HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN DIỆU LINH

PHÂN LOẠI CÂU HỎI PHÁP QUY TIẾNG VIỆT SỬ DỤNG MÔ HÌNH BERT

Chuyên ngành: Khoa học máy tính

Mã số: **8.48.01.04**

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI – 2021

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: TS. Phùng Văn Ổn

Phản biện 2: PGS.TS. Trần Đình Quế

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

Vào lúc: 08 giờ 40 ngày 09 tháng 01 năm 2021

Có thể tìm hiểu luận văn tại:

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

MỞ ĐẦU

Với nhu cầu trao đổi và tìm kiếm thông tin của con người ngày càng cao, đồng nghĩ với việc người dùng mong muốn kết quả tìm kiếm trả về một cách ngắn gọn, súc tích, chính xác nhất. Vì vậy, hệ thống hỏi đáp tự động ra đời nhằm đáp ứng nhu cầu này.

Hệ thống hỏi-đáp tự động là hệ thống được xây dựng nhằm mục đích 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.

Phân loại câu hỏi là pha đầu tiên trong kiến trúc chung của một hệ thống hỏi đáp, có nhiệm vụ tìm ra các thông tin cần thiết làm đầu vào cho quá trình xử lý của các pha sau (trích chọn tài liệu, trích xuất câu trả lời, v.v).

Văn bản pháp quy là văn bản có các quy phạm pháp luật do các cơ quan quản lý nhà nước, ở trung ương, cơ quan quyền lực nhà nước, cơ quan quản lý nhà nước ở địa phương ban hành theo thẩm quyền lập quy của mình. Muốn hỏi đáp một vấn đề pháp luật cần phải tra cứu tìm kiếm rất nhiều tài liệu văn bản pháp luật liên quan. Vì vậy, để giúp cho việc rút ngắn thời gian tìm kiếm thì cần phân loại câu hỏi pháp quy theo các lĩnh vực pháp luật.

Phân loại đa nhãn là phân loại văn bản, trong đó mỗi văn bản có thể thuộc một số chủ đề được xác định trước cùng một lúc. Một câu hỏi pháp quy thông thường có thể sẽ liên quan đến nhiều loại lĩnh vực pháp luật. Việc phân loại câu hỏi pháp quy tiếng Việt đặt ra là mỗi câu hỏi có thể thuộc một số lĩnh vực. Vì vậy, bài toán phân loại câu hỏi pháp quy tiếng Việt là bài toán phân loại đa nhãn câu hỏi pháp quy tiếng Việt.

Các phương pháp phổ biến hiện nay có rất nhiều phương pháp và cách tiếp cận để giải quyết bài toán phân loại câu hỏi. Gần đây có nhiều phương pháp học sâu sử dụng mạng nơ-ron phổ biến và cho kết quả tốt hơn do có thể tự động trích chọn được những thông tin cần thiết và học được ngữ nghĩa từ dữ liệu.

Mô hình BERT bản chất là một dạng mô hình huấn luyện trước, tận dụng các nguồn dữ liệu không có nhãn để học, sau đó dùng vào các bài toán khác.

Phân loại câu hỏi pháp quy tiếng Việt là bài toán phân loại câu hỏi về pháp luật thành các lĩnh vực pháp lý.

Luận văn "*Phân loại câu hỏi pháp quy tiếng Việt sử dụng mô hình BERT*" thực hiện mô hình hóa bài toán dưới dạng một bài toán phân lớp đa nhãn. Trong đó mỗi câu

hỏi có thể thuộc một hoặc nhiều lĩnh vực khác nhau. Luận văn thực hiện phân loại câu hỏi sử dụng cách tiếp cận học máy giám sát, cụ thể là sử dụng một số mô hình truyền thống SVM và mô hình BERT[18, 6]. Kết quả thực nghiệm tốt nhất đạt được khi sử dụng mô hình BERT là 89.47% (độ đo F1).

Nội dung chính của luận văn được trình bày trong ba chương như sau:

- **Chương 1: Giới thiệu bài toán phân loại câu hỏi pháp quy tiếng Việt** : Trong chương này, luận văn giới thiệu bài toán phân loại câu hỏi, đặc điểm dữ liệu câu hỏi pháp quy, một số nghiên cứu liên quan, các phương pháp phân loại câu hỏi và kết luận chương.
- **Chương 2**: **Phân loại câu hỏi pháp quy tiếng Việt sử dụng mô hình BERT** : Trong chương 2, luận văn giới thiệu về bái toán phân loại đa nhãn câu hỏi tiếng Việt, giới thiệu một số mô hình học sâu, giới thiệu phương pháp BERT và trình bày mô hình phân loại câu hỏi pháp quy tiếng Việt sử dụng mô hình BERT.
- **Chương 3**: **Thực nghiệm đánh giá** : Chương này, luận văn trình bày tổng quan về kho ngữ liệu, cách thu thập, tiền xử lý, xây dựng tập nhãn và thống kê kho ngữ liệu; sử dụng các thư viện có sẵn cài đặt hệ thống phân loại câu hỏi và áp dụng phương pháp được đề xuất ở Chương 2; thực hiện huấn luyện hệ thống với bộ dữ liệu và tập nhãn đã xây dựng và tthống kê và đánh giá kết quả thực nghiệm.

CHƯƠNG 1: BÀI TOÁN PHÂN LOẠI CÂU HỎI

1.1 Giới thiệu bài toán phân loại câu hỏi

Hệ thống hỏi đáp là một hệ thống đóng vai trò phổ biến trong việc tìm kiếm thông tin nhanh chóng, chính xác và hiệu quả. Nhiệm vụ của nó là đưa ra câu trả lời đầy đủ và chính xác ứng với yêu cầu mong muốn của người dùng và câu trả lời được thể hiện bằng ngôn ngữ tự nhiên. Một trong các yếu tố đóng vai trò quan trọng trong hệ thống hỏi đáp là phân loại câu hỏi.

Bài toán phân loại câu hỏi thực chất có thể xem là bài toán phân lớp. Phân loại câu hỏi là việc gán các nhãn phân loại cho các câu hỏi dựa trên mức độ tương tự của câu hỏi đó so với các câu hỏi đã được gán nhãn trong tập huấn luyện. Việc phân loại câu hỏi thường được thể hiện bằng cách gán cho câu hỏi một nhãn có sẵn theo tập nhãn cho trước.

Bài toán phân loại câu hỏi có thể được mô tả như sau:

- **Input:**
  - Cho trước một các câu hỏi q.
  - Tập các chủ đề (phân loại) được định nghĩa $C = \{c_1, c_2, \dots , c_n\}$.

  ➔ Tìm câu hỏi q thuộc chủ đề nào?
- **Output:**
  - Nhãn c_i của câu hỏi q_j.

1.2 Đặc điểm dữ liệu câu hỏi pháp quy

Văn bản pháp quy là văn bản có các quy phạm pháp luật do các cơ quan quản lý nhà nước, ở trung ương, cơ quan quyền lực nhà nước, cơ quan quản lý nhà nước ở địa phương ban hành theo thẩm quyền lập quy của mình.

Câu hỏi pháp quy có đặc điểm ý hỏi có thể liên quan đến một hoặc nhiều điều luật. Thông thường, câu hỏi chỉ phân theo một nhãn nhất định, nhưng với câu hỏi pháp quy thì một câu hỏi có thể có một hoặc nhiều hơn một nhãn do ý hỏi của câu hỏi có liên quan đến nhiều điều luật khác nhau mà không thể ghép chung làm một.

Ví dụ: câu hỏi "*Chi phí cho tổ chức công chứng với giao dịch về quyền sử dụng đất gắn liền với nhà ở?*" có ý hỏi thuộc lĩnh vực "công chứng" và lĩnh vực "phí và lệ phí".

1.3 Một số nghiên cứu liên quan

1.3.1 Một số nghiên cứu cho phân loại đa nhãn

Nhóm nghiên cứu David Vilar, Maria Jose Castro và Emilio Sanchis[17] đã có nghiên cứu về phân loại đa nhãn sử dụng mô hình đa thức. Áp dụng các quy tắc phân loại đa nhãn, nhóm nghiên cứu đã xem xét nhiệm vụ phân loại văn bản. Trong đó, mỗi văn bản được gán một vectơ W chiều ứng với số lượng từ, trong đó W là kích thước của từ vựng. Biểu diễn này được gọi là túi của từ (bag-of-words). Nhóm nghiên cứu đã sử dụng phân loại Naive Bayes trong phần khởi tạo mô hình sự kiện đa thức của nó.

Kết quả về phân loại văn bản với kho ngữ liệu Reuters-21578 của họ cho thấy cách tiếp cận xác suất tích lũy sau thực hiện tốt hơn các bộ phân loại nhị phân được sử dụng rộng rãi nhất.

1.3.2 Một số nghiên cứu cho phân loại câu hỏi tiếng Việt

Hiện nay đã có rất nhiều nghiên cứu phân loại câu hỏi tiếng Việt và đạt được một số thành tựu nhất định. Điển hình là một số các nghiên cứu về học sâu đạt kết quả khá tốt như:

Phân loại câu hỏi không thành thật[8] được xuất bản năm 2019 sử dụng kiến trúc mạng nơ-ron hồi quy Recurrent Neural Network (RNN) như một Long Short-Term Memory (LSTM) và một Gated Recurrent Units (GRU). Họ sử dụng LSTM trên một vec-tơ từ được đào tạo để nắm bắt thông tin ngữ nghĩa và cú pháp. LSTM được sử dụng để tránh vấn đề vanishing gradient (gradient có giá trị nhỏ dần theo từng lớp khi thực hiện lan truyền ngược).

Bên cạnh đó cũng có nghiên cứu về phân loại câu hỏi chuyên sâu sử dụng mạng thần kinh tích chập Convolutional Neural Networks (CNNs)[11] được xuất bản năm 2017. Ý tưởng chính của họ trong nghiên cứu này là mở rộng dựa trên công việc hiện có để tạo ra một CNN hai lớp đó là phân loại câu hỏi thành các danh mục chính và phụ của chúng.

1.4 Các phương pháp phân loại câu hỏi

Hầu hết các cách tiếp cận bài toán phân loại câu hỏi thuộc 2 loại : tiếp cận dựa trên luật và tiếp cận dựa trên học máy.

Tiếp cận dựa trên luật[3] là cách tiếp cận được cho là đơn giản nhất để phân loại câu hỏi. Trong cách tiếp cận này, việc phân loại câu hỏi dựa vào các luật ngữ pháp viết tay.

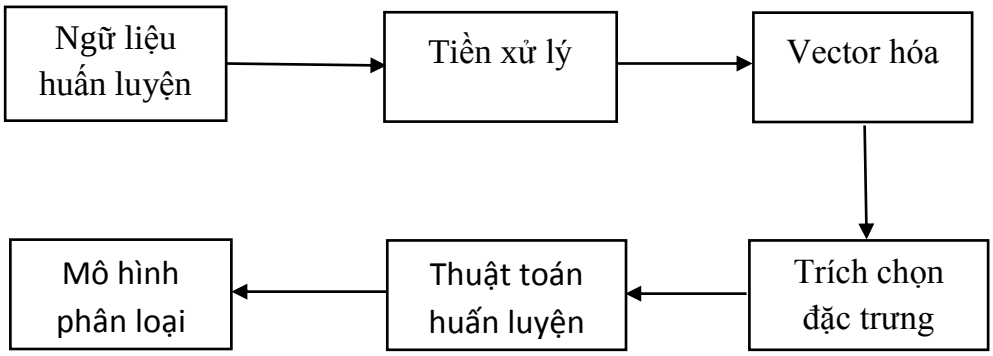
Tiếp cận dựa trên học máy[3] là cách tiếp cận được sử dụng phổ biến rộng rãi để giải quyết bài toán phân loại câu hỏi.

Cách tiếp cận dựa trên học máy chia làm hai nhóm là nhóm các phương pháp học máy truyền thống và nhóm các phương pháp sử dụng mạng nơ-ron (Neural NetWork). Nhóm các phương pháp học máy truyền thống thường được sử dụng như là tính xác suất Naïve Bayes, Maximum Entropy, cây quyết định (decision Tree), lân cận (Nearest-Neighbors), Máy Vector hỗ trợ (Support Vector machine - SVM), K-nearest neighbors (KNN), v.v.

1.4.1 Phương pháp học máy truyền thống

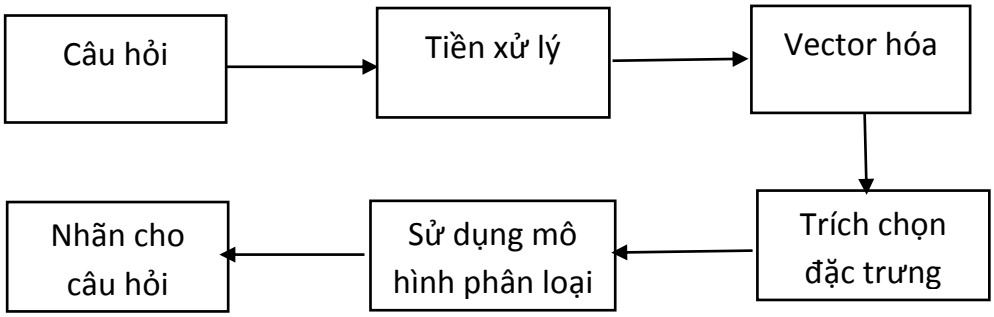
Với các phương pháp học máy truyền thống như SVM, KNN, cây quyết định, v.v thì quá trình phân loại dữ liệu văn bản (document, câu) thường gồm hai giai đoạn sau:

- Giai đoạn huẩn luyện:



Hình 1-1 Mô hình giai đoạn huấn luyện [2]

- Giai đoạn phân lớp:



Hình 1-2 Mô hình giai đoạn phân lớp [2]

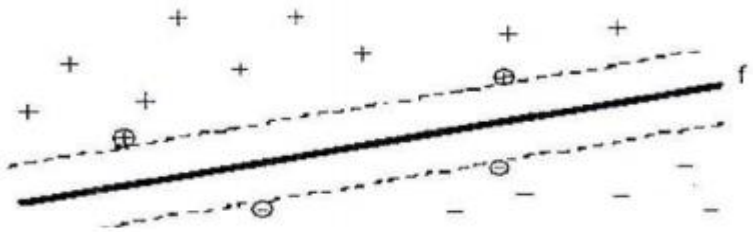
❖ **Mô hình SVM[3]**

Giải thuật máy vector hỗ trợ SVM ra đời từ lý thuyết học thống kê do Vapnik và Chervonenkis xây dựng năm 1995[4]. Đây là một giải thuật phân lớp phổ biến, có hiệu quả cao và đã được áp dụng nhiều trong lĩnh vực khai phá dữ liệu và nhận dạng.

Phương pháp này thực hiện phân lớp dựa trên nguyên lý Cực tiểu hóa rủi ro có cấu trúc SRM (Structural Risk Minimization) [5], được xem là một trong các phương pháp phân lớp giám sát không tham số tinh vi.

SVM cho trước một tập dữ liệu huấn luyện bao gồm dữ liệu cùng với nhãn của chúng thuộc các lớp cho trước, được biểu diễn trong không gian vector, trong đó mỗi dữ liệu là một điểm, phương pháp này tìm ra một siêu phẳng quyết định tốt nhất có thể chia các điểm trên không gian này thành hai lớp riêng biệt tương ứng là lớp (+) và lớp (-). Chất lượng của siêu phẳng được quyết định bởi khoảng cách (gọi là biên hay lề) của điểm dữ liệu gần nhất của mỗi lớp đến mặt phẳng này. Khi đó, khoảng cách biên càng lớn thì mặt phẳng quyết định càng tốt, đồng thời việc phân loại càng chính xác.

Mục đích của phương pháp SVM là tìm được khoảng cách biên lớn nhất, điều này được mình họa như sau:



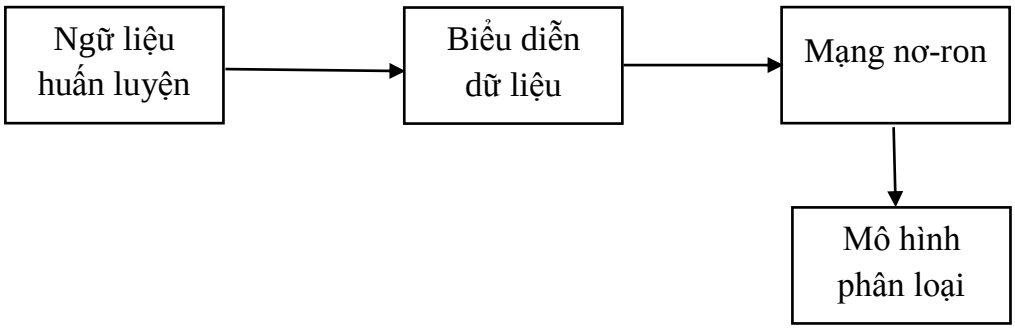
Hình 1-3 Siêu phẳng phân chia dữ liệu học thành 2 lớp (+) và (-) với khoảng cách biên là lớn nhất. Các biên gần nhất (điểm được khoanh tròn) là các Support Vector[5]

Đây là mô hình mạnh và chính xác nhất trong một số các mô hình nổi tiếng về phân lớp dữ liệu.

1.4.2 Phương pháp sử dụng mạng nơ-ron

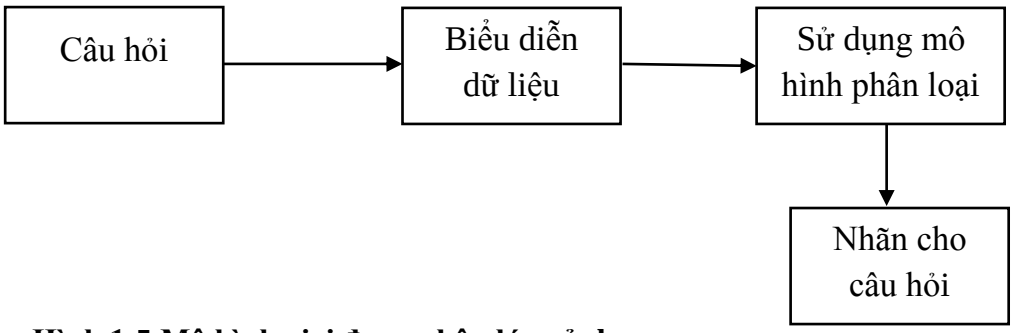
Với phương pháp sử dụng mạng nơ-ron như LSTM, CNN, RNN, v.v thì quá trình phân loại dữ liệu văn bản cũng gồm hai giai đoạn:

- Giai đoạn huấn luyện:



Hình 1-4 Mô hình giai đoạn huấn luyện sử dụng mạng nơ-ron.

- Giai đoạn phân lớp:

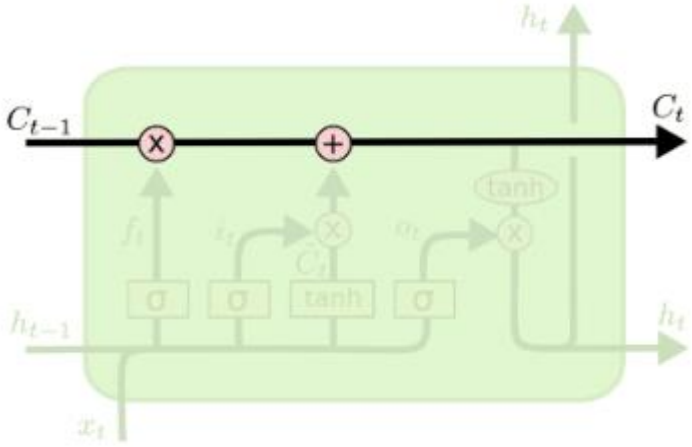


Hình 1-5 Mô hình giai đoạn phân lớp sử dụng mạng nơ-ron.

❖ **Mô hình LSTM[22]**

LSTM (Long short term memory) là mô hình có khả năng học các phụ thuộc dài hạn tức là có khả năng ghi nhớ thông tin quá khứ và trong khi dự đoán các giá trị tương lai.

Chìa khóa của LSTM là trạng thái tế bào (cell state) - chính đường nằm ngang C_{t-1} đến C_t phía trên của sơ đồ hình vẽ, nó như một dạng bang chuyền. Trạng thái tế bào sử dụng để lưu trữ và lan truyền các thông tin có ích trong mạng, nó tương tự như một bộ nhớ cục bộ của mạng.



Hình 1-6 Tế bào trang thái LSTM giống như một băng chuyền [22]

Mấu chốt của LSTM là trạng thái ô, đường ngang chạy dọc theo đỉnh của sơ đồ. Trạng thái tế bào giống như một băng chuyền. Nó chạy thẳng qua toàn bộ chuỗi, chỉ một vài tương tác tuyến tính nhỏ được thực hiện. Điều này làm cho thông tin ít có khả năng thay đổi trong suốt quá trình lan truyền.

❖ **Mô hình BERT[23]**

BERT (Bidirectional Encoder Representations from Transformers) được hiểu là một mô hình học trước hay còn gọi là pre-train model, học các vector đại diện theo ngữ cảnh hai chiều của từ, được sử dụng để chuyển sang các bài toán khác trong lĩnh

vực xử lý ngôn ngữ tự nhiên. BERT đã thành công trong việc cải thiện những công việc trong việc tìm ra đại diện của từ trong không gian số (không gian mà máy tính có thể hiểu được) thông qua ngữ cảnh của nó.

Mô hình BERT đã tạo các biểu diễn theo ngữ cảnh dựa trên các từ trước và sau đó để dẫn đến một mô hình ngôn ngữ với ngữ nghĩa phong phú hơn. Điều này cho thấy mô hình BERT mở rộng khả năng của các phương pháp trước đây.

Các mô hình ngôn ngữ dựa trên LSTM (Long Short Term Memory) hai chiều đào tạo một mô hình ngôn ngữ tiêu chuẩn từ trái sang phải và cũng đào tạo một mô hình ngôn ngữ từ phải sang trái (đảo ngược) dự đoán các từ trước, các từ tiếp theo. Sự khác biệt quan trọng là không LSTM nào đưa cả hai mã thông báo trước và sau vào tài khoản cùng một lúc.

Vì vậy, luận văn chọn mô hình BERT để thực hiện nghiên cứu lần này.

1.5 Kết luận chương

Chương này đã giới thiệu tổng quan bài toán phân loại câu hỏi, nêu bật được đặc điểm của dữ liệu câu hỏi pháp quy, đưa ra được các nghiên cứu phân loại câu hỏi liên quan và giới thiệu được một số phương pháp phân loại câu hỏi.

CHƯƠNG 2: PHÂN LOẠI CÂU HỎI PHÁP QUY TIẾNG VIỆT SỬ DỤNG MÔ HÌNH BERT

2.1 Bài toán phân loại đa nhãn câu hỏi tiếng Việt

Phân loại đa nhãn[15] là nhiệm vụ gán mỗi cá thể trong số các cá thể đã cho vào một tập hợp các lớp được xác định trước, trong một miền mà một cá thể có thể đồng thời thuộc một số lớp.

Bài toán phân loại đa nhãn là bài toán phân loại mà mục tiêu cho một mẫu suy nhất từ tập dữ liệu là danh sách n nhãn nhị phân riêng biệt.

Trong phân loại nhiều lớp, mỗi mẫu được gán cho một và chỉ một nhãn, tức mỗi mẫu chỉ có thể thuộc một trong các lớp C. Trong trường hợp đa nhãn, mỗi mẫu có thể thuộc một hoặc nhiều loại.

Bài toán phân loại câu hỏi đa nhãn có thể được mô tả như sau:

- **Input:**
  - Cho trước một câu hỏi tiếng Việt Q.
  - Tập các nhãn (phân loại) được định nghĩa $C = \{c_1, c_2, \dots, c_n\}$.

  ➔ Tìm Q thuộc những nhãn nào?
- **Output:**
  - Tập nhãn $\{c_i\}$ của câu hỏi Q.

Cách tiếp cận phổ biến để phân loại đa nhãn dựa trên việc chuyển đổi bài toán thành một hoặc nhiều cách phân loại đơn nhãn. Phương pháp biến đổi đơn giản nhất là liên quan nhị phân bao gồm các bộ phân loại khác nhau cho các nhãn khác nhau. Nói cách khác, bài toán ban đầu được chuyển thành n phân loại đơn nhãn hai lớp, trong đó n là số nhãn có thể có. Một trong những nhược điểm lớn của phân loại nhị phân là nó có thể loại trừ sự phụ thuộc giữa các nhãn.

2.2 Giải pháp cho bài toán phân loại đa nhãn

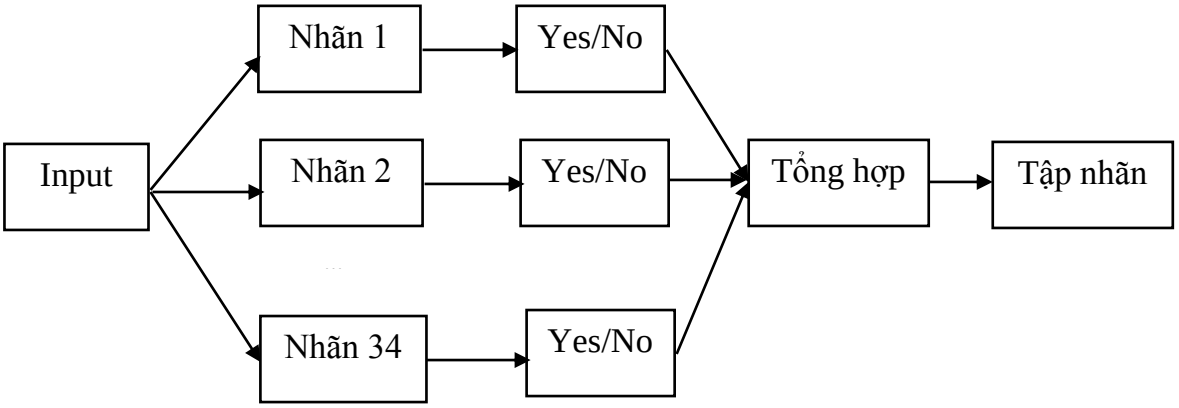
Luận văn mô hình hóa bài toán phân loại đa nhãn dưới dạng bài toán phân lớp. Đầu vào là câu hỏi, đầu ra là các nhãn thuộc vào tập nhãn đã có.

Có hai cách giải quyết cho bài toán phân loại đa nhãn đó là:

- Xây dựng nhiều bộ phân loại nhị phân. Mỗi bước một nhãn thì có một bộ phân loại nhị phân và kiểm tra Yes/No nó có thuộc vào lớp đấy không.
- Xây dựng bộ phân loại đa nhãn.

2.2.1 Giải pháp theo phân loại nhị phân

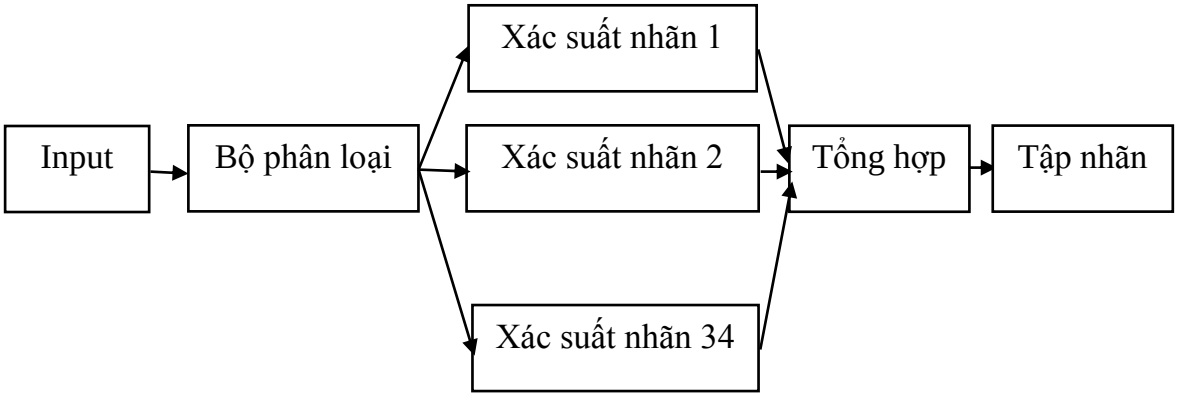
Luận văn xây dựng 34 bộ phân loại nhị phân. Mục đích của bộ phân loại nhị phân là xác định xem câu hỏi đó có chứa nhãn thuộc loại đó hay không. Mỗi bộ phân loại nhị phân có một nhãn. Cần xác định nhãn cho một câu hỏi mới thì luận văn cho chạy qua 34 bộ phân loại. Cái nào trả lời Yes thì nó là nhãn cho câu hỏi đó.



Hình 2-1 Mô hinh giải pháp phân loại theo phân loại nhị phân

2.2.2 Giải pháp theo phân loại đa nhãn

Luận văn xây dựng bộ phân lớp 34 nhãn. Để xác định nhãn cho một câu hỏi mới thì luận văn cho chạy một lần phân lớp lấy xác suất rồi so sánh các xác suất đó với ngưỡng (chọn ngưỡng là 0.5). Lớp nào có xác suất lớn hơn hoặc bằng ngưỡng thì nó là nhãn cho câu hỏi đó. Nếu trong trường hợp các lớp đều có xác suất nhỏ hơn ngưỡng thì coi đó là bài toán phân loại đa lớp, chọn lớp có xác suất lớn nhất là nhãn của câu hỏi đó.



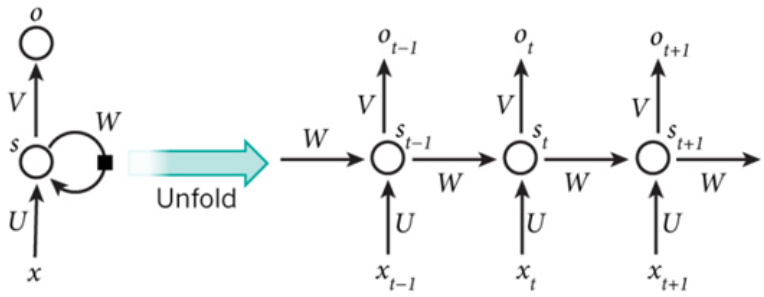
Hình 2-2 Mô hình giải pháp phân loại theo phân loại đa nhãn

2.3 Một số mô hình học sâu

2.3.1 Mô hình mạng nơ-ron hồi quy (RNN - Recurrent Neural Network)

RNN[24] là một chuỗi các khối mạng nơ-ron được liên kết với nhau như một chuỗi. Mỗi một khối sẽ chuyển tin nhắn đến khối tiếp theo. RNN coi dữ liệu đầu vào là một chuỗi (sequence) liên tục, nối tiếp nhau theo thứ tự thời gian.

Mô hình hoạt động của RNN có thể được mô tả trong hình dưới đây:



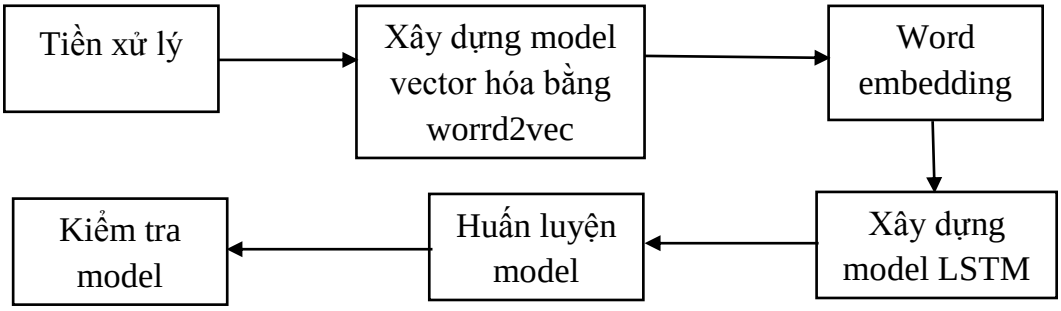
Hình 2-3 Mô hình một mạng nơ-ron hồi quy

RNN là một mô hình mạng nơ-ron có "bộ nhớ" để lưu trữ thông tin của phần xử lý trước đó. RNN chỉ tỏ ra hiệu quả với chuỗi dữ liệu có độ dài không quá lớn (short-term memory hay còn gọi là long-term dependency problem). Nguyên nhân của vấn đề này là do vanishing gradient problem (gradient có giá trị nhỏ dần theo từng lớp khi thực hiện lan truyền ngược).

Ứng dụng trong bài toán phân lớp

Việc giải bài toán phân loại sẽ bao gồm việc giải quyết một chuỗi các bài toán nhỏ hơn. Chuỗi các bài toán nhỏ hơn này được gọi là pipline của mô hình học máy.

Phân loại văn bản sử dụng mô hình mạng RNN gồm các bước sau:

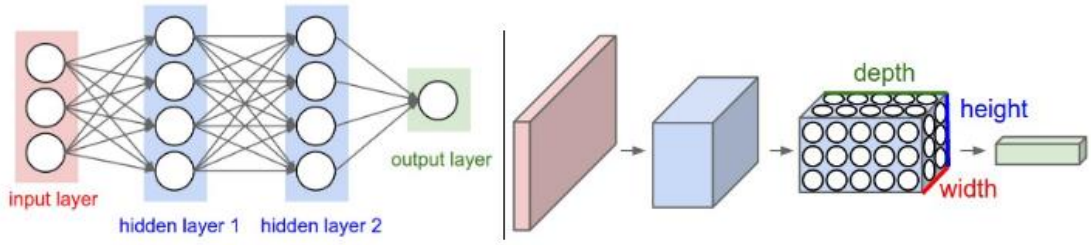


Hình 2-4 Các bước của bài toán phân loại văn bản sử dụng mạng nơ-ron RNN.

2.3.2 Mô hình mạng nơ-ron tích chập (Convolutional Neural Network – CNN)

Mạng CNN[25] là một tập hợp các lớp Convolution chồng lên nhau và sử dụng các hàm nonlinear activation như ReLU và tanh để kích hoạt các trọng số trong các node.

CNN đơn giản là một chuỗi các lớp và mỗi lớp của ConvNet chuyển đổi một lượng kích hoạt thành một lượng kích hoạt khác thông qua một chức năng có thể phân biệt. CNN sử dụng ba loại lớp chính để xây dựng kiến trúc: Lớp Convolutions (Convolutional Layer), Lớp tổng hợp (Poolong Layer) và Lớp được kết nối đầy đủ (Fully-Connected Layer) (chính xác như được thấy trong các Mạng thần kinh thông thường). Các lớp này sẽ được xếp chồng để tạo thành một kiến trúc CNN đầy đủ.



Hình 2-5 Bên trái: Mạng nơ-ron ba lớp thông thướng. Bên phải: Một CNN sắp xếp theo nơ-ron của nó theo ba chiều (chiều rộng, chiều cao, chiều sâu).

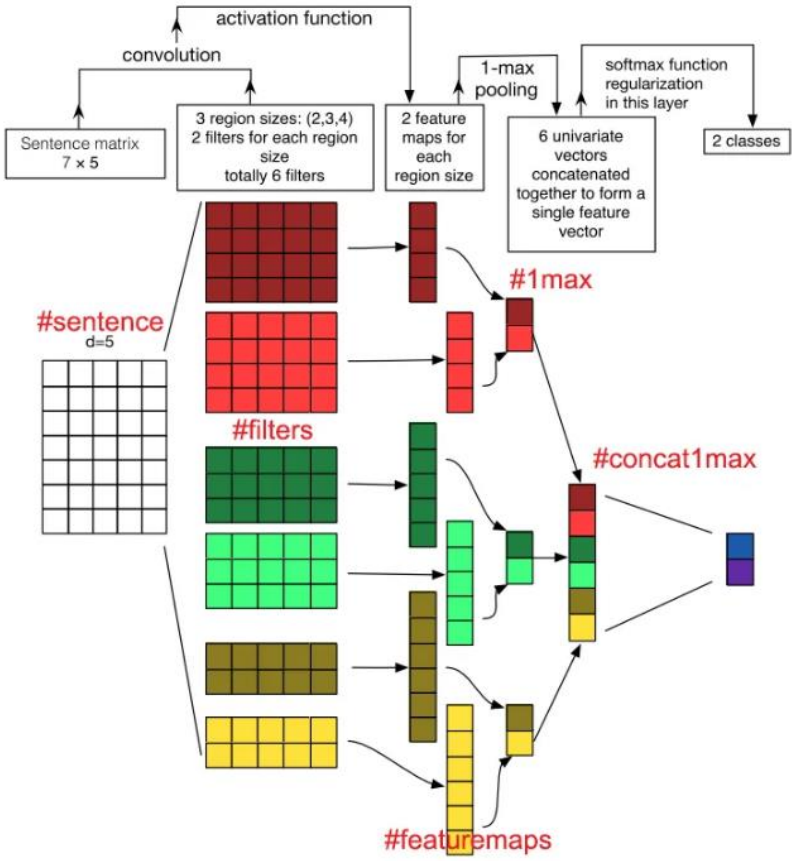
Trong mô hình CNN có 2 khía cạnh cần quan tâm là tính bất biến (Location Invariance) và tính kết hợp (Compositionality). Với cùng một đối tượng, nếu đối tượng này được chiếu theo các gốc độ khác nhau (translation, rotation, scaling) thì độ chính xác của thuật toán sẽ bị ảnh hưởng đáng kể.

CNNs cho ra mô hình với độ chính xác rất cao. Cũng giống như cách con người nhận biết các vật thể trong tự nhiên.

Ứng dụng trong bài toán phân lớp

Trong bài toán phân lớp văn bản, mô hình CNN sử dụng bộ lọc có các kích thước khác nhau và mỗi kích thước có 2 bộ lọc khác nhau. Các bộ lọc thực hiện nhân tích chập (convolution) lên ma trận của câu văn bản đầu vào và mỗi bộ lọc tạo ra một map lưu trữ các đặc trưng (featues map). Các map đặc trưng này từng map qua sẽ đi qua 1-max pooling. Tức là giá trị lớn nhất trong mỗi map đặc trưng sẽ được lưu lại.

Do vậy, một vector có một phần tử được tạo ra ở mỗi map đặc trưng. Sau đó, các giá trị này được nối lại với nhau tạo nên lớp áp chót. Và cuối cùng, kết quả này đi qua một hàm softmax và nhận được là vector đặc trưng và dùng nó để dự đoán nhãn cho văn bản.



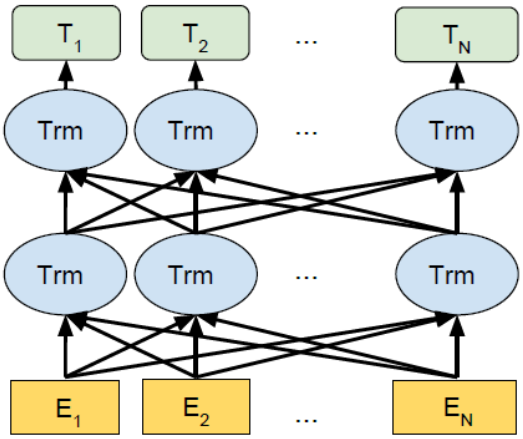
Hình 2-6 Kiến trúc mô hình CNN dùng trong phân loại văn bản.

2.4 Giới thiệu phương pháp BERT

BERT[26](Bidirectional Encoder Representations from Transformers) (tạm dịch: Mô hình mã hóa hai chiều dữ liệu từ các khối Transformer), là một phương pháp kỹ thuật được xây dựng dựa trên mô hình mạng mô phỏng theo hệ thống nơ-ron thần kinh của con người (neural network) dùng để đào tạo trước (pre-train) quá trình xử lý ngôn ngữ tự nhiên.

Điểm đột phá của BERT nằm ở khả năng huấn luyện các mô hình ngôn ngữ dựa trên toàn bộ tổ hợp các từ trong một câu hoặc truy vấn (huấn luyện hai chiều), thay vì cách thức huấn luyện truyền thống dựa trên thứ tự xuất hiện của các từ (từ trái qua phải hoặc kết hợp giữa trái qua phải và phải qua trái).

Kiến trúc mô hình BERT là một bộ mã hóa Transformer hai chiều (bidirectional Transformer encoder). Bộ mã hóa hai chiều (bidirectional encoder) là một tính năng nổi bật giúp phân biệt BERT với OpenAI GPT (sử dụng từ trái sang phải Transformer) và ELMo (kết hợp giữa huấn luyện từ trái sang phải và một mạng riêng rẽ phải sang trái LSTM).



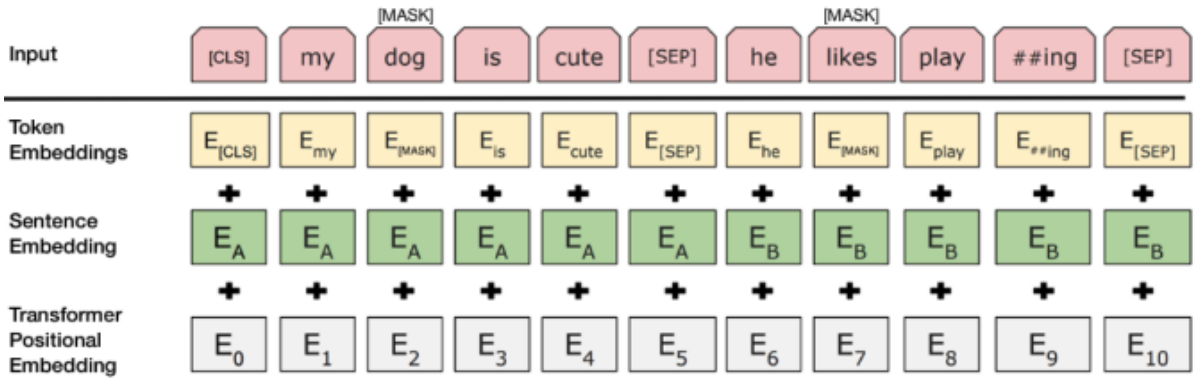
Hình 2-7 Kiến trúc của mô hình BERT [28]

2.5 Mô hình phân loại câu hỏi pháp quy tiếng Việt sử dụng mô hình BERT

2.5.1 Biểu diễn đầu vào

Đầu vào có thể là biểu diễn của một câu văn bản đơn hoặc một cặp câu văn bản(ví dụ: [Câu hỏi, câu trả lời]) được đặt thành 1 chuỗi tạo bởi các từ.

Chuỗi đầu vào BERT biểu diễn một cách tường minh cả văn bản đơn và cặp văn bản. Với văn bản đơn, chuỗi đầu vào BERT là sự ghép nối của token phân loại đặc biệt “<cls>”, token của chuỗi văn bản, và token phân tách đặc biệt “<sep>”. Với cặp văn bản, chuỗi đầu vào BERT là sự ghép nối của “<cls>”, token của chuỗi văn bản đầu, “<sep>”, token của chuỗi văn bản thứ hai, và “<sep>”.

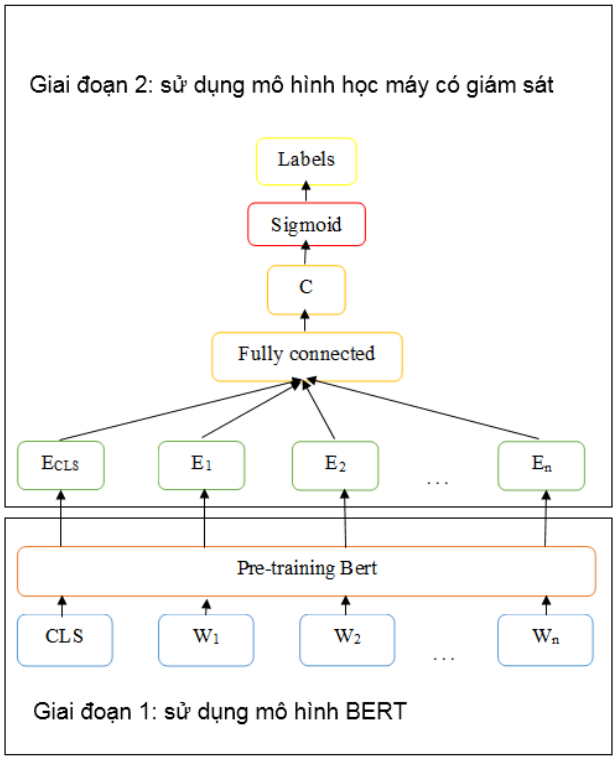


Hình 2-8 Mô hình đại diện đầu vào của BERT [28].

Trong trường hợp các cặp câu được gộp lại với nhau thành một chuỗi duy nhất, chúng ta phân biệt các câu theo 2 cách là tách chúng bởi một token đặc biệt [SEP] và thêm một segment embedding cho mỗi câu.

2.5.2 Mô hình huấn luyện

Mô hình huấn luyện gồm hai giai đoạn chính là học mô hình huấn luyện trước sử dụng mô hình BERT và học có giám sát để đào tạo lớp cuối cho nhiệm vụ phân loại.



Hình 2-3 Mô hình huấn luyện phân loại đa nhãn sử dụng mô hình Bert.

Các token của câu sẽ được đưa vào mô hình huấn luyện trước Bert tạo ra các Embedding. Các Embedding này được đưa vào Fine-tuning sử dụng mô hình học có giám sát để phân loại.

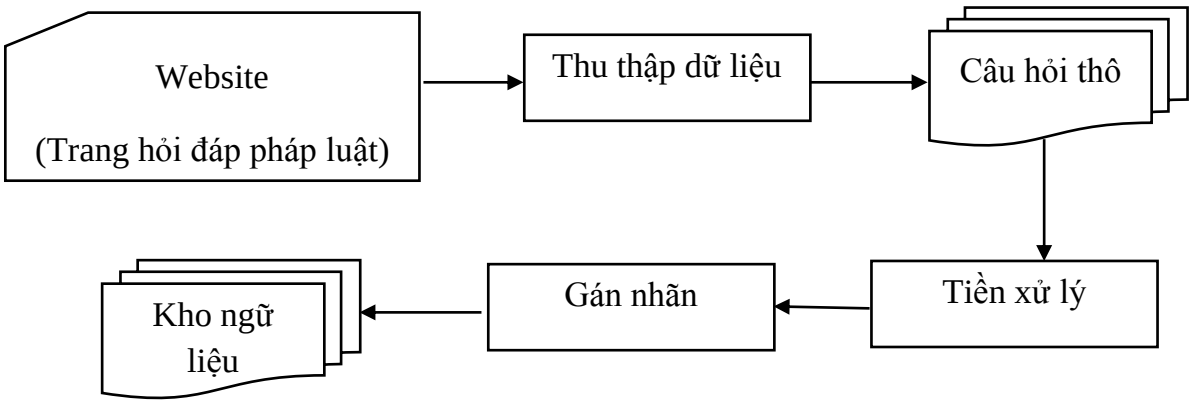
2.6 Kết luận chương

Nội dung chương đã giới thiệu được bài toán phân loại đa nhãn câu hỏi tiếng Việt, giới thiệu được một số mô hình học sâu, giới thiệu phương pháp BERT và đưa ra được mô hình phân loại câu hỏi pháp quy tiếng Việt sử dụng mô hình BERT.

CHƯƠNG 3: THỰC NGHIỆM ĐÁNH GIÁ

3.1 Xây dựng kho ngữ liệu

Việc thực hiện xây dựng kho ngữ liệu luận văn đã thực hiện theo từng giai đoạn trong mô hình dưới đây:



Hình 3-1 Mô hình xây dựng kho ngữ liệu.

3.1.1 Thu thập dữ liệu

Luận văn lấy dữ liệu từ 3 trang web:

- Hỏi đáp và tư vấn pháp luật: https://hdpl.moj.gov.vn/Pages/home.aspx
- Hỏi đáp pháp luật: https://hoidapphapluat.net/
- Hệ thống pháp luật Việt Nam, chuyên trang pháp luật và tư vấn: http://hethongphapluatvietnam.com/hoi-dap-phap-luat.html

Dữ liệu gồm hơn 5000 câu hỏi lĩnh vực pháp luật. Nội dung về những hỏi đáp về quy định, thủ tục và điều luật của pháp luật.

3.1.2 Tiền xử lý

Dữ liệu sau khi thu thập được từ 3 trang web sẽ được tiến hành tiền xử lý. Luận văn thực hiện tiền xử lý dữ liệu bằng cách loại bỏ một số nhiễu như: câu sai chính tả, lỗi font.

3.1.3 Gán nhãn

Tập nhãn luận văn xây dựng gồm 34 nhãn.

Bảng 3-1 Bảng nhãn và ví dụ

Nhãn	Ví dụ
Ban hành văn bản quy phạm pháp luật	Văn bản quy phạm pháp luật hết hiệu lực trong trường hợp nào?
Bảo hiểm	Quy định của pháp luật về thời gian nghỉ hưởng chế độ thai sản?
Bảo vệ môi trường	Tập trung chăn nuôi quy mô lớn có phải đáp ứng điều kiện về môi trường gì không?
Cán bộ, công chức, viên chức	Pháp luật quy định về nghĩa vụ của công chức khi thi hành công vụ như thế nào?
Công chứng	Công chứng viên thành lập văn phòng công chứng cần làm thế nào?
Công dân	Người nhà có thể xin hộ giấy xác nhận tình trạng độc thân để đăng ký kết hôn với người nước ngoài không hay phải là người trực tiếp?
Cư trú	Chủ hộ muốn tách hộ khẩu cho thành viên có được không?
Dân sự	Xin cho biết, pháp luật có quy định về vấn đề trổ cửa sổ sang nhà hàng xóm không?
Giao thông đường bộ	Mua chiếc xe ô tô cũ, mua qua nhiều người phải làm những thủ tục gì để được sang tên chính chủ, việc đăng ký là khác tỉnh?
Giám định tư pháp	Quy định của pháp luật về văn phòng giám định tư pháp?
Hình sự	Bị phạt tù nhưng được hưởng án treo về tội đánh bạc, nay tiếp tục vi phạm về tội đánh bạc thì bị xử lý như thế nào?
Hôn nhân và gia đình	Tài sản được mua từ tài sản riêng của vợ/chồng trong thời kỳ hôn nhân có phải là tài sản chung của vợ chồng không?
Khiếu nại, tố cáo	Công dân được quyền kiếu nại quyết định hành chính của cơ quan hành chính không?
Kinh tế	Thời hạn gửi giấy đòi nợ của chủ nợ khi doanh nghiệp phá sản là bao lâu?
Lao động	Hợp đồng thử việc có thời gian bao lâu?
Lý lịch tư pháp	Cập nhật thông tin lý lịch tư pháp trong trường hợp người bị kết án được xoá án tích thực hiện như thế nào?
Nhà ở	Có được thế chấp nhà ở hình thành trong tương lai tại tổ chức không phải là tổ chức tín dụng không?

Nuôi con nuôi	Trẻ bị bỏ rơi được hiểu như thế nào?
Phí và lệ phí	Lệ phí cấp giấy chứng nhận đăng ký xe?
Phòng, chống ma túy	Muốn được cai nghiện ma túy tại gia đình thì cần đăng ký như thế nào?
Quản lý, sử dụng	Tài sản công tại cơ quan nhà nước được bán thanh lý trong trường hợp nào?
Quốc phòng	Đã đăng ký nghĩa vụ quân sự mà thay đổi nơi cư trú thì có phải làm thủ tục đăng ký thay đổi không?
Quốc tịch Việt Nam	Hồ sơ xin nhập quốc tịch Việt Nam đối với trường hợp nhập quốc tịch việt nam của người không có quốc tịch đã cư trú ổn định ở việt nam?
Thi hành án	Tài sản chung của vợ chồng bị cưỡng chế thi hành án thì xử lý như thế nào?
Thuế	Lệ phí trước bạ đối khi cấp giấy chứng nhận về đất?
Trách nhiệm bồi thường của Nhà nước	Phạm vi trách nhiệm bồi thường của nhà nước trong hoạt động quản lý hành chính?
Tố tụng	Hết thời hiệu khởi kiện về thừa kế và các thừa kế có tranh chấp thì giải quyết thế nào?
Tổ chức chính phủ	Người có quyền yêu cầu cấp bản sao học bạ?
Tổ chức cơ quan, chính quyền	Những nhiệm vụ quyền hạn của chủ tịch ủy ban nhân dân xã?
Xây dựng	Đề nghị cho biết những công trình xây dựng nào phải xin cấp Giấy phép xây dựng?
Xử lý vi phạm hành chính	Pháp luật quy định như thế nào về hành vi vi phạm hành chính, hình thức xử phạt và biện pháp khắc phục hậu quả trong hoạt động trọng tài thương mại?
Đất đai	Được Nhà nước giao đất theo diện giãn dân có được xem xét để được cấp giấy chứng nhận quyền sử dụng đất không?
Đấu giá tài sản	Các tài sản phải thông qua bán đấu giá?
Đầu tư	Những dự án đầu tư ra nước ngoài như thế nào thì phải được Quốc Hội quyết định chủ trương đầu tư?

3.1.4 Cách gán nhãn thủ công

Giai đoạn gán nhãn thủ công luận văn thực hiện hai người gán nhãn. Luận văn sử dụng độ đo Cohen's kappa tính toán độ tương đồng gán nhãn giữa hai người.

Áp dụng vào bộ dữ liệu, kết quả đo độ tương đồng phân loại giữa hai người là 0.99. Kết quả cho thấy hai người gán nhãn khá tương đồng với nhau.

3.1.5 Thống kê kho ngữ liệu

Dữ liệu gồm 5896 câu lĩnh vực pháp luật. Nội dung về những câu hỏi về pháp luật. Kho ngữ liệu bao gồm 5896 câu, tổng 324095 từ, số từ trung bình trên một câu là 54 từ, số từ (không tính lặp) trên toàn bộ kho ngữ liệu là 1285 từ. Tổng số nhãn là 34.

Bảng 3-2 Thống kê tần suất các nhãn trong kho ngữ liệu

STT	Nhãn	Số câu hỏi	Tỉ lệ trong kho ngữ liệu (%)	STT	Nhãn	Số câu hỏi	Tỉ lệ trong kho ngữ liệu (%)
1	Ban hành văn bản quy phạm pháp luật	18	0,31	18	Nuôi con nuôi	135	2,29
2	Bảo hiểm	29	0,49	19	Phí và lệ phí	83	1,41
3	Bảo vệ môi trường	12	0,20	20	Phòng, chống ma túy	47	0,80
4	Cán bộ, công chức, viên chức	14	0,24	21	Quản lý, sử dụng	13	0,22
5	Công chứng	327	5,55	22	Quốc phòng	16	0,27
6	Công dân	405	6,87	23	Quốc tịch Việt Nam	67	1,14
7	Cư trú	162	2,75	24	Thi hành án	636	10,79
8	Dân sự	1234	20,93	25	Thuế	30	0,51
9	Giao thông đường bộ	65	1,10	26	Trách nhiệm bồi thường	120	2,04

					của Nhà nước		
10	Giám định tư pháp	22	0,37	27	Tố tụng	317	5,38
11	Hình sự	484	8,21	28	Tổ chức chính phủ	193	3,27
12	Hôn nhân và gia đình	552	9,36	29	Tổ chức cơ quan, chính quyền	20	0,34
13	Khiếu nại, tố cáo	42	0,71	30	Xây dựng	24	0,41
14	Kinh tế	114	1,93	31	Xử lý vi phạm hành chính	263	4,46
15	Lao động	90	1,53	32	Đất đai	469	7,95
16	Lý lịch tư pháp	91	1,54	33	Đấu giá tài sản	30	0,51
17	Nhà ở	75	1,27	34	Đầu tư	28	0,47

Bảng 3-3 Thống kê câu hỏi theo lượng nhãn

Số nhãn	Số câu hỏi
1	5579
2	307
3	6
4	4

3.2 Thiết lập thực nghiệm

Với dữ liệu chuẩn bị cho thực nghiệm, luân văn lấy được 5896 câu hỏi pháp quy tiếng Việt. Từ dữ liệu này, luận văn chia thành 10 bộ dữ liệu, trong đó mỗi bộ dữ liệu xây dựng bằng cách ngẫu nhiên trong tập dữ liệu có. Kết quả thu được ở 10 lần thực nghiệm sẽ được tính trung bình để ra được kết quả của thực nghiệm.

Để đánh giá kết quả của việc xác định thực thể và thuộc tính ta đánh giá thông qua độ chính xác (precision), độ bao phủ (recall) và F1.

3.3 Công cụ thực nghiệm

Luận văn sử dụng 2 công cụ thực nghiêm là sklearn svm Linear SVC sử dụng cho mô hình SVM và simpletransformers sử dụng cho hai mô hình còn lại là BERT multilingual và PHOBERT.

Cả 3 mô hình đều sử dụng công cụ python.

3.4 Các mô hình thực nghiệm

Phương pháp phân loại dựa trên học máy đươc chia làm 2 nhóm chính là phương pháp học máy truyền thống và phương pháp học máy sử dụng mạng nơ-ron. Do vậy, luận văn đã lựa chọn thực nghiệm hai mô hình chính đại diện cho hai nhóm phương pháp đó là mô hình SVM đại diện cho nhóm phương pháp học máy truyền thống, mô hình BERT đại diện cho nhóm phương pháp học máy sử dụng mạng nơ-ron.

❖ Mô hình SVM

Mô hình SVM luận văn thực nghiệm sử dụng pipeline để thực hiện các bước theo trình tư với một đối tượng, dùng TfidfVectorizer để thay đổi vectơ văn bản được tạo bởi bộ vectơ đếm và dùng hỗ trợ máy vector LinearSVC.

❖ Mô hình BERT multilingual

BERT multilingual là một mô hình của google BERT đa ngôn ngữ. Mô hình được đào tạo trước trên 104 ngôn ngữ hàng đầu có Wikipedia lớn nhất bằng cách sử dụng mục tiêu tạo mô hình ngôn ngữ bị che (masked language modeling - MLM). Mô hình này phân biệt chữ hoa chữ thường.

Luận văn sử dụng mô hình huấn luyện trước bert-base-multilingual-cased. Trong mô hình huấn luyện, luận văn sử dụng ClassificationModel của simpleTransformer để tạo mô hình huấn luyện. Luận văn thực hiện huấn luyện với số lượng train epochs là 10.

❖ Mô hình PHOBERT

PHOBERT[27] là mô hình huấn luyện trước, đặc biệt chỉ huấn luyện dành riêng cho tiếng Việt. PHOBERT huấn luyện dựa trên kiến trúc và cách tiếp cận giống RoBERTa.

Tương tự như BERT, PHOBERT cũng có hai phiên bản là PHOBERT base với 12 transformers block và PHOBERT large với 24 transformers block.

Trong nghiên cứu này, luận văn thử nghiệm với mô hình PHOBERT base. Luận văn sử dụng bpe của mô hình để encode một câu hỏi thành một danh sách các

subword. Mô hình có dict chứa từ điển sẵn có của PHOBERT. Luận văn sẽ sử dụng từ điển này để giúp ánh xạ ngược từ subword về id của nó trong bộ từ vựng được cung cấp sẵn.

Xây dựng model huấn luyện PHOBERT có hai lựa chọn là Fairseq và Transformer. Ở đây luân văn lựa chọn thử nghiệm với Transformer và sử dụng BertForSequenceClassification để tạo model. Trong phân loại binary luận văn thực hiện huấn luyện với số lượng epochs là 10, batch_size là 32, hidden_dropout_prob là 0.1.

Với mỗi mô hình luận văn đều thực nghiệm hai phương pháp là phân loại nhị phân và phân loại đa nhãn.

3.5 Kết quả thực nghiệm

3.5.1 Phân loại binary

Bảng 1-4 Kết quả thực nghiệm phân loại binary của 3 mô hình

Mô hình	PRECISION(%)	RECALL(%)	F1(%)
SVM	92,68	83,64	87,93
BERT multilingual	88,14	85,59	86,85
PHOBERT	88,79	75,28	81,48

Từ bảng kết quả nhận thấy với độ đo F1 mô hình SVM cho kết quả tốt nhất (87,93%), cao hơn mô hình BERT multilingual (86,85%) là 1,08% và cao hơn 6,45% so với mô hình PHOBERT(81,48%).

Mô hình PHOBERT cho kết quả thấp nhất.

3.5.2 Phân loại đa nhãn

Bảng 3-5 Bảng kết quả thực nghiệm của 3 mô hình

Phương pháp	SVM	BERT	PHOBERT
Đa nhãn	87,39	89,47	86,65
Binary	87,93	86,85	81,48

Từ bảng kết quả nhận thấy:

- Kết quả phân loại đa nhãn sử dụng mô hình BERT multilingual đạt kết quả tốt nhất (89,47%).

- Kết quả thu được từ mô hình SVM theo phương pháp phân loại nhị phân là 87,93% với mô hình SVM theo phương pháp phân loại đa nhãn cao hơn 0,54%. Kết quả thu được từ mô hình PHOBERT theo phương pháp phân loại nhị phân là 81,48% thấp hơn 5,17% so với phương pháp phân loại đa nhãn (86,65%).
- SVM ổn định cho cả hai phương pháp đều trên 87%. Với các mô hình dùng BERT thì phân loại đa nhãn tốt hơn binary. Có thể mạng nơ-ron này đủ phức tạp để nó mô hình hóa được vấn đề học đa nhãn nên nó tốt hơn trong trường hợp đa nhãn.

3.6 Kết luận chương

Chương này đã trình bày được cách thiết lập thực nghiệm, mô tả được các mô hình thực nghiệm, giới thiệu được các công cụ thực nghiệm, đưa ra kết quả và phân tích đánh giá được kết quả thực nghiệm.

KẾT LUẬN

Phân loại câu hỏi tiếng Việt không còn là một vấn đề mới, nhưng phân loại câu hỏi pháp quy tiếng Việt là một nghiên cứu mới mà hiện nay ít có nghiên cứu về vấn đề này.

Khác với phân loại câu hỏi thông thường, câu hỏi pháp quy có đặc điểm ý hỏi có thể liên quan đến một hoặc nhiều điều luật. Thông thường, câu hỏi chỉ phân theo một nhãn nhất định, nhưng với câu hỏi pháp quy thì một câu hỏi có thể có một hoặc nhiều hơn một nhãn do ý hỏi của câu hỏi có liên quan đến nhiều điều luật khác nhau mà không thể ghép chung làm một. Vì vậy việc giải quyết bài toán phân loại câu hỏi pháp quy tiếng Việt có phần phức tạp hơn phân loại câu hỏi thông thường. Từ việc giải quyết bài toán này giúp góp phần đem lại sự thuận tiện cho người dùng trong việc thu thập và tìm kiếm thông tin về pháp luật.

Nhìn chung, luận văn đã đạt được:

- Nghiên cứu cho bài toán phân loại câu hỏi pháp quy Tiếng Việt là bài toán còn ít được nghiên cứu.
- Xây dựng được bộ dữ liệu cho bài toán.
- Nghiên cứu này chỉ là nghiên cứu ban đầu có thể đóng góp bộ dữ liệu cho các nghiên cứu tiếp theo.
- Nghiên cứu một số phương pháp phân loại dưa trên học máy sử dụng mô hình BERT là một mô hình huấn luyện sẵn mà hiện tại đang đạt kết quả phương pháp hiện đại trong xử lý ngôn ngữ tự nhiên.
- Thực nghiệm, phân tích, đánh giá kết quả và tìm ra được trường hợp cho kết quả tốt nhất.

Về hướng phát triển tương lai, luận văn sẽ tiến hành phát triển một tập dữ liệu câu hỏi pháp quy tiếng Việt lớn hơn và nghiên cứu sử dụng thêm nhiều phương pháp, góp phần cải thiện tốt hơn khả năng phân loại. Ngoài ra luận văn sẽ nghiên cứu và thử nghiệm với một số mô hình khác để tìm ra mô hình phù hợp nhất với bài toán phân loại câu hỏi pháp quy tiếng Việt.