TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN MÔN**

**KHAI KHÁC DỮ LIỆU VÀ KHAI PHÁ TRI THỨC**

**EMOTION RECOGNITION FOR VIETNAMESE SOCIAL MEDIA TEXT**

*Người hướng dẫn*: **TIẾN SĨ LÊ CUNG TƯỞNG**

*Người thực hiện*: **HUỲNH TẤN LỢI – 51800574**

**VÕ TẤN LỰC - 51800900**

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 MÔN**

**KHAI KHÁC DỮ LIỆU VÀ KHAI PHÁ TRI THỨC**

**EMOTION RECOGNITION FOR VIETNAMESE SOCIAL MEDIA TEXT**

*Người hướng dẫn*: **TIẾN SĨ LÊ CUNG TƯỞNG**

*Người thực hiện*: **HUỲNH TẤN LỢI**

**VÕ TẤN LỰC**

Khoá  **: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

LỜI CẢM ƠN

Trong thời gian làm đồ án này, chúng em đã nhận được nhiều sự giúp đỡ, đóng góp ý kiến và chỉ bảo nhiệt tình của thầy và bạn bè.

Nhóm em xin gửi lời cảm ơn chân thành đến thầy Lê Cung Tưởng, giảng viên Bộ môn Khai thác dữ liệu và khai phá tri thức ĐH Tôn Đức Thắng , người đã tận tình hướng dẫn, chỉ bảo nhóm em trong suốt quá trình làm đồ án cũng như trong việc giảng dạy. Nhóm em cũng xin chân thành cảm ơn các thầy cô giáo trong trường ĐH Tôn Đức Thắng nói chung, các thầy cô trong khoa CNTT nói riêng đã dạy dỗ cho em kiến thức về các môn đại cương cũng như các môn chuyên ngành, giúp chúng em có được cơ sở lý thuyết vững vàng và tạo điều kiện giúp đỡ nhóm em trong suốt quá trình học tập. Cuối cùng, nhóm em xin chân thành cảm ơn gia đình và bạn bè, đã luôn tạo điều kiện, quan tâm, giúp đỡ, động viên em trong suốt quá trình học tập và hoàn thành bài đồ án này.

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Chúng tôi xin cam đoan đây là sản phẩm đồ án của riêng chúng tôi và được sự hướng dẫn của thầy Lê Cung Tưởng. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào chúng tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do chúng tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 15 tháng 5 năm 2021*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

TÓM TẮT

Nhận biết cảm xúc đổi với các văn bản trên mạng xã hội Việt Nam (Emotion Recognition for Vietnamese Social Network Text) là một đề tài được đề xuất trong đồ án lần này. Cụ thể, ở đây chúng ta sẽ tìm hiểu cảm xúc trong các đoạn văn bản trên mạng xã hội Việt Nam như: Youtube, Facebook, Instagram,…

Đề tài được cung cấp 3 tập dataset:

* test\_nor\_811.xlsx
* train\_nor\_811.xlsx
* valid\_nor\_811.xlsx

Với đề tài “Nhận biết cảm xúc đổi với các văn bản trên mạng xã hội Việt Nam”, việc nhận biết cảm xúc là một cách tiếp cận cao hơn trong việc phân tích tình cảm. Trong nhiệm vụ của đề tài được đặt ra, không phải kết quả theo 2 hướng positive hay negative hay là đánh giá mức độ cảm xúc theo dạng xếp hạng (1 đến 5) mà ở đây chúng ta sẽ phân tích mức độ cảm xúc chi tiết hơn bằng các biểu hiện như: Disgust (chán ghét), Enjoyment (thích thú), Anger (giận dữ), Surprise (ngạc nhiên), Sadness (buồn bã), Fear (sợ hãi) và Other (khác). Sau khi dùng tập dataset được cung cấp để train model, dữ liệu test sẽ được crawl từ các comment trên các trang mạng xã hội và cụ thể ở đây sẽ là Youtube. Chúng ta sẽ chọn 3 video tương ứng, crawl data là các comment trong 3 video và xuất thảnh 3 tập dataset:

* GFG\_1.csv
* GFG\_2.csv
* GFG\_3.csv

Cuối cùng, thực hiện đánh giá với 3 tập dữ liệu ta thu được và đưa ra bảng so sánh kết quả độ chính xác giữa các dữ liệu kiểm thử với nhau.

MỤC LỤC

[LỜI CẢM ƠN 1](#_Toc72067458)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN 3](#_Toc72067459)

[TÓM TẮT 4](#_Toc72067460)

[MỤC LỤC 1](#_Toc72067461)

[DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 3](#_Toc72067462)

[CHƯƠNG 1 – THU THẬP DỮ LIỆU 6](#_Toc72067463)

[1.1 Giới thiệu 6](#_Toc72067464)

[1.2 Kỹ thuật sử dụng 6](#_Toc72067465)

[1.2.1 Selenium 6](#_Toc72067466)

[1.2.2 Cài đặt selenium và chrome drive 7](#_Toc72067467)

[1.3 Thực thi 8](#_Toc72067468)

[CHƯƠNG 2 – TIỀN XỬ LÝ VÀ HUẤN LUYỆN MÔ HÌNH (PREPROCESSING AND TRANING MODEL) 10](#_Toc72067469)

[2.1. Tiền xử lý dữ liệu (preprocessing data) 10](#_Toc72067470)

[2.1.1 Loại bỏ các từ kéo dài 10](#_Toc72067471)

[2.1.2 Dấu câu (punctuation) 10](#_Toc72067472)

[2.1.3 Tách từ (Word segmentation) 11](#_Toc72067473)

[2.2. Huấn luyện mô hình (Training model) 13](#_Toc72067474)

[2.2.1 Đọc file và xử lý dữ liệu 13](#_Toc72067475)

[2.2.2 TF-IDF vectorizer 16](#_Toc72067476)

[2.2.3 Huấn luyện mô hình – Logistic Regression 17](#_Toc72067477)

[2.3 Dữ liệu thử nghiệm (Testing data) 18](#_Toc72067478)

[CHƯƠNG 3 - THỰC NGHIỆM TRÊN DỮ LIỆU CRAWL 20](#_Toc72067479)

[3.1. Tổng quan về tập dữ liệu 20](#_Toc72067480)

[3.2 Thực nghiệm 20](#_Toc72067481)

[3.3 Bảng so sánh kết quả 24](#_Toc72067482)

[TÀI LIỆU THAM KHẢO 25](#_Toc72067483)

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

**DANH MỤC HÌNH**

[Hình 1: Ví dụ về việc gắn label cho từng comment sau khi crawl thành công 6](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066507)

[Hình 2: Lệnh cài đặt selenium và chrome drive trên colab 7](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066508)

[Hình 3: Khai báo và config cái library 8](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066509)

[Hình 4: Thẻ youtube 8](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066510)

[Hình 5: Thực thi crawl comment 9](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066511)

[Hình 6: ví dụ crawl data 9](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066512)

[Hình 7: Loại bỏ từ kéo dài 10](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066513)

[Hình 8: Punctuation 11](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066514)

[Hình 9: Vietnamese word segmentation 12](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066515)

[Hình 10: Kết quả preprocessing data 12](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066516)

[Hình 11: Khai báo link video và import thư viện 13](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066517)

[Hình 12: Đọc file dataset – traning data 14](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066518)

[Hình 13: Preprocessing training data 15](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066519)

[Hình 14: Gộp train và validate dataset thành 1 tập mới 16](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066520)

[Hình 15: Trích xuất đặc trưng (feature) 16](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066521)

[Hình 16: Label encoder 17](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066522)

[Hình 17: Logistic Regression model 17](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066523)

[Hình 18: Testing data 18](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066524)

[Hình 19: Classification report for test\_nor\_811.xlsx 19](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066525)

[Hình 20: Đọc file GFG\_1.csv 20](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066526)

[Hình 21: Kết quả sau khi đọc file GFG\_1.csv 20](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066527)

[Hình 22: Trích xuất feature và encoder label GFG\_1.csv 21](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066528)

[Hình 23: Kết quả trích xuất feature và encoder label GFG\_1.csv 21](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066529)

[Hình 24: Áp dụng GFG\_1.csv vào model 22](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066530)

[Hình 25: Performance metrics – GFG\_1.csv 22](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066531)

[Hình 26: Performance metrics GFG\_2.csv 23](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066532)

[Hình 27: Performance metrics GFG\_3.csv 23](file:///E:\DATA%20MINING\ĐỒ%20ÁN%20CUỐI%20KỲ%20-%20DATA%20MINING\demo.docx#_Toc72066533)

**DANH MỤC BẢNG**

[Table 1: Bảng phân công công việc …………………………………………………………..5](#_Toc72066534)

[Table 2: Bảng so sánh kết quả …………………………………………………………24](#_Toc72066535)

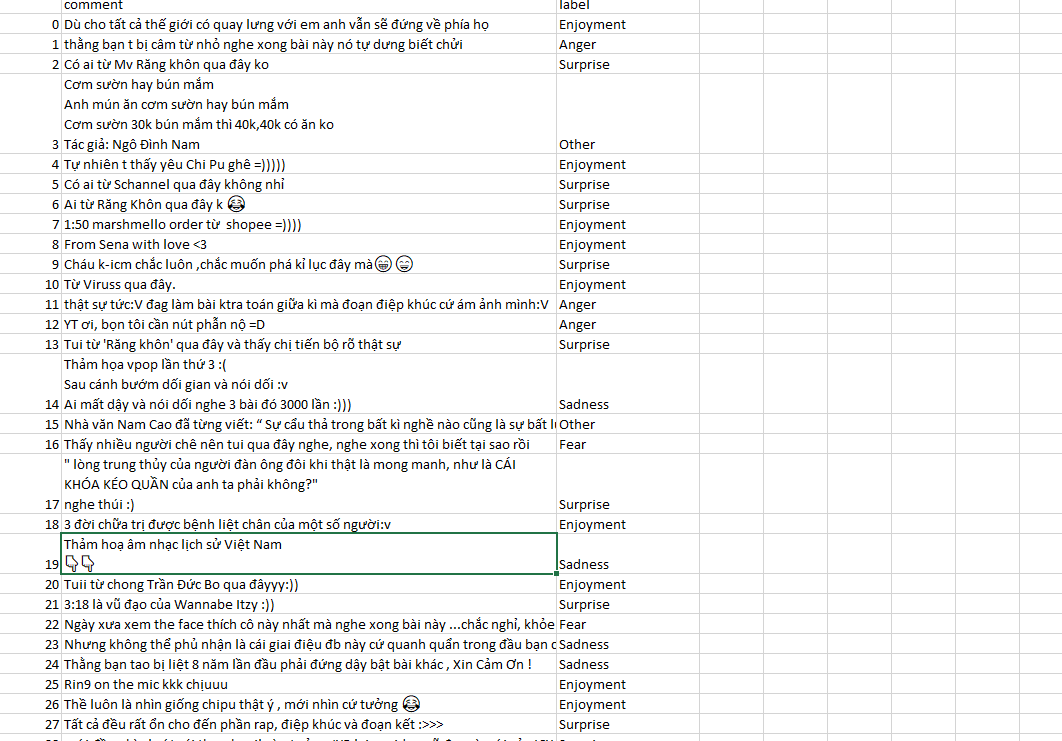
**BẢNG PHÂN CÔNG**

|  |  |  |
| --- | --- | --- |
| **Thành viên** | **Công việc** | **Tiến độ** |
| **Huỳnh Tấn Lợi** | Preprocessing data | Hoàn thành |
| Gắn label - crawl dataset | Hoàn thành |
| Báo cáo | Hoàn thành |
| **Võ Tấn Lực** | Web crawling | Hoàn thành |
| Hiện thực mô hình | Hoàn thành |
| Silde thuyết trình | Hoàn thành |

Table 1: Bảng phân công công việc

CHƯƠNG 1 – THU THẬP DỮ LIỆU

1.1 Giới thiệu

Thu thập dữ liệu dựa trên các comment của Youtube bằng cách crawl dữ liệu. Dữ liệu được lưu trữ dưới dạng file csv hơn 300 comment của 3 video khác nhau trong Youtube. Sau khi crawl dữ liệu thành công, phân tích mức độ cảm xúc chi tiết của từng comment bằng các biểu hiện như: Disgust (chán ghét), Enjoyment (thích thú), Anger (giận dữ), Surprise (ngạc nhiên), Sadness (buồn bã), Fear (sợ hãi) và Other (khác) và gắn label vào cho từng comment tương ứng.

Hình : Ví dụ về việc gắn label cho từng comment sau khi crawl thành công

1.2 Kỹ thuật sử dụng

1.2.1 Selenium

Selenium là một trong những công cụ kiểm thử phần mềm tự động mã nguồn mở (open source test automation tool) cho việc kiểm thử ứng dụng web. Chạy được trên hầu hết các trình duyệt hiện nay như: IE, Mozilla, FireFox, Chrome,… Được sử dụng cho nhiều ứng dụng như tự động hoá các tác vụ quản trị dựa trên Web hay tương tác với các nền tảng không cung cấp Api, cũng như phương hướng tiếp cận trong đề tài này là thu thập thông tin Web (Web crawling).

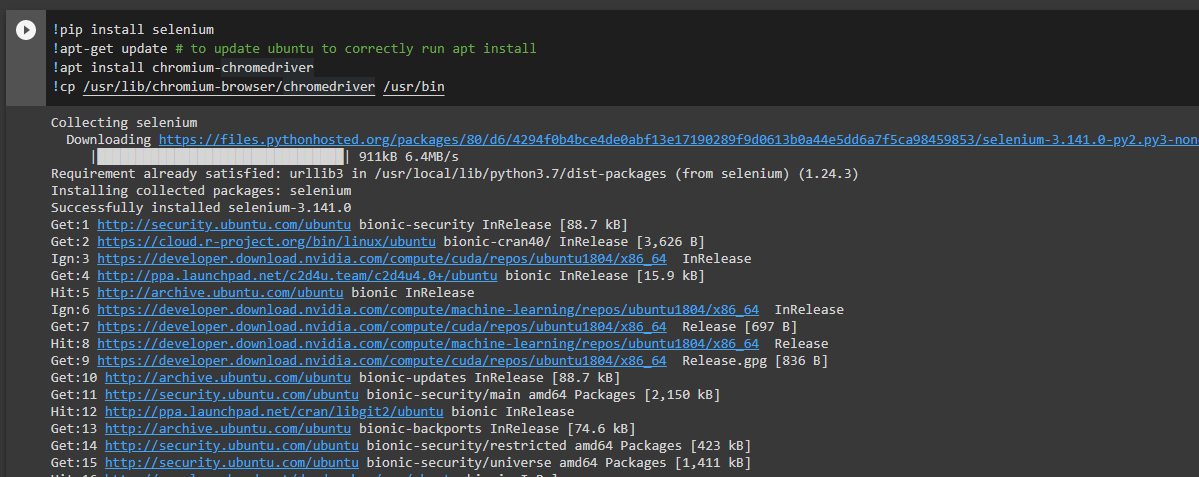
Selenium WebDriver (Selenium 2): là thư viện cho phép người dung lập trình (scripting) test script trên các ngôn ngữ lập trình khác nhau như Python, Java, C#, Ruby…

Selenium Python bindings cung cấp một API đơn giảm để viết functional/acceptance test sử dụng selenium webdriver.

Thông qua Selenium Python API có thể truy cập tất cả các chức năng của selenium webdriver một cách trực quan.

Selenium Python bindings cung cấp một API thuận tiện để truy cập Webdrivers như Firefox, IE, Chrome, Remote. Hiện tại hỗ trợ Python version 2.7 và 3.2.

1.2.2 Cài đặt selenium và chrome drive

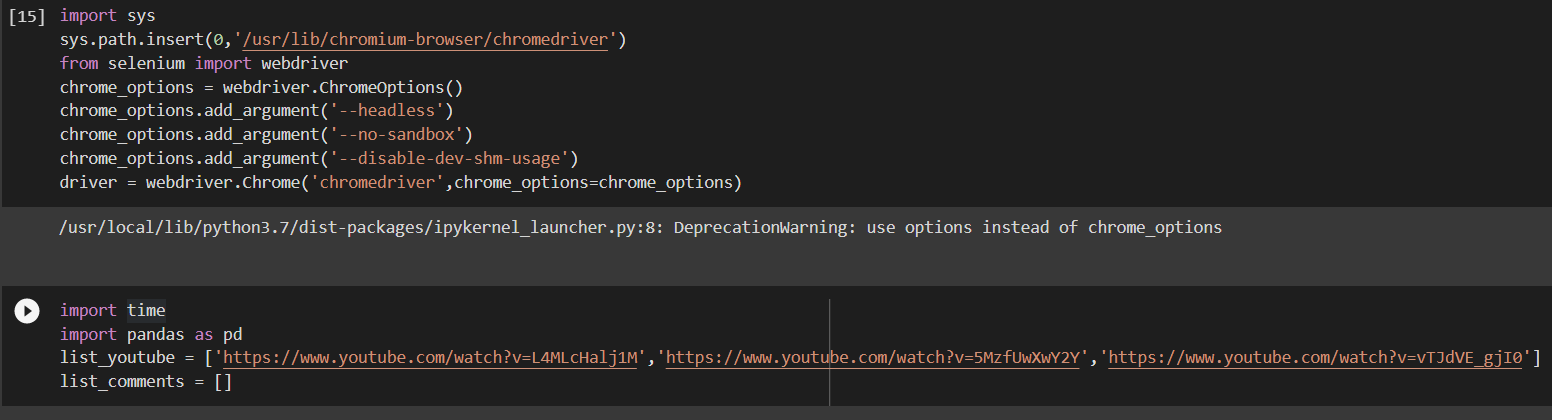
Web crawling sẽ được thực thi trên Google colab, để có thể sử dụng selenium và chrome drive ta chạy lệnh cài đặt:

Hình 2: Lệnh cài đặt selenium và chrome drive trên colab

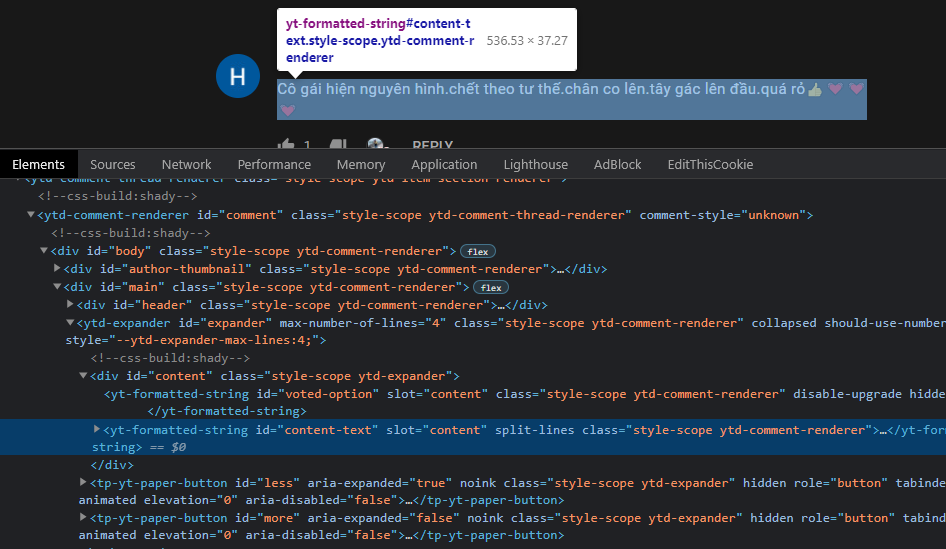
1.3 Thực thi

Dữ liệu comment sẽ được crawl trên 3 link youtube với mỗi tập dữ liệu xấp xỉ hơn 100 comment tương ứng :

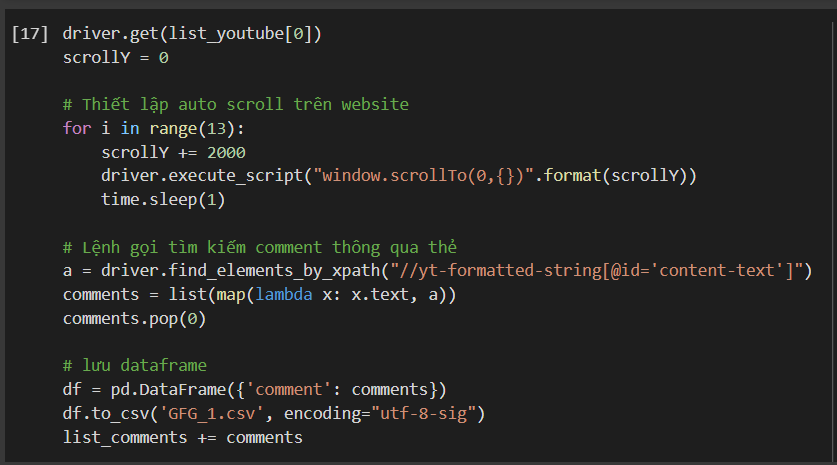
* <https://www.youtube.com/watch?v=L4MLcHalj1M>
* <https://www.youtube.com/watch?v=5MzfUwXwY2Y>
* <https://www.youtube.com/watch?v=vTJdVE_gjI0>

Khai báo và config các library cần thiết:

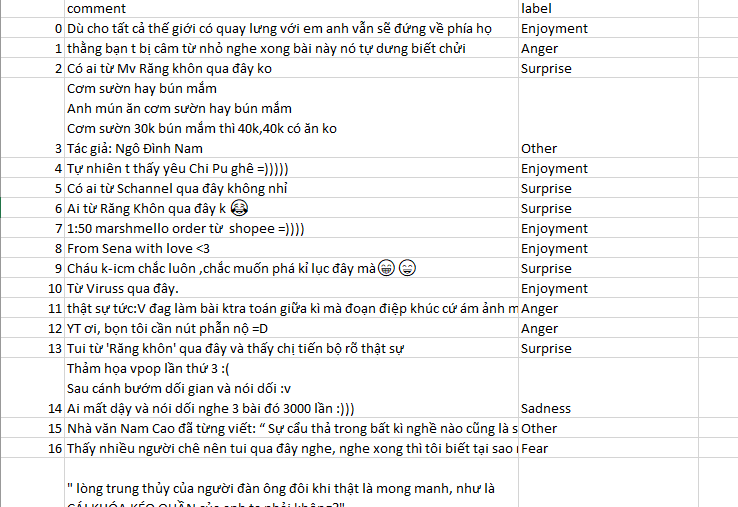
Hình 3: Khai báo và config cái library

Sử dụng selenium truyền link youtube cần crawl comment, sau đó thiết lập scroll chrome để lấy các thẻ chứa comment tương ứng với link youtube đó:

Hình 4: Thẻ youtube

Sau khi xác định được thẻ cần chỏ đến, ta thực hiện scroll và dùng lệnh *find\_elements\_by\_xpath* để tìm kiếm và lấy nội dung comment:

Hình 5: Thực thi crawl comment

Sau khi crawl data thành công, lưu data với tên GFG\_1 với định dạng csv. Tương tự như 2 video khác, ta được 3 tập dataset với hơn 340 comments của 3 video youtube khác nhau, sau đó thực hiện gắn label cảm xúc cho từng comment:

Hình 6: ví dụ crawl data

CHƯƠNG 2 – TIỀN XỬ LÝ VÀ HUẤN LUYỆN MÔ HÌNH (PREPROCESSING AND TRANING MODEL)

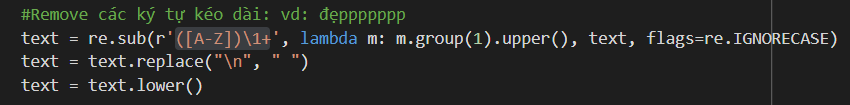
2.1. Tiền xử lý dữ liệu (preprocessing data)

Để cải tiến được mô hình khi train, việc tiền xử lý dữ liệu trước khi đem vào train là một việc hết sức quan trọng. Đây là một kỹ thuật khai thác dữ liệu thô thành một định dạng dễ hiểu. Dự liệu thực tế như trong đề tài này là các comment trên video youtube, chúng thường không đầy đủ, không nhất quán, nhất là trong tập dữ liệu này là Tiếng Việt.

### *2.1.1 Loại bỏ các từ kéo dài*

Trong Tiếng Việt, việc xuất hiện những từ kéo dài thường hay xảy ra nhất là đối với dữ liệu từ nguồn mạng xã hội (Youtube). Từ kéo dài ở đây có nghĩa là những từ bị lặp đi 1 chữ cái duy nhất ở phía sau 1 từ hay cụm từ. Ví dụ: đẹpppppppppppppp

Việc xuất hiện từ kéo dài trong dữ liệu thô sẽ gây ảnh hưởng đến độ chính xác do mô hình dự đoán sau này.

Sử dụng Regular Expression để có thể loại bỏ các từ kéo dài:

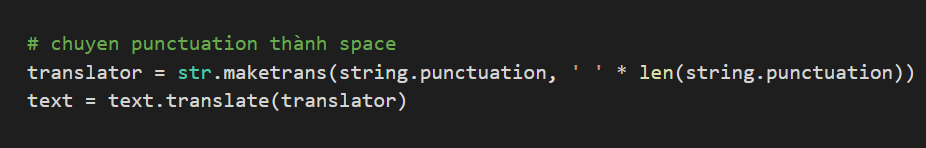
Hình 7: Loại bỏ từ kéo dài

Chuỗi Regular Expression *([A-Z])\1+* : Khớp các từ có character từ A-Z và character cuối không quá 1 ký tự giống nhau.

Sau khi loại bỏ từ kéo dài, tiếp tục replace các đoạn comment xuống hàng thành 1 dòng.

Chuyển comment thành chữ thường.

### *2.1.2 Dấu câu (punctuation)*

Punctuation (phép đánh dấu câu) là hệ thống các ký hiệu (! ; “ ...). Chúng ta sử dụng các ký hiệu này để tách câu và các phần trong câu, và để làm rõ nghĩa hơn. Mỗi ký hiệu được gọi là một “punctuation mark” (dấu câu). Trong quá trình huấn luyện mô hình, việc xuất hiện dấu câu sẽ gây ảnh hưởng đến độ chính xác của mô hình. Vì vậy phải loại bỏ nó :

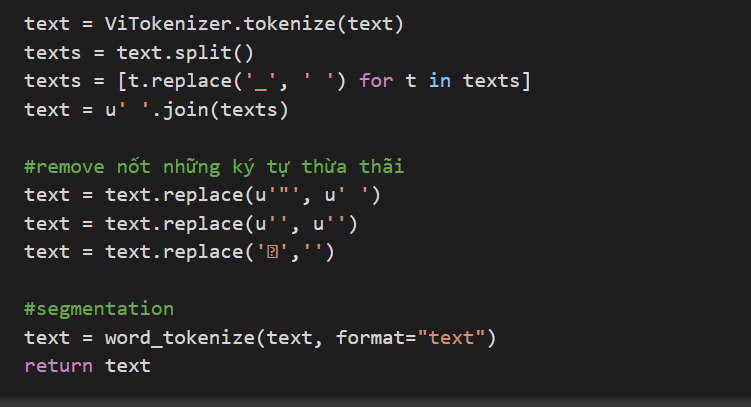
Hình 8: Punctuation

### *2.1.3 Tách từ (Word segmentation)*

Là một bước quan trọng bậc nhất trong xử lý ngôn ngữ tự nhiên. Tiếng Việt không đơn giản như tiếng anh vì nó có thêm các từ ghép. Ví dụ : lý do, hành động, thoải mái, … Đây là các từ ghép, chúng đi với nhau tạo thành 1 từ hoàn chỉnh.

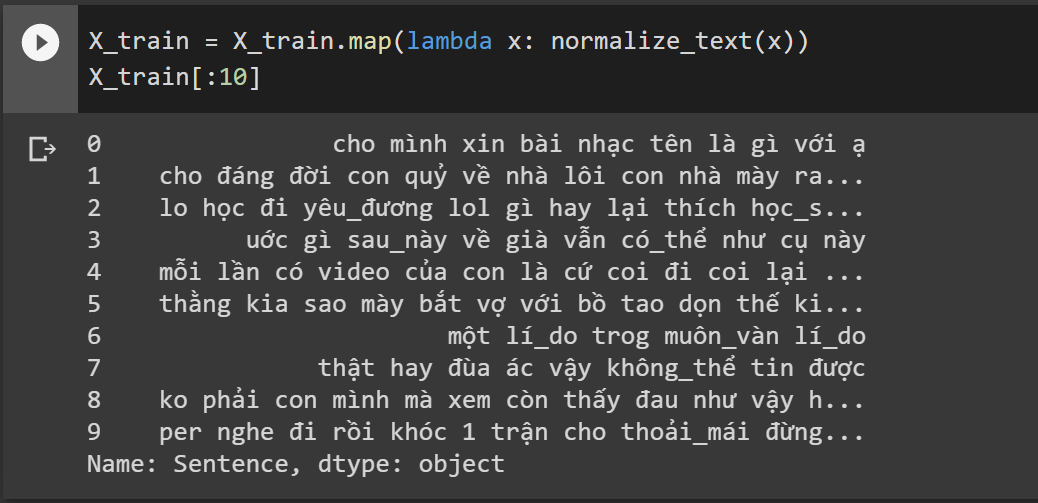
VnTokenizer là gói thư viện hỗ trợ việc phân tách ngữ nghĩa của từ, đưa ra chính xác các cặp từ trong 1 câu :

Lệnh cài đặt: *pip install pyvi*

Áp dụng thư viện vào việc tiền xử lý dữ liệu:

Hình 9: Vietnamese word segmentation

Sau khi áp dụng thư viện ra được các cặp từ ghép, nói chúng với nhau bằng dấu “\_”. Tiếp theo loại bỏ mấy ký tự thừa và áp dụng *work\_tokenize* của thư viện undertheseađể tìm kiếm 1 lần nữa các từ kép.

Sau khi thực hiện các bước trên, dữ liệu nhận được cuối cùng:

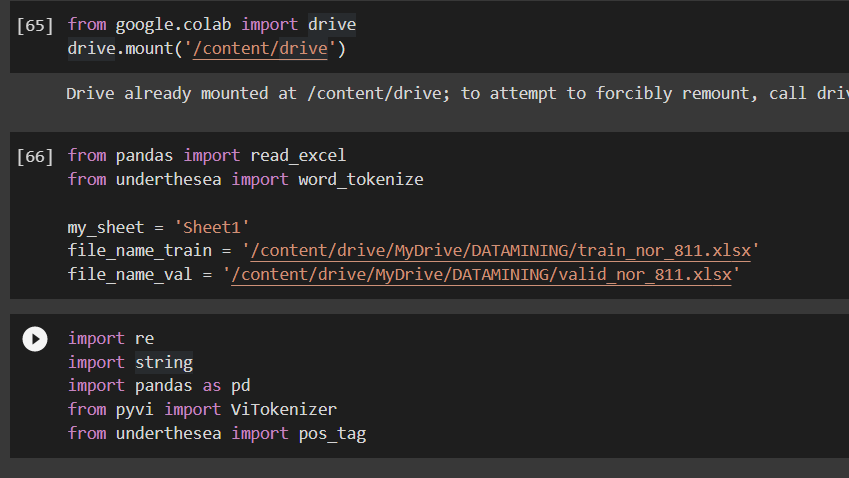
Hình 10: Kết quả preprocessing data

2.2. Huấn luyện mô hình (Training model)

*2.2.1 Đọc file và xử lý dữ liệu*

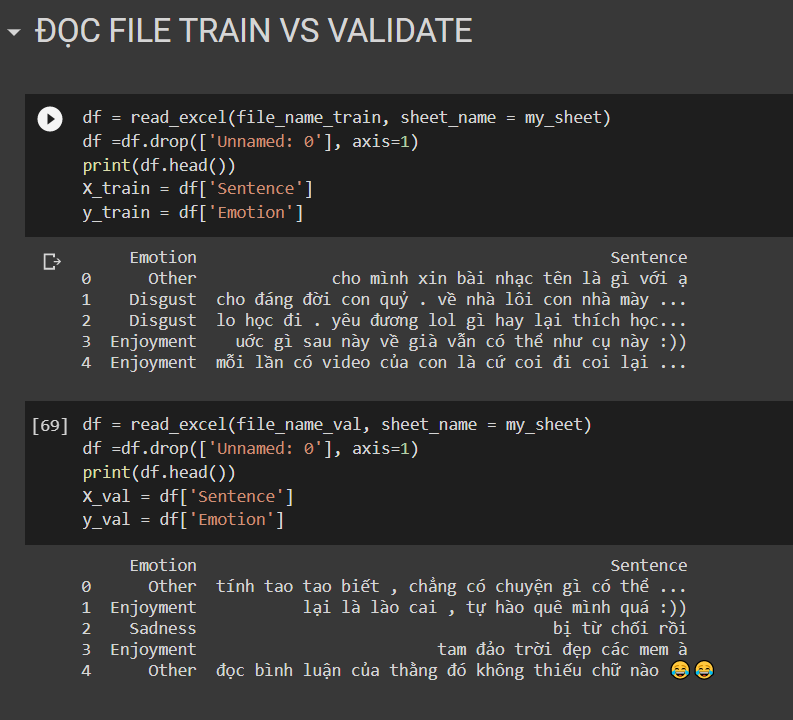
Khai báo đường dẫn lưu file và import các thư viện cần thiết, cụ thể dataset được chọn làm data training:

* train\_nor\_811.xlsx
* valid\_nor\_811.xlsx



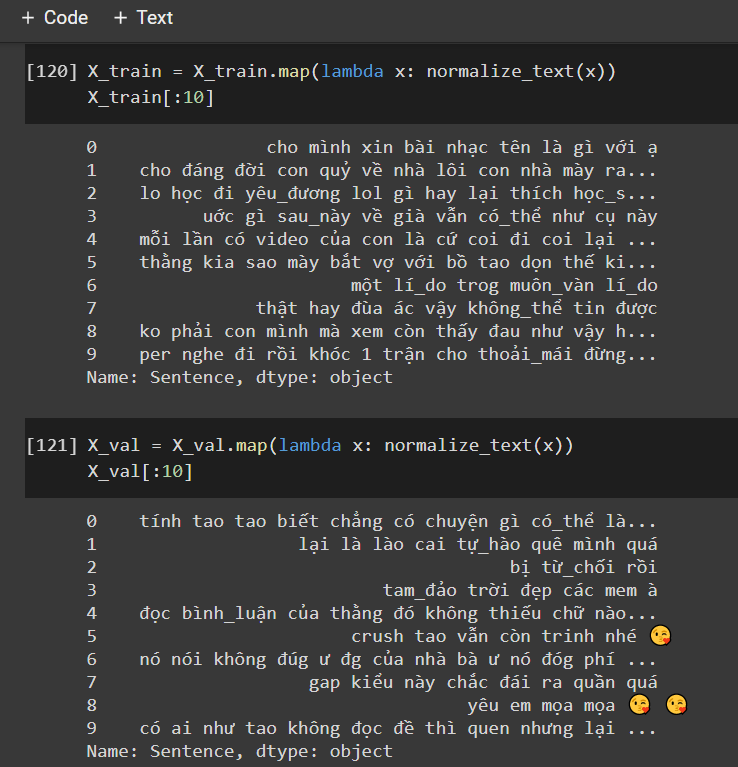
Hình 11: Khai báo link video và import thư viện

Đọc dataset, phân tách label và feature trong 2 tập dataset:



Hình 12: Đọc file dataset – traning data

Tiến hành thực hiện bước tiền xử lý dữ liệu cho tập training dataset:



Hình 13: Preprocessing training data

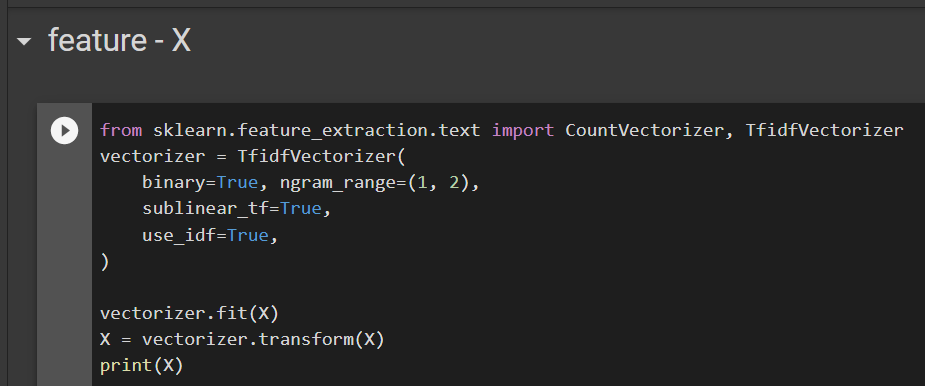
Ở đây áp dụng hàm *normalize\_text(x)* cho từng comment để tiền xử lý dữ liệu đưa ra dữ liệu trực quan nhất cho huấn luyện mô hình (training model).

Sau khi hoàn thành phân label, feature cũng như tiền xử lý dữ liệu cho ra được dữ liệu trực quan nhất, tiến hành gộp 2 dataset lại với nhau thành 1 tập training dataset hoàn chỉnh:

### *2.2.2 TF-IDF vectorizer*

Hình 14: Gộp train và validate dataset thành 1 tập mới

TF-IDF (Term Frequency – Inverse Document Frequency) là 1 kĩ thuật sử dụng trong khai phá dữ liệu văn bản. Trọng số này được sử dụng để đánh giá tầm quan trọng của một từ trong một văn bản. Giá trị cao thể hiện độ quan trọng cao và nó phụ thuộc vào số lần từ xuất hiện trong văn bản nhưng bù lại bởi tần suất của từ đó trong tập dữ liệu.

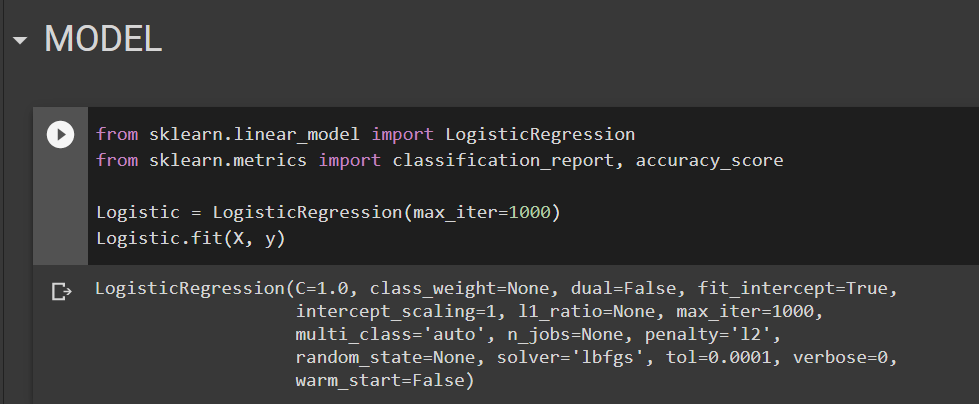
Áp dụng TF-IDF vào data để chuyển data thành dạng vecor:

Hình 15: Trích xuất đặc trưng (feature)

Sau khi hoàn thành, tiếp tục encoder label thành vector:

### *2.2.3 Huấn luyện mô hình – Logistic Regression*

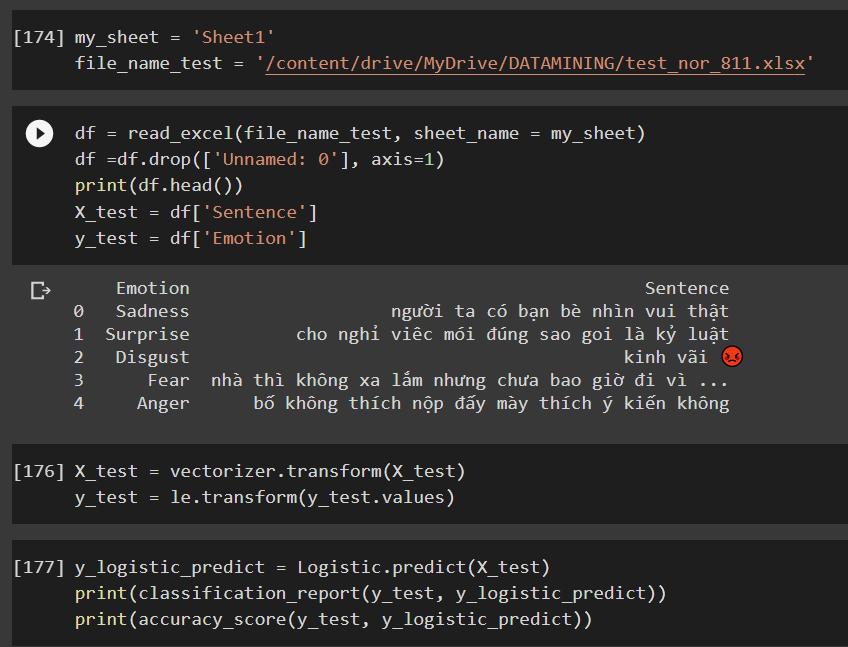
Hình 16: Label encoder

Khai báo thư viện sklearn – Logistic Regression, tiến hành training dựa trên tập dữ liệu sau khi được xử lý và trích xuất đặc trưng:

Hình 17: Logistic Regression model

## 2.3 Dữ liệu thử nghiệm (Testing data)

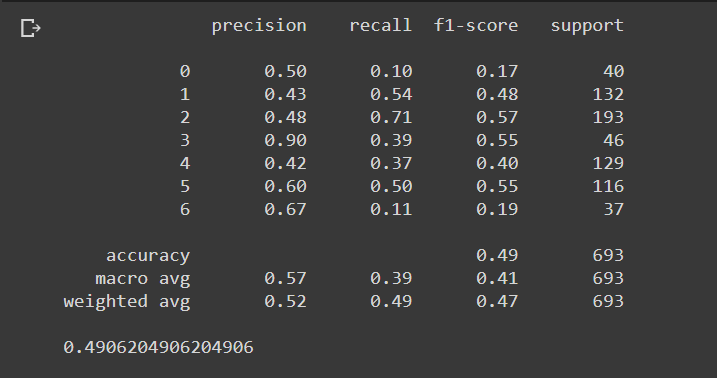
Dữ liệu thử nghiệm được cung cấp sẳn trong đề tài này, tiến hành đọc file và thực hiện các bước phân tách feature với label, trích xuất feature và encoder label tương tự các mục trên:

* test\_nor\_811.xlsx

Hình 18: Testing data

Thực hiện truyền feature *X* testing data vào model => *y\_logistic\_predict*

Kết quả:

Độ chính xác sau khi test data đạt 49%

Hình 19: Classification report for test\_nor\_811.xlsx

CHƯƠNG 3 - THỰC NGHIỆM TRÊN DỮ LIỆU CRAWL

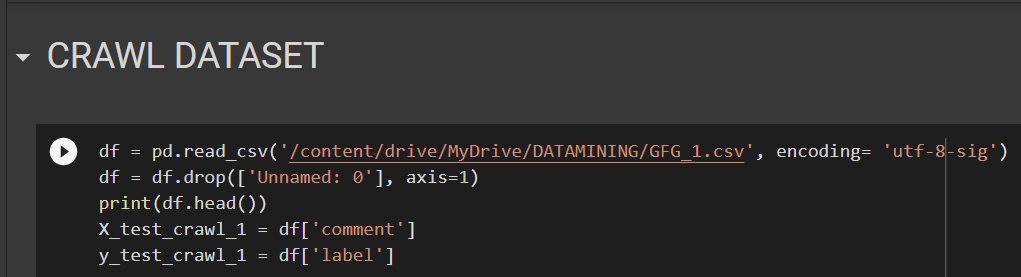
3.1. Tổng quan về tập dữ liệu

Tập dữ liệu được crawl từ nguồn các comment video trên Youtube, phân thành 3 tập dữ liệu ứng với 3 video trên youtube:

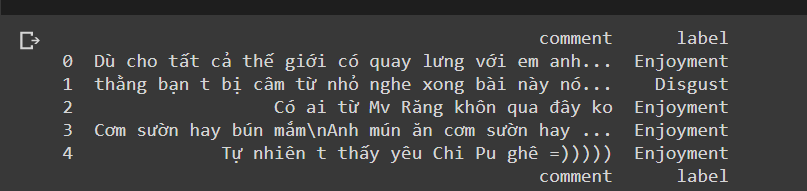
* GFG\_1.csv: 119 comment
* GFG\_2.csv: 108 comment
* GFG\_3.csv: 110 comment

Tất cả tập dữ liệu đã được gán label cảm xúc cho tương ứng từng nội dung comment.

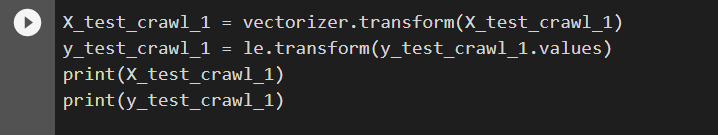
3.2 Thực nghiệm

**Tương tự như tập dữ liệu thực nghiệm ở Chương 2, chúng ta bắt đầu đọc file GFG\_1.csv thông qua google drive và sau đó phân feature với label:

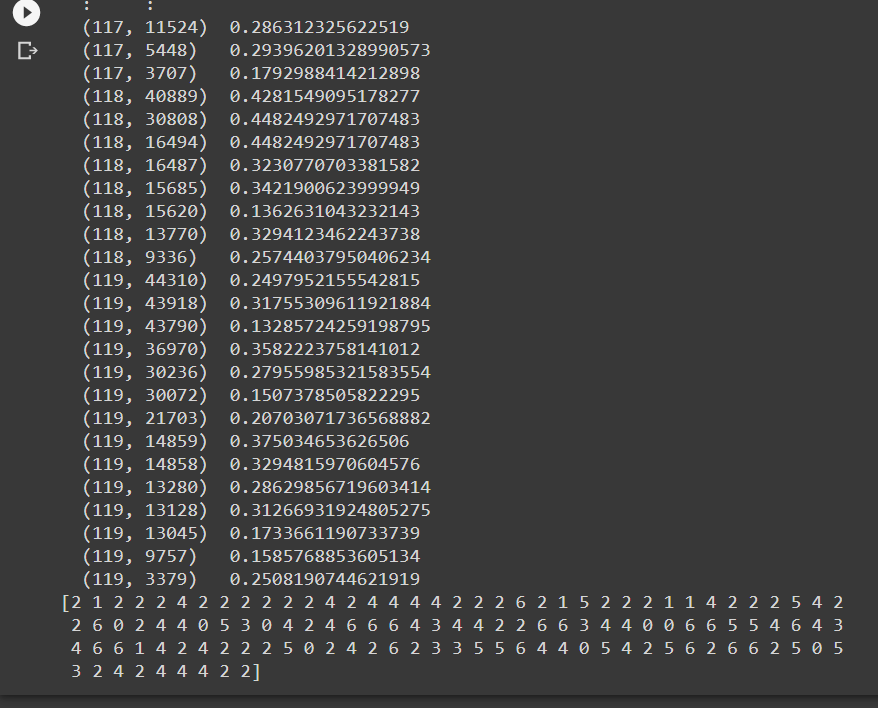
Hình : Đọc file GFG\_1.csv

Kết quả:

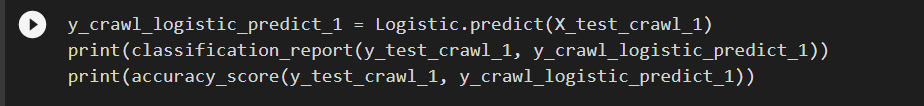
Hình : Kết quả sau khi đọc file GFG\_1.csv

Sau khi hoàn thành đọc file và phân chia dữ liệu, tiến hành trích xuất feature và encoder label thành các vector:

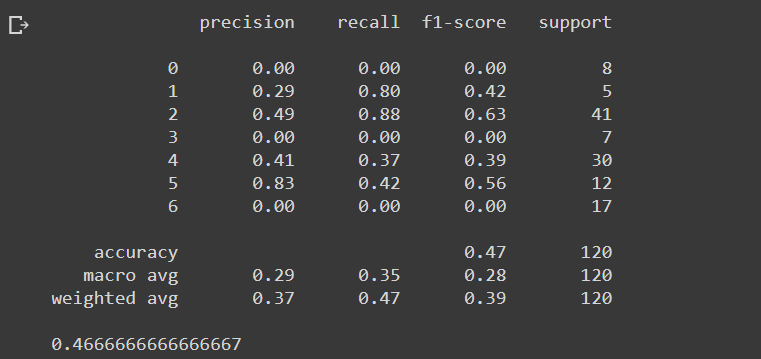
Hình 22: Trích xuất feature và encoder label GFG\_1.csv

Kết quả nhận được:

Hình 23: Kết quả trích xuất feature và encoder label GFG\_1.csv

Sau khi hoàn thành tất cả các bước trên, bước cuối cùng chúng ta thêm dữ liệu test vào mô hình:

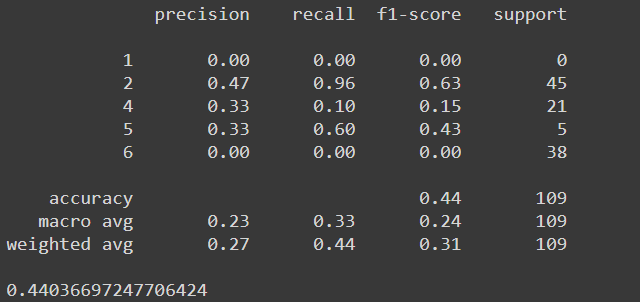
Hình : Áp dụng GFG\_1.csv vào model

Kết quả: độ chính xác xấp xỉ 47%

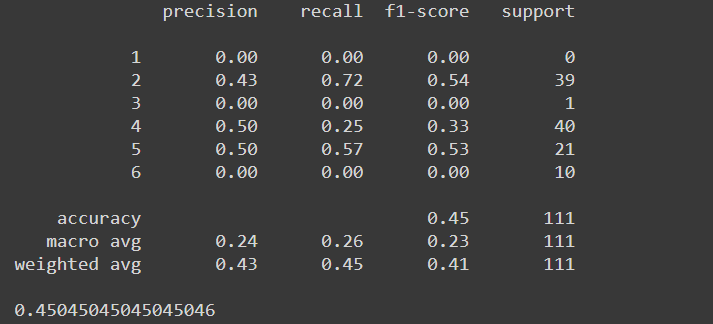
Hình : Performance metrics – GFG\_1.csv

Tương tự như GFG\_1.csv, chúng ta chạy lần lượt GFG\_2.csv và GFG\_3.csv cho ra được kết quả.

GFG\_2.csv:



Hình : Performance metrics GFG\_2.csv

GFG\_3.csv:

Hình : Performance metrics GFG\_3.csv

## 3.3 Bảng so sánh kết quả

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **test\_nor\_811.xlsx** | **GFG\_1.csv** | **GFG\_2.csv** | **GFG\_3.csv** |
| **Logistic Regression** | 49% | 47% | 44% | 45% |

Table : Bảng so sánh kết quả

# TÀI LIỆU THAM KHẢO

1. https://www.geeksforgeeks.org/string-punctuation-in-python
2. https://pypi.org/project/pyvi
3. https://underthesea.readthedocs.io/en/develop/package\_reference.html
4. <https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html>
5. <https://viblo.asia/p/bai-1-gioi-thieu-ve-selenium-aWj538VwK6m>
6. <https://ieeexplore.ieee.org/abstract/document/9067730>
7. <https://www.programiz.com/python-programming/methods/string/maketrans>
8. https://viblo.asia/p/phan-loai-van-ban-tieng-viet-tu-dong-phan-1-yMnKM3bal7P