64	68.40
Sức khỏe	62.12	76.87	***78.87***	66.24	73.37	18.33	68.16	68.25
Trung bình các nhãn	58.15	65.22	**70.05**	62.35	52.98	22.91	64.45	39.93



Hình PL2.3: So sánh Accuracy và F1- score trên bộ dữ liệu chủ đề của luận án

Tổng hợp kết quả từ bộ dữ liệu chủ đề của luận án thì thuật toán MNB cho kết quả cao nhất, tiếp theo là các thuật toán cho kết quả xếp xỉ nhau là W2V và C45. Thuật toán C45 và thuật toán SVM lần lượt cho kết quả tốt trong các bộ dữ liệu 20 NewsGroups và SemEval-2017 nhưng lại không cho kết quả cao trong bộ dữ liệu chủ

đề của luận án. So sánh kết quả thực nghiệm giữa độ chính xác Accuracy và F1-score của các thuật toán trên bộ ngữ liệu chủ đề của luận án được minh họa trong Hình 3.6

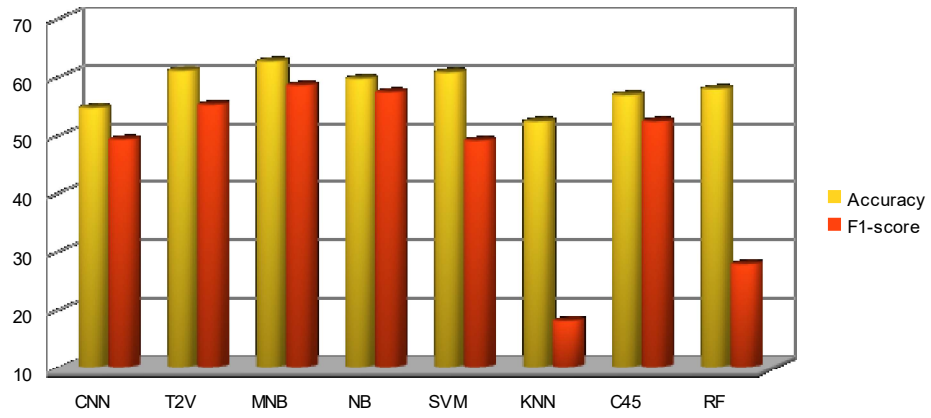
Bảng PL2.8: Độ chính xác các thuật toán trên bộ ngữ liệu cảm xúc của luận án

Nhãn	**CNN**	**W2V**	**MNB**	**NB**	**SVM**	**K-NN**	**C4.5**	**RF**
Anger	56.36	59.39	***62.73***	62.42	62.12	54.85	60.91	55.76
Disgust	55.76	58.48	***63.03***	60.61	62.73	54.85	56.36	54.85
Fear	53.64	***60.79***	56.67	56.36	58.18	55.45	52.12	55.76
Joy	47.88	60.03	***67.58***	58.48	61.82	50.30	52.12	59.70
Love	66.97	71.21	***72.73***	69.39	70.30	46.97	66.67	69.70
Sad	51.21	55.76	***57.58***	54.55	53.33	47.58	54.24	53.03
Other	49.39	***57.27***	56.67	54.25	55.15	54.85	53.94	54.85
Trung bình các nhãn	54.46	60.85	**62.42**	59.44	60.52	52.12	56.62	57.66

Kết quả thu được độ chính xác Accuracy từ bộ ngữ liệu cảm xúc của luận án trình bày trong Bảng PL2.8. Kết quả cho thấy thuật toán MNB đạt giá trị Accuracy cao nhất trong 5/7 nhãn. Thuật toán W2V cho giá trị Accuracy cao nhất trên hai nhãn còn lại là "*Fear"* và *"Other"*. Xét kết quả trung bình trên tất cả các nhãn, thuật toán MNB cho giá trị Accuracy cao nhất, tiếp theo lần lượt là các thuật toán W2V và thuật toán SVM.

Bảng PL2.9: F1- score các thuật toán trên bộ ngữ liệu cảm xúc của luận án

Nhãn	**CNN**	**W2V**	**MNB**	**NB**	**SVM**	**K-NN**	**C4.5**	**RF**
Anger	41.11	***66.68***	63.79	60.81	42.07	1.29	57.84	9.70
Disgust	***64.23***	60.93	62.36	58.82	47.12	1.29	50.87	6.21
Fear	27.61	47.63	***53.19***	47.26	36.82	3.87	48.13	7.46
Joy	53.16	57.99	58.95	58.96	***62.83***	15.87	45.77	61.65
Love	61.18	64.61	68.34	68.30	72.24	50.00	63.51	***72.46***
Sad	***59.36***	49.30	51.53	52.27	38.21	50.13	47.44	19.03
Other	36.37	37.89	49.98	***53.37***	42.18	2.58	51.37	16.48
Trung bình các nhãn	49.00	55.00	**58.30**	57.11	48.78	17.86	52.13	27.57



Hình PL2.4: So sánh Accuracy và F1- score trên bộ dữ liệu cảm xúc của luận án

Tổng hợp các kết quả từ bộ dữ liệu cảm xúc của luận án cho thấy thuật toán MNB cho kết cao nhất, tiếp theo là thuật toán NB và thuật toán W2V. Tương tự như kết quả trên bộ dữ liệu chủ đề của luận án, thuật toán C45 và SVM lần lượt cho kết quả tốt trong các bộ dữ liệu 20 NewsGroups và SemEval-2017 nhưng lại không cho kết quả cao trong bộ dữ liệu cảm xúc của luận án. So sánh kết quả thực nghiệm giữa độ chính xác Accuracy và F1-score của các thuật toán trên bộ ngữ liệu cảm xúc của luận án được minh họa trong Hình PL2.4.

Từ các kết quả Accuracy và F1-score thu được trên 4 bộ ngữ liệu thực nghiệm, xét trên các kết quả thực nghiệm, luận án lựa chọn sử dụng thuật toán MNB để gán nhãn hay xác định các đặc trưng chủ đề, đặc trưng cảm xúc và đặc trưng quan điểm của các bài viết trong các mô hình đề xuất của luận án ở các phần sau

PHỤ LỤC C: DANH MỤC CÁC TỪ DỪNG SỬ DỤNG TRONG LUẬN ÁN

a lô	a ha	ai	ai ai	ai nấy
ai đó	alô	amen	anh	anh ấy
ba	ba ba	ba bản	ba cùng	ba họ
ba ngày	ba ngôi	ba tăng	bao giờ	bao lâu
bao nhiêu	bao nả	bay biến	biết	biết bao
biết bao nhiêu	biết chắc	biết chừng nào	biết mình	biết mấy
biết thế	biết trước	biết việc	biết đâu	biết đâu chừng
biết đâu đấy	biết được	buổi	buổi làm	buổi mới
buổi ngày	buổi sớm	bà	bà ấy	bài
bài bác	bài bỏ	bài cái	bác	bán
bán cấp	bán dạ	bán thế	bây bẩy	bây chừ
bây giờ	bây nhiêu	bèn	béng	bên
bên bị	bên có	bên cạnh	bông	bước
bước khỏi	bước tới	bước đi	bạn	bản
bản bộ	bản riêng	bản thân	bản ý	bất chợt
bất cứ	bất giác	bất kì	bất kể	bất kỳ
bất luận	bất ngờ	bất nhược	bất quá	bất quá chỉ
bất thình lình	bất tử	bất đồ	bấy	bấy chầy
bấy chừ	bấy giờ	bấy lâu	bấy lâu nay	bấy nay
bấy nhiêu	bập bà bập bõm	bập bõm	bắt đầu	bắt đầu từ
bằng	bằng cứ	bằng không	bằng người	bằng nhau
bằng như	bằng nào	bằng nấy	bằng vào	bằng được
bằng ấy	bển	bệt	bị	bị chú
bị vì	bỏ	bỏ bà	bỏ cha	bỏ cuộc
bỏ không	bỏ lại	bỏ mình	bỏ mất	bỏ mẹ
bỏ nhỏ	bỏ quá	bỏ ra	bỏ riêng	bỏ việc
bỏ xa	bỗng	bỗng chốc	bỗng dưng	bỗng không
bỗng nhiên	bỗng nhưng	bỗng thấy	bỗng đâu	bộ
bộ thuộc	bộ điều	bội phần	bớ	bởi
bởi ai	bởi chưng	bởi nhưng	bởi sao	bởi thế
bởi thế cho nên	bởi tại	bởi vì	bởi vậy	bởi đâu
bức	cao	cao lâu	cao ráo	cao răng
cao sang	cao số	cao thấp	cao thế	cao xa
cha	cha chả	chao ôi	chia sẻ	chiếc
cho	cho biết	cho chắc	cho hay	cho nhau
cho nên	cho rằng	cho rồi	cho thấy	cho tin
cho tới	cho tới khi	cho về	cho ăn	cho đang
cho được	cho đến	cho đến khi	cho đến nỗi	choa
chu cha	chui cha	chung	chung cho	chung chung
chung cuộc	chung cục	chung nhau	chung qui	chung quy
chung quy lại	chung ái	chuyển	chuyển tự	chuyển đạt
chuyện	chuẩn bị	chành chạnh	chí chết	chính
chính bản	chính giữa	chính là	chính thị	chính điểm
chùn chùn	chùn chũn	chú	chú dẫn	chú khách
chú mày	chú mình	chúng	chúng mình	chúng ta
chúng tôi	chúng ông	chăn chắn	chăng	chăng chắc
chăng nữa	chơi	chơi họ	chưa	chưa bao giờ
chưa chắc	chưa có	chưa cần	chưa dùng	chưa dễ
chưa kể	chưa tính	chưa từng	chầm chập	chậc
chắc	chắc chắn	chắc dạ	chắc hẳn	chắc lòng
chắc người	chắc vào	chắc ăn	chẳng lẽ	chẳng những
chẳng nữa	chẳng phải	chết nỗi	chết thật	chết tiệt

chỉ	chỉ chính	chỉ có	chỉ là	chỉ tên
chỉn	chị	chị bộ	chị ấy	chịu
chịu chưa	chịu lời	chịu tốt	chịu ăn	chọn
chọn bên	chọn ra	chốc chốc	chớ	chớ chi
chớ gì	chớ không	chớ kể	chớ như	chợt
chợt nghe	chợt nhìn	chủn	chứ	chứ ai
chứ còn	chứ gì	chứ không	chứ không phải	chứ lại
chứ lị	chứ như	chứ sao	coi bộ	coi mòi
con	con con	con dạ	con nhà	con tính
cu cậu	cuối	cuối cùng	cuối điểm	cuốn
cuộc	càng	càng càng	càng hay	cá nhân
các	các cậu	cách	cách bức	cách không
cách nhau	cách đều	cái	cái gì	cái họ
cái đã	cái đó	cái ấy	câu hỏi	cây
cây nước	còn	còn như	còn nữa	còn thời gian
còn về	có	có ai	có chuyện	có chăng
có chăng là	có chứ	có cơ	có dễ	có họ
có khi	có ngày	có người	có nhiều	có nhà
có phải	có số	có tháng	có thể	có thể
có vẻ	có ý	có ăn	có điều	có điều kiện
có đáng	có đâu	có được	cóc khô	cô
cô mình	cô quả	cô tăng	cô ấy	công nhiên
cùng	cùng chung	cùng cực	cùng nhau	cùng tuổi
cùng tột	cùng với	cùng ăn	căn	căn cái
căn cắt	căn tính	cũng	cũng như	cũng nên
cũng thế	cũng vậy	cũng vậy thôi	cũng được	cơ
cơ chỉ	cơ chừng	cơ cùng	cơ dẫn	cơ hồ
cơ hội	cơ mà	cơn	cả	cả nghe
cả nghĩ	cả ngày	cả người	cả nhà	cả năm
cả thảy	cả thể	cả tin	cả ăn	cả đến
cảm thấy	cảm ơn	cấp	cấp số	cấp trực tiếp
cần	cần cấp	cần gì	cần số	cật lực
cật sức	cậu	cổ lai	cụ thể	cụ thể là
cụ thể như	của	của ngọt	của tin	cứ
cứ như	cứ việc	cứ điểm	cực lực	do
do vì	do vậy	do đó	duy	duy chỉ
duy có	dài	dài lời	dài ra	dành
dành dành	dào	dì	dù	dù cho
dù dì	dù gì	dù rằng	dù sao	dùng
dùng cho	dùng hết	dùng làm	dùng đến	dưới
dưới nước	dạ	dạ bán	dạ con	dạ dài
dạ dạ	dạ khách	dần dà	dần dần	dầu sao
dẫn	dẫu	dẫu mà	dẫu rằng	dẫu sao
dễ	dễ dùng	dễ gì	dễ khiến	dễ nghe
dễ ngươi	dễ như chơi	dễ sợ	dễ sử dụng	dễ thường
dễ thấy	dễ ăn	dễ đâu	dở chừng	dữ
dữ cách	em	em em	giá trị	giá trị thực tế
giảm	giảm chính	giảm thấp	giảm thế	giống
giống người	giống nhau	giống như	giờ	giờ lâu
giờ này	giờ đi	giờ đây	giờ đến	giữ
giữ lấy	giữ ý	giữa	giữa lúc	gây
gây cho	gây giống	gây ra	gây thêm	gì
gì gì	gì đó	gần	gần bên	gần hết
gần ngày	gần như	gần xa	gần đây	gần đến

gặp	gặp khó khăn	gặp phải	gồm	hay
hay biết	hay hay	hay không	hay là	hay làm
hay nhỉ	hay nói	hay sao	hay tin	hay đâu
hiểu	hiện nay	hiện tại	hoàn toàn	hoặc
hoặc là	hãy	hãy còn	hơn	hơn cả
hơn hết	hơn là	hơn nữa	hơn trước	hầu hết
hết	hết chuyện	hết cả	hết của	hết nói
hết ráo	hết rồi	hết ý	họ	họ gần
họ xa	hỏi	hỏi lại	hỏi xem	hỏi xin
hỗ trợ	khi	khi khác	khi không	khi nào
khi nên	khi trước	khiến	khoảng	khoảng cách
khoảng không	khá	khá tốt	khác	khác gì
khác khác	khác nhau	khác nào	khác thường	khác xa
khách	khó	khó biết	khó chơi	khó khăn
khó làm	khó mở	khó nghe	khó nghĩ	khó nói
khó thấy	khó tránh	không	không ai	không bao giờ
không bao lâu	không biết	không bán	không chỉ	không còn
không có	không có gì	không cùng	không cần	không cứ
không dùng	không gì	không hay	không khỏi	không kể
không ngoài	không nhận	không những	không phải	không phải không
không thể	không tính	không điều kiện	không được	không đầy
không để	khẳng định	khỏi	khỏi nói	kể
kể cả	kể như	kể tới	kể từ	liên quan
loại	loại từ	luôn	luôn cả	luôn luôn
luôn tay	là	là cùng	là là	là nhiều
là phải	là thế nào	là vì	là ít	làm
làm bằng	làm cho	làm dần dần	làm gì	làm lòng
làm lại	làm lấy	làm mất	làm ngay	làm như
làm nên	làm ra	làm riêng	làm sao	làm theo
làm thế nào	làm tin	làm tôi	làm tăng	làm tại
làm tắp lự	làm vì	làm đúng	làm được	lâu
lâu các	lâu lâu	lâu nay	lâu ngày	lên
lên cao	lên cơn	lên mạnh	lên ngôi	lên nước
lên số	lên xuống	lên đến	lòng	lòng không
lúc	lúc khác	lúc lâu	lúc nào	lúc này
lúc sáng	lúc trước	lúc đi	lúc đó	lúc đến
lúc ấy	lý do	lượng	lượng cả	lượng số
lượng từ	lại	lại bộ	lại cái	lại còn
lại giống	lại làm	lại người	lại nói	lại nữa
lại quả	lại thôi	lại ăn	lại đây	lấy
lấy có	lấy cả	lấy giống	lấy làm	lấy lý do
lấy lại	lấy ra	lấy ráo	lấy sau	lấy số
lấy thêm	lấy thế	lấy vào	lấy xuống	lấy được
lấy để	lần	lần khác	lần lần	lần nào
lần này	lần sang	lần sau	lần theo	lần trước
lần tìm	lớn	lớn lên	lớn nhỏ	lời
lời chú	lời nói	mang	mang lại	mang mang
mang nặng	mang về	muốn	mà	mà cả
mà không	mà lại	mà thôi	mà vẫn	mình
mạnh	mất	mất còn	mọi	mọi giờ
mọi khi	mọi lúc	mọi người	mọi nơi	mọi sự
mọi thứ	mọi việc	mối	mỗi	mỗi lúc
mỗi lần	mỗi một	mỗi ngày	mỗi người	một
một cách	một cơn	một khi	một lúc	một số

một vài	một ít	mới	mới hay	mới rồi
mới đây	mở	mở mang	mở nước	mở ra
mợ	mức	nay	ngay	ngay bây giờ
ngay cả	ngay khi	ngay khi đến	ngay lúc	ngay lúc này
ngay lập tức	ngay thật	ngay tức khắc	ngay tức thì	ngay từ
nghe	nghe chừng	nghe hiểu	nghe không	nghe lại
nghe nhìn	nghe như	nghe nói	nghe ra	nghe rõ
nghe thấy	nghe tin	nghe trực tiếp	nghe đâu	nghe đâu như
nghe được	nghen	nghiễm nhiên	nghĩ	nghĩ lại
nghĩ ra	nghĩ tới	nghĩ xa	nghĩ đến	nghỉm
ngoài	ngoài này	ngoài ra	ngoài xa	ngoải
nguồn	ngày	ngày càng	ngày cấp	ngày giờ
ngày ngày	ngày nào	ngày này	ngày nọ	ngày qua
ngày rày	ngày tháng	ngày xưa	ngày xửa	ngày đến
ngày ấy	ngôi	ngôi nhà	ngôi thứ	ngõ hầu
ngăn ngắt	ngươi	người	người hỏi	người khác
người khách	người mình	người nghe	người người	người nhận
ngọn	ngọn nguồn	ngọt	ngồi	ngồi bệt
ngồi không	ngồi sau	ngồi trệt	ngộ nhỡ	nhanh
nhanh lên	nhanh tay	nhau	nhiên hậu	nhiều
nhiều ít	nhiệt liệt	nhung nhăng	nhà	nhà chung
nhà khó	nhà làm	nhà ngoài	nhà ngươi	nhà tôi
nhà việc	nhân dịp	nhân tiện	nhé	nhìn
nhìn chung	nhìn lại	nhìn nhận	nhìn theo	nhìn thấy
nhìn xuống	nhóm	nhón nhén	như	như ai
như chơi	như không	như là	như nhau	như quả
như sau	như thường	như thế	như thế nào	như thể
như trên	như trước	như tuồng	như vậy	như ý
nhưng	nhưng mà	nhược bằng	nhất	nhất loạt
nhất luật	nhất là	nhất mực	nhất nhất	nhất quyết
nhất sinh	nhất thiết	nhất thì	nhất tâm	nhất tề
nhất đán	nhất định	nhận	nhận biết	nhận họ
nhận làm	nhận nhau	nhận ra	nhận thấy	nhận việc
nhận được	nhằm	nhằm khi	nhằm lúc	nhằm vào
nhằm để	nhỉ	nhỏ	nhỏ người	nhớ
nhớ bập bõm	nhớ lại	nhớ lấy	nhớ ra	nhờ
nhờ chuyển	nhờ có	nhờ nhờ	nhờ đó	nhỡ ra
những	những ai	những khi	những là	những lúc
những muốn	những như	nào	nào cũng	nào hay
nào là	nào phải	nào đâu	nào đó	này
này nọ	nên	nên chi	nên chăng	nên làm
nên người	nên tránh	nó	nóc	nói
nói bông	nói chung	nói khó	nói là	nói lên
nói lại	nói nhỏ	nói phải	nói qua	nói ra
nói riêng	nói rõ	nói thêm	nói thật	nói toẹt
nói trước	nói tốt	nói với	nói xa	nói ý
nói đến	nói đủ	năm	năm tháng	nơi
nơi nơi	nước	nước bài	nước cùng	nước lên
nước nặng	nước quả	nước xuống	nước ăn	nước đến
nấy	nặng	nặng căn	nặng mình	nặng về
nếu	nếu có	nếu cần	nếu không	nếu mà
nếu như	nếu thế	nếu vậy	nếu được	nền
nọ	nớ	nức nở	nữa	nữa khi
nữa là	nữa rồi	oai oái	oái	pho

phè	phè phè	phía	phía bên	phía bạn
phía dưới	phía sau	phía trong	phía trên	phía trước
phóc	phót	phù hợp	phăn phắt	phương chi
phải	phải biết	phải chi	phải chăng	phải cách
phải cái	phải giờ	phải khi	phải không	phải lại
phải lời	phải người	phải như	phải rồi	phải tay
phần	phần lớn	phần nhiều	phần nào	phần sau
phần việc	phắt	phỉ phui	phỏng	phỏng như
phỏng nước	phỏng theo	phỏng tính	phốc	phụt
phứt	qua	qua chuyện	qua khỏi	qua lại
qua lần	qua ngày	qua tay	qua thì	qua đi
quan trọng	quan trọng vấn đề	quan tâm	quay	quay bước
quay lại	quay số	quay đi	quá	quá bán
quá bộ	quá giờ	quá lời	quá mức	quá nhiều
quá tay	quá thì	quá tin	quá trình	quá tuổi
quá đáng	quá ư	quả	quả là	quả thật
quả thế	quả vậy	quận	ra	ra bài
ra bộ	ra chơi	ra gì	ra lại	ra lời
ra ngôi	ra người	ra sao	ra tay	ra vào
ra ý	ra điều	ra đây	ren rén	riu ríu
riêng	riêng từng	riệt	rày	ráo
ráo cả	ráo nước	ráo trọi	rén	rén bước
rích	rón rén	rõ	rõ là	rõ thật
rút cục	răng	răng răng	rất	rất lâu
rằng	rằng là	rốt cuộc	rốt cục	rồi
rồi nữa	rồi ra	rồi sao	rồi sau	rồi tay
rồi thì	rồi xem	rồi đây	rứa	sa sả
sang	sang năm	sang sáng	sang tay	sao
sao bản	sao bằng	sao cho	sao vậy	sao đang
sau	sau chót	sau cuối	sau cùng	sau hết
sau này	sau nữa	sau sau	sau đây	sau đó
so	so với	song le	suýt	suýt nữa
sáng	sáng ngày	sáng rõ	sáng thế	sáng ý
sì	sì sì	sất	sắp	sắp đặt
sẽ	sẽ biết	sẽ hay	số	số cho biết
số cụ thể	số loại	số là	số người	số phần
số thiếu	sốt sột	sớm	sớm ngày	sở dĩ
sử dụng	sự	sự thế	sự việc	tanh
tanh tanh	tay	tay quay	tha hồ	tha hồ chơi
tha hồ ăn	than ôi	thanh	thanh ba	thanh chuyển
thanh không	thanh thanh	thanh tính	thanh điều kiện	thanh điểm
thay đổi	thay đổi tình trạng	theo	theo bước	theo như
theo tin	thi thoảng	thiếu	thiếu gì	thiếu điểm
thoạt	thoạt nghe	thoạt nhiên	thoắt	thuần
thuần ái	thuộc	thuộc bài	thuộc cách	thuộc lại
thuộc từ	thà	thà là	thà rằng	thành ra
thành thử	thái quá	tháng	tháng ngày	tháng năm
tháng tháng	thêm	thêm chuyện	thêm giờ	thêm vào
thì	thì giờ	thì là	thì phải	thì ra
thì thôi	thình lình	thích	thích cứ	thích thuộc
thích tự	thích ý	thím	thôi	thôi việc
thúng thắng	thương ôi	thường	thường bị	thường hay
thường khi	thường số	thường sự	thường thôi	thường thường
thường tính	thường tại	thường xuất hiện	thường đến	thảo hèn

thảo nào	thấp	thấp cơ	thấp thỏm	thấp xuống
thấy	thấy tháng	thấy	thậm	thậm chí
thậm cấp	thậm từ	thật	thật chắc	thật là
thật lực	thật quả	thật ra	thật sự	thật thà
thật tốt	thật vậy	thế	thế chuẩn bị	thế là
thế lại	thế mà	thế nào	thế nên	thế ra
thế sự	thế thì	thế thôi	thế thường	thế thế
thế à	thế đó	thếch	thỉnh thoảng	thỏm
thốc	thốc tháo	thốt	thốt nhiên	thốt nói
thốt thôi	thộc	thời gian	thời gian sử dụng	thời gian tính
thời điểm	thục mạng	thứ	thứ bản	thứ đến
thửa	thực hiện	thực hiện đúng	thực ra	thực sự
thực tế	thực vậy	tin	tin thêm	tin vào
tiếp theo	tiếp tục	tiếp đó	tiện thể	toà
toé khói	toẹt	trong	trong khi	trong lúc
trong mình	trong ngoài	trong này	trong số	trong vùng
trong đó	trong ấy	tránh	tránh khỏi	tránh ra
tránh tình trạng	tránh xa	trên	trên bộ	trên dưới
trước	trước hết	trước khi	trước kia	trước nay
trước ngày	trước nhất	trước sau	trước tiên	trước tuổi
trước đây	trước đó	trả	trả của	trả lại
trả ngay	trả trước	trếu tráo	trển	trệt
trệu trạo	trỏng	trời đất ơi	trở thành	trừ phi
trực tiếp	trực tiếp làm	tuy	tuy có	tuy là
tuy nhiên	tuy rằng	tuy thế	tuy vậy	tuy đã
tuyệt nhiên	tuần tự	tuốt luốt	tuốt tuồn tuột	tuốt tuột
tuổi	tuổi cả	tuổi tôi	tà tà	tên
tên chính	tên cái	tên họ	tên tự	tênh
tênh tênh	tìm	tìm bạn	tìm cách	tìm hiểu
tìm ra	tìm việc	tình trạng	tính	tính cách
tính căn	tính người	tính phỏng	tính từ	tít mù
tò te	tôi	tôi con	tông tốc	tù tì
tăm tắp	tăng	tăng chúng	tăng cấp	tăng giảm
tăng thêm	tăng thế	tại	tại lòng	tại nơi
tại sao	tại tôi	tại vì	tại đâu	tại đây
tại đó	tạo	tạo cơ hội	tạo nên	tạo ra
tạo ý	tạo điều kiện	tấm	tấm bản	tấm các
tấn	tấn tới	tất cả	tất cả bao nhiêu	tất thảy
tất tần tật	tất tật	tập trung	tắp	tắp lự
tắp tắp	tọt	tỏ ra	tỏ vẻ	tốc tả
tối ư	tốt	tốt bạn	tốt bộ	tốt hơn
tốt mối	tốt ngày	tột	tột cùng	tớ
tới	tới gần	tới mức	tới nơi	tới thì
tức thì	tức tốc	từ	từ căn	từ giờ
từ khi	từ loại	từ nay	từ thế	từ tính
từ tại	từ từ	từ ái	từ điều	từ đó
từ ấy	từng	từng cái	từng giờ	từng nhà
từng phần	từng thời gian	từng đơn vị	từng ấy	tự
tự cao	tự khi	tự lượng	tự tính	tự tạo
tự vì	tự ý	tự ăn	tựu trung	veo
veo veo	việc	việc gì	vung thiên địa	vung tàn tán
vung tán tàn	và	vài	vài ba	vài người
vài nhà	vài nơi	vài tên	vài điều	vào
vào gặp	vào khoảng	vào lúc	vào vùng	vào đến

vâng	vâng chịu	vâng dạ	vâng vâng	vâng ý
vèo	vèo vèo	vì	vì chưng	vì rằng
vì sao	vì thế	vì vậy	ví bằng	ví dù
ví phỏng	ví thử	vô hình trung	vô kể	vô luận
vô vàn	vùng	vùng lên	vùng nước	văng tê
vượt	vượt khỏi	vượt quá	vạn nhất	vả chăng
vả lại	vấn đề	vấn đề quan trọng	vẫn	vẫn thế
vậy	vậy là	vậy mà	vậy nên	vậy ra
vậy thì	vậy ư	về	về không	về nước
về phần	về sau	về tay	vị trí	vị tất
vốn dĩ	với	với lại	với nhau	vở
vụt	vừa	vừa khi	vừa lúc	vừa mới
vừa qua	vừa rồi	vừa vừa	xa	xa cách
xa gần	xa nhà	xa tanh	xa tắp	xa xa
xa xả	xem	xem lại	xem ra	xem số
xin	xin gặp	xin vâng	xiết bao	xon xón
xoành xoạch	xoét	xoắn	xoẹt	xuất hiện
xuất kì bất ý	xuất kỳ bất ý	xuể	xuống	xăm xúi
xăm xăm	xăm xắm	xảy ra	xềnh xệch	xệp
xử lý	yêu cầu	à	à này	à ơi
ào	ào vào	ào ào	á	á à
ái	ái chà	ái dà	áng	áng như
âu là	ít	ít biết	ít có	ít hơn
ít khi	ít lâu	ít nhiều	ít nhất	ít nữa
ít quá	ít ra	ít thôi	ít thấy	ô hay
ô hô	ô kê	ô kìa	ôi chao	ôi thôi
ông	ông nhỏ	ông tạo	ông từ	ông ấy
ông ổng	úi	úi chà	úi dào	ý
ý chừng	ý da	ý hoặc	ăn	ăn chung
ăn chắc	ăn chịu	ăn cuộc	ăn hết	ăn hỏi
ăn làm	ăn người	ăn ngồi	ăn quá	ăn riêng
ăn sáng	ăn tay	ăn trên	ăn về	đang
đang tay	đang thì	điều	điều gì	điều kiện
điểm	điểm chính	điểm gặp	điểm đầu tiên	đành đạch
đáng	đáng kể	đáng lí	đáng lý	đáng lẽ
đáng số	đánh giá	đánh đùng	đáo để	đâu
đâu có	đâu cũng	đâu như	đâu nào	đâu phải
đâu đâu	đâu đây	đâu đó	đây	đây này
đây rồi	đây đó	đã	đã hay	đã không
đã là	đã lâu	đã thế	đã vậy	đã đủ
đó	đó đây	đúng	đúng ngày	đúng ra
đúng tuổi	đúng với	đơn vị	đưa	đưa cho
đưa chuyện	đưa em	đưa ra	đưa tay	đưa tin
đưa tới	đưa vào	đưa về	đưa xuống	đưa đến
được	được cái	được lời	được nước	được tin
đại loại	đại nhân	đại phàm	đại để	đạt
đảm bảo	đầu tiên	đầy	đầy năm	đầy phè
đầy tuổi	đặc biệt	đặt	đặt làm	đặt mình
đặt mức	đặt ra	đặt trước	đặt để	đến
đến bao giờ	đến cùng	đến cùng cực	đến cả	đến giờ
đến gần	đến hay	đến khi	đến lúc	đến lời
đến nay	đến ngày	đến nơi	đến nỗi	đến thì
đến thế	đến tuổi	đến xem	đến điều	đến đâu
đều	đều bước	đều nhau	đều đều	để

để cho	để giống	để không	để lòng	để lại
để mà	để phần	để được	để đến nỗi	đối với
đồng thời	đủ	đủ dùng	đủ nơi	đủ số
đủ điều	đủ điểm	ơ	ơ hay	ơ kìa
ơi	ơi là	ư	ạ	ạ ơi
ấy	ấy là	ầu ơ	ắt	ắt hẳn
ắt là	ắt phải	ắt thật	ối dào	ối giời
ối giời ơi	ồ	ồ ồ	ổng	ớ
ớ này	ờ	ờ ờ	ở	ở lại
ở như	ở nhờ	ở năm	ở trên	ở vào
ở đây	ở đó	ở được	ủa	ứ hự
ứ ừ	ừ	ừ nhé	ừ thì	ừ ào
ừ ừ	ử			