BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC CẦN THƠ**

**KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**

**🙞 🕮 🙜**



**ĐỒ ÁN HỌC PHẦN**

**KHAI KHOÁNG DỮ LIỆU**

**ĐỀ TÀI**

**Phân lớp dữ liệu bình luận sản phẩm trên website Tiki**

**Sinh viên thực hiện:**

**Trần Bảo Duy - B1812257**

**Trần Bùi Lý Đức - B1812262**

Cần Thơ, 12/2021

BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC CẦN THƠ**

**KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**

**🙞 🕮 🙜**



**ĐỒ ÁN HỌC PHẦN**

**KHAI KHOÁNG DỮ LIỆU**

**ĐỀ TÀI**

**Phân lớp dữ liệu bình luận sản phẩm trên website Tiki**

**Giảng viên hướng dẫn Sinh viên thực hiện**

**TS. Lưu Tiến Đạo Trần Bảo Duy – B1812257**

**Trần Bùi Lý Đức – B1812262**

Cần Thơ, 12/2021

**NHẬN XÉT CỦA GIẢNG VIÊN**

------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Cần Thơ, ngày tháng 12 năm 2021

(GVHD ký và ghi rõ họ tên)

**LỜI CẢM ƠN**

Để có thể hoàn thành đồ án học phần Khai Khoáng Dữ Liệu một cách hoàn chỉnh, chúng em xin được bày tỏ lòng biết ơn chân thành và sâu sắc đến giảng viên TS. Lưu Tiến Đạo – người đã trực tiếp tận tình hướng dẫn, giúp đỡ chúng em. Trong suốt quá trình thực hiện đề tài, nhờ những sự chỉ bảo và hướng dẫn quý giá đó mà đồ án đã được hoàn thành một cách tốt nhất.

Tuy có nhiều cố gắng trong quá trình thực hiện đồ án học phần, nhưng không thể tránh khỏi những sai sót. Chúng em rất mong nhận được sự đóng góp ý kiến quý báu của Thầy và các bạn để đồ án hoàn thiện tốt hơn.

Cần Thơ, ngày 26 tháng 12 năm 2021

Người viết

Trần Bùi Lý Đức, Trần Bảo Duy

**MỤC LỤC**

Nội dung

[1. Bài toán phân lớp văn bản 5](#_Toc91168954)

[2. Chuẩn bị dữ liệu 7](#_Toc91168955)

[3. Tiền xử lý dữ liệu văn bản 10](#_Toc91168956)

[Xóa HTML code 10](#_Toc91168957)

[Chuẩn hóa Unicode 11](#_Toc91168958)

[Thay thế các từ viết tắt 12](#_Toc91168959)

[Xóa từ dừng (StopWord) 12](#_Toc91168960)

[Thực hiện tách từ tiếng Việt 13](#_Toc91168961)

[Đưa văn bản thành chữ thường 13](#_Toc91168962)

[Chuẩn hóa câu 14](#_Toc91168963)

[4. Vector hoá văn bản ( Word Embedding ) 15](#_Toc91168964)

[5. Xây dựng mô hình phân loại văn bản 16](#_Toc91168965)

[5.1. Xử lý mất cân bằng dữ liệu và gộp nhãn 16](#_Toc91168966)

[5.2. Huấn luyện mô hình bằng các giải thuật máy học với nghi thức đánh giá K-Fold cross validation: 18](#_Toc91168967)

[5.3. Kết quả huấn luyện 19](#_Toc91168968)

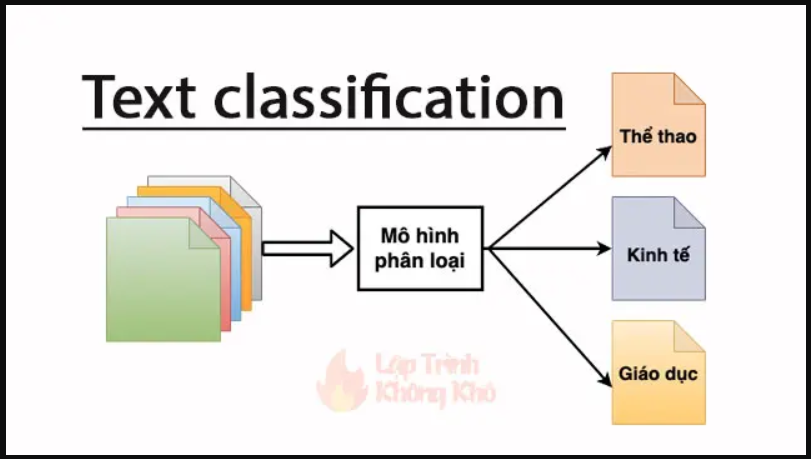
[5.4. Cải thiện mô hình máy học với PhoBERT 19](#_Toc91168969)

[6. Ứng dụng mô hình máy học vừa huấn luyện để xây dựng website dự đoán đánh giá từ người dùng 25](#_Toc91168970)

[7. Tài liệu tham khảo 27](#_Toc91168971)

## Bài toán phân lớp văn bản

Phân lớp văn bản (text classification) là bài toán thuộc nhóm học có giám sát (supervised learning) trong máy học. Ta xây dựng mô hình từ tập dữ liệu được gán nhãn, sau đó dùng mô hình đó để dự đoán nhãn cho đoạn văn bản mới [2].

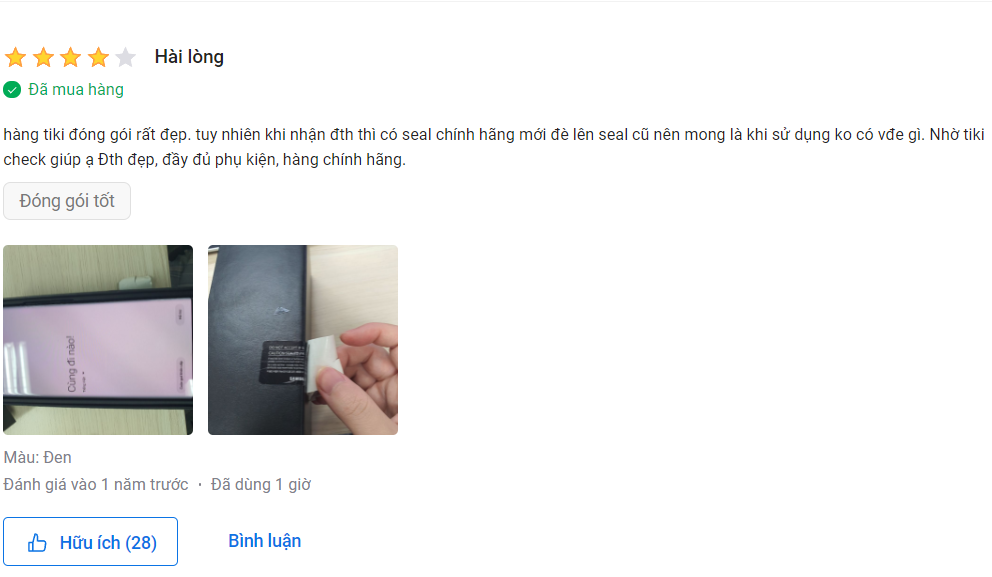


*Mô hình phân loại văn bản*

## Chuẩn bị dữ liệu

Dữ liệu là một trong những yếu tố quan trọng nhất quyết định sự thành công của một mô hình. Muốn mô hình đạt độ hiệu quả cao cần có một tập dữ liệu đủ lớn, có tính tổng quát cao.

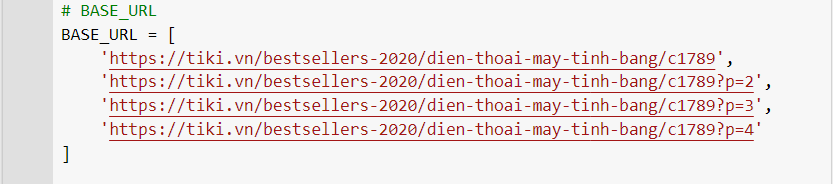
Dữ liệu được thu thập từ những bình luận đánh giá về điện thoại di động của khách hàng trên website tiki.vn, cuối cùng thu được tập dữ liệu gồm có 16115 mẫu, thuộc tính content là nội dung bình luận của khách hàng, nhãn rating là số sao mà khách hàng đã đánh giá cho sản phẩm mà họ mua.



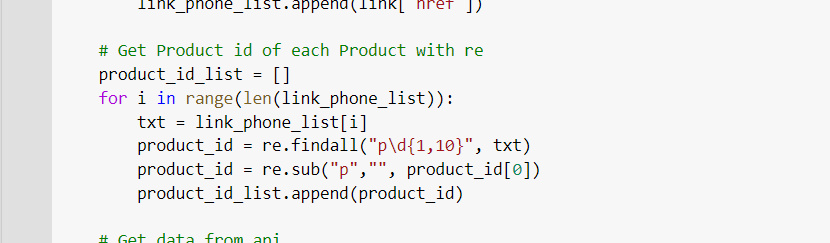
*Ví dụ về đánh giá từ người dùng trên website Tiki*

Việc tải dữ liệu trên được thực hiện bằng đoạn code Python như sau:













Dữ liệu được lưu vào file json với 1 cột nhãn và 1 cột thuộc tính

* Content : Nội dung bình luận
* Rating: Số lượng sao tương ứng với nội dung bình luận

## Tiền xử lý dữ liệu văn bản

Sau khi đã có được dữ liệu thì bước tiếp theo đó là tiền xử lý dữ liệu. Đây là bước quan trọng nhất, có ảnh hưởng lớn nhất với chất lượng của mô hình máy học mà chúng ta xây dựng. Bước này cần được thực hiện trước khi trích xuất đặc trưng.

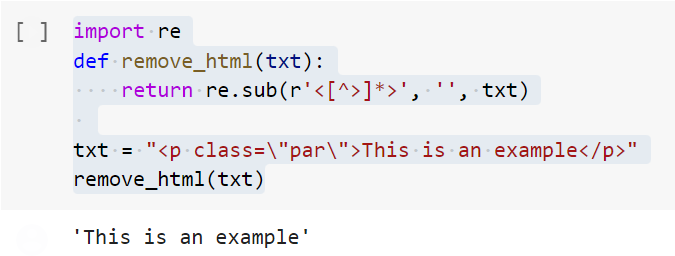
Tiền xử lý dữ liệu bao gồm các công việc như chuẩn hóa dữ liệu, loại bỏ các thành phần không cần thiết cho việc phân loại văn bản, có thể là:

* Xóa HTML code
* Chuẩn hóa bảng mã Unicode
* Thay thế từ viết tắt
* Xóa từ dừng
* Thực hiện tách từ tiếng Việt
* Đưa văn bản thành chữ thường
* Chuẩn hóa câu

### Xóa HTML code

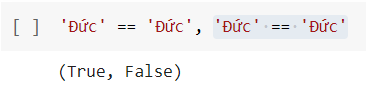
Dữ liệu thu thập trên mạng đôi khi vẫn còn tồn lại đoạn mã HTML, các đoạn mã này không có ý nghĩa cho việc phân loại văn bản, chúng làm cho mô hình chúng ta không còn được chính xác.

Việc xóa các mã HTML có thể được thực hiện như sau:

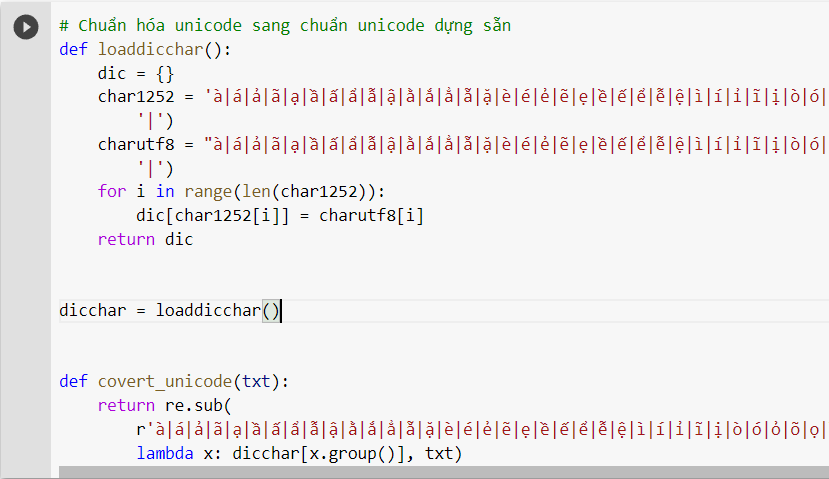


### Chuẩn hóa Unicode

Hiện nay, có 2 loại mã Unicode thường được sử dụng đó là Unicode tổ hợp và Unicode dựng sẵn. Dẫn đến việc dù là 2 từ giống nhau nhưng lại bị xem như là 2 từ khác nhau. Nếu không xử lý vấn đề này thì mô hình sẽ xem như là 2 từ khác nhau, mặc dù chúng là một [2].

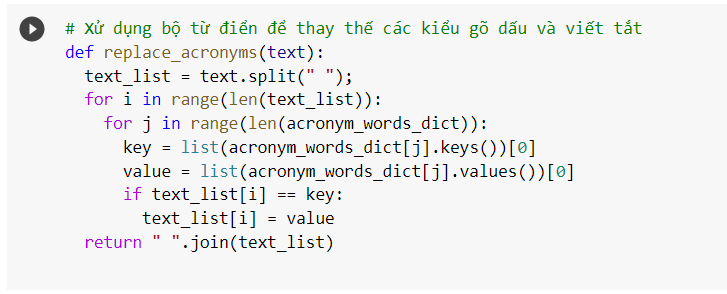


Xử lý vấn đề này bằng cách đưa toàn bộ Unicode tổ hợp về định dạng Unicode dựng sẵn phổ biến hơn:



### Thay thế các từ viết tắt

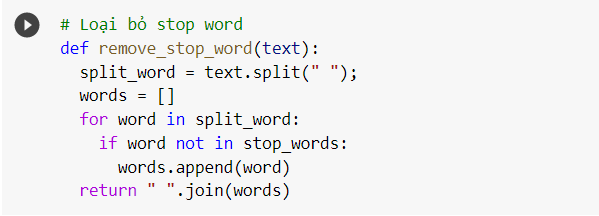
****



### Xóa từ dừng (StopWord)

StopWord là các từ thường xuất hiện nhiều ở ngôn ngữ tự nhiên ví dụ như (nhiều, trên, đó, những ,..) những từ này thường không mang lại ý nghĩa nên chúng ta cần phải loại bỏ chúng đi [2].

Để loại bỏ các StopWord cần chuẩn bị một danh sách các từ StopWord có sẵn. Việc cần làm bây giờ là duyệt qua từng bản ghi của dữ liệu và loại bỏ tất cả các từ trong dữ liệu mà có trong danh sách StopWord

****

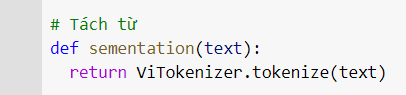
### Thực hiện tách từ tiếng Việt

Đơn vị trong tiếng việt bao gồm từ đơn và từ ghép. Nên chúng ta cần phải nói cho mô hình máy học biết đâu là từ đơn, đâu là từ ghép.



Bởi vì mô hình sẽ coi các từ đặc trưng tách nhau theo dấu cách. Do đó cần phải nối các từ ghép lại thành một từ để không bị tách sai.



****

### Đưa văn bản thành chữ thường

Việc đưa dữ liệu về chữ viết thường là rất cần thiết. Bởi vì đặc trưng này không có tác dụng ở bài toán phân loại văn bản. Đưa về chữ viết thường giúp giảm số lượng đặc trưng (vì máy tính hiểu hoa thường là 2 từ khác nhau) và tăng độ chính xác hơn cho mô hình. [2]

****

### Chuẩn hóa câu

Tiền xử lý bao gồm việc loại bỏ các dữ liệu không có tác dụng cho việc phân loại văn bản. Việc này giúp:

* Giảm số chiều đặc trưng, tăng tốc độ học và xử lý
* Tránh làm ảnh hưởng xấu đến kết quả mô hình

Các dấu ngắt câu, số đếm và các ký tự đặc biệt không giúp cho việc phân loại trở nên tốt hơn. Do đó chúng ta nên loại bỏ nó.

****

## Vector hoá văn bản ( Word Embedding )

Thông thường, máy tính không thể hiểu được ý nghĩa các từ. Như vậy, để xử lý được ngôn ngữ tự nhiên, ta cần có một phương pháp để biểu diễn văn bản dưới dạng mà máy tính có thể hiểu được. Phương pháp tiêu chuẩn để biểu diễn văn bản đó là biểu diễn các văn bản theo vector. Trong đó, các từ, cụm từ thuộc kho tài liệu ngôn ngữ được ánh xạ thành những vector trên hệ không gian số thực[2].

Ta có thể chia các phương pháp Vector hóa văn bản thành hai nhóm chính: Phương pháp Word Embedding cổ điển và Neural Embedding (Vector hóa văn bản theo phương pháp mạng nơ-ron)[2].

Có nhiều phương pháp biểu diễn văn bản như là mô hình không gian vector (vector space model) dựa trên việc đánh trọng số của từ theo tần suất xuất hiện của từ đó, mô hình túi từ (bag of words model), mô hình hóa văn bản thành đồ thị (graph-based model). Trong bài toán này, chúng ta sẽ sử dụng không gian vector để biểu diễn văn bản.

Ví dụ về việc sử dụng không gian vector để biểu diễn văn bản

* văn bản 1: “sản phẩm chính hãng”
* văn bản 2: “sản phẩm trầy xước”

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | sản | phẩm | chính | hãng | trầy | xước |
| văn bản 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| văn bản 2 | 1 | 1 | 0 | 0 | 1 | 1 |

## 

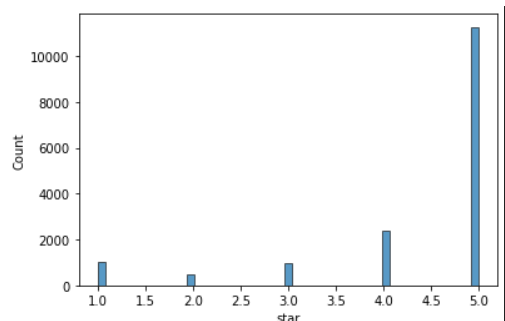
## Xây dựng mô hình phân loại văn bản

### Xử lý mất cân bằng dữ liệu và gộp nhãn

Mất cân bằng dữ liệu sẽ dẫn tới mô hình dự báo kém chính xác và đa phần kết quả dự báo bị thiên về nhãn đa số.

Do dữ liệu không căn bằng, các nhãn 1, 2, 3 rất ít so với các nhãn 4 và 5 nên cần phải xóa dữ liệu nhãn 4 và 5 sau đó gộp các nhãn 1 2 3 thành một nhãn là 0 và gộp các nhãn 4, 5 thành một nhãn là 1.

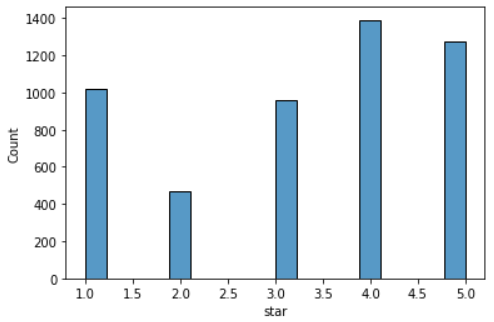
Chuyển dữ liệu có 5 nhãn thành dữ liệu 2 nhãn: 0 (hài lòng) và 1 (không hài lòng) nhằm để cải thiện hiệu năng của mô hình đồng thời để bài toán mang tính trực quan hơn.

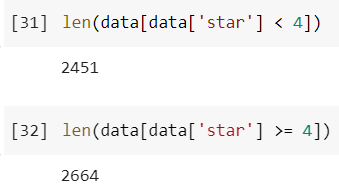


Đoạn code để cân bằng dữ liệu sau đây xóa 10 ngàn phần tử nhãn 5 sao và 1 ngàn phần tử nhãn 4 sao



Sau khi xóa thì dữ liệu mới đã tương đối cân bằng

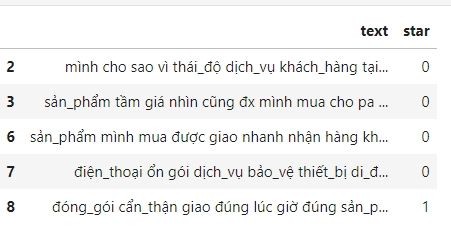




Tiếp theo ta gộp các nhãn 1, 2, 3 thành 0 (không hài lòng) và 4, 5 thành 1 (hài lòng).



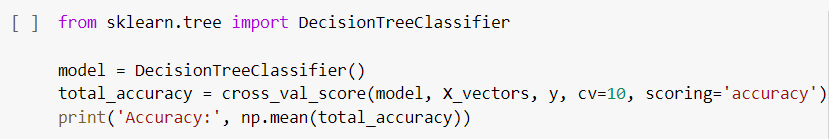
Dữ liệu sau khi gộp các nhãn lại



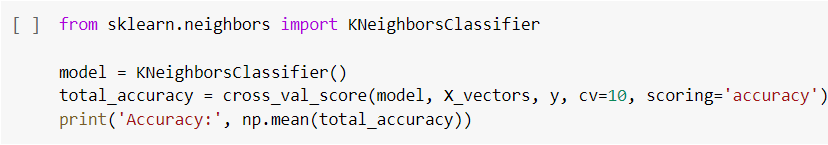
### Huấn luyện mô hình bằng các giải thuật máy học với nghi thức đánh giá K-Fold cross validation

Cross validation là một phương pháp thống kê được sử dụng để ước lượng hiệu quả của các mô hình học máy. Nó thường được sử dụng để so sánh và chọn ra mô hình tốt nhất cho một bài toán. Kỹ thuật này dễ hiểu, dễ thực hiện và cho ra các ước lượng tin cậy hơn so với các phương pháp khác.

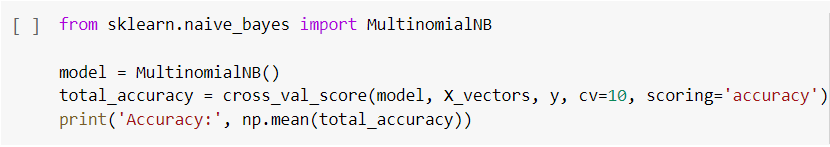
#### Huấn luyện mô hình bằng giải thuật Cây quyết định

****

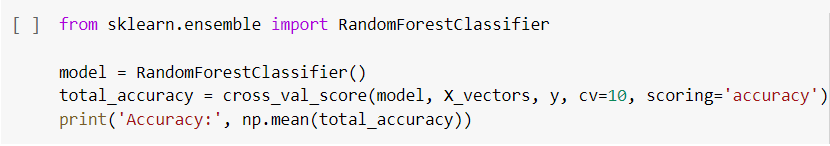
#### Huấn luyện mô hình bằng giải thuật K láng giềng



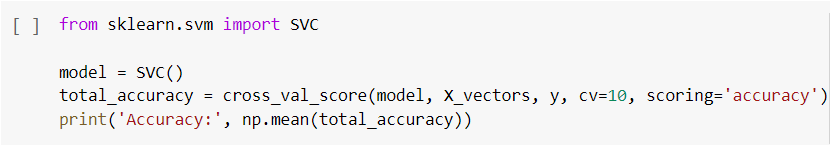
#### Huấn luyện mô hình bằng giải thuật Naïve Bayes



#### Huấn luyện mô hình bằng giải thuật Random Forest



#### Huấn luyện mô hình bằng giải thuật SVM



### Kết quả huấn luyện

* **Lần 1**
* DecisionTreeClassifier: 71.04%
* KNeighborsClassifier: 65.37%
* MultinomialNB: 78.47%
* RandomForestClassifier: 78.43%
* SVM: 78.61%
* Giải thuật Random Forest cho kết quả tốt nhất với độ chính xác **78.61%**
* **Lần 2 Khi không loại bỏ từ dừng**
* DecisionTreeClassifier: 72.70%
* KNeighborsClassifier: 69.05%
* MultinomialNB: 80.48%
* RandomForestClassifier: 80.72%
* SVM: 81.44%
* Giải thuật SVM cho kết quả tốt nhất với độ chính xác **81.44%**
* Khi không loại bỏ từ dừng thì độ chính xác thu được sẽ cao hơn

### Cải thiện mô hình máy học với PhoBERT

Tham khảo từ Phạm Hữu Quang trên viblo [3]

Ta sẽ bắt đầu huấn luyện mô hình bằng PhoBERT sau bước tiền xử lý văn bản

Đầu tiên ta cài đặt thư viện transformers:



Tải về pre-trained PhoBERT base transformers

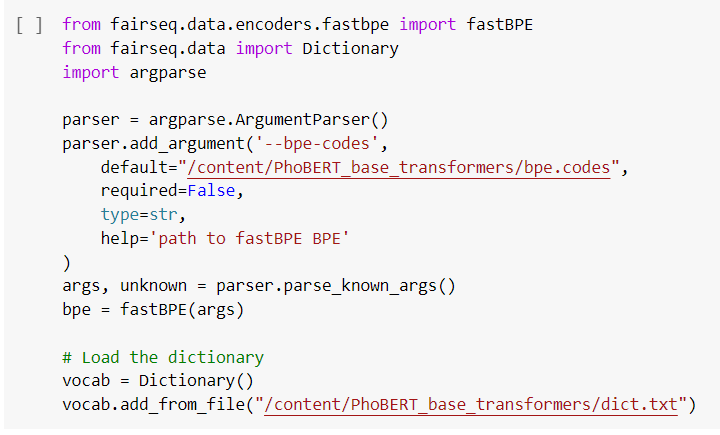


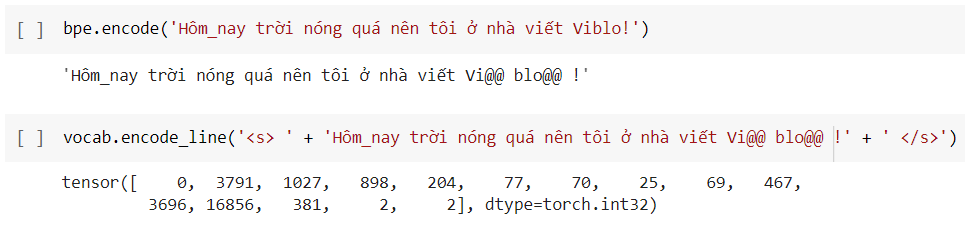
Khi giải nén PhoBERT base transformers, chúng ta sẽ thấy thư mục này gồm 4 file nhỏ bao gồm config.json chứa config của model, model.bin lưu trữ pre-trained weight của model, bpe.codes và dict.txt chứa từ điển sẵn có của PhoBERT [3]

Kế tiếp, chúng ta cài đặt fastBPE và FAIRSeq



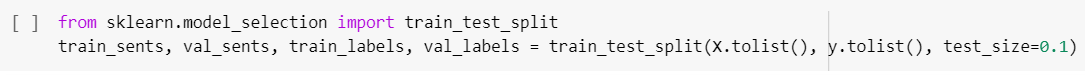
Chúng ta có thể load model và bpe này lên theo hướng dẫn của PhoBERT. [3]



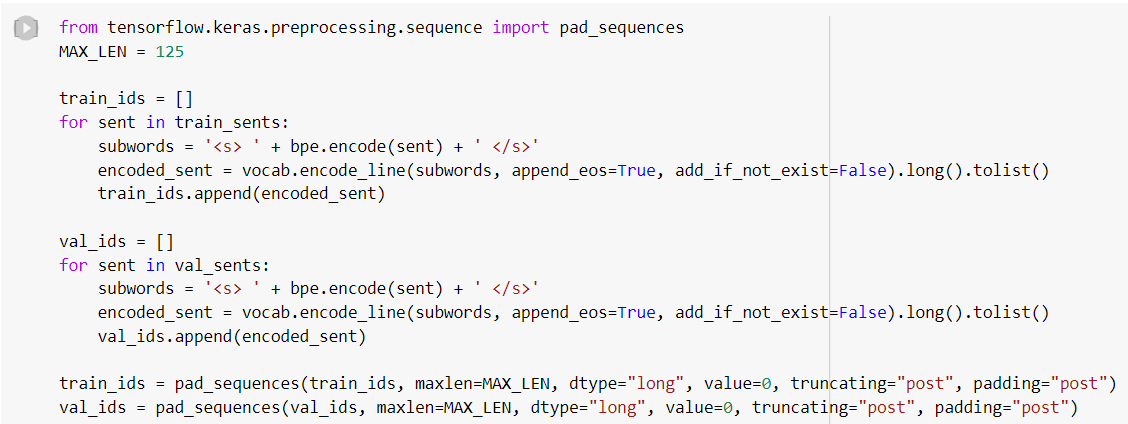


Với mỗi câu chúng ta cần phải thêm <s> là token đánh dấu vị trí bắt đầu câu và </s> đánh dấu kết thúc mỗi câu. [3]

Ta sử dụng phương thức của scikit learn để tách dữ liệu thành 2 tập, tập train và tập kiểm thử để đánh giá mô hình [3]

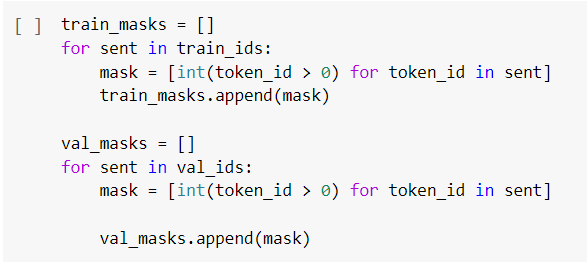


Tiếp theo, từ dữ liệu thô này, chúng ta sử dụng bpe đã load ở trên để đưa text đầu vào dưới dạng subword và ánh xạ các subword này về dạng index trong từ điển [3]

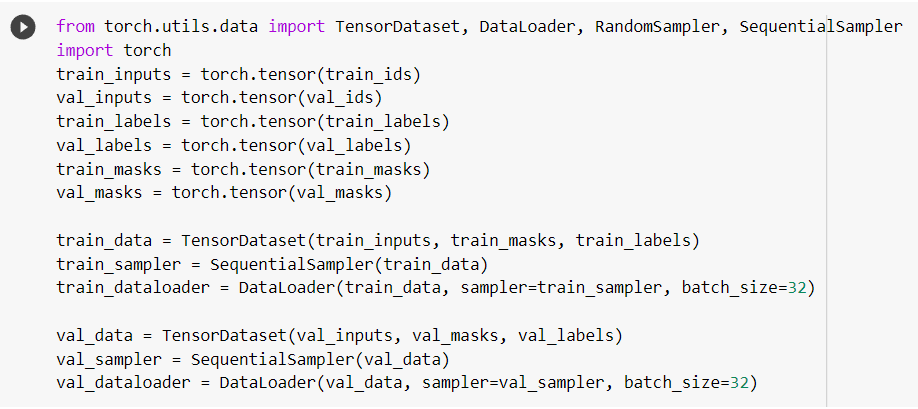


*train\_ids* bây giờ đã trở thành một list dữ liệu mẫu trong đó mỗi mẫu là một list id của các subword có trong từ điển. Các câu ngắn hơn 125 subword được padding 0 ở cuối, những câu dài hơn được cắt đi cho đủ 125. [3]

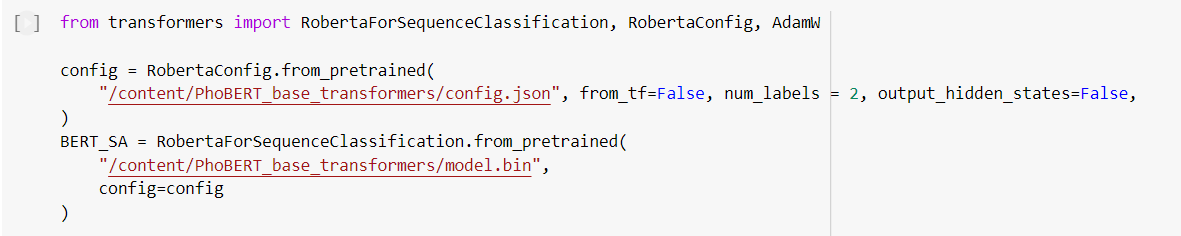
Tiếp theo, mình tạo một mask gồm các giá trị 0, 1 để làm đầu vào cho thư viện transformers, mask này cho biết các giá trị nào của chuỗi đã được padding. [3]

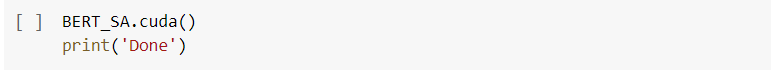


Và giờ, dữ liệu đầu vào cho mô hình đã gần như chuẩn bị xong, chỉ cần chuyển về tensor và sử dụng DataLoader của torch để tạo dataloader nữa thôi. [3]

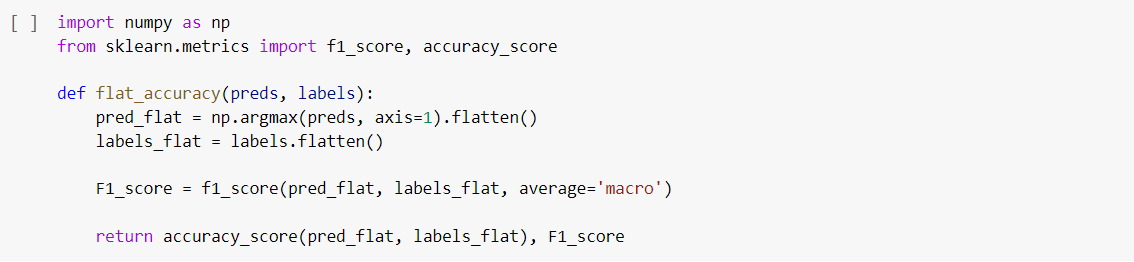


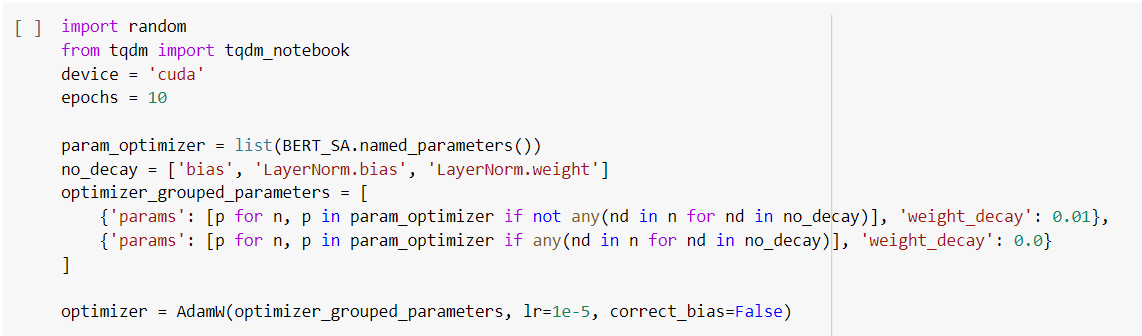
Cuối cùng, việc chuẩn bị dữ liệu cũng đã xong. Chúng ta quay lại với việc load model PhoBERT. [3]

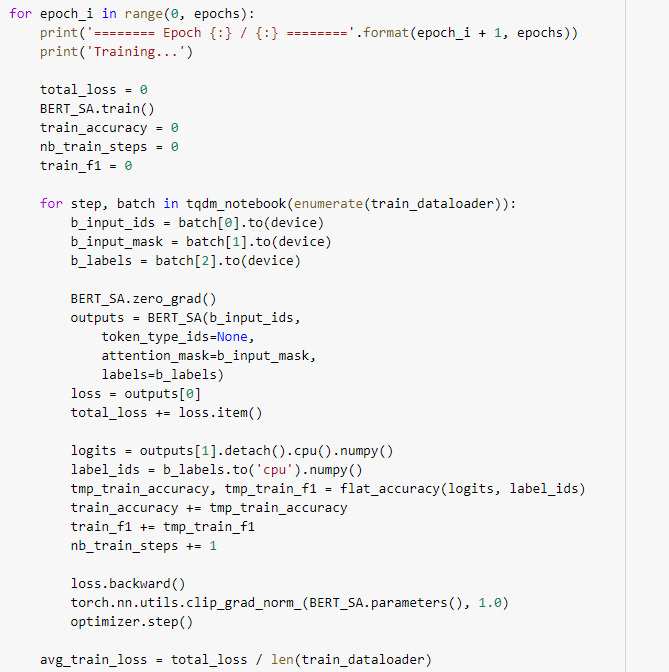


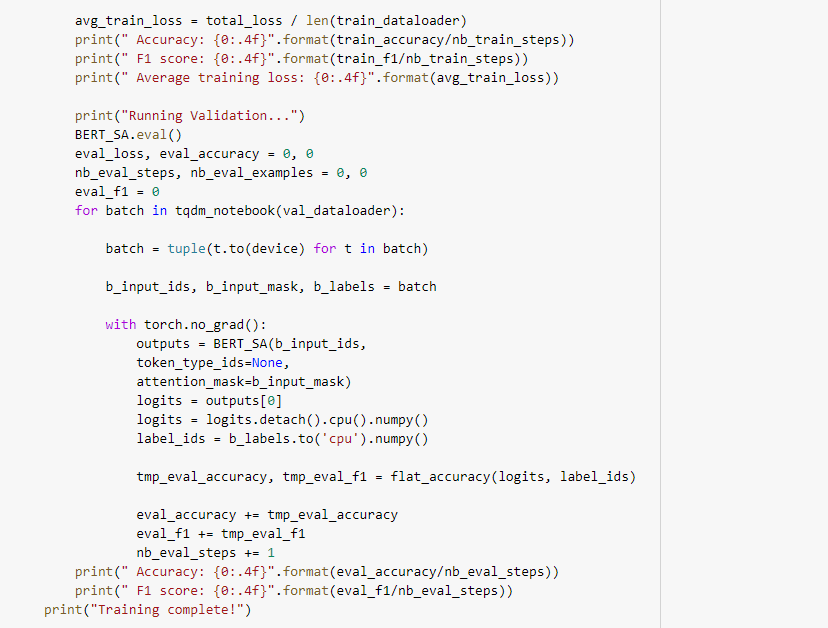


Dưới đây là phần code hoàn thiện cho phần training mô hình

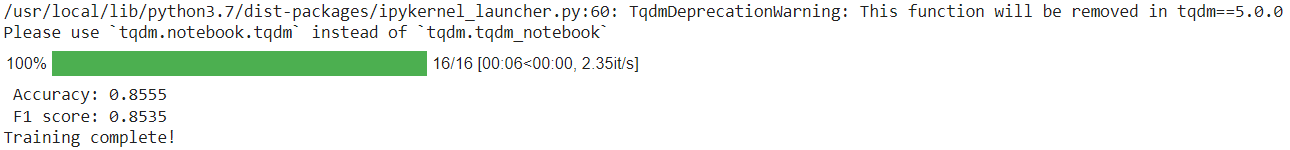








Khi áp dụng PhoBERT vào để huấn luyện thì mô hình cũng không được cải thiện đáng kể. Độ chính xác chỉ đạt khoảng 76%. Còn nếu ta không xóa từ dừng thì độ chính xác có thể tăng lên 85%.



Kết quả huấn luyện

Lần 1:

Accuracy: 76.95% F1 score: 76.16%

Lần 2 khi không xóa từ dừng:

Accuracy: 85.55% F1 score: 85.35%

Do mô hình PhoBERT không cải thiện được độ chính xác đáng kể nên ta sẽ sử dụng mô hình máy học để xây dựng website. Điều này giúp việc xây dựng website đơn giản hơn và website cũng không quá nặng.

## Ứng dụng mô hình máy học vừa huấn luyện để xây dựng website dự đoán đánh giá từ người dùng



*Giao diện ứng dụng phân lớp bình luận sản phẩm trên Website Tiki*

Giao diện gồm 3 thành phần:

* Header: Gồm logo, mã nguồn và thông tin liên hệ
* Body: Phần chính của ứng dụng. Chứa một text box để người dùng nhập bình luận và một button dùng để dự đoán bình luận.
* Footer: Chứa thông tin website

Trên thực tế website cho dự đoán khá chính xác

Với mặt hàng điện thoại di động website dự đoán đúng 8 trên 10 đánh giá, trong đó dự đoán hài lòng đúng 3 trên 5, không hài lòng đúng 5 trên 5. Điều này có thể là do trong những đánh giá hài lòng xuất hiện nhiều từ mang nghĩa tiêu cực.

## Tài liệu tham khảo

[1] <https://drive.google.com/file/d/1nWIBj1_OHZ6k6gvGJAxXjK7IEPNLA_h-/view>

[2] https://nguyenvanhieu.vn/phan-loai-van-ban-tieng-viet/#chuan-bi-du-lieu

[3] Tham Khảo từ Phạm Hữu Quang trên viblo: <https://viblo.asia/p/bert-roberta-phobert-bertweet-ung-dung-state-of-the-art-pre-trained-model-cho-bai-toan-phan-loai-van-ban-4P856PEWZY3>