HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Nguyễn Tiến Đạt

PHÁT HIỆN Ý ĐỊNH NGƯỜI DÙNG TRONG HỆ THỐNG HỎI ĐÁP SỬ DỤNG MẠNG NƠRON

Chuyên ngành: Hệ thống thông tin

Mã số: 8.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI - NĂM 2019

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: TS. Ngô Xuân Bách

Phản biện 1: ……

Phản biện 2: ……..

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ..... giờ ..... ngày ..... tháng .... năm .....

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

MỞ ĐẦU

Nghiên cứu về hệ thống hỏi đáp tự động (Q&A) đã được quan tâm từ rất lâu trên thế giới. Ngay từ những năm 1960, các hệ thống hỏi đáp đầu tiên sử dụng cơ sở dữ liệu đã được ra đời. Với mục đích hệ thống được xây dựng để thực hiện việc tìm kiếm tự động câu trả lời từ một tập lớn các tài liệu cho câu hỏi đầu vào một cách chính xác.

Hiện nay, số lượng hệ thống hỏi đáp ngày càng tăng, số lượng câu hỏi gửi về các hệ thống mỗi ngày ngày càng nhiều và việc phát hiện được ý định câu hỏi của người dùng là một trong những bước đầu tiên để lựa chọn được câu trả lời đúng với mong muốn người dùng quan tâm.

Ở các trường Đại học, hệ thống hỏi đáp đang được áp dụng phổ biến và từng bước phát triển, điều này giúp các học sinh THPT muốn tiếp cận, tìm hiểu thông tin cũng như bản thân các sinh viên trong trường muốn biết rõ hơn về các khóa học, lợi ích mà trường Đại học đang có một cách thuận tiện, nhanh chóng. Tuy nhiên, để giải quyết số lượng câu hỏi lớn trong một thời gian thì việc xây dựng đề xuất giải pháp phát hiện thông tin người dùng muốn hỏi trong hệ thống hỏi đáp là tiền đề để xác định và tìm kiếm được câu trả lời phù hợp với ý định người dùng.

Vì những lý do trên nên tôi quyết định lựa chọn đề tài ***"Phát hiện ý định người dùng trong hệ thống hỏi đáp sử dụng mạng nơron"*** để nghiên cứu và đưa ra một giải pháp sử dụng học máy để phát hiện ý định người dùng trong hệ thống hỏi đáp. Từ đó các hệ thống hỏi đáp sẽ tiết kiệm được thời gian, giải quyết được các câu hỏi nhanh chóng và đúng vấn đề mà các học sinh THPT hay Đại học đang có nhu cầu muốn hỏi. Cùng với đó, những nghiên cứu trong khóa luận có thể coi là tiền đề cho các nghiên cứu tiếp theo để đưa ra các câu trả lời và phân loại câu hỏi theo ý định người dùng cho một hệ thống hỏi đáp ngày một hoàn thiện.

Luận văn được tổ chức gồm ba chương gồm:

Chương 1: Giới thiệu tổng quan về bài toán xử lý ngôn ngữ tự nhiên. Tìm hiểu bài toán phân loại văn bản và giới thiệu bài toán phát hiện ý định người dùng trong hệ thống hỏi đáp.

Chương 2: Trình bày phương pháp giải quyết bài toán và các phương pháp biểu diễn đặc trưng cho văn bản cùng phương pháp học máy mà đề tài lựa chọn: sử dụng mạng nơron và so sánh với Support Vector Machine (SVM).

Chương 3: Trình bày về kịch bản thực nghiệm cho các trường hợp xác định ý định người dùng trên bộ dữ liệu thực nghiệm được thu thập từ: *Kênh thông tin trực tuyến, Khoa Quốc tế, Đại học quốc gia Hà Nội.*

CHƯƠNG 1: TỔNG QUAN BÀI TOÁN PHÁT HIỆN Ý ĐỊNH

NGƯỜI DÙNG

1.1 Xử lý ngôn ngữ tự nhiên

Xử lý ngôn ngữ tự nhiên (natural language processing - NLP) là một nhánh của trí tuệ nhân tạo tập trung vào các ứng dụng trên ngôn ngữ của con người. Trong trí tuệ nhân tạo thì xử lý ngôn ngữ tự nhiên là một trong những phần khó nhất vì nó liên quan đến việc phải hiểu ý nghĩa ngôn ngữ - công cụ hoàn hảo nhất của tư duy và giao tiếp.

Xử lý ngôn ngữ tự nhiên là lĩnh vực đã được nghiên cứu từ nhiều năm nay và đạt được nhiều bước tiến quan trọng trong những năm gần đây với các ứng dụng về bài toán trong thực tế như:

- Nhận dạng chữ viết (bao gồm chữ in và chữ viết tay),
- Nhận dạng tiếng nói,
- Dịch tự động,
- Tìm kiếm thông tin,
- Tóm tắt văn bản,
- Khai phá dữ liệu,
- Phát hiện tri thức, v.v.

1.2 Bài toán phát hiện ý định người dùng trong hệ thống hỏi đáp

1.2.1 Phân loại văn bản

Phân loại văn bản là quá trình phân lớp một đối tượng dữ liệu vào một hay nhiều lớp cho trước nhờ một mô hình phân lớp mà mô hình này được xây dựng dựa trên một tập hợp các đối tượng dữ liệu đã được gán nhãn từ trước gọi là tập dữ liệu học (tập huấn luyện).

Quá trình phân lớp còn được gọi là quá trình gán nhãn cho các đối tượng dữ liệu. Các bài toán phân loại văn bản thường thấy là:

- Phân cụm văn bản,
- Tóm tắt văn bản,
- Xác định quan điểm,
- Phát hiện ý định,
- Phân tích cảm xúc, hành vi của người dùng, v.v.

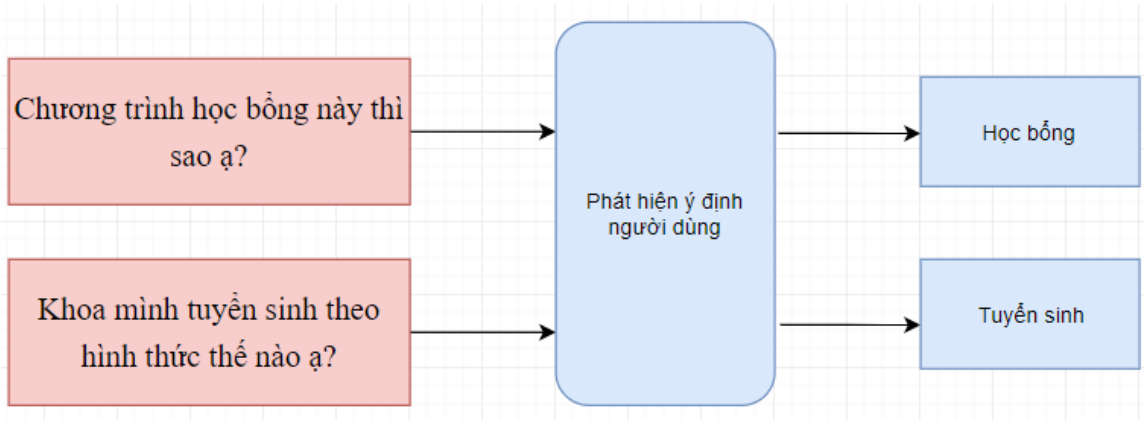
Trong nội dung luận văn này sẽ tập trung vào bài toán phát hiện ý định của người dùng trong hệ thống hỏi đáp của trường Đại học.

1.2.2 Phát biểu bài toán

Bài toán xây dựng hệ thống hỏi đáp là một bài toán khó thuộc lĩnh vực xử lý ngôn ngữ tự nhiên. Ngôn ngữ tự nhiên vốn nhập nhằng, đa nghĩa, việc xác định được ngữ nghĩa của câu hỏi cũng như phát hiện ra câu trả lời là một thách thức không nhỏ. Không những vậy, các câu hỏi có thể mang theo các thói quen, phong cách gõ chữ của cá nhân người hỏi như "em muốn hỏi mã dăng ký cảu htttt ạ" (Em muốn hỏi mã đăng ký của HTTT ạ?), "Mã ngahf kh quản lí là bao nhiêu ak" (Mã ngành quản lí là bao nhiêu ạ?). Ngoài ra, "Ý định người dùng còn 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" [9].

Ý tưởng của luận văn là sẽ đi sâu vào giải quyết bài toán xác định ý định người dùng (học sinh, sinh viên) với:

- Đầu vào: Một câu hỏi của người dùng(học sinh, sinh viên)
- Đầu ra: Ý định của người dùng(thông tin mà học sinh, sinh viên muốn hỏi)



Hình 1.1 Bài toán phát hiện ý định người dùng

Chẳng hạn như ví dụ tại hình 1.1, với đầu vào câu hỏi trong hệ thống hỏi đáp là "Chương trình học bổng này thì sao ạ?" hệ thống sẽ đưa ra được ý định của người dùng là muốn hỏi về học bổng, hay với câu hỏi "Khoa mình tuyển sinh theo hình thức thế nào ạ?" thì hệ thống sẽ phát hiện được ý định của người dùng là muốn hỏi về vấn đề tuyển sinh.

1.2.3 Ý nghĩa bài toán

Ý định là một khái niệm quan trọng, được coi như chìa khóa để xây dựng các hệ thống hỏi đáp hiện nay. Luận văn mong muốn sẽ đưa ra được ý định người dùng dựa trên các ý định

cho trước làm tiền đề cho các hệ thống gợi ý, giới thiệu,… vấn đề mà người dùng đang quan tâm.

Ví dụ: người dùng đặt câu hỏi *"Ngành quản lí thì cơ hội nghề nghiệp ntn ạ?"*; hệ thống sẽ đưa ra được ý định của người dùng là: *cơ hội nghề nghiệp*; từ đó làm tiền đề cho các hệ thống gợi ý, giới thiệu, đưa ra các lời mời về cơ hội việc làm liên quan đến thông tin nghề nghiệp người dùng muốn hỏi.

1.3 Các nghiên cứu liên quan

Trong những năm gần đây, đã có nhiều đề tài về phát hiện ý định người dùng với các phương pháp khác nhau được áp dụng ví dụ như đề tài *"Identifying Intention Posts in Discussion Forums"*[18] về xác định ý định người dùng dựa trên các bài viết đăng trong 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. 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. 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).

Ngoài ra, tác giả Ahmed Husseini Orabi cùng cộng sự đã thực hiện một đề tài rất thiết thực và có ý nghĩa về việc sử dụng học sâu để phát hiện trầm cảm của người dùng Twitter: *"Deep Learning for Depression Detection of Twitter Users"* [6]. Công trình trình bày việc xử lý ngôn ngữ tự nhiên trên mạng xã hội twitter, thực hiện đánh giá và so sánh trên một số mô hình học sâu, cụ thể là 3 mô hình CNN và 1 mô hình RNN và đưa ra kết quả về vấn đề rối loạn tâm thần và làm tiền đề cho hệ thống phát hiện các hành vi, cảm xúc tiêu cực của người dùng cá nhân trên mạng xã hội.

Không chỉ có vậy, đề tài *"Supervised Clustering of Questions into Intents for Dialog System Applications"* [12], của Iryna Haponchyk và cộng sự đề cập đến việc phân cụm các câu hỏi của các hệ thống hỏi đáp thành các ý định khác nhau. Cụ thể, công trình tập trung vào các ý định của người dùng hệ thống hỏi đáp thông dụng về các phân cụm như: thời tiết, giảm cân, địa điểm,… Công trình đã một phần nào đó chứng minh được "ý định" là chìa khóa quan trọng để xây dựng hệ thống hỏi đáp thông minh, xác định nhanh mục đích trong mỗi ngữ cảnh. Trong công trình này, nhóm tác giả cũng đã đề xuất một mô hình để tự động phân cụm

các câu hỏi thành các mục đích của người dùng với độ chính xác của phân cụm khá cao (khoảng 80%), có thể giúp thiết kế các hệ thống hỏi đáp sau này.

Bên cạnh đó, với sức hút và sự phát triển nhanh chóng của lĩnh vực xử lý ngôn ngữ tự nhiên trong những năm gần đây, đã có rất nhiều công trình nghiên cứu của các tác giả [7], [8], [13], [14], [15] liên quan đến việc khai phá quan điểm, phân tích ý định từ nhiều nguồn dữ liệu với các phương pháp khác nhau như sử dụng phương pháp SVM, sử dụng mô hình mạng nơron hồi quy, mô hình mạng nơron tích chập,… với kết quả rất khả quan và hứa hẹn sẽ phát triển và bùng nổ trong những năm tới.

Qua việc nghiên cứu, khảo sát các đề tài liên quan đến vấn đề phát hiện ý định người dùng trong hệ thống hỏi đáp của trường Đại học còn hạn chế và chưa có nhiều. Bên cạnh đó, luận văn nhận thấy nhu cầu xử lý và phát hiện ý định người dùng trong hệ thống hỏi đáp dành cho học sinh, sinh viên mỗi kỳ tuyển dụng của trường Đại học ngày một lớn nên việc học hỏi, tiếp thu các đề tài phát hiện ý định người dùng để áp dụng với hệ thống hỏi đáp của trường Đại học là cần thiết.

Luận văn sẽ tham khảo, tìm hiểu và giới thiệu về các phương pháp phổ biến, sau đó sẽ áp dụng và đưa ra kết quả đánh giá cũng như đề xuất giải pháp để xây dựng phát triển hệ thống hỏi đáp cho các trường Đại học. Những đóng góp ban đầu của luận văn như: xử lý tiền dữ liệu, phân lớp dữ liệu trên các phương pháp khác nhau sẽ làm cơ sở ban đầu trong việc đánh giá và lựa chọn các phương pháp, mô hình học máy sao cho phù hợp với hệ thống hỏi đáp trong trường Đại học, làm tiền đề cho các ứng dụng tự động, phân tích sử dụng dữ liệu từ hệ thống hỏi đáp sau này.

1.4Kết luận chương

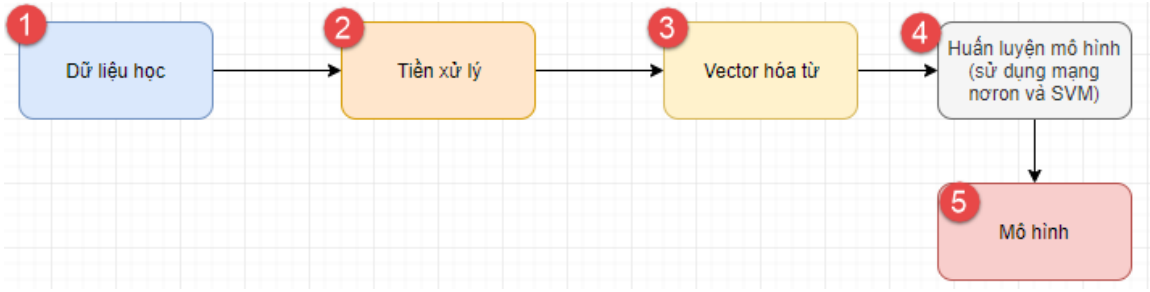
Chương 1 đã giới thiệu tổng quan về bài toán xử lý ngôn ngữ tự nhiên. Tìm hiểu bài toán phân loại văn bản và giới thiệu bài toán phát hiện ý định người dùng trong hệ thống hỏi đáp dành cho học sinh, sinh viên của trường Đại học, từ đó đưa ra những vấn đề cần làm rõ và giải quyết trong luận văn.

Trong chương 2, luận văn sẽ trình bày về hướng giải quyết cho bài toán phát hiện ý định người dùng, và đi sâu hơn trình bày về các phương pháp sẽ áp dụng để giải quyết bài toán.

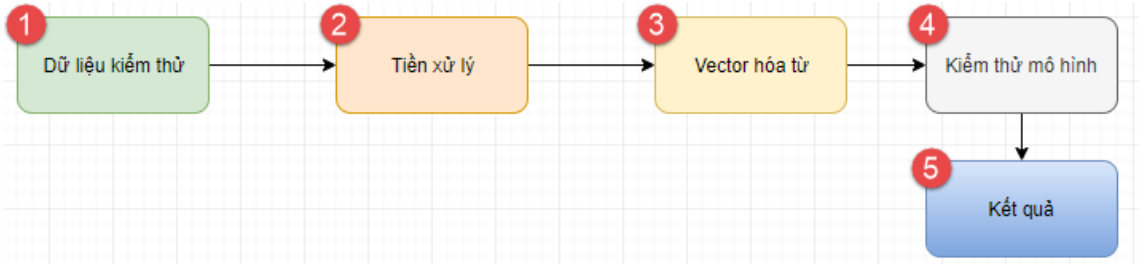
CHƯƠNG 2: PHƯƠNG PHÁP PHÁT HIỆN Ý ĐỊNH NGƯỜI DÙNG SỬ DỤNG HỌC MÁY

2.1 Phương pháp giải quyết bài toán

Để giải quyết bài toán phát hiện ý định người dùng trong hệ thống hỏi đáp của trường Đại học, từ những câu hỏi được tổng hợp từ hệ thống hỏi đáp ví dụ như: *"các chủ đề NCKH năm nay là như thế nào ạ?", "thủ tục đăng kí NCKH ?";* ta sẽ phân lớp và đưa được về nhóm *"Nghiên cứu khoa học"*. Luận văn đã tham khảo và tìm hiểu sau đó đưa ra được các bước thực hiện để xây dựng phương pháp giải quyết cho bài toán xác định ý định người dùng được chia làm 2 giai đoạn: huấn luyện và kiểm thử. Hai giai đoạn được mô tả như trong hình 2.1 và 2.2 dưới đây:



Hình 2.1 Giai đoạn huấn luyện mô hình



Hình 2.2 Giai đoạn kiểm thử mô hình

Áp dụng phương pháp chia làm 2 giai đoạn như đã trình bày ở trên, bài toán phát hiện ý định người dùng trong hệ thống hỏi đáp, luận văn sẽ thực hiện các bước sau:

1. Chia dữ liệu thành 2 phần: dữ liệu học và dữ liệu kiểm thử
2. Tiền xử lý dữ liệu đầu vào: Loại bỏ các ký tự đặc biệt, các tiền tố dư thừa, các từ stopwords
3. Vector hóa từ cho tập dữ liệu
4. Áp dụng mô hình học máy để giải quyết bài toán, bao gồm mô hình mạng nơron và so sánh với phương pháp SVM

5. Đưa ra mô hình huấn luyện và kết quả kiểm thử.

Tại bước 1, luận văn sẽ áp dụng phương pháp K-fold cross validation và chia dữ liệu thành 3 phần bằng nhau. Cụ thể về phương pháp K-fold cross validation sẽ được luận văn trình bày tại mục 3.2 về thiết lập thực nghiệm.

Trong bước 2, tiền xử lý dữ liệu, chẳng hạn với dữ liệu đầu vào mẫu như trên, ta phải loại bỏ các tiền tố dư thừa của việc đánh số thứ tự như *"1767.", "1768."* và các *khoảng trắng* cùng với các stopwords: *"ạ", "gì", "thì", …*

Các phần tiếp theo của chương 2 sẽ trình bày chi tiết hơn về các phương pháp, mô hình và đưa ra đề xuất lựa chọn và áp dụng vào việc phát hiện ý định của người dùng trong hệ thống hỏi đáp.

2.2 Các phương pháp biểu diễn đặc trưng của văn bản

2.2.1 Phương pháp N-Gram

2.2.2 Phương pháp TF-IDF

2.2.3 Phương pháp Word Vectors

2.3 Các phương pháp học máy xây dựng mô hình phân lớp

2.3.1 Phương pháp SVM

2.3.2 Kiến trúc mạng nơron tích chập (CNN)

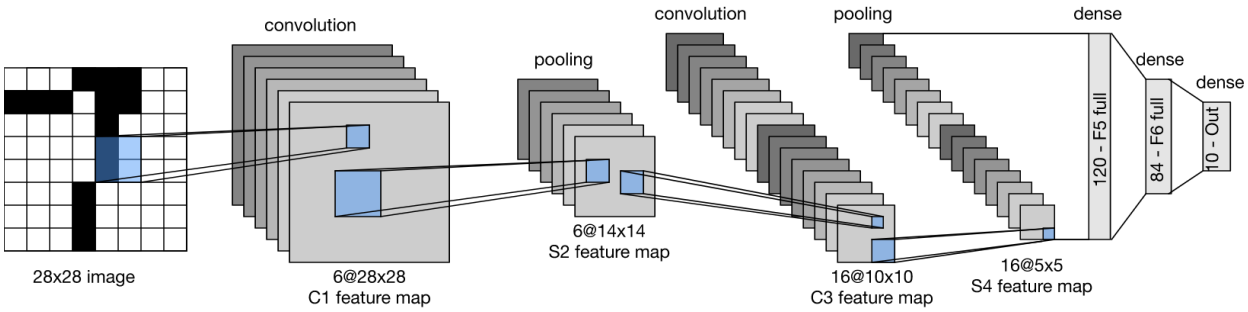
Mạng nơron tích chập [19] là một trong những mạng truyền thẳng đặc biệt. Mạng nơron tích chập là một mô hình học sâu phổ biến và tiên tiến nhất hiện nay. Hầu hết các hệ thống nhận diện và xử lý ảnh hiện nay đều sử dụng mạng nơron tích chập vì tốc độ xử lý nhanh và độ chính xác cao. Trong mạng nơron truyền thống, các tầng được coi là một chiều, thì trong mạng nơron tích chập, các tầng được coi là 3 chiều, gồm: chiều cao, chiều rộng và chiều sâu. Mạng nơron tích chập có hai khái niệm quan trọng: kết nối cục bộ và chia sẻ tham số. Những khái niệm này góp phần giảm số lượng trọng số cần được huấn luyện, do đó tăng nhanh được tốc độ tính toán.

Có ba tầng chính để xây dựng kiến trúc cho một mạng nơron tích chập:

1. Tầng tích chập
2. Tầng gộp (pooling layer)
3. Tầng được kết nối đầy đủ (fully-connected).

Tầng kết nối đầy đủ giống như các mạng nơron thông thường, và tầng chập thực hiện tích chập nhiều lần trên tầng trước. Tầng gộp có thể làm giảm kích thước mẫu trên từng khối

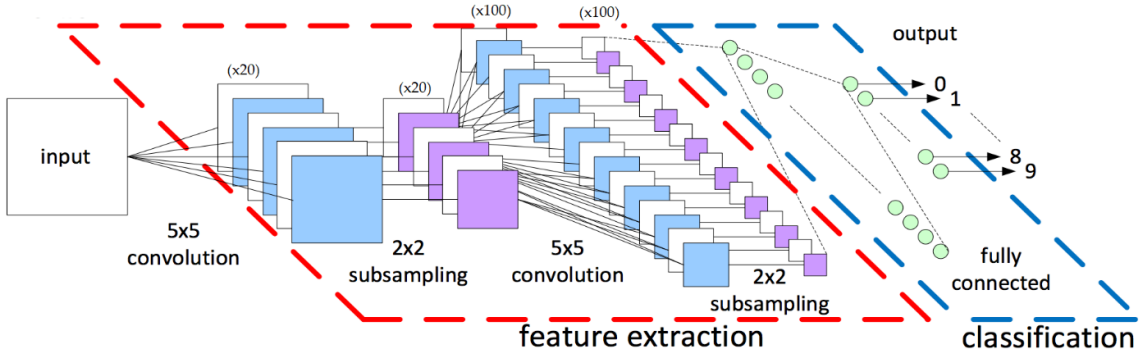
2x2 của tầng trước đó. Ở các mạng nơron tích chập, kiến trúc mạng thường chồng ba tầng này để xây dựng kiến trúc đầy đủ. Ví dụ minh họa về một kiến trúc mạng nơron tích chập đầy đủ:



Hình 2.3 Kiến trúc mạng LeNet [19]

Sau quá trình tìm hiểu và tham khảo, với điều kiện thiết bị thực nghiệm còn hạn chế, với kiến trúc CNN, luận văn quyết định áp dụng 2 convolutional layers với các thông số sau:

- Convolutional layer 1:
  - 20 Feature maps
  - Patch size 5x5
  - Pool size 2x2
- Convolutional layer 2:
  - 100 Feature maps
  - Patch size 5x5
  - Pool size 2x2



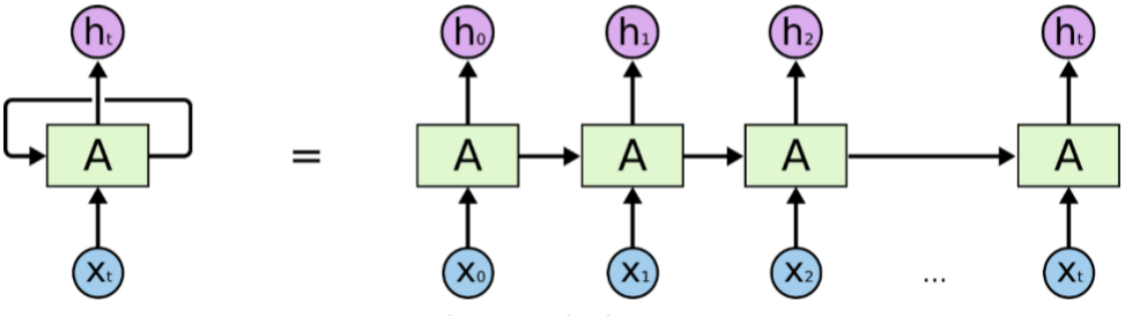
Hình 2.4 Mô hình CNN luận văn sử dụng

2.3.3 Kiến trúc mạng nơron hồi quy (RNN)

a. Giới thiệu mạng nơron hồi quy RNN

Mạng nơron hồi quy RNN được mô hình để giải quyết vấn đề mô phỏng về mặt thời gian của dữ liệu chuỗi. Do đó, mạng RNN rất phù hợp cho việc mô hình hóa xử lý ngôn ngữ.

Trong đó, mỗi từ trong chuỗi đầu vào sẽ được liên kết với một bước thời gian cụ thể. Trong thực tế, số bước thời gian sẽ bằng với độ dài tối đa của chuỗi.



Hình 2.5 Mô hình mạng RNN [18]

Hình 2.4 là mô tả cơ bản của mạng RNN. Hàm A nhận đầu vào x_t tại thời điểm t và đầu ra là giá trị vector ẩn h_t. Nhận thấy, hàm A cho phép thông tin được lặp lại truyền từ một bước của mạng tới bước tiếp theo. Sử dụng mạng RNN có rất nhiều ứng dụng như nhận dạng giọng nói, mô hình hóa ngôn ngữ, dịch, nhận dạng ảnh.

Tuy nhiên, mạng RNN có vấn đề lưu trữ thông tin ngữ cảnh phụ thuộc lâu dài. Xét 2 trường hợp ví dụ sau đây:

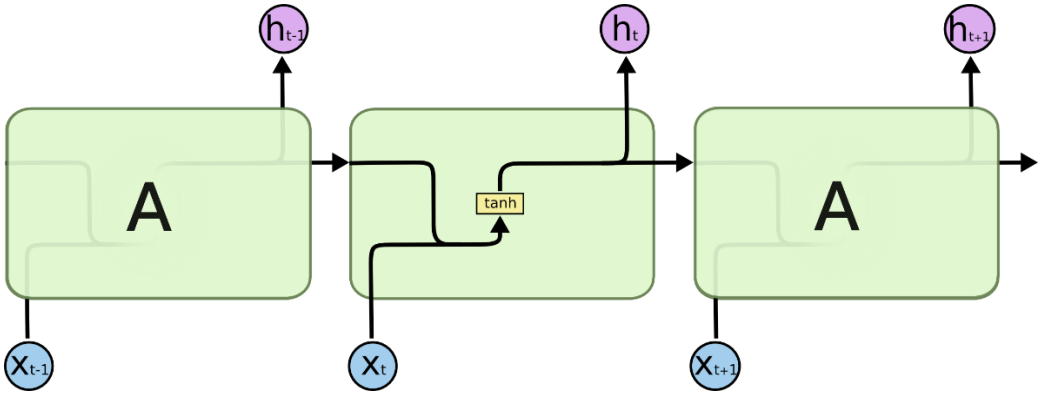
1. Trên đường nhiều xe cộ.
2. Tôi lớn lên ở Hà Nội, tôi có thể nhớ hết danh lam thắng cảnh tại Hà Nội.

Với ví dụ 1, ta không cần thông tin ngữ cảnh, nhưng trong trường hợp 2, các thông tin phía trước đó gợi ý rằng từ tiếp theo có thể liên quan đến tên của một thành phố. Trong trường hợp 2, khoảng cách giữa 2 phụ thuộc này là lớn hơn. Để đưa ra dự đoán này, bắt buộc mạng RNN phải lưu trữ toàn bộ các từ vào trong bộ nhớ. Trong phạm vi khoảng cách phụ thuộc này thấp thì có thể khả thi, nhưng nếu với khoảng cách cực lớn, đoạn văn dài thì việc lưu trữ của RNN trở nên nặng nề và không hợp lý. Đây chính là vấn đề lưu trữ thông tin phụ thuộc lâu dài.

Trên lý thuyết, mạng RNN có thể phát sinh bộ nhớ đủ để xử lý vấn đề lưu trữ phụ thuộc dài. Tuy nhiên, trong thực tế thì không phải vậy. Vấn đề này đã được Hochreiter (1991) đưa ra như thách thức của mạng RNN. Và mạng Long short-term memory (LSTM) được phát biểu năm 1997 đã giải quyết được vấn đề này.

b. Mạng Long short-term memory (LSTM)

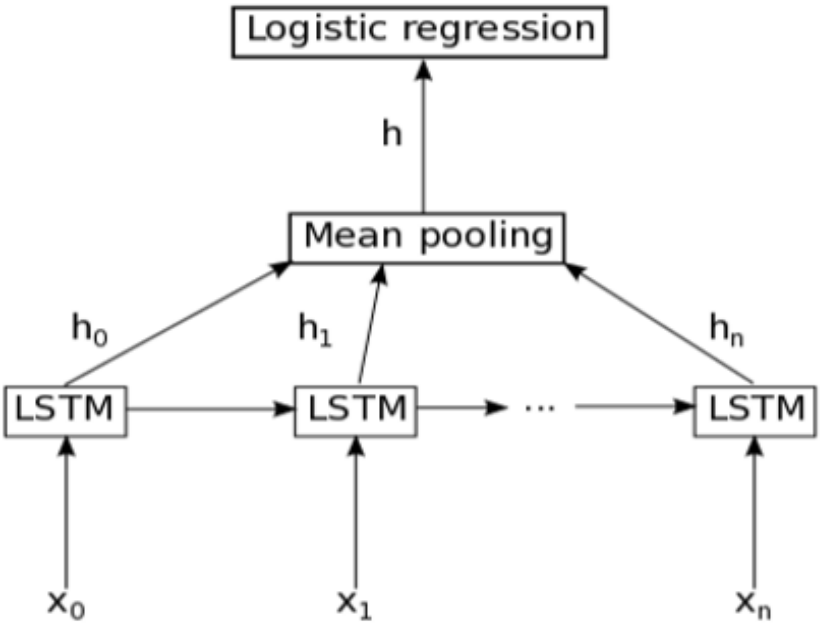
Long short term memory là cải tiến của mạng RNN nhằm giải quyết vấn đề học, lưu trữ thông tin ngữ cảnh phụ thuộc dài. tôi cùng xem xét cách LSTM [11] cải tiến hơn so với mạng RNN. Trong mô hình RNN, tại thời điểm t thì giá trị của vector ẩn ht chỉ được tính bằng một hàm tanh



Hình 2.6 Module xử lý h_t của RNN [18]

LSTM cũng có cấu trúc mắt xích tương tự, nhưng các module lặp có cấu trúc khác hẳn. Thay vì chỉ có một layer neural network, thì LSTM có tới bốn layer, tương tác với nhau theo một cấu trúc cụ thể.

Với ưu điểm về lưu trữ phụ thuộc dài, model sử dụng để huấn luyện trong luận văn này là model LSTM. Mô hình được luận văn sử dụng được mô tả trong hình 2.17 gồm một lớp LSTM duy nhất sau đó là một lớp tổng hợp trung bình (full-connection) và một lớp hồi quy logistic. Các từ được vector hóa sử dụng mô hình Word2vec.



Hình 2.7 Mô hình LSTM luận văn sử dụng

2.4 Kết luận chương

Chương 2 đã trình bày về quá trình tìm hiểu và áp dụng thuật toán TF-IDF, N-Gram để trích xuất đặc trưng. Bên cạnh đó, chương này cũng đã trình bày giới thiệu về thuật toán SVM, mạng nơron tích chập, mạng nơron hồi quy để phân lớp dữ liệu.

Với những kiến thức đã tìm hiểu và trình bày tại chương, luận văn sẽ áp dụng kiến trúc mạng nơron hồi quy – LSTM, kiến trúc mạng CNN và so sánh với SVM.

Chương 3 sẽ tiến hành thiết lập thực nghiệm dữ liệu với phương pháp đã đề xuất trên các kịch bản khác nhau, sau đó sẽ đánh giá độ chính xác và đưa ra đề xuất định hướng tiếp theo.

CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ

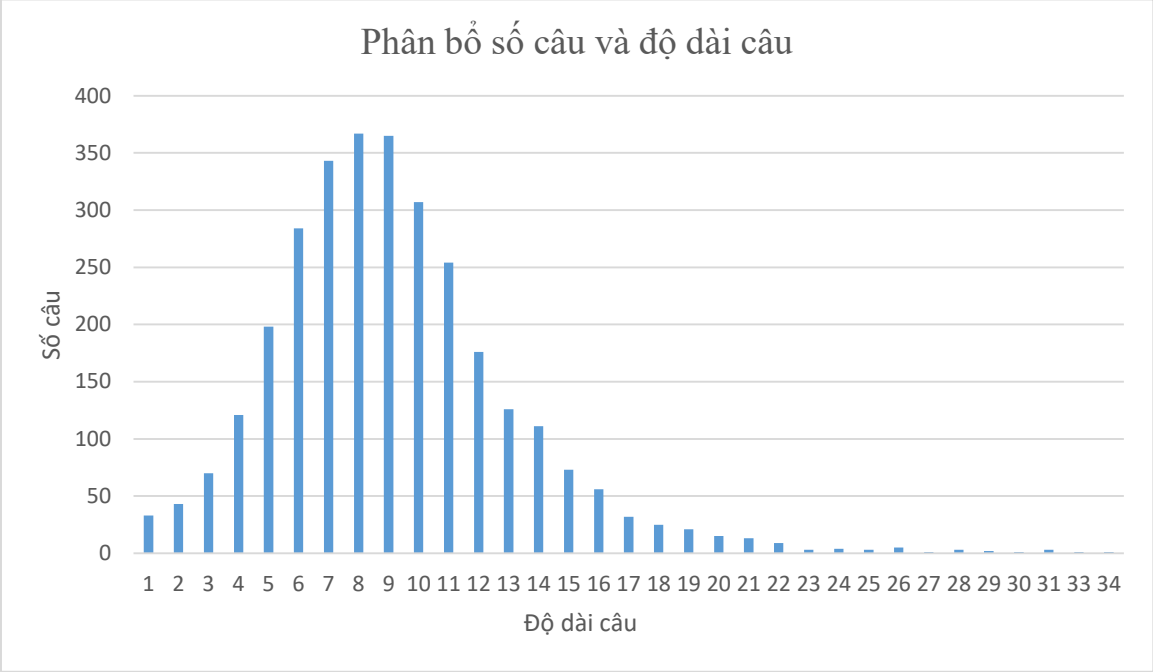
3.1 Dữ liệu thực nghiệm

Luận văn sử dụng dữ liệu thực nghiệm được thu thập từ: ***Kênh thông tin trực tuyến, Khoa Quốc tế, Đại học quốc gia Hà Nội*** với tổng số lượng là **3069** câu hỏi. Quá trình gán nhãn cho tệp dữ liệu gồm 3 bạn tham gia, 2 bạn gán nhãn và 1 bạn kiểm tra lại việc gán nhãn. Sau khi thực hiện gán nhãn, các câu hỏi được đưa về các lớp ý định sau: *Thông tin về trường, thông tin liên lạc, thông tin về khoa, cơ hội nghề nghiệp, điều kiện tiếng Anh, học phí, điểm chuẩn, nhập học, thủ tục, học bổng, nghiên cứu khoa học, tài liệu, từ chối/ không đồng ý, đồng ý, khác*. Số lượng cụ thể thu được sau quá trình gán nhãn ý định được mô tả tại bảng 3.1.

Nội dung ý định	Số lượng
Thông tin về trường	150
Thông tin liên lạc	91
Thông tin về khoa	569
Cơ hội nghề nghiệp	73
Điều kiện tiếng Anh	84
Học phí	192
Điểm chuẩn	83
Nhập học	275
Thủ tục	502
Học bổng	379
Nghiên cứu khoa học	300
Tài liệu	86
Từ chối, không đồng ý	100
Đồng ý	100
Khác	85

Bảng 3.1 Bảng mô tả dữ liệu thực nghiệm

Làm khảo sát với tập dữ liệu này, luận văn có được biểu đồ phân bố số lượng từ trong câu như biểu đồ 3.1



Hình 3.1 Biểu đồ phân bố số câu và độ dài câu

Dựa vào biểu đồ trên ta có thể thấy:

- Số lượng câu tập trung phần lớn khoảng 5 đến 12 từ
- Số lượng câu trên 100 từ khá nhiều, toàn bộ các câu có số lượng từ từ 4 đến 14 đều trên 100.
- Số lượng câu có độ dài 8 từ là nhiều nhất: 367 câu.
- Không có câu nào có độ dài 32 từ.
- Số lượng câu có độ dài 27, 30, 33, 34 thấp nhất: 1 câu.

3.2 Thiết lập thực nghiệm

Quá trình thực nghiệm thuật toán gồm 3 giai đoạn chính:

- Tiền xử lý dữ liệu: Loại bỏ các dư thừa, các từ vô nghĩa trong câu.
- Vector hóa và trích chọn đặc trưng: Sử dụng 2 thuật toán TF-IDF, N-Grams với n lần lượt chọn các giá trị 1, 2, 3.
- Xây dựng bộ phân lớp dữ liệu: Sử dụng LSTM, CNN và SVM.

Tiền xử lý dữ liệu: Luận văn sử dụng ngôn ngữ python để xử lý các dữ liệu dư thừa, loại bỏ các stopwords.

Vector hóa: Luận văn sử dụng filter StringToVector có sẵn trong Weka để thiết lập và trích chọn đặc trưng của dữ liệu.

Mô hình phân lớp: Mô hình mà luận văn sử dụng được mô tả trong phần 2.3.2 về mô hình CNN và phần 2.3.3 về mô hình LSTM.

Thiết lập tham số với Weka:

Sau quá trình nghiên cứu và tìm hiểu các phương pháp đánh giá thực nghiệm, luận văn đề xuất sử dụng phương pháp K-fold Cross Validation. K-fold cross validation có các đặc điểm sau:

- Tập toàn bộ các ví dụ D được chia ngẫu nhiên thành k tập con không giao nhau (gọi là "fold") có kích thước xấp xỉ nhau.
- Mỗi lần (trong số k lần) lặp, một tập con được sử dụng làm tập kiểm thử, và (k-1) tập con còn lại được dùng làm tập huấn luyện.
- k giá trị lỗi (mỗi giá trị tương ứng với một fold) được tính trung bình cộng để thu được giá trị lỗi tổng thể.

Để đánh giá chính xác hơn chất lượng của mô hình ta sử dụng thêm 2 độ đo là Precision và Recall.

- Precision được định nghĩa là tỉ lệ số điểm true positive trong số những điểm được phân loại là positive (TP + FP).

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

Công thức (3. 1) Tính Precision

- Recall được định nghĩa là tỉ lệ số điểm true positive trong số những điểm thực sự là positive (TP + FN).

$$Recall = \frac{TP}{TP + FN}$$

Công thức (3. 2) Tính Recall

Thực tế thì hai độ đo trên không phải lúc nào cũng tăng giảm tương ứng với nhau, có trường hợp Recall cao còn Precision thấp và ngược lại, để cho đánh giá tổng quát hơn ta dùng độ đo F-measure là trung bình điều hòa của 2 độ đo trên với hệ số 0.5 (tầm quan trọng của 2 hệ số ngang nhau):

$$F_1 = \frac{2}{\frac{1}{pricision} + \frac{1}{recall}} = 2\frac{pricision\ .recall}{pricision + recall}$$

Công thức (3. 3) Tính F_1

3.3 Công cụ thực nghiệm

3.3.1 Môi trường thực nghiệm

Thành phần	Thông số
CPU	CPU Intel Core i5 3.3GHz
RAM	RAM 8GB
Hệ điều hành (OS)	Windows 10 Professional 64bit

Bảng 3.2 Môi trường thực nghiệm

3.3.2 Công cụ phần mềm

Tên	Mô tả
PyCharm	IDE sử dụng Python để tiền xử lý dữ liệu. https://www.jetbrains.com/pycharm/
Weka 3.8	Công cụ tích hợp hỗ trợ các thuật toán học máy. https://www.cs.waikato.ac.nz/ml/weka/
Package WekaDeeplearning4j	Gói thư viện deep learning dành cho Weka. https://deeplearning.cms.waikato.ac.nz/user-guide/getting-started/
Package LibSVM	Gói thư viện thuật toán SVM cho Weka. http://weka.sourceforge.net/doc.stable/weka/classifiers/functions/LibSVM.html
Packge NeuralNetwork	Gói thư viện hỗ trợ Neural Network cho Weka. https://github.com/amten/NeuralNetwork

Bảng 3.3 Công cụ phần mềm

3.4Kết quả thực nghiệm

3.4.1 Kết quả

LSTM	**Unigrams**			**Bigrams**			**Trigrams**			**TF-IDF**		
Acc (%)	**85.14**			72.47			54.58			85.04		
Độ đo / Ý định	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**
Thông tin về trường	84.1	74.0	78.7	84.1	64.0	67.1	70.6	24.0	35.8	82.8	74.0	78.2
Thông tin liên lạc	83.7	79.1	81.4	83.7	46.2	56.8	73.3	12.1	20.8	83.7	79.1	81.4
Thông tin về khoa	85.0	85.8	85.4	85.0	87.0	68.9	39.0	**81.5**	52.8	84.9	85.8	85.3

Cơ hội nghề nghiệp	71.8	76.7	74.2	71.8	34.2	43.9	50.0	9.6	16.1	71.4	75.3	73.3
Điều kiện tiếng Anh	88.4	90.5	89.4	88.4	61.9	69.8	**91.7**	26.2	40.7	88.4	90.5	89.4
Học phí	83.4	89.1	86.1	83.4	68.2	70.8	61.0	43.2	50.6	83.4	89.1	86.1
Điểm chuẩn	70.4	60.2	64.9	70.4	33.7	44.4	55.0	13.3	21.4	70.4	60.2	64.9
Nhập học	81.1	87.3	84.1	81.1	77.1	71.5	66.5	68.7	67.6	81.1	87.3	84.1
Thủ tục	89.8	93.4	91.6	89.8	85.3	84.9	48.8	69.7	57.4	89.8	93.4	91.6
Học bổng	94.3	91.0	92.6	94.3	82.3	83.1	81.9	57.3	67.4	94.3	91.0	92.6
Nghiên cứu khoa học	**96.6**	**94.3**	**95.4**	**96.6**	**87.3**	**88.7**	87.1	74.0	**80.0**	**96.6**	**94.3**	**95.4**
Tài liệu	82.2	86.0	84.1	82.2	54.7	63.1	91.4	37.2	52.9	82.0	84.9	83.4
Từ chối, không đồng ý	80.8	59.0	68.2	80.8	46.0	56.1	63.0	17.0	26.8	79.7	59.0	67.8
Đồng ý	79.4	81.0	80.2	79.4	43.0	56.6	48.0	12.0	19.2	78.6	81.0	79.8
Khác	40.0	44.7	42.2	40.0	5.9	9.9	16.7	2.4	4.1	40.2	43.5	41.8

Bảng 3.4 Kết quả mô hình LSTM

CNN	**Unigrams**			**Bigrams**			**Trigrams**			**TF-IDF**		
Acc (%)	**85.76**			82.37			72.79			81.23		
Độ đo / Ý định	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**
Thông tin về trường	68.9	81.3	74.6	73.5	76.0	74.8	88.5	92.7	90.6	67.6	78.0	72.4
Thông tin liên lạc	87.2	90.1	88.6	96.7	95.6	96.1	98.9	98.9	98.9	76.1	76.9	76.5
Thông tin về khoa	82.5	83.8	83.2	67.1	83.0	74.2	81.8	59.9	69.2	80.5	77.9	79.2
Cơ hội nghề nghiệp	88.9	76.7	82.4	90.3	89.0	89.7	93.2	93.2	93.2	74.6	72.6	73.6
Điều kiện tiếng Anh	93.1	96.4	94.7	**98.8**	**96.4**	**97.6**	**100**	**100**	**100**	**91.6**	90.5	91.0
Học phí	91.0	89.6	90.3	78.5	79.7	79.1	81.8	56.3	66.7	85.1	86.5	85.8
Điểm chuẩn	85.1	75.9	80.3	89.3	80.7	84.8	82.8	92.8	87.5	64.3	75.9	69.6
Nhập học	85.5	90.2	87.8	77.3	72.0	74.6	81.5	69.1	74.8	84.5	85.5	85.0
Thủ tục	92.1	90.8	91.5	86.8	84.1	85.4	50.7	80.7	62.3	90.7	89.0	89.8
Học bổng	94.0	**95.8**	94.9	92.2	81.5	86.6	53.4	66.2	59.1	90.9	92.1	91.5
Nghiên cứu khoa học	**96.9**	95.3	**96.1**	93	88.0	90.4	93.4	70.7	80.5	87.1	**97.0**	**91.8**
Tài liệu	92.9	90.7	91.8	95.3	94.2	94.7	98.8	95.3	97.0	71.0	76.7	73.7
Từ chối, không đồng ý	61.6	69.0	65.1	81.2	69.0	74.6	90.3	56.0	69.1	52.7	58.0	55.2
Đồng ý	61.4	62.0	61.7	88.4	76.0	81.7	83.6	61.0	70.5	59.0	46.0	51.7
Khác	38.6	20.0	74.6	87.5	82.4	84.8	93.3	82.4	87.5	35.1	15.3	21.3

Bảng 3.5 Kết quả mô hình CNN

SVM	Unigrams			Bigrams			Trigrams			TF-IDF		
Acc (%)	**88.89**			70.22			51.48			87.59		
Độ đo / Ý định	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**
Thông tin về trường	84.7	81.3	83.0	69.5	60.7	64.8	90.6	19.3	31.9	78.4	80.0	79.2
Thông tin liên lạc	96.3	86.8	91.3	94.1	35.2	51.2	**100**	8.8	16.2	97.3	80.2	88.0
Thông tin về khoa	86.2	91.2	88.6	46.6	**88.8**	61.1	29.9	**91.9**	45.1	87.9	89.5	88.7
Cơ hội nghề nghiệp	88.7	75.3	81.5	95.2	27.4	42.6	80.0	5.5	10.3	90.0	74.0	81.2
Điều kiện tiếng Anh	97.5	91.7	94.5	**97.0**	76.2	85.3	97.9	56.0	71.2	97.4	90.5	93.8
Học phí	92.3	93.2	92.7	82.8	57.8	68.1	71.1	33.3	45.4	91.3	92.7	92.0
Điểm chuẩn	92.9	78.3	85.0	83.3	24.1	37.4	83.3	6.0	11.2	92.5	74.7	82.7
Nhập học	87.1	88.7	87.9	74.8	71.3	73.0	80.5	55.6	65.8	85.2	90.2	87.6
Thủ tục	94.5	95.0	94.7	84.2	82.7	83.4	71.0	57.6	63.6	94.7	92.0	93.3
Học bổng	**98.4**	94.7	96.5	78.8	84.4	81.5	79.1	52.0	62.7	95.2	**93.9**	94.6
Nghiên cứu khoa học	97.3	**96.7**	**97.0**	96.5	82.7	**89.0**	92.9	70.0	**79.8**	98.0	96.3	97.1
Tài liệu	97.3	82.6	89.3	92.7	59.3	72.3	**100**	26.7	42.2	**98.8**	93.0	**95.8**
Từ chối, không đồng ý	54.6	89.0	67.7	76.3	45.0	56.6	78.9	15.0	25.2	47.2	91.0	62.1
Đồng ý	84.0	79.0	81.4	90.0	36.0	51.4	81.3	13.0	22.4	78.7	70.0	74.1
Khác	39.7	27.1	32.2	25.0	1.2	2.2	00.0	00.0	00.0	39.2	23.5	29.4

Bảng 3.6 Kết quả phương pháp SVM

3.4.2 Đánh giá kết quả

a. So sánh độ chính xác của các phương pháp trích chọn đặc trưng

b. So sánh đặc trưng unigrams và bigrams LSTM và SVM

3.5 Kết luận chương

Nội dung chương này trình quá trình thực nghiệm luận văn phát hiện ý định người dùng trong hệ thống hỏi đáp trên bộ dữ liệu thu tập được từ "*Kênh thông tin trực tuyến, Khoa Quốc tế, Đại học quốc gia Hà Nội"*. Dựa trên số liệu kết quả thực nghiệm ở chương này luận văn đưa ra phân tích đánh giá về phương pháp thực hiện. Các kết quả cho thấy việc sử dụng các đặc trưng khác nhau mang lại độ chính xác khác nhau. Sau khi quan sát bộ dữ liệu, có rất nhiều từ được viết theo văn phong riêng và sai chính tả (Ví dụ: "add" – ý hỏi admin, ad) hay viết tắt (Ví dụ: k thay cho không) dù đã loại bỏ stopwords. Đây thực sự là thách thức trong việc xây dựng hệ thống phát hiện ý định với ngôn ngữ tự nhiên, đặc biệt bằng tiếng Việt.

KẾT LUẬN

Nghiên cứu về xử lý ngôn ngữ tự nhiên nói chung, về bài toán phát hiện ý định người dùng nói riêng với tôi là công nghệ mới, thời gian nghiên cứu còn ngắn nên vẫn còn nhiều vấn đề chưa thực sự nắm bắt tốt. Tuy nhiên qua quá trình nghiên cứu, luận văn đã tìm hiểu sâu về các giai đoạn từ tiền xử lý dữ liệu đến việc chọn các phương pháp biểu diễn đặc trưng của văn bản (N-grams, TF-IDF), phương pháp học máy để xây dựng mô hình phân lớp dữ liệu mạng nơron (kiến trúc LSTM và CNN trong luận văn đề xuất) và so sánh với phương pháp SVM.

Sử dụng mạng nơron nói chung hay mô hình LSTM và CNN nói riêng trong Deep Learning là một hướng đi có kỹ thuật và hiệu quả trong bài toán xử lý chuỗi và hiện đang được các nhà nghiên cứu sử dụng rất nhiều. Tuy nhiên, LSTM và CNN không phải là một kỹ thuật vạn năng mà cứ bài toán về NLP là lại áp dụng được. Nó còn căn cứ vào nhiều yếu tố như tập ngữ liệu, đặc tính của tập ngữ liệu. Vì đôi khi sử dụng một thuật toán SVM lại cho ra kết quả tốt hơn.

Trong tương lai, luận văn có thể được phát triển nghiên cứu các mô hình khác, thay đổi cấu trúc mạng nơron nhiều lớp hơn hoặc kết hợp các loại mạng nơron với nhau để nâng cao độ chính xác và cải thiện tốc độ xử lý đối với việc phát hiện ý định người dùng chính xác hơn. Luận văn cũng là tiền đề xây dựng hệ thống tư vấn, quảng cáo trong hệ thống hỏi đáp của trường Đại học phù hợp, với lượng người quan tâm cao và hỗ trợ nhanh chóng giải đáp đúng các vấn đề trong hệ thống hỏi đáp.

TÀI LIỆU THAM KHẢO

Tiếng Việt

[1] Ngo Xuan Bach, Tu Minh Phuong, "Leveraging User Ratings for Resource-Poor Sentiment Classification", In Proceedings of the 19th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES), Procedia Computer Science, pp. 322–331, 2015.

[2] Nguyen Thi Duyen, Ngo Xuan Bach, Tu Minh Phuong, "An Empirical Study on Sentiment Analysis for Vietnamese". In Proceedings of the International Conference on Advanced Technologies for Communications (ATC), Special session on Computational Science and Computational Intelligence (CSCI), pp. 309-314, 2014.

[3] Vũ Hữu Tiệp, Blog Machine Learning Cơ bản tại địa chỉ https://machinelearningcoban.com.

[4] Kim Đình Sơn, Đặng Ngọc Thuyên, Phùng Văn Chiến, Ngô Thành Đạt, Các mô hình ngôn ngữ N-gram và Ứng dụng, 2013.

[5] https://vi.wikipedia.org/wiki/Ng%C3%B4n_ng%E1%BB%AF, truy nhập ngày 18/10/2019.

Tiếng Anh

[6] Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, Diana Inkpen, "*Deep Learning for Depression Detection of Twitter Users*", 2018.

[7] Awais Athar, Simone Teufel, *"Detection of Implicit Citations for Sentiment Detection"*, 2012.

[8] B. Liu (2009), Handbook Chapter: Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing, Handbook of Natural Language Processing. Marcel Dekker, Inc. New York, NY, USA.

[9] Bratman, Michael, "Intention, plans, and practical reason.", 1987.

[10] Google (2013), Word2vec model https://code.google.com/archive/p/word2vec/.

[11] Hochreiter and Schmidhuber (1997), Long short-term memory.

[12] Iryna Haponchyk, Antonio Uva1, Seunghak Yu, Olga Uryupina, Alessandro Moschitti, *"Supervised Clustering of Questions into Intents for Dialog System Applications"*, 2018.

[13] Maria Karanasou, Christos Doulkeridis, Maria Halkidi, *"DsUniPi: An SVM-based Approach for Sentiment Analysis of Figurative Language on Twitter"*, 2015.

[14] Peng Chen, Zhongqian Sun Lidong Bing, Wei Yang, *"Recurrent Attention Network on Memory for Aspect Sentiment Analysis"*, 2017.

[15] Peng Zhou, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, Bo Xu, *"Text Classification Improved by Integrating Bidirectional LSTM with Two-dimensional Max Pooling"*, 2016.

[16] Zheng Chen, Fan Lin, Huan Liu, Yin Liu, Wei-Ying Ma and Liu Wenyin, "User Intention Modeling in Web Applications Using Data Mining", 2002.

[17] Zhiyuan Chen, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh, "Identifying Intention Posts in Discussion Forums", 2013.

[18] http://colah.github.io/posts/2015-08-Understanding-LSTMs/, truy nhập ngày 18/10/2019.

[19] https://d2l.ai/chapter_convolutional-neural-networks/lenet.html, truy nhập ngày 18/10/2019.

[20] http://karpathy.github.io/2015/05/21/rnn-effectiveness/, truy nhập ngày 18/10/2019.