Đồ án cuối kì
Đề tài: Phân loại tin tức nhạy cảm

Bình Gia Huy, Mai Phúc Minh
Giảng viên hướng dẫn: Lê Đình Duy, Phạm Nguyễn Trường An

Máy học

Trường Đại học Công nghệ Thông tin
Ngày 25 tháng 1 năm 2024

Mục lục

0 Update sau vấn đáp **1**

1 Bài toán **1**

2 Dữ liệu **1**
2.1 Về bộ dữ liệu 1
2.2 Chia dữ liệu 1
2.3 Tiền xử lí dữ liệu 2

3 Đặc trưng **2**
3.1 Xử lí dữ liệu 2
3.1.1 Logistic Regression 2
3.1.2 PhoBert 2

4 Thuật toán **2**
4.1 Logistic Regression 2
4.2 PhoBERT 2

5 Cài đặt, tinh chỉnh tham số **2**
5.1 Logistic Regression 2
5.2 PhoBERT 3

6 Đánh giá kết quả, kết luận **3**

7 Tài liệu tham khảo **4**

0 Update sau vấn đáp

1 Bài toán

- Hiện nay, có nhiều phụ huynh ở Việt Nam đang rèn luyện khả năng đọc, tiếp thu thông tin của con em mình bằng cách cho trẻ đọc báo. Tuy nhiên, không phải bài báo nào cũng chứa các nội dung thân thiện với bạn đọc nhỏ tuổi. Chính vì vậy, bọn em đã quyết định thử sức với bài toán phân loại những tin tức nhạy cảm (chiến tranh, bạo lực,...) để lọc các bài báo không phù hợp với trẻ em.
- Input: Bài báo dạng văn bản (gồm nội dung, tiêu đề, tóm tắt).
- Output: Nhãn "Không nhạy cảm" hoặc "Nhạy cảm".
  Link github của nhóm

2 Dữ liệu

2.1 Về bộ dữ liệu

- Bộ dữ liệu do nhóm tự xây dựng bằng cách crawl các bài báo từ VNExpress, Baomoi.
- VNExpress: 357 bài báo (Nhạy cảm: 52, Không nhạy cảm: 305), Baomoi: 732 bài báo (Nhạy cảm: 80, Không nhạy cảm: 652).

2.2 Chia dữ liệu

- Mô hình Logistic Regression với: chia dữ liệu thành 2 tập train/test với tỉ lệ 80:20.
- Mô hình PhoBert: chia dữ liệu thành 3 tập train/dev/test với tỉ lệ 60:20:20.

2.3 Tiền xử lí dữ liệu

- Chuyển văn bản thành dạng in thường.
- Lọc từ dừng (những từ xuất hiện nhiều, không bổ trợ nhiều cho ý nghĩa của câu).
- Lọc dấu câu.
- Lọc khoản trắng dư.

3 Đặc trưng

3.1 Xử lí dữ liệu

3.1.1 Logistic Regression

1. Mỗi văn bản được tách thành một danh sách từ bằng thư viện underthesea.
2. Thực hiện word embedding bằng Word2Vec của thư viện gensim.
3. Mỗi văn bản là một vector trung bình cộng của các vector của các từ trong văn bản đó theo chiều từ trên xuống dưới.

3.1.2 PhoBert

1. Mỗi văn bản được tách thành một danh sách từ bằng thư viện underthesea.
2. Thay thế dấu cách của mỗi từ trong danh sách thành dấu ngạch dưới và gộp danh sách các từ lại thành một chuỗi.
3. Gộp văn bản và label thành một Dataset theo thư viện của Huggingface.
4. Cho văn bản vào hàm tokenizer để chuyển văn bản thành token mà model có thể hiểu được.

4 Thuật toán

4.1 Logistic Regression

Công thức Logistic Regression:

$$f_{\vec{w},b}(\vec{x}) = \frac{1}{1 + e^{-(\vec{w} \cdot \vec{x} + b)}}$$

Logistic Regression là một mô hình đơn giản nhưng hiệu quả cho việc phân loại văn bản.

4.2 PhoBERT

PhoBERT là mô hình pretrained tiếng Việt dựa trên RoBERTa - một mô hình tối ưu quy trình pretraining của BERT - một model dựa trên kiến trúc transformers.
PhoBERT là một công nghệ state-of-the-art trong Xử lý ngôn ngữ tự nhiên; nó tối ưu và hiệu quả hơn nhiều công nghệ đi trước trên nhiều tác vụ.

5 Cài đặt, tinh chỉnh tham số

5.1 Logistic Regression

Tham số của mô hình Logistic Regression được lựa chọn bằng GridSearchCV của thư viện sklearn.
GridSearchCV gồm:

- Grid Search là một thuật toán vét cạn để chọn ra tổ hợp tham số tốt nhất.

- KFold Cross Validation (ở đây là 5 fold) là một phương pháp resampling để đánh giá model trên một dataset hạn chế.

Tham số sau khi tinh chỉnh:

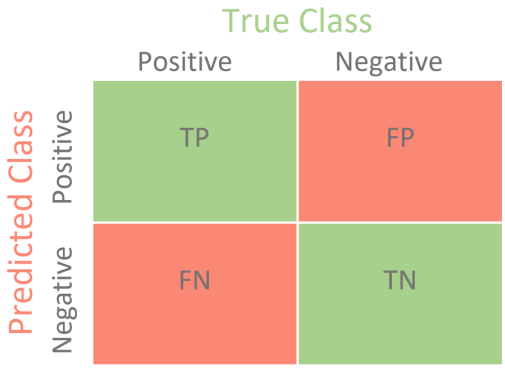
- 'C': 1.1 (tham số regularization).
- 'class_weight': None (mỗi class có weight là 1).
- 'max_iter': 200 (số vòng lặp tối đa để solver hội tụ).
- 'solver': 'liblinear' (thuật toán tối ưu).

5.2 PhoBERT

max_length của mỗi example là 256. Mô hình được train với 10 epochs.

6 Đánh giá kết quả, kết luận

Các metrics được sử dụng là: Precision, Recall, F1



Hình 1: Confusion matrix

Precision là tỷ lệ số lần nhận dạng positive đúng so với số lần được nhận dạng positive:

$$Precision = \frac{TP}{TP + FP}$$

Recall là tỷ lệ số lần nhận dạng positive đúng so với số kết quả positive thật sự:

$$Recall = \frac{TP}{TP + FN}$$

F1 là số dung hòa Precision và Recall:

$$F1 = 2\frac{P * R}{P + R}$$

Kết quả của các mô hình:

		Precision	Recall	F1
Logistic Regression	Không nhạy cảm	0.94	0.98	0.96
	Nhạy cảm	0.76	0.50	0.60
PhoBERT	Không nhạy cảm	0.97	0.94	0.96
	Nhạy cảm	0.65	0.77	0.70

Nhìn chung, PhoBERT có điểm số nhỉnh hơn Logistic Regression. Do có quá ít dữ liệu có nhãn "nhạy cảm" nên cả hai mô hình đều không có điểm F1 thật sự tốt khi nhận dạng nhãn "nhạy cảm".
Mô hình Logistic Regression có khuynh hướng bỏ sót các nhãn "nhạy cảm" với điểm Recall thấp hơn. Ngược lại, PhoBERT có khuynh hướng nhận nhầm các nhãn "nhạy cảm" với điểm Precision thấp. Vì vậy, nhóm đánh giá mô hình PhoBERT phù hợp để sử dụng hơn vì việc lọc nhầm một số bài báo "không nhạy cảm" là không nghiêm trọng bằng việc bỏ sót các bài báo "nhạy cảm".

7 Tài liệu tham khảo

https://github.com/VinAIResearch/PhoBERT
https://scikitlearn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
http://scikitlearn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
https://en.wikipedia.org/wiki/BERT_(language_model)
https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf
https://arxiv.org/abs/2003.00744