

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

DOÃN THỊ HUYỀN TRANG

**TRÍCH XUẤT Ý ĐỊNH NGƯỜI DÙNG MUA HÀNG
TRÊN MẠNG XÃ HỘI
SỬ DỤNG PHƯƠNG PHÁP SUY LUẬN CÁC MÔ HÌNH**

LUẬN VĂN THẠC SĨ

Chuyên ngành : Quản lý Hệ thống thông tin

.

TÓM TẮT LUẬN VĂN THẠC SĨ

Hà Nội, 04 tháng 12 năm 2016

MỞ ĐẦU

1. Tính cấp thiết của đề tài

Sức nóng và độ lan tỏa của mạng xã hội (Social Network - SN) đã và đang phát triển dữ dội và không hề thấy dấu hiệu thuyên giảm. Sự tăng trưởng nhanh chóng của mạng xã hội đã thu hút một lượng lớn số nhà nghiên cứu khám phá và nghiên cứu về miền lĩnh vực rộng lớn này.

Nhận diện, trích xuất ý định nói chung và ý định mua hàng của người dùng nói riêng đã và đang là một đề tài nghiên cứu thời sự [16], dự đoán được ý định của người dùng từ những hành vi của họ là chủ đề nghiên cứu nhận được sự quan tâm đặc biệt các nhóm nghiên cứu của các tác giả Xiao Ding cùng cộng sự [16], Fu cùng cộng sự [15]. Với doanh nghiệp, các nhà cung cấp dịch vụ việc biết được ý định, mong muốn của người dùng sẽ giúp họ cải tiến tốt hơn sản phẩm, hệ thống của mình để đảm bảo cung cấp đúng nội dung khách hàng cần, mở rộng số lượng người dùng quan tâm, quảng bá thương hiệu, hình ảnh. Bên cạnh đó, việc phát hiện ý định người dùng trên mạng xã hội được doanh nghiệp, cá nhân quan tâm để đưa ra những tư vấn dịch vụ, sản phẩm phù hợp. Hơn thế nữa, kết quả của bài toán khai thác ý định người dùng có thể được ứng dụng làm đầu vào cho các bài toán khác như xây dựng hệ tư vấn xã hội dựa trên ý định người dùng, dự đoán sở thích người dùng, dự đoán xu hướng tương lai,

2. Mục tiêu nghiên cứu

Trong bài viết của mình, tôi tập trung vào việc nhận diện và trích xuất ra nhu cầu, mong muốn, ý định mua hàng của người dùng trên mạng xã hội từ hành vi của họ. Hành vi người dùng trên mạng xã hội bao gồm nhiều hoạt động, chẳng hạn như thiết lập các mối quan hệ: bạn bè, gia đình, thần tượng...; đăng tải hoặc bình luận các nội dung hay thông tin; thiết lập nhu cầu sở thích bằng việc thích (like) hoặc tham gia vào các trang (page) hoặc các nhóm (group).... Đáng chú ý, không phải tất cả các hoạt động hay hành vi của người dùng đều được thể hiện rõ ràng và là nguồn dữ liệu, tài nguyên có ích. Do vậy, luận văn này tập trung vào hành vi đăng tải bài viết và bình luận, một trong những hành vi phổ biến và thể hiện rõ nhất mong muốn, ý định của một người dùng bất kỳ.

Dựa trên những hướng tiếp cận đã đề cập ở trên, trong luận văn này, tôi tiến hành áp dụng phương pháp suy luận các mô hình vào bài toán khai thác ý định mua hàng người dùng trên mạng xã hội cụ thể là trên Facebook dựa vào hành vi đăng tải và bình luận của họ trên các trang bán hàng (fanpage).

TÓM TẮT LUẬN VĂN

1. Tầm quan trọng của ý định người dùng trên mạng xã hội

Việc nhận diện ra những ý định từ người dùng là một yếu tố quan trọng cho các nhà cung cấp dịch vụ hay các doanh nghiệp tổ chức thông qua những bài viết, lượt thích (Like) hay những bình luận của họ Nelson-Field và các đồng nghiệp. (2012) đã công nhận rằng tiềm năng của Facebook đạt một phần ba dân số thế giới, và vì vậy Facebook đang trở thành một công cụ ngày càng quan trọng đối với các nhà tiếp thị thông qua việc nắm rõ nhu cầu, mong muốn của người dùng hoặc gọi chung là ý định người dùng. Bujega (2006) chỉ ra lợi ích của việc định hướng tiếp thị và quảng cáo của họ đến đúng những người dùng từ hành vi, thói quen mua sắm mà họ đã từng có.

Người sử dụng không phải lúc nào cũng biết chính xác những gì họ muốn. Đôi khi, họ chỉ biết rằng họ muốn được giúp đỡ để tìm ra những gì họ muốn. Điều này sẽ thuộc về trách nhiệm của các nhà tiếp thị và các nhà sản xuất sản phẩm. Facebook đã trở thành một trong những nền tảng mà người dùng có thể trình bày quan điểm, nhu cầu, ý định của họ về sản phẩm, cuộc sống và những gì trong tâm trí họ. Do vậy, những thông tin được đưa ra nếu được khai thác sẽ là một kho thông tin quý báu cho các bên liên quan. Vậy, ý định người dùng có tầm quan trọng như thế nào? Theo Long Jin và cộng sự, ý định, mong muốn hay nhu cầu người dùng trên mạng xã hội quan trọng khác nhau với các đối tượng Internet khác nhau ở nhiều khía cạnh:

- **Đối với nhà cung cấp dịch vụ Internet (Internet Service Provider - ISP):** Họ sẽ biết được sự phát triển của mạng xã hội, từ đó họ có thể có những nghiên cứu nhằm phát triển hoặc cải thiện mô hình giao thông, luồng giao tiếp trên mạng xã hội chẳng hạn như việc thiết lập một hành động của cơ sở hạ tầng.
- **Đối với các nhà cung cấp dịch vụ mạng xã hội:** Nó giúp họ hiểu thái độ của khách hàng hướng tới việc cải thiện dịch vụ. Hơn thế nữa, từ quan điểm của việc đầu tư cơ sở hạ tầng, chẳng hạn như những vị trí nào hiệu quả chi phí nhất để xây dựng các trung tâm dữ liệu hoặc cụm mạng lưới phân phối nội dung (Content Delivery Network - CDN) có thể được khai thác để cung cấp dữ liệu được truy cập một cách thường xuyên, hiệu biết, nắm được phân bố địa lý và hoạt động giao thông của người sử dụng cũng là những nguồn thông tin quan trọng.

- **Với các nhà nghiên cứu:** Phát hiện được ý định người dùng sẽ là bài toán con cho các nghiên cứu quan trọng. Ví dụ, để xây dựng hệ tư vấn người dùng, trước hết họ cần phải xác định được người dùng thích gì, người dùng mong muốn gì từ những hành vi của họ để từ đó mới có thể tư vấn cho họ theo đúng hướng họ muốn. Vậy thì bài toán nhận diện hay trích xuất ra ý định người dùng là một trong những bài toán con quan trọng của chủ đề này. Hay, với đề tài dự đoán tính cách người dùng, để biết được người dùng có tính cách như nào, sở thích ra sao thì họ cũng cần biết được là người dùng thường có những thói quen gì, họ hay làm gì, họ hay nghĩ gì và mong muốn gì. Tất cả đều liên quan đến việc hiểu ý định hay nhu cầu của người dùng.
- **Với các nhà kinh doanh, công ty, tổ chức cung cấp sản phẩm, dịch vụ:** Khi nắm được ý định người dùng, phần nào đó họ sẽ biết được về thị hiếu người dùng, thống kê được mức độ tiêu thụ hàng hóa, biết và cải thiện được chiến lược kinh doanh nếu cần,

2. Định nghĩa ý định người dùng

Với từng miền ứng dụng khác nhau sẽ có những định nghĩa khác nhau về ý định người dùng. Theo Bratman (1987): “Ý định là một trạng thái đại diện cho suy nghĩ thực hiện một hoặc nhiều hành động trong tương lai. Ý định bao gồm những hành động như kế hoạch hoặc suy nghĩ tính trước. Ý định có thể ở trạng thái rõ ràng – explicitly hoặc tiềm ẩn/không rõ ràng – implicitly, trực tiếp hoặc gián tiếp. Ý định rõ ràng là một tuyên bố rõ ràng và trực tiếp của người dùng về những gì người đó có kế hoạch làm.”. Theo Zhiyuan Chen, Bing Liu và cộng sự ý định có hai loại là ý định ẩn và ý định rõ ràng. Ý định rõ ràng tức là mong muốn của người dùng được thể hiện rõ ràng không cần kết hợp. Những trường hợp ý định kết hợp được xếp vào ý định ẩn. Jinpeng Wang **Error! Reference source not found.** đã đề xuất việc thống kê các bài viết ý định thành 6 loại là: Đồ ăn và Nước Uống (Food & Drink), Du lịch (Travel), Sức khỏe và Giáo dục (Career & Education), Hàng hóa và Dịch vụ (Goods & Services), Sự kiện và Hoạt động (Event & Activities), các loại khác (Trifle). Trong luận văn này, tôi sẽ chỉ tập trung vào ý định rõ ràng về việc mua một sản phẩm/dịch vụ của người dùng qua các bình luận của họ trên các trang bán hàng của facebook.

3. Phát biểu bài toán và một số khó khăn thách thức

Mục đích của bài toán: Nhận diện được ý định người dùng về việc mua một sản phẩm, dịch vụ bất kỳ từ những bài đăng/bình luận của họ.

Đầu vào:

- Các bài đăng, bình luận trong các trang bán hàng (fanpage) trên mạng xã hội

Đầu ra:

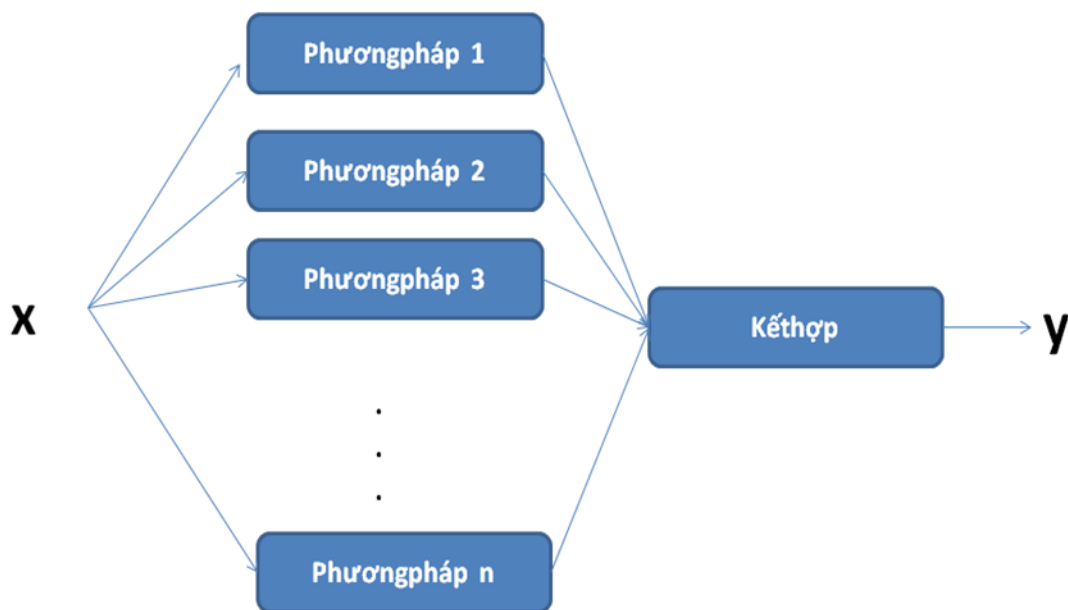
- Các ý định trong bài đăng, bình luận được phát hiện.

Không giống như cách sử dụng từ ngữ trong các văn bản chính thống, từ ngữ trên các diễn đàn trực tuyến hoặc mạng xã hội được sử dụng rất thoải mái tùy theo sở thích và ý đồ của tác giả do vậy mà việc nhận diện hay tìm ra được ý định rõ ràng của người dùng gặp khá nhiều khó khăn. Nói chung, các câu/từ được sử dụng theo thể tự do của mỗi người dùng: bài viết, bình luận có thể quá dài hoặc quá ngắn, người dùng sử dụng tiếng lóng, teen code, sai chính tả, các từ viết tắt, hashtags Thực tế là có thể được khai thác để xây dựng một bộ phân loại ý định rõ ràng và ý định ẩn dựa trên các dữ liệu đã được gán nhãn trong một số lĩnh vực và áp dụng nó vào một miền/mục tiêu mới mà không cần phải gán nhãn cho bất kỳ dữ liệu huấn luyện trong miền mục tiêu. Tuy nhiên, như vậy sẽ làm dữ liệu bị bó hẹp trong một miền nào đó. Bởi vậy, trong bài toán của mình, tôi đã quyết định xây dựng tập dữ liệu test hoàn toàn mới. Chính vì thế bài toán tìm và nhận diện được chính xác ý định của người dùng trên mạng xã hội gặp khá nhiều khó khăn và thách thức

4. Phương pháp suy luận các mô hình

Kể từ năm 1990, phương pháp suy luận các mô hình hay còn được gọi là phương pháp kết hợp các mô hình đã trở thành một chủ đề nóng trong cộng đồng nghiên cứu. Nhiều tác giả, nhà nghiên cứu từ nhiều lĩnh vực như học máy, nhận dạng mẫu, khai phá dữ liệu, mạng thần kinh và thống kê đã khám phá và khai thác phương pháp này trên nhiều khía cạnh khác nhau.

Trái ngược với các hướng tiếp cận học thông thường chỉ sử dụng một phương pháp từ dữ liệu huấn luyện, phương pháp suy luận các mô hình sẽ sử dụng một tập các phương pháp và kết hợp chúng lại để giải quyết cùng một vấn đề. Do vậy, phương pháp suy luận các mô hình còn được gọi là hệ thống phân loại đa phương pháp. Bên dưới là hình ảnh về một kiến trúc về kết hợp mô hình.



Hình 1. Một kiến trúc suy luận chung

Theo tác giả Zhou và cộng sự, một phương pháp kết hợp sẽ chứa nhiều mô hình được gọi là các mô hình cơ sở. Các mô hình cơ sở được xây dựng từ tập dữ liệu huấn luyện bởi một giải thuật học cơ sở như cây quyết định, mạng xoắn hay một giải thuật học nào đó. Phần lớn, các phương pháp kết hợp sẽ sử dụng duy nhất một giải thuật học để xây dựng các mô hình cơ sở đồng nhất. tức là các mô hình có kiểu giống nhau sẽ được sử dụng để xây dựng các phương pháp kết hợp đồng nhất. Tuy nhiên, có một số phương pháp sử dụng nhiều giải thuật học (mô hình có kiểu khác nhau) để xây dựng các mô hình không đồng nhất. Có ba chủ đề đóng góp nên sự bao phủ của phương pháp suy luận các mô hình. Đó là: Kết hợp các bộ phân lớp, kết hợp các mô hình yếu và trộn các mô hình mạnh.

Kết hợp các bộ phân lớp hầu như phổ biến trong cộng đồng nhận diện mẫu. Trong chủ đề này, các nhà nghiên cứu thường làm việc trên các bộ phân lớp mạnh, sau đó cố gắng tạo ra các luật kết hợp để có được các bộ phân lớp kết hợp mạnh hơn. Bởi vậy, trong nhiệm vụ trước hết để thực hiện chủ đề này là cần phải hiểu sâu về các thiết kế và cách sử dụng các luật kết hợp.

Kết hợp các mô hình yếu hầu hết được áp dụng trong cộng đồng học máy. Trong chủ đề này, các nhà nghiên cứu sẽ làm việc với các mô hình yếu và cố gắng thiết kế các thuật toán để tăng hiệu suất từ yếu lên mạnh, bởi vậy đã dẫn tới sự ra đời của nhiều phương pháp lai các mô hình như Boosting, Bagging, ... và các lý thuyết cần phải được hiểu tại sao, như thế nào mà các mô hình yếu có thể được tăng lên như vậy.

Chủ đề thứ ba là pha trộn các mô hình mạnh, chủ đề này thường được áp dụng trong cộng đồng các mạng xoắn. Pha trộn các mô hình mạnh, các nhà nghiên cứu cần phải xem xét

tới chiến lược chia để trị, cố gắng học một hỗn hợp các mô hình tham số và sử dụng các luật kết hợp để tìm ra một giải pháp tổng thể.

5. Tư tưởng đề xuất mô hình

Bài toán phát hiện ý định mua hàng của người dùng sử dụng phương pháp kết hợp thực chất là việc giải quyết bài toán phân lớp sử dụng kết hợp nhiều mô hình phân lớp. Phân lớp (hay phân loại) là một tiến trình xử lý nhằm xếp các mẫu dữ liệu hay các đối tượng vào một trong các lớp đã được định nghĩa trước.

Các thuật toán phân lớp tiêu biểu bao gồm như mạng xoắn (Neural Network), cây quyết định (Decision Tree), suy luận gộp (Joint Inference), mạng Bayesian, máy vector hỗ trợ (Support Vector Machine - SVM), K người láng giềng gần nhất (K-NN), ... Trong công trình này, tôi đề xuất một giải pháp nhằm tập trung vào cải thiện chất lượng bài toán dựa trên dữ liệu Facebook. Tư tưởng xây dựng mô hình đề xuất xoay quanh những khía cạnh sau:

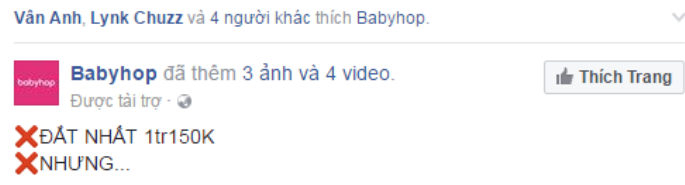
- **Về mặt dữ liệu:** Công trình đề xuất sử dụng nguồn dữ liệu là các bài viết, các bình luận của người dùng trong các trang bán hàng của mạng xã hội Facebook bởi nguyên nhân:
 - *Thứ nhất là dữ liệu lớn:* Theo thống kê, Facebook được coi là một trong các mạng xã hội được sử dụng phổ biến nhất trên thế giới. Thông thường, người dùng chia sẻ trạng thái, cảm xúc và tán gẫu với mọi người qua các bài đăng, bình luận của họ. Nhưng trong khoảng 5 năm trở lại đây kể từ khi kinh doanh trên Facebook xuất hiện, người dùng truy cập facebook không chỉ để nói chuyện, chia sẻ cảm xúc mà họ còn sử dụng mạng xã hội này như một kênh mua sắm. Họ có thể đăng tải, tìm kiếm, mua bán, trao đổi hàng hóa, sản phẩm hay dịch vụ qua trang cá nhân, nhóm (group) hoặc các trang bán hàng (fanpage), do đó thông tin trên mạng xã hội này sẽ nhanh chóng, kịp thời. Đồng thời, hiện tại số người dùng Việt Nam sử dụng mạng xã hội Facebook đã lên tới hơn 37 triệu người dùng (so với cùng thời gian năm 2015 đã tăng lên 7 triệu người dùng), điều này đồng nghĩa với việc dữ liệu trên Facebook là rất lớn và luôn nóng hổi. Tuy nhiên, vấn đề khó khăn gặp phải khi sử dụng dữ liệu trên Facebook đó là việc chuẩn hóa dữ liệu. Cụ thể, với mỗi người dùng họ sẽ có cách sử dụng, diễn tả từ ngữ khác nhau theo phong cách và sở thích của họ. Do vậy, việc gặp phải các bài viết, bình luận gặp vấn đề về từ viết tắt, từ lóng, hashtag, ... là điều dễ thấy. Ngoài ra, một hạn chế khác khi sử dụng dữ liệu trên Facebook đó là việc thông tin bị nhiễu. Bên dưới là một bình luận chưa được chuẩn hóa của người dùng: “*Áo này bao tiền 1 chiếc vậy*”. Trường hợp này sẽ rất khó để nhận diện ra ý định của họ.



Hình 1. Một ví dụ về dữ liệu chưa chuẩn hóa

- *Thứ hai là dữ liệu rất đa dạng và phong phú:* Theo như thống kê, hiện tại Facebook Việt Nam có khoảng hơn 350.000 trang bán hàng có tần suất truy cập thường xuyên và ổn định với nhiều chủng loại mặt hàng đa dạng, có những trang bán hàng có số lượng người theo dõi/yêu thích lên đến hàng triệu người. Trung bình, cứ mỗi bài đăng sẽ có khoảng hơn 50 lượt bình luận bao gồm những bình luận mang thông tin, ý định mua hàng và không có ý định gì. Bởi lý do này mà dữ liệu được thu thập từ những trang bán hàng là nguồn tài nguyên vô cùng phong phú và có ích cho bài toán phát hiện ý định người dùng
- *Trang (Page) mở hơn là Nhóm (Group):* Facebook có 2 loại kênh marketing là Trang - Fanpage và Nhóm - Group. Nhóm được xây dựng cho cộng đồng của các cá nhân có cùng sở thích, cùng quan điểm, cùng đam mê, chia sẻ ý tưởng, chia sẻ ý kiến, chia sẻ lợi ích chung, nơi để thảo luận bàn tán giữa các thành viên Tuy nhiên, nhóm có 2 loại là Nhóm đóng (private group) và Nhóm mở (public group). Nếu Nhóm mở, nó sẽ có vai trò gần giống như một trang fanpage. Nếu là Nhóm đóng, Nhóm sẽ có một người Admin đóng vai trò là chủ của Nhóm, nếu bạn muốn tham gia hay đăng bài viết sẽ phải qua kiểm duyệt của Admin và chỉ ai tham gia vào Nhóm mới có thể xem được bài đăng, các bình luận hay các hoạt động của Nhóm. Một hạn chế của Nhóm đó là thành viên trong Nhóm có thể là yêu thích, đam mê thực sự, tham gia Nhóm một cách chủ động hoặc không quan tâm tới hoạt động Nhóm nhưng vẫn nằm trong Nhóm này bởi họ có thể bị cho vào Nhóm bởi một người khác. Ngược lại, Trang thì mở hơn rất nhiều so với nhóm. Chỉ cần nhớ địa chỉ Trang là bất cứ ai cũng có thể truy cập. Vì Trang là công khai nên được hiển thị cho công cụ tìm kiếm bao gồm các bài viết, hình ảnh của bạn, và video mà Trang chia sẻ. Tất cả mọi người trên Facebook, có thể kết

nối với các trang bằng cách trở thành một fan hâm mộ và sau đó nhận được thông tin cập nhật của họ trong dòng tin tức của chính người đó và tương tác với họ. Một ưu điểm nữa của Trang đó là, khi một người dùng trở thành một fan hâm mộ của 1 Trang, thông tin này sẽ được hiện lên trên dòng tin tức của tất cả bạn bè của người đó, do vậy, bất cứ ai nếu có chung sở thích hay mối quan tâm cũng có thể trở thành fan hâm mộ.



Hình 2. Một ví dụ về tính mở của Trang.

Một vấn đề khác có thể tạm coi là một khó khăn của tôi trong quá trình trích xuất dữ liệu, đó là trên mạng xã hội Facebook có rất nhiều các Trang với nhiều chủ đề khác nhau: Mua bán, Học tập, Sức Khỏe, Làm đẹp, ... nên để có thể thu thập được các Trang thuộc chủ đề mua bán, tôi phải truy vấn dữ liệu dựa trên tập danh sách các từ khóa là các sản phẩm, mặt hàng liên quan đến chủ đề mà luận văn quan tâm. Ví dụ: xịt khoáng Caudalie, chân váy xòe, thắt lưng, đồng hồ, áo sơ mi, ... Việc xây dựng một tập các danh sách chứa đầy đủ các từ khóa về tất cả các loại hàng trên Trang là hoàn toàn không thể, bởi vậy mà việc thu thập được tất cả các Trang bán hàng để lấy được đa dạng các bình luận phục vụ cho dữ liệu học sẽ dừng lại ở một giới hạn nhất định do vậy, việc dự đoán ý định người dùng có thể sẽ không thể bao phủ hết dữ liệu của tất cả các Trang.

Quá trình phát hiện ý định người dùng trên các trang bán hàng: Để tập trung vào việc phát hiện ý định người dùng mua hàng, từ những dữ liệu đã được thu thập là các bài viết của các trang bán hàng, tôi thực hiện một bước tiền xử lý nhằm lọc ra những bài viết bán hàng. Bởi trong một trang bán hàng sẽ có những bài viết bán hàng và bài viết không bán hàng. Đây là một bài toán phân lớp nhị phân, tôi sử dụng ba mô hình phân lớp như đã giới thiệu nhằm tìm ra được bài viết mong muốn. Từ những bài viết này đó tôi thu lấy những bình luận thực sự liên quan đến chủ đề luận văn, tránh nhiễu dữ liệu. Sau đó, các tài liệu lại một lần nữa được phân lớp để tìm ra được ý định tiềm ẩn bên trong.

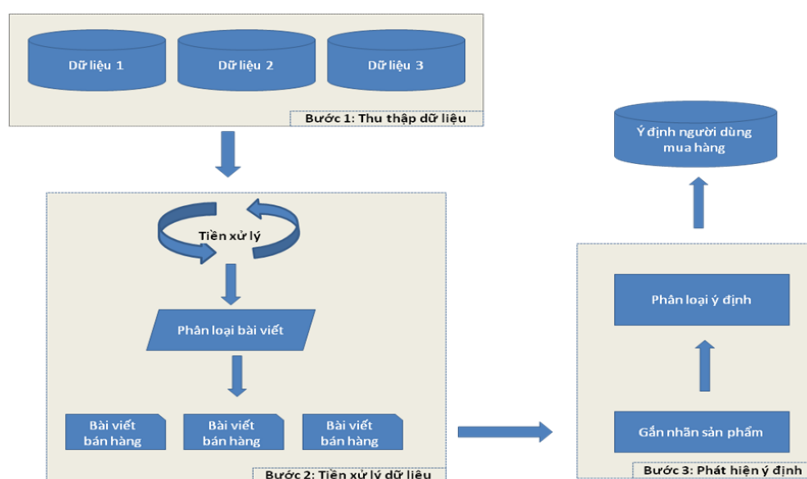
6. Mô hình đề xuất

Trong luận văn của mình, tôi tiến hành kết hợp ba thuật toán tương đương ba mô hình phân lớp tiêu biểu là: Support Vector Machine, K – Nearest Neighbors, Maximum Entropy

Model sau đó sử dụng kỹ thuật Bagging đã được miêu tả trong phần trước để tạo ra bộ phân lớp tốt nhất.

Công trình nghiên cứu này đề xuất mô hình nhằm giải quyết bài toán phát hiện ý định hành vi người dùng trên mạng xã hội Facebook. Mô hình như ảnh minh họa bên dưới có 03 bước chính:

- Bước 1: Thu thập dữ liệu
- Bước 2: Phân loại bài viết
- Bước 3: Phân tích và phát hiện ý định



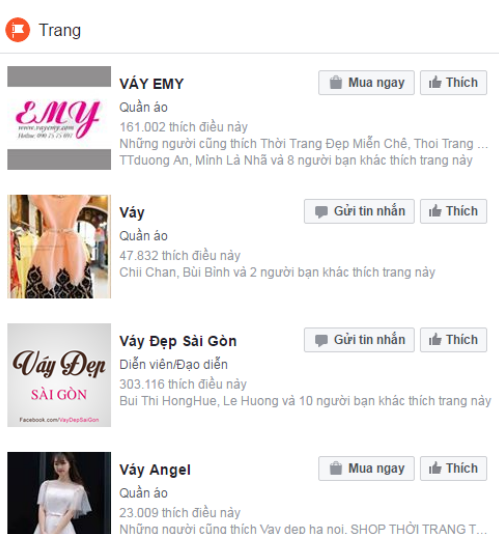
Hình 2. Mô hình đề xuất

Chi tiết về các bước trong mô hình đề xuất sẽ được trình bày chi tiết trong các phần bên dưới.

a. Thu thập dữ liệu

Trong bước này, tôi có hai nhiệm vụ chính:

- **Nhiệm vụ 1**: Thu thập các trang bán hàng.
 - Tôi sử dụng danh sách chứa các từ khóa sản phẩm, dịch vụ, ... để tìm kiếm các Trang bán hàng tương ứng. Cụ thể, tôi truy vấn tên các sản phẩm sử dụng tính năng tìm kiếm của Facebook, chỉ lựa chọn lấy những Trang có nhiều fan hâm mộ và hoạt động ổn định trong khoảng thời gian hiện tại.



Hình 3. Hình ảnh về quá trình thu thập Trang bán hàng

- **Nhiệm vụ 2:** Trong danh sách Trang bán hàng thu được, tôi lựa chọn ngẫu nhiên các Trang bất kỳ và thu thập nội dung các bài đăng, bình luận (comment) của nó. Để thực hiện được nhiệm vụ này, tôi sử dụng công cụ Facebook Graph API

```

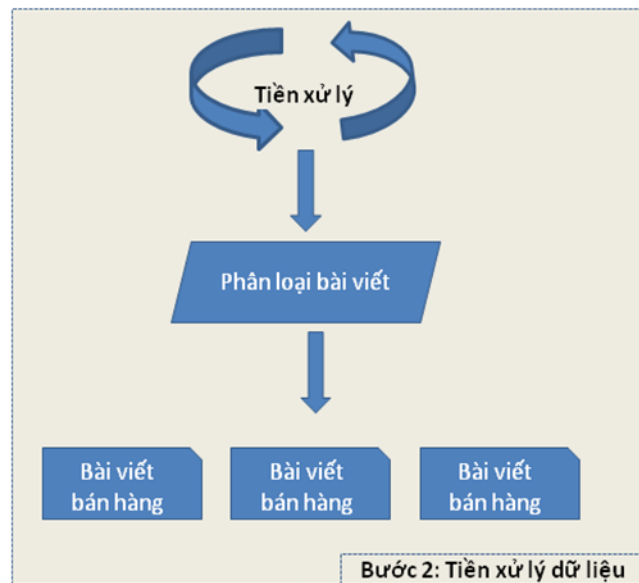
803 {
804   "id": "1855978744648203_1855982961314448",
805   "postId": "1855978744648203",
806   "actorProfileId": "100003132260513",
807   "actorProfileDisplayName": "Hải Yến",
808   "content": "Áo bn b",
809   "commentTime": 1476423773000,
810 },
811 {
812   "id": "1855978744648203_1855985564647521",
813   "postId": "1855978744648203",
814   "actorProfileId": "100003857448765",
815   "actorProfileDisplayName": "Tuyet Tuyet Ngo",
816   "content": "Shop inbox mình xin cái giá!",
817   "commentTime": 1476424482000
818 },
819 {
820   "id": "1855978744648203_1856132781299466",
821   "postId": "1855978744648203",
822   "actorProfileId": "100004025581213",
823   "actorProfileDisplayName": "Thuy Duong",
824   "content": "Bn chị ơi có để lại cho e 1 cái",
825   "commentTime": 1476444802000
826 },

```

Hình 4. Hình ảnh về quá trình thu thập dữ liệu sử dụng Facebook Graph API

b. Tiền xử lý dữ liệu

Nhằm hạn chế mức tối đa dữ liệu nhiễu trong tập dữ liệu của mình, tôi thực hiện phân lớp cho các bài viết thành hai loại: Bài viết bán hàng và bài viết không bán hàng.



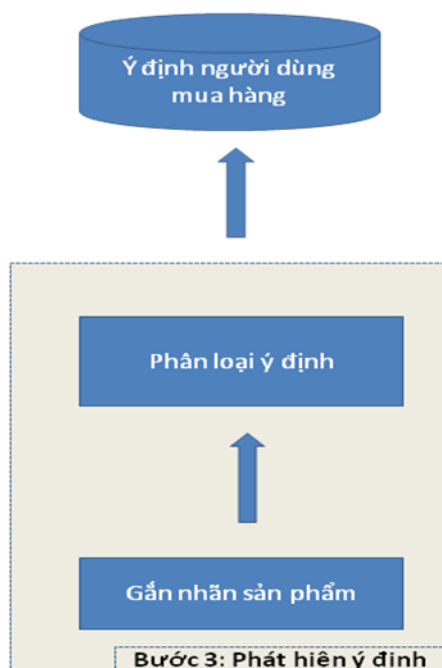
Hình 3. Bước 2: Tiền xử lý dữ liệu

Trong bước này, tôi sử dụng ba mô hình phân lớp đã được đề cập ở các chương trước để thu lấy những bài viết bán hàng cho nhiệm vụ tiếp theo. Bởi, những bài viết không bán hàng thì lượng người dùng bình luận mua hàng rất ít và hầu như là không có. Sau khi thu thập và phân lớp dữ liệu bài đăng, tiến hành gán nhãn sản phẩm cho các bài đăng bán hàng dựa trên cây danh mục sản phẩm đã được xây dựng.

Ví dụ:

- Bài viết 1: “**Sáng nay shop vừa về lô quần legging uniqlo xin đét từng đường kim mũi chỉ. Hàng của uniqlo thì mọi người khỏi phải lo về độ bền rồi nhé. Dùng vài năm có khi mòn mỏng mà không sút chỉ ý ạ. Hihhihi. Giá 230k/bé. Sỉ ib để được giá tốt ạ.**” sẽ được nhận diện là bài viết bán hàng.
- Bài viết 2: “**Hôm nay lạnh quá mọi người ạ. Mời cả nhà cùng Mint nghe nhạc nha. Hôm nay mát trời, chỉ nằm ăn bắp rang bơ và nghe nhạc thôi ý.**” Sẽ được nhận diện là bài viết không bán hàng.

c. Phân tích và phát hiện ý định



Hình 4. Hình ảnh về quá trình phân tích và phát hiện ý định người dùng.

Sau khi đã có các bài viết bán hàng, tôi thu thập các bình luận của các bài viết tương ứng và phát hiện ý định mua hàng của người dùng trong các bình luận này. Trước khi tiến hành quá trình phát hiện ý định, tôi xây dựng định nghĩa về mức độ ý định của người dùng. Ý định người dùng có thể chia làm 4 loại: Chắc chắn mua hàng; đang có nhu cầu, nếu có sẵn hàng sẽ mua ngay; có nhu cầu nhưng chưa thực sự cần mua và không có ý định.

- **Loại 1 – Chắc chắn mua:** Các bình luận sẽ là các câu trần thuật chứa các từ khóa như: lấy, ship cho, bán cho, ib giá, inbox,
- **Loại 2 – Đang có nhu cầu, nếu có hàng sẽ mua luôn:** Loại này thường là các câu chứa từ khóa: vừa không, còn không, giá giờ bao nhiêu, ...
- **Loại 3 – Có nhu cầu nhưng chưa thực sự cần:** Các bình luận của loại ý định này sẽ có thể là: chậm, hóng giá, đặt gạch, x tạm, đẹp quá, made in, ...
- **Loại 4 – không có ý định:** Bình luận sẽ chứa từ khóa: đẹp nhưng đắt, cũng bình thường, tạm, xấu, ...

7. Thực nghiệm

a. Dữ liệu

Như đã miêu tả trong phần mô hình đề xuất, dữ liệu được thu thập về là các trang bán hàng chứa các bài viết bán hàng hoặc không bán hàng. Trong phần này, tôi thực hiện xây dựng tập

dữ liệu huấn luyện cho mục đích phân loại các bài viết. Mục đích của việc này nhằm nâng cao chất lượng các bình luận chứa các ý định trong các bài viết.

Đầu tiên, tôi thực hiện xây dựng tập danh sách các sản phẩm với mục tiêu càng đa dạng càng tốt. Dữ liệu này sau đó sẽ được sử dụng để truy vấn lấy các Trang bán hàng. Những từ khóa về số lượng sản phẩm bao gồm 814 từ về các chủ đề: nội thất gia đình, thời trang may mặc, mỹ phẩm, thiết bị điện tử, thực phẩm. Để thu thập danh sách các Trang bán hàng, tôi truy cập miền trang mạng xã hội Facebook: <https://www.facebook.com> và sử dụng tính năng tìm kiếm nâng cao để tìm ra các Trang bán hàng từ danh sách kết quả trả về. Kết quả thu được 350.000 Trang bán hàng.

Tiếp theo, tập trung vào định nghĩa các bài viết bán hàng là các bài viết chứa tên các sản phẩm do vậy mà từ những tên sản phẩm đã thu thập, tôi thực hiện xây dựng cây danh mục sản phẩm nhằm phát hiện ra nhiều hơn các bài viết bán hàng. Bên dưới là số liệu thống kê dữ liệu bài viết theo hai lớp được biểu diễn trong bảng bên dưới:

Bảng 1. Bảng thống kê số lượng dữ liệu bài viết phân lớp

Số lượng sản phẩm	Bài viết bán hàng	Bài viết không bán hàng	Tổng
814	9.264	3.924	13.188

Như đã nhắc tới trong phần mô tả các bước của mô hình đề xuất, tôi sẽ sử dụng tập dữ liệu chứa các trường hợp có thể xảy ra trong các bình luận cũng như danh sách các mức độ mua sắm tương ứng.

Ví dụ:

- Theo mức độ chắc chắn chứa ý định mua hàng của người dùng: “**Cho mình** hai hộp chè vằng **tới địa chỉ 112 Xuân Thủy. SĐT** mình là: 01651651656”
- Theo mức độ đang có nhu cầu, có ý định mua nếu có sẵn hàng: “ **Thích quá mà hết hàng**, shop ơi, khi nào **hàng lại về nữa** ạ? **Giữ cho mình** một chiếc màu đen size xs nhé.”
- Theo mức độ có nhu cầu nhưng chưa cần mua ngay: “**Chăm**, chờ khi nào có lương sẽ qua mức.”

Sau khi áp dụng mức độ mua sắm của người dùng, tôi thống kê dữ liệu ý định trong ý kiến theo 2 lớp như sau :

Bảng 2. Bảng thống kê số lượng dữ liệu ý định trong bình luận

Bình luận có ý định	Bình luận không có ý định	Tổng
23.181	11.788	34.969

b. Thực nghiệm và kết quả

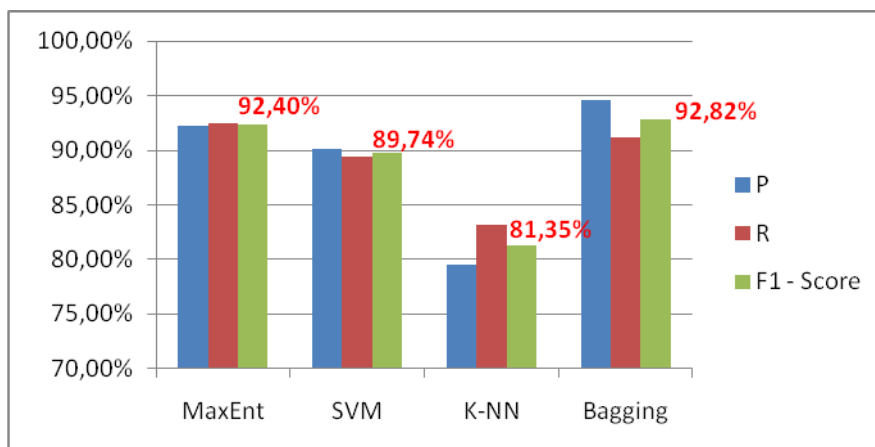
Trong hệ thống của mình, tôi tiến hành phân loại các bài viết bán hàng và phân loại ý định trong ý kiến của cá nhân. Các bài viết sẽ được tiến hành phân loại bài viết có phải có nội dung bán hàng hay không, nếu một bài viết có liên quan đến bán hàng sẽ thực hiện lấy các ý kiến bình luận trong bài viết đó tiến hành phân loại xem có chứa ý định mua hàng hay không.

Các thực nghiệm đánh giá mô hình sử dụng phương pháp kiểm thử chéo 10 folds (10-folds cross validation) nghĩa là chia làm 10 phần bằng nhau, lần lượt huấn luyện 9 phần để đánh giá 1 phần sau đó sử dụng độ đo đã được nêu trước đó. Để huấn luyện mô hình tôi sử dụng kết hợp ba mô hình Support Vector Machine, K – Nearest Neighbors và Maximum Entropy Model và sử dụng phương pháp Voting (bỏ phiếu) để lựa chọn kết quả đúng nhất. Ưu điểm của phương pháp này là việc kết hợp nhiều mô hình phân lớp sẽ đưa kết quả chính xác hơn khi sử dụng một mô hình.

- Kết quả phân lớp bài viết bán hàng và không bán hàng:

Bảng 3. Kết quả phân lớp bài viết bán hàng/không bán hàng

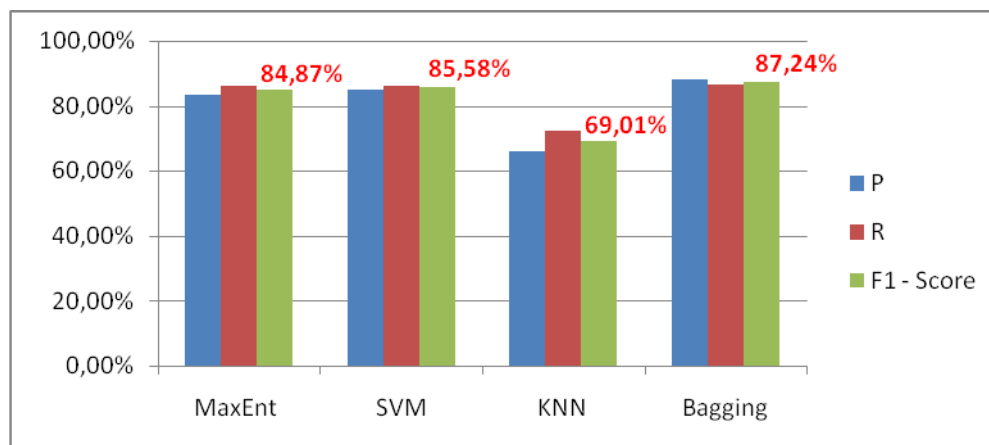
	Độ chính xác (P)	Độ hồi tưởng (R)	F ₁ – score
MaxEnt	92,29%	92,51%	92,40%
SVM	90,12%	89,36%	89,74%
K-NN	79,58%	83,21%	81,35%
Bagging	94,58%	91,13%	92,82%



- Kết quả phân lớp ý định: Có ý định và không có ý định, spam

Bảng 4. Bảng kết quả phân lớp ý định

	Độ chính xác (P)	Độ hồi tưởng (R)	F ₁ - score
MaxEnt	83,57%	86,22%	84,87%
SVM	84,87%	86,31%	85,58%
K-NN	65,93%	72,40%	69,01%
Bagging	88,12%	86,37%	87,24%



KẾT LUẬN

Trong công trình này, luận văn tiến hành nghiên cứu phương pháp nhằm cải thiện độ chính xác cho bài toán phân lớp dữ liệu, cụ thể là cải thiện độ chính xác cho bài toán nhận diện, phát hiện ý định người dùng mua hàng trên mạng xã hội facebook qua các bình luận của họ. Bài toán này được xác định là một bài toán có độ phức tạp cao và là nền tảng của nhiều nghiên cứu thực tế. Phương pháp giải quyết của luận văn tập trung vào việc tăng cường chất lượng nhằm nhận diện được nhiều và chính xác các ý định nằm ẩn trong các bình luận của tập dữ liệu đầu vào.

Dựa vào các nghiên cứu về phương pháp suy luận các mô hình (Ensemble Methods) bằng việc kết hợp các mô hình phân lớp quen thuộc Support Vector Machine, k – Nearest Neighbors và Maximum Entropy Model cùng với miền dữ liệu phong phú và rộng lớn Facebook, luận văn đã đưa ra một mô hình để giải quyết cho bài toán đề ra. Quá trình thực nghiệm đạt được kết quả khả quan, cho thấy tính đúng đắn của việc lựa chọn cũng như kết hợp các phương pháp, đồng thời hứa hẹn nhiều tiềm năng phát triển hoàn thiện.