APPLICATION OF BERT ARCHITECTURE FOR STORAGE TIME OF RECORD CLASSIFICATION PROBLEM

Ton Nu Thi Sau*, Tran Quoc Toanh
Hanoi University of Home Affairs Campus in HCM City

ARTICLE INFO

Received: 06/02/2021

Revised: 19/4/2021

Published: 04/5/2021

KEYWORDS

BERT architecture
Machine learning
Deep learning
Record classification
Text classification

ABSTRACT

Record storage at the competent agencies and organizations is an essential problem in the management and organization of document preservation. However, with the increasing number of archives and many different types of documents, leading to overloading documents during the archiving process. Therefore, the classification of records according to the preservation period is a very important step in preservation, contributing to optimize the composition of the archive fonts, and save the cost of document Therefore, in this paper, we present a study evaluating the effectiveness of the BERT model compared with traditional machine learning and deep learning algorithms on a real-world dataset to solve this task automatically. Experimental results show that the BERT model achieved the best results with 93.10% of precision, 90.68% of recall and 91.49% of F1-score. This result shows that the BERT model can be applied to build systems to support record classification in the real-world application is completely feasible.

ỨNG DỤNG MÔ HÌNH BERT CHO BÀI TOÁN PHÂN LOẠI HỒ SƠ THEO THỜI HẠN BẢO QUẢN

Tôn Nữ Thị Sáu*, Trần Quốc Toanh
Phân hiệu Trường Đại học Nội vụ Hà Nội tại TP. Hồ Chí Minh

THÔNG TIN BÀI BÁO

Ngày nhận bài: 06/02/2021

Ngày hoàn thiện: 19/4/2021

Ngày đăng: 04/5/2021

TỪ KHÓA

Kiến trúc BERT
Máy học
Học sâu
Phân loại hồ sơ
Phân loại văn bản

TÓM TẮT

Công tác lưu trữ hồ sơ tại các cơ quan, tổ chức có thẩm quyền là một vấn đề cần thiết trong việc quản lý và tổ chức bảo quản tài liệu. Tuy nhiên, hiện nay với số lượng hồ sơ lưu trữ ngày càng nhiều và có nhiều loại văn bản quy định lưu trữ khác nhau dẫn đến việc tình trạng quá tải tài liệu trong quá trình lưu trữ. Do đó, việc phân loại hồ sơ theo thời hạn bảo quản là một công đoạn rất quan trọng trong việc bảo quản, góp phần tối ưu hóa thành phần trong các phòng lưu trữ, tiết kiệm chi phí bảo quản tài liệu. Để góp phần giải quyết được vấn đề trên, trong bài báo này, chúng tôi trình bày nghiên cứu đánh giá sự hiệu quả của mô hình BERT so sánh với các thuật toán máy học truyền thống và mô hình học sâu trên các bộ dữ liệu thực tế hồ sơ lưu trữ theo thời hạn bảo quản ở các cơ quan. Kết quả nghiên cứu cho thấy rằng, mô hình BERT đạt kết quả tốt nhất với độ chính xác là 93,10%, độ phủ là 90,68% và độ đo F1 là 91,49%. Kết quả này cho thấy rằng, mô hình BERT có thể được áp dụng để xây dựng các hệ thống hỗ trợ phân loại hồ sơ theo thời hạn bảo quản là hoàn toàn khả thi.

DOI: https://doi.org/10.34238/tnu-jst.3990

* Corresponding author. *Email: sauvtc@gmail.com*

1. Giới thiệu

Trong những năm trở lại đây, các doanh nghiệp, cơ quan quản lý nhà nước đều ứng dụng công nghệ thông tin vào hoạt động hàng ngày và các ứng dụng đó đã trở thành công cụ quen thuộc của người dân. Hiện nay, cán bộ, công chức đang xác định thời hạn bảo quản tài liệu theo cách thủ công. Cách này làm mất nhiều thời gian, công sức, dễ nhầm lẫn do số lượng hồ sơ nhiều, đa dạng về lĩnh vực [1]. Mặt khác, có một số hồ sơ hình thành trong quá trình giải quyết công việc không có trong quy định của nhà nước, với cách làm thủ công thì phải tham vấn các chuyên gia chỉnh lý tài liệu có kinh nghiệm, nhưng các ý kiến của chuyên gia thường không đồng nhất. Cho nên, việc xác định thời hạn bảo quản cho hồ sơ tại các Ủy ban nhân dân (UBND) cấp xã chưa được thực hiện một cách triệt để [2]. Trong đó việc phân loại theo thời hạn bảo quản có vai trò rất quan trọng, buộc thực hiện và phải được thực hiện bởi một đội ngũ có chuyên môn về nghiệp vụ văn thư lưu trữ. Bởi vì, mục đích của việc phân loại theo thời hạn bảo quản góp phần tối ưu hóa thành phần trong các phông lưu trữ: Tiết kiệm chi phí bảo quản tài liệu (kho tàng, trang thiết bị, điện. v,v); khắc phục tình trạng hồ sơ, tài liệu tích đống và đặc biệt là việc tiêu hủy hồ sơ, tài liệu tùy tiện. Hiện nay, việc áp dụng các kỹ thuật công nghệ ứng dụng vào giải quyết các bài toán thực tế trong xã hội ngày càng được quan tâm. Điển hình như tác giả N. T. T. Huong và D. M. Trung [3] đã áp dụng thuật toán Random Forest để phân loại bản đồ sử dụng đất, hay tác giả T. C. De and P. N. Khang [4] đã áp dụng phương pháp Support Vector Machine và Cây quyết định để phân loại các văn bản. Gần đây, tác giả D. T. Thanh và các cộng sự [5] đã trình bày một công trình nghiên cứu bài toán phân loại văn bản ứng dụng trong việc phân loại chủ đề cho bài khoa học sử dụng các kỹ thuật máy học như SVM, KNN và Naive Bayes. Kết quả nghiên cứu được thử nghiệm cho tạp chí Đại học Cần Thơ. Đối với bài toán phân loại tên hồ sơ theo thời hạn bảo quản, tác giả T. N T Sau và cộng sự [6] đã nghiên cứu các mô hình máy học truyền thống như SVM kết hợp với các đặc trưng khác nhau. Tuy nhiên, các kết quả vẫn chưa được thử nghiệm ở các kỹ thuật hiệu quả khác.

Nhận thấy được vấn đề trên, chúng tôi thực hiện công trình nghiên cứu các phương pháp, kỹ thuật xử lý dữ liệu văn bản và các mô hình máy học truyền thống cũng như mô hình học sâu trên bộ dữ liệu thực tế về phân loại tên hồ sơ theo thời hạn bảo quản. Mục đích của chúng tôi là nghiên cứu và áp dụng các trí tuệ nhân tạo cho việc hỗ trợ cán bộ, công chức, viên chức thực hiện công việc phân loại hồ sơ theo thời hạn bảo quản. Do đó, trong bài báo này, chúng tôi đề xuất phương pháp dựa trên kiến trúc BERT so sánh với các phương pháp máy học và học sâu trong bài toán phân loại tự động tên hồ sơ tiếng Việt của các UBND cấp xã theo thời hạn bảo quản. Đóng góp của chúng tôi trong bài báo này là ra tìm phương pháp tốt nhất để phân loại tự động tên hồ sơ tiếng Việt của các UBND cấp xã theo thời hạn bảo quản để đánh giá sự khả thi trong việc nghiên cứu các phương pháp trí tuệ nhân tạo áp dụng vào các bài toán trong thực tế.

2. Công trình nghiên cứu liên quan

Với sự phát triển của công nghệ thông tin và ngành trí tuệ nhân tạo nhiều kỹ thuật phân loại học có giám sát đã được phát triển và triển khai trong phần mềm để phân loại dữ liệu chính xác. Công trình nghiên cứu [7] đã trình bày các kết quả thực nghiệm các phương pháp máy học truyền thống như Naive Bayes, SVM cho các bài toán phân loại và đạt kết quả tốt trên nhiều bộ dữ liệu khác nhau. Gần đây với sự phát triển của các mô hình học sâu, tác giả Y.Kim [8] đã áp dụng và đề xuất sử dụng mạng tích chập Convolutional Neural Network (CNN) cho các bài toán phân loại văn bản khác nhau. Kết quả thực nghiệm cho thấy được sự hiệu quả của mô hình CNN trong lĩnh vực Xử lý ngôn ngữ tự nhiên. Sau đó, K.Kowsari và cộng sự [9] đã giới thiệu phương pháp học sâu phân cấp cho phân loại văn bản (HDLTex), đây là sự kết hợp của tất cả các kỹ thuật học sâu trong cấu trúc phân cấp để phân loại tài liệu, mô hình này đã cải thiện độ chính xác so với các mô hình truyền thống. Tiếp theo đó, tác giả K. Kowsari [10] đã đề xuất mô hình Học sâu đa mô hình ngẫu nhiên (RMDL) dành cho phân lớp. Mô hình RMDL giải quyết được vấn đề tìm ra cấu trúc,

kiến trúc học sâu tốt nhất và đồng thời cải thiện sự vững chắc cũng như độ chính xác thông qua quần thể kiến trúc học sâu. Mô hình RMDL có thể chấp nhận dữ liệu đầu vào đa dạng bao gồm văn bản, video, hình ảnh và biểu tượng. Gần đây hơn, một mô hình ngôn ngữ đã huấn luyện từ dữ liệu Bidirectional Encoder Representations from Transformer (BERT) [11] đạt nhiều kết quả tốt cho các bài toán phân loại văn bản.

Đối với sự phát triển của tiếng Việt, các nhà nghiên cứu cũng quan tâm đến các bài toán phân loại văn bản trong những năm gần đây [4]. Các tác giả P. T. Ha và cộng sự [12] sử dụng hai mô hình SVM, Naive Bayes để phân loại tự động tin tức tiếng Việt. Họ thử nghiệm dữ liệu lấy từ các trang tin tức (vietnamnet.vn và vnexpress.net) với mô hình SVM cho độ chính xác 94%. Tiếp theo sau đó, tác giả N. T. Hai và các cộng sự [13] đã nghiên cứu để đánh giá hiệu suất của ba mô hình được sử dụng rộng rãi: Chi-square (CHI), Information Gain (IG), Document Frequency (DF) và đề xuất một mô hình lựa chọn tính năng lai, được gọi là SIGCHI, kết hợp giữa mô hình Chi-square và Information Gain. Kết quả thử nghiệm của họ cho thấy, mô hình họ đề xuất tốt hơn so với các mô hình khác. Độ chính xác của SIGCHI cao hơn 18,65% so với CHI và cao hơn 27,72% so với mô hình DF. Gần đây, tác giả D. T. Thanh và các cộng sự [5] đã trình bày một công trình nghiên cứu bài toán phân loại cho bài báo khoa học, kết quả thực nghiệm cho thấy rằng phương pháp SVM đạt kết quả tốt nhất với độ chính xác lớn hơn 91%. Đối với các mô hình học sâu, tác giả P. Le-Hong and A.-C. Le [14] đã đánh giá hiệu quả của bốn mô hình mạng trí tuệ nhân tạo trên các bộ dữ liệu câu tiếng Việt và Tiếng Anh. Kết quả nghiên cứu đưa ra một số đề xuất khi áp dụng các mô hình mạng nhân tạo cho bài toán phân loại câu. Kế đến, K. D. T. Nguyen và các cộng sự [15] đã trình bày một công trình nghiên cứu đánh giá sự hiệu quả của kiến trúc mạng HAN (Hierarchical Attention Networks) đối với bài toán phân loại chủ đề các bài báo tin tức tiếng Việt. Kết quả so sánh với các mô hình máy học truyền thống cho thấy phương pháp HAN đạt hiệu quả tốt với chỉ số F1 là 86,37%.

Bảng 1. *Thống kê số lượng dữ liệu trong từng tập huấn luyện, kiểm tra và phát triển*

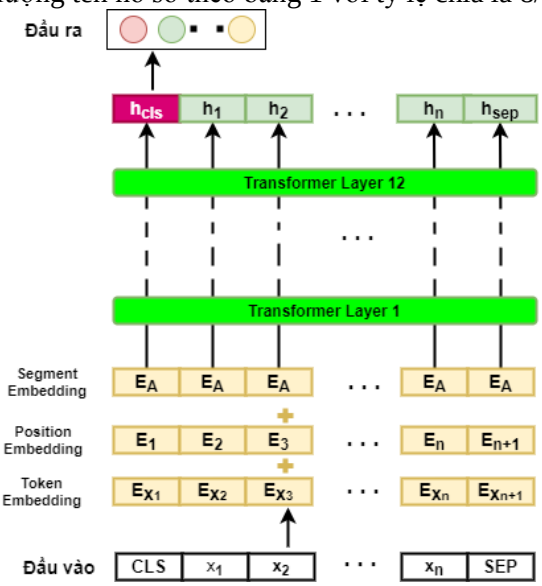
Thời gian lưu trữ	Tập huấn luyện	Tập phát triển	Tập kiểm tra
2 năm	838	90	91
5 năm	140	16	18
10 năm	3 747	431	445
15 năm	748	78	99
20 năm	4 018	437	532
50 năm	140	25	17
70 năm	105	11	7
Vĩnh viễn	4 523	497	549
Theo tuổi thọ công trình	337	37	45

Dựa vào các công trình nghiên cứu liên quan, trong bài báo này chúng tôi nghiên cứu đề xuất sử dụng kiến trúc BERT so sánh với các mô hình máy học truyền thống và mô hình học sâu cho bài toán phân loại tên hồ sơ theo thời hạn bảo quản trên dữ liệu thực tế. Nghiên cứu trong bài báo này có thể được áp dụng vào các hệ thống quản lý lưu trữ tại các cơ quan quản lý hồ sơ để tăng chất lượng quản lý và số hóa thông tin lưu trữ.

3. Thông tin dữ liệu

Để đảm bảo sự khách quan, tính thực tiễn và tính khả thi trong nghiên cứu, chúng tôi thu thập được 18.021 tên hồ sơ từ hai nguồn. Một là, hồ sơ hình thành trong quá trình hoạt động thuộc lĩnh vực địa chính, kế toán, tư pháp, văn phòng và hộ tịch của các UBND phường tại TP. Hồ Chí Minh. Hai là, một số tên nhóm hồ sơ, tài liệu có trong Thông tư số 09/2011/TT-BNV, Thông tư số 46/2016/TTBTNM. Những văn bản này quy định nhóm hồ sơ, tài liệu chung hình thành trong quá trình hoạt động của các UBND cấp xã. Để ứng dụng kết quả nghiên cứu vào thực tế, những tên hồ sơ này đã được các chuyên gia với nhiều năm kinh nghiệm trong ngành lưu trữ gán nhãn

thời hạn bảo quản phù hợp. Đồng thời, các UBND cấp xã chấp nhận thời hạn bảo quản đối với các hồ sơ của họ. Sau khi tiến hành thu thập và gán nhãn theo các thời hạn bảo quản, bộ dữ liệu được trình bày cụ thể số lượng tên hồ sơ theo bảng 1 với tỷ lệ chia là 8/1/1.



Hình 1. *Kiến trúc mô hình BERT cho bài toán phân loại hồ sơ theo thời hạn bảo quản*

Nhìn vào bảng 1, chúng ta dễ dàng nhận thấy được sự mất cân bằng giữa các nhãn dữ liệu với nhau, cụ thể nhãn các tên hồ sơ được lưu vĩnh viễn có tần số nhiều nhất là 4523 tên hồ sơ, tiếp theo là các nhãn 20 năm và 10 năm. Trong khi đó, các nhãn 5 năm, 50 năm xuất hiện tương đối ít. Điều này có thể giải thích được, bởi vì chúng tôi thu thập dữ liệu thực tế ở các trung tâm xử lý lưu trữ tại cơ quan, do đó tỷ lệ các hồ sơ lưu trữ có sự chênh lệch cao. Tuy nhiên, đây cũng chính là thách thức của bộ dữ liệu mà chúng tôi thu thập.

4. Kiến trúc mô hình

Trong bài báo này, chúng tôi đề xuất một mô hình dựa trên kiến trúc BERT. Chúng tôi sử dụng kiến trúc BERT được công bố bởi nghiên cứu của Viện VinAI [16]. Mô hình PhoBERT được tối ưu hoá sử dụng quá trình huấn luyện RoBERTa và được huấn luyện trên 20GB dữ liệu văn bản tiếng Việt. Kết quả được công bố trong bài báo [16] đã chứng tỏ rằng việc sử dụng mô hình BERT như là lớp nhúng từ đem lại kết quả tốt hơn so với các phương pháp học sâu khác. Bởi vì BERT cho phép chúng ta biểu diễn của từ vựng theo ngữ cảnh tốt hơn so với các phương pháp nhúng từ truyền thống trước đây như là word2vec hay Glove. Chính vì lý do đó, chúng tôi tiến hành thử nghiệm đề xuất kiến trúc BERT kết hợp với hàm tuyến tính để áp dụng trong bài toán phân loại hồ sơ theo thời hạn bảo quản. Mô hình được trình bày như ở hình 1. Mô hình bao gồm ba thành phần chính như sau:

Đầu vào: Mỗi tên hồ sơ đầu vào đã được tiền xử lý X với n từ vựng có dạng như sau: $X_{1:n}$ = $x_1, x_2, \ldots, x_n$ với x_i là vị trí thứ i trong chuỗi đầu vào sẽ được tách thành các từ vựng và được biểu diễn thành các giá trị số dựa trên tập từ điển đã huấn luyện của mô hình phoBERT [15]. Bên cạnh đó, vị trí của từng mẫu từ cũng được lấy để làm đầu vào cho mô hình BERT. Chúng tôi lựa chọn tên hồ sơ dài nhất trong tập huấn luyện là giá trị độ dài đầu vào, đối với các câu có độ dài ngắn hơn sẽ tự động được thêm giá trị <pad>.

BERT mã hóa: Trong bài báo này, chúng tôi sử dụng kiến trúc $BERT_{base}$ với 12 khối Transformer và 12 self-attention để lấy đặc trưng biểu diễn cho chuỗi đầu vào với kích thước không quá 512 từ vựng. Đầu ra của mô hình này là một lớp ẩn H = $\{h_1, h_2, \ldots, h_n\}$ tương ứng

với chuỗi đầu vào. Để phân loại tên hồ sơ theo thời hạn bảo quản, chúng tôi rút trích lớp đặc trưng biểu diễn của từ vựng [CLS] làm vectơ đặc trưng biểu diễn cho chuỗi tên hồ sơ đầu vào.

Đầu ra: Với vectơ đại diện cho chuỗi đầu vào, chúng tôi sử dụng một bộ phân lớp với hàm kích hoạt softmax để tính toán giá trị phân bố xác suất của từng nhãn phân loại theo thời hạn bảo hành của hồ sơ**.**

$$p(c|h) = softmax(Wh) \quad (1)$$

trong đó, W là trọng số của lớp tuyến tính.

Bởi vì chúng tôi thu thập tên hồ sơ từ các UBND cấp xã, đây là một nguồn dữ liệu tương đối sạch. Tuy nhiên, để tăng độ chính xác cho mô hình phân lớp, chúng tôi tiến hành tiền xử lý dữ liệu trước khi đưa vào mô hình để huấn luyện. Các bước tiền xử lý được trình bày như sau:

+ **Bước 1:** Loại bỏ các thành phần gây nhiễu trong đầu vào như ký tự đặc biệt, khoảng trắng thừa, dấu chấm, dấu phẩy hay dấu gạch ngang.

+ **Bước 2:** Chúng tôi sử dụng biểu thức chính quy để thay thế các dữ liệu số thành ký từ "num", ngày tháng thành "date", năm thành "year".

+ **Bước 3:** Đưa các từ viết tắt thành các cụm từ có nghĩa tương ứng ví dụ như "QSDĐ"= "Quyền sử dụng đất" hay "GCNQSDĐ" = "Giấy chứng nhận quyền sở dụng đất".

+ **Bước 4:** Đưa các từ đồng nghĩa về một định dạng từ duy nhất để thống nhất ý nghĩa của dữ liệu. Ví dụ như là "thiếu số", "khuyết số".

+ **Bước 5:** Tiếp theo sau đó, chúng tôi sử dụng thư viện VNCoreNLP [17] để tách đầu vào thành các từ vựng bởi vì cấu tạo của một từ vựng trong tiếng Việt bao gồm một hoặc nhiều âm tiết kết hợp với nhau.

+ **Bước 6:** Bước cuối cùng là chuyển tất cả các từ vựng trong chuỗi đầu vào thành chữ thường.

5. Kết quả thí nghiệm

5.1. Mô hình so sánh

Để đánh giá hiệu quả của mô hình đề xuất sử dụng kiến trúc BERT, trong bài báo này, chúng tôi cũng nghiên cứu và cài đặt lại các phương pháp máy học khác như Support Vector Machine, Naive Bayes, Random Forrest, Decision Tree, K Nearest Neighbor hay Neural Network kết hợp với các đặc trưng thủ công được rút trích. Ngoài ra, chúng tôi cũng cài đặt so sánh các phương pháp học sâu như mạng hồi quy Long short-term Memory, mạng tích chập Convolution Neural Network trên bộ dữ liệu đã thu thập nhằm đánh giá tổng quan hiệu quả so sánh thực nghiệm. Chi tiết thông số các mô hình so sánh được chúng tôi trình bày như sau:

- Mô hình Support Vector Machine (SVM): Chúng tôi sử dụng mô hình Linear SVM với thông số C=0,1.

- Mô hình Naive Bayes (NB): Bởi vì các đặc trưng rút trích của chúng tôi sau khi biểu diễn sẽ trở thành các vec-tơ rời rạc, do đó chúng tôi sử dụng mô hình Naive Bayes đa thức.

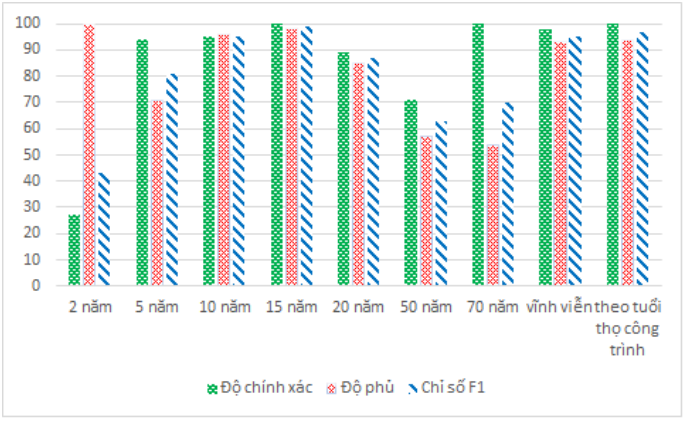
- Mô hình Decision Tree (DT): Chúng tôi sử dụng thuật toán Decision Tree với các tham số mặc định đề xuất.

- Mô hình K Nearest Neighbor (KNN): Chúng tôi sử dụng 3 neighbor, độ đo Euclidean và trọng lượng đồng nhất.

- Mô hình Neural Network (NN): Một lớp ẩn duy nhất với 100 node, sử dụng hàm kích hoạt ReLu, hàm tối ưu hóa Adam, $\alpha = 0{,}001$ và tối đa 200 lần lặp.

- Mô hình mạng tích chập CNN: Kiến trúc CNN được trình bày bởi Kim [8] đã thể hiện tính hiệu quả trên các bộ dữ liệu khác nhau trong các bài toán phân loại văn bản. Chúng tôi cài đặt lại các thông số như đề xuất của tác giả.

- Mô hình mạng hồi quy LSTM: Tương tự như mô hình CNN, chúng tôi cài đặt mô hình mạng hồi quy LSTM [18].



Hình 2. *Kết quả các độ đo của từng nhãn lưu trữ của mô hình BERT trên tập kiểm tra*

Đối với các mô hình máy học truyền thống, chúng tôi sẽ tiến hành rút trích các đặc trưng n-gram (2,3,4 grams) kết hợp với nhãn từ loại và các từ vựng (danh từ, động từ và tính từ) trong tên các hồ sơ lưu trữ. Sau khi rút trích đặc trưng xong, chúng tôi sẽ sử dụng kỹ thuật TF-IDF để biểu diễn các đặc trưng thành các vec-tơ số đại diện cho từng tên hồ sơ và đưa vào các mô hình huấn luyện.

5.2. Chi tiết cài đặt

Đối với mô hình BERT, chúng tôi sử dụng PhoBERT [16] với kích thước lớp ẩn là 768 chiều và tổng số lớp biến đổi (tranformer layer) là 12. Giá trị tốc độ học của mô hình được thực nghiệm theo tập giá trị 2e-5, 3e-5, 4e-5 và lựa chọn giá trị tốt nhất là 2e-5. Giá trị batch size được gán là 16. Đối với mô hình học sâu CNN thì chúng tôi sử dụng 3 bộ lọc tích chập khác nhau với kích thước của kernel là 2, 3, 4 và mỗi bộ lọc có 128 chiều với hàm kích hoạt ReLU. Còn đối các mô hình LSTM thì số units có giá trị là 256. Cả hai mô hình CNN và LSTM đều sử dụng một bộ nhúng từ word2vec[1] đã huấn luyện trên tập dữ liệu các bài báo tin tức với số chiều của mỗi véc-tơ là 300 chiều. Đối với các mô hình máy học truyền thống, chúng tôi sử dụng kỹ thuật Grid Search để lựa chọn ra các tham số tốt nhất trên tập phát triển của chúng tôi. Để có kết quả tổng quan, mỗi thí nghiệm trong bài báo của chúng tôi được thực nghiệm lặp lại 5 lần với các giá trị số ngẫu nhiên khác nhau.

Bảng 2. *Kết quả thí nghiệm các phương pháp máy học, học sâu so với mô hình BERT*

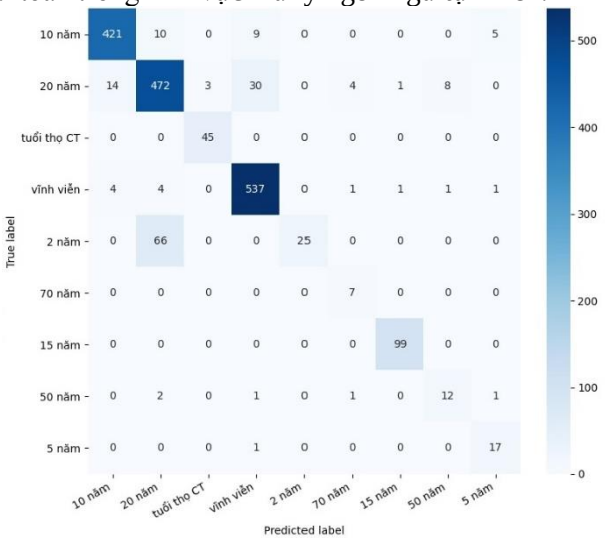
Phương pháp	Độ chính xác	Độ phủ	Chỉ số F1
SVM	89,18	90,46	89,82
NB	84,60	85,19	83,48
KNN	88,24	87,97	88,01
DT	86,65	86,25	86,36
RF	90,04	89,63	89,77
NN	88,57	89,96	89,26
CNN	91,05	90,02	90,30
LSTM	91,09	89,96	89,15
BERT	**93,10**	**90,68**	**91,49**

5.3. Kết quả thực nghiệm

Bảng 2 trình bày kết quả thực nghiệm các mô hình trên tập kiểm tra theo các độ đo như: độ chính xác, độ phủ và chỉ số F1. Nhìn vào bảng 2, chúng ta dễ dàng thấy rằng, đối với các phương pháp máy học truyền thống thì mô hình SVM đạt kết quả tốt nhất so với các phương pháp còn lại với độ đo F1. Kết quả của phương pháp SVM cao hơn các phương pháp còn lại khoảng từ +0,56% đến +6,34%. Điều đó chứng tỏ rằng phương pháp SVM vẫn là phương pháp được sử dụng hiệu

[1] https://github.com/sonvx/word2vecVN

quả cho các bài toán phân loại. Tiếp theo sau đó là phương pháp mạng nhân tạo kết hợp dựa trên các đặc trưng thủ công đạt kết quả với độ đo F1 là 89,26%. Tiếp theo chúng ta sẽ so sánh giữa hai phương pháp học sâu là mạng tích chập CNN và mạng hồi quy LSTM thì chúng ta dễ dàng nhận thấy rằng phương pháp CNN đạt hiệu quả tốt hơn phương pháp LSTM là +1,15% về độ đo F1. Còn so sánh với mô hình máy học SVM, thì mô hình CNN cao hơn phương pháp SVM +0,48%. Điều này chứng tỏ rằng các phương pháp học sâu cho hiệu suất tốt hơn các phương pháp máy học truyền thống trong bài toán phân loại tên hồ sơ theo thời gian lưu trữ. Tuy nhiên, kết quả cao nhất trong thực nghiệm của chúng tôi là phương pháp dựa trên mô hình BERT, kết quả mô hình này đạt độ chính xác là 93,10%, độ phủ là 90,68% và chỉ số F1-score là 91,49%. Mô hình này cao hơn phương pháp máy học truyền thống tốt nhất SVM về độ đo F1 là +1,67% và phương pháp học sâu CNN là +1,19%. Điều này chứng minh rằng BERT hiện tại đang là một mô hình hiệu quả đối với các bài toán trong lĩnh vực Xử lý ngôn ngữ tự nhiên.



Hình 3. *Ma trận nhầm lẫn giữa các nhãn lưu trữ của mô hình BERT*

Hình 2 mô tả kết quả chi tiết các độ đo của từng nhãn lưu trữ trên tập kiểm tra. Chúng ta có thể thấy rằng, các nhãn lưu trữ có kết quả F1 thấp lần lượt là nhãn "2 năm", "50 năm" và "70 năm". Nếu xét về số lượng dữ liệu cho mỗi nhãn trong tập huấn luyện thì các nhãn "50 năm" và "70 năm" có số lượng mẫu huấn luyện thấp nhất trong toàn bộ dữ liệu, tuy nhiên đối với nhãn "2 năm" có số lượng dữ liệu tương đối nhưng kết quả lại thấp nhất trong tất cả các nhãn. Để trả lời câu hỏi này, chúng tôi kiểm tra sự phân loại của mô hình thông qua ma trận nhầm lẫn. Nhìn vào Hình 3, chúng ta có thể thấy rằng, nhãn "2 năm" bị dự đoán hầu hết thành nhãn "20 năm" với 66 mẫu dữ liệu, để trả lời câu hỏi này, chúng tôi tiến hành phân tích lại dữ liệu huấn luyện đã được gán nhãn bởi các chuyên gia lưu trữ. Chúng tôi nhận ra được vấn đề như sau: (1) Dữ liệu chưa có sự đồng nhất cao do chúng tôi thu thập dữ liệu thực tế từ nhiều UBND khác nhau cho nên các chuyên gia gán nhãn cho hồ sơ chưa có đồng thuận cao, ví dụ như nhãn hồ sơ "chứng thực chữ ký" các chuyên gia có lúc gán nhãn "2 năm", có lúc gán nhãn "20 năm", hồ sơ "hợp đồng chuyển nhượng quyền sử dụng đất" các chuyên gia khi thì gán nhãn "70 năm", khi thì gán "Vĩnh viễn", v.v. Cho nên, khi huấn luyện mô hình cho kết quả phân lớp giữa cặp nhãn "2 năm" và nhãn "20 năm" cũng như nhãn "70 năm" và nhãn "Vĩnh viễn" thường tỷ lệ cao dự đoán sai lệch với nhau. Do đó, khi đưa vào thực tế, chúng ta nên kiểm tra lại các dữ liệu gán nhãn bởi các chuyên gia và đánh giá độ đồng thuận, sau đó xây dựng mô hình và áp dụng cho các cơ quan. Từ đó, kết quả lưu trữ sẽ đồng nhất giữa các cơ quan quản lý văn thư - lưu trữ.

6. Kết luận và hướng phát triển

Trong bài báo này, chúng tôi đã nghiên cứu các giải pháp tự động phân loại tên hồ sơ bảo quản sử dụng các phương pháp máy học nhằm hỗ trợ cán bộ, công chức làm việc tại các UBND cấp xã góp phần vào ứng dụng công nghệ thông tin trong công tác văn thư, lưu trữ. Hiện nay, nhu cầu về việc tra cứu và gán nhãn thời hạn bảo quản cho số lượng lớn hồ sơ tại các UBND cấp xã rất cần thiết. Do đó, việc sử dụng các mô hình máy học để phân loại tự động tên hồ sơ theo thời hạn bảo quản giúp nâng cao ý thức bảo vệ hồ sơ của cán bộ, công chức. Mặt khác, còn hỗ trợ cán bộ, công chức trong việc đưa ra quyết định tiêu hủy hồ sơ hết thời hạn bảo quản một cách chính xác. Kết quả thực nghiệm minh chứng mô hình BERT cho kết quả phân loại hiệu quả hơn so với các mô hình khác với độ chính xác là 93,10%, độ phủ là 90,68% và chỉ số F1 là 91,49%. Điều này cho thấy sự hiệu quả vượt trội của kiến trúc BERT đối với các bài toán phân loại hồ sơ theo thời hạn bảo quản. Các kết quả nghiên cứu trong đề tài này cho thấy các mô hình máy học có thể dễ dàng áp dụng vào các bài toán thực tế trong mô hình quản lý.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] N. V. Ket, "Clerical - archive 4.0": premise, scientific - legal basis and basic features," *Proceedings of scientific seminars: Management and confidentiality of electronic documents in the context of the industrial revolution 4.0: Current situation - Solutions*, HCM City National University Publisher, 2018, pp. 41-52.

[2] H. Q. Cuong, "Identify documents archived during the operation of the commune-level government in Ho Chi Minh City," Master thesis, Ho Chi Minh City University of Science and Humanities, 2017.

[3] N. T. T. Huong and D. M. Trung, "Applying the random forest classification algorithm to develop land cover map of Dak Lak based on 8-olive landsat satellite image," *Journal of Agriculture and Rural Development*, vol. 13, pp. 122-129, 2018.

[4] T. C. De and P. N. Khang, "Text classification with Support Vector Machine and Decision Tree," *Can Tho University Journal of Science*, vol. 21a, pp. 52–63, 2012.

[5] D. T. Thanh, N. Thai-Nghe, and T. Thanh, "Solutions to classify scientific articles by machine learning," *Can Tho University Journal of Science*, vol. 55, pp. 29-37, 2019.

[6] T. N. T. Sau, D. V. Thin, and N. L. T Nguyen, "Classification of file names in Vietnamese according to the preservation period," *The conference on Information Technology and Its Applications*, 2019, pp. 198-206.

[7] S. Xu, "Bayesian naıve bayes classifiers to text classification," *Journal of Information Science*, vol. 44, no. 1, pp. 48-59, 2018.

[8] Y. Kim, "Convolutional neural networks for sentence classification," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1746-1751.

[9] K. Kowsari, D. E. Brown, M. Heidarysafa, K. J. Meimandi, M. S. Gerber, and L. E. Barnes, "Hdltex: Hierarchical deep learning for text classification," *Conference on machine learning and applications (ICMLA)*, 2017, pp. 364-371.

[10] K. Kowsari, M. Heidarysafa, D. E. Brown, K. J. Meimandi, and L. E. Barnes, "Rmdl: Random multimodel deep learning for classification," *International Conference on Information System and Data Mining*, 2018, pp. 19-28.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", arXiv preprint arXiv:1810.04805, 2018.

[12] P. T. Ha and N. Q. Chi, "Automatic classification for vietnamese news," *Advances in Computer Science: an International Journal*, vol. 4, no. 4, pp. 126-132, 2015.

[13] N. T. Hai, N. H. Nghia, T. D. Le, and V. T. Nguyen, "A hybrid feature selection method for vietnamese text classification," *Conference on Knowledge and Systems Engineering (KSE), IEEE*, 2015, pp. 91-96.

[14] P. Le-Hong and A.-C. Le, "A comparative study of neural network models for sentence classification," *5th NAFOSTED Conference on Information and Computer Science (NICS), IEEE*, 2018, pp. 360-365.

[15] K. D. T. Nguyen, A. P. Viet, and T. H. Hoang, "Vietnamese document classification using

hierarchical attention networks," *Frontiers in Intelligent Computing: Theory and Applications, Springer*, 2020, pp. 120-130.

[16] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: Pre-trained language models for Vietnamese", *arXiv preprint*, vol. arXiv:2003.00744, 2020.

[17] T. Vu, D. Q. Nguyen, D. Q. Nguyen, M. Dras, and M. Johnson, "VnCoreNLP: A Vietnamese natural language processing toolkit," *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, Jun. 2018, pp. 56-60.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735-1780, 1997.