ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Lương Thái Lê

BIỂU DIỄN VÀ PHÂN TÍCH DỮ LIỆU TRÊN ĐỒ THỊ LỚN CHO MÔ HÌNH HÓA NGƯỜI DÙNG VÀ HỆ TƯ VẤN

Chuyên ngành: Hệ thống Thông tin

Mã số: 9480104.01

TÓM TẮT LUẬN ÁN TIẾN SĨ HỆ THỐNG THÔNG TIN

Hà Nội - 2019

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Lương Thái Lê

BIỂU DIỄN VÀ PHÂN TÍCH DỮ LIỆU TRÊN ĐỒ THỊ LỚN CHO MÔ HÌNH HÓA NGƯỜI DÙNG VÀ HỆ TƯ VẤN

Chuyên ngành: Hệ thống Thông tin

Mã số: 9480104.01

Cán bộ hướng dẫn chính: **PGS.TS. Phan Xuân Hiếu**

Cán bộ hướng dẫn phụ: **PGS.TS. Trần Văn Long**

TÓM TẮT LUẬN ÁN TIẾN SĨ HỆ THỐNG THÔNG TIN

Hà Nội - 2019

Mục lục

Mở đầu

Phân tích ý định từ các văn bản trực tuyến là một bài toán có nhiều ý nghĩa về cả khoa học và thực tiễn. Một phân tích đầy đủ ý định của người dùng khi nó mới chỉ ở dưới dạng các bài đăng/bình luận trên các phương tiện truyền thông trực tuyến là chìa khóa quan trọng để các doanh nghiệp, các dịch vụ kinh doanh có thể kịp thời nắm bắt được thị hiếu và nhu cầu khách hàng, dự báo tiêu dùng, tìm kiếm khách hàng tiềm năng và định hướng tiếp thị, cung ứng. Về mặt khoa học, phân tích ý định từ văn bản được xếp vào lớp bài toán hiểu ngôn ngữ tự nhiên (natural language understanding - NLU) vốn đòi hỏi các phân tích sâu về ngôn ngữ như phân tích cú pháp, phân tích ngữ nghĩa. Chính vì vậy, từ đầu những năm 2000, các cộng đông nghiên cứu khoa học trên thế giới đã có nhiều công bố về bài toán này. Hầu hết các nghiên cứu ban đầu chủ yếu tập trung theo hướng tiếp cận phân lớp ý định vào một lớp ngữ nghĩa nào đó, điển hình là các nghiên cứu của các nhóm tác giả Broder (2002)[12], Chen (2013)[21], Gupta (2014)[40], Wang (2015)[113]. Bên cạnh đó, một số ít nghiên cứu đề xuất cách tiếp cận hiểu sâu hơn về ngữ nghĩa, nội dung của ý định, điển hình là các nghiên cứu của vác tác giả và cộng sự: Li (2010)[73], Castellanos (2012)[16], Zhang (2017)[120].

Tuy vậy vấn đề phân tích và hiểu ý định từ các văn bản trực tuyến vẫn còn nhiều khía cạnh chưa được khai thác triệt để như: một định nghĩa đặc tả được cấu trúc ý định một cách tổng quát, một quy trình xuyên suốt để hiểu ý định… Đây cũng chính là một trong những thách thức mà luận án cần tiếp cận giải quyết.

Các vấn đề và nội dung nghiên cứu

Phân tích và xác định một cách chính xác, đầy đủ, trọn vẹn ý định của người viết từ văn bản là một vấn đề khó và nhiều thử thách trong lĩnh vực xử lý ngôn ngữ tự nhiên (những khó khăn này sẽ được trình bày chi tiết ở Chương 2 của luận án). Luận án xem những thử thách này là những nhiệm vụ cần giải quyết và vượt qua, từ đó luận án đặt trọng tâm vào việc tiếp cận và giải quyết năm vấn đề quan trọng sau:

1, *Định nghĩa, biểu diễn ý định và tiến trình phân tích ý định*: Việc tìm được một cách định nghĩa ý định sao cho phù hợp với mục tiêu và phạm vi nghiên cứu là rất quan trọng.

2, *Phát hiện sự hiện diện của ý định*: Việc xác định sự tồn tại của ý định trong văn bản là khâu quan trọng cần thực hiện trước khi tiến hành các phân tích cụ thể hơn. Về mặt khoa học, việc phân tích trực tiếp trên tập các văn bản mang ý định sẽ tránh được phần lớn vấn đề về dữ liệu thưa và không cân bằng.

3, *Xác định miền quan tâm của ý định*: Việc xác định trước miền của ý định giúp chúng ta có thể giới hạn những thông tin về ý định cũng như làm giảm sự phong phú về từ vựng, từ đó giúp cho việc phân tích đạt độ chính xác cao hơn.

4, *Xác định thông tin ý định theo tiếp cận phân tích nông*: Các kỹ thuật phân tích sâu như phân tích cú pháp, ngữ nghĩa đối với tiếng Việt còn là vấn đề khó và chưa đạt được độ chính xác mong muốn. Vì thế, luận án đặt vấn đề theo một hướng tiếp cận khác: xác định ý định dựa trên phân tích ngôn ngữ ở mức nông, hay gọi tắt là phân tích nông.

5, *Phân tích và xác định ý định độc lập miền*: Một trong những khía cạnh quan trọng trong xử lý ngôn ngữ tự nhiên nói chung và trong bài toán này nói riêng là vấn đề về miền dữ liệu. Liệu chúng ta có thể phân

tích ý định ở mức độc lập miền? Liệu chúng ta có thể sử dụng dữ liệu và tri thức từ một miền đã có để phân tích trên các miền mới? Một phần quan trọng của luận án sẽ tìm kiếm câu trả lời cho những câu hỏi trên.

Phạm vi và phương pháp nghiên cứu

Trong khuôn khổ luận án này, chúng tôi hạn chế phạm vi và nội dung nghiên cứu của mình ở một số điểm sau:

- *Dạng ý định*: Luận án chỉ quan tâm ý định tường minh hay còn gọi là ý định rõ (explicit intent). Luận án chưa xem xét phân tích các ý định ẩn (implicit intent). Luận án có thể xử lý vấn đề đa ý định trong văn bản nhưng không xử lý trường hợp đa ý định trong một câu hoặc các ý định có tính lồng nhau. Luận án cũng không xem xét khía cạnh về tính hiệu lực của ý định. Nghĩa là một ý định có thể đề cập trong quá khứ và có thể đã hết hiệu lực nhưng vẫn được xem là một ý định hợp lệ.

- *Dạng dữ liệu*: Luận án tập trung phân tích ý định từ các bài đăng, bình luận của người dùng trên các phương tiện truyền thông xã hội trực tuyến. Trong luận án này chúng tôi sử dụng thuật ngữ *văn bản* cho ngắn gọn. Độ dài các văn bản cần từ hai từ trở lên và không dài quá 800 từ.

Nghiên cứu lý thuyết đề xuất mô hình, phương pháp giải quyết các bài toán xác định ý định người dùng từ văn bản cũng như nghiên cứu thực nghiệm để kiểm chứng đánh giá các đề xuất của luận án.

Kết quả và đóng góp chính của luận án

- *Thứ nhất*, luận án đề xuất một định nghĩa về ý định hướng miền quan tâm phù hợp cho văn bản truyền thông xã hội trực tuyến, đồng thời đề xuất tiến trình ba pha gồm ba bài toán phân tích và xác định

thông tin ý định. Trong đó, bài toán một (*lọc ý định*) và bài toán hai (*xác định miền quan tâm*) lần lượt được mô hình hóa thành bài toán phân lớp nhị phân và phân lớp đa lớp. Các nội dung và kết quả nghiên cứu này được trình bày trong công trình [LTLe1], [LTLe2].

- *Thứ hai*, luận án đề xuất mô hình hóa bài toán ba (*trích chọn nội dung của ý định*) dưới dạng trích chọn thông tin trên dữ liệu chuỗi. Các mô hình học máy thống kê cho dữ liệu chuỗi như CRFs, mô hình học sâu Bi-LSTM-CRFs được đề xuất để giải quyết bài toán này. Luận án cũng đề xuất tập nhãn đặc trưng tương ứng những nội dung ý định cần trích xuất trên từng miền dữ liệu. Các nội dung và kết quả này được trình bày trong công trình [LTLe3]. Hơn nữa, luận án đề xuất một phương pháp hiệu quả để nâng cao độ chính xác của bài toán trích chọn nội dung ý định dựa trên các mô hình học kết hợp (ensemble learning) mà cụ thể ở đây là kỹ thuật học bộ ba (tri-training). Nội dung và kết quả nghiên cứu này được trình bày trong [LTLe4].

- *Thứ ba*, luận án đề xuất mô hình phân tích và xác định ý định độc lập miền (domain-independent) dựa trên ý tưởng xây dựng tập nhãn chung cho các miền dữ liệu. Luận án đã tiến hành phân tích thực nghiệm, so sánh, đánh giá hiệu quả của hai cách tiếp cận phụ thuộc miền và độc lập miền cũng như thảo luận về ưu nhược điểm của mỗi cách tiếp cận. Nội dung và kết quả này được trình bày trong công trình [LTLe5].

Cấu trúc luận án

Toàn thể nội dung luận án bao gồm:

- Phần *Mở đầu*, phần này đề cập ý nghĩa và tính cấp thiết của luận án, tổng quan về bối cảnh nghiên cứu, động lực, mục tiêu, phạm vi, nội dung nghiên cứu, cùng những đóng góp chính của luận án.

- Chương 1, *Tổng quan về ý định và phân tích ý định*. Chương này giới thiệu về khái niệm ý định, thể hiện ý định trong văn bản, đồng thời giới thiệu về bài toán phân tích ý định từ văn bản trực tuyến cùng một khảo sát về những nghiên cứu liên quan. Phần cuối của chương nhắc lại sơ lược các kiến thức cơ sở được sử dụng trong luận án.

- Chương 2, *Phân tích ý định từ văn bản trực tuyến*. Chương này đưa ra khái niệm *miền quan tâm* và *ý định hướng miền quan tâm* của luận án. Từ đó phân tích và đề xuất tiến trình ba pha giải quyết bài toán phân tích ý định.

- Chương 3, *Phát hiện ý định và xác định miền quan tâm của ý định*. Chương này đề xuất các phương pháp học máy hiệu quả để giải quyết pha một (tức là bài toán phát hiện ý định), và pha hai (tức là bài toán xác định miền quan tâm của ý định).

- Chương 4, *Trích chọn ý định từ văn bản trực tuyến theo tiếp cận học máy*. Chương này đề xuất việc mô hình hóa pha ba của tiến trình ba pha về bài toán trích chọn thông tin trên dữ liệu chuỗi. Sau đó, lần lượt tiếp cận giải quyết bài toán nhờ phương pháp CRFs và Bi-LSTM-CRFs. Chương này cũng đề xuất một phương pháp hiệu quả dựa vào kỹ thuật học kết hợp để nâng cao độ chính xác của bài toán trích chọn ý định.

- Chương 5, *Thích nghi miền trong xác định ý định người dùng*. Chương này trình bày phương pháp trích chọn ý định độc lập miền dựa vào một bộ nhãn tổng quát do luận án đề xuất. Phần cuối của chương đưa ra những nhận định về ưu nhược điểm của bộ nhãn chung và bộ nhãn riêng.

- Phần *Kết luận*, phần này tổng hợp các kết quả chính mà luận án đóng góp.

Chương 1.

Tổng quan về ý định và phân tích ý định

1.1 Ý định và thể hiện ý định trong ngôn ngữ

Có rất nhiều quan điểm về định nghĩa "ý định" trên thế giới. Theo Bratman (1987) [13], "*ý định là một trạng thái tinh thần thể hiện sự cam kết thực hiện một hay nhiều hành động trong tương lai*". Hay theo Scheer (2004) [100], "*ý định là một hướng hành động được ai đó lựa chọn*". Trong đó, với cách định nghĩa của Scheer thì không cần có sự *cam kết* đối với ý định.

Có nhiều cách để thể hiện ý định: qua cử chỉ, hành động, lời nói, văn bản…

1.2 Phân tích và hiểu ý định: bối cảnh khoa học

Phân tích và hiểu ý định từ văn bản trực tuyến gồm một số hướng nghiên cứu chính sau:

1.2.1 *Phân tích và xác định ý định từ truy vấn tìm kiếm*.

Các truy vấn tìm kiếm thường là các văn bản rất ngắn, đa dạng, đa nghĩa và nhập nhằng. Điển hình cho hướng nghiên cứu này là những nghiên cứu của Broder(2002)[12], Dai (2006)[26], Hu (2009)[49], Li (2010)[73].

1.2.2 *Phân tích ý định trong ngôn ngữ nói*.

Ngôn ngữ nói ở đây chỉ các câu nói trong các đoạn hội thoại giữa người dùng với nhau trên các phương tiện truyền thông xã hội, hoặc giữa người dùng với một hệ thống hội thoại tự động nào đó. Các nghiên cứu điển hình theo hướng này là Kimura(1998)[63], K.Yao(2015)[116], Kim (2016)[62].

1.2.3 *Phân tích ý định trong các bài đăng*.

Một bài đăng (post/comment/tweet) trên các phương tiện truyền thông xã hội trực tuyến thường dài hơn và mang nhiều nội dung thông tin hơn các truy vấn. Điển hình cho hướng nghiên cứu này là các công bố của Castellanos (2012)[16], Chen (2013)[21], Wang (2015)[113], Ngo (2017)[84].

1.3 Một số kỹ thuật khai phá dữ liệu và mô hình học máy

Phần này giới thiệu cơ bản về một số kiến thức cơ bản liên quan đến luận án như kỹ thuật phân lớp, kỹ thuật trích xuất thông tin, mạng nơ ron.

1.4 Kết luận chương

Chương này giới thiệu về khái niệm ý định và thể hiện ý định trong văn bản. Bên cạnh đó, một khảo sát về các hướng nghiên cứu liên quan và các cách tiếp cận giải quyết bài toán xác định ý định người dùng trên thế giới cũng được trình bày trong chương này.

Chương 2.

Phân tích ý định từ văn bản trực tuyến

2.1 Phân tích ý định từ văn bản xã hội trực tuyến tiếng Việt

Luận án hướng tới mục tiêu xây dựng một quá trình xuyên suốt để phân tích và hiểu ý định người dùng từ các văn bản tiếng Việt, tức là các bài đăng (posts) và các bình luận (comments), trên các phương tiện truyền thông xã hội trực tuyến.

Phần này cũng đề ra các khó khăn của bài toán và các vấn đề nghiên cứu chính của luận án.

1) *Sự đa dạng của ý định*: Sự đa dạng về lĩnh vực, về đặc điểm của ý định tạo nên sự phân bố rộng khắp về mặt từ vựng lẫn nội dung thông tin.

2) *Đa ý định*: Một bài đăng của người dùng có thể chứa nhiều hơn một ý định và các ý định này lại thuộc những lĩnh vực khác nhau.

3) *Tính nhập nhằng*: người viết có ý định "bán hoa quả" nhưng mô hình có thể xác định nhầm thành ý định "mua".

4) *Ý định ẩn*: người viết không đề cập một cách tường minh nhu cầu hay mục tiêu hành động cụ thể mà để người đọc tự suy diễn.

5) *Sự phong phú của ngôn ngữ văn bản truyền thông trực tuyến*: có thể chứa từ địa phương, tiếng lóng, từ viết tắt, ngôn ngữ ``teen'', và đặc biệt có nhiều lỗi chính tả lẫn ngữ pháp.

6) *Dữ liệu thưa và không cân bằng*: ý định của người viết nếu có thường chỉ thể hiện trong một vài câu nằm rải rác trong văn bản. Hầu hết các câu còn lại không mang ý định.

7) *Tính hiệu lực của ý định*: Có những bài đăng chứa ý định nhưng rất khó xác định được ý định đó còn hiệu lực hay đã là quá khứ.

8) *Sự hạn chế về dữ liệu thực nghiệm*: chưa có bất cứ một tập dữ liệu chuẩn nào cho văn bản tiếng Việt đối với bài toán phân tích và xác định ý định. Đây là trở ngại không nhỏ trong quá trình nghiên cứu và thực hiện luận án.

2.2 Định nghĩa ý định người dùng: bối cảnh khoa học

2.2.1 Định nghĩa ý định người dùng theo tiếp cận từ điển

2.2.2 Định nghĩa ý định người dùng theo hướng cấu trúc

2.3 Định nghĩa ý định hướng miền quan tâm

2.3.1 Định nghĩa về quan điểm của BingLiu

2.3.2 Định nghĩa ý định của BingLiu

Ý định là một cấu trúc gồm 5 thành phần bao gồm *hành động ý định* (intended-action), *đích của ý định* (intention-target), *độ mạnh của ý định* (intention-intensity), *chủ thể của ý định* (holder), và *thời điểm phát biểu ý định* (time).

2.3.3 Định nghĩa ý định hướng miền quan tâm của luận án

Luận án đề xuất ý định rõ hướng miền quan tâm là một bộ năm

$$I_u^e = (u, \boldsymbol{c}, d, w, \boldsymbol{p}) \qquad (1.1)$$

trong đó:

- u là thành phần xác định người dùng như nickname, id
- **c** là thành phần chỉ ngữ cảnh, tức là hoàn cảnh hay tình huống liên quan ảnh hưởng đến ý định như: người dùng đang có thai, vừa mới kết hôn, có con nhỏ, đang bị ngân hàng siết nợ…
- d là thành phần chỉ miền quan tâm của ý định, ví dụ miền *Bất động sản, Du lịch, Tài chính*…
- **p** là danh sách các thuộc tính, thông tin liên quan đến ý định. Nó có thể được biểu diễn bởi một danh sách các bộ đôi *thuộc tính – giá trị*. Ví dụ **p** có thể là {*địa điểm* = "373 đường Trần Xuân Soạn", *diện tích* = "80m2", *giá* = "3.5 tỷ"…}

2.4 Tiến trình ba pha phân tích và xác định ý định

Luận án đề xuất chiến lược giải quyết bài toán hiểu ý định người dùng gồm ba pha chính. Ba pha đó lần lượt là:

(1). *Lọc bài đăng mang ý định người dùng* (User intent filtering): Pha này sẽ giúp phát hiện và lấy về những văn bản mang ý định rõ của người dùng từ vô vàn những văn bản trên các phương tiện truyền thông xã hội trực tuyến. Pha này sẽ giúp xác định thành phần "u"

(2). *Xác định miền quan tâm của ý định* (User intent domain and category identification): với một văn bản mang ý định của người dùng, pha này sẽ xác định xem ý định đó thuộc lĩnh vực nào (*Bất động sản*, *Tài chính,* hay *Du lịch*…). Pha này giúp xác định thành phần "d".

(3). *Phân tích và trích xuất ý định* (User intent parsing and extraction): với đầu vào là một đoạn văn bản trực tuyến mang ý định người dùng và lĩnh vực của ý định đó, pha này giúp phân tích và trích xuất tất cả những thông tin cần thiết liên quan đến ý định người dùng. Pha này giúp xác định các thành phần: "**c**", "w", "**p**".

2.5 Tiền xử lý dữ liệu

Dữ liệu sau khi thu thập được làm sạch với các thao tác: bỏ các biểu tượng cảm xúc, các ký tự lạ, tách các dấu câu thành các từ…

2.6 Kết luận chương

Chương này trình bày định nghĩa ý định hướng miền quan tâm mà luận án đề xuất, đồng thời đề xuất tiến trình ba pha giải quyết bài toán phân tích và hiểu ý định người dùng. Những đề xuất này đã được công bố trong công trình [LTLe1].

Chương 3.

Phát hiện ý định và xác định miền quan tâm của ý định

3.1 Giới thiệu

Chương này tập trung giải quyết pha 1 và pha 2 trong tiến trình ba pha đề xuất ở chương 1.

3.2 Nghiên cứu liên quan

3.2.1 Phát hện bài đăng trực tuyến mang ý định:

Một số nghiên cứu điển hình theo hướng tiếp cận này là các nghiên cứu của: Chen (2013) [21], Gupta (2014) [40], Ngo (2017) [84].

3.2.2 Xác định miền quan tâm của ý định:

Một số nghiên cứu điển hình theo hướng tiếp cận này là của các nhóm tác giả sau: Wang (2015) [113], Hashemi (2016) [43].

3.3 Phát hiện ý định

3.3.1 Phát biểu bài toán:

Xây dựng mô hình để xác định một văn bản trực tuyến tiếng Việt (bài đăng/bình luận trên các phương tiện truyền thông xã hội) có chứa ý định rõ của người dùng hay không

3.3.2 Mô hình thực nghiệm:

- Sử dụng thuật toán cực đại hóa entropy (ME) phân lớp nhị phân các văn bản trực tuyến tiếng Việt vào 2 lớp: *Mang ý định rõ* (EI); *Không mang ý định* (NI)
- Sử dụng hai loại đặc trưng, đó là n-grams (1-gram, 2-gram, 3-gram) và từ điển chỉ mục (chứa các cụm từ như *muốn mua*, *tính*, *dự định*, *cần tìm*…).

3.3.3 Dữ liệu thực nghiệm:

Dữ liệu gồm 1315 văn bản trực tuyến được lấy từ các nguồn như: webtretho.com, lamchame.com, facebook.com…, sau đó được gán nhãn dưới sự đồng thuận của một nhóm 5 sinh viên và thu được 588 văn bản có nhãn EI và 727 văn bản có nhãn NI.

3.3.4 Thiết kế thực nghiệm

Thực nghiệm theo phương pháp đánh giá chéo 4-fold với lần lượt từng loại đặc trưng.

3.3.5 Kết quả thực nghiệm

Kết quả thực nghiệm cho thấy độ chính xác F1 khá cao và ổn định đối với tất cả các fold (đều hơn 88%). Điều đó chứng tỏ mô hình và các đặc trưng mà chúng tôi đề xuất phù hợp để giải quyết bài toán đặt ra. Fold 4 đạt độ chính xác cao nhất với độ chính xác trung bình mịn của F1 là 92.07%, trong đó lớp NI và lớp EI lần lượt đạt độ chính xác F1 là 92.9% và 91.03%.

3.4 Xác định miền quan tâm của ý định

3.4.1 Phát biểu bài toán

Cho văn bản trực tuyến tiếng Việt (bài đăng/bình luận tiếng Việt trên các phương tiện truyền thông xã hội) chứa ý định rõ của người dùng. Hãy xây dựng mô hình xác định miền quan tâm của ý định đó

3.4.2 Mô hình thực nghiệm

- Luận án đề xuất mô hình hóa bài toán xác định miền ý định về bài toán phân lớp đa lớp (13 lớp như bên dưới) và đề xuất sử dụng hai mô hình phân lớp là cực đại hóa entropy (ME) và máy hỗ trợ véc tơ (SVMs) để tiến hành thực nghiệm.
- Sử dụng 2 loại đặc trưng là n-grams và từ điển chỉ mục. Từ điển chỉ mục được tạo tự động bằng cách lựa chọn 10-30 n-grams có đặc trưng cao nhất cho mỗi miền ý định.

3.4.3 Dữ liệu thực nghiệm

Dữ liệu gồm 7009 văn bản mang ý định rõ thu được từ các diễn đàn nổi tiếng và các trang facebook công khai. Luận án đề xuất xây dựng một phân hoạch gồm 13 lớp miền quan tâm. Sau khi thực hiện gán nhãn thu được số lượng bài đăng tương đương với mỗi miền như sau: *Thiết bị điện tử* (546), *Thời trang & Phụ kiện* (586), *Tài chính* (314), *Dịch vụ ăn uống* (424), *Nội thất & Tạp hóa* (699), *Sức khỏe & Làm đẹp* (322), *Công việc & Giáo dục* (1296), *Vật nuôi & Cây trồng* (385), *Bất động sản* (750), *Thể thao & Giải trí* (456), *Giao thông Vận tải* (649), *Du lịch & Khách sạn* (354), *Khác* (228).

3.4.4 Thiết kế thực nghiệm

Dữ liệu được chia thành 5 phần với tỉ lệ 4 train :1 test. Sau đó chúng tôi tiến hành thực nghiệm đánh giá chéo 5-fold với mỗi mô hình.

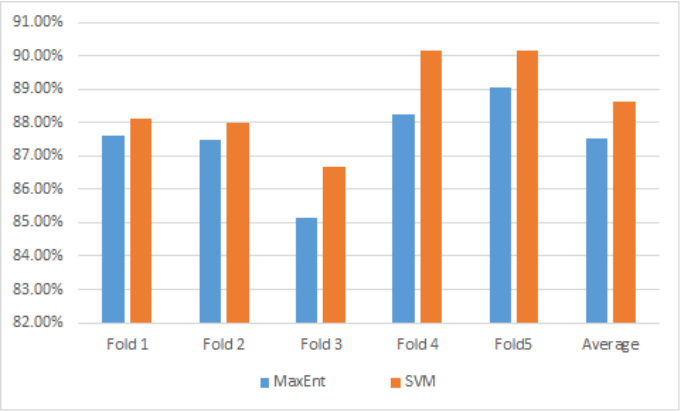
3.4.5 Kết quả thực nghiệm

Hình 3.3 thể hiện kết quả trung bình F1 của mỗi fold khi thực nghiệm lần lượt với mô hình ME và mô hình SVMs. Có thể thấy kết quả thực nghiệm khá ổn định trên cả 5 fold và đều đạt độ chính xác F1 trên 85%. Đặc biệt, mô hình SVMs luôn đạt độ chính xác cao hơn mô hình ME trong mọi thực nghiệm. Kết quả độ chính xác F1 đối với từng lớp miền ý định tương ứng với fold tốt nhất được trình bày trong hình 3.4. Độ chính xác F1 của mỗi lớp hầu hết đều cao hơn 80%, trừ lớp *Khác*. Một số lý do có thể lý giải cho kết quả đó là: (i) lớp *Khác* có số lượng bài đăng ít nhất; (ii) các bài đăng thuộc lớp K*hác* rất đa dạng nên khó tìm được đặc trưng riêng phân biệt tốt.

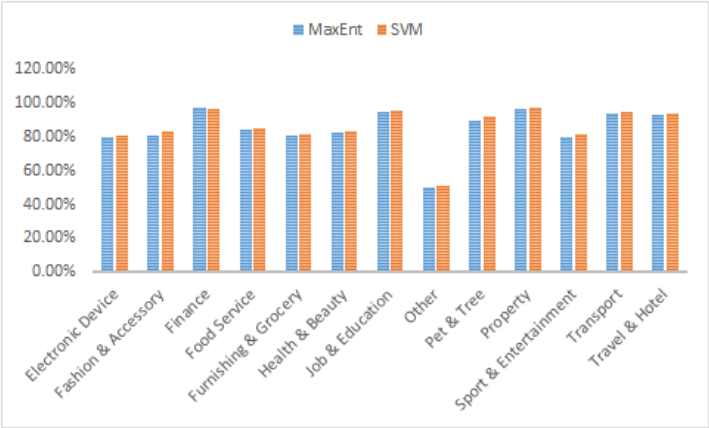
3.5 Kết luận chương

Chương 2 đề xuất mô hình hóa bài toán *Phát hiện ý định* về bài toán phân lớp nhị phân, và bài toán *Xác định miền quan tâm của ý định* về bài toán phân lớp đa lớp. Hai phương pháp phân lớp đơn giản

nhưng hiệu quả được đề xuất để tiến hành thực nghiệm cho hai bài toán trong chương này là ME và SVM. Kết quả của các thực nghiệm chứng tỏ phương pháp đề xuất của luận án phù hợp và hiệu quả. Nội dung và kết quả nghiên cứu của chương 2 được công bố trong [LTLe1] và [LTLe2].



Hình 3.3 Độ chính xác F_1 khi đánh giá chéo 5-fold với ME và SVMs



Hình 3.4 Độ chính xác trung bình F_1 đối với mỗi miền quan tâm của ý định

Chương 4.

Phân tích và trích chọn nội dung ý định

4.1 Giới thiệu

Chương này tập trung giải quyết pha 3 (trích chọn nội dung ý định) của tiến trình ba pha theo tiếp cận học máy và học sâu. Luận án lựa chọn hai miền ý định là *Bất động sản* và *Mỹ phẩm & Làm đẹp* để thực nghiệm. Đặc biệt chương này đề xuất một phương pháp hiệu quả để nâng cao độ chính xác của bài toán trích chọn ý định nhờ sử dụng kỹ thuật học kết hợp các mô hình học sâu.

4.2 Nghiên cứu liên quan

4.2.1 Trích chọn ý định

Một số nghiên cứu điển hình liên quan đến bài toán phân tích và trích chọn nội dung ý định là Li (2010) [73], Castellanos (2012) [16], Hamroun (2015) [42].

4.2.2 Kỹ thuật huấn luyện bộ ba (tri-training)

4.2.3 Phương pháp học kết hợp (ensemble learning)

4.3 Phát biểu bài toán

Cho văn bản trực tuyến tiếng Việt mang ý định rõ thuộc miền quan tâm "d" đã được xác định trước. Hãy xây dựng mô hình trích chọn những thông tin quan trọng về ý định đó

4.4 Trích chọn ý định theo tiếp cận học máy thống kê và học sâu

4.4.1 Xây dựng bộ nhãn thực nghiệm

Luận án đề xuất mô hình hóa bài toán trích chọn ý định về bài toán xác định thực thể được nhắc đến (EMD – entity mentioned detection). Vì vậy, đầu tiên chúng tôi cần xây dựng bộ nhãn tương ứng với các thực thể cần được trích chọn. Luận án đề xuất bộ nhãn gồm 13 nhãn

cho miền *Bất động sản* (bảng 4.1) và bộ nhãn gồm 9 nhãn cho miền *Mỹ phẩm & Làm đẹp* (bảng 4.2).

4.4.2 Trích chọn ý định với phương pháp CRFs

Với bài toán trích chọn ý định và dữ liệu thu được, luận án đề xuất sử dụng 3 loại đặc trưng cho mô hình CRFs: n-gram; biểu thức chính quy; từ điển chỉ mục.

4.4.3 Trích chọn ý định với phương pháp học sâu Bi-LSTM

Luận án kế thừa mô hình Bi-LSTM-CRFs được đề xuất bởi Lample và cộng sự (2016) [68]. Chúng tôi sử dụng kỹ thuật FastText để tạo véc tơ mã hóa từ cho đầu vào của mô hình, mỗi véc tơ có kích thước 100. Để thực nghiệm cho bài toán trích chọn ý, chúng tôi sử dụng một số kỹ thuật kết hợp với mô hình Bi-LSTM-CRFs. Thứ nhất là kỹ thuật *biểu diễn từ dựa vào mã hóa ký tự (Character-based Embedding)*, được ký hiệu là "Char". Với kỹ thuật này chúng tôi tạo ra véc tơ biểu diễn từ dựa vào ký tự với kích thước 25. Thứ hai là kỹ thuật *Tiền huấn luyện (Pre-trained)*, được ký hiệu là "Pre". Với kỹ thuật này, chúng tôi sử dụng phương pháp Skip-gram để tạo véc tơ biểu diễn từ cho bảng tham chiếu (look-up table). Thứ ba là kỹ thuật *Cắt tỉa (Dropout)*, được ký hiệu là "Drop". Kỹ thuật này được sử dụng để làm giảm hiện tượng *quá khớp* của mô hình với dữ liệu huấn luyện bằng cách bỏ đi ngẫu nhiên một số *đơn vị (unit)* theo một tỉ lệ cho trước. Trong thực nghiệm của mình, chúng tôi sử dụng tỉ lệ *cắt tỉa* là p = 0.3

4.4.4 Độ đo đánh giá mô hình thực nghiệm

Sử dụng độ chính xác (precision), độ hồi tưởng (recall), và độ đo F_1 được tính theo mức chunk-based (cụm từ được phân đoạn)

4.4.5 Dữ liệu thực nghiệm

Dữ liệu thực nghiệm được thu thập chủ yếu từ các diễn đàn và facebook. Chúng tôi thu được 712 văn bản cho lĩnh vực *Bất động sản*

và 1500 văn bản cho lĩnh vực *Mỹ phẩm & Làm đẹp*. Sau đó dữ liệu được gán nhãn theo hướng dẫn trong bảng 3.1 và bảng 3.2 trong luận án. Cuối cùng dữ liệu được chuyển sang chuẩn BIO để làm đầu vào cho các mô hình học máy. Với mô hình Bi-LSTM-CRFs, dữ liệu được chia theo tỷ lệ 3:1:1 (train:validation:test); còn với mô hình CRFs, dữ liệu được chia theo tỷ lệ 3:1 (train:test).

4.4.6 Thiết kế thực nghiệm

Với mỗi miền ý định, luận án lần lượt thực nghiệm 5 mô hình sau: (i) LSTM-CRF(Char): huấn luyện mô hình Bi-LSTM-CRFs kết hợp với kỹ thuật CHAR; (ii) LSTM-CRF(Char + Drop): huấn luyện mô hình Bi-LSTM-CRFs kết hợp với các kỹ thuật Char và Drop; (iii) LSTM-CRF(Char + Pre): huấn luyện mô hình Bi-LSTM-CRFs kết hợp với các kỹ thuật Char và Pre; (iv) LSTM-CRF(Char + Pre + Drop): huấn luyện mô hình Bi-LSTM-CRFs kết hợp với các kỹ thuật Char, Pre và Drop; (v) CRFs: huấn luyện mô hình CRFs với các đặc trưng đã xây dựng.

4.4.7 Kết quả thực nghiệm

Bảng 4.6 và 4.7 lần lượt thể hiện kết quả thực nghiệm lần lượt 5 mô hình trên miền *Mỹ phẩm & Làm đẹp* và miền *Bất động sản*. Mỗi miền ý định đạt độ chính xác cao nhất với một mô hình khác nhau. Điều này có thể được lý giải bởi sự khác nhau về đặc trưng dữ liệu từng miền, hơn nữa, miền *Bất động sản* có ít ví dụ thực nghiệm nên việc sử dụng kỹ thuật Pre có thể không hiệu quả.

	Precision	**Recall**	**F1**
LSTM-CRF (Char)	90.99%	87.19%	89.01%
LSTM-CRF (Char+Drop)	92.08%	89.37%	90.71%
LSTM-CRF (Char+Pre)	90.14%	89.25%	89.69%
LSTM-CRF (Char+Pre+Drop)	92.79%	89.60%	**91.17%**
CRFs	92.15%	73.49%	81.76%

Bảng 4.6 Trung bình F1-score với mỗi mô hình thực nghiệm thuộc lĩnh vực Mỹ phẩm & Làm đẹp

	Precision	Recall	F1
LSTM-CRF (Char)	91.94%	90.83%	**91.37%**
LSTM-CRF (Char+Drop)	90.39%	89.37%	89.87%
LSTM-CRF (Char+Pre)	89.98%	89.90%	89.94%
LSTM-CRF (Char+Pre+Drop)	90.23%	89.00%	89.53%
CRFs	87.21%	85.68%	86.43%

Bảng 4.7 Trung bình F1-score với mỗi mô hình thực nghiệm thuộc lĩnh vực Bất động sản

4.5 Trích chọn ý định dựa trên kết hợp các mô hình học sâu

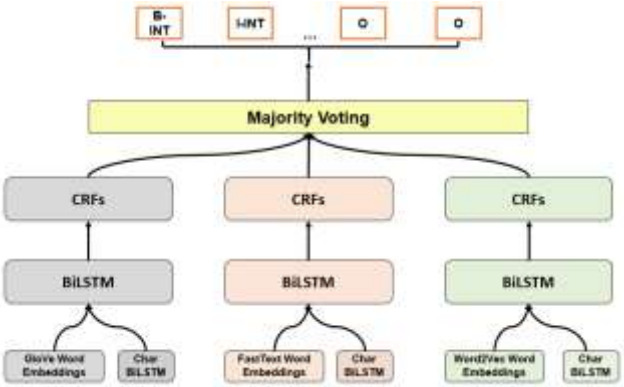
4.5.1 Xây dựng bộ nhãn thực nghiệm

Để tiến hành thực nghiệm mô hình này, chúng tôi lựa chọn 3 miền ý định là *Bất động sản, Du lịch* và *Xe cộ* để thu thập dữ liệu và trích chọn ý định. Chúng tôi lần lượt đề xuất 3 bộ nhãn tương ứng với 3 bộ thông tin ý định cần trích chọn, trong đó bộ nhãn *Bất động sản* gồm 18 nhãn, bộ nhãn *Du lịch* gồm 15 nhãn, và bộ nhãn *Xe cộ* gồm 17 nhãn. Các bộ nhãn này lần lượt được trình bày trong các bảng 5.1, 5.2, 5.3 trong luận án.

4.5.2 Mô hình thực nghiệm

4.5.2.1 Mô hình học kết hợp không chia sẻ tài nguyên

Luận án đề xuất một mô hình học kết hợp ba thành phần học sâu để nâng cao hiệu quả bài toán trích chọn thông tin ý định. Trong mô hình này, mỗi thành phần học sâu là một mô hình Bi-LSTM-CRFs được khởi tạo với các kỹ thuật biểu diễn từ khác nhau: GloVe, FastText, Word2Vec. Biểu diễn đầu vào khác nhau này đảm bảo sự đa dạng cần thiết của 3 mô hình học kết hợp theo kỹ thuật tri-training. Kết quả đoán nhận cuối cùng của mô hình học kết hợp nhận được bởi một thủ tục *bình chọn theo đa số (majority voting)* dựa trên các kết quả của các thành phần con. Trường hợp hệ cân bằng, tức là giá trị nhãn thu được từ 3 thành phần khác nhau từng đôi một thì nhãn có kết quả Viterbi cao nhất khi huấn luyện mô hình CRFs sẽ được chọn làm kết quả đoán nhận của mô hình cuối. Mô hình này được trình bày trong hình 4.14.



Hình 4.14 Mô hình trích chọn ý định dựa trên kết hợp các mô hình học sâu (Mô hình luận án đề xuất)

4.5.2.2 Mô hình chia sẻ tài nguyên

Để làm giảm thời gian huấn luyện mô hình, luận án đề xuất một mô hình học kết hợp chia sẻ tầng *Biểu diễn từ dựa vào ký tự*. Mô hình này được trình bày trong hình 4.16 trong luận án.

4.5.3 Dữ liệu thực nghiệm

Dữ liệu được lấy chủ yếu từ các diễn đàn và facebook nổi tiếng ở Việt Nam. Chúng tôi thu được khoảng 9000 văn bản, trong đó mỗi miền ý định có khoảng 3000 văn bản. Dữ liệu sau đó được chia theo tỉ lệ 3:1:1 (train:validation:test) để tiến hành thực nghiệm.

4.5.4 Thiết kế thực nghiệm

Chúng tôi lần lượt thực nghiệm 6 mô hình trên mỗi miền ý định: (i) Mô hình đề xuất không chia sẻ tài nguyên (OUR PROPOSED ENSEMBLE MODEL); (ii) Mô hình đề xuất có chia sẻ tài nguyên (SHARING CHAR LAYER MODEL); (iii) Mô hình Bi-LSTM-CRFs đơn có biểu diễn từ được kết nối từ 3 véc tơ biểu diễ từ được tạo bởi lần lượt 3 kỹ thuật Glove, FastText, Word2Vec (3-EMBEDDING); (iv),(v),(vi) lần lượt là các mô hình Bi-LSTM-CRFs đơn có đầu vào là

véc tơ được biểu diễn bởi lần lượt các kỹ thuật Glove, FastText, Word2Vec.

4.5.5 Kết quả thực nghiệm

Mô hình đề xuất của chúng tôi đã nâng kết quả trích chọn ý định lên khá cao so với các mô hình đơn. Điển hình, trong miền *Xe cộ*, mô hình đề xuất có độ chính xác cao hơn mô hình đơn WORD2VEC gần 3%, hình 4.19. Mô hình chia sẻ tài nguyên mà chúng tôi đề xuất tuy có độ chính xác thấp hơn mô hình không chia sẻ tài nguyên nhưng vẫn cao hơn tất cả các mô hình đơn.

Hình 4.20 Trung bình F1 qua 5 lần chạy khác nhau của mỗi mô hình đối với miền xe cộ (transportation)

4.6 Kết luận chương

Chương 3 đề xuất mô hình hóa bài toán trích chọn thông tin ý định về bài toán xác định thực thể được nhắc đến, đồng thời đề xuất sử dụng phương pháp học máy CRFs và mô hình học sâu Bi-LSTM-CRFs để giải quyết bài toán. Đặc biệt, chương này đề xuất một mô hình hiệu quả để nâng cao độ chính xác của bài toán trích chọn thông tin, đó là mô hình học kết hợp ba thành phần học sâu. Các kết quả nghiên cứu này được trình bày trong [LTLe3] và [LTLe4].

Chương 5.

Phân tích và trích chọn ý định độc lập miền

5.1 Giới thiệu

Chương này đề xuất một cách tiếp cận không phụ thuộc vào miền ý định cho bài toán phân tích ý định.

5.2. Nghiên cứu liên quan

Một số ít nghiên cứu về phân tích ý định dựa trên kỹ thuật học thích nghi miền điển hình là nghiên cứu của Chen (2013) [21], Ding (2015) [30] và Ngo (2017) [84].

5.3 Trích xuất ý định theo tiếp cận độc lập miền

5.3.1 Phát biểu bài toán

Cho văn bản trực tuyến tiếng Việt mang ý định rõ thuộc một miền quan tâm bất kỳ chưa được xác định trước. Hãy xây dựng mô hình trích chọn những thông tin quan trọng về ý định đó

5.3.2 Xây dựng bộ nhãn độc lập miền

Dựa vào 3 bộ nhãn của 3 miền ý định *Bất động sản, Du lịch, Xe cộ,* luận án đề xuất bộ nhãn chung gồm 10 nhãn : *intent, brand, contact, context, description, location, number of object, object, other, price*. Sự tương quan giữa bộ nhãn chung và 3 bộ nhãn riêng được trình bày trong bảng 5.4 trong luận án. Ngoài ra, chúng tôi cũng đã thử sử dụng bộ nhãn này để trích chọn thông tin với một số miền ý định khác nữa, kết quả cho thấy bộ nhãn chung là phù hợp.

5.3.3 Mô hình trích xuất ý định độc lập miền

Chúng tôi lần lượt xây dựng các mô hình CRFs, Bi-LSTM, Bi-LSTM-CRFs để thực nghiệm trích xuất ý định độc lập miền. Trong đó, mô hình CRFs sử dụng các đặc trưng: n-grams, biểu thức chính quy, từ điển chỉ mục, gán nhãn từ loại, cấu tạo từ (chứa chữ số, chữ

cái đầu viết hoa). Hai mô hình còn lại dùng bộ tham số chung với: kích thước véc tơ biểu diễn từ là 100, phương pháp tối ưu Adam, tỉ lệ *cắt tỉa* là p = 0.5.

5.3.4 Dữ liệu thực nghiệm

Dùng bộ dữ liệu huấn luyện mô hình học kết hợp đề xuất trong chương 3. Dữ liệu được gán nhãn bằng cả nhãn chung và nhãn riêng.

5.3.5 Thiết kế thực nghiệm

Luận án tiến hành 42 thực nghiệm với lần lượt 3 mô hình CRFs, Bi-LSTM, Bi-LSTM-CRFs bao gồm:

- Thực nghiệm với bộ nhãn độc lập miền và bộ nhãn riêng trên mỗi một miền ý định cụ thể riêng biệt.
- Thực nghiệm với bộ nhãn độc lập miền và bộ nhãn riêng trên mỗi tổ hợp 2 trong số 3 miền ý định.
- Thực nghiệm với bộ nhãn độc lập miền và bộ nhãn riêng trên tổ hợp của cả 3 miền ý định.

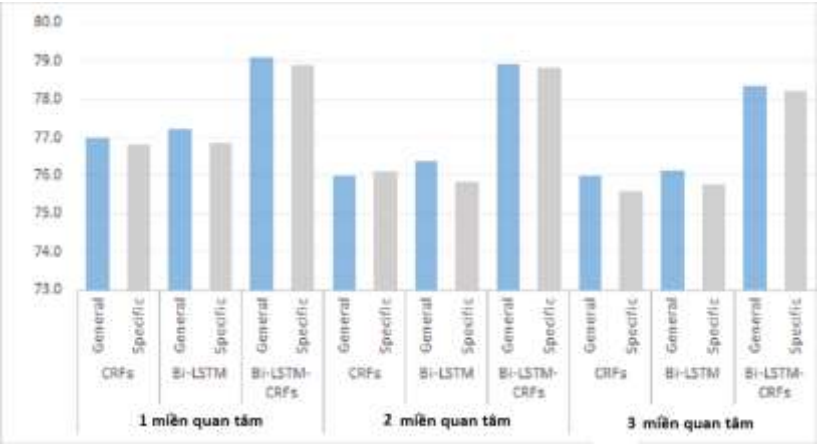
5.3.6 Kết quả thực nghiệm

Kết quả thực nghiệm cho thấy khi số miền quan tâm tăng lên thì bộ nhãn độc lập miền thể hiện tốt hơn bộ nhãn riêng trong việc trích chọn ý định người dùng đối với cả 3 mô hình. Điều đó được thể hiện trong hình 5.3.

Kết quả thực nghiệm cũng cho thấy rằng, nếu cần trích chọn ý định trên một miền ý định cụ thể thì việc dùng bộ nhãn riêng lại tốt hơn bộ nhãn chung. Lý do vì độ chính xác đạt được của bộ nhãn riêng khi đó chỉ thấp hơn bộ nhãn độc lập miền một chút nhưng hiệu quả chính xuất thông tin lại đầy đủ và chi tiết hơn.

5.3.7 Mô phỏng mô hình trích xuất ý định độc lập miền

Luận án đã xây dựng một trang web ở địa chỉ *www.ydinhviet.com* để mô phỏng mô hình trích chọn ý định độc lập miền



Hình 5.3 Kết quả F1 trung bình khi áp dụng các mô hình CRFs, Bi-LSTM, Bi-LSTM-CRFs lần lượt trên 1, 2 và 3 miền quan tâm với bộ nhãn độc lập miền (General) và bộ nhãn riêng (Specific) tương ứng.

5.4 Kết luận chương

Trong chương 5, luận án đề xuất một mô hình trích chọn ý định độc lập miền dựa trên ý tưởng xây dựng bộ nhãn độc lập miền. Nội dung và kết quả nghiên cứu của chương này được trình bày trong công trình [LTLe5].

Kết luận

Như đã đề cập xuyên suốt trong luận án, phân tích và xác định ý định từ văn bản là bài toán khó trong lĩnh vực khai phá văn bản và xử lý ngôn ngữ tự nhiên. Đã có những nghiên cứu tiếp cận bài toán này ở các góc độ khác nhau và phạm vi khác nhau. Luận án này đã trình bày những đề xuất về việc mô hình hoá và giải quyết các vấn đề xoay quanh bài toán phát hiện và phân tích, xác định nội dung ý định từ văn bản truyền thông xã hội trực tuyến tiếng Việt. Tựu trung lại, luận án đạt được những kết quả và đóng góp chính như sau:

Thứ nhất, luận án đề xuất một định nghĩa về ý định rõ hướng miền quan tâm phù hợp cho văn bản truyền thông xã hội trực tuyến đồng thời đề xuất tiến trình ba pha gồm ba bài toán phân tích và xác định thông tin ý định. Trong đó, bài toán một (lọc ý định) và bài toán hai (xác định miền ý định) lần lượt được mô hình hóa thành bài toán phân lớp nhị phân và phân lớp đa lớp. Các nội dung và kết quả nghiên cứu này được trình bày trong công trình [LTLe1], [LTLe2].

Thứ hai, luận án đề xuất mô hình hóa bài toán ba (trích chọn thông tin cụ thể của ý định) dưới dạng trích chọn thông tin trên dữ liệu chuỗi. Các mô hình học máy thống kê cho dữ liệu chuỗi như CRFs, mô hình học sâu Bi-LSTM-CRFs được đề xuất để giải quyết bài toán này. Luận án cũng đề xuất tập nhãn đặc trưng tương ứng những nội dung ý định cần trích xuất trên từng miền dữ liệu. Các nội dung và kết quả này được trình bày trong công trình [LTLe3]. Hơn nữa, luận án đề xuất một phương pháp hiệu quả để nâng cao độ chính xác của bài toán trích chọn thông tin ý định dựa trên các mô hình học kết hợp (ensemble learning) và kỹ thuật huấn luyện bộ ba (tri-training). Nội dung và kết quả nghiên cứu này được trình bày trong công trình [LTLe4].

Thứ ba, luận án đề xuất tiếp cận việc phân tích và xác định ý định độc lập miền (domain-independent) dựa trên ý tưởng xây dựng tập nhãn chung cho các miền dữ liệu. Luận án đã tiến hành phân tích thực nghiệm, so sánh, đánh giá hiệu quả của hai cách tiếp cận phụ thuộc miền và độc lập miền cũng như thảo luận về ưu nhược điểm của mỗi cách tiếp cận. Nội dung và kết quả này được trình bày trong công trình [LTLe5].

Bên cạnh đó, luận án cũng cung cấp một khảo sát tổng quan về hướng nghiên cứu phân tích và xác định ý định từ văn bản. Có thể nói các đóng góp của luận án có ý nghĩa trong việc bổ sung và hoàn thiện các kết quả nghên cứu về phân tích ý định trên thế giới và đặc biệt là cho tiếng Việt. Các kết quả của luận án đã công bố trong các công trình khoa học được đăng tải trên các tạp chí, hội nghị chuyên ngành trong nước và quốc tế có phản biện.

Mặc dù luận án đã đạt được một số kết quả nghiên cứu tích cực, nhưng vẫn còn tồn tại những hạn chế chưa giải quyết được như: (i) luận án mới chỉ sử dụng hai loại đặc trưng với bài toán phân lớp ở pha thứ nhất và pha thứ hai trong khi có rất nhiều loại đặc trưng hiệu quả khác chưa được khai thác; (ii) luận án chưa xử lý được trường hợp bài đăng mang ý định ẩn; (iii) chưa giải quyết được vấn đề một bài đăng mang nhiều ý định rõ cùng một lúc; (iv) bộ dữ liệu còn khiêm tốn đối với thực nghiệm theo phương pháp học sâu. Trong tương lai gần, NCS sẽ tiếp tục tập trung giải quyết các vấn đề vừa nêu.

DANH MỤC CÁC CÔNG TRÌNH KHOA HỌC CỦA TÁC GIẢ LIÊN QUAN ĐẾN LUẬN ÁN

[1.] [LTLe1] **Thai-Le Luong**, Thi-Hanh Tran, Quoc-Tuan Truong, Thi-Minh-Ngoc Truong, Thi-Thu Phi and Xuan-Hieu Phan (2016). *Learning to Filter User Explicit Intents in Online Vietnamese Social Media Texts*. The Eighth Asian Conference on Intelligent Information and Database Systems (ACIIDS), pp.13-24, Springer, 2016. **[SCOPUS, DBLP]**

[2.] [LTle2] **Thai-Le Luong**, Quoc-Tuan Truong, Hai-Trieu Dang and Xuan-Hieu Phan (2016). *Domain Identification for Intention Posts on Online Social Media*. In Proceedings of the Seventh Symposium on Information and Communication Technology (SoICT), pp. 52-57, ACM, 2016. **[SCOPUS, DBLP]**

[3.] [LTLe3] **Thai-Le Luong**, Minh-Son Cao, Duc-Thang Le and Xuan-Hieu Phan (2017). *Intent Extraction from Social Media Texts Using Sequential Segmentation and Deep Learning Models*. In Proceedings of the 9th International Conference on Knowledge and Systems Engineering (KSE), pp. 215-220, Springer LNCS, IEEE, 2017. **[SCOPUS, DBLP]**

[4.] [LTLe4] **Thai-Le Luong**, Nhu-Thuat Tran and Xuan-Hieu Phan (2019). *Improving Intent Extraction Using Ensemble Neural Network*. In Proceedings of the 19th International Symposium on Communications and Information Technologies (ISCIT), pp. 58-63, IEEE, 2019. **[DBLP]**

[5.] [LTLe5] **Thai-Le Luong**, Nhu-Thuat Tran, Tien-Son Dang, Quoc-Long Tran and Xuan-Hieu Phan (2019). *Domain-independent Intent Extraction from Online Texts*. Computación y Sistemas, Accepted, 2019. **[SCOPUS Journal]**