DSKT Lab: Báo cáo kết quả bài toán phân loại

Ngô Tùng Lâm

1 Tổng quan bài toán

Nhiệm vụ của bài toán là phân loại các đoạn tin điện từ theo 10 chủ đề: Chính trị xã hội, Khoa học, Kinh doanh, Pháp luật, Sức khỏe, Thế giới, Thể thao, Vi tính, Văn hóa, Đời sống.
Tập dữ liệu được chia thành tập dữ liệu huấn luyện và tập dữ liệu kiểm tra. Ở cả 2 tập huấn luyện đều có 5000 đoạn tin điện tử, với mỗi chủ đề đều gồm 500 mẫu văn bản.

2 Hướng tiếp cận và kết quả

Em sẽ tiếp cận bài toán theo các bước sau: Preprocessing, Feature Extraction, Training và Evaluation. Đây cũng là các bước cơ bản trong việc xử lý các bài toán NLP.
Trong bài toán này, em đã thử hai hướng tiếp cận: sử dụng mô hình xác suất Naive Bayes và sử dụng mô hình LSTM.

2.1 Preprocessing

Bước tiền xử lý dữ liệu được xử lý giống nhau ở cả 2 hướng tiếp cận

2.1.1 Normalization

Trong phần chuẩn hóa, em thực hiện các bước sau:

- Đưa tất cả kí tự về in thường (lowercase)
- Loại bỏ dấu câu, các kí tự đặc biệt Loại bỏ các stopword (các từ không mang ý nghĩa quan trọng trong câu). Ta sẽ sử dụng danh sách vietnamese_stopwords có sẵn

2.1.2 Tokenization

Tokenization là tách văn bản gốc thành các phần nhỏ hơn, ở đây em sẽ tách thành các từ và cụm từ Tiếng Việt. Đối với các văn bản tiếng Việt, ta có thể sử dụng tokenizer có sẵn như pyvi hay underthesea. Trong bài toán này, em sẽ sử dụng tokenizer của pyvi do đem lại hiệu quả cao hơn trong tác vụ tokenization so với underthesea

2.2 Mô hình Naive Bayes

2.2.1 Hướng tiếp cận

Naive Bayes Classifier (NBC) là một thuật toán phân loại dựa trên tính toán xác suất áp dụng định lý Bayes. Thuật toán sử dụng phương pháp Bag of Words (BOW), biểu diễn văn bản thành tập các từ và số lần xuất hiện của chúng. Ở hướng tiếp cận này, ta giả định rằng vị trí các từ xuất hiện trong văn bản không quan trọng trong việc phân loại chúng vào các chủ đề

2.2.2 Kết quả

Mô hình Naive Bayes đạt kết quả khá tốt, macro average precision, recall và F1-score đều đạt 0.89. Phân tích kĩ hơn vào các nhãn, ta có thể đưa ra nhận xét sau:

- Các đoạn tin thuộc chủ để Chính trị Xã hội được mô hình dự đoán đúng ít nhất. Có thể nhận thấy các từ ngữ trong các văn bản Chính trị Xã hội khá đa dạng, có thể nhầm lẫn sang các lớp khác
- Mô hình có thể dự đoán các đoạn tin thuộc chủ đề Thể thao khá tốt, precision đạt 1.0 (các dự đoán positive đều chính xác), tuy nhiên còn bỏ lỡ một vài điểm.
- Các chủ đề khác đều được mô hình dự đoán khá tốt

Label	Precision	Recall	F1-score	Support
Chính trị Xã hội	0.79	0.73	0.75	500
Đời sống	0.86	0.87	0.86	500
Khoa học	0.88	0.84	0.86	500
Kinh doanh	0.83	0.91	0.87	500
Pháp luật	0.86	0.95	0.90	500
Sức khỏe	0.89	0.90	0.89	500
Thế giới	0.90	0.87	0.88	500
Thể thao	1.00	0.95	0.97	500
Vi tính	0.92	0.93	0.93	500
Văn hóa	0.93	0.93	0.93	500
Macro avg	0.89	0.89	0.89	5000

Table 1: Kết quả phân loại sử dụng Naive Bayes

2.3 Mô hình LSTM

2.3.1 Hướng tiếp cận

LSTM là một kiến trúc mạng nơ-ron đặc biệt thuộc họ mạng nơ-ron hồi quy (Recurrent Neural Network - RNN). Điểm khác biệt chính so với RNN truyền thống là LSTM có khả năng học và ghi nhớ thông tin trong khoảng thời gian dài hơn, khắc phục vấn đề "quên lãng" thông tin khi xử lý các chuỗi dài.

Ở hướng tiếp cận này, em sử dụng count-based word embedding method là TF_IDF. TF-IDF là một phương pháp thống kê được sử dụng để đánh giá tầm quan trọng của một từ trong một tài liệu hoặc một tập hợp tài liệu, dựa vào 2 phần: TF - tần suất xuất hiện của một từ trong một tài liệu và IDF - nghịch đảo mức độ phổ biến của một từ trong toàn bộ tập dữ liệu.

Dưới đây là kiến trúc chung của mô hình LSTM được sử dụng:

Layer	Size	Parameters
Input layer	(None, 150)	0
Reshape	(None, 1, 150)	0
LSTM	(None, 256)	416,768
Dense	(None, 512)	131,584
Dropout	(None, 512)	0
Dense	(None, 128)	65,664
Dropout	(None, 128)	0
Dense	(None, 256)	33,024
Dropout	(None, 256)	0
Dense	(None, 64)	16,448
Dropout	(None, 64)	0
Dense (output)	(None, 10)	650

Table 2: Kiến trúc mô hình LSTM layers

2.3.2 Kết quả

Label	Precision	Recall	F1-score	Support
Chính trị xã hội	0.80	0.70	0.74	500
Đời sống	0.86	0.82	0.84	500
Khoa học	0.81	0.86	0.84	500
Kinh doanh	0.85	0.87	0.86	500
Pháp luật	0.89	0.92	0.90	500
Sức khỏe	0.87	0.92	0.90	500
Thế giới	0.91	0.85	0.88	500
Thể thao	0.99	0.97	0.98	500
Văn hóa	0.91	0.93	0.92	500
Vi tính	0.90	0.95	0.93	500
Macro avg	0.88	0.88	0.88	5000

Table 3: Kết quả phân loại sử dụng LSTM

Mô hình LSTM đạt kết quả khá tương đồng với mô hình Naive Bayes, với F1-score đạt 0.88. Mô hình vẫn nhận diện tốt nhất các điểm thuộc nhãn Thể thao, kém nhất các điểm thuộc nhãn Chính trị xã hội. Kết quả nhận diện của 2 mô hình ở các nhãn khác cũng khá tương đồng.

3 Kết luận

Cả hai hướng tiếp cận đều đưa ra được hiệu quả nhận diện khá tốt. Tuy nhiên, ta mới sử dụng các phương pháp word embedding count base, vì vậy mô hình chưa nắm bắt được ngữ nghĩa của câu từ. Thay vì đó, sử dụng các phương pháp predictive word embedding có thể đem lại hiệu quả nhận diện cao hơn, cụ thể với mô hình LSTM.