HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN DIỆU LINH

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

LUẬN VĂN THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

HÀ NỘI – 2021

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 tinh**

Mã số : **8.48.01.01**

LUẬN VĂN THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC:
PGS. TS. NGÔ XUÂN BÁCH

HÀ NỘI – 2021

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của bản thân. Các số liệu, kết quả trình bày trong luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào trước đây.

Tác giả

Nguyễn Diệu Linh

LỜI CẢM ƠN

Em xin chân thành cảm ơn PGS.TS.Ngô Xuân Bách, bộ môn Khoa học máy tính, Khoa Công nghệ thông tin đã tận tình chỉ dạy và hướng dẫn cho em trong việc lựa chọn đề tài, thực hiện đề tài và viết báo cáo luận văn, giúp cho em có thể hoàn thành tốt luận văn này.

Em xin chân thành cảm ơn các thầy cô giáo Khoa Công nghệ thông tin là những người giảng dạy em, đặc biệt các thầy cô trong khoa Sau đại học đã tận tình dạy dỗ và chỉ bảo em trong suốt 2 năm học .

Em xin chân thành cảm ơn em Nguyễn Thị Minh Phương đã tham gia xây dựng kho ngữ liệu cho bài toán.

Cuối cùng em xin cảm ơn gia đình, bạn bè, những người đã luôn bên cạnh động viên em những lúc khó khăn và giúp đỡ em trong suốt thời gian học tập và nghiên cứu, tạo mọi điều kiện tốt nhất cho em để có thể hoàn thành tốt luận văn của mình.

Mặc dù đã cố gắng hoàn thành nghiên cứu trong phạm vi và khả năng cho phép nhưng chắc chắn sẽ không tránh khỏi những thiếu sót. Em kính mong nhận được sự thông cảm của thầy cô và các bạn.

Em xin chân thành cảm ơn!

Hà Nội, 12/2020

Nguyễn Diệu Linh

MỤC LỤC

LỜI CAM ĐOAN ..... **i**
LỜI CẢM ƠN ..... **ii**
MỤC LỤC ..... **iii**
DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT ..... **v**
DANH MỤC HÌNH VẼ ..... **vi**
DANH MỤC BẢNG BIỂU ..... **vii**
MỞ ĐẦU **1**
CHƯƠNG 1: BÀI TOÁN PHÂN LOẠI CÂU HỎI **4**
1.1 Giới thiệu bài toán phân loại câu hỏi 4
1.2 Đặc điểm dữ liệu câu hỏi pháp quy 5
1.3 Một số nghiên cứu liên quan 6
1.3.1 Một số nghiên cứu cho phân loại đa nhãn 6
1.3.2 Một số nghiên cứu cho phân loại câu hỏi tiếng Việt 7
1.4 Các phương pháp phân loại câu hỏi 8
1.4.1 Phương pháp học máy truyền thống 9
1.4.2 Phương pháp sử dụng mạng nơ-ron 11
1.5 Kết luận chương 16
CHƯƠNG 2: PHÂN LOẠI CÂU HỎI PHÁP QUY TIẾNG VIỆT SỬ DỤNG MÔ HÌNH BERT **17**
2.1 Bài toán phân loại đa nhãn câu hỏi tiếng Việt 17
2.2 Giải pháp cho bài toán phân loại đa nhãn 18
2.2.1 Giải pháp theo phân loại nhị phân 19
2.2.2 Giải pháp theo phân loại đa nhãn 21
2.3 Một số mô hình học sâu 24
2.3.1 Mô hình mạng nơ-ron hồi quy (RNN - Recurrent Neural Network) 24
2.3.2 Mô hình mạng nơ-ron tích chập (Convolutional Neural Network – CNN) 27
2.4 Giới thiệu phương pháp BERT 31
2.5 Mô hình phân loại câu hỏi pháp quy tiếng Việt sử dụng mô hình BERT 33
2.5.1 Biểu diễn đầu vào 33

2.5.2 Mô hình huấn luyện 35
2.6 Kết luận chương 37
CHƯƠNG 3: THỰC NGHIỆM ĐÁNH GIÁ **38**
3.1 Xây dựng kho ngữ liệu 38
3.1.1 Thu thập dữ liệu 39
3.1.2 Tiền xử lý 39
3.1.3 Gán nhãn 39
3.1.4 Thống kê kho ngữ liệu 42
3.2 Thiết lập thực nghiệm 45
3.3 Công cụ thực nghiệm 45
3.4 Các mô hình thực nghiệm 46
3.5 Kết quả thực nghiệm 47
3.5.1 Phân loại binary 47
3.5.2 Phân loại đa nhãn 53
3.6 Kết luận chương 61
KẾT LUẬN **62**
TÀI LIỆU THAM KHẢO **63**

DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
BERT	Bidirectional Encoder Representations from Transformers	Biểu diễn mã hóa hai chiều từ Transformer
CNN	Convolutional Neural Network	Mạng nơ-ron tích chập
GRU	Gated Recurrent Units	Cổng tái Unit
LSTM	Long-Short Term Memory	Mạng bộ nhớ dài-ngắn
MLM	Masked language modeling	Mô hình ngôn ngữ bị che
RNN	Recurrent Neural Network	Mạng nơ-ron hồi quy
SRM	Structural Risk Minimization	Cực tiểu hóa rủi ro có cấu trúc
SVM	Support Vector machine	Máy vector hỗ trợ

DANH MỤC HÌNH VẼ

Hình 1-1 Mô hình giai đoạn huấn luyện 9
Hình 1-2 Mô hình giai đoạn phân lớp 10
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.....11
Hình 1-4 Mô hình giai đoạn huấn luyện sử dụng mạng nơ-ron.....12
Hình 1-5 Mô hình giai đoạn phân lớp sử dụng mạng nơ-ron. 12
Hình 1-6 Tế bào trang thái LSTM giống như một băng chuyền 13
Hình 1-7 LSTM focus f 14
Hình 1-8 LSTM focus I 14
Hình 1-9 LSTM focus c 15
Hình 1-10 LSTM focus o 15
Hình 2-1 Mô hinh giải pháp phân loại theo phân loại nhị phân 19
Hình 2-2 Mô hình giải pháp phân loại theo phân loại đa nhãn.....22
Hình 2-3 Mô hình một mạng nơ-ron hồi quy.....25
Hình 2-4 Vanilla RNN, LSTM, GRU.....26
Hình 2-5 Các bước của bài toán phân loại văn bản sử dụng mạng nơ-ron RNN. ..27
Hình 2-6 Bên trái: Mạng nơ-ron ba lớp thông thướng. Bên phải: Một CNN sắp xếp theo nơ-ron của nó theo ba chiều 28
Hình 2-7 Kiến trúc mô hình CNN dùng trong phân loại văn bản.....30
Hình 2-8 Kiến trúc của mô hình BERT 33
Hình 2-9 Mô hình đại diện đầu vào của BERT 34
Hình 2-10 Mô hình huấn luyện phân loại đa nhãn sử dụng mô hình Bert.....35
Hình 3-1 Mô hình xây dựng kho ngữ liệu. 38
Hình 3-2 Biểu đồ kết quả thực nghiệm phân loại binary của 3 mô hình. 48
Hình 3-3 Biểu đồ kết quả thực nghiệm phân loại đa nhãn của 3 mô hình.....54

DANH MỤC BẢNG BIỂU

Bảng 3-1 Bảng nhãn và ví dụ 39
Bảng 3-2 Thống kê tần suất các nhãn trong kho ngữ liệu 43
Bảng 3-3 Thống kê câu hỏi theo lượng nhãn 44
Bảng 3-4 Kết quả thực nghiệm phân loại binary của 3 mô hình 48
Bảng 3-5 Kết quả thực nghiệm phân loại binary sử dụng mô hình SVM 48
Bảng 3-6 Kết quả thực nghiệm phân loại binary sử dụng mô hình BERT 50
Bảng 3-7 Kết quả thực nghiệm phân loại binary sử dụng mô hình PHOBERT 52
Bảng 3-8 Bảng kết quả thực nghiệm phân loại đa nhãn của 3 mô hình 54
Bảng 3-9 Bảng kết quả thực nghiệm các nhãn phân loại đa nhãn sử dụng mô hình SVM.....56
Bảng 3-10 Bảng kết quả thực nghiệm các nhãn phân loại đa nhãn sử dụng mô hình BERT 57

MỞ ĐẦU

Ngày nay công nghệ thông tin phát triển mạnh mẽ, hầu như đã xâm nhập toàn bộ các lĩnh vực đời sống xã hội. Xã hội ngày càng phát triển thì nhu cầu áp dụng các tiến bộ của công nghệ thông tin vào cuộc sống ngày càng cao để giải quyết những vấn đề phức tạp như y tế, giáo dục, pháp luật. Với nhu cầu trao đổi và tìm kiếm thông tin của con người ngày càng cao, thông tin tràn ngập trên mọi phương tiện truyền thông, đặc biệt là sự phát triển rộng rãi của mạng Internet, hằng ngày con người phải xử lý một lượng thông tin khổng lồ. Những hỏi đáp của người dùng dưới dạng truy vấn sẽ được tìm kiếm và trả về một cách ngắn gọn, súc tích, chính xác nhất những gì mà họ mong muốn. 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. Hệ thống hỏi-đáp tự động liên quan đến 3 lĩnh vực lớn là xử lý ngôn ngữ tự nhiên (Natural Language Processing), tìm kiếm thông tin (Information Retrieval) và rút trích thông tin (Information Extraction).

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). Vì vậy phân loại câu hỏi là một bước quan trọng trong hệ thống hỏi đáp, ảnh hưởng trực tiếp đến hoạt động của toàn bộ hệ thống. Nếu phân loại câu hỏi không tốt thì sẽ không thể tìm ra được câu trả lời.

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. Văn bản pháp quy có vai trò quan trọng trong cuộc sống. 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 như mạng nơ-ron nhân chập (Convolutional Neural Network – CNN), mạng nơ-ron hồi quy (Recurrent Neural Network –RNN) cùng các biến thể của nó như LSTM (Long-Short Term Memory) và mô hình BERT.

Mạng nơ-ron có ưu điểm là có thể tự động trích chọn được những thông tin cần thiết và học được ngữ nghĩa từ dữ liệu. Thông thường các mô hình truyền thống phải trích chọn đặc trưng một cách thủ công, mạng nơ-ron sẽ thực hiện tự động chọn ra các đặc trưng cần thiết. Điều này giúp việc chọn được các đặc trưng tốt hơn và đưa ra được kết quả tốt hơn.

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. Mô hình BERT đã thành công trong việc cải thiện những công việc gần đây 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.

Với mục đích đưa những tiến bộ công nghệ vào phục vụ cho cuộc sống, chúng tôi xin chọn đề tài nghiên cứu "*Phân loại câu hỏi pháp quy tiếng Việt sử dụng mô hình BERT*". Phân loại câu hỏi pháp quy tiếng Việt là bài toán phân loại câu hỏi về pháp luật thành các lĩnh vực pháp lý, được phân vào một số loại ví dụ như: Công chứng, Dân sự, Hôn nhan và gia đình, Quốc tịch Việt Nam, v.v.

Luận văn 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

Trong chương 1, luận văn trình bày cái nhìn tổng quan về bài toán phân loại câu hỏi, bao gồm giới thiệu cơ bản về bài toán phân loại câu hỏi; đặc điểm của dữ liệu câu hỏi pháp quy; các nghiên cứu liên quan về phân loại câu hỏi; các phương pháp 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.

Trước khi tìm ra được câu trả lời cho câu hỏi, hệ thống cần phải xác định được câu hỏi đó thuộc loại nào, hỏi về cái gì.

Ví dụ:

- Câu hỏi "*Ai là chủ tích nước Việt Nam năm 2010*" là câu hỏi về "*người*" hay câu "*Việt Nam có bao nhiêu người mặc bệnh covid-19*" là câu hỏi về số lượng.
- Câu hỏi "*Quy định của pháp luật về nghĩa vụ tài sản của vợ chồng đối với người thứ ba khi ly hôn?*" là câu hỏi về "*hôn nhân và gia đình*".

Xác định được loại câu hỏi không chỉ có thể thu gọn phạm vi được không gian tìm kiếm cần tìm câu trả lời, nó còn có thể tìm kiếm chính xác câu trả lời trong một tập lớn các ứng viên trả lời. Như ở ví dụ trên, hệ thống trả lời có thể chỉ quan tâm đến các ứng viên là tên các thực thể là "*người*" hay "*số lượng*" mà không cần phải kiểm tra toàn bộ các đoạn văn bản để tìm ở đâu có thể chứa câu trả lời hoặc không. Vì vậy, phân loại câu hỏi đóng vai trò quan trọng trong hệ thống trả lời tự động.

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. Nó ánh xạ một câu hỏi vào một chủ đề đã biết trong một tập hữu hạn các chủ đề dựa trên

các đặc trưng của câu hỏi. Phân loại câu hỏi[1] nhận đầu vào là câu hỏi dưới dạng ngôn ngữ tự nhiên của người dùng, đưa ra nhãn phân loại cho câu hỏi đó, xem câu hỏi đó thuộc loại nào. 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?

Ví dụ: đối với tiếng Việt: "Cơ sở Học viện Công nghệ Bưu Chính Viễn Thông ở Hà Nội nằm ở đâu?" và tập nhãn đã có.

- **Output:**
  - Nhãn c_i của câu hỏi q_j.

Ví dụ: Địa điểm

Có rất nhiều kỹ thuật học máy và khái phá dữ liệu đã được áp dụng vào bài toán phân loại câu hỏi như: cây quyết định (decision tree)[19], Naïve Bayes[20], K-láng giềng gần nhất (KNN)[21], mạng nơron (neural network)(như mạng thần kinh tích chập Convolutional Neural Networks (CNNs)[7], mạng nơ-ron hồi quy Recurrent Neural Network (RNN)[17], v.v), v.v.

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. Văn bản pháp quy có vai trò quan trọng trong cuộc sống. 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.

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.

Trong mô hình, họ đưa ra giả định rằng xác suất của sự kiện xảy ra (sự xuất hiện của từ) độc lập với ngữ cảnh và vị trí của từ trong văn bản mà nó xuất hiện, và do đó cách biểu diễn được chọn là phù hợp. Họ sử dụng phương pháp tiếp cận theo xác suất tích lũy sau bằng cách làm việc với xác suất thực sau để xử lý ngưỡng theo cách chính xác. Một khả năng để tính toán xác suất này theo cách có thể thống kê được số là đưa ra một phép toán tối đa trong quy tắc Bayes và sau đó đưa ra một hàm logarit và một hàm lũy thừa cho phép tính toán các xác suất một cách đáng tin cậy.

Họ thực nghiệm đo hiệu quả các phương pháp theo thức tự tăng dần độ khó của nhiệm vụ. Trước tiên, họ xem xét vấn đề phân loại đơn nhãn đơn giản, tức là chỉ xem xét các mẫu có một nhãn lớp duy nhất. Họ nhận được tỷ lệ lỗi là 8,56% trong trường hợp này. Nếu họ đưa các mẫu không có nhãn vào để ước tính tốt hơn các thông số làm mịn, họ không nhận được bất kỳ sự cải thiện nào về tỷ lệ lỗi. Ngoài tỷ lệ lỗi, trong bài toán phân loại đa nhãn họ cũng tính đến độ đo precision/recall.

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. Việc sử dụng một vec-tơ từ được đào tạo trước cung cấp một số lợi thế. Một từ tương tự được nhóm lại với nhau. 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).

Họ đã thực hiện nghiên cứu của mình bằng 7 bước: Bước 1: Khai thác dữ liệu. Bước 2: Mô tả dữ liệu. Phân tích dữ liệu được thực hiện bằng cách vẽ đồ thị và sử dụng pandas. Bước 3: Tiền xử lý dữ liệu. Bước 4: Word embedding. Họ thực hiện embedding layer như một sự kết hợp hai word embedding được đào tạo trước, GloVe, paragram, cùng với mạng nơ-ron. GloVe[9] là mô hình song tuyến tính với các mục tiêu bình phương nhỏ nhất có trọng số. Đào tạo mô hình này dựa trên quan sát đơn giản tỷ lệ của xác suất xảy ra đồng từ-từ. Paragram[10] là mô hình thành phần. Paragram mã hóa các chuỗi từ tùy ý thành một vectơ như Glove. Bước 5: Thử nghiệm với mô hình học giám sát. Họ thử nghiệm 3 thuật toán học giám sát: Multinomial Naïve Bayes, K-nearest, Logistic Regression. Bước 6: Thiết kế mạng nơ-ron. Họ sử dụng RNN để làm mô hình. RNN là một loại mạng thần kinh trong đó đầu ra từ bước trước được đưa vào làm đầu vào cho bước hiện tại. Mạng lưới nơ-ron của họ bao gồm lớp đầu vào, 5 lớp ẩn và 1 lớp đầu ra. Lớp đầu vào bao gồm 65 nút. Lớp đầu vào này được kết nối với lớp nhúng. Lớp nhúng này được sử dụng để tạo đại diện vec-tơ của các từ. Trọng lượng của lớp nhúng được khởi tạo bằng cách

ghép các phần nhúng của bên thứ ba (GloVe và paragram). Bước 7: Đào tạo mạng nơ-ron. Kết quả tốt nhất sử dụng mô hình RNN của họ là 69,13% với độ đo là F1. Nghiên cứu này cho kết quả thực nghiệm không quá tốt.

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. Vì đối số là các kết quả rất nhanh, thay vì tạo một mạng duy nhất có thể phân loại một ví dụ thành 50 lớp, họ tạo mạng riêng cho mỗi lớp chính và điều này giúp cung cấp cho lớp thứ cấp CNN một số thông tin trước về danh mục chính. Kiến trúc được đề xuất cho mạng nơ-ron tích chập bao gồm một lớp convolutional để học tìm hiểu một số bộ lọc để đạt được chiều cao nhất định.

Trong mạng này, họ lấy từ bigram đến pent-gram. Điều này giúp họ tìm hiểu ý định của câu hỏi ở một mức độ lớn hơn. Tiếp theo, họ đã thêm một lớp gộp k-max (Kalchbrenner et al., 2014)[12]. Họ đã sử dụng nhóm tối đa 2 cho mạng của mình để tích lũy thêm thông tin từ các bộ lọc tích chập. Sau đó, họ hợp nhất tất cả các đầu ra gộp chung này để tạo thành một lớp được kết nối đầy đủ. Các CNN có xu hướng hoạt động tốt hơn khi các lớp được kết nối đầy đủ hơn được thêm vào cuối trước khi lớp softmax đầu ra [13, 14]. Do đó, họ thêm hai lớp với các nút ẩn N và N / 2 với các tiếp tuyến hyperbol là các hàm kích hoạt của chúng. Dropout 0,5 đã được sử dụng trong hai lớp này để tránh quá mức trong khi đào tạo. Họ đang sử dụng hai tầng CNN để phân loại các câu hỏi ở các cấp độ khác nhau - chính và phụ. Các câu hỏi được phân loại thành các loại chính của chúng theo CNN cấp 1 được chuyển đến CNN intier 2 thích hợp để xác định danh mục phụ của chúng. Nghiên cứu của họ cho kết quả tốt nhất với độ đo Accuracy là 90.43% với câu hỏi chính và 76,52% với câu hỏi phụ. Nhận thấy rằng kết quả nghiên cứu của họ khá tốt.

Hiện nay có ít nghiên cứu về phân loại câu hỏi pháp quy tiếng Việt.

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. Các luật này có được là do nghiên cứu và đề xuất từ các chuyên gia. Đối với cách tiếp cận này, một loạt các biểu thức chính quy (regular expression) được tạo ra để so khớp với câu hỏi từ đó quyết định phân loại của câu hỏi và loại câu trả lời.

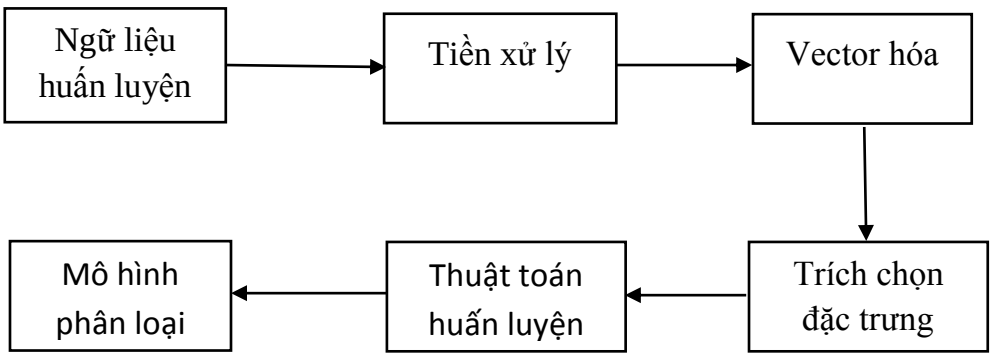
Tiếp cận dựa trên học máy[3] là cách tiếp cận được sử dụng phổ biến rộng rãi để giải quyết bài toán phân loại câu hỏi. Cách tiếp cận này sẽ thay thế các kiến thức chuyên môn bằng một tập lớn các câu hỏi được gán nhãn (tập dữ liệu mẫu). Sử dụng tập này, một bộ phân lớp sẽ được huấn luyện có giám sát.

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. Cách tiếp cận bằng học máy đã giải quyết được các hạn chế trong cách tiếp cận dựa trên luật.

1.4.1 Phương pháp học máy truyền thống

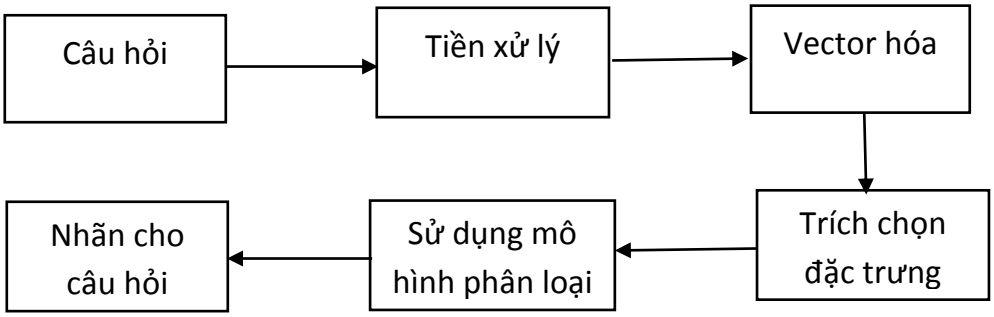
Với các phương pháp học máy truyền thống như SVM, KNN, cây quyết định, v.v thì quá trình phân loại dữ liệu văn bản (document, câu) thường gồm hai giai đoạn sau:

- Giai đoạn huẩn luyện: Giai đoạn huấn luyện nhận đầu vào là tập ngữ liệu huấn luyện gồm các câu hỏi đã được gán nhãn, sau khi xử lý tập ngữ liệu và áp dụng các thuật toán huấn luyện sẽ cho ra đầu ra là một mô hình phân loại.



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

- Giai đoạn phân lớp: Giai đoạn phân lớp nhận đầu vào là câu hỏi của người dùng dưới dạng ngôn ngữ tự nhiên, sau quá trình tiền xử lý và áp dụng mô hình phân loại sẽ cho ra nhãn phân loại của câu hỏi đầu vào.



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.

Giải thuật SVM thuộc nhóm giải thuật học máy có giám sát và được sử dụng trong các bài toán phân lớp và hồi quy, chủ yếu là bài toán phân lớp. SVM là một thuật toán phân loại nhị phân nhận dữ liệu đầu vào và phân loại chúng thành hai loại khác nhau. Với một bộ các dữ liệu huấn luyện thuộc hai loại cho trước, thuật toán huấn luyện SVM xây dựng một mô hình SVM để phân loại các dữ liệu khác vào hai thể loại đó.

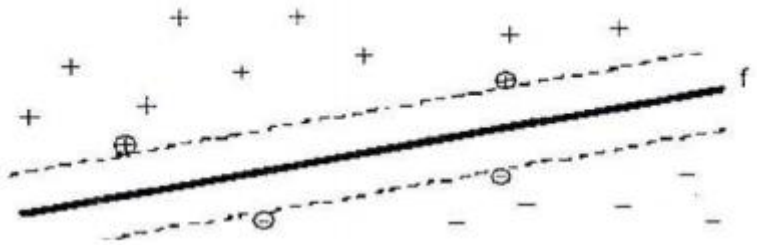
Phương pháp này thực hiện phân lớp dựa trên nguyên lý Cực tiểu hóa rủi ro có cấu trúc SRM (Structural Risk Minimization) [5], được xem là một trong các phương pháp phân lớp giám sát không tham số tinh vi. Các hàm công cụ của SVM cho phép tạo không gian chuyển đổi để xây dựng mặt phẳng phân lớp để phân chia các lớp ra thành các phần riêng biệt.

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.

Trong nhiều trường hợp, không thể phân chia các lớp dữ liệu một cách tuyến tính trong một không gian ban đầu được dùng để mô tả một vấn đề. Vì vậy, nhiều khi cần phải ánh xạ các điểm dữ liệu trong không gian ban đầu vào một không gian mới nhiều chiều hơn, để việc phân tách chúng trở nên dễ dàng hơn trong không gian mới.

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



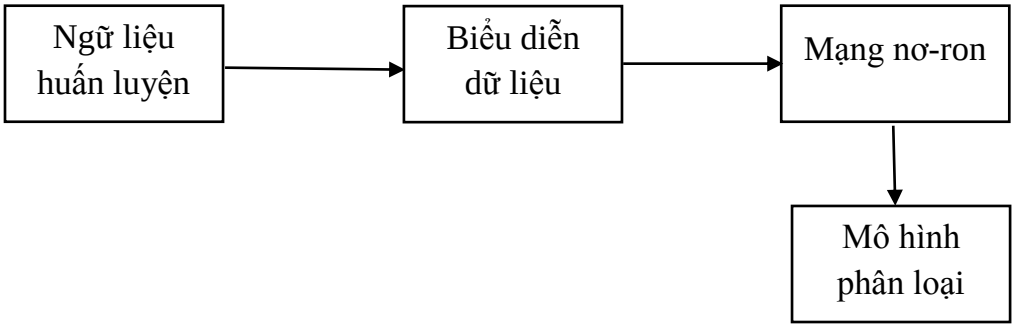
Hình 1-3 Siêu phẳng phân chia dữ liệu học thành 2 lớp (+) và (-) với khoảng cách biên là lớn nhất. Các biên gần nhất (điểm được khoanh tròn) là các Support Vector[5]

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

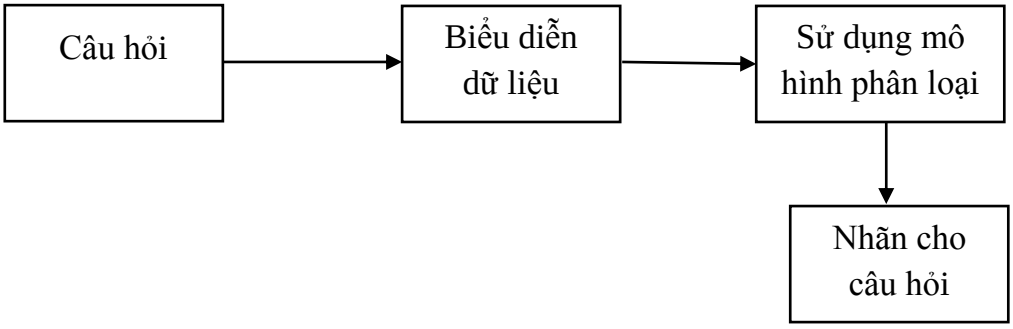
Với phương pháp sử dụng mạng nơ-ron như LSTM, CNN, RNN, v.v thì quá trình phân loại dữ liệu văn bản cũng gồm hai giai đoạn:

- Giai đoạn huấn luyện: Giai đoạn huấn luyện nhận đầu vào là tập ngữ liệu huấn luyện gồm các câu hỏi đã được gán nhãn, sau khi biểu diễn dữ liệu và đưa vào mạng nơ-ron sẽ cho ra đầu ra là một mô hình phân loại.



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

- Giai đoạn phân lớp: Giai đoạn phân lớp nhận đầu vào là câu hỏi của người dùng dưới dạng ngôn ngữ tự nhiên, sau quá trình biểu diễn dữ liệu và áp dụng mô hình phân loại sẽ cho ra nhãn phân loại của câu hỏi đầu vào.



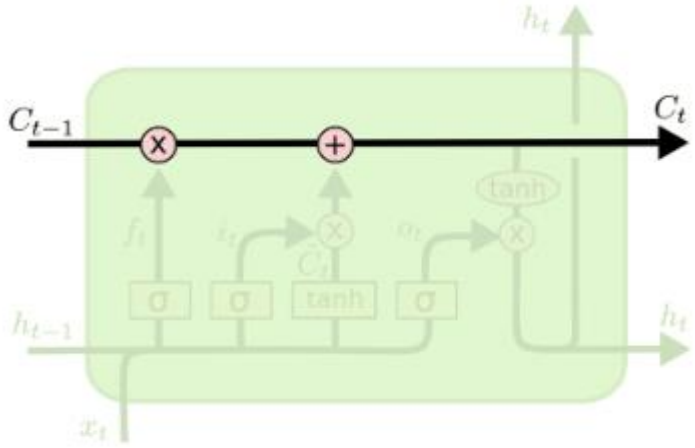
Hình 1-5 Mô hình giai đoạn phân lớp sử dụng mạng nơ-ron.

❖ **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. LSTM được giới thiệu bởi Hochreiter & Schmidhuber (1997), và sau đó đã được cải tiến và phổ biến rộng rãi. Mô hình này tương thích với nhiều bài toán, vì vậy nó được sử dụng rộng rãi trong nhiều ngành liên quan.

LSTM được thiết kế để giải quyết được vấn đề phụ thuộc xa (long-term dependency). Việc ghi nhớ thông tin trong suốt thời gian dài là đặc tính mặc định của chúng, chứ ta không cần phải huấn luyện nó để có thể nhớ được. Tức là ngay nội tại của nó đã có thể ghi nhớ được mà không cần bất kì can thiệp nào.

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.

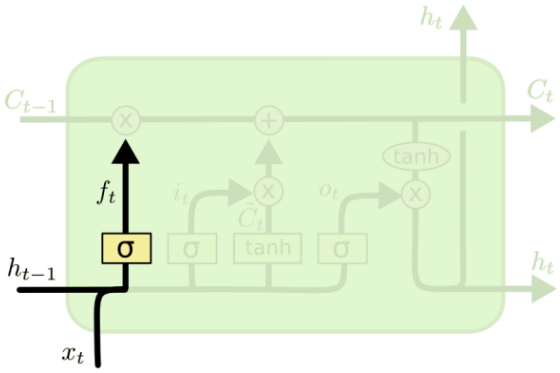


Hình 1-6 Tế bào trang thái LSTM giống như một băng chuyền [22]

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.

Cổng là một cách để cho thông tin đi qua. Một LSTM có 3 cổng để bảo vệ và điều khiển trạng thái tế bào. Mỗi cổng gồm một lớp mạng sigmoid và một toán tử nhân. Sigmoid có đầu ra là 0 và 1, thể hiện bao nhiêu thông tin sẽ được đưa qua cổng.

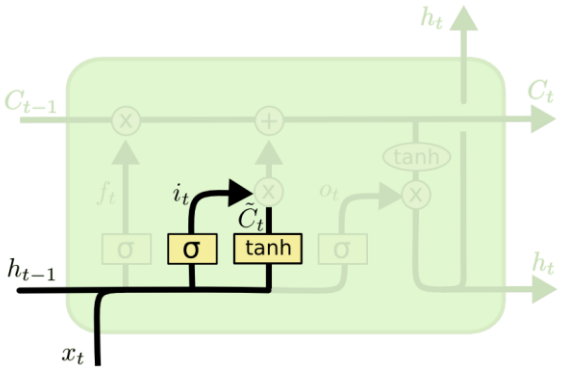
Bước đầu tiên trong mô hình LSTM là việc quyết định thông tin nào sẽ được đưa đến trạng thái tế bào thông qua cổng. Quá trình này được thực hiện thông qua một lớp sigmoid gọi là "lớp cổng chặn" - cổng chặn với hai đầu vào là h_{t-1} và x_t, và cho đầu ra là một giá trị trong phạm vi [0, 1] cho mỗi đầu vào trạng thái ô C_{t-1}. 1 tương đương với "lưu giữ thông tin", 0 tương đương với "xóa thông tin".



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

Hình 1-7 LSTM focus f [22]

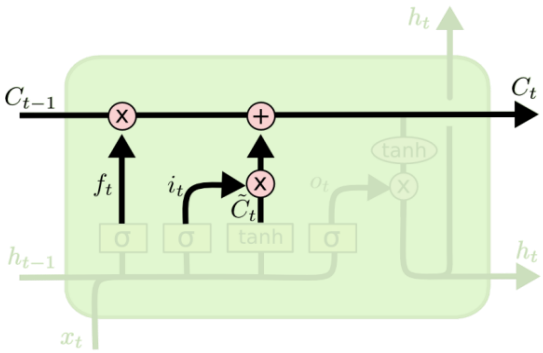
Bước tiếp theo là xác định loại thông tin mới nào cần được lưu lại trong cell state. Ta có hai phần. Một là single sigmoid layer được gọi là "input gate layer" quyết định các giá trị nào cần được cập nhật. Tiếp theo, một $tanh$ layer tạo ra một vector với giá trị mới có thể đưa vào cell state, C_t được thêm vào trong ô trạng thái.



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

Hình 1-8 LSTM focus I [22]

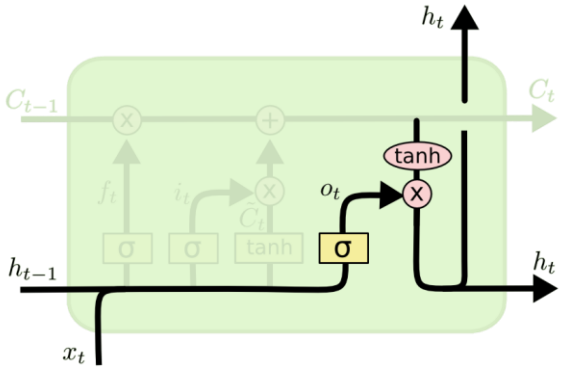
Ở bước tiếp theo, kết hợp hai thành phần này lại để cập nhật vào cell state. Lúc cập nhật vào cell state cũ C_{t-1} vào cell state mới C_t. Tại bước này thực hiện nhân trạng thái cũ với f_t, để cần nhớ hoặc quên đi những gì trước đó hay không. Sau đó, bổ sung $i_t * \tilde{C}_t$. Đây là giá trị ứng viên mới, co giãn (scale) số lượng giá trị mà ta muốn cập nhật cho mỗi state.



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Hình 1-9 LSTM focus c [22].

Cuối cùng, cần quyết định xem thông tin output là gì. Output này cần dựa trên trạng thái của cell state, nhưng sẽ là giá trị được lọc bớt một số thông tin. Đầu tiên, chạy qua một single sigmoid layer để quyết định xem phần tử nào của cell state sẽ tác động đến output. Sau đó, ta sẽ đẩy cell state đi qua một function tanh giá trị khoảng [-1, 1] và nhân với một output sigmoid gate, để giữ lại những phần ta muốn output ra ngoài.



$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

Hình 1-10 LSTM focus o [22]

❖ **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.

Các kỹ thuật quen thuộc phổ biến như Word2vec, FastText hay Glove cũng tìm ra đại diện của từ thông qua ngữ cảnh chung của chúng. Tuy nhiên, những ngữ cảnh của các kỹ thuật này là đa dạng trong dữ liệu tự nhiên. Ví dụ các từ như "con chuột" có ngữ nghĩa khác nhau ở các ngữ cảnh khác nhau như "Con chuột máy tính này thật đẹp!" và "con chuột này to thật." Trong khi các mô hình như Word2vec, fastText tìm ra 1 vector đại diện cho mỗi từ dựa trên 1 tập ngữ liệu lớn nên không thể hiện được sự đa dạng của ngữ cảnh. Việc biểu diễn mỗi từ dựa trên các từ khác trong câu thành một đại diện sẽ mang lại kết quả ý nghĩa rất nhiều. Như trong ví dụ trên, ý nghĩa của từ "con chuột" sẽ được biểu diễn cụ thể dựa vào các từ phía trước hoặc sau nó trong câu. Nếu đại diện của từ "con chuột" được xây dựng dựa trên những ngữ cảnh cụ thể này thì sẽ có được biểu diễn tốt hơn.

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

Trong chương này, luận văn giới thiệu bài 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.

2.1 Bài toán phân loại đa nhãn câu hỏi tiếng Việt

Phân loại đa nhãn (Multi label classification) đã thu hút nhiều sự chú ý, nhờ tính hữu dụng và tính toàn diện của nó trong các ứng dụng thế giới thực, trong đó các đối tượng có thể được đặc trưng bởi nhiều hơn một nhãn như trong cách tiếp cận truyền thống.

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. Phân loại đa nhãn ngày càng nhận được sự chú ý và đã xuất hiện phổ biến trong nhiều lĩnh vực ứng dụng như phân loại web, đề xuất thẻ, dự đoán chức năng gen, chẩn đoán y tế và lập chỉ mục video (Elisseeff & Weston, 2001; Rousu, Saunders, Szedmák, & Shawe-Taylor, 2006 ; Silla & Freitas, 2011; Trohidis, Tsoumakas, Kalliris, & Vlahavas, 2008; Tsoumakas, Katakis, & Vlahavas, 2010; Zhang & Zhou, 2007).

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?

Ví dụ: Câu hỏi "*Hồ sơ đăng ký thay đổi tên của bên nhận thế chấp?*" và tập nhãn {"Công chứng", "Dân sự", "Tổ chức chính phủ", "Bảo hiểm", "Cư trú", "Nuôi con nuôi", "Thi hành án", "Quản lý, sử dụng", "Hôn nhân và gia đình", "Quốc tịch Việt Nam", "Đầu tư", "Ban hành văn bản quy phạm pháp luật", "Bảo vệ môi trường", "Xây dựng", "Tổ chức cơ quan, chính quyền", "Tố tụng", "Công dân", "Quốc phòng", "Hình sự", "Giao thông đường bộ", "Thuế", "Đất đai", "Đấu giá tài sản", "Phòng, chống ma túy", "Cán bộ, công chức, viên chức", "Khiếu nại, tố cáo", "Kinh tế", "Xử lý vi phạm hành chính", "Phí và lệ phí", "Lao động", "Nhà ở", "Lý lịch tư pháp", "Trách nhiệm bồi thường của Nhà nước", "Giám định tư pháp"}

- **Output:**
  - Tập nhãn $\{c_i\}$ của câu hỏi Q.

Ví dụ: Câu hỏi ở input phía trên có nhãn là: {Dân sự, Đất đai}.

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

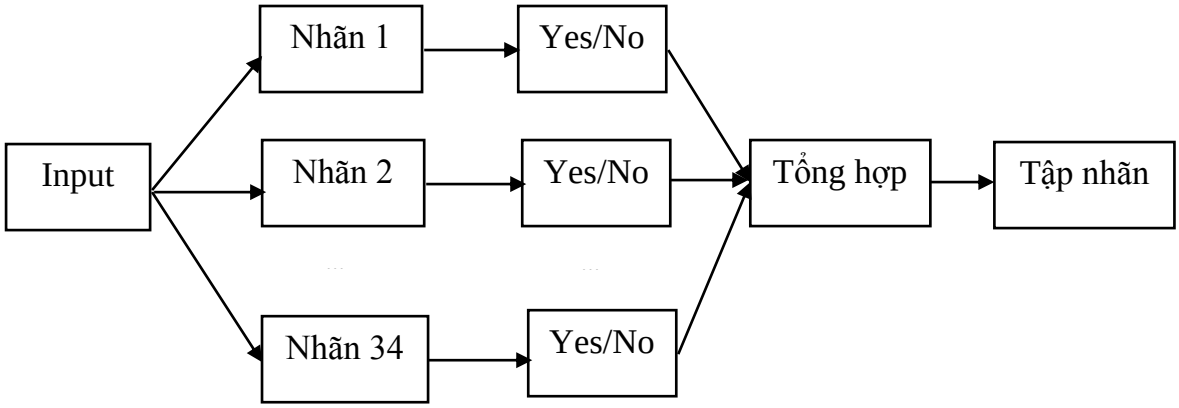
Luận văn mô hình hóa bài toán phân loại đa nhãn dưới dạng bài toán phân lớp. Đầu vào là câu hỏi, đầu ra là các nhãn thuộc vào tập nhãn đã có.

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

Ví dụ:

- Câu hỏi: "*Trả lại xe ô tô vi phạm giao thông gây chết người cho chủ sở hữu khi nào?*"
- Tổng hợp phân loại nhị phân của câu hỏi sau khi chạy qua 34 bộ phân loại như sau:

Nhãn	Trả lời
Công chứng	0
Dân sự	0
Tổ chức chính phủ	0
Bảo hiểm	0
Cư trú	0

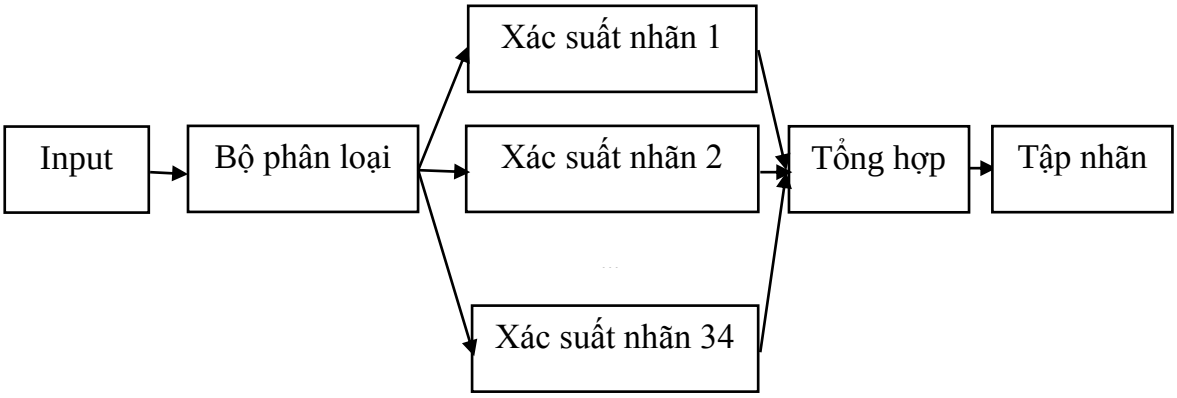
Nuôi con nuôi	0
Thi hành án	0
Quản lý, sử dụng	0
Hôn nhân và gia đình	0
Quốc tịch Việt Nam	0
Đầu tư	0
Ban hành văn bản quy phạm pháp luật	0
Bảo vệ môi trường	0
Xây dựng	0
Tổ chức cơ quan, chính quyền	0
Tố tụng	1
Công dân	0
Quốc phòng	0
Hình sự	1
Giao thông đường bộ	0
Thuế	0
Đất đai	0
Đấu giá tài sản	0
Phòng, chống ma túy	0

Cán bộ, công chức, viên chức	0
Khiếu nại, tố cáo	0
Kinh tế	0
Xử lý vi phạm hành chính	0
Phí và lệ phí	0
Lao động	0
Nhà ở	0
Lý lịch tư pháp	0
Trách nhiệm bồi thường của Nhà nước	0
Giám định tư pháp	0

➔ Câu hỏi có nhãn là {Tố tụng, Hình sự}.

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

Ví dụ:

- Câu hỏi: "*Trả lại xe ô tô vi phạm giao thông gây chết người cho chủ sở hữu khi nào?*"
- Xác suất các nhãn của câu hỏi sau khi chạy qua bộ phân lớp 34 nhãn như sau:

Nhãn	Xác suất
Công chứng	0,01
Dân sự	0,02
Tổ chức chính phủ	0,01
Bảo hiểm	0,01
Cư trú	0,01
Nuôi con nuôi	0,01
Thi hành án	0,01
Quản lý, sử dụng	0,01
Hôn nhân và gia đình	0

Quốc tịch Việt Nam	0
Đầu tư	0,01
Ban hành văn bản quy phạm pháp luật	0,01
Bảo vệ môi trường	0
Xây dựng	0
Tổ chức cơ quan, chính quyền	0
Tố tụng	0,51
Công dân	0,01
Quốc phòng	0
Hình sự	0,96
Giao thông đường bộ	0,01
Thuế	0,01
Đất đai	0,01
Đấu giá tài sản	0
Phòng, chống ma túy	0
Cán bộ, công chức, viên chức	0
Khiếu nại, tố cáo	0
Kinh tế	0,01
Xử lý vi phạm hành chính	0,01

Phí và lệ phí	0
Lao động	0
Nhà ở	0,01
Lý lịch tư pháp	0
Trách nhiệm bồi thường của Nhà nước	0
Giám định tư pháp	0

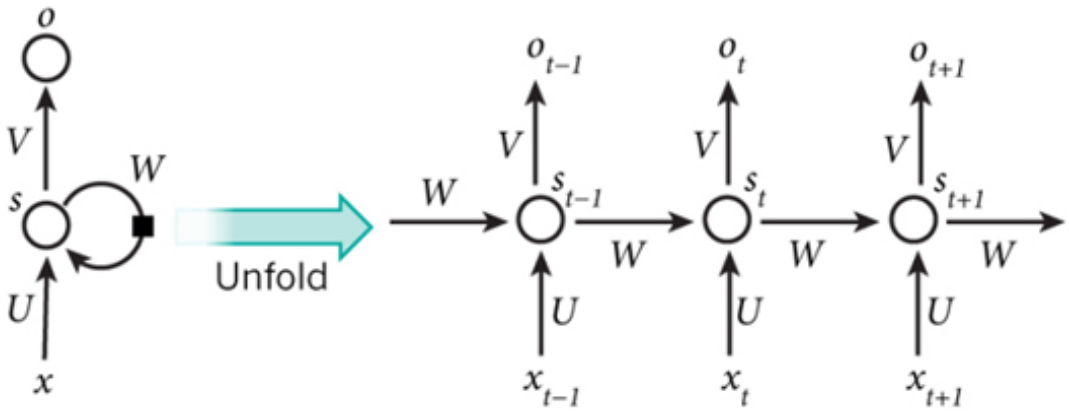
➔ Nhãn của câu hỏi là {Tố tụng, Hình sự}.

Vấn đề còn lại bây giờ là quy về bài toán xây dựng các bộ phân lớp. Có nhiều mô hình để xây dựng các bộ phân lớp, một số mô hình sẽ được trình bày ở mục tiếp theo.

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. Ví dụ như một đoạn text có thể được coi là một chuỗi các từ vựng (words) hoặc là một chuỗi các ký tự (character). Tại thời điểm t, với dữ liệu đầu vào xt ta có kết quả đầu ra là y_t. Tuy nhiên, khác với mạng Feed Forward Network, y_t lại được sử dụng là đầu vào để tính kết quả đầu ra cho thời điểm (t+1). Điều này cho phép RNN có thể lưu trữ và truyền thông tin đến thời điểm tiếp theo. 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[24]

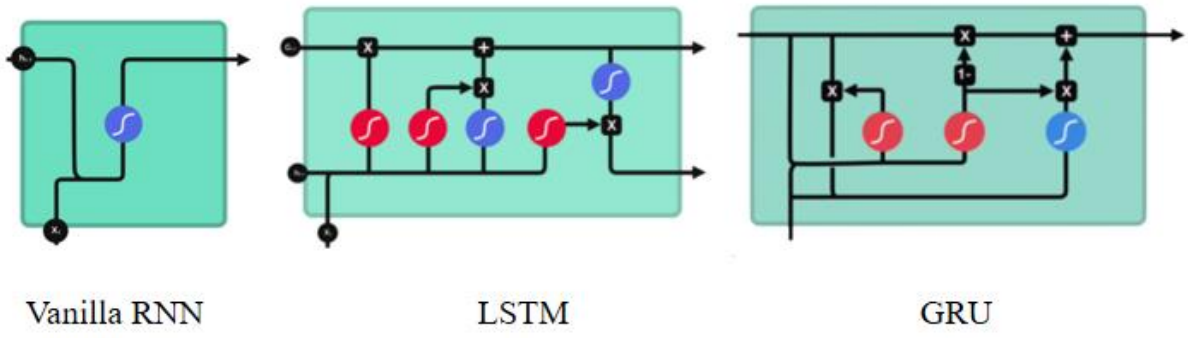
RNN chỉ sử dụng một mạng nơ-ron duy nhất để tính giá trị đầu ra của từng timestep. Do đó các đầu ra khi trở thành đầu vào sẽ được nhân với cùng một ma trận trọng số. Đây cũng chính là lý do tại sao có từ Recurrent trong tên của RNN. Recurrent có nghĩa là mô hình sẽ thực hiện các phép tính toán giống hệt nhau cho từng phần tử của chuỗi dữ liệu đầu vào và kết quả đầu ra sẽ phụ thuộc vào kết quả của các tính toán ở phần trước.

Như vậy, có thể hiểu một cách đơn giản rằng RNN là một mô hình mạng nơ-ron có "bộ nhớ" để lưu trữ thông tin của phần xử lý trước đó. Về mặt lý thuyết thì RNN có thể xử lý và lưu trữ thông tin của một chuỗi dữ liệu với độ dài bất kỳ. Tuy nhiên trong thực tế thì 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). Khi gradient trở nên rất nhỏ (có giá trị gần bằng 0) thì giá trị của ma trận trọng số sẽ không được cập nhật thêm và do đó mạng Neuron sẽ dừng việc học tại lớp này. Đây cũng chính là lý do khiến cho RNN không thể lưu trữ thông tin của các timesteps đầu tiên trong một chuỗi dữ liệu có độ dài lớn.

Quan sát về nhược điểm của RNN, nhận thấy kiến trúc này không hề có cơ chế lọc những thông tin không cần thiết. Bộ nhớ của kiến trúc có hạn, nếu lưu tất cả những chi tiết không cần thiết thì sẽ dẫn đến quá tải, từ đó quên những thứ ở xa

trong quá khứ. Từ suy nghĩ đó, người ta phát triển các kiến trúc để khắc phục các nhược điểm của RNN là LSTM (Long Short-Term Memory) và GRU (Gated Recurrent Units) với việc sử dụng cơ chế "cổng" nhằm bổ sung thông tin mới và loại bỏ thông tin không cần thiết từ "bộ nhớ", từ đó giúp tăng khả năng lưu trữ thông tin quan trọng của RNN.

LSTM và GRU đều có nguyên tắc hoạt động giống như Vanilla RNN, tuy nhiên điểm khác nhau cơ bản giữa chúng là về cấu trúc của các Cell. Cấu trúc này được mô tả như hình dưới đây:



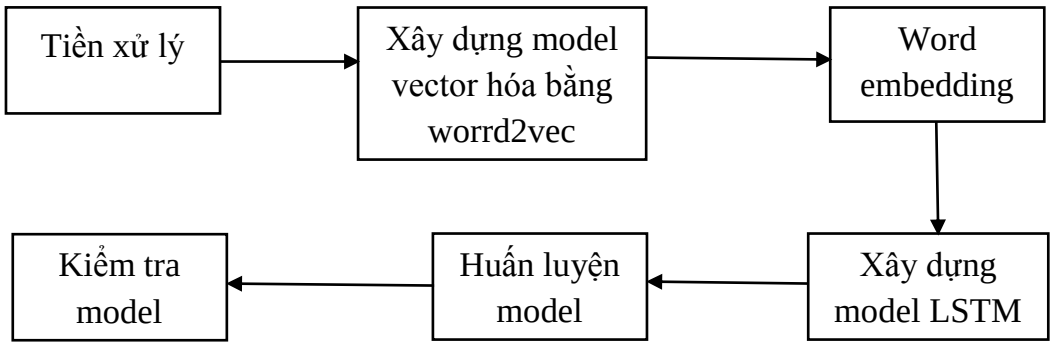
Hình 2-4 Vanilla RNN, LSTM, GRU[24]

Trong Vanilla RNN chỉ sử dụng tanh function với dữ liệu đầu vào là Current input (x_t) và thông tin lưu trữ từ timestep trước (Hidden state h_{t-1}). Tuy nhiên trong LSTM và GRU, sử dụng kết hợp tanh và sigmoid function cùng với các thuật toán để quyết định thông tin nào nên được lưu trữ và thông tin nào nên được loại bỏ.

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 2-5 Các bước của bài toán phân loại văn bản sử dụng mạng nơ-ron RNN.

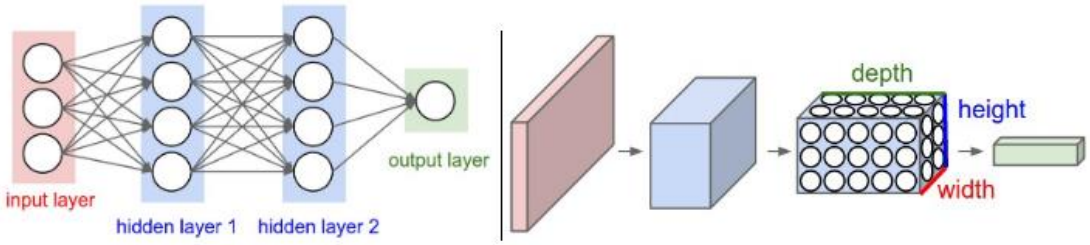
- Tiền xử lý kho ngữ liệu.
- Xây dựng model vector hóa bằng Word2vec cho tập ngữ liệu văn bản đã được tiền xử lý. Mô hình Word2vec bản chất là việc huấn luyện một mạng nơ-ron nhân tạo - Artificial Neural Network (ANN) với một lớp ẩn. Các cặp từ được tách theo skip-gram và dựa trên xác suất để tính độ tương quan giữa các từ.
- Word embedding sử dụng mô hình kết quả của Word2vec để vector từng câu trong tập ngữ liệu.
- Áp dụng mạng nơ-ron RNN để giải quyết bài toán bao gồm các bước nhờ: Xây dựng model RNN, huấn luyện model RNN, kiểm tra model 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. Mỗi một lớp sau khi thông qua các hàm kích hoạt sẽ tạo ra các thông tin trừu tượng hơn cho các lớp tiếp theo.

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 2-6 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)[28]

Mỗi lớp của CNN chuyển đổi khối lượng đầu vào 3D thành khối lượng đầu ra 3D của các kích hoạt nơ-ron. Trong hình vẽ trên, lớp đầu vào màu đỏ giữ hình ảnh, vì vậy chiều rộng và chiều cao của nó sẽ là kích thước của hình ảnh và chiều sâu sẽ là 3 (Đỏ, Xanh lục, Xanh lam).

Mỗi một lớp sau khi thông qua các hàm kích hoạt sẽ tạo ra các thông tin trừu tượng hơn cho các lớp tiếp theo. Trong mô hình mạng truyền ngược (feedforward neural network) thì mỗi neural đầu vào (input node) cho mỗi neural đầu ra trong các lớp tiếp theo.

Mô hình này gọi là mạng kết nối đầy đủ (fully connected layer) hay mạng toàn vẹn (affine layer). Còn trong mô hình CNNs thì ngược lại. Các layer liên kết được với nhau thông qua cơ chế convolution.

Layer tiếp theo là kết quả convolution từ layer trước đó, nhờ vậy mà có được các kết nối cục bộ. Như vậy mỗi neuron ở lớp kế tiếp sinh ra từ kết quả của filter áp đặt lên một vùng ảnh cục bộ của neuron trước đó.

Mỗi một lớp được sử dụng các filter khác nhau thông thường có hàng trăm hàng nghìn filter như vậy và kết hợp kết quả của chúng lại. Ngoài ra có một số layer

khác như pooling/subsampling layer dùng để chắt lọc lại các thông tin hữu ích hơn (loại bỏ các thông tin nhiễu).

Trong quá trình huấn luyện mạng (traning) CNN tự động học các giá trị qua các lớp filter dựa vào cách thức mà bạn thực hiện.

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.

Pooling layer sẽ cho bạn tính bất biến đối với phép dịch chuyển (translation), phép quay (rotation) và phép co giãn (scaling). Tính kết hợp cục bộ cho ta các cấp độ biểu diễn thông tin từ mức độ thấp đến mức độ cao và trừu tượng hơn thông qua convolution từ các filter.

Thông thường, định kỳ chèn một lớp Pooling vào giữa các lớp chuyển đổi liên tiếp trong kiến trúc CNN. Chức năng của nó là giảm dần kích thước không gian của biểu diễn để giảm lượng tham số và tính toán trong mạng, và do đó cũng kiểm soát việc trang bị quá mức. Lớp Pooling hoạt động độc lập trên mọi lát cắt sâu của đầu vào và thay đổi kích thước của nó theo không gian, sử dụng phép toán MAX.

Lớp fully connected được kết nối đầy đủ có kết nối đầy đủ với tất cả các hoạt động trong lớp trước đó, như được thấy trong các Mạng nơ-ron thông thường. Do đó, kích hoạt của chúng có thể được tính bằng phép nhân ma trận theo sau là phần bù lệch.

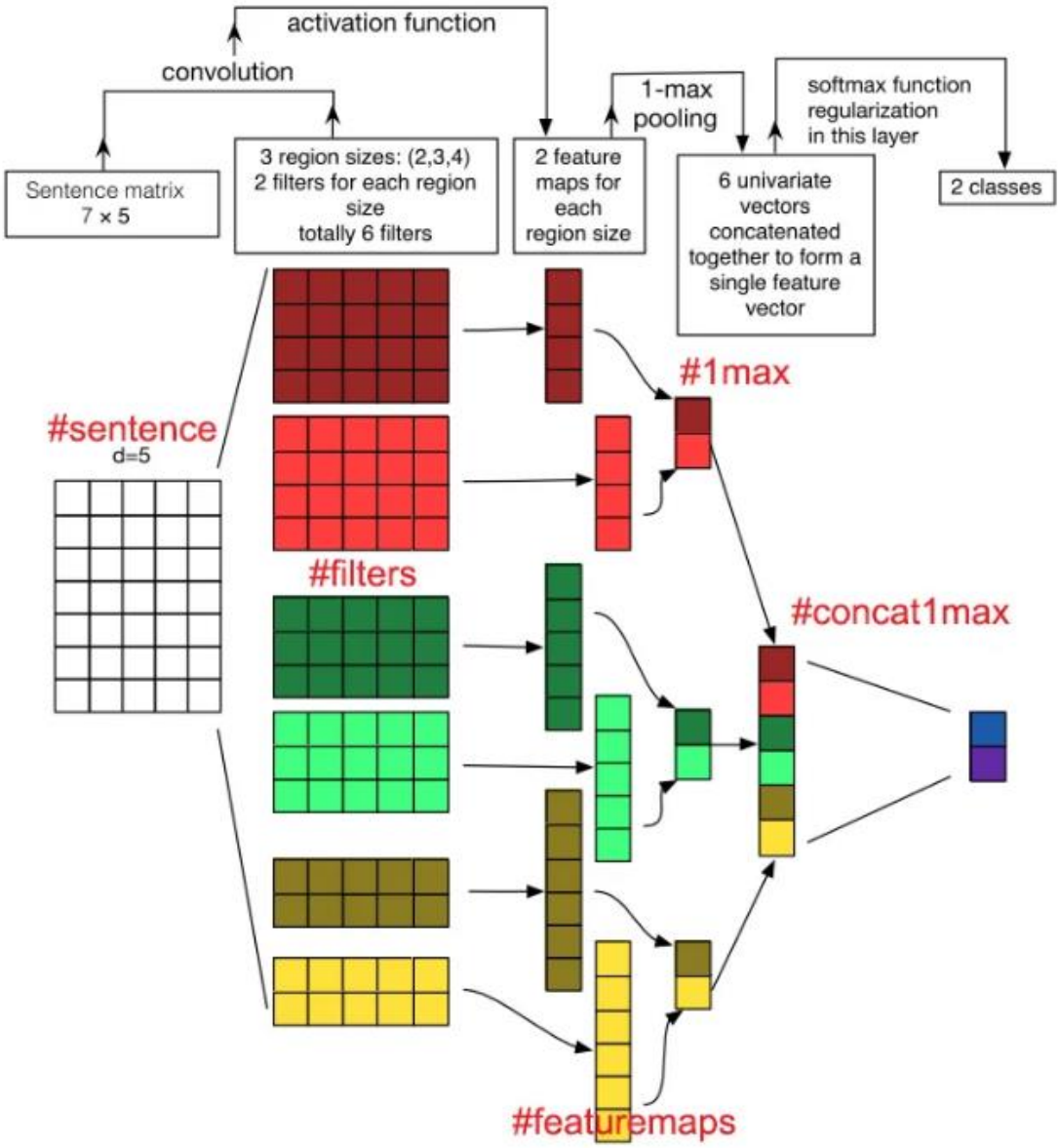
Đó là lý do tại sao CNNs cho ra mô hình với độ chính xác rất cao. Cũng giống như cách con người nhận biết các vật thể trong tự nhiên.

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 2-7 Kiến trúc mô hình CNN dùng trong phân loại văn bản[28]

Trong ví dụ mô hình kiến trúc trên, mô hình sử dụng ba bộ lọc với các kích thước khác nhau: 2, 3, 4. Do vậy sẽ có 6 map đặc trưng.

Các bước phân loại văn bản sử dụng mô hình CNN được mô tả như sau:

- Văn bản được tiền xử lý tách từ. Kích thước ma trận của cả câu văn phụ thuộc vào số lượng từ và số chiều của word vector (bằng số lượng từ x số chiều word vector).
- Câu văn sau khi được biểu diễn dưới dạng ma trận sẽ được đưa vào bộ lọc (filters). Một tính chất mong muốn của thuật toán CNN trong phân loại văn bản là giữ được chuỗi từ liên tiếp. Do vậy, cần cố định số chiều của bộ lọc cho phù hợp với số chiều của từ. Filters ở đây không phải là bộ lọc để lọc bỏ các phần tử khỏi ma trận bị lọc.
- Features map: trong bước này bộ lọc nhân tích chập tạo ra một map lưu trữ các đặc trưng.
- Để đưa ma trận đặc trưng về kích thước như nhau hoặc trong nhiều trường hợp chỉ muốn giữ lại các đặc trưng tiêu biểu có thể sử dụng max-pooling để lấy ra các giá trị lớn nhất trong map đặc trưng. Điều này giúp giảm chiều dữ liệu, tăng tốc độ tính toán.
- #concat1max thực hiện việc kết hợp tất cả các đặc trưng ở bước trước lại. Đưa ra một vector đặc trưng cuối cùng để đưa vào fully-connected. Sau khi áp dụng 1-max pooling, đã có những vector có kích thước cố định. Vector cố định kích thước này sau đó được đưa vào một hàm softmax (lớp fully-connected) để giải quyết việc phân loại.

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

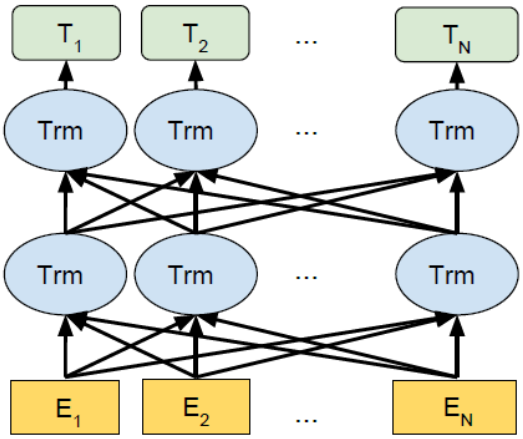
trình xử lý ngôn ngữ tự nhiên. Nói một cách đơn giản, thì nó có thể được sử dụng để giúp Google phân biệt rõ hơn ngữ cảnh của các từ xuất hiện trong truy vấn tìm kiếm.

Ví dụ, trong các cụm từ "nine to five" (từ 9 giờ sáng đến 5 giờ chiều) và "a quarter to five" (5 giờ kém 15 phút) thì từ "to" có hai ý nghĩa khác nhau, sự khác biệt này có thể rõ ràng đối với con người chúng ta nhưng không phải đối với các máy tìm kiếm. BERT được thiết kế để phân biệt những sắc thái ngữ nghĩa như thế, từ đó giúp đưa ra những kết quả phù hợp và có liên quan hơn.

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). BERT cho phép mô hình ngôn ngữ học về ngữ cảnh của từ vựng dựa trên các từ xung quanh nó, thay vì chỉ dựa vào từ ngữ đứng trước hoặc ngay sau nó.

Google gọi BERT là công nghệ "có tính hai chiều rất sâu" bởi vì sự diễn giải ngữ cảnh của các từ bắt đầu từ "tầng đáy thấp nhất trong một mạng lưới neural network gồm rất nhiều tầng".

Kiến trúc mô hình BERT là một bộ mã hóa Transformer hai chiều (bidirectional Transformer encoder). Việc sử dụng Transformer không có gì đáng ngạc nhiên vì đây là một xu hướng gần đây do tính hiệu quả và hiệu suất vượt trội của huấn luyện Transformers trong việc phát hiện các phụ thuộc với khoảng cách xa (long-distance dependencies) so với kiến trúc Recurrent neural network. Trong khi đó, 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 2-8 Kiến trúc của mô hình BERT [26]

Sử dụng bộ mã hóa Transformer được tiền huấn luyện, BERT có thể biểu diễn bất kỳ token nào dựa trên ngữ cảnh hai chiều của nó. Trong quá trình học có giám sát trên các tác vụ xuôi dòng, BERT tương tự như GPT ở hai khía cạnh. Đầu tiên, các biểu diễn BERT sẽ được truyền vào một tầng đầu ra được bổ sung, với những thay đổi tối thiểu tới kiến trúc mô hình tùy thuộc vào bản chất của tác vụ, chẳng hạn như dự đoán cho mỗi token hay dự đoán cho toàn bộ chuỗi. Thứ hai, tất cả các tham số của bộ mã hóa Transformer đã tiền huấn luyện đều được tinh chỉnh, trong khi tầng đầu ra bổ sung sẽ được huấn luyện từ đầu.

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

Tổng quan phương pháp của luận văn bao gồm hai giai đoạn chính. Giai đoạn đầu tiên là huấn luyện trên mô hình huấn luyện trước sử dụng mô hình BERT. Sau đó, luận văn dùng mô hình BERT được huấn luyện ở giai đoạn trước để đưa vào mô hình học có giám sát tạo một model để đào tạo đánh giá và dự đoán nhiệm vụ phân loại đa nhãn.

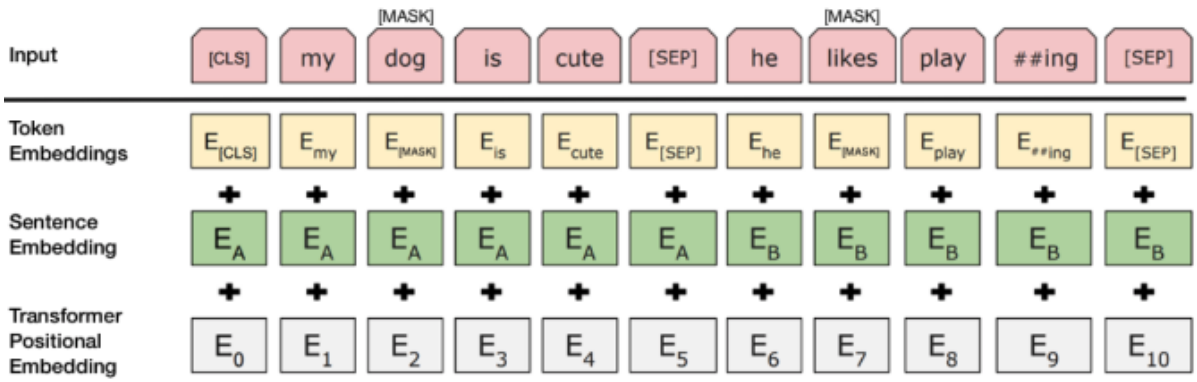
2.5.1 Biểu diễn đầu vào

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>”. Ta sẽ phân biệt nhất quán thuật ngữ “chuỗi đầu vào BERT” với các kiểu “chuỗi” khác. Chẳng hạn, một chuỗi đầu vào BERT có thể bao gồm cả một chuỗi văn bản hoặc hai chuỗi văn bản.

Khi có một chuỗi đầu vào cụ thể, biểu diễn đầu vào được xây dựng bằng cách tính tổng các token đó với vector phân đoạn và vị trí tương ứng của các từ trong chuỗi.

Cho dễ hình dung, biểu diễn đầu vào được trực quan hóa trong hình dưới đây:



Hình 2-9 Mô hình đại diện đầu vào của BERT [26].

Token đầu tiên cho mỗi chuỗi được mặc định là một token đặc biệt có giá trị là [CLS]. Đầu ra của Transformer (hidden state cuối cùng) tương ứng với token này sẽ được sử dụng để đại diện cho cả câu trong các nhiệm vụ phân loại. Nếu không trong các nhiệm vụ phân loại, vector này được bỏ qua.

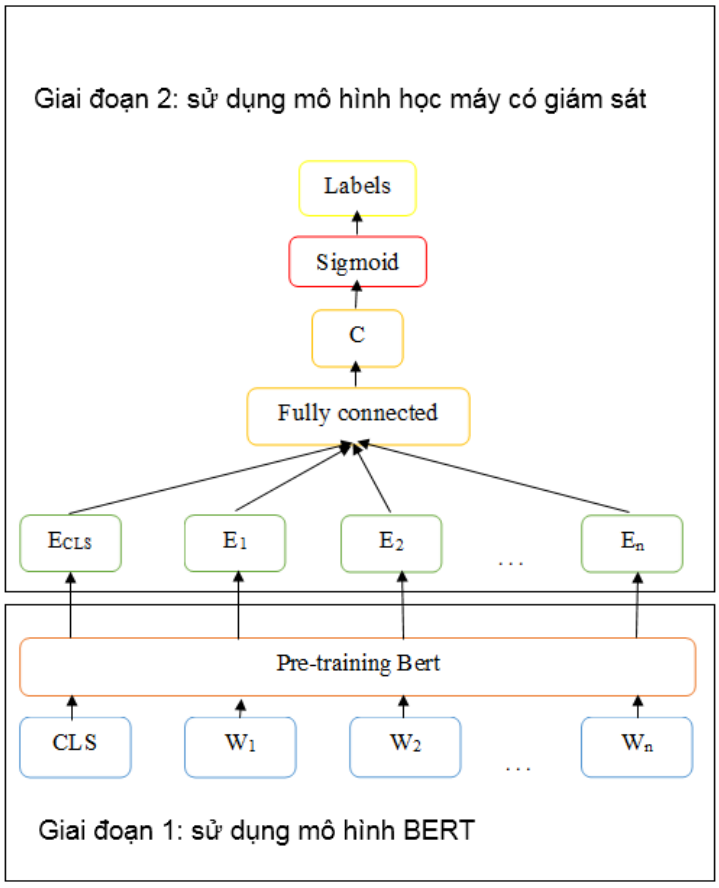
Trong trường hợp các cặp câu được gộp lại với nhau thành một chuỗi duy nhất, chúng ta phân biệt các câu theo 2 cách. Đầu tiên, chúng ta tách chúng bởi một token đặc biệt [SEP]. Thứ hai, chúng ta thêm một segment embedding cho câu A và một segment embedding khác cho câu B như hình vẽ.

Khi chỉ có 1 câu đơn duy nhất, segment embedding chỉ có cho câu A.

Kiến trúc hai chiều của BERT là bộ mã hóa Transformer. Thông thường trong bộ mã hóa Transformer, các embedding vị trí được cộng vào mỗi vị trí của chuỗi đầu vào BERT. Tuy nhiên, khác với bộ mã hóa Transformer nguyên bản, BERT sử dụng các embedding vị trí có thể học được cho thấy các embedding của chuỗi đầu vào BERT là tổng các embedding của token, embedding đoạn và embedding vị trí.

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 2-10 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.

Fine-tuning

Gọi L là số lớp Transformer (blocks) được sử dụng với kích thước của các lớp ẩn là H và số heads ở lớp attention là A.

Đối với các nhiệm vụ phân loại câu, BERT được fine-tuning rất đơn giản. Để có được biểu diễn của một chuỗi đầu vào với số chiều cố định chỉ cần lấy hidden state ở lớp cuối cùng, tức là đầu ra của lớp Transformer cho token đầu tiên (token đặc biệt [CLS] được xây dựng cho đầu chuỗi). Luận văn gọi véc-tơ này là C ($C \in R^H$).

Chỉ có một tham số được thêm vào quá trình fine-tuning là W ($W \in R^{KxH}$) với K là số nhãn lớp phân loại.

Xác suất của nhãn P là một phân phối với P ∈ R^K được tính toán bởi một hàm sigmoid:

$$P = sigmoid(CW^T)$$

Tất cả các tham số của BERT và W được fine-tuning để tối ưu hóa hàm lỗi.

Trong phân loại đa nhãn thay vì softmax, luận văn sử dụng sigmoid để lấy xác suất. Ký hiệu tập câu hỏi $X = x_1, x_2, \dots x_n$ với n là số câu hỏi, và tập nhãn $y = y_1, y_2, \dots y_m$ với m là số nhãn.

Trong phân loại nhị phân đơn giản, không có sự phân biệt lớn giữa hai loại, tuy nhiên trong trường hợp đa nhãn, sigmoid cho phép xử lý các nhãn không độc lập, trong khi softmax xử lý các lớp độc lập.

Hàm sigmoid được biểu diễn theo công thức:

$$\sigma(z) = \frac{1}{1 + \exp(-z)}, z \in R$$

Trong đó z là vec-tơ k chiều của giá trị thực xác suất nhãn.

Với hàm kích hoạt sigmoid ở lớp đầu ra, mô hình xác suất mạng nơ-ron của một lớp c_j là phân phối Bernoulli:

$$P(c_j|x_i) = \frac{1}{1 + \exp(-z_j)}$$

Sau khi sử dụng hàm sigmoid thì xác suất của mỗi lớp là độc lập với xác suất của lớp khác.

Giải pháp theo phân loại nhị phân

Theo phân loại nhị phân, mô hình sẽ chạy qua từng nhãn. Tại các lần xét duyệt từng nhãn, so sánh xác suất là nhãn hay không là nhãn đó để đưa ra dự đoán. Xác suất trường hợp nào lớn hơn thì nó là nhãn của câu hỏi đó.

Nhãn của câu hỏi sẽ được xác định bởi:

$$c = argmax\{P(c_i|x_i)\}$$

Giải pháp theo phân loại đa nhãn

Theo phân loại đa nhãn sẽ chọn ra ngưỡng để so sánh các xác suất của các nhãn. Nếu nhãn có xác suất vượt qua nhưỡng thì nhãn đó được lựa chọn là nhãn phù hợp của câu hỏi đó. Thông thường các bài toán lựa chọn ngưỡng là 0,5.

2.6 Kết luận chương

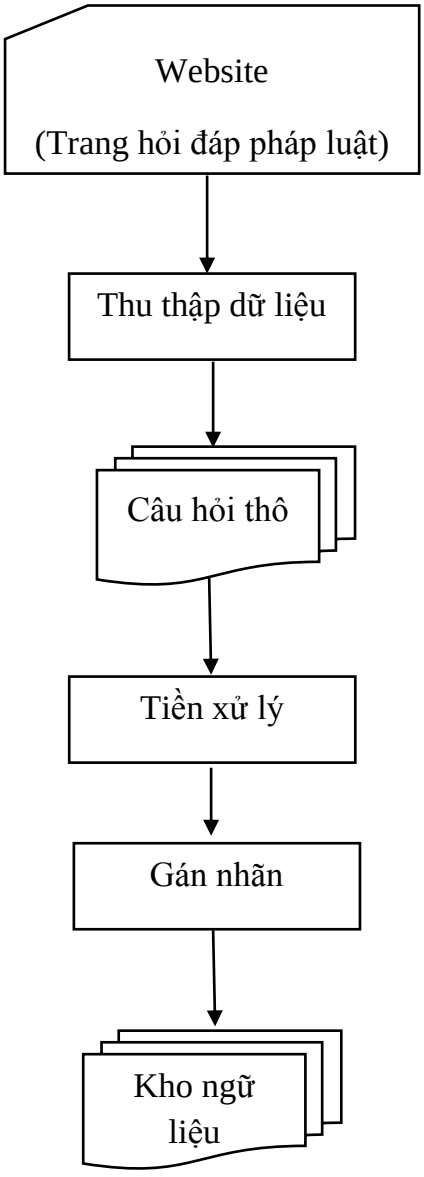
Nội dung chương đã giới thiệu được bài toán phân loại đa nhãn câu hỏi tiếng Việt, giới thiệu được một số mô hình học sâu, giới thiệu phương pháp BERT và đưa ra được mô hình phân loại câu hỏi pháp quy tiếng Việt sử dụng mô hình BERT.

CHƯƠNG 3: THỰC NGHIỆM ĐÁNH GIÁ

Chương này trình bày cách xây dựng kho ngữ liệu, mô tả cách thiết lập thực nghiệm, đưa ra các mô hình thực nghiệm, giới thiệu các công cụ được sử dụng trong bài toán và đáh giá kết quả thực nghiệm.

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

Dữ liệu câu hỏi pháp quy thuộc về các lĩnh vực của pháp luật về xã hội. Do đó, luận văn tiến gán nhãn cho dữ liệu câu hỏi pháp quy dựa trên nội dung câu hỏi.

Luận văn gán nhãn dựa theo nội dung câu hỏi và câu trả lời. Câu trả lời của câu hỏi là các bộ luật, thông tư, điều luật. Dựa vào nội dung câu hỏi và câu trả lời, luận văn phân loại câu hỏi thuộc vào bộ luật nào. Khi các câu hỏi đã được xác định thuộc bộ luật nào thì tiến hành nhóm các bộ luật thuộc cũng một lĩnh vực làm một và đặt tên chung cho nhóm bộ luật đó. Từ đó xây dựng được tập nhãn cho bộ dữ liệu.

Tập nhãn phụ thuộc vào miền bộ luật. Vì vậy, luận văn lựa chọn nhãn phụ thuộc vào bộ luật liên quan của câu hỏi trong bộ dữ liệu thử nghiệm.

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	Văn bản quy phạm pháp luật hết hiệu lực trong trường hợp nào?

luật	
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?

Giai đoạn gán nhãn thủ công luận văn thực hiện hai người gán nhãn. Vì vậy, luận văn cần biết được xem kết quả gán nhãn của hai người có tương đồng với nhau không. Để kiểm tra được điều đó, luận văn sử dụng độ đo Cohen's kappa tính toán độ tương đồng gán nhãn giữa hai người.

Công thức:

$$K = \frac{p_o - p_e}{1 - p_e}$$

Trong đó: p_0 là xác suất tương đối giữa 2 người.

p_e là xác suất ngẫu nhiên giữa 2 người.

Ví dụ: Có 2 người A và B cùng duyệt một tập hồ sơ gồm 50 bộ, mỗi kết quả được đọc bởi 2 người, mỗi người nói "đủ" hoặc "thiếu" ám chỉ hồ sơ đủ giấy tờ hoặc thiếu giấy tờ. Ta có kết quả duyệt của 2 người như sau:

		B	
		Đủ	Thiếu
A	Đủ	20	5
	Thiếu	10	15

Khi đó: p_0 = (20 + 15) / 50 = 0.70

Xác suất người A đọc "Đủ" là 50%

Xác suất người B đọc "Đủ" là 60%

Xác suất cả 2 người đọc "Đủ" là : 0.5*0.6=0.3

Xác suất cả 2 người đọc "Thiếu" là : 0.5*0.4=0.2

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.

3.1.4 Thống kê kho ngữ liệu

Dữ liệu gồm 5896 câu lĩnh vực pháp luật. Nội dung về những câu hỏi về pháp luật.

- Tổng số câu: 5896.
- Tổng số từ: 324095.
- Tổng từ trung bình trên câu: 54.
- Số từ (không tính lặp) trên toàn bộ kho ngữ liệu: 1285.

Tổng tag: 34.

Phân bố nhãn từ loại được trình bày trong bảng 3-2.

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

Nhãn	Số câu hỏi	Tỉ lệ trong kho ngữ liệu (%)
Ban hành văn bản quy phạm pháp luật	18	0,31
Bảo hiểm	29	0,49
Bảo vệ môi trường	12	0,20
Cán bộ, công chức, viên chức	14	0,24
Công chứng	327	5,55
Công dân	405	6,87
Cư trú	162	2,75
Dân sự	1234	20,93
Giao thông đường bộ	65	1,10
Giám định tư pháp	22	0,37
Hình sự	484	8,21
Hôn nhân và gia đình	552	9,36
Khiếu nại, tố cáo	42	0,71
Kinh tế	114	1,93
Lao động	90	1,53
Lý lịch tư pháp	91	1,54

Nhà ở	75	1,27
Nuôi con nuôi	135	2,29
Phí và lệ phí	83	1,41
Phòng, chống ma túy	47	0,80
Quản lý, sử dụng	13	0,22
Quốc phòng	16	0,27
Quốc tịch Việt Nam	67	1,14
Thi hành án	636	10,79
Thuế	30	0,51
Trách nhiệm bồi thường của Nhà nước	120	2,04
Tố tụng	317	5,38
Tổ chức chính phủ	193	3,27
Tổ chức cơ quan, chính quyền	20	0,34
Xây dựng	24	0,41
Xử lý vi phạm hành chính	263	4,46
Đất đai	469	7,95
Đấu giá tài sản	30	0,51
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

Ví dụ:

- Câu hỏi có 1 nhãn là: Quy định của pháp luật về văn phòng giám định tư pháp? (câu hỏi mang nhãn Giám định tư pháp).
- Câu hỏi có 2 nhãn là: Pháp luật có cho phép thay đổi họ cho con theo họ của ông nội không? (câu hỏi mang nhãn Dân sư, Công dân).
- Câu hỏi có 3 nhãn là: Hợp đồng tặng cho đất hiệu lực pháp luật kể từ thời điểm có đầy đủ chữ ký của các bên có đúng không? (câu mang nhãn Công chứng, Dân sự, Đất đai).
- Câu hỏi có 4 nhãn là: Thủ tục đăng ký kết hôn với người Việt Nam nhưng lại định cư ở nước ngoài? (câu hỏi mang nhãn Hôn nhân và gia đình, Quốc tịch Việt Nam, Tổ chức cơ quan, chính quyền, Công dân).

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 được xác định như sau:

$$precision = \frac{số\ nhãn\ gán\ đúng}{tổng\ số\ nhãn\ được\ gán}$$

$$recall = \frac{số\ nhãn\ gán\ đúng}{tổng\ số\ nhãn\ thực\ tế}$$

$$F1 = \frac{2 * precision * recall}{precision + recall}$$

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.

Sklearn svm Linear SVC

Sklearn svm Linear SVC tương tự như SVC với tham số kernel = “linear”, nhưng được triển khai dưới dạng liblinear chứ không phải libsvm, nó linh hoạt hơn trong việc lựa chọn các hàm penalties và hàm loss và nên mở rộng quy mô tốt hơn đến số lượng lớn dữ liệu.

Để cài đặt công cụ dùng lệnh:

Pip install sklearn

Simpletransformer

Simpletransformer model được xây dựng với một nhiệm vũ xử lý ngôn ngữ tự nhiên cụ thể. Mỗi mô hình như vậy được trang bị các tính năng và chức năng được thiết kế để phù hợp nhất với nhiệm vụ mà chúng dự định thực hiện.

Để cài đặt sử dụng lệnh:

Pip install simpletransformers

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.

Nghiên cứu đã thực hiện 2 loại phân loại là phân loại nhị phân cho từng nhãn và phân loại đa nhãn sử dụng 3 mô hình SVM, BERT multi language và PHOBERT.

❖ 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.

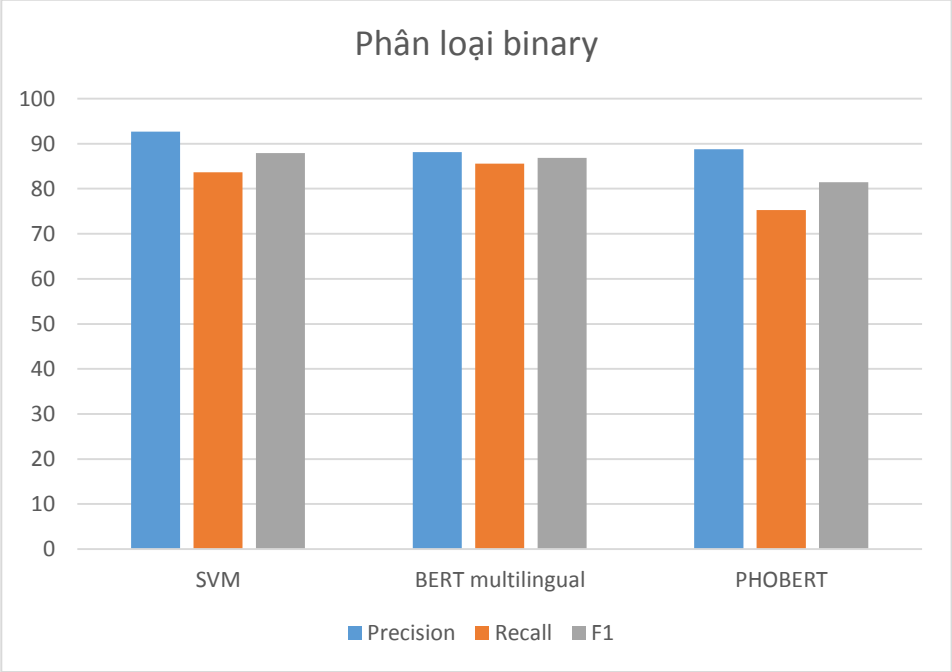
Xây dựng model huấn luyện PHOBERT có hai lựa chọn là Fairseq và Transformer. Ở đây luân văn lựa chọn thử nghiệm với Transformer và sử dụng BertForSequenceClassification để tạo model. Trong phân loại binary luận văn thực hiện huấn luyện với số lượng epochs là 10, batch_size là 32, hidden_dropout_prob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

Luận văn tiến hành làm thực nghiệm với từng nhãn. Kết quả thực nghiệm từng phương pháp khá khả quan. Dưới đây là bảng kết quả các mô hình luận văn làm thực nghiệm.



Hình 3-2 Biểu đồ kết quả thực nghiệm phân loại binary của 3 mô hình.

Bảng 3-4 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.

Kết quả chi tiết cho từng nhãn được trình bày ở dưới đây:

Bảng *3-5* Kết quả thực nghiệm phân loại binary sử dụng mô hình SVM

Nhãn	Precision(%)	Recall(%)	F1(%)
Ban hành văn bản quy phạm pháp luật	99,66	99,66	99,66
Bảo hiểm	99,75	99,75	99,75

Bảo vệ môi trường	99,92	99,92	99,92
Cán bộ, công chức, viên chức	99,58	99,58	99,58
Công chứng	99,24	99,24	99,24
Công dân	98,14	98,14	98,14
Cư trú	99,24	99,24	99,24
Dân sự	95,76	95,76	95,76
Giao thông đường bộ	99,58	99,58	99,58
Giám định tư pháp	99,83	99,83	99,83
Hình sự	98,81	98,81	98,81
Hôn nhân và gia đình	97,37	97,37	97,37
Khiếu nại, tố cáo	99,66	99,66	99,66
Kinh tế	99,58	99,58	99,58
Lao động	99,66	99,66	99,66
Lý lịch tư pháp	99,58	99,58	99,58
Nhà ở	99,58	99,58	99,58
Nuôi con nuôi	99,58	99,58	99,58
Phí và lệ phí	99,58	99,58	99,58
Phòng, chống ma túy	99,83	99,83	99,83
Quản lý, sử dụng	99,83	99,83	99,83
Quốc phòng	99,75	99,75	99,75
Quốc tịch Việt Nam	99,83	99,83	99,83
Thi hành án	98,64	98,64	98,64
Thuế	99,58	99,58	99,58

Trách nhiệm bồi thường của Nhà nước	99,92	99,92	99,92
Tố tụng	96,44	96,44	96,44
Tổ chức chính phủ	99,15	99,15	99,15
Tổ chức cơ quan, chính quyền	99,83	99,83	99,83
Xây dựng	99,66	99,66	99,66
Xử lý vi phạm hành chính	99,07	99,07	99,07
Đất đai	98,05	98,05	98,05
Đấu giá tài sản	99,66	99,66	99,66
Đầu tư	99,75	99,75	99,75

Bảng 3-6 Kết quả thực nghiệm phân loại binary sử dụng mô hình BERT

Nhãn	Precision(%)	Recall(%)	F1(%)
Ban hành văn bản quy phạm pháp luật	99,58	99,58	99,58
Bảo hiểm	99,75	99,75	99,75
Bảo vệ môi trường	99,75	99,75	99,75
Cán bộ, công chức, viên chức	99,66	99,66	99,66
Công chứng	98,64	98,64	98,64
Công dân	98,39	98,39	98,39
Cư trú	98,22	98,22	98,22
Dân sự	96,44	96,44	96,44
Giao thông đường bộ	99,83	99,83	99,83
Giám định tư pháp	99,41	99,41	99,41
Hình sự	98,64	98,64	98,64
Hôn nhân và gia đình	97,46	97,46	97,46

Khiếu nại, tố cáo	99,75	99,75	99,75
Kinh tế	99,75	99,75	99,75
Lao động	99,66	99,66	99,66
Lý lịch tư pháp	99,41	99,41	99,41
Nhà ở	99,41	99,41	99,41
Nuôi con nuôi	99,66	99,66	99,66
Phí và lệ phí	99,83	99,83	99,83
Phòng, chống ma túy	99,83	99,83	99,83
Quản lý, sử dụng	99,92	99,92	99,92
Quốc phòng	99,83	99,83	99,83
Quốc tịch Việt Nam	99,49	99,49	99,49
Thi hành án	98,81	98,81	98,81
Thuế	99,58	99,58	99,58
Trách nhiệm bồi thường của Nhà nước	100,0	100,0	100,0
Tố tụng	97,63	97,63	97,63
Tổ chức chính phủ	99,15	99,15	99,15
Tổ chức cơ quan, chính quyền	99,83	99,83	99,83
Xây dựng	99,41	99,41	99,41
Xử lý vi phạm hành chính	98,64	98,64	98,64
Đất đai	97,97	97,97	97,97
Đấu giá tài sản	99,66	99,66	99,66
Đầu tư	99,75	99,75	99,75

Bảng 3-7 Kết quả thực nghiệm phân loại binary sử dụng mô hình PHOBERT

Nhãn	Precision(%)	Recall(%)	F1(%)
Ban hành văn bản quy phạm pháp luật	99,58	99,58	99,58
Bảo hiểm	99,58	99,58	99,58
Bảo vệ môi trường	99,75	99,75	99,75
Cán bộ, công chức, viên chức	99,66	99,66	99,66
Công chứng	98,98	98,98	98,98
Công dân	97,54	97,54	97,54
Cư trú	98,90	98,90	98,90
Dân sự	95,08	95,08	95,08
Giao thông đường bộ	99,49	99,49	99,49
Giám định tư pháp	99,41	99,41	99,41
Hình sự	98,81	98,81	98,81
Hôn nhân và gia đình	96,95	96,95	96,95
Khiếu nại, tố cáo	99,66	99,66	99,66
Kinh tế	99,49	99,49	99,49
Lao động	99,75	99,75	99,75
Lý lịch tư pháp	99,41	99,41	99,41
Nhà ở	99,41	99,41	99,41
Nuôi con nuôi	99,49	99,49	99,49
Phí và lệ phí	98,39	98,39	98,39
Phòng, chống ma túy	99,92	99,92	99,92
Quản lý, sử dụng	99,92	99,92	99,92

Quốc phòng	99,83	99,83	99,83
Quốc tịch Việt Nam	99,66	99,66	99,66
Thi hành án	98,39	98,39	98,39
Thuế	99,24	99,24	99,24
Trách nhiệm bồi thường của Nhà nước	99,92	99,92	99,92
Tố tụng	95,42	95,42	95,42
Tổ chức chính phủ	98,47	98,47	98,47
Tổ chức cơ quan, chính quyền	99,83	99,83	99,83
Xây dựng	99,41	99,41	99,41
Xử lý vi phạm hành chính	98,47	98,47	98,47
Đất đai	97,20	97,20	97,20
Đấu giá tài sản	99,66	99,66	99,66
Đầu tư	99,32	99,32	99,32

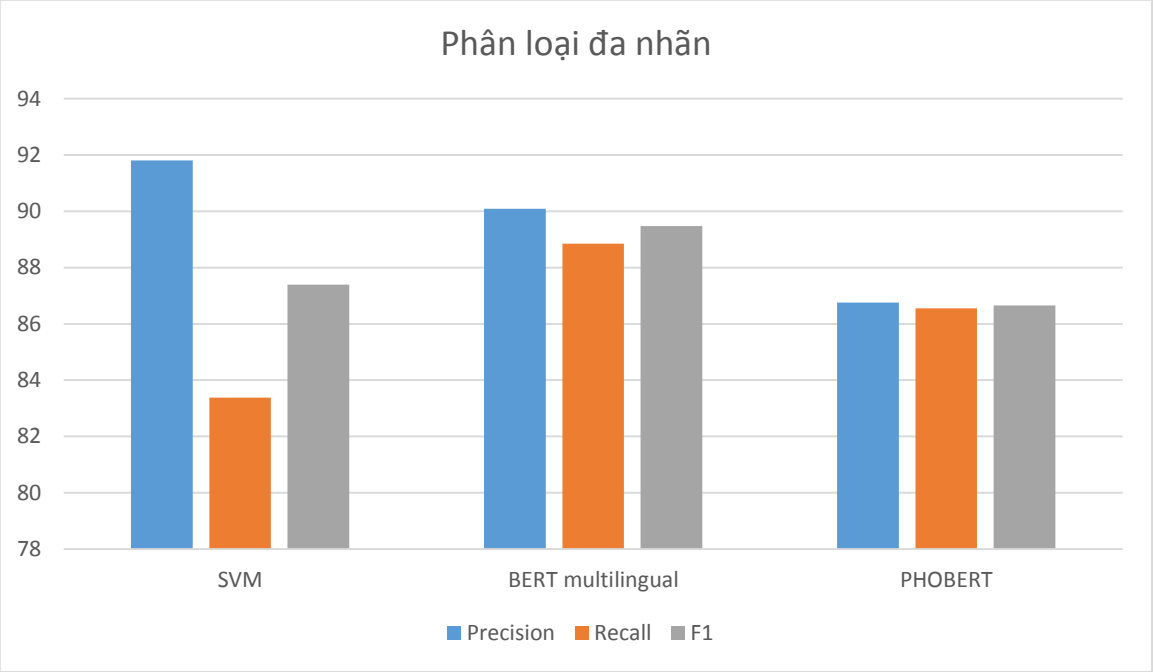
Từ các bảng kết quả trên nhận thấy:

- Kết quả phân loại nhị phân từng nhãn của các mô hình khá tương đồng. Các nhãn được phân loại đạt kết quả khá tốt, đều trên 95%.

- Nhãn "Trách nhiệm bồi thường của Nhà nướ" đạt kết quả chính xác nhất (100%) với mô hình BERT.

3.5.2 Phân loại đa nhãn

Luận văn tiến hành thực nghiệm phân loại đa nhãn sử dụng các mô hình được trình bày ở phần 3.3.



Hình 3-3 Biểu đồ kết quả thực nghiệm phân loại đa nhãn của 3 mô hình.

Bảng 3-*8* Bảng kết quả thực nghiệm phân loại đa nhãn của 3 mô hình

Mô hình	PRECISION(%)	RECALL(%)	F1(%)
SVM	91,81	83,38	87,39
BERT multilingual	90,09	88,85	89,47
PHOBERT	86,76	86,55	86,65

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.

Kết quả chi tiết các nhãn được trình bày ở dưới đây:

Bảng 3-9 Bảng kết quả thực nghiệm các nhãn phân loại đa nhãn sử dụng mô hình SVM

Nhãn	Precision(%)	Recall(%)	F1(%)
Ban hành văn bản quy phạm pháp luật	60,0	31,0	38,71
Bảo hiểm	100,0	79,74	87,90
Bảo vệ môi trường	90,0	90,0	90,0
Cán bộ, công chức, viên chức	50,0	30,0	35,52
Công chứng	92,93	85,58	89,06
Công dân	90,67	85,14	87,65
Cư trú	94,67	83,71	88,60
Dân sự	89,57	86,15	87,80
Giao thông đường bộ	90,62	79,06	83,63
Giám định tư pháp	100,0	79,09	87,0
Hình sự	94,59	89,91	92,17
Hôn nhân và gia đình	88,66	80,16	84,18
Khiếu nại, tố cáo	92,78	90,56	91,20
Kinh tế	95,88	83,15	88,88
Lao động	95,10	86,07	90,09
Lý lịch tư pháp	89,67	95,18	92,27
Nhà ở	87,23	64,97	74,15
Nuôi con nuôi	97,0	92,10	94,37
Phí và lệ phí	95,03	84,67	89,20
Phòng, chống ma túy	94,79	92,90	93,43
Quản lý, sử dụng	80,0	54,17	62,90

Quốc phòng	65,0	41,33	47,56
Quốc tịch Việt Nam	95,20	84,67	89,37
Thi hành án	96,53	92,57	94,49
Thuế	90,64	63,46	73,19
Trách nhiệm bồi thường của Nhà nước	99,30	96,24	97,70
Tố tụng	85,81	57,37	68,52
Tổ chức chính phủ	89,05	79,01	83,33
Tổ chức cơ quan, chính quyền	30,0	7,78	11,52
Xây dựng	100,0	67,19	79,28
Xử lý vi phạm hành chính	96,39	84,79	90,03
Đất đai	87,03	81,11	83,87
Đấu giá tài sản	92,50	63,27	74,78
Đầu tư	93,42	76,89	83,01

Bảng 3-10 Bảng kết quả thực nghiệm các nhãn phân loại đa nhãn sử dụng mô hình BERT

Nhãn	Precision(%)	Recall(%)	F1(%)
Ban hành văn bản quy phạm pháp luật	76,67	40,50	51,33
Bảo hiểm	84,94	88,78	85,95
Bảo vệ môi trường	50,0	30,0	36,33
Cán bộ, công chức, viên chức	0,0	0,0	0,0
Công chứng	90,36	90,48	90,39
Công dân	89,67	92,95	91,20
Cư trú	95,26	89,36	92,08

Dân sự	89,93	90,75	90,32
Giao thông đường bộ	81,54	83,36	82,02
Giám định tư pháp	84,52	78,31	80,10
Hình sự	93,56	95,74	94,62
Hôn nhân và gia đình	86,67	86,42	86,44
Khiếu nại, tố cáo	92,63	94,78	93,32
Kinh tế	90,53	87,94	89,17
Lao động	92,16	91,61	91,38
Lý lịch tư pháp	97,70	93,94	95,71
Nhà ở	74,41	85,36	78,74
Nuôi con nuôi	93,79	93,17	93,25
Phí và lệ phí	83,57	86,45	84,47
Phòng, chống ma túy	91,25	100,0	95,25
Quản lý, sử dụng	45,0	27,5	32,0
Quốc phòng	68,33	54,83	58,10
Quốc tịch Việt Nam	94,23	88,05	90,88
Thi hành án	95,56	94,99	95,24
Thuế	97,50	72,02	81,28
Trách nhiệm bồi thường của Nhà nước	97,39	99,57	98,45
Tố tụng	84,36	76,27	79,93
Tổ chức chính phủ	89,52	88,62	88,95
Tổ chức cơ quan, chính quyền	10,0	5,0	6,67
Xây dựng	97,50	47,35	62,34

Xử lý vi phạm hành chính	91,26	89,07	90,04
Đất đai	87,88	87,34	87,56
Đấu giá tài sản	76,24	68,04	70,64
Đầu tư	80,42	83,49	79,55

Bảng 3-11 Bảng kết quả thực nghiệm các nhãn phân loại đa nhãn sử dụng mô hình PHOBERRT

Nhãn	Precision(%)	Recall(%)	F1(%)
Ban hành văn bản quy phạm pháp luật	66,67	29,33	39,43
Bảo hiểm	82,67	83,37	80,96
Bảo vệ môi trường	60,0	40,0	46,33
Cán bộ, công chức, viên chức	45,0	34,17	36,38
Công chứng	88,63	88,41	88,42
Công dân	84,86	90,33	87,30
Cư trú	87,90	86,38	86,83
Dân sự	86,85	88,43	87,61
Giao thông đường bộ	74,88	80,13	75,85
Giám định tư pháp	100,0	80,75	88,76
Hình sự	90,41	92,82	91,47
Hôn nhân và gia đình	85,34	83,25	83,99
Khiếu nại, tố cáo	88,39	89,42	88,44
Kinh tế	89,04	88,77	88,58
Lao động	92,28	85,87	88,46
Lý lịch tư pháp	89,35	95,36	92,08
Nhà ở	70,10	67,02	67,55

Nuôi con nuôi	92,0	92,75	92,22
Phí và lệ phí	92,26	74,59	82,14
Phòng, chống ma túy	91,53	95,33	93,10
Quản lý, sử dụng	80,0	48,33	58,0
Quốc phòng	74,17	56,83	63,0
Quốc tịch Việt Nam	88,71	89,08	88,70
Thi hành án	92,86	94,69	93,74
Thuế	88,56	71,56	77,19
Trách nhiệm bồi thường của Nhà nước	94,66	97,48	95,83
Tố tụng	79,56	69,18	73,64
Tổ chức chính phủ	85,75	86,84	86,18
Tổ chức cơ quan, chính quyền	35,0	16,43	20,83
Xây dựng	100,0	59,44	72,99
Xử lý vi phạm hành chính	86,22	86,22	85,52
Đất đai	82,21	87,35	84,35
Đấu giá tài sản	73,49	59,11	64,19
Đầu tư	83,54	87,92	85,11

Từ bảng kết quả thực nghiệm các nhãn nhận thấy:

- Nhãn "Cán bộ, Công chức, Viên chức" của hai mô hình SVM và PHOBERT có kết quả thấp như nhau (36,38%). Với mô hình BERT không có kết quả dự đoán nào chính xác. Điều này có thể do lượng nhãn này trong kho ngữ liệu còn khác ít (chiếm 0,24% kho ngữ liệu) nên việc huấn luyện chưa được tốt dẫn đến kết quả dự đoán chưa được tốt.

- Tương tự nhãn “Tổ chức cơ quan, chính quyền” cũng có kết quả thấp, kết quả sử dụng mô hình PHOBERT (đạt 20,83%) cao hơn hai mô hình SVM (11,52%) và mô hình BERT (6,67%).
- Nhãn “Trách nhiệm bồi thường của Nhà nước” cho kết quả dự đoán tốt nhất với các mô hình, trong đó mô hình BERTcho kết quả cao nhất (98,45%), cao hơn mô hình SVM (97,70%) và mô hình PHOBERT (95,83%). Nhãn này chiếm 2,04% kho ngữ liệu.
- Kết quả dự đoán các nhãn cho thấy các nhãn được dự đoán thấp có thể do lượng dữ liệu nhãn đó trong bộ dữ liệu chưa được nhiều để huấn luyện tốt, hoặc do lượng phân bổ dữ liệu trong bộ train/test chưa được đồng đều, lượng dữ liệu huấn luyện ít còn lượng dữ liệu test chiếm đa số hoặc ngược lại.

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.

TÀI LIỆU THAM KHẢO

Tiếng Việt

[1] Nguyễn Đức Vinh, Phân tích câu hỏi trong hệ thống hỏi đáp tiếng Việt, Khóa luận tốt nghiệp đại học, Đại học quốc gia Hà Nội, 2009.

[2] Nguyễn Minh Thành, Phân loại văn bản, Đồ án môn học Xử lý ngôn ngữ tự nhiên, Đại học quốc gia Thành phố Hồ Chí Minh, 01/2011.

[3] Vu Thi Tuyen, Một số mô hình học máy trong phân loại câu hỏi, Đại học Công nghê, 2016

[4] Nguyễn Thị Hương Thảo. Phân lớp phân cấp Taxonomy văn bản Web và ứng dụng. Khóa luận tốt nghiệp đại học, Đại học Công nghệ, 2006.

[5] Phạm Văn Sơn, Tìm hiểu về Support Vector Machine cho bài toán phân lớp quan điểm

Tiếng Anh

[6] Jacob, Devlin Ming-Wei Chang, Kenton Lee, Kristina Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019)

[7] Yoon Kim: Convolutional Neural Networks for Sentence Classification, New York University (2014)

[8] Bishal Gaire, Bishal Rijal, Dilip Gautam, Nabin Lamichhane, Saurav Sharma, Insincere Question Classification Using Deep Learning, Nhà xuất bản Viện Kỹ thuật đại học Tribhuvan, Nepal.

[9] J. Pennington, R. Socher, and C. Manning, ―Glove: Global Vectors for Word Representation,‖ Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014.

[10] Wieting, John Bansal, Mohit Gimpel, Kevin Livescu, Karen, ―Towards universal paraphrastic sentence embeddings,‖ arXiv preprint arXiv:1511.08198, 2015.

[11] Prudhvi Raj Dachapally, In-depth Question classification using Convolutional Neural Networks, Trường Tin học và máy tính Bloomington, U.S.A.

[12] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," CoRR, vol. abs/1404.2188, 2014. [Online]. Available: http://arxiv.org/abs/1404.2188

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/ 4824-imagenet-classification-with-deep-convolutional-neural-networks. Pdf.

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, November 1998.

[15] Thi-Ngan Pham, Van-Quang Nguyen, Van-Hien Tran, Tri-Thanh Nguyen, Quang-Thuy Ha: A semi-supervised multi-label classification framework with feature reduction and enrichment, JOURNAL OF INFORMATION AND TELECOMMUNICATION, 2017 VOL. 1, NO. 2, 141–154

[16] David Vilar, Maria Jose Castro và Emilio Sanchis, Multi-label text classification using multinomial models(2004)

Trang Web

[17] itechseeker.com/tutorials/nlp-with-deep-learning/ly-thuyet-chung/cac-bien-the-cua-rnn/

[18] https://towardsdatascience.com/transformers-for-multilabel-classification

[19] https://machinelearningcoban.com/2018/01/14/id3/

[20] https://machinelearningcoban.com/2017/08/08/nbc/

[21] https://machinelearningcoban.com/2017/01/08/knn/

[22]https://dominhhai.github.io/vi/2017/10/what-is-lstm/

[23] https://viblo.asia/p/bert-buoc-dot-pha-moi-trong-cong-nghe-xu-ly-ngon-ngu-tu-nhien-cua-google-RnB5pGV7lPG

[24] http://itechseeker.com/tutorials/nlp-with-deep-learning/ly-thuyet-chung/recurrent-neural-network/

[25] https://nttuan8.com/bai-6-convolutional-neural-network/

[26] https://viblo.asia/p/hieu-hon-ve-bert-buoc-nhay-lon-cua-google-eW65GANOZDO

[27] https://viblo.asia/p/bert-roberta-phobert-bertweet-ung-dung-state-of-the-art-pre-trained-model-cho-bai-toan-phan-loai-van-ban

[28] http://itechseeker.com/tutorials/nlp-with-deep-learning/ly-thuyet-chung/convolutional-neural-network/