

THU THẬP DỮ LIỆU

<https://drive.google.com/drive/folders/1SgZFRAWr5106SFBE60-E0UQoGRAPFt-8>

Bộ dữ liệu gồm 3 folder: train, dev, test. Mỗi folder chứa các file sents.txt, sentiments.txt, topic.txt.

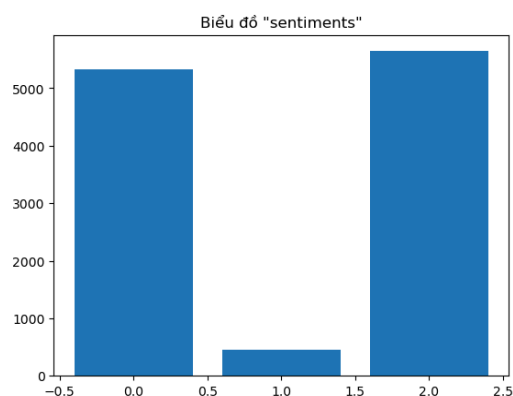
MÔ TẢ DỮ LIỆU:

Trong dữ liệu này có những phản hồi của sinh viên Việt Nam. Những phản hồi của họ là những văn bản ngắn được học sinh viết tự do, được phân loại theo cột “topics” và “sentiments”

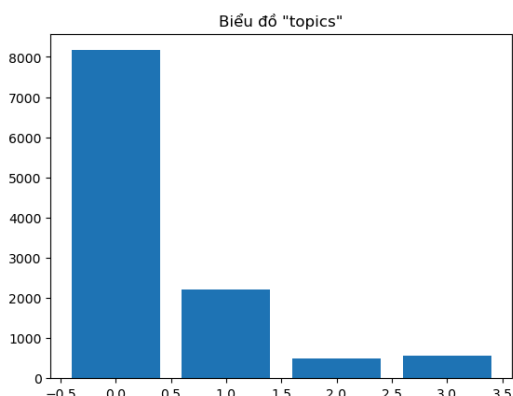
	Sents	sentiments	topics
0	slide giáo trình đầy đủ.	2	1
1	nhật tính giảng dạy , gần gũi với sinh viên .	2	0
2	đi học đầy đủ full điểm chuyên cần .	0	1
3	chưa áp dụng công nghệ thông tin và các thiết bị hỗ trợ cho việc giảng dạy .	0	0
4	thầy giảng bài hay , có nhiều bài tập ví dụ ngay trên lớp .	2	0
...
11421	chỉ vì môn game mà em học hai lần mà không qua thật em rất không hài lòng vì những lý do vô cùng thiếu chuyên nghiệp như thế này .	0	1
11422	em cảm ơn cô nhiều .	2	0
11423	giao bài tập quá nhiều .	0	0
11424	giáo viên dạy dễ hiểu , nhiệt tình	2	0
11425	gợi gọn doubledot hay , tận tình , phù hợp với mọi trình độ cũng như nhu cầu môn học .	2	0

11426 rows x 3 columns

Phân tích tập dữ liệu:



Cột “sentiments” có chứa tổng cộng 3 giá trị: 0, 1, 2



Cột “topics” có chứa tổng cộng 4 giá trị: 0, 1, 2, 3

Sau khi lọc những dòng có dữ liệu số tương ứng, đưa ra được nhãn mà các giá trị số biểu diễn

sentiments

- 0: tiêu cực **5325**
- 1: trung lập **458**
- 2: tích cực **5643**

Topics

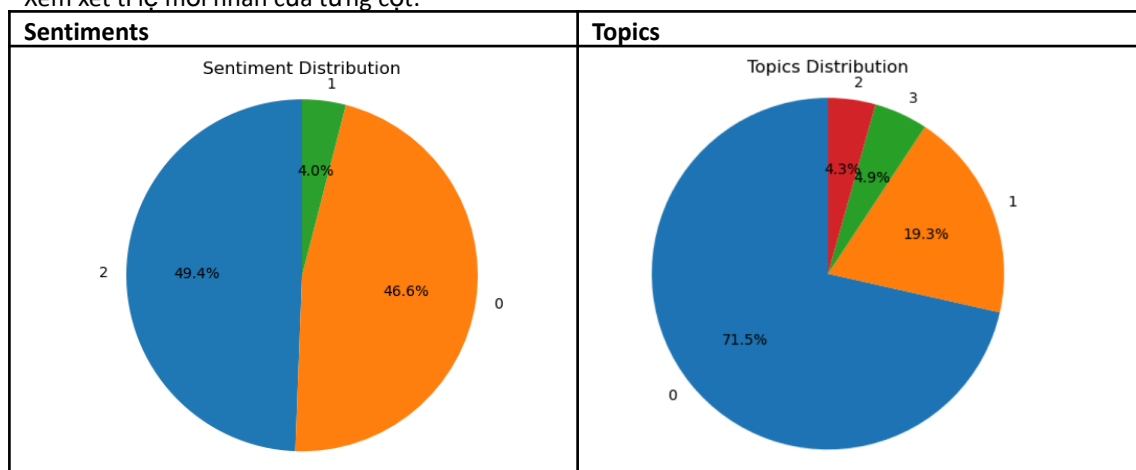
- 0: giảng viên **8166**
- 1: chương trình **2201**
- 2: môi trường **497**
- 3: tào lao **562**

Sinh viên phản hồi về các vấn đề bao gồm giảng viên(0), chương trình giảng dạy(1), cơ sở vật chất(2) và những vấn đề khác(3) tương ứng với dữ liệu của cột (topics)

Đối với mỗi vấn đề nêu trên, phản hồi của sinh viên có 3 loại trạng thái: tiêu cực(0), trung lập(1), tích cực(2) tương ứng với dữ liệu của cột (sentiments)

sentiments	0	1	2	Total
topics				
0	2909	186	5071	8166
1	1716	101	384	2201
2	474	10	13	497
3	226	161	175	562
Total	5325	458	5643	11426

Phản hồi của học sinh chủ yếu là phản hồi về giảng viên, và chủ yếu là phản hồi tích cực (tiêu cực cũng không kém)
 Về chương trình giảng dạy và cơ sở, chủ yếu là phản hồi tiêu cực.
 Ở total của mỗi cột, nhận thấy sự mất cân bằng khá lớn.
 Xem xét tỉ lệ mỗi nhãn của từng cột:



TIỀN XỬ LÝ:

những đối tượng cần chú ý:

"doubledot"

"double."

"dot"

tên riêng đã được mã hóa dưới dạng "wjzwj***"

một số chữ kéo dài lên 2 đơn vị: "ii","aa","ee"

"Colon***"

"(y)"

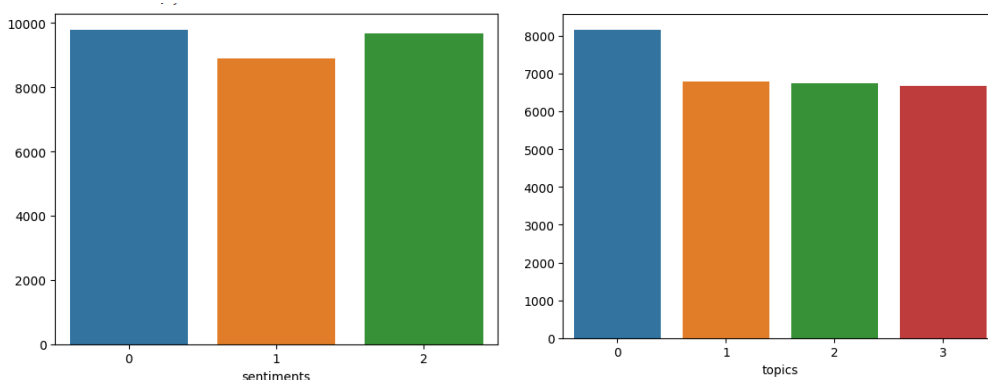
Thực hiện chuyển đổi

Từ	Chuyển đổi thành	Xóa
doubledot	:	
double.	:	
dot	.	
jzwj***	Đồng	
ii	i	
aa	a	
ee	e	
colon***		Xóa

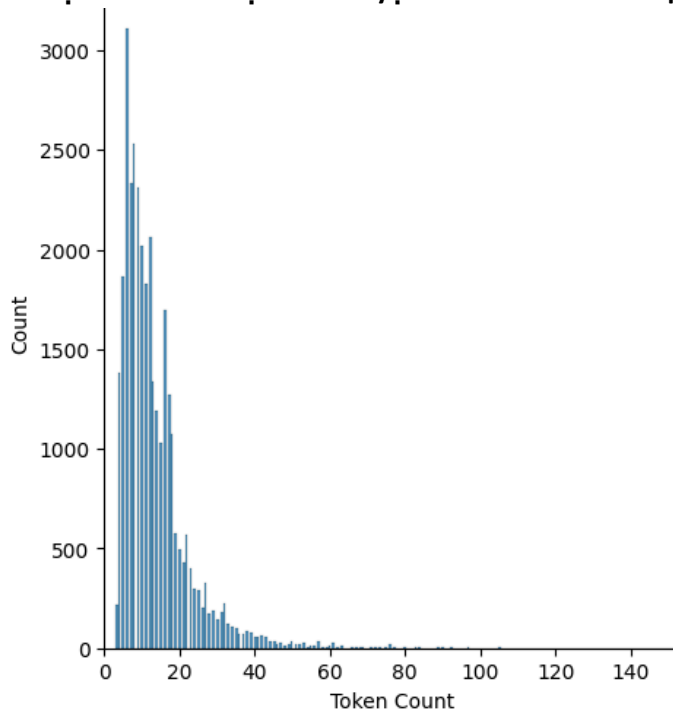
Đối với ký tự "(y)", sau khi xem xét tập dữ liệu, chỉ có một dòng duy nhất chứa "(y)", thực hiện xóa ngay trên tập dữ liệu.

Cân bằng dữ liệu: Hiện tại đã dùng phương pháp đơn giản nhất là **over Sampling**, logic nhân các nhãn bị thiếu

Sau khi cân bằng:



Tải bộ tokenizer được huấn luyện trước cho mô hình phoBert (base)



sau khi vẽ biểu đồ phân phối số lượng token của văn bản được mã hóa, thấy có rất ít câu có độ dài trên 100. Chọn max token (độ dài tối đa) là 100.

Xử lí:

lấy dữ liệu 1 dòng văn bản, sử dụng tokenizer của phoBert để mã hóa:

-cắt câu = max token

-thêm mã thông báo đặc biệt (CLS ở đầu chuỗi và SEP cuối chuỗi, để mô hình có thể biết vị trí bắt đầu xử lí và kết thúc xử lí)

-lấy độ dài tối đa của chuỗi đầu ra = max token

-padding để đồng bộ hóa độ dài các chuỗi đầu ra.

-attention mask để phân biệt phần dữ liệu cần xử lí và phần dữ liệu padding

-trả về văn bản được mã hóa và attention mask đều đã được làm phẳng thành vector 1 chiều.

Triển khai MODEL :

Nhiệm vụ:

Dự đoán dựa trên cảm xúc và dựa trên chủ đề.

- Nhiệm vụ dựa trên **cảm xúc**: Cho một câu phản hồi của học sinh Việt Nam, xác định xem câu đó thể hiện tình cảm tích cực, tiêu cực hay trung lập.

(Ví dụ: các câu có được hiển thị trong bảng và các phân cực tình cảm tương ứng được hiển thị trong “**sentiments**” của bảng dưới.)

- Nhiệm vụ theo **chủ đề**: Cho một câu phản hồi của học sinh Việt Nam, xác định xem câu đó thể hiện thông tin liên quan đến Giảng viên, Chương trình giảng dạy, Cơ sở vật chất hay Khác.

(Ví dụ: các câu có được hiển thị trong bảng và các chủ đề liên quan được trình bày tương ứng ở cột “**topics**” của bảng dưới)

	Sents	sentiments	topics
0	slide giáo trình đầy đủ .	2	1
1	nhiệt tình giảng dạy , gần gũi với sinh viên .	2	0
2	đi học đầy đủ full điểm chuyên cần .	0	1
3	chưa áp dụng công nghệ thông tin và các thiết ...	0	0
4	thầy giảng bài hay , có nhiều bài tập ví dụ ng...	2	0

Model coding:

Dựa trên code mẫu của model train phân loại emotion với data social feedback, thực hiện một số chỉnh sửa để model có thể phân loại đa nhiệm (sentiments và topics)

###.....

ĐÁNH GIÁ

Thực hiện train trên tập dữ liệu chưa cân bằng (a) và đã cân bằng (b):

Gộp data train+dev, Dùng K-fold chia data gộp thành 5 fold, fold đầu để đánh giá, 4 fold sau train

Dùng data train để huấn luyện, data dev để đánh giá và data test để kiểm tra

Giá trị trong mỗi ô:

Sentiments
Topics

Đánh giá kết quả trên tập test:

Model	Accuracy	Precision	Recall	F1-score
(a)	0.9283006	0.92	0.93	0.92
	0.8774478	0.88	0.88	0.88
(b)	0.9283006	0.92	0.93	0.92
	0.8755527	0.88	0.88	0.88
(a)	0.9330385	0.93	0.93	0.93
	0.8878711	0.89	0.89	0.89
(2) (b)	0.9276689	0.92	0.93	0.92
	0.8843967	0.88	0.88	0.88

Thực hiện kiểm tra trên input 27 câu ngẫu nhiên:

Feedback	Sentiments	Topics
Thầy Đồng cho bài tập mà không chỉ cách làm, bắt sinh viên tự mò	Tiêu cực	Giảng viên
Phòng học quá đông, khó tập trung.	Tiêu cực	Cơ sở
Bài kiểm tra quá khó so với nội dung đã học.	Tiêu cực	Cơ sở
Phản hồi từ giảng viên rất hữu ích.	Tích cực	Giảng viên
Em không có ý kiến gì thêm.	Trung lập	Tào lao
Khó khăn trong việc liên lạc với giáo viên.	Tiêu cực	Giảng viên
cần cải thiện chất lượng nhà vệ sinh	Tiêu cực	Cơ sở
Bài giảng có tính ứng dụng cao.	Tích cực	Giáo án
Phòng lab cần được nâng cấp.	Tiêu cực	Cơ sở
Thầy Đồng hiền và dễ thương, mỗi tội giảng khó hiểu	Tiêu cực	Giảng viên
Cần cải thiện chất lượng nhà vệ sinh.	Tiêu cực	Cơ sở
Bài giảng dễ hiểu, slide đẹp mắt.	Tích cực	Giáo án
Cô Y dạy rất hay. Em hy vọng được cô dạy thêm nhiều môn nữa.	Tích cực	Giảng viên
Cần tăng cường an ninh trong khuôn viên trường.	Tiêu cực	Cơ sở
Thiếu sự tận tâm từ một số giáo viên.	Tiêu cực	Giảng viên
Em muốn tìm hiểu thêm về ngành công nghệ thông tin.	Trung lập	Giáo án
Hệ thống wifi cần được cải thiện.	Tiêu cực	Cơ sở
Bãi giữ xe như lộn!	Tiêu cực	Cơ sở
Căn tin rất đẹp, đồ ăn rất ngon.	Tích cực	Cơ sở
tiết học nên được record để hỗ trợ việc học sau giờ.	Tiêu cực	Giáo án
Cần có thêm hỗ trợ tâm lý cho sinh viên.	Tiêu cực	Tào lao
Giảng viên không chỉ chia sẻ kiến thức mà còn khuyến khích sự tò mò và sáng tạo.	Tích cực	Giảng viên
Chương trình học có cấu trúc rõ ràng, giúp tôi theo dõi tiến trình học tập một cách hiệu quả.	Tích cực	Giáo án
Cô Châu rất thân thiện và sẵn sàng hỗ trợ, tạo môi trường học tốt.	Tích cực	Giảng viên
Một tiết học 2h15 phút mà thầy Vĩnh dạy có 1 tiếng.	Tiêu cực	Giảng viên
Một số giảng viên cần cải thiện khả năng sư phạm :)	Tiêu cực	Giảng viên
Đóng tiền để học mà lên toàn phải dạy lại cho giảng viên, chứ giảng viên chẳng dạy gì.	Tiêu cực	Giảng viên

Tỉ lệ dự đoán đúng:

	Sentiments	Sentiments sai	Topics	Topics sai
2a	27/27		26/27	21
2b	26/27	16	26/27	21
1a	25/27	10,16	22/27	7,14,18,19,21
1b	25/27	10,16	24/27	14,16,21

Hiện tại là 2a đang tốt nhất dựa trên accuracy và mẫu input 27 câu.

27 câu ngẫu nhiên có khá ít câu trung lập và tào lao -> mô hình chưa cân bằng có tỉ lệ dự đoán tốt hơn, nhưng kiểm tra xem có tốt với các lớp thiếu số không?

Thử 10 câu (Trung lập - tào lao) lấy trên tập test với data chưa cân bằng

tiếp thu bài cũng ổn .
hoàn toàn bình thường .
tạm thời chưa có .
với lại đặc thù sinh viên trường em nó khác cô ạ .
vì thời gian trong lớp tập trung hơn .
lớp có 2 giảng viên hướng dẫn là cô Đồng và thầy Đồng .
em không có ý kiến gì .
dạ em không có ý kiến về vấn đề này
chú thích những phần cần phải tập trung .
phản ý kiến của em trong phần khảo sát thầy Đồng em nói có liên quan đến thầy luôn .

Kết quả:

2a (đúng 3/10)

Student Feedback	Sentiment	Topic
tiếp thu bài cũng ổn .	Tích cực	Giảng viên
hoàn toàn bình thường .	Tích cực	Tào lao
tạm thời chưa có .	Trung lập	Tào lao
với lại đặc thù sinh viên trường em nó khác cô ạ .	Trung lập	Giảng viên
vì thời gian trong lớp tập trung hơn .	Trung lập	Giáo án
lớp có 2 giảng viên hướng dẫn là cô Đồng và thầy Đồng .	Tích cực	Giảng viên
em không có ý kiến gì .	Trung lập	Tào lao
dạ em không có ý kiến về vấn đề này	Trung lập	Tào lao
chú thích những phần cần phải tập trung .	Tích cực	Tào lao
phản ý kiến của em trong phần khảo sát thầy Đồng em nói có liên quan đến thầy luôn .	Trung lập	Giảng viên

1a (đúng 4/10)

Student Feedback	Sentiment	Topic
tiếp thu bài cũng ổn .	Trung lập	Giáo án
hoàn toàn bình thường .	Trung lập	Tào lao
tạm thời chưa có .	Trung lập	Tào lao
với lại đặc thù sinh viên trường em nó khác cô ạ .	Trung lập	Giảng viên
vì thời gian trong lớp tập trung hơn .	Tiêu cực	Giáo án
lớp có 2 giảng viên hướng dẫn là cô Đồng và thầy Đồng .	Trung lập	Giảng viên
em không có ý kiến gì .	Trung lập	Tào lao
dạ em không có ý kiến về vấn đề này	Trung lập	Tào lao
chú thích những phần cần phải tập trung .	Tích cực	Giảng viên
phản ý kiến của em trong phần khảo sát thầy Đồng em nói có liên quan đến thầy luôn .	Trung lập	Giảng viên

Có thể thấy, cả 2 đều sai nhiều lắm.

2(a) dự đoán sentiments (đúng 6/10) và topic (đúng 5/10) .

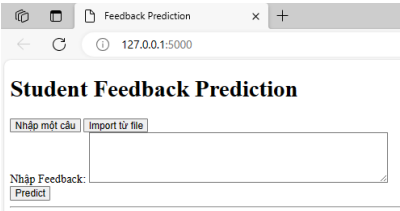
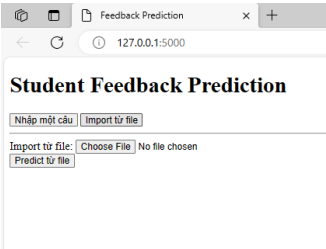
1(a) dự đoán sentiments (đúng 7/10) và topics (đúng 4/10).

Có thể nhận thấy, các thiếu số (trung lập-tào lao) bị dự đoán nhầm là các lớp đa số (tích cực/tiêu cực - giảng viên) khá nhiều

Thử trên tập data đã cân bằng thì kết quả có khả quan hơn, nhưng độ khả quang không thể đánh giá là có cải thiện được.

DEPLOYMENTS:

Đã dựng khuôn web đơn giản nhất có các tính năng phục vụ nhiệm vụ của model.

Cho phép nhập một câu và in ra dự đoán	Cho phép import một file (hiện tại là txt) và in ra dự đoán cho từng câu trong file	
		trả lại dạng bảng và cho phép tải về dưới dạng file excel <div><div>Một số giảng viên cần cả</div><div>Đồng tiền để học mà lên</div><div>Download Predictions</div><div>Quay lại</div></div>