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 SỸ NGÀNH CÔNG NGHỆ THÔNG TIN

HÀ NỘI– 2016

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

Ngành: Công nghệ thông tin
Chuyên ngành: Quản lý hệ thống thông tin

LUẬN VĂN THẠC SỸ NGÀNH CÔNG NGHỆ THÔNG TIN

Cán bộ hướng dẫn: PGS. TS. Hà Quang Thụy

HÀ NỘI – 2016

VIETNAM NATIONAL UNIVERSITY, HANOI
UNIVERSITY OF ENGINEERING AND TECHNOLOGY

Doan Thi Huyen Trang

USER CONSUMPTION INTENT IDENTIFICATION FROM SOCIAL NETWORK USING ENSEMBLE METHODS

Major: Information Technology

Supervisor: Assoc. Prof. Ha Quang Thuy

HA NOI –2016

Lời cảm ơn

Trước tiên, em xin bày tỏ lời cảm ơn sâu sắc tới Phó giáo sư Tiến sĩ Hà Quang Thụy người đã tận tình chỉ bảo, hướng dẫn em trong quá trình tìm hiểu, nghiên cứu để hoàn thành luận văn tốt nghiệp của mình.

Đặc biệt, em xin gửi lời cảm ơn chân thành nhất tới Thạc sĩ Trần Mai Vũ - người đã tận tình hỗ trợ về kiến thức chuyên môn, giúp đỡ em rất nhiều để hoàn thành luận văn.

Đồng thời, xin cảm ơn các thầy, các anh chị và các bạn trong Phòng Thí nghiệm DS&KTLab và Đề tài QG.15.22 đã chia sẻ những kinh nghiệm, kiến thức quý báu cho em trong quá trình nghiên cứu.

Cuối cùng, em muốn gửi lời cảm ơn tới gia đình, những người thân yêu luôn bên cạnh, động viên, giúp đỡ em trong suốt quá trình học tập và trong thời gian thực hiện luận văn thạc sỹ.

Xin chân thành cảm ơn!

Hà Nội, ngày 1 tháng 11 năm 2016

Tác giả

Doãn Thị Huyền Trang

Tóm tắt

Tóm tắt:

Vài năm trở lại đây, nhu cầu sử dụng mạng xã hội của người dùng không ngừng tăng. Con người sử dụng mạng xã hội không chỉ để giải trí như: cập nhật trạng thái, kết bạn, tán gẫu, nói chuyện mà họ còn dùng mạng xã hội như một công cụ tìm kiếm thông tin hay sản phẩm, dịch vụ và là nơi mua bán, trao đổi hàng hóa. Đoán được nhu cầu này của đông đảo người dùng, bài toán phát hiện ý định mua hàng của người dùng trên mạng xã hội ra đời nhằm tìm ra các ý định, mong muốn mua một sản phẩm được người dùng thể hiện trong các bài đăng, các bình luận trên mạng xã hội Facebook để từ đó làm kết quả đầu vào cho nhiều bài toán quan trọng, mang lại nhiều giá trị không nhỏ cho cộng đồng nghiên cứu như: hệ tư vấn người dùng – giúp hỗ trợ người dùng tìm kiếm hàng hóa, sản phẩm dịch vụ đúng địa chỉ với thời gian nhanh nhất, bài toán dự đoán sở thích người dùng qua những hành vi của họ và nhiều bài toán có ý nghĩa khác nữa. Bài toán hiện đã và đang nhận được sự quan tâm đặc biệt trong nhiều hướng nghiên cứu mới bởi nó có sức ảnh hưởng không nhỏ và là nguồn tài nguyên quan trọng cho các bên liên quan như các công ty, tổ chức, chính phủ, … . Mặc dù có tiềm năng lớn cho các ứng dụng nhưng việc xác định các ý định rõ ràng của người dùng thực sự là một bài toán, một hướng nghiên cứu khó trong xử lý ngôn ngữ tự nhiên.

Luận văn với đề tài "***Trích xuất ý định người dùng mua hàng trên mạng xã hội sử dụng phương pháp suy luận các mô hình***" nghiên cứu nội dung, các thuộc tính, các thuật toán nhằm giải quyết bài toán phân lớp. Luận văn thực hiện tiến hành thực nghiệm trên bộ dữ liệu là các bài đăng, các bình luận trên Facebook, sử dụng phương pháp lai ghép các mô hình phân lớp: Support Vector Machine (SVM), K – Nearest Neighbors (KNN) và Maximum Entropy (Maxent) mang lại kết quả tốt hơn so với việc chỉ sử dụng một mô hình phân lớp. Kết quả trả về với độ chính xác P là 88,12%, độ hồi tưởng R là 86,37% và độ đo F1 là **87,24%**

Từ khóa: ý định, phương pháp lai ghép mô hình, Support Vector Machine, K- Nearest Neighbors, Maximum Entropy,…

Abstract

Abstract:

Social media platforms are often used by people to express their needs and desires. Such data offer great opportunities to identify users' consumption intention from user-generated contents, so that better tailored products or services can be recommended. However, there have been few efforts on mining commercial intents from social media contents. In this thesis, I investigate the use of social media data to identify consumption intentions for individuals. I use ensemble methods based on three classification models: Support Vector Machine, K- Nearest Neighbors, Maximum Entropy Model for identifying whether the user has a consumption intention on your comment.

Experiment results have show that the proposed method is quite well with Precision: 88,12%, Recall: 86,37% and F1- score: **87,24%**.

Keywords: intent, ensemble methods, Support Vector Machine, K- Nearest Neighbors, Maximum Entropy

Lời cam đoan

Tôi xin cam đoan rằng đây là công trình nghiên cứu của mình, có sự giúp đỡ từ giáo viên hướng dẫn là Phó giáo sư, Tiến sỹ Hà Quang Thụy và Thạc sĩ Trần Mai Vũ.

Các nội dung nghiên cứu và kết quả trong đề tài này là trung thực, không sao chép từ bất cứ nguồn nào có sẵn. Tất cả những tham khảo từ các nghiên cứu liên quan đều được nêu nguồn gốc một cách rõ ràng từ danh mục tài liệu tham khảo trong khóa luận. Trong luận văn, không có việc sao chép tài liệu, công trình nghiên cứu của người khác mà không chỉ rõ về tài liệu tham khảo.

Nếu phát hiện có bất kỳ sự gian lận nào, tôi xin hoàn toàn chịu trách nhiệm trước hội đồng, cũng như kết quả luận văn tốt nghiệp của mình.

Tác giả

DoãnThị Huyền Trang

Mục lục

Danh sách bảng

Danh sách hình vẽ

Bảng các ký hiệu

Từ viết tắt	Thuật ngữ
SVM	Support Vector Machine
KNN	K – Nearest Neighbors
MEM	Maximum Entropy Model
SN	Social Network
ISP	Internet Service Provider
IG	Information Gain

Mở đầu

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.

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.

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 hay 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 nhiều nghiên cứu 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, ….

Dựa trên những hướng tiếp cận đã đề cập ở trên, trong luận văn này, tôi tiến hành áp dụng phương pháp lai ghép 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 bình luận của họ trên các trang bán hàng (fanpage).

Sau khi thu được kết quả của ba mô hình phân lớp Support Vector Machine (SVM), K – Nearest Neighbors (KNN) và Maximum Entropy (Maxent), luận văn sử dụng phương pháp bình chọn theo biếu bầu - Voting để lựa chọn được kết quả phân lớp tốt nhất. Thực nghiệm trả về với độ đo chính xác là **88,12%,** độ hồi tưởng là **86,37%** và độ đo F_1 là **87,24%** phần nào chứng minh được độ hiệu quả của phương pháp áp dụng.

Nội dung của luận văn gồm 03 chương:

Chương 1: Giới thiệu chung mô tả tầm quan trọng của ý định mua hàng và khái quát bài toán. Sau đó nêu định nghĩa về ý định mua hàng của người dùng, các loại ý định người dùng và cuối cùng là hướng tiếp cận nhằm giải quyết bài toán đề ra.

Chương 2: **Phương pháp lai ghép các mô hình** trình bày về phương pháp lai ghép các mô hình và kỹ thuật Bagging nhằm cải thiện chất lượng bài toán phân lớp. Đây cũng chính là phương pháp sẽ được áp dụng cho bài toán đã đề xuất trong chương một.

Chương 3: **Mô hình đề xuất**, **thực nghiệm, kết quả và đánh giá** nhằm nêu rõ và chi tiết các bước trong quá trình giải quyết bài toán. Trong chương này cũng sẽ trình bày quá trình thực hiện và hoàn thành thực nghiệm, đưa ra một số đánh giá, nhận xét các kết quả thu được.

Phần kết luận: Tóm lược những kết quả đạt được của luận văn. Đồng thời đưa ra những hạn chế, những điểm cần khắc phục và đưa ra định hướng nghiên cứu trong thời gian sắp tới.

Chương 1. Giới thiệu chung

Những năm qua, sự phát triển không ngừng của mạng Internet và sự ra đời của các thiết bị kết nối thông minh như máy tính bảng, điện thoại thông minh đã kéo theo sự phát triển của các phương tiện truyền thông xã hội cũng như các trang mạng xã hội như Facebook, Twitter, Google+, … Tuy nhiên, điển hình nhất là Facebook. Tính trên toàn thế giới, Việt Nam là quốc gia mà Facebook có thị phần tăng trưởng nhanh nhất, với tốc độ 146% trong 6 tháng (từ tháng 5 - 10/2012), trung bình cứ 3 giây thì Facebook có 1 người dùng Việt Nam mới (Socialbakers & SocialTimes.Me 2013). Theo thống kê[1] 2015, ở Việt Nam có khoảng 30 triệu tài khoản Facebook và đến tháng 7 năm 2016 thì con số này đã tăng lên tới 37 triệu. Trung bình, người Việt dành khoảng 2,5 tiếng mỗi ngày trên Facebook cho việc trò chuyện với bạn bè và theo dõi thương hiệu sản phẩm. Bảng bên dưới là một vài con số thống kê về số lượng người sử dụng các trang mạng xã hội.

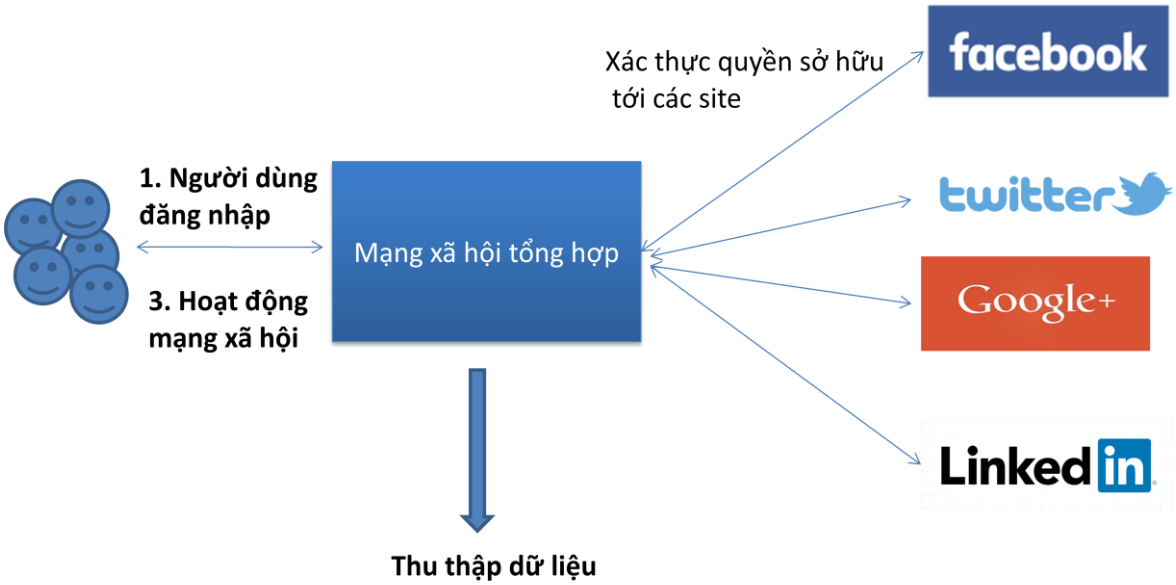
Bảng 1. Thống kê về số người sử dụng các kênh mạng xã hội

Kênh mạng xã hội	Số người sử dụng
Facebook	1.01 tỷ người (Tháng 10/2012)
Twitter	500 triệu người (Tháng 4/2012)
Google+	400 triệu người (Tháng 9/2012)
Linkedln	175 triệu người (Tháng 6/2012)

Trong không gian này, người dùng có xu hướng thể hiện bản thân và sẵn sàng chia sẻ các hoạt động, cảm xúc, suy nghĩ, mong muốn của mình bởi vậy việc chia sẻ một bài viết, hay gửi một vài bình luận thông qua các trang mạng xã hội trở thành một phần tất yếu trong cuộc sống hàng ngày của rất nhiều người. Kết quả là, những bài đăng, những bình luận của người dùng trên diễn đàn và mạng xã hội có thể phản ánh rất nhiều quan điểm, ý kiến và cả ý định của họ. Các bài viết trên đó được xem như là một nguồn tài nguyên quan trọng cho việc phân tích ý định người dùng [6] (Hollerit, Krollm và Strohmaier 2013; Zhao cùng cộng sự 2014). Ví dụ, một bài viết ý định trên Facebook: "*Ib mình gửi địa chỉ ship hàng nhé*" hay một bình luận "*Áo*

[1]http://vtv.vn/thi-truong/viet-nam-co-hon-30-trieu-nguoi-dung-facebook-2015061710512952.htm

pull này có size XS không bạn? Ship cho mình 1 chiếc tới địa chỉ số 4 Hồ Tùng Mậu, sđt: 0973999119 sau 5h chiều nhé" chỉ ra một cách rõ ràng về ý định về việc mua một cái gì đó người dùng. Hình 1 là sơ đồ về việc thu thập dữ liệu của người dùng trên mạng xã hội.



Hình 1. Thu thập dữ liệu thông qua mạng xã hội tổng hợp.

Nhận thức được xu hướng quan trọng này, đã có khá nhiều công trình nghiên cứu tập trung vào việc dự đoán, kết hợp hay phân loại ý định người dùng từ những hoạt động trực tuyến của họ như xác định ý định người dùng từ các bài viết trên diễn đàn trực tuyến 44[5], tương tác giữa các thiết bị như máy tính, điện thoại khi tìm kiếm web, ... Hầu hết, các nghiên cứu đều cố gắng đoán hoặc xác định ý định ẩn sau các truy vấn tìm kiếm của người dùng và hành vi của họ trên trình duyệt. Việc hiểu ý định tìm kiếm sẽ giúp cải thiện chất lượng tìm kiếm của người dùng một cách đáng kể. Tuy nhiên, bài toán trích xuất ý định của người dùng vẫn còn nhiều thách thức. Các bài đăng của người dùng rất nhiễu và thường chứa các từ lóng, lỗi chính tả, cảm xúc và hashtags, ... Ngoài ra, sẽ rất tốn thời gian để tạo ra dữ liệu được gắn nhãn nếu áp dụng hướng tiếp cận giám sát.

Phần đầu chương 1 của luận văn khái quát về tầm quan trọng của bài toán khai thác ý định người dùng, nêu một vài định nghĩa về ý định đã được các nhà nghiên cứu phát biểu và cuối cùng là một vài mô tả về một vài công trình liên quan.

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

Người dùng mạng Facebook đã thừa nhận họ tin tưởng trang mạng xã hội này và sẵn sàng chia sẻ nhiều thông tin trên Facefook hơn là các trang khác như MySpace hay Twitter (Dwyer cùng cộng sự., 2007) [46]. Bởi vậy, 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. Năm 2012, Nelson-Field và các đồng nghiệp [47] 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) [48] 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. Cũng bởi nhu cầu đó, khoảng năm năm trở lại đây, Facebook đã dần 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 cùng 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 hạnh 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, hiểu 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, ...

1.2. 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) [4]: "***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 cùng cộng sự [2][3] 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. Ví dụ, một người dùng viết, "***Tôi đang tìm kiếm một thương hiệu xe mới để thay thế cũ Ford Focus của tôi***" - "***I am looking for a brand new car to replace my old Ford Focus***". Đây là một ví dụ cho một ý định rõ ràng. Theo Jinpeng Wang cùng cộng sự [1] định nghĩa trong mạng xã hội twitter thì một bài tweet sẽ là 1 ý định tweet nếu (1) nó chứa ít nhất 1 động từ và (2) một mô tả 1 cách rõ ràng ý định của người dùng để thực thi 1 hành động (3) trong 1 cách nào đó dễ nhận biết. Ví dụ: Tweet: "***Nếu được điểm A trong kỳ thi này, tôi muốn mua 1 xbox, xin hãy ban phước***" - "***I want to buy an xbox, if get A in this***

examination. Bless me!!!" là một ý định tweet và có đủ ba điều kiện trong định nghĩa.

Bảng 2 là một vài các bài đăng của người dùng trên diễn đàn trực tuyến và mạng xã hội trong đó có bài đăng chứa ý định rõ ràng và bài đăng không chứa ý định hoặc chứa ý định không rõ ràng được đề xuất bởi nhóm tác giả Le cùng cộng sự [5]:

Bảng 2. Một vài ví dụ về các bài đăng chứa/không chứa ý định

Bài đăng trên diễn đàn trực tuyến/mạng xã hội	**Loại ý định**
Chào các bác, vợ chồng em mới chuyển qua nhà mới, tính đầu tư mua cái máy lọc nước để uống và sử dụng nấu nướng luôn cho an toàn vì thỉnh thoảng mở nước em thấy có cặn cặn. Trước ở cùng cụ thì toàn đun nước cho cả nhà uống nên ko quan tâm tới mấy loại máy lọc này. Sớt google thì thấy có 2 dòng RO và Nano, em đọc thông tin mà rối tung tù mù, chả biết nên chọn của loại nào. Các bác tư vấn giúp em phát nhé	Ý định rõ ràng
Tình hình là con e71 của mình nếu sạc bằng cục sạc AC 3E (cục sạc đi kèm với con 1200 dòng ra thấp - khoảng 350mah) thì chỉ mất khoảng 3 tiếng 20p. Trong khi con của thằng bạn thì mất tới 4,5 tiếng gì đấy mới đầy. Các bạn cho mình hỏi là nếu sạc bằng cục sạc AC 3E này thì các bạn sạc mất bao nhiêu tiếng? Mình nghi ngờ cục pin của mình có vấn đề rồi hix thấy dùng bình thường nghịch ngợm chút ít thì chỉ đc 3 ngày	Không có ý định/Ý định ẩn
Em sinh viên đang định thay máy e thấy thích con s4 e330 k biết giá con đó h khoảng bn? K biết ở hà nội thì có chỗ nào bán k?	Ý định rõ ràng
Bạn ơi, gửi cho mình 2 thỏi son 11, 12 về địa chỉ 111 Sơn Tây nhé.	Ý định rõ ràng
Tôi yêu những đóa cẩm tú cầu, Sắc mầu linh hoạt biết dường nào. Lúc trắng, lúc xanh, lúc hồng nhạt. Thanh thản trong sương không biết sầu.	Không có ý định/Ý định ẩn

Trong công trình của mình, để có thể phân biệt các loại bài viết ý định khác nhau, nhóm tác giả Jinpeng Wang [1] 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):

- **Đồ ăn và Nước uống (Food and Drink)**: Các tác giả bài viết lên kế hoạch để có một số đồ ăn hoặc nước uống.
- **Du lịch (Travel)**: Các tác giả bài viết hứng thú/quan tâm với các buổi thăm các địa điểm đặc biệt.
- **Sức khỏe và Giáo dục (Career and Education)**: Các tác giả bài viết muốn có 1 công việc, 1 chứng chỉ/bằng cấp hoặc tự thực hiện 1 điều gì đó. Loại này xuất hiện trong Twellow[5] cái mà tổ chức người dùng twitter vào 1 taxonomy.
- **Hàng hóa và Dịch vụ (Goods and Services):** Các tác giả quan tâm hoặc muốn có 1 số loại hàng hóa không phải thực phẩm, hay nước uống (Ví dụ: ô tô) hoặc các dịch vụ (Ví dụ: cắt tóc). Loại này tương ứng với sự kết hợp của 4 loại trong Groupon, cụ thể là Beauty and Spa, Health and Fitness, Automotive, Shopping and Apparel. Chúng được kết hợp bởi chúng đều thuộc về Goods and Services và mỗi loại này đều chỉ là 1 tỉ lệ rất nhỏ trên mạng xã hội.
- **Sự kiện và Hoạt động (Event and Activities)**: Các tác giả muốn tham gia một số hoạt động không thuộc các loại nói trên (Ví dụ: hòa nhạc). Loại này tương ứng với loại Event và Activities của Groupon.
- **Khác (Trifle)**: Loại này của ý định các bài viết nói về thói quen hàng ngày hoặc một số tâm trạng lặt vặt (Java cùng cộng sự 2007).

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.

Trong phần tiếp theo, luận văn sẽ đi tới khái quát bài toán phát hiện ý định người dùng nhằm mô tả rõ hơn về các nhiệm vụ để đi tới giải quyết bài toán.

1.3. Phát biểu bài toán phát hiện ý định người dùng

Mọi người thường đăng những nhu cầu và mong muốn trên phương tiện truyền thông xã hội. Những dữ liệu này cung cấp cơ hội lớn để xác định ý định của người sử dụng từ nội dung mà người dùng tạo ra, để đo đếm sản phẩm hoặc dịch vụ một cách tốt hơn. Phát hiện và khai thác ý định của người sử dụng trên mạng xã hội có lợi ích rất lớn đến nhà cung cấp sản phẩm/dịch vụ, chẳng hạn như công ty đại chúng, chính phủ hoặc các tổ chức phi lợi nhuận, để giúp họ hiểu rõ hơn về khách hàng tiềm năng của họ và do đó cải thiện dịch vụ của họ hoặc chiến lược quảng cáo cho công chúng nói chung.

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 Facebook

Đầu ra:

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

Ví dụ:

Đầu vào: *Shop ơi, ship cho tớ 1 set Serum Caudalie dòng trị nám tới địa chỉ 202 Xuân Thủy nhé. SĐT: 0972001287*

Đầu ra: Có ý định

Hình 2. Ví dụ về một bình luận có ý định.

1.4. Khó khăn và thách thức

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 toántìm và nhận diện được chính xác ý định của người dùng trên mạng xã hội gặp khá nhiều khó khăn và thách thức. Cụ thể:

- ***Bài viết chứa từ viết tắt, teen code, tối nghĩa***. Với lối diễn đạt vô cùng đơn giản, người dùng luôn thể hiện và diễn tả cảm xúc, mong muốn bằng từ ngữ theo sở thích, thói quen của mình, những trường hợp như này sẽ gặp thường xuyên trên mạng xã hội. Ví dụ: "*Tôi **ms** một chiếc túi*" hoặc "***Mih mún** lấy 1 **th0j s0n BJ***"
- ***Bài viết chứa các từ bị sai chính tả, lẫn cả tiếng nước ngoài***. Ví dụ: "*Anh em nào đã **xử** dụng dịch vụ truyển phát nhanh cho mình **sin reiew** với? Mình muốn truyển một món đồ từ Hồ Tùng Mậu đi Long Biên. Cảm ơn*" hoặc "*Cho em **nấy** một cái với ạ."*
- ***Bài viết chứa tiếng lóng, từ địa phương***. Trường hợp xảy ra với các đối tượng là giới trẻ và với từng địa phương. Ứng với mỗi vùng miền hay với mỗi lứa tuổi, họ có cách sử dụng ngôn từ khác nhau. Ví dụ: "*Mày lại muốn ăn **gạch/hành** phải không*?". Ở đây không có nghĩa là người dùng có nhu cầu ăn uống "gạch" hay "hành" mà là cách diễn đạt, cách thể hiện xô bồ mang ý nghĩa "mày lại muốn ăn đánh phải không?" hoặc "***Mần** gì ăn **bi** giờ*? *Hay để tau xuống **gò** mua gì về nhá*?" trong câu này "Mần" sẽ mang nghĩa "làm", "tau" mang nghĩa "tao", "gò" có nghĩa là "cái chợ". Hay một ví dụ khác: "*Em muốn mua 1 miếng **kiếng** để làm khung cửa sổ, shop cắt cho em theo khổ 1m*0.5m và ship cho e về khu tập thể trường Đại học Công Nghiệp. Em cảm ơn*" thì trong ví dụ này "kiếng" là từ địa phương, được người dùng sử dụng thay từ "kính".
- ***Bài viết chứa nhiều hashtag, dẫn đến khó hiểu***. Ví dụ: *"#muon #an #mycay #huhuhu #themqua*". Thông thường những câu như này hệ thống sẽ khó phát hiện ra ý đồ, mong muốn của người dùng. Bởi thứ nhất: Các từ không được viết tách nhau, dẫn đến không thể tìm được từ loại chính xác. Thứ hai: Các từ viết không dấu dễ gây hiểu nhầm.

- Trong một bài viết hay bình luận, ý định thường thể hiện trong chỉ một hoặc hai câu trong khi rất nhiều câu khác không biểu lộ ý định. Điều này sẽ tạo ra nhiễu cho việc phân loại.
- Từ/cụm từ dùng để diễn tả ý định khá hạn chế so với các loại biểu thức chính quy. Nghĩa là tập các đặc trưng chia sẻ ở các miền/lĩnh vực khác nhau là rất nhỏ.
- Trong các lĩnh vực khác nhau, cách để bày tỏ ý định tương tự nhau thường giống nhau. Điều này có nghĩa rằng chỉ có những đặc trưng tích cực (bài viết có ý định) được chia sẻ giữa các lĩnh vực khác nhau, trong khi các đặc trưng chỉ ra những lớp tiêu cực (bài viết không có ý định) trong các lĩnh vực khác nhau lại rất đa dạng.

1.5. Các hướng tiếp cận giải quyết bài toán

1.5.1. Khai phá ý định người dùng trên trang mạng xã hội Twitter

Tác giả Jinpeng Wang cùng cộng sự [1] đã đề xuất việc nghiên cứu bài toán về **xác định và thống kê các bài đăng trên Twitter của một ai đó thành các loại ý định**.

Jinpeng Wang cùng cộng sự cho đưa ra định nghĩa một bài Tweet sẽ chứa ý định nếu (1) nó chứa ít nhất một động từ và (2) một mô tả một cách rõ ràng ý định của người dùng để thực thi một hành động (3) trong một cách nào đó dễ nhận biết. Nhóm tác giả đề xuất một đồ thị dựa trên hướng tiếp cận bán giám sát để kết hợp các loại ý định cho các bài Tweet và xây dựng tập dữ liệu test bằng phương thức Bootstrap - phương pháp không giám sát hiệu quả cho việc lấy các tweet có ý định. Sau đó, họ xây dựng đồ thị ý định intent-graph để biểu thị mối quan hệ của các tweet với nhau, mối quan hệ giữa các tweet với các từ khóa và mối quan hệ giữa các từ khóa để từ đó xây dựng và giải quyết bài toán kết hợp hay khai phá các loại ý định. Kết quả cho thấy rằng phương pháp được áp dụng có hiệu quả trong việc kết hợp các loại ý định cho các bài viết trên twitter so với các phương pháp khác: SVM-Multi, phương pháp của Hollerit’s, phương pháp của Velikovich và phương pháp của Hassan.

1.5.2. Xác định ý định các bài viết trên các diễn đàn thảo luận

Zhiyuan Chen, Bing Liu cùng cộng sự đã nghiên cứu một vấn đề không những mới lạ mà còn có giá trị lớn, cụ thể là xác định các bài viết thảo luận bày tỏ ý định của người dùng trên các diễn đàn thảo luận trực tuyến [2]. Công trình tập trung vào việc xác định những bài đăng (post) của người dùng với ý định rõ ràng. "Rõ ràng" nghĩa là ý định được nêu rõ ràng trong các văn bản, không cần phải suy luận. Dựa vào đặc trưng của bài toán đưa ra, công trình đưa ra giải pháp sử dụng phương pháp học chuyển dịch bởi đối với một loại ý định đặc biệt chẳng hạn như mua bán, những cách bày tỏ ý định trong các lĩnh vực khác nhau thường rất giống nhau do vậy mà họ có thể xây dựng một bộ phân loại 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.

Tác giả thực hiện giải quyết vấn đề đặt ra như giải một bài toán phân loại 2 lớp lớp tích cực (bài viết chứa ý định) và lớp tiêu cực (bài viết không có ý định). Họ đề xuất một phương pháp học chuyển dịch mới áp dụng chung cho các loại ý định khác nhau là Co-Class. Co-Class làm việc như sau: Trước tiên họ xây dựng một bộ phân loại h bằng cách sử dụng dữ liệu đã được gắn nhãn từ những lĩnh vực hiện có, được gọi là dữ liệu nguồn – source data, và sau đó áp dụng bộ phân loại để phân loại những dữ liệu đích – target data (những dữ liệu chưa có nhãn). Dựa trên các dữ liệu đích được dán nhãn bởi bộ phân loại h, họ thực hiện lựa chọn đặc trưng trên dữ liệu đích. Tập các đặc trưng được chọn sẽ được sử dụng để xây dựng 2 bộ phân loại, một là h_S - từ các dữ liệu nguồn đã được gắn nhãn và một h_T từ các dữ liệu đích đã được dán nhãn. Hai bộ phân loại h_S và h_T sẽ làm việc cùng nhau để thực hiện phân loại các dữ liệu đích. Quá trình chạy lặp đi lặp lại cho đến khi nhãn được gán ổn định cho các dữ liệu đích. Trong mỗi lần lặp cả hai bộ phân loại sử dụng cùng một tập đặc trưng đã được lựa chọn từ miền đích để tập trung vào miền đích.

Thực nghiệm của họ sử dụng bốn tập dữ liệu được trích xuất từ bốn diễn đàn thảo luận thuộc bốn lĩnh vực khác nhau là Cellphone[2], Electronics[3], Camera[4], TV[5] và chỉ quan tâm tới những ý định về mua sắm - "buy". Mỗi tập dữ liệu chứa 1.000 bài post được gắn nhãn bằng tay bởi hai chuyên gia gắn nhãn. Đầu tiên, hai chuyên gia độc lập gắn nhãn cho 1/5 số bài post và họ cảm thấy kết quả tương đối khớp

[2] http://www.howardforums.com/forums.php,
[3] http://www.avsforum.com/avs-vb/
[4] http://forum.digitalcamerareview.com/
[5] http://www.avforums.com/forums/tvs/

nhau do vậy mà 4/5 số bài post còn lại chỉ được gắn nhãn bởi một chuyên gia. Sử dụng độ chính xác, hồi tưởng và độ đo F_1 họ kết luận rằng phương pháp Co-Class phù hợp để xác định các bài post chứa ý định.

1.5.3. Xây dựng mô hình ý định người dùng trên mạng xã hội sử dụng khai phá dữ liệu

Theo Zheng Chen cùng cộng sự [3], ý định của người sử dụng có thể được phân thành hai cấp độ: Ý định hành động và ý định ngữ nghĩa. Ý định hành động là mức độ thấp hơn, chẳng hạn như click chuột, đánh máy trên bàn phím và hành động cơ bản khác được thực hiện trên một máy tính. Ý định ngữ nghĩa tương ứng những gì người dùng muốn đạt được ở mức cao, trong đó có thể bao gồm một số hoạt động cơ bản trên một máy tính để thực hiện nó. Ví dụ: "***Tôi muốn mua một quyển sách từ Amazon***" - "*I want to buy a book from Amazon*"; "***Tôi muốn tìm một vài tài liệu về khai phá dữ liệu***" - "*I want to find some papers on data mining*" [12] là những ý định ngữ nghĩa.

Công trình tập trung vào việc dự đoán ý định hành động dựa trên các tính năng mà nhóm tác giả trích xuất từ sự tương tác người dùng. Ví dụ, trong khi lướt web, người dùng có thể tiến hành một loạt các hành động bao gồm cả cách nhấp (siêu liên kết), lưu(các trang), và đóng (các trình duyệt). Giả sử người dùng muốn mua một máy ảnh kỹ thuật số mà là ý định ngữ nghĩa, ông có thể làm như sau:

- Bước 1: Mở một trình duyệt Web bất kỳ
- Bước 2: Nhập www.amazon.com vào thanh địa chỉ
- Bước 3: Sau khi trang được trả về, một loạt các máy ảnh kỹ thuật số sẽ được hiển thị trên màn hình tìm kiếm
- Bước 4: Nhấp chuột vào một trong các đối tượng được chứa trong trang
- Bước 5: Click chuột vào nút mua để xác nhận
- Bước 6: Sau khi giao dịch xong, đóng trình duyệt

Trong ví dụ này, mục tiêu của tác giả là dự đoán chuỗi các hành động cơ bản mà người dùng sẽ tiến hành trong một hệ thống để hoàn thành ý định mua một thiết bị camera trên web.

Để khai phá ý định người dùng web, đầu tiên các tác giả tiến hành phân tích các đặc trưng ngôn ngữ. Theo Zheng Chen cùng cộng sự, ngôn ngữ có hai loại đặc

trưng ngôn ngữ: đặc trưng từ khóa – "keyword" và đặc trưng khái niệm. Một từ khóa "keyword" là một từ đơn được trích xuất từ các văn bản loại trừ các từ dừng – "stop-word" [17]. Ví dụ một câu "*Attached word file is a map of our office*" được so sánh với keyword "attach word file map office". Sau khi có các định nghĩa về đặc trưng ngôn ngữ, nhóm tác giả sử dụng giải thuật nhằm trích xuất đăc trưng và thuật toán Apriori được đề xuất bởi Agrawal cùng cộng sự [2] phục vụ sinh luật kết hợp. Sau đó, nhóm tác giả sử dụng Naïve Bayes phân loại để xây dựng mô hình ý định. Các thuật toán đã được sửa đổi các tham số để hỗ trợ gia tăng việc học.

Ngoài ra, để lựa chọn đặc trưng tốt, tác giả sử dụng IG (Information Gain) [18] mục tiêu để giảm kích thước của từ điển và nâng cao khả năng thực thi khi tập huấn luyện nhỏ. IG là thước đo độ quan trọng của thông tin trong dự đoán ý định bằng cách biết được sự có mặt hay vắng mặt của một đặc trưng trong bản ghi.

Sau khi được huấn luyện xong, mô hình ý định thu được cho người dùng sử dụng để dự đoán ý định của người dùng trong tương lai. Quy trình dự đoán như sau: Một tập các đặc trưng ngôn ngữ đại diện (< f_1, f_2,…, f_n>) được trích xuất từ văn bản của người dùng. Giả định các đặc trưng này độc lập với nhau, mô hình dự đoán tính toán các khả năng của tất cả ý định người dùng (V) và chọn một với khả năng lớn nhất (V_{NB}) dựa vào hàm sau :

$$V_{NB} = \arg\max P(v_j | f_1, f_2, \ldots, f_n) \text{ với } v_j \in V$$

$$= \arg\max P(v_j)\, P(f_1, f_2, \ldots, f_n \mid v_j) \text{ với } v_j \in V$$

$$= \arg\max P(v_j)\, P(f_1 \mid v_j)\, P(f_2 \mid v_j) \ldots P(f_n \mid v_j) \text{ với } v_j \in V$$

Bảng 3. Những phân bố có thể của mô hình huấn luyện. P(chọn) = 0.5, P(lưu) = 0.2, P(đóng) = 0.3.

	Chọn (**Click**)	Lưu (**Save**)	Đóng (**Close**)
Học (Learn**)**	0.5	0.2	0.3
Nghiên cứu (Research**)**	0.6	0.2	0.2
Nhận diện (Recognition**)**	0.7	0.1	0.2

Theo công thức được định nghĩa phía trên, có thể tính toán được:

- P(chọn)P(f_1 = "học" | chọn) * P(f_2 = "nghiên cứu" | "chọn") * P(f_3 = "nhận diện" | chọn) = 0.105

- P(lưu)P(f_1 = “học” | lưu) * P(f_2 = “nghiên cứu” | “lưu”) * P(f_3 = “nhận diện” | lưu) = 0.0008
- P(đóng)P(f_1 = “học” | đóng) * P(f_2 = “nghiên cứu” | “đóng”) * P(f_3 = “nhận diện” | đóng) = 0.0036

Từ kết quả trên, với những đăc trưng trên thì sẽ được phân lớp vào “click” .

Áp dụng vào công trình của mình, nhóm tác giả đã phát triển một công cụ tự động thu thập dữ liệu đăng nhập của người dùng trong môi trường IE, ghi lại năm hoạt động chính của người dùng. Lựa chọn ngẫu nhiên một số trang làm dữ liệu huấn luyện và phần còn lại là các kiểm tra dữ liệu theo một tỷ lệ đào tạo và lặp lại chia 10 lần trong một thử nghiệm.

Kết quả cho thấy, với α, β, ngưỡng IG tương ứng là 0,005; 0,6 và 0,02 thuật toán dự đoán có độ chính xác đạt 85%. Cụ thể “Browse” là hành động dễ dự đoán vì nó có phân hoạch lớn (60%) trong tập dữ liệu huấn luyện. “Query” là hành động được dự đoán với độ chính xác cao vì đặc trưng khái niệm đưa ra mô tả tốt trong kết quả tìm kiếm nội dung trang web. Tuy nhiên “Close” có độ chính xác thấp chỉ ra rằng nó có lẽ là không thể đoán trước chỉ dựa vào những thông tin văn bản.

1.5.4. Lọc ý định rõ ràng người dùng trong các bài viết Tiếng Việt trên phương tiện xã hội

Lọc ý định rõ ràng là mục tiêu được đặt ra trong công trình của nhóm tác giả Lương Thái Lê cùng cộng sự [5]. Đây là một bài toán con trong quá trình phân tích và trích xuất ý định. Theo các tác giả, quá trình phân tích và hiểu ý định người dùng gồm 3 pha chính: (1) Lọc ý định người dùng, (2) Xác định miền ý định, (3) Trích xuất và phân tích ý định.

Để xây dựng mô hình và phân tích ý định người dùng, các tác giả đưa ra định nghĩa: Ý định rõ ràng của người dùng có thể được định nghĩa qua 5 cấp/tính chất:

$$I_e^u = (u, c, d, w, p) \qquad \textbf{(1)}$$

Trong đó:

- **u**: xác định người dùng. *Ví dụ*: tên người dùng hoặc ID trên dịch vụ phương tiện xã hội
- **c**: ngữ cảnh hiện tại hoặc điều kiện xung ngữ cảnh hiện tại hoặc điều kiện quanh ý định này. *Ví dụ*: một người dùng có thể hiện tại có thai, ốm hoặc

có em bé. Ngữ cảnh c cũng có thể chứa thời gian tại ý định được bày tỏ hoặc bài đăng trên trực tuyến.

- **d**: miền của ý định.
- **w**: keyword hoặc cụm đại diện ý định. Nó có thể là tên một cái gì hoặc một hành động.
- **p**: danh sách tính chất hoặc ràng buộc liên quan tới một ý định. Nó là một danh sách cặp tính chất – giá trị liên quan đến ý định.

Trong công trình này, tác giả sử dụng phương pháp Maximum Entropy (MaxEnt) để xây dựng bộ phân lớp và sử dụng n-gram để định nghĩa các mẫu trưng/thuộc tính (với n = 2 và n = 3) kết hợp với bộ từ điển để sinh ra một vài mẫu kiểu như: muốn mua, cần tìm, đang cần, định vay, cần bán, muốn thuê, ..

Đánh giá mô hình phân lớp là bộ dữ liệu gồm 1.315 bài đăng/bình luận được lấy từ những bài đăng và bình luận bằng Tiếng việt của người dùng trên kênh phương tiện xã hội trực tuyến Facebook và Webtretho (một diễn đàn hoạt động mạnh ở Việt Nam) và được gắn nhãn bởi một nhóm sinh viên. Kết quả tập dữ liệu chứa 588 bài đăng/bình luận có ý định rõ ràng và 727 bài đăng/bình luận không có ý định. Kết quả dữ liệu chia ngẫu nhiên thành 4 phần. Tiến hành kiểm thử chéo 4 –fold cross – validation và kết quả đặt được với độ trung bình hơn 90%, đây cũng là một kết quả khả quan vì công trình chỉ dùng n-gram và từ điển để tìm kiếm đặc trưng.

1.6. Tóm tắt chương 1

Chương 1 của luận văn đã trình bày về tầm quan trọng của bài toán phát hiện ý định người dùng, khái quát về đầu vào cũng như đầu ra của bài toán, nêu lên những khó khăn gặp phải trong quá trình tìm hiểu dữ liệu và cuối cùng là hướng tiến cận giải quyết bài toán mà luận văn đang hướng tới. Trong chương 2, luận văn sẽ hướng tới tìm hiểu về phương pháp suy luận được sử dụng để áp dụng cho bài toán của mình, các phương pháp cải thiện chất lượng của bài toán và một số công trình đã sử dụng phương pháp suy luận để từ đó biết được mức độ, phạm vi sử dụng mô hình.

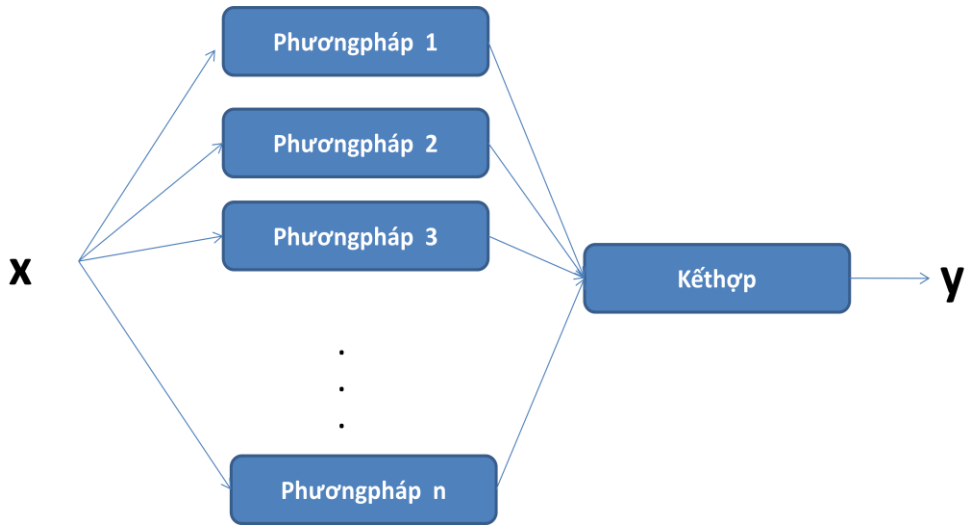
Chương 2. Phương pháp suy luận các mô hình và áp dụng nó cho các bài toán phân lớp

Trong chương này, báo cáo trình bày chi tiết về phương pháp suy luận hay còn được gọi là phương pháp lai ghép, kết hợp các mô hình. Tiếp sau đó trong phần 2.2, mô hình sẽ trình bày mức độ hiệu quả của phương pháp này để biết được mức độ cũng như phạm vi sử dụng của mô hình. Trong phần 2.3, báo cáo trình bày về một vài kỹ thuật và thuật toán nhằm cái thiện chất lượng của phương pháp được nghiên cứu.

2.1. Phương pháp suy luận các mô hình

Kể từ năm 1990, phương pháp suy luận hay còn 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[11][12][13][15][16][22][30][35][36].

Trái ngược với các hướng tiếp cận học thông thường chỉ sử dụng một phương pháp từ dữ liệu huấn luyện, phương pháp kết hợp 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 kết hợp 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 3. Một kiến trúc kết hợp chung

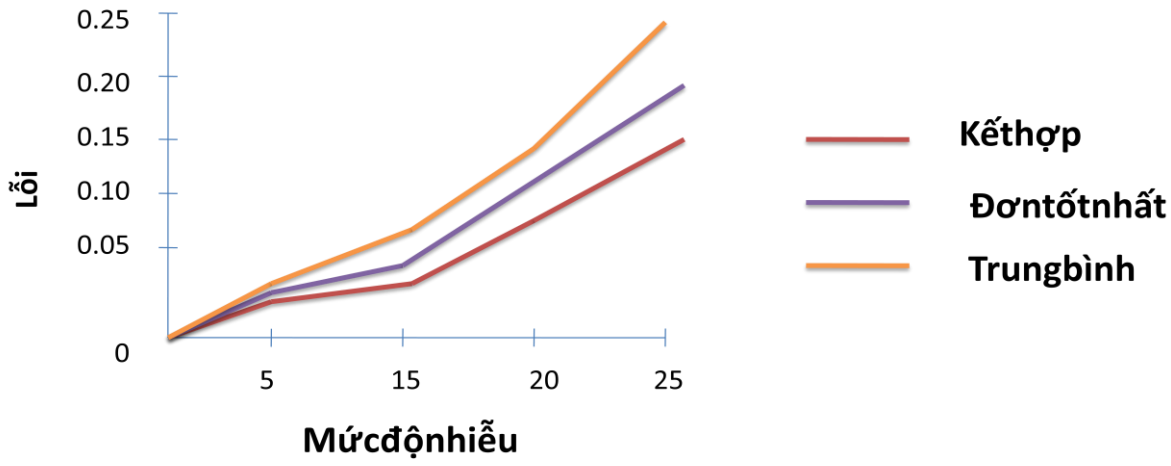
Theo tác giả Zhou [22], một phương pháp kết hợp sẽ chứa nhiều mô hình được gọi là các mô hình cơ sở. Các mô hình cơ sở được xây dựng từ tập dữ liệu huấn luyện bởi một giải thuật học cơ sở như cây quyết định, mạng xoắn hay một giải thuật học nào đó. Phần lớn, các phương pháp kết hợp sẽ sử dụng duy nhất 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 lai ghép 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 đưa ra các giải thích về việc tăng chất lượng của các mô hình yếu. [14][18][32][33]

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 [19][23].

Kể từ những năm 1990, phương pháp lai ghép các mô hình đã trở thành một mô hình học lớn với sự tiên phong của hai công trình lớn. Một là thực nghiệm của tác giả Hansen và Salamon (1990) [19], thực nghiệm cho thấy rằng những dự đoán được tạo ra bởi sự kết hợp một tập các bộ phân lớp thường chính xác hơn so với những dự đoán được tạo bởi một bộ phân lớp đơn tốt nhất [Hình 4]. Thực nghiệm thứ hai của tác giả [Schapire, 1990] đưa ra một chứng minh rằng những mô hình yếu có thể được phát triển thành mô hình mạnh. Trong thực tế, việc tìm kiếm ra các mô hình mạnh thường khá khó, tuy nhiên để tìm ra các mô hình yếu có lẽ dễ hơn. Do vậy, thực nghiệm này có lẽ đã mở ra một hướng đi đầy hứa hẹn cho việc tạo ra các mô hình mạnh từ phương pháp kết hợp các mô hình.



Hình 4. Một thực nghiệm chứng minh của Hasen và Salamon: Kết hợp thì thường tốt hơn mô hình đơn tốt nhất.

Nói chung, một phương pháp kết hợp được tạo thành từ hai bước. Bước một, tạo ra các mô hình cơ sở. Bước hai, kết hợp chúng lại. Để tạo ra một phương pháp lai tốt thì các mô hình cơ sở cần được xây dựng càng chính xác, càng đa dạng càng tốt.

2.2. Đánh giá hiệu quả của phương pháp lai ghép các mô hình

KDD-Cup[6] là cuộc thi khai phá dữ liệu nổi tiếng. Kể từ năm 1997, hàng năm, cuộc thi này tổ chức và thu hút nhiều sự quan tâm của các đội khai phá dữ liệu trên toàn thế giới. Các bài toán trong cuộc thi hầu hết là các chủ đề thực tiễn như phát hiện xâm nhập mạng (2001), dự đoán vị trí của protein và hoạt động của các phân tử sinh học (2001), phát hiện sự tắc nghẽn trong phổi (2006), quản lý mối quan hệ khách hàng (2009), khai phá dữ liệu giáo dục (2010), hệ tư vấn âm nhạc (2011), ... Trong các cuộc thi KDD-Cup vừa qua, trong số các kỹ thuật khác nhau được sử dụng trong các giải pháp, các phương pháp lai ghép các mô hình đã gây nhiều chú ý nhất và giành chiến thắng trong hầu hết các cuộc đua. Ví dụ, KDD-Cups trong 3 năm (2009-2011), những người chiến thắng ở vị trí thứ nhất và vị trí thứ hai đều sử dụng phương pháp lai ghép, kết hợp mô hình.

Một cuộc thi nổi tiếng khác, NetFlix Prize[7] được tổ chức bởi DVD trực tuyến nhằm tìm kiếm cải thiện độ chính xác của các dự đoán về số lượng người sẽ thích một bộ phim dựa trên sở thích của họ. Nếu một đội tham gia cải thiện thuật toán

[6]http://www.sigkdd.org/kddcup/.
[7]http://www.netflixprize.com/.

riêng của Netflix độ chính xác tăng lên 10% thì sẽ giành được giải thưởng trị giá 1,000,000 USD. Vào 21/9/2009, Nexflix đã trao giải 1,000,000 USD cho đội BellKor's Pragatic Chaos, giải pháp của họ dựa trên sự kết hợp của nhiều bộ phân lớp bao gồm các mô hình nhân tố bất đối xứng (anymetric factor model), mô hình hồi quy, học máy Boltzmann được hạn chế (restricted Boltzmann machines), ma trận thừa (matrix factorization), k người láng giềng gần nhất (K-NN), ...

Ngoài những kết quả ấn tượng trong cuộc thi, phương pháp lai kết hợp các mô hình cũng đã được áp dụng một cách thành công cho nhiều nhiệm vụ thực tiễn đa dạng. Đây là một phương pháp rất hữu ích trong các bài toán sử dụng kỹ thuật học.

Năm 2000, Huang cùng cộng sự [20] đã thiết kế một kiến trúc lai cho bài toán nhận diện khuôn mặt người tư thế bất biến, đặc biệt là nhận diện các khuôn mặt với phép quay sâu. Ý tưởng cơ bản là kết hợp một lượng các mạng xoắn được xem cụ thể với một mô đun kết hợp được thiết kế đặc biệt. Trái ngược với những kỹ thuật thông thường yêu cầu đầu vào là thông tin của tư thế, khung làm việc này không yêu cầu thông tin về tư thế, nó thậm trí có thể đưa ra các ước tính tư thế cho kết quả nhận diện. Huang cùng cộng sự báo cáo rằng khung làm việc này thậm trí vượt trội so với các kỹ thuật thông thường. Ngay sau đó, năm 2001 Li cùng cộng sự [26] cũng đã sử dụng phương pháp tương tự để áp dụng cho bài toán nhận diện khuôn mặt người đa cảm xúc.

Thêm nữa, phương pháp lai ghép các mô hình còn được đánh giá cao cho việc mô tả các bài toán bảo mật máy tính bởi mỗi hành động thực thi trên hệ thống máy tính có thể được quan sát ở nhiều mức độ trừ tượng và các thông tin liên quan có thể được thu thập từ nhiều nguồn thông tin [27]. Giacinto cùng cộng sự đã áp dụng phương pháp lai ghép các mô hình vào việc phát hiện xâm nhập [30][31]. Giacinto cho rằng khi phát hiện xác tấn công đã biết, phương pháp này cho hiệu quả rất tốt.

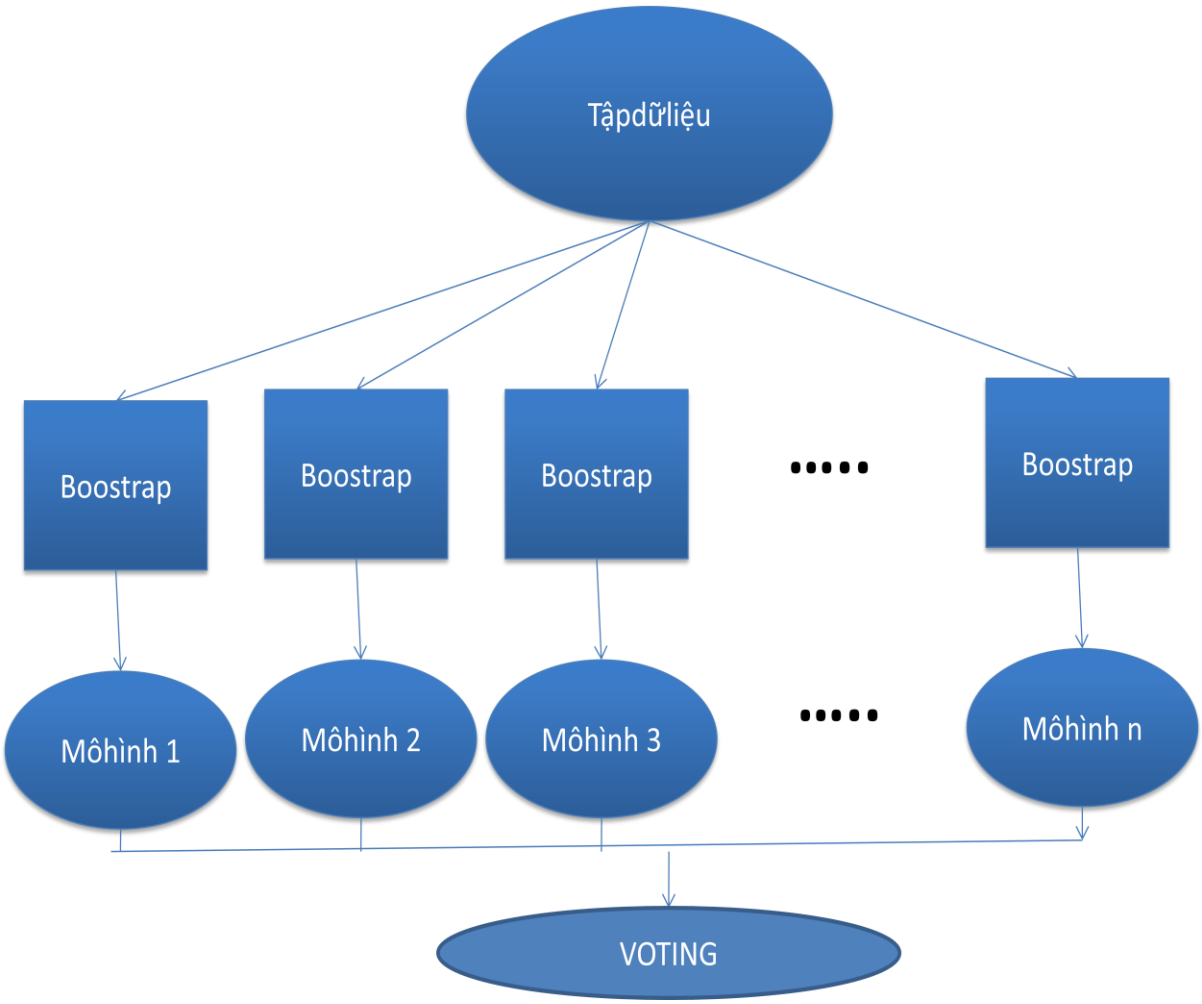
2.3. Bagging - Kỹ thuật nâng cao độ chính xác của phương pháp lai ghép các mô hình trong bài toán phân lớp

Boosting và bagging là hai trong số những tiếp cận gần đây cho phép nâng cao độ chính xác của các giải thuật phân lớp. Ở phần này, tôi mô tả kỹ hơn về phương pháp bagging – phương pháp được sử dụng trong luận văn của mình. Cụ thể, tôi thực hiện việc áp dụng phương pháp bagging trong việc giải quyết bài toán phân lớp nhận diện ý định người dùng mua hàng.

Bagging xuất phát từ tên viết tắt Bootstrap AGGregatING [Breiman, 1996] có nghĩa là 2 thành phần chính của Bagging là Bootstrap và Aggregation. Chúng ta đã biết rằng sự kết hợp của các mô hình cơ sở độc lập sẽ dẫn tới việc giảm đáng kể các lỗi. Bởi vậy, mục tiêu là lấy càng nhiều mô hình cơ sở càng độc lập càng tốt.

Bagging tạo ra các bộ phân lớp từ các tập mẫu con có hoàn lại tập mẫu ban đầu (mẫu boostrap) và một thuật toán học máy, mỗi tập mẫu sẽ tạo ra một bộ phân lớp cơ bản. Các bộ phân lớp sẽ được kết hợp bằng phương pháp biểu quyết theo số đông – phương pháp Voting. Tức là khi có một ví dụ cần được phân lớp, mỗi bộ phân lớp sẽ cho ra một kết quả. Và kết quả nào xuất hiện nhiều nhất sẽ được lấy làm kết quả của bộ kết hợp.

Cụ thể: Bagging tạo ra N tập huấn luyện được chọn có lặp từ tập dữ liệu huấn luyện ban đầu. Trong đó các ví dụ huấn luyện có thể được chọn hơn một lần hoặc không được chọn lần nào. Từ mỗi tập huấn luyện mới, Bagging cho chạy với một thuật toán học máy L_b để sinh ra M bộ phân loại cơ bản h_m. Khi có một ví dụ phân lớp mới, kết quả của bộ kết hợp sẽ là kết quả nhận được nhiều nhất khi chạy M bộ phân lớp cơ bản. Thuật toán tiêu biểu cho kỹ thuật này là Random Forest. Hình vẽ bên dưới mô tả về phương pháp Bagging.



Hình 5. Hình ảnh về phương pháp Bagging.

2.4. Phương pháp suy luận các mô hình trong việc giải quyết bài toán phân lớp và ý tưởng áp dụng

Phương pháp suy luận các mô hình từ lâu đã nhận được nhiều quan tâm từ cộng đồng nghiên cứu. Có khá nhiều tác giả đã sử dụng phương pháp này cho các nghiên cứu nhằm giải quyết và cải thiện chất lượng bài toán của họ, chẳng hạn [29][30][31][35]

Liên quan tới việc áp dụng phương pháp suy luận (kết hợp) cho bài toán phân lớp, nhóm các tác giả Wei Wu, Zheng Liu và Yan He đã sử dụng phương pháp này cho bài toán phân loại lỗi của đường ống xử lý nước thải một cách tự động [44]. Trong công trình này, nhóm tác giả đã chứng minh tính hiệu quả của việc sử dụng phương pháp kết hợp bốn mô hình: AdaBoost, Rừng ngẫu nhiên (Random Forest), Rừng xoay (Rotation Forest), và RotBoost trong bài toán phát hiện tự động lỗi có thể thay thế sức người. Michiel van Wezel cùng cộng sự [45]

đưa ra cách cải thiện bài toán dự đoán trong các lựa chọn của khách hàng sử dụng phương pháp lai ghép. Sajid Yousuf Bhat cùng cộng sự [42] thì sử dụng phương pháp này cho bài toán phát hiện thư rác. Trong bài báo này, tác giả đánh giá việc thực hiện một số phương pháp học kết hợp sử dụng đặc điểm cấu trúc dựa vào nội dung của thư nhằm phát hiện thư rác trên các trang mạng xã hội trực tuyến. Các tác giả đánh giá hiệu suất của ba bộ phân loại bao gồm J48 (cây quyết định), IBK (k-NN sử dụng k = 5 hàng xóm gần nhất), và NaïveBayes sau đó sử dụng các kỹ thuật bagging, boosting và stacking để đánh giá hiệu quả. David Optiz cùng cộng sự [40] cũng đã đưa ra nhận định về hiệu quả của việc sử dụng kết hợp các bộ phân lớp luôn tốt hơn chỉ sử dụng một bộ phân lớp duy nhất. Trong bài báo, tác giả đã đánh giá hai kỹ thuật Bagging và Boosting trên 23 bộ dữ liệu sử dụng kết hợp hai giải thuật phân lớp mạng thần kinh và cây quyết định. Kết quả cho thấy rằng Bagging hầu như luôn luôn chính xác hơn một bộ phân lớp đơn và nó đôi khi chính xác hơn Boosting. Mặt khác, Boosting có thể tạo ra các kết hợp ít chính xác hơn một bộ phân lớp đơn–cụ thể là sử dụng mạng thần kinh. Ngoài ra, tác giả cũng nhận định rằng, hiệu quả của các phương pháp Boosting còn phụ thuộc vào đặc điểm của bộ dữ liệu được kiểm tra. Cũng sử dụng phương pháp này cho bài toán phát hiện ý định nhưng trên một miền lĩnh vực khác là trình duyệt Web, Alejandro Figueroa cùng cộng sự [35] đề xuất sử dụng 3 hướng kết hợp khác nhau: Kết hợp thường, Kết hợp hướng ngữ nghĩa, Kết hợp hướng độ dài cho kho dữ liệu truy vấn Web AOL chứa khoảng 21 triệu truy vấn từ 650.000 người dùng tìm kiếm. Đánh giá kết quả thu được, các tác giả nhận xét rằng việc kết hợp các bộ phân loại hỗ trợ rất nhiều trong việc cải thiện chất lượng bài toán xác định ý định người dùng.

2.5. Tóm tắt chương 2

Trong chương này, luận văn đã giới thiệu khái quát về phương pháp lai ghép các mô hình và kỹ thuật Bagging nhằm cải thiện chất lượng của phương pháp. Luận văn cũng trình bày một số công trình đã áp dụng phương pháp lai ghép nhằm giải quyết bài toán phân lớp dữ liệu.

Từ những chứng minh và các phương pháp thực hiện, tôi đã nghiên cứu và đề xuất một mô hình nhằm cải thiện chất lượng bài toán nhận diện, phát hiện ý định mua hàng của người dùng trên mạng xã hội Facebook sử dụng kết hợp ba phương pháp SVM – KNN– MaxEnt.

Trong chương tiếp theo, luận văn sẽ giới thiệu mô hình đề xuất giải quyết bài toán và các bước cụ thể.

Chương 3. Mô hình và thực nghiệm

3.1. Tư tưởng đề xuất mô hình

Bài toán phát hiện ý định mua hàng của người dùng 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 Beyesian, 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 sử dụng kết hợp ba thuật toán phân lớp tiêu biểu SVM – KNN – MaxEnt. 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 dùngchia 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 tintrên 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 làmột bình luận chưa được chuẩn hóa của người dùng: "*Áo này bao tjền 1 chjếc vậy*". Trường hợp này sẽ rất khó để nhận diện ra ý định của họ.

Hình 6. 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 7. 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.

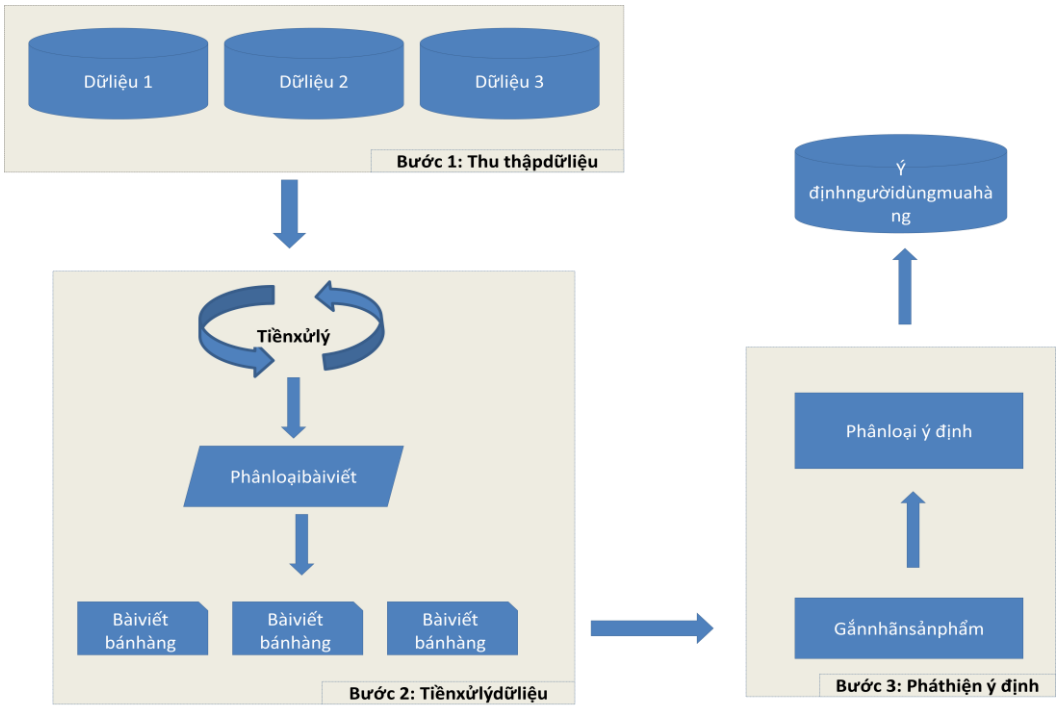
- **Thứ hai, về quá trình phát hiện ý định người dùng trên các trang bán hàng**: Để tập trung vào việc phát hiện ý định người dùng mua hàng, từ những dữ liệu đã được thu thậplà các bài viết của các trang bán hàng, tôi thực hiện một bước tiền xử lý nhằm lọc ra những bài viết bán hàng. Bởi trong một trang 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.

3.2. 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 (SVM), K – Nearest Neighbors (KNN) và Maximum Entropy Model (MaxEnt) 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 nàyđề xuất mô hìnhnhằm 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 8. 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.

3.2.1. Thu thập dữ liệu

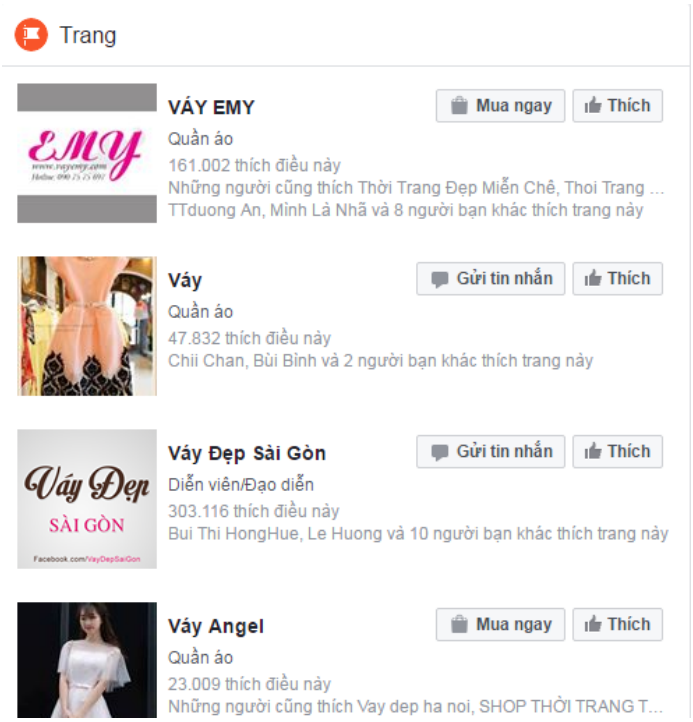
Trong bước này, tôi có hai nhiệm vụ chính:

- ***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. Bên dưới là một ví dụ về cây sản phẩm được xây dựng từ từ khóa sản phẩm nhằm phục vụ cho bài toán phân loại bài viết.

Danh mục sản phẩm	Mục con 1	Mục con 2	Mục con 3
Thời trang	Thời trang nữ	Đầm, váy	Váy đuôi cá
			Váy xòe
		Áo	Áo sơ mi
			Áo rớt vai
	Thời trang nam	Áo	Áo thun
			Áo body
		Quần	Quần tây
			Quần jean

Hình 9. Ví dụ về cây danh mục sản phẩm.

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 10. 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

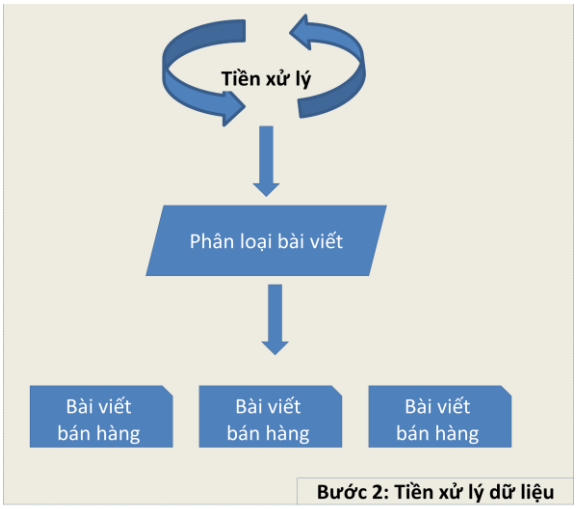
(comment) của nó. Để thực hiện được nhiệm vụ này, tôi sử dụng công cụ Facebook Graph API


```
{
  "id": "1855978744648203_1855982961314448",
  "postId": "1855978744648203",
  "actorProfileId": "100003132260513",
  "actorProfileDisplayName": "Hải Yến",
  "content": "Áo bn b",
  "commentTime": 1476423773000,
},
{
  "id": "1855978744648203_1855985564647521",
  "postId": "1855978744648203",
  "actorProfileId": "100003857448765",
  "actorProfileDisplayName": "Tuyet Tuyet Ngo",
  "content": "Shop inbox mình xin cái giá!",
  "commentTime": 1476424482000
},
{
  "id": "1855978744648203_1856132781299466",
  "postId": "1855978744648203",
  "actorProfileId": "100004025581213",
  "actorProfileDisplayName": "Thuy Duong",
  "content": "Bn chị ơi có để lại cho e 1 cái",
  "commentTime": 1476444802000
},
```


Hình 11. Hình ảnh về quá trình thu thập dữ liệu sử dụng Facebook Graph API

3.2.2. 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 12. Bước 2: Tiền xử lý dữ liệu

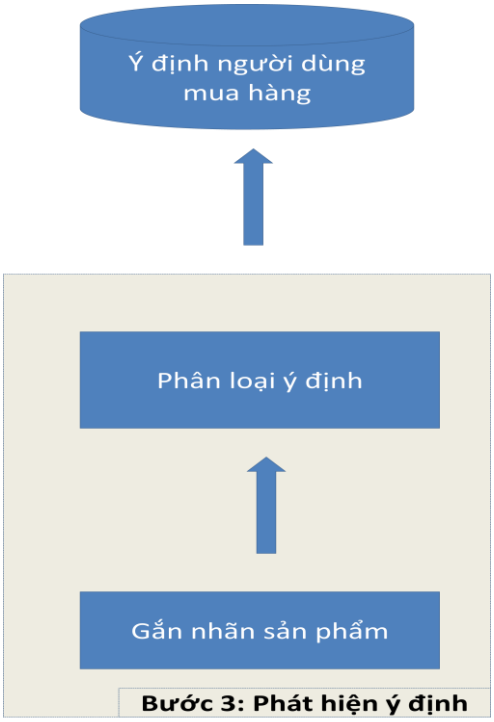
Trong bước này, tôi sử dụng ba mô hình phân lớp đã được đề cập ở các chương trước để thu lấy những bài viết bán hàng cho nhiệm vụ tiếp theo. Bởi, những bài viết không bán hàng thì lượng người dùng bình luận mua hàng rất ít và hầu như là không có. Sau khi thu thập và phân lớp dữ liệu bài đăng, tiến hành 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 mông mà không sứt chỉ ý ạ. Hihihi. 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.

3.2.3. Phân tích và phát hiện ý định



Hình 13.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 như: 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: chấm, hóng giá, đặt gạch, x tạm, đẹp quá, made in, …
- ***Loại 4 – không có ý định***: Bình luận sẽ chứa từ khóa như: đẹp nhưng đắt, cũng bình thường, tạm, xấu, …

3.3. Các độ đo đánh giá

Hầu hết các hệ thống sử dụng độ chính xác, độ hồi tưởng, độ đo F-score (F_1) để tính hiệu năng mô hình học máy, chúng tôi hướng tới cũng sử dụng những độ đo chuẩn này, cụ thể:

- *Độ hồi tưởng (Recall*): Số ý định được hệ thống đã nhận dạng đúng / Tổng số ý định đúng chứa trong văn bản đầu vào.
- *Độ chính xác (Precision):* Số ý định được hệ thống nhận dạng đúng / Tổng số ý định được xác định bởi hệ thống.
- *F-score (F_1):* Độ đo hài hòa giữa độ chính xác và độ hồi tưởng.

$$\textit{Công thức: } F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Phương pháp thường được sử dụng để đánh giá là *kiểm thử chéo* (cross validation). Phương pháp này tiến hành chia ngẫu nhiên tập dữ liệu thành N phần bằng nhau, mỗi một thực nghiệm sẽ tiến hành học trên N-1 phần và đánh giá mô hình học trên phần còn lại. Kết quả cuối cùng sẽ được thống kê từ N thực nghiệm ở trên.

3.4. Kết quả thực nghiệm và đánh giá

Dựa vào mô hình đề xuất ở phần trên, luận văn tiến hành thực nghiệm việc phát hiện ý định mua hàng của người dùng trên các Trang bán hàng dựa vào bình luận của họ. Để làm rõ mô hình đề xuất cũng như 3 pha chính trong mô hình, dữ liệu thực nghiệm được tiến hành xây dựng trong trên khoảng thời gian 3 tháng, từ ngày 01/01/2016 đến ngày 01/03/2016

3.4.1. Môi trường thực nghiệm

Các thông số phần cứng của hệ thống cài đặt thử nghiệm phương pháp được thực hiện thể hiện trong bảng sau đây:

Bảng 4. Môi trường thực nghiệm

Thành phần	**Chỉ số**
Bộ vi xử lý	Intel Core i3-3110M 2.53GHz
Bộ nhớ trong	2.00G
Bộ nhớ ngoài	320GB
Hệ điều hành	Windows 7 Ultimate

Các phần mềm, công cụ được sử dụng:

Bảng 5. Bảng tên các phần mềm được sử dụng.

STT	**Tên phần mềm**	**Nguồn**
1	Facebook Graph API	https://developers.facebook.com/tools/explorer
2	Eclipse-SDK	http://www.eclipse.org/downloads
3	Liblinear	http://www.csie.ntu.edu.tw/~cjlin/liblinear/
4	JDK 1.8	http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html
5	OpenNLP	https://opennlp.apache.org/
6	LuaText 1.0	https://www.latex-project.org/

Bảng 6. Bảng danh sách các module trong thực nghiệm.

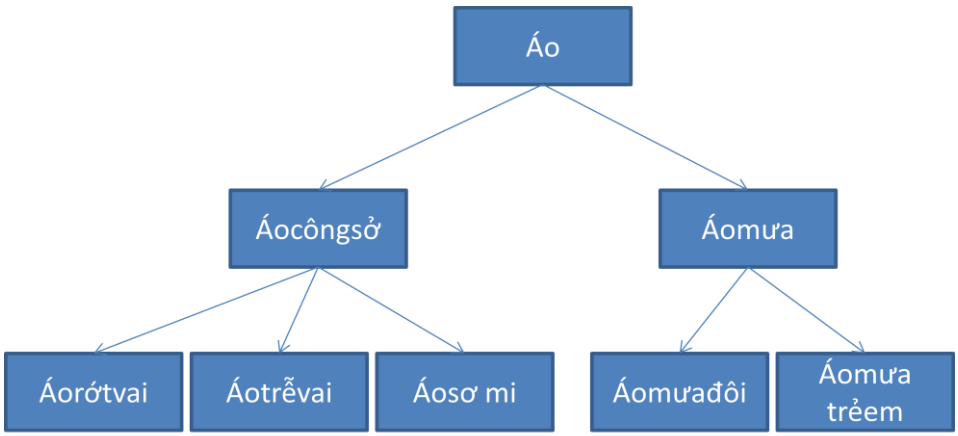
STT	**Gói chính**	**Mục đích**
1	org.sonat.analysis.classify	Các lớp thực huấn luyện, kiểm thử và thực thi các thành phần xác định ý định: - CommentClassifier.java - CommentModelCrossValidator.java - CommentTrainer.java - PostClassifier.java - PostModelCrossValidator.java - PostTrainer.java
2	org.sonat.analysis.feature	Các đặc trưng sử dụng trong quá trình phần lớp: Bigram, bow, length, regex, trigram
3	org.sonat.analysis.classify	Chứa các class là các kỹ thuật phân lớp Maxent, SVM, Bagging: - MaxentClassification.java - SVMClassification.java - KNNClassification.java - BaggingEnsemble.java
4	org.sonat.analysis.vlsp	Công cụ xử lý ngôn ngữ như tách câu, tách từ…
5	Resources	Lưu các mô hình phân loại văn bản

3.4.2. Dữ liệu huấn luyện bài viết

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.



Hình 14. Ví dụ về cây danh mục sản phẩm.

Thống kê dữ liệu bài viết theo hai lớp được biểu diễn trong bảng bên dưới:

Bảng 7. 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

3.4.3. Dữ liệu phân loại ý định trong bình luận

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: "***Chấm***, chờ khi nào có lương sẽ qua múc."

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 8. 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

3.5. Thực nghiệm đánh giá mô hình phân lớp

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.

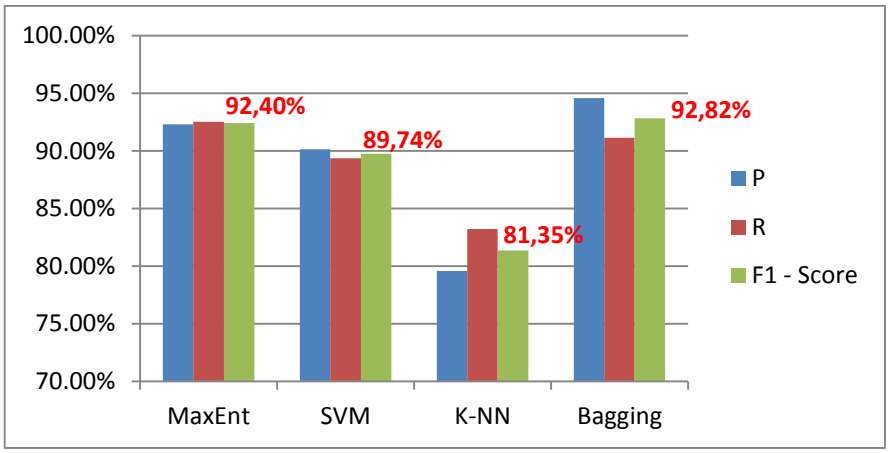
Luận văn so sánh đánh giá hiệu quả của các phân lớp riêng biệt cùng với các mô hình sử dụng kỹ thuật lai ghép. Danh sách các kỹ thuật được sử dụng như sau:

- Phương pháp K người láng giếng gần nhất (KNearest Neighbors - KNN)
  - Sử dụng tham số K = 3
  - Độ đo tính sự tương tự là độ đo Cosine
- Phương pháp máy vector hỗ trợ (Support Vector Machine)
  - Công cụ Liblinear
  - Sử dụng tham số L2-loss linear SVM, L1-loss linear SVM
- Phương pháp cực đại entropy (Maximum Entropy - Maxent)
  - Công cụ OpenNLP 1.6.0
  - Tham số iteration = 200, cutoff= 1
- Phương pháp lai ghép 3 mô hình sử dụng kỹ thuật Bagging
  - Sử dụng kỹ thuật ghép bình chọn (voting), nhãn chiếm đa số sẽ là nhãn cuối cùng của dữ liệu

Dưới đây là kết quả phân lớp bài viết bán hàng:

Bảng 9. Bảng kết quả phân lớp bài viết 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%**

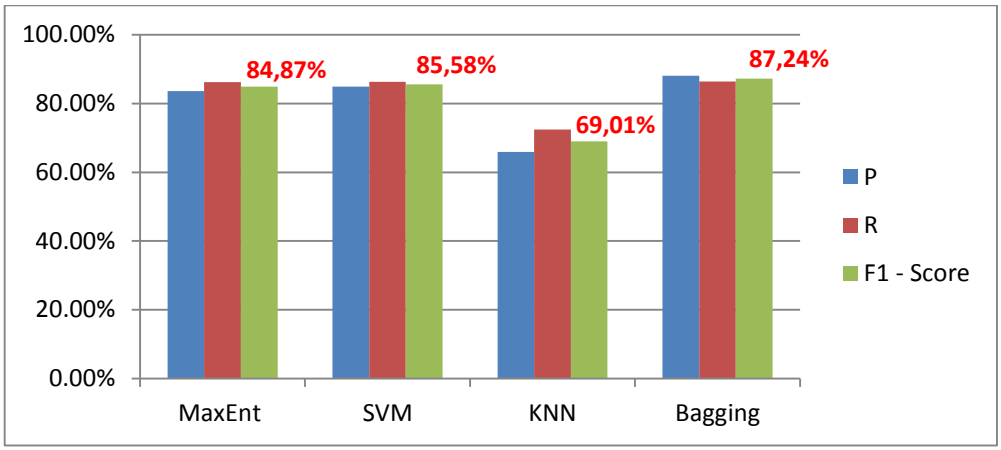


Hình 15. Kết quả phân lớp bài viết bán hàng.

- Kết quả phân lớp ý định: Có ý định

Bảng 10. Bảng kết quả phân lớp các 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%**



Hình 16. Kết quả phân lớp ý định.

Nhận xét: Kết hợp kết quả từ 2 bảng 7, 8 và biểu đồ 16 và 17 cho thấy: Với bài toán phân lớp các bài viết bán hàng, đa phần các mô hình đều đạt kết quả tốt (khoảng 90%). Trong đó là hai mô hình phân lớp MaxEnt (92,40%) và Bagging (92,82%) trong khi hai mô hình phân lớp còn lại đưa ra độ chính xác thấp hơn là SVM xấp xỉ 89% và K – NN với số người hàng xóm là K = 3 thì đạt kết quả khoảng 81%. Với bài toán phân lớp ý định mua hàng của người dùng, độ chính xác thấp nhất vẫn là mô hình K – NN (69,01%) trong khi MaxEnt và SVM đưa ra kết quả gần ngang nhau lần lượt là 84,87% và 85,58%. Đưa ra kết quả tốt nhất vẫn là Bagging 87,24%.

Kết quả trên cho thấy bài toán phân loại ý định người dùng sử dụng phương pháp lai ghép, kết hợp các mô hình thực sự tốt hơn khi chỉ sử dụng một mô hình phân lớp duy nhất.

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.

Tài liệu tham khảo

[1] Wang, J., Cong, G., Zhao, W. X., & Li, X. (2015, January). Mining User Intents in Twitter: A Semi-Supervised Approach to Inferring Intent Categories for Tweets. In *AAAI* (pp. 318-324).

[2] Chen, Z., Liu, B., Hsu, M., Castellanos, M., & Ghosh, R. (2013, June). Identifying Intention Posts in Discussion Forums. In *HLT-NAACL* (pp. 1041-1050).

[3] Chen, Z., Lin, F., Liu, H., Liu, Y., Ma, W. Y., & Wenyin, L. (2002). User intention modeling in web applications using data mining. *World Wide Web*, *5*(3), 181-191.

[4] Bratman, Michael. "Intention, plans, and practical reason." (1987).

[5] Luong, T. L., Tran, T. H., Truong, Q. T., Phi, T. T., & Phan, X. H. (2016, March). Learning to Filter User Explicit Intents in Online Vietnamese Social Media Texts. In *Asian Conference on Intelligent Information and Database Systems*(pp. 13-24). Springer Berlin Heidelberg.

[6] Kröll, M., & Strohmaier, M. (2009, September). Analyzing human intentions in natural language text. In *Proceedings of the fifth international conference on Knowledge capture* (pp. 197-198). ACM.

[7] Purohit, H., Dong, G., Shalin, V., Thirunarayan, K., & Sheth, A. (2015, December). Intent Classification of Short-Text on Social Media. In *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*(pp. 222-228). IEEE..

[8] Khademi, G., Mohammadi, H., Simon, D., & Hardin, E. C. (2015, October). Evolutionary optimization of user intent recognition for transfemoral amputees. In *Biomedical Circuits and Systems Conference (BioCAS), 2015 IEEE* (pp. 1-4). IEEE.

[9] Jansen, B. J., Booth, D. L., & Spink, A. (2007, May). Determining the user intent of web search engine queries. In *Proceedings of the 16th international conference on World Wide Web* (pp. 1149-1150). ACM.

[10] Andrei Broder. A Taxonomy of Web Search. SIGIR Forum, 36(2):3–10, September 2002

[11] Sewell, Martin. "Ensemble learning." *RN* 11.02 (2008).

[12] Ho, Tin Kam, Jonathan J. Hull, and Sargur N. Srihari. "Decision combination in multiple classifier systems." *IEEE transactions on pattern analysis and machine intelligence* 16.1 (1994): 66-75.

[13] Larkey, Leah S., and W. Bruce Croft. "Combining classifiers in text categorization." *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1996.

[14] Skurichina, Marina, and Robert PW Duin. "Bagging, boosting and the random subspace method for linear classifiers." *Pattern Analysis & Applications* 5.2 (2002): 121-135.

[15] Xu, Lei, Adam Krzyzak, and Ching Y. Suen. "Methods of combining multiple classifiers and their applications to handwriting recognition." *IEEE transactions on systems, man, and cybernetics* 22.3 (1992): 418-435.

[16] Fu, B., and Liu, T. 2013. Weakly-supervised consumption intent detection in microblogs. Journal of Computational Information Systems 6(9):2423–2431.

[17] Ding, X., Liu, T., Duan, J., & Nie, J. Y. (2015, January). Mining User Consumption Intention from Social Media Using Domain Adaptive Convolutional Neural Network. In *AAAI* (pp. 2389-2395).

[18] Kuncheva, L. I. (2004). *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons.

[19] Hansen, Lars Kai, and Peter Salamon. "Neural network ensembles." *IEEE transactions on pattern analysis and machine intelligence* 12 (1990): 993-1001.

[20] Huang, Z.-H. Zhou, H.-J. Zhang, and T. Chen. Pose invariant face recognition. In Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition, pages 245–250, Grenoble, France, 2000

[21] Zhou and Y. Jiang. Medical diagnosis with C4.5 rule preceded by artificial neural network ensemble. IEEE Transactions on Information Technology in Biomedicine, 7(1):37–42, 2003

[22] Zhou. When semi-supervised learning meets ensemble learning. Frontiers of Electrical and Electronic Engineering in China, 6(1):6–16, 2011

[23] Giacinto and F. Roli. Design of effective neural network ensembles for image classification purposes. Image and Vision Computing, 19(9-10): 699–707, 2001

[24] Gneiting and A. E. Raftery. Atmospheric science: Weather forecasting with ensemble methods. Science, 310(5746):248–249, 2005

[25] K. Ho, J. J. Hull, and S. N. Srihari. Decision combination in multiple classifier systems. IEEE Transaction on Pattern Analysis and Machine Intelligence, 16(1):66–75, 1994.

[26] Z. Li, Q. Fu, L. Gu, B. Scholkopf, and H. J. Zhang. Kernel machine based ¨ learning for multi-view face detection and pose estimation. In Proceedings of the 8th International Conference on Computer Vision, pages 674– 679, Vancouver, Canada, 2001

[27] Corona, G. Giacinto, C. Mazzariello, F. Roli, and C. Sansone. Information fusion for computer security: State of the art and open issues. Information Fusion, 10(4):274–284, 2009

[28] Giacinto, F. Roli, and G. Fumera. Design of effective multiple classi- fier systems by clustering of classifiers. In Proceedings of the 15th International Conference on Pattern Recognition, pages 160–163, Barcelona, Spain, 2000

[29] Giacinto and F. Roli. Design of effective neural network ensembles for image classification purposes. Image and Vision Computing, 19(9-10): 699–707, 2001

[30] Giacinto, F. Roli, and L. Didaci. Fusion of multiple classifiers for intrusion detection in computer networks. Pattern Recognition Letters, 24(12): 1795–1803, 2003

[31] Giacinto, R. Perdisci, M. D. Rio, and F. Roli. Intrusion detection in computer networks by a modular ensemble of one-class classifiers. Information Fusion, 9(1):69–82, 2008

[32] Freund. Boosting a weak learning algorithm by majority. Information and Computation, 121(2):256–285, 1995

[33] Freund. An adaptive version of the boost by majority algorithm. Machine Learning, 43(3):293–318, 2001

[34] Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1):119–139, 1997

[35] Figueroa, Alejandro, and John Atkinson. "Ensembling Classifiers for Detecting User Intentions behind Web Queries." *IEEE Internet Computing*20, no. 2 (2016): 8-16.

[36] Ponti Jr, M. P. (2011, August). Combining classifiers: from the creation of ensembles to the decision fusion. In *Graphics, Patterns and Images Tutorials (SIBGRAPI-T), 2011 24th SIBGRAPI Conference on* (pp. 1-10). IEEE.

[37] Moon, Hojin, Hongshik Ahn, Ralph L. Kodell, Songjoon Baek, Chien-Ju Lin, and James J. Chen. "Ensemble methods for classification of patients for personalized medicine with high-dimensional data." *Artificial intelligence in medicine* 41, no. 3 (2007): 197-207.

[38] Bar, Ariel, Lior Rokach, Guy Shani, Bracha Shapira, and Alon Schclar. "Improving simple collaborative filtering models using ensemble methods." In *International Workshop on Multiple Classifier Systems*, pp. 1-12. Springer Berlin Heidelberg, 2013.

[39] Zamora, Juan, Marcelo Mendoza, and Héctor Allende. "Query Intent Detection Based on Query Log Mining." *J. Web Eng.* 13.1&2 (2014): 24-52.

[40] Opitz, David, and Richard Maclin. "Popular ensemble methods: An empirical study." *Journal of Artificial Intelligence Research* 11 (1999): 169-198.

[41] Abbasian, H., Drummond, C., Japkowicz, N., & Matwin, S. (2013, September). Inner ensembles: Using ensemble methods inside the learning algorithm. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 33-48). Springer Berlin Heidelberg.

[42] Bhat, Sajid Yousuf, Muhammad Abulaish, and Abdulrahman A. Mirza. "Spammer classification using ensemble methods over structural social network features." *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 02*. IEEE Computer Society, 2014.

[43] Shalaby, Walid, Khalifeh Al Jadda, Mohammed Korayem, and Trey Grainger. "Entity Type Recognition using an Ensemble of Distributional Semantic Models to Enhance Query Understanding." *arXiv preprint arXiv:1604.00933*(2016).

[44] Wu, W., Liu, Z., & He, Y. (2015). Classification of defects with ensemble methods in the automated visual inspection of sewer pipes. *Pattern Analysis and Applications*, *18*(2), 263-276.

[45] van Wezel, Michiel, and Rob Potharst. "Improved customer choice predictions using ensemble methods." *European Journal of Operational Research* 181.1 (2007): 436-452.

[46] Dwyer, Catherine, Starr Hiltz, and Katia Passerini. "Trust and privacy concern within social networking sites: A comparison of Facebook and MySpace." *AMCIS 2007 proceedings* (2007): 339.

[47] Nelson-Field, Karen, Erica Riebe, and Byron Sharp. "What's not to "Like?"." *Journal of Advertising Research* 52.2 (2012): 262-269.

[48] Bugeja, Martin, and Natalie Gallery. "Is older goodwill value relevant?." *Accounting & Finance* 46.4 (2006): 519-535.