TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KỲ**

**MÔN KHAI THÁC DỮ LIỆU VÀ KHAI PHÁ TRI THỨC**

**NHẬN BIẾT CẢM XÚC VỚI VĂN BẢN TRÊN MẠNG XÃ HỘI VIỆT NAM**

*Người hướng dẫn*: **LÊ CUNG TƯỞNG**

*Người thực hiện*: **NGUYỄN MINH ĐĂNG KHOA – 51800882**

**TRẦN NGỌC DŨNG – 51800187**

Lớp **: 18050302**

Khoá  **: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KỲ**

**MÔN KHAI THÁC DỮ LIỆU VÀ KHAI PHÁ TRI THỨC**

**NHẬN BIẾT CẢM XÚC VỚI VĂN BẢN TRÊN MẠNG XÃ HỘI VIỆT NAM**

Người hướng dẫn: **LÊ CUNG TƯỞNG**

Người thực hiện: **NGUYỄN MINH ĐĂNG KHOA – 51800882**

**TRẦN NGỌC DŨNG – 51800187**

Lớp **: 18050302**

Khoá  **: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

LỜI CẢM ƠN

Chúng em xin cảm ơn thầy Lê Cung Tưởng – người trực tiếp đảm nhiệm trọng trách giảng dạy cũng như cho chúng em những lời khuyên vô cùng hữu ích trong quá trình học tập. Kiến thức của thầy không chỉ nằm ở trong việc giảng dạy lý thuyết mà còn ở những dẫn chứng cụ thể, những trường hợp xảy ra trong quá trình đi làm ở những công ty. Từ những việc đó có thể giúp cho chúng em dễ dàng định hướng cũng như tránh những sai lầm không đáng có. Chúng em kính chúc thầy luôn dồi dào sức khoẻ và gặt hái được nhiều thành công hơn trong sự nghiệp.

MỤC LỤC

[LỜI CẢM ƠN 1](#_Toc72088148)

[MỤC LỤC 2](#_Toc72088149)

[DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 3](#_Toc72088150)

[CHƯƠNG 1 – GIỚI THIỆU 5](#_Toc72088151)

[1.1 Thu thập dữ liệu 5](#_Toc72088152)

[1.2 Nhận biết cảm xúc với văn bản 5](#_Toc72088153)

[1.3 Áp dụng mô hình 5](#_Toc72088154)

[1.4 Phân tích cảm xúc trong ngôn ngữ tự nhiên là gì? 5](#_Toc72088155)

[1.5 Ứng dụng của phân tích cảm xúc trong văn bản 6](#_Toc72088156)

[CHƯƠNG 2 – GIẢI PHÁP 7](#_Toc72088157)

[2.1 Thu thập dữ liệu 7](#_Toc72088158)

[2.1.1 Sử dụng selenium 7](#_Toc72088159)

[2.1.2 Xử lý dữ liệu - preprocessing 10](#_Toc72088160)

[2.2 Nhận biết cảm xúc với văn bản 12](#_Toc72088161)

[2.2.1 Chuẩn bị dữ liệu 13](#_Toc72088162)

[2.2.2 Tiền xử lý dữ liệu - Preprocessing 14](#_Toc72088163)

[2.2.3 Chuẩn hóa dữ liệu - Normalize 15](#_Toc72088164)

[2.2.4 Vector hóa - Vectorizer 16](#_Toc72088165)

[2.2.5 Xây dựng Model – Neural Network 16](#_Toc72088166)

[2.2.6 Đào tạo model 17](#_Toc72088167)

[CHƯƠNG 3 – KẾT QUẢ 19](#_Toc72088168)

[3.1 Dùng model predict tập test 19](#_Toc72088169)

[3.2 Dùng model predict tập data crawl trên Youtube 20](#_Toc72088170)

[TÀI LIỆU THAM KHẢO 22](#_Toc72088171)

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

**DANH MỤC HÌNH**

[Hình 1 Sử dụng Selenium và ChromeDriver để crawl dữ liệu 7](#_Toc72088172)

[Hình 2 Làm sạch dữ liệu crawl trên Youtube 7](#_Toc72088173)

[Hình 3 Các thư viện cần thiết 7](#_Toc72088174)

[Hình 4 Link Youtube muốn crawl 8](#_Toc72088175)

[Hình 5 Thiết lập WebDriver 8](#_Toc72088176)

[Hình 6 Vòng lặp load trang 8](#_Toc72088177)

[Hình 7 Tìm phần tử ‘content-text’ 9](#_Toc72088178)

[Hình 8 Đưa dữ liệu bình luận vào data\_crawl 9](#_Toc72088179)

[Hình 9 Hàm xuất file excel từ data\_crawl 9](#_Toc72088180)

[Hình 10: Đọc dữ liệu và bỏ cột dư thừa 10](#_Toc72088181)

[Hình 11: Chuyển chữ trong bình luận về chữ thường 10](#_Toc72088182)

[Hình 12: Thay thế từ viết tắt, từ chung nghĩa 11](#_Toc72088183)

[Hình 13: Xóa dấu câu và phân đoạn từ 11](#_Toc72088184)

[Hình 14: Xóa icon trong bình luận 12](#_Toc72088185)

[Hình 15: Xuất ra file Excel sau khi preprocessing 12](#_Toc72088186)

[Hình 16 Các bước thực hiện task 2 12](#_Toc72088187)

[Hình 17 Hàm vẽ phân bố Emotion trong data 13](#_Toc72088188)

[Hình 18 Biểu đồ phân bố Emotion trong data train 13](#_Toc72088189)

[Hình 19 Xóa emoji trong các bình luận 14](#_Toc72088190)

[Hình 20 Hàm text\_process 14](#_Toc72088191)

[Hình 21 Chuẩn hóa Emotion trong data train 15](#_Toc72088192)

[Hình 22 Tách ra làm 2 tập data và label 15](#_Toc72088193)

[Hình 23 Khởi tạo TfidfVectorizer 16](#_Toc72088194)

[Hình 24 Vector hóa các câu bình luận 16](#_Toc72088195)

[Hình 25 One-hot encoding cho Emotion 16](#_Toc72088196)

[Hình 26 Neural Network Model với 2 Layer Dense 17](#_Toc72088197)

[Hình 27 EarlyStopping 17](#_Toc72088198)

[Hình 28 Training model 17](#_Toc72088199)

[Hình 29 Vẽ loss và accuracy của train và valid 18](#_Toc72088200)

[Hình 30 Predict test 19](#_Toc72088201)

[Hình 31 Accuray, F1-score trên data\_test 19](#_Toc72088202)

[Hình 32 Đọc data crawl 20](#_Toc72088203)

[Hình 33 Chuẩn hóa Emotion trong data crawl 20](#_Toc72088204)

[Hình 34 Vectorizer cho data crawl 21](#_Toc72088205)

[Hình 35 Độ chính xác gán nhãn trên data crawl 21](#_Toc72088206)

CHƯƠNG 1 – GIỚI THIỆU

Trong bài đồ án này, thực hiện ba công việc:

1.1 Thu thập dữ liệu

Chọn ba video ngôn ngữ Việt Nam trên nền tảng Youtube có ít nhất 100 bình luận và thu thập ít nhất 100 bình luận cho mỗi video. Sau đó tự gán nhãn cho những bình luận thu thập được. Các nhãn bao gồm: Disgust(chán ghét), Enjoyment(thích thú), Anger (giận dữ), Surprise (ngạc nhiên), Sadness (buồn bã), Fear (sợ hãi), Other (khác).

1.2 Nhận biết cảm xúc với văn bản

Sử dụng bộ dữ liệu training và validation trong thư mục UIT-VSMEC của thầy Lê Cung Tưởng đã cung cấp sẵn để xây dựng và đào tạo một mô hình (mô hình phân loại) để gán nhãn cho những bình luận. Sau khi đào tạo mô hình, sử dụng bộ dữ liệu test để predict và báo cáo tỉ lệ phần trăm gán nhãn đúng bằng các độ đo Accuracy, F1-score,…

1.3 Áp dụng mô hình

Cuối cùng là áp dụng mô hình đã xây dựng và đạo tạo bên trên để predict dữ liệu đã thu thập và báo cáo tỉ lệ phần trăm gán nhãn đúng bằng các độ đo Accuracy, F1-score,…

1.4 Phân tích cảm xúc trong ngôn ngữ tự nhiên là gì?

Bài toán phân tích cảm xúc thuộc dạng bài toán phân tích ngữ nghĩa văn bản. Vì vậy, ta cần phải xây dựng một mô hình hiểu được ý nghĩa của câu văn, đoạn văn để quyết định xem câu văn đó hoặc đoạn văn đó mang cảm xúc nào.

Phát biểu theo góc nhìn của máy học (Machine Learning) thì phân tích cảm xúc là bài toán phân lớp cảm xúc dựa trên văn bản ngôn ngữ tự nhiên. Đầu vào của bài toán là một câu hay một đoạn văn bản, còn đầu ra là các giá trị xác suất (điểm số) của N lớp cảm xúc mà ta cần xác định.

1.5 Ứng dụng của phân tích cảm xúc trong văn bản

Việc phân tích cảm xúc trong văn bản được ứng dụng trong hàng loạt các vấn đề như: Quản trị thương hiệu doanh nghiệp, thương hiệu sản phẩm, quản trị quan hệ khách hàng, khảo sát ý kiến xã hội học, phân tích trạng thái tâm lý con người, đánh giá bình luận người dùng...

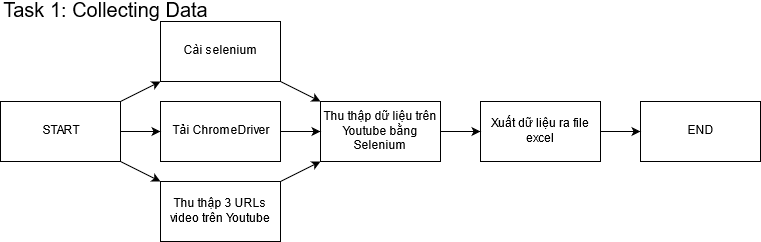
Chúng ta đang sống trong kỷ nguyên số, đặc biệt những năm gần đây nổi lên với mạng xã hội, với hàng triệu người dùng trên thế giới, với lượng thông tin nội dung được người dùng tạo ra hằng ngày cực kỳ lớn, với đa dạng các hình thức như dòng trạng thái, hình ảnh, video. Mạng xã hội có những đặc điểm là: thông tin do người dùng tạo ra, mang tính cá nhân cho nên chất lượng nội dung hay tính đúng đắn, xác thực là tương đối; một thông tin mới được tạo lại có sức lan tỏa nhanh đến đông đảo các người dùng khác, so với các kênh thông tin truyền thống như truyền hình, truyền thành, báo chí, diễn đàn, blog...

Điều này đặt ra cho các doanh nghiệp lớn giải quyết bài toán quản trị thương hiệu doanh nghiệp, quản trị thương hiệu sản phẩm trước các dư luận không tốt trên mạng xã hội rất khó khăn, cả về nguồn xuất phát thông tin, cả về khối lượng thông tin cần xử lý. Chưa kể việc các đối thủ cạnh tranh trên thương trường lợi dụng mạng xã hội để cố ý tạo các thông tin bất lợi cho nhau.

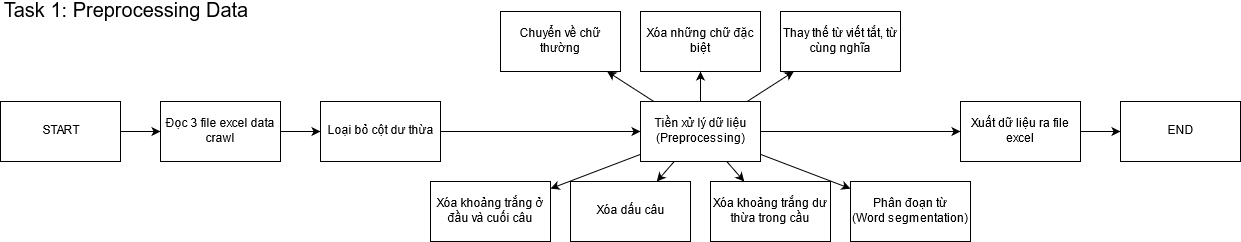
Các doanh nghiệp mua các dữ liệu thời gian thực (real time) từ các công ty mạng xã hội về để xử lý các thông tin liên quan đến doanh nghiệp hay các sản phẩm mà doanh nghiệp đó kinh doanh, nhằm phát hiện và ngăn chặn sớm sự lan rộng các thông tin bất lợi trên mạng xã hội, có hình thức đính chính phản hồi đến các khách hàng của mình, đồng thời thương lượng, ngăn chặn tận gốc những người tạo ra các nội dung đó. Điều cốt yếu của giải pháp này chính là phân tích cảm xúc của các dòng trạng thái trên mạng xã hội nhằm lọc ra các thông tin bất lợi để xử lý.

CHƯƠNG 2 – GIẢI PHÁP

2.1 Thu thập dữ liệu



Hình Sử dụng Selenium và ChromeDriver để crawl dữ liệu

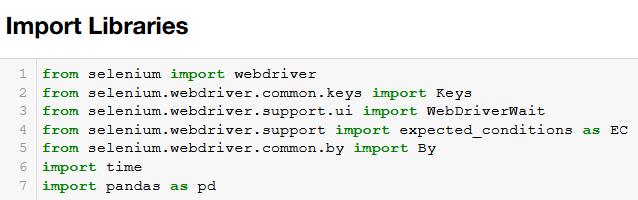


Hình Làm sạch dữ liệu crawl trên Youtube

2.1.1 Sử dụng selenium

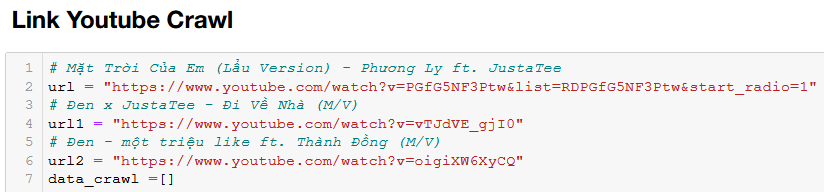
Selenium là một công cụ phổ biến để tự động hóa các trình duyệt. Selenium yêu cầu ChromeDriver với trình duyệt Chrome trước khi thu thập dữ liệu.

Đầu tiên, import một số thư viện cần thiết:



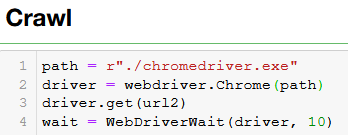
Hình Các thư viện cần thiết

Tạo biến chứa các link Youtube đã chọn và tập data\_crawl:



Hình Link Youtube muốn crawl

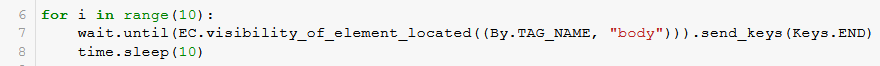
Bắt đầu crawl dữ liệu tuần tự với các url trên. Bắt đầu thiết lập webdriver để tìm nạp nội dung của URL từ Youtube:



Hình Thiết lập WebDriver

Dán url vào driver.get(url) – thao tác này sẽ mở ra một cửa số trình duyệt Chrome mới cho liên kết đó. WebDriverWait yêu cầu webdriver tạm dừng một khoảng thời gian (tính theo giây), rồi sau đó mới thực hiện tiếp các thao tác tiếp theo, nếu không có wait có thể dẫn đến lỗi NotFoundException do không tìm thấy phần tử muốn crawl. Lí do là do trang web chưa tải được xong, vấn đề là do mạng kém hoặc dữ liệu cần load lên quá lớn, trong một khoảng thời gian ngắn không load kịp.

Chromedriver kéo xuống cuối trang để load trang 10 lần, mỗi lần sẽ sleep(10) để trình duyệt load.



Hình Vòng lặp load trang

Sau khi load xong, tạo biến comments để tìm những phần tử có id là ‘content-text’.



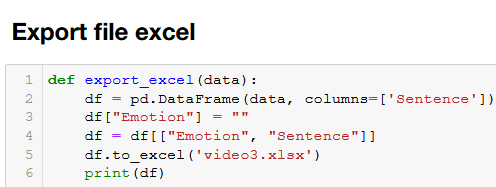
Hình Tìm phần tử ‘content-text’

Cuối cùng là dùng for để đưa comment vào tập data\_crawl đã tạo trước đó.



Hình Đưa dữ liệu bình luận vào data\_crawl

Sau khi đã thu thập xong dữ liệu, viết một hàm export data\_crawl ra file excel để tiến hành xử lý data và gán nhãn.

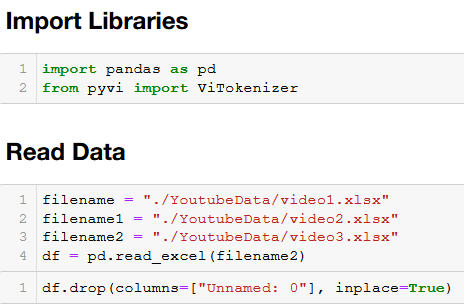


Hình Hàm xuất file excel từ data\_crawl

Lặp lại các thao tác trên với 3 url tạo ra 3 file excel data của 3 video.

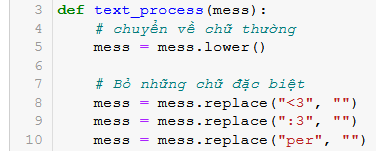
2.1.2 Xử lý dữ liệu - preprocessing

Đầu tiên là import thư viện cần thiết, sau đó đọc dữ liệu vào bằng thư viện pandas và loại bỏ cột “Unnamed: 0” ra khỏi dataframe.



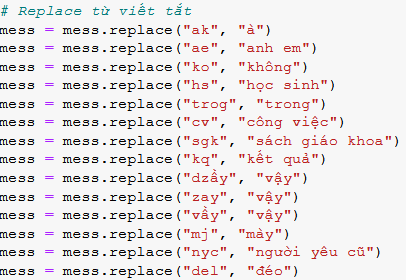
Hình : Đọc dữ liệu và bỏ cột dư thừa

Tạo một hàm text\_process để preprocessing. Đầu tiên, chuyển tất cả về chữ thường, sau đó bỏ những từ đặc biệt.



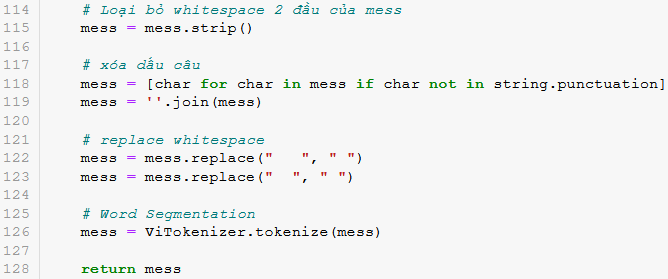
Hình : Chuyển chữ trong bình luận về chữ thường

Tiếp theo sẽ replace một số từ viết tắt, một từ chung nghĩa nhưng lại được viết khác nhau được liệt kê từ bộ dữ liệu train của thầy Lê Cung Tưởng.



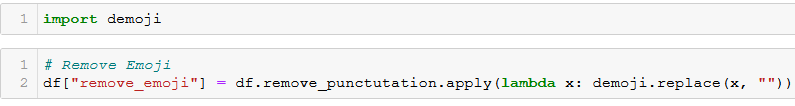
Hình : Thay thế từ viết tắt, từ chung nghĩa

Cuối cùng là loại bỏ các khoảng trắng ở đầu cuối của mỗi câu (strip), xóa dấu câu trong câu (string.punctuation), thay thế các khoảng trắng dư thừa trong câu và quan trọng của tiếng Việt là phân đoạn từ (word segmentation) trong tiếng Việt (sử dụng thư viện pyvi).



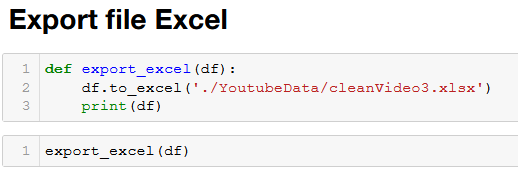
Hình : Xóa dấu câu và phân đoạn từ

Áp dụng hàm text\_process cho dataframe Sentence. Sau đó tiếp tục sử dụng thư viện demoji để xóa bỏ các icon trong bình luận.



Hình : Xóa icon trong bình luận

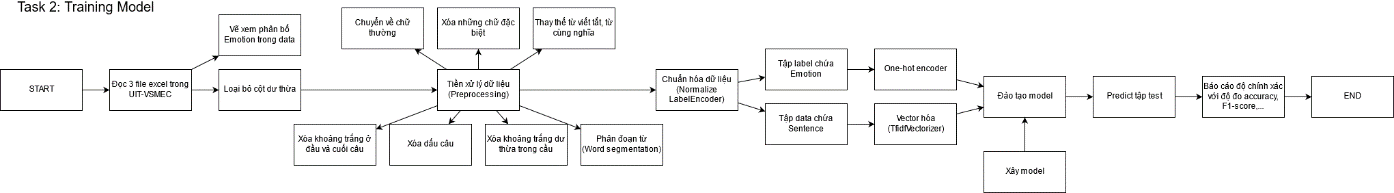
Sau tất cả, viết hàm xuất những bình luận đã preprocessing ra file excel và bắt đầu gán nhãn thủ công.



Hình : Xuất ra file Excel sau khi preprocessing

Lặp lại các thao tác trên với 3 file excel đã crawl trước đó và tạo ra 3 file excel data đã preprocessing.

2.2 Nhận biết cảm xúc với văn bản

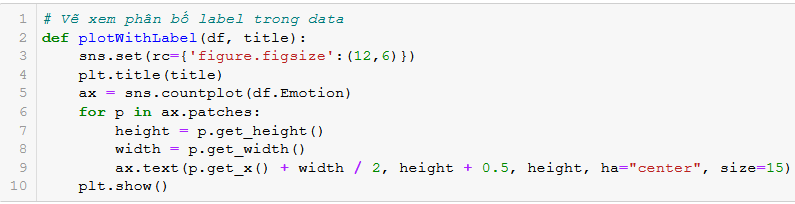


Hình 16 Các bước thực hiện task 2

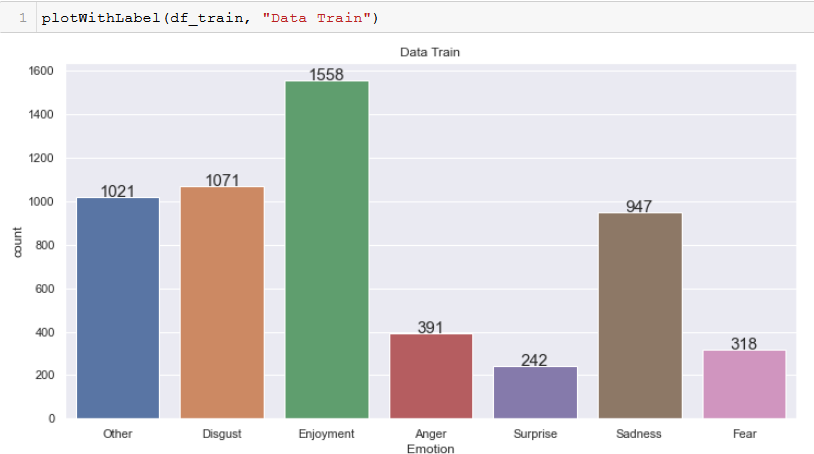
2.2.1 Chuẩn bị dữ liệu

Import các thư viện cần dùng, sau đó đọc 3 excel trong thư mục UIT-VSMEC vào 3 dataframe df\_train, df\_test, df\_valid. Loại bỏ cột dư thừa.

Tạo 1 hàm vẽ để xem phân bố Emotion trong 3 tập data với thư viện seaborn.



Hình 17 Hàm vẽ phân bố Emotion trong data

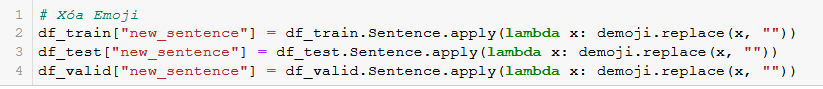


Hình 18 Biểu đồ phân bố Emotion trong data train

2.2.2 Tiền xử lý dữ liệu - Preprocessing

Những bước xử lý dữ liệu giống với lúc xử lý dữ liệu trên data crawl.

Xóa emoji trong bình luận bằng thư viện demoji.



Hình 19 Xóa emoji trong các bình luận

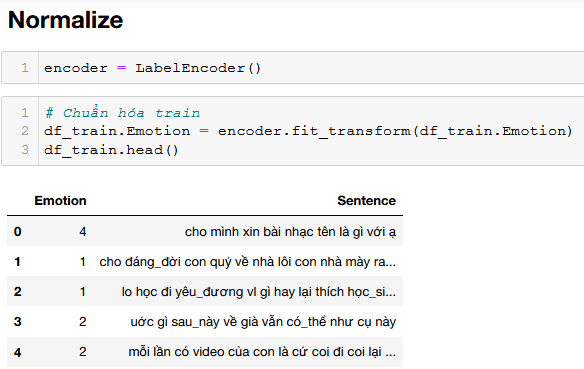
Tạo hàm text\_process gồm chuyển về chữ thường, bỏ những chữ đặc biệt, thay thế từ viết tắt, loại bỏ khoảng trắng đầu cuối câu, xóa dấu câu (string.punctuation), thay thế khoảng trắng dư và quan trọng phân đoạn từ (word segmentation) bằng thư viện pyvi.



Hình 20 Hàm text\_process

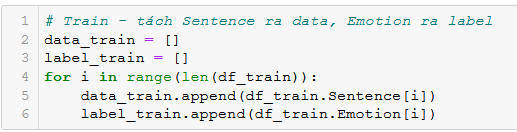
2.2.3 Chuẩn hóa dữ liệu - Normalize

Chuẩn hóa dữ liệu bằng LabelEncoder của sklearn – chuyển đổi 7 phân lớp cảm xúc thành số trong 3 tập data.



Hình 21 Chuẩn hóa Emotion trong data train

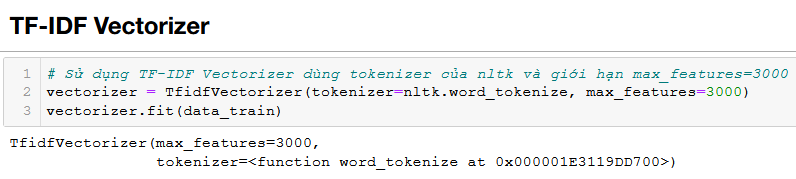
Sau đó tạo ra 2 tập gồm data và label để chứa Sentence và Emotion.



Hình 22 Tách ra làm 2 tập data và label

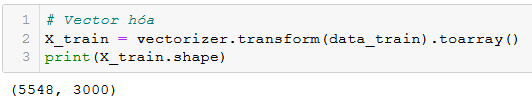
2.2.4 Vector hóa - Vectorizer

Sử dụng TF-IDF Vectorizer của sklearn để vector hóa data của 3 tập train, test, valid. Khởi tạo TfidfVectorizer dùng tokenizer của nltk.word\_tokenize và giới hạn max\_features là 3000. Sau đó fit trên data\_train để học từ vựng.



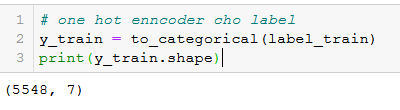
Hình 23 Khởi tạo TfidfVectorizer

Sau đó, dùng vectorizer.transform và toarray() để vector hóa cho từng bình luận trong data\_train, data\_test, data\_valid.



Hình 24 Vector hóa các câu bình luận

Còn phần Emotion thì dùng to\_categorical của keras để one-hot encoding.



Hình 25 One-hot encoding cho Emotion

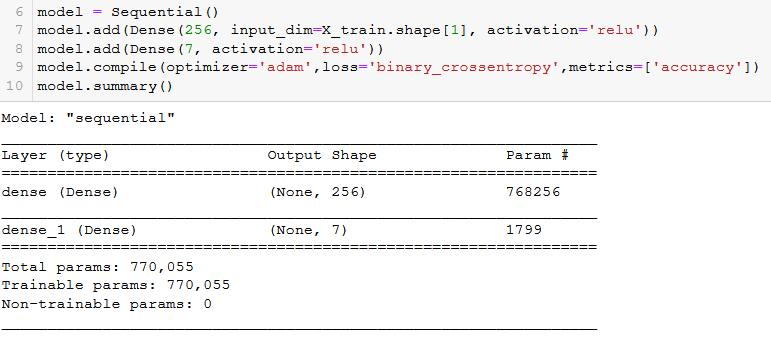
2.2.5 Xây dựng Model – Neural Network

Đầu tiên, khởi tạo model bằng Sequential của Keras.

Layer Dense đầu tiên với unit=256, input\_dim là features của train, hàm kích hoạt dùng relu.

Thêm một Layer Dense với unit=7 – đầu ra của model, hàm kích hoạt là relu.

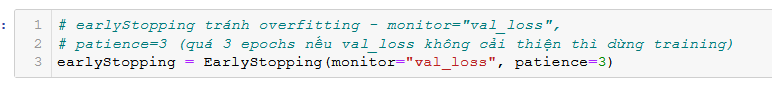
Model compile với optimizer=adam, hàm loss=binary\_crossentropy, metrics=accuracy.



Hình 26 Neural Network Model với 2 Layer Dense

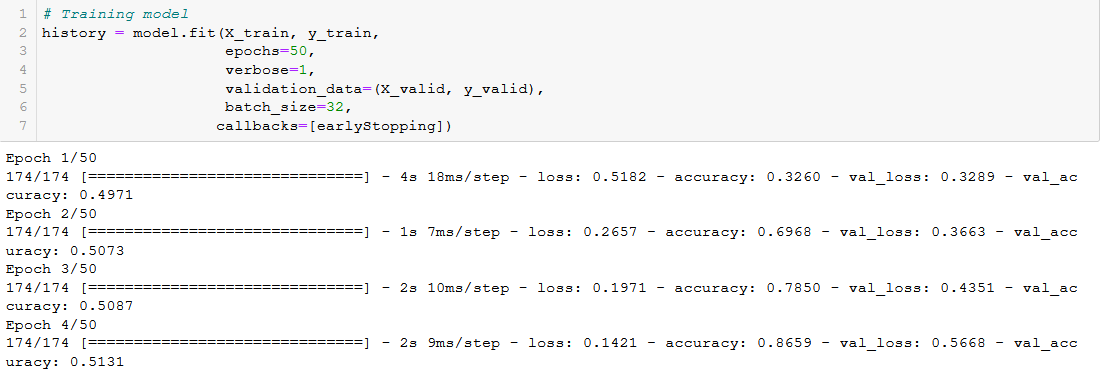
2.2.6 Đào tạo model

Khởi tạo earlyStopping để tránh overfitting với monitor là val\_loss, patience là 3 (nếu quá 3 epochs val\_loss không cải thiện thì sẽ dừng training model).



Hình 27 EarlyStopping

Đào tạo model với EPOCHS=50, BATCH\_SIZE=32.



Hình 28 Training model

Sau đó vẽ biểu diễn loss và accuracy của train và valid.

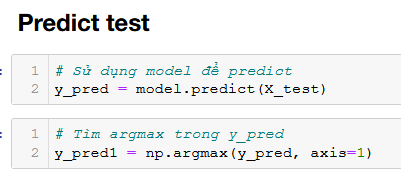


Hình 29 Vẽ loss và accuracy của train và valid

CHƯƠNG 3 – KẾT QUẢ

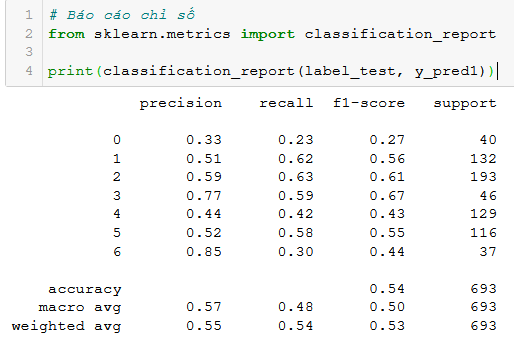
3.1 Dùng model predict tập test

Sử dụng model.predict lên tập test.



Hình 30 Predict test

Báo cáo độ chính xác với classification\_report của sklearn.



Hình 31 Accuray, F1-score trên data\_test

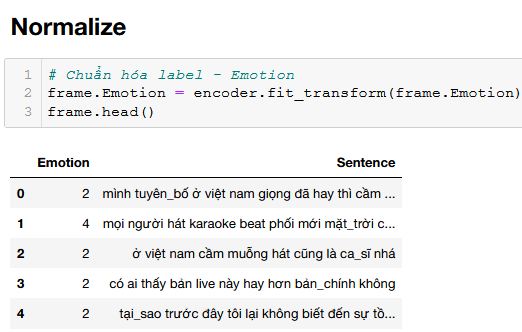
3.2 Dùng model predict tập data crawl trên Youtube

Đầu tiên, đọc 3 file excel crawl đã preprocessing vào và gom thành 1 dataframe với pandas. Sau đó loại bỏ cột dư thừa.



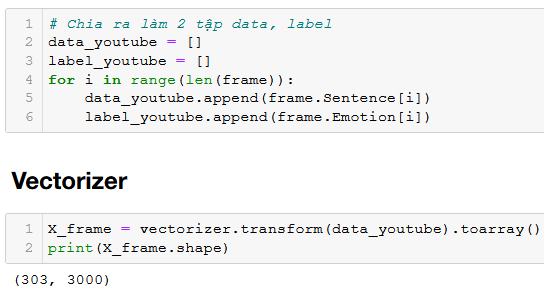
Hình 32 Đọc data crawl

Tiếp theo là chuẩn hóa Emotion với LabelEncoder.



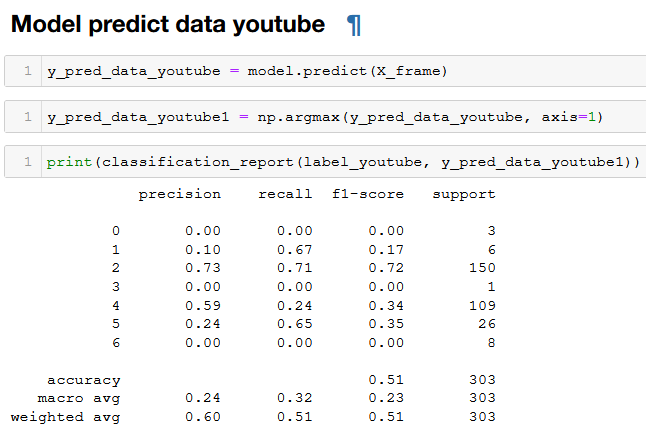
Hình 33 Chuẩn hóa Emotion trong data crawl

Chia dataframe ra làm 2 tập data và label. Sau đó, vector hóa cho data.



Hình 34 Vectorizer cho data crawl

Dùng model đã training để predict data crawl và báo cáo độ chính xác với accuracy, F1-score,….



Hình 35 Độ chính xác gán nhãn trên data crawl

TÀI LIỆU THAM KHẢO

1. <https://streetcodevn.com/blog/sav>