NGHIÊN CỨU MÔ HÌNH PHÂN LOẠI VĂN BẢN TIẾNG VIỆT ỨNG DỤNG CHO PHÂN LOẠI BÁO ĐIỆN TỬ

RESEARCH MODEL OF VIETNAMESE TEXT CLASSIFICAITION FOR CLASSIFING ELECTRONIC NEWS

SVTH: Võ Hoàng An[1]

[1]Lớp: 19146CL2A, MSSV: 19146147, Khoa: Đào tạo Chất lượng cao

[1]Trường Đại học Sư phạm Kỹ Thuật TP.HCM, Việt Nam

TÓM TẮT

Xử lý ngôn ngữ tự nhiên, một lĩnh vực quan trọng của Trí tuệ nhận tạo và Học máy cũng như được ứng dụng rộng rãi trong đời sống con người. Trong đó bài toán phân loại văn bản là nhu cầu cấp thiết hiện nay trong thời kỳ mà tất cả dữ liệu văn bản đang được số hoá. Bài báo trình bày cụ thể quá trình xử lý các dữ liệu văn bản và xây dựng mô hình phân loại văn bản tiếng Việt, đồng thời ứng dụng vào thực tế để phân loại chủ đề các bài báo điện tử. Mô hình sử dụng tập dữ liệu với hơn 4000 văn bản trên mỗi chủ đề, giải thích phương pháp tách từ, vector hoá văn bản bằng từ điển và xây dựng mô hình Neural Network để phân loại văn bản.

Từ khoá: Phân loại văn bản, Từ điển, Neural Network, Xử lý ngôn ngữ, Tách từ

1. Đặt vấn đề

Xử lý ngôn ngữ là một nhánh nghiên cứu quan trọng của trí tuệ nhân tạo được phát triển nhằm xây dựng các chương trình máy tính có khả năng phân tích, xử lý và hiểu ngôn ngữ con người. Mục tiêu của lĩnh vực này là giúp máy tính hiểu và thực hiện hiệu quả những nhiệm vụ liên quan đến ngôn ngữ của con người. Rất nhiều ứng dụng của xử lý ngôn ngữ đã và đang được sử dụng trong đời sống con người ngày nay, một số ứng dụng phổ biến được chỉ ra như: Các hệ thống máy dịch ngôn ngữ (Google translation); Hệ thống tóm tắt văn bản; Hệ thống sửa lỗi chính tả, lỗi cú pháp; Hệ thống chatbot; Hệ thống tổng đài tự động;…

Trong các ứng dụng của xử lý ngôn ngữ thì bài toán phân loại là một trong những bài toán quan trọng. Bài toán phân loại văn bản là một bài toán học có giám sát (supervied learning) trong học máy được thực hiện bằng cách huấn luyện máy tính học theo tập dữ liệu các văn bản đã được gán nhãn, từ đó có thể phân loại các văn bản mới và gán nhãn cho các văn bản đó dựa trên mức độ tương tự của văn bản mới so với các văn bản đã được gán nhãn trong tập dữ liệu. Các ứng dụng của phân loại văn bản rất đa dạng như: Phân loại mail spam;

Phân tích cảm xúc các bình luận, đánh giá của người dùng; Phân loại tin tức theo chủ đề;…

Trên thế giới đã có nhiều công trình nghiên cứu đạt những kết quả khả quan, tuy nhiên các nghiên cứu này chủ yếu vào văn bản tiếng Anh do ngôn ngữ này tương đối thuận lợi khi xử lý. Hiện nay, việc nghiên cứu và ứng dụng cho các văn bản tiếng Việt thì còn hạn chế do gặp nhiều khó khăn trong vấn đề tách từ và câu.

Trong bài báo này, tác giả sẽ trình bày về bài toán phân loại văn bản tiếng Việt và xây dựng mô hình Neural Network ứng dụng trong phân loại chủ đề báo điện tử. Do sự gia tăng của số lượng tin tức hằng ngày trên mạng Internet và nhu cầu tìm kiếm văn bản của người dùng cũng tăng nên việc phân loại các bài báo điện tử là điều cần thiết, nó sẽ giúp chúng ta quản lý lượng tin tức lớn và tìm kiếm thông tin một cách nhanh chóng hơn.

2. Phương pháp và mô hình

Để giải quyết một bài toán phân loại thông thường sẽ trãi qua các bước cơ bản:

- Chuẩn bị dữ liệu đã gán nhãn
- Tiền xử lí dữ liệu
- Trích xuất đặc trưng
- Xây dựng mô hình
- Tinh chỉnh mô hình

2.1. Chuẩn bị dữ liệu

Trong mô hình này tác giả sử dụng tập dữ liệu là các tập tin dạng text có sẵn trên Internet trong đó dữ liệu là các đoạn văn bản được trích xuất từ các bài báo điện tử và đã được gán nhãn chủ đề. Tập dữ liệu được sử dụng bao gồm 10 chủ đề và khoảng 4000 tập tin trên mỗi chủ đề với các lĩnh vực: Chính trị Xã hội, Đời sống, Khoa học, Kinh doanh, Pháp Luật, Sức Khoẻ, Thế giới, Thể thao, Văn hoá và Công nghệ.

Bảng 1. Tập dữ liệu 10 chủ đề

Số thứ tự	Tên chủ đề	Số lượng
0	Chính trị Xã hội	4,472
1	Công nghệ	4,173
2	Đời sống	4,365
3	Khoa học	3,916
4	Kinh doanh	4,327
5	Pháp luật	4,345
6	Sức khoẻ	4,483

7	Thế giới	4,307
8	Thể thao	4,287
9	Văn hoá	4,300
Tổng		42,975

2.2. Tiền xử lí dữ liệu

Ở bước tiền xử lí dữ liệu, đối với văn bản thường sẽ thực hiện các công việc: Xoá ký tự, chuẩn hoá dấu, tách từ, loại bỏ từ dừng… Trong đó yếu tố quyết định đến độ chính xác của mô hình là phần tách từ. Khác với tiếng Anh trong tiếng Việt các từ không tách nhau bởi khoảng trắng mà có thể 2 hoặc 3 từ liên tiếp mới tạo thành một tiếng có nghĩa và tuỳ thuộc vào ngữ cảnh của câu thì các từ sẽ có nghĩa khác nhau nên việc tách từ tiếng Việt sẽ khó khăn hơn nhiều. Bài toán tách từ gồm có 3 phương pháp tiếp cận: Dựa vào từ điển, dựa vào thống kê và kết hợp cả hai phương pháp. Hiện nay có các phương pháp thực hiện phổ biến như: So khớp cực đại, mô hình Markov ẩn, so khớp dài nhất, độ hỗn loạn cực đại,…

Trong mô hình này, sẽ thực hiện tách từ bằng cách dùng một bộ từ điển tiếng Việt (có sẵn) và sử dụng phương pháp Longest Matching (So khớp dài nhất) để tách các từ có nghĩa.

Phương pháp được thực hiện như sau: với một chuỗi kí tự [A1 A2 A3…An] bằng cách xét từ dài nhất trong chuỗi có thuộc từ điển hay không, nếu có xem nó là một từ có nghĩa, nếu không giảm chiều dài chuỗi kí tự đi 1 và tiếp tục xét cho đến khi hết từ trong chuỗi.

Các bước tiền xử lí khác tương đối đơn giản nên sẽ không đề cập đến trong bài báo này.

Thuật toán tách từ

```
(1) text = [A1 A2 A3 … An]
(2) start = 0; word = []; dict = (từ điển)
(3) While end > start:
  (3.1) end =len(text);
  (3.2) if text[start:end] không thuộc dict then end ← end - 1
  (3.3) if text[start:end] thuộc dict then words ← text[start:end];
        start = end + 1
  (3.4) if start >= end: break
(4) return words
```

Kết quả áp dụng thuật toán tách từ có thể tách được đa số các từ có nghĩa.

Bảng 2. Kết quả thuật toán tách từ

Câu trước khi tách	Câu sau khi tách
TP HCM Phố cà phê Tân Bình nóng theo từng đường bóng đá	Phố cà_phê Tân Bình nóng theo từng đường bóng_đá
Nhiều tỉnh lộ, quốc lộ từ Bắc vào Nam hư hỏng nghiêm trọng do xe chở quá tải	Nhiều tỉnh quốc_lộ từ Bắc vào Nam hư_hỏng nghiêm_trọng do xe chở quá_tải
Trong ba ngày đầu, các biểu hiện đau họng và nuốt vướng của người bệnh giảm	Trong ba_ngày các biểu_hiện đau họng và nuốt vướng của người_bệnh giảm

Trong quá trình xử lý dữ liệu, mỗi tập tin sẽ lấy 2 đoạn văn bản nên tổng dữ liệu đưa vào huấn luyện sẽ gấp đôi dữ liệu ban đầu, khoảng hơn 80.000 đoạn văn bản thuộc 10 chủ đề.

2.3. Trích xuất đặc trưng

Sau khi đã làm sạch các dữ liệu, cần trích xuất các đặc trưng của dữ liệu bằng cách chuyển các đoạn văn bản thành các vector số mà máy tính có thể hiểu và thực hiện huấn luyện. Có rất nhiều phương pháp để thực hiện trích trọn đặc trưng, một số phương pháp chuyển đổi văn bản sang vector dữ liệu số phổ biến như:

- Frequency-based Embedding: TF-IDF, Count vector, Co-occurrence Matrix, Glove,…
- Prediction-based Embedding: Word2vec: gồm 2 model chính đó là Continuous Bag of Words Model và Skip-Gram Model, …
- Các kỹ thuật sử dụng các mạng học sâu như Transformer, BERT, …

Việc trích chọn đặc trưng trong mô hình này được thực hiện theo cách sử dụng bộ từ điển chứa các từ được đánh trọng số, với mỗi câu cần vector hoá sẽ chuyển các từ trong câu thành các trọng số tương ứng trong từ điển.

Ví dụ: Ta có bộ từ điển {‘sinh’: 0, ‘viên’: 1, ‘sư_phạm’: 2, ‘kỹ_thuật’: 3, ‘tôi’: 4, ‘học’: 5}. Xét một câu [tôi học sư_phạm kỹ_thuật], vậy vector tương ứng của câu là [4 5 2 3].

Để thực hiện, tác giả đã tạo ra một bộ từ điển mới dựa trên các từ có trong tập dữ liệu đã được xử lý. Kết quả thu được là bộ từ điển mới khoảng 26.719 từ được đánh trọng số theo thứ tự của từ trong từ điển. Đối với tên chủ đề của các bài báo sẽ được one-hot encoding để mỗi chủ đề tương ứng với một vector duy nhất.

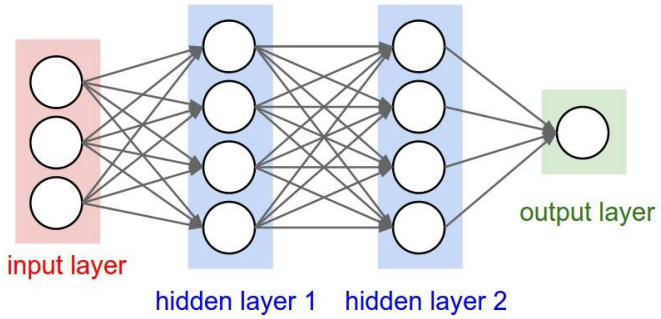
2.4. Xây dựng mô hình

Tác giả chọn cách xây dựng mô hình bằng mạng Neural Network để huấn luyện dữ liệu. Neural Network hay còn gọi là mạng Neural nhân tạo, sử

dụng các mô hình toán học phức tạp để xử lý thông tin. Chúng dựa trên mô hình hoạt động của các tế bào thần kinh và khớp thần kinh trong não của con người.

Một mạng Neural Network có 3 thành phần bao gồm:

- Lớp đầu vào đại diện cho các dữ liệu đầu vào.
- Lớp ẩn đại diện cho các nút trung gian phân chia không gian đầu vào thành các vùng có ranh giới.
- Lớp đầu ra đại diện cho đầu ra của mạng neural.



Hình 1. Mạng Neural Network

Với các dữ liệu đã được vector hoá, chia dữ liệu thành hai phần train và test với tỉ lệ là 75% dữ liệu cho huấn luyện và 15% dữ liệu cho kiểm thử.

Sau đó xây dựng mô hình với đầu vào sẽ qua một lớp Embedding, sau đó qua một lớp GlobalAveragePooling1D, Dropout và đầu ra là 10 ngõ ra tương ứng với 10 chủ đề.

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_1 (Embedding)     (None, 3987, 300)         8015700

 global_average_pooling1d_1   (None, 300)              0
 (GlobalAveragePooling1D)

 dropout_1 (Dropout)         (None, 300)               0

 dense_1 (Dense)             (None, 10)                3010

=================================================================
Total params: 8,018,710
Trainable params: 8,018,710
Non-trainable params: 0
_________________________________________________________________
```

Hình 2. Các lớp mô hình Neural

Trong đó:

- Embedding layer sẽ học cách mã hoá mỗi số tự nhiên thành một embedding vector có 300 chiều.
- GlobalAveragePooling1D sẽ đảm bảo rằng đầu ra của các vector là như nhau.

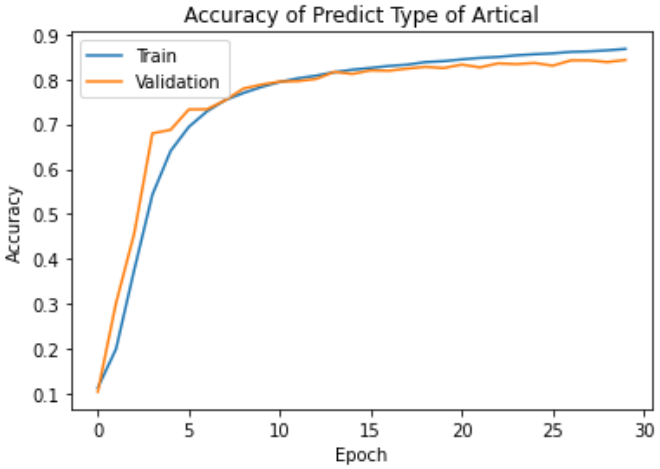
- Dropout để tránh bị overfitting.

Biên dịch mô hình với 3 thành phần là Loss: Categorical_crossentropy; Optimizer: Adam; Metrics: Accuracy.

Tiến hành thực hiện huấn luyện mô hình trên tập huấn luyện và kiểm tra trên tập kiểm thử. Sau khoảng 30 lần huấn luyện thì độ chính xác trên tập huấn luyện là 86.7% và độ chính xác kiểm tra trên tập kiểm thử là 84.2%.

3. Đánh giá mô hình

Mô hình đã thực hiện kiểm tra 15% dữ liệu tương đương khoảng 12.570 mẫu dữ liệu, quá trình huấn luyện không xảy ra hiện tượng overfitting và sau 20 lần thì độ chính xác kiểm tra có xu hướng bão hoà.



Hình 3. Đồ thị độ chính xác quá trình huấn luyện mô hình

Để tổng quát hơn về kết quả độ chính xác trên tập kiểm thử, tác giả sử dụng Confusion matrix để trực quan hoá các kết quả.

Bảng 3. Confusion matrix đánh giá hiệu quả mô hình

Chủ đề	CH XH	Công nghệ	Đời sống	Khoa học	Kinh doanh	Pháp luật	Sức khoẻ	Thế giới	Thể thao	Văn hoá
CH XH	**962 7.65%**	39 0.31%	53 0.42%	27 0.21%	59 0.47%	104 0.83%	15 0.19%	24 0.19%	13 0.1%	25 0.2%
Công nghệ	13 0.1%	**1172 9.32%**	9 0.07%	14 0.11%	24 0.19%	7 0.06%	2 0.02%	14 0.11%	4 0.03%	13 0.1%
Đời sống	34 0.27%	35 0.28%	**920 7.32%**	62 0.49%	17 0.14%	23 0.18%	42 0.33%	17 0.14%	19 0.15%	110 0.88%
Khoa học	13 0.1%	53 0.42%	75 0.6%	**900 7.16%**	15 0.12%	3 0.02%	57 0.45%	31 0.25%	11 0.09%	14 0.11%
Kinh doanh	34 0.27%	63 0.5%	14 0.11%	8 0.06%	**1110 8.83%**	21 0.17%	3 0.02%	39 0.31%	8 0.06%	3 0.02%
Pháp luật	47 0.37%	16 0.13%	24 0.19%	1 0.01%	28 0.22%	**1094 8.7%**	5 0.04%	11 0.09%	16 0.13%	9 0.07%
Sức khoẻ	10 0.08%	11 0.09%	45 0.36%	54 0.43%	6 0.05%	6 0.05%	**1107 8.81%**	12 0.1%	5 0.04%	7 0.06%

Thế giới	17 0.14%	25 0.2%	25 0.2%	31 0.25%	29 0.23%	13 0.1%	14 0.11%	**1024** **8.15%**	8 0.06%	19 0.15%
Thể thao	5 0.04%	9 0.07%	14 0.11%	2 0.02%	0 0.0%	9 0.07%	4 0.03%	10 0.08%	**1155** **9.19%**	12 0.1%
Văn hoá	8 0.06%	14 0.11%	65 0.52%	14 0.11%	1 0.01%	6 0.05%	6 0.05%	17 0.14%	5 0.04%	**1148** **9.13%**

Có tổng cộng 10.592 đoạn văn bản được dự đoán đúng trên tổng 12.570 đoạn văn bản kiểm thử. Từ Confusion matrix có thể suy ra các thông số Precision, Recall và F1-Score.

Bảng 4. Thông số Precision, Recall và F1-scrore đánh giá mô hình

Tên chủ đề	Precision	Recall	F1-score
CH-XH	84%	73%	78%
Công nghệ	82%	92%	87%
Đời sống	74%	72%	73%
Khoa học	81%	77%	79%
Kinh doanh	86%	85%	86%
Pháp luật	85%	87%	86%
Sức khoẻ	88%	88%	88%
Thế giới	85%	85%	85%
Thể thao	93%	95%	94%
Văn hoá	84%	89%	87%
Tổng	84%		

Qua các thông số chủ đề được dự đoán đúng nhiều nhất là chủ đề Thể thao với chỉ số F1-Score là 94%. Chủ đề dự đoán thấp nhất là chủ đề Đời sống với chỉ số F1-Score là 73%.

4. Kết luận

Trong xử lý văn bản đầu vào phần quan trọng nhất là phần tách từ tiếng Việt, do tiếng Việt sẽ gồm các từ hoặc các liên từ mới hình thành một tiếng có nghĩa, vì vậy việc xử lý phần tách từ này khá khó khăn. Việc sử dụng phương pháp tách từ bằng từ điển khá dễ thực hiện và tách các từ có nghĩa cũng tương đối ổn định. Tuy nhiên nó cũng có nhược điểm là bộ từ điển của chưa đầy đủ và có những từ không hoàn toàn chính xác dẫn đến việc không tách được những từ mới hoặc những từ không có trong từ điển.

Về mô hình Neural với ít lớp ẩn nhưng đã cho ra kết quả khả quan với tỉ lệ dự đoán trung bình là 84%.

Tài liệu tham khảo

[1]. Chris Albon (2018). *Machine Learning with Python Cookbook*, Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472, United States of America

[2]. Lê Vĩnh Phú, Diệp Minh Hoàng (2014). *Phân loại tin tức tiếng việt sử dụng các phương pháp học máy*, Luận văn tốt nghiệp Đại học, Trường Đại học Bách khoa.
[3]. Vũ Hữu Tiệp (2016 - 2020). *Machine learning cơ bản*
[4]. Nguyễn Trần Thiên Thanh, Trần Khải Hoàng (2005). *Tìm hiểu các phương pháp tiếp cận bài toán phân loại văn bản và xây dựng phân mềm phân loại tin tức báo điện tử*, Luận văn Cử nhân tin học, Trường Đại học Khoa học Tự nhiên.
[2]. Lưu Tuấn Anh. *Natural Language Processing*. <http://viet.jnlp.org/>, xem ngày 17/6/2022

Nguồn dữ liệu:

[1]. Cong Duy Vu Hoang (2019). <https://github.com/duyvuleo/VNTC>, xem ngày 5/6/2022
[2]. Van-Duyet Le (2015). < https://github.com/stopwords/vietnamese-stopwords >, xem ngày 7/6/2022
[3]. Vũ Anh (2018), <https://github.com/undertheseanlp/dictionary>, xem ngày 16/6/2022