**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN MÔN TÍNH TOÁN ĐA PHƯƠNG TIỆN**

*ĐỀ TÀI:* **PHÂN TÍCH BÌNH LUẬN CỦA MỘT TRANG THƯƠNG MẠI ĐIỆN TỬ**

GVPT: Lê Đình Duy Lớp: CS114.K21

Phạm Nguyễn Trường An 18521060 – Trịnh Hưng Long

18521062 – Hà Văn Luân

18521274 – Lỡ Đình Phương

*TPHCM, ngày 6 tháng 8 năm 2020*

**MỤC LỤC**

[1 Tổng quan 1](#_Toc47843680)

[1.1 Thông tin nhóm 1](#_Toc47843681)

[1.2 Giới thiệu đề tài 1](#_Toc47843682)

[2 Đồ án 2](#_Toc47843683)

[2.1 Mô tả đồ án 2](#_Toc47843684)

[2.1.1 Ý tưởng 2](#_Toc47843685)

[2.2.2 Input & Output 2](#_Toc47843686)

[2.2 Chuẩn bị dữ liệu *(Prepare Dataset)* 2](#_Toc47843687)

[2.2 Tiền xử lý dữ liệu *(Data Preprocessing)* 4](#_Toc47843688)

[2.3 Xây dựng và huấn luyện model (Choosing and Training model) 8](#_Toc47843689)

[2.4 Thiết kế giao diện cho người dùng 13](#_Toc47843690)

[2.5 Kết quả thực nghiệm 15](#_Toc47843691)

[3 Tổng kết 21](#_Toc47843692)

[3.1 Những điểm hạn chế 21](#_Toc47843693)

[3.2 Hướng dẫn sử dụng source code 21](#_Toc47843694)

[4 Tài liệu tham khảo 24](#_Toc47843695)

[5 Bảng phân công 25](#_Toc47843696)

# Tổng quan

## **Thông tin nhóm**

|  |  |  |
| --- | --- | --- |
| STT | Mã số sinh viên | Họ và Tên |
| 1 | 18521060 | Trịnh Hưng Long |
| 2 | 18521062 | Hà Văn Luân |
| 3 | 18521274 | Lỡ Đình Phương |

## **Giới thiệu đề tài**

* Hiện nay nhu cầu mua hàng qua mạng của người dùng ngày càng trở nên phát triển mạnh hơn do những lợi ích mà nó mang lại như tiện lợi, chi phí rẻ, có nhiều chương trình khuyến mãi hấp dẫn, có thể ngồi ở nhà để xem sản phẩm mà không cần phải đến tận nơi để xem, … Tuy nhiên, việc mua hàng qua mạng cũng có những nhược điểm, trong đó có việc người dùng không thể tận mắt đánh giá sản phẩm của mình như mua trực tiếp tại các cửa hàng được. Vì vậy, các mục bình luận về sản phẩm của những người dùng đã sử dụng qua sản phẩm đóng vai trò quan trọng trong việc đánh giá chất lượng các sản phẩm tương ứng, các bình luận chủ yếu gồm 3 loại: tích cực, trung tính hoặc tiêu cực. Tuy nhiên, số lượng các bình luận trên các trang thương mại điện tử rất nhiều, gây khó khăn khi đánh giá từng bình luận bằng tay, vì vậy, các thuật toán máy học sẽ hỗ trợ đắc lực cho việc phân loại này.
* Để minh chứng cho đề tài này, chúng em sẽ tập trung phân loại các sản phẩm là "điện thoại" từ các bình luận trên một trang thương mại điện tử cụ thể đó là trang “*Thế Giới Di Động”*, hỗ trợ cho việc mua sắm điện thoại của người tiêu dùng ngày nay.
* Ngôn ngữ sử dụng: Python.

# 

# Đồ án

## **2.1 Mô tả đồ án**

### **2.1.1 Ý tưởng**

* Đưa một đoạn bình luận của 1 sản phẩm bất kỳ và đưa ra kết quả dự đoán bình luận (là tích cực, tiêu cực hoặc trung tính).

### **2.2.2 Input & Output**

* + - Input: Một bình luận về sản phẩm điện thoại của khách hàng bằng tiếng Việt.
    - Output: Bình luận đó là tiêu cực, trung tính hay tích cực (3 class).

## **2.2 Chuẩn bị dữ liệu *(Prepare Dataset)***

* Crawl data (các bình luận sản phẩm trên trang) từ trang thương mại điện tử: *'*[*https://www.thegioididong.com/dtdd*](https://www.thegioididong.com/dtdd)*’  (Thế Giới Di Động)* bằng thư viện BeautifulSoup – 1 package Python dùng để phân tích cú pháp các tài liệu HTML và XML.
* Ta thu về được 4679 bình luận (file new\_data\_machinelearning.json) với 4 features (‘TYPE’, ‘NAME’, ‘COMMENT’, ‘RATING COMMENT’) các nhãn được gán tự động theo các tiêu chí với:
* Các bình luận có số lượng đạt 5 sao là 5 sẽ cho nhãn bằng 1.
* Các bình luận có số lượng đạt 5 sao là 4 và 3 sẽ cho nhãn bằng 0.
* Các bình luận có số lượng đạt 5 sao là 1 và 2 sẽ cho nhãn bằng -1.
* Sau đó, tụi em sẽ thực hiện kiểm tra lại các nhãn các bình luận của bộ data trên bằng tay (với mỗi bạn kiểm tra xấp sỉ khoảng 1500 bình luận), thu gọn và làm cân bằng bộ data trên -> thu về được 3000 bình luận (1000 bình luận tích cực, 1000 bình luận trung tính và 1000 bình luận tiêu cực) – file new\_data-1-1000 (1).json.

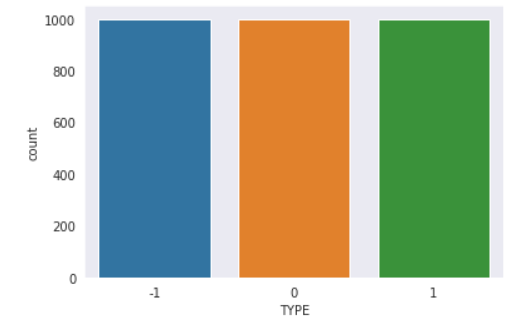
 

*File chứa 4679 bình luận sau khi File chứa 3000 bình luận sau*

*thực hiện crawl data với các khi thực hiện kiểm tra lại*

*nhãn được gán tự động theo bằng tay các nhãn, cân bằng*

*các tiêu chí trên. thu gọn lại (1000 tích cực, 1000 tiêu cực, 1000 trung tính).*



*Biểu đồ minh họa bộ data hoàn chỉnh*

\* Link colab: <https://colab.research.google.com/drive/1ApzLOcKpevmINQXFR9llo15pFWLjjMHa?usp=sharing>

## **2.2 Tiền xử lý dữ liệu *(Data Preprocessing)***

* Import các thư viện cần thiết hỗ trợ cho việc xử lí dữ liệu:

Thư viện re (Regular Expression) dùng để so khớp các chuỗi hoặc một tập các chuỗi.

Thư viện underthesea được phát triển bởi nhóm nghiên cứu xử lý ngôn ngữ tự nhiên tiếng Việt của tác giả chính là *Vũ Anh*.



* Đưa đoạn text về chữ thường (lower):

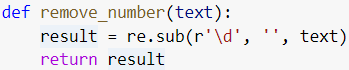


* Loại bỏ các con số có trong đoạn text:

re.sub: 1 phương thức có tác dụng so khớp và thay thế chuỗi so khớp được.

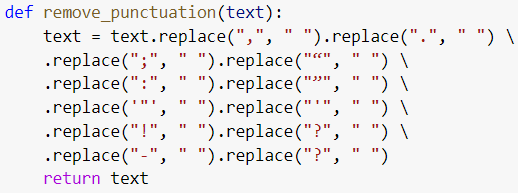
Với: r’\d’ tương ứng với bất kỳ chữ số thập phân Unicode nào [0-9].

* + Thay thế các chữ số tìm được bằng ‘’.



* Loại bỏ các dấu câu có trong đoạn text:

Dùng hàm .replace có sẵn trong python để thay thế các dấu câu tìm được bằng “ ”.



* Loại bỏ các khoảng trắng thừa có trong đoạn text:

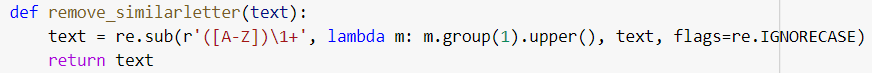
Dùng hàm .split() để chuyển chuỗi text thành một list cắt theo separator (dấu phân tách) – separator để trống mặc định là khoảng cách.

Sau đó dùng hàm .join() để chuyển list về chuỗi - các phần tử cách nhau bởi một khoảng cách “ ”.



* Loại bỏ các kí tự giống nhau liên tiếp có trong đoạn text:

VD: quaaaa -> qua, kkkkkkkk -> k,…



* Tách từ Tiếng Việt sử dụng thư viện *underthesea* :

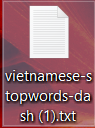


* Loại bỏ stopwords tiếng việt có trong đoạn text:

Stopwords là những từ xuất hiện rất nhiều trong các bài viết, các đoạn text nhưng lại không hề liên quan gì đến nội dung và ý nghĩa của bài viết, gây mơ hồ, làm quá trình máy học, phân loại giảm đi độ chính xác.

Trong quá trình tìm kiếm, chúng em thu thập được 1 file vietnamese-stopwords-dash (1).txt chứa một list những từ stopwords Tiếng Việt của tác giả *Lê Văn Duyệt* và chúng em có tinh chỉnh lại file cho phù hợp với đồ án lần này.

Link : [*https://github.com/stopwords/vietnamese-stopwords/blob/master/vietnamese-stopwords-dash.txt*](https://github.com/stopwords/vietnamese-stopwords/blob/master/vietnamese-stopwords-dash.txt)

 A screenshot of text

Description automatically generated

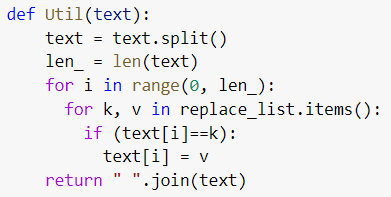
A screenshot of a cell phone

Description automatically generated

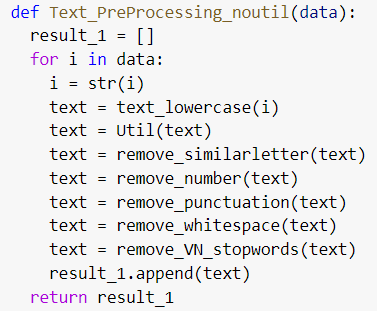
* Chuẩn hóa dữ liệu:
* Việc chuẩn hóa là một công đoạn hết sức cần thiết, vì bộ data chúng em thu thập là các bình luận khá là thông thường, ngẫu hứng (dữ liệu chưa sạch) trên trang thương mại điện tử, việc xuất hiện các teencode, viết tắt,… là một chuyện hết sức bình thường.
* Trong lúc thu thập dữ liệu, chúng em thu thập được một dict chứa các teencode, viết tắt,… Trong quá trình xử lí dữ liệu, sẽ thực hiện tìm trong các bình luận nếu chứa các từ giống với key của phần tử trong replace\_list ,ta gán giá trị từ đó bằng value của key tương ứng.



*replace\_list() thu thập được*



* Sau đó chúng em tổng hợp các hàm về một hàm xử lý dữ liệu (Text\_PreProcessing\_noutil) để thuận tiện cho việc sử dụng:



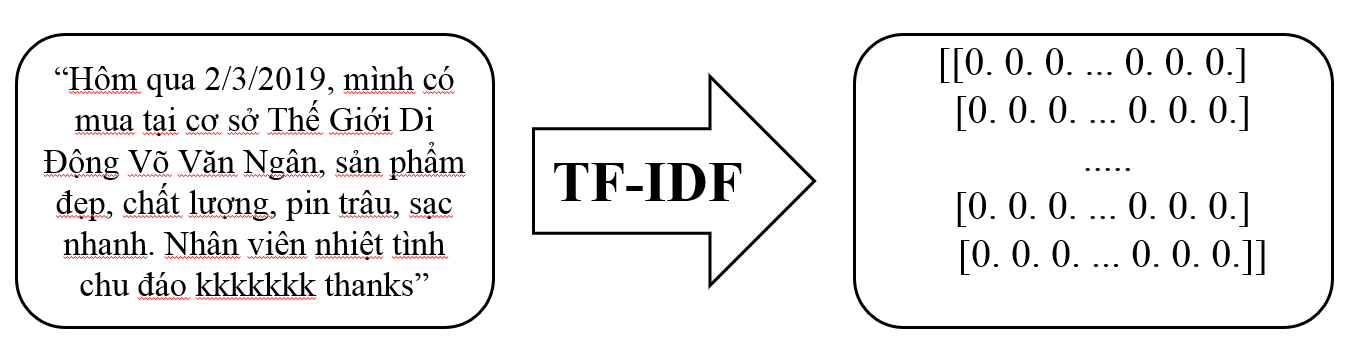
* TF-IDF (Term Frequency - Inverse Document Frequency):

TF: tần số xuất hiện của 1 từ trong 1 văn bản

IDF: tần số nghịch của 1 từ trong một tập các văn bản

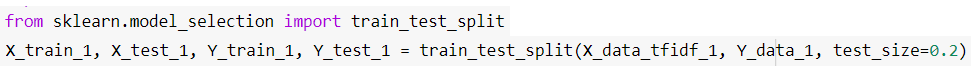
Kĩ thuật TF-IDF dùng để tính toán mức độ quan trọng của từ trong một văn bản.

TfidfVectorizer dùng để chuyển đổi dữ liệu văn bản sang ma trận các features TF-IDF.



## **2.3 Xây dựng và huấn luyện model (Choosing and Training model)**

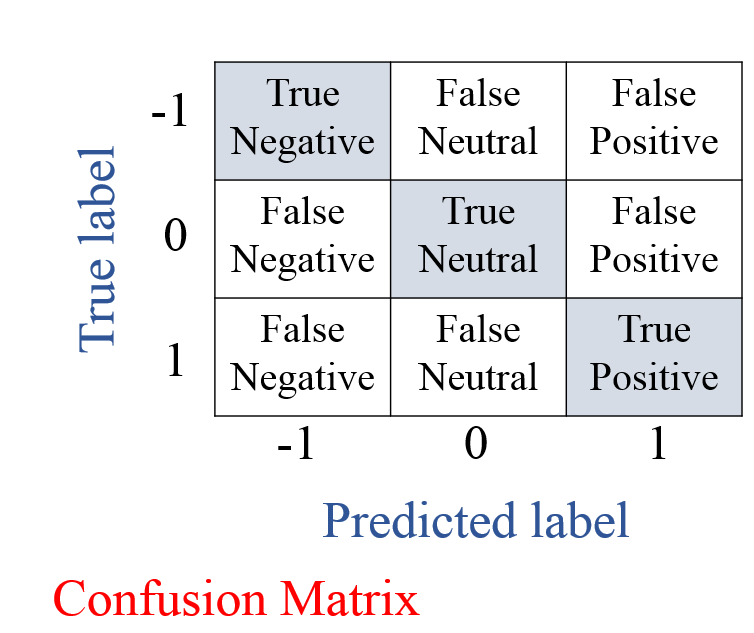
* Train\_test\_split: Chia dữ liệu (dataset) thành train set và test set để huấn luyện và thử nghiệm trên tập dữ liệu thu thập được theo tỉ lệ train/test ứng với 80/20.



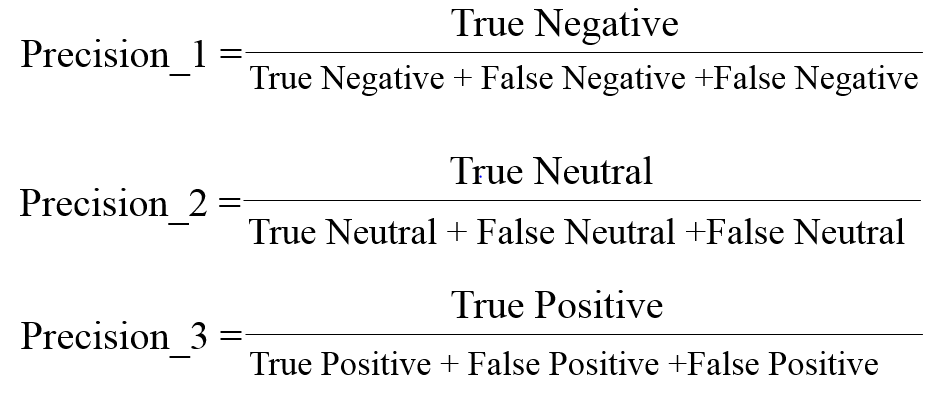
* Đánh giá model:



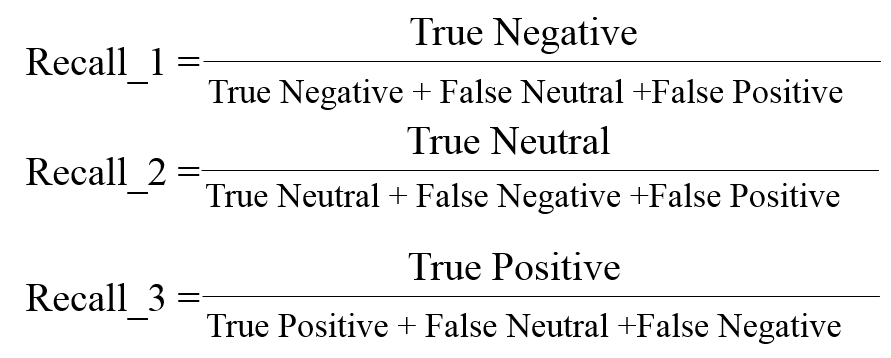
* Cách tính độ chính xác của model bằng **score** thông thường chỉ cho ta biết phần trăm dữ liệu được phân loại đúng mà không chỉ ra được dữ liệu được phân loại như thế nào, nên ta sử dụng một ma trận được gọi là **confusion matrix.**
* **Confusion matrix** giúp ta có cái nhìn chi tiết hơn trong quá trình chọn lọc model dựa trên tập dữ liệu có sẵn.
* Để đánh giá chất lượng của model, ta sử dụng khái niệm *F1-score*, khái niệm này dựa trên 2 khái niệm khác là *Precision* và *Recall.*
* Bài toán lần này có 3 class (tích cực, tiêu cực và trung tính) nên sẽ có *True/False Positive, True/False Negative, True/False Neutral.*



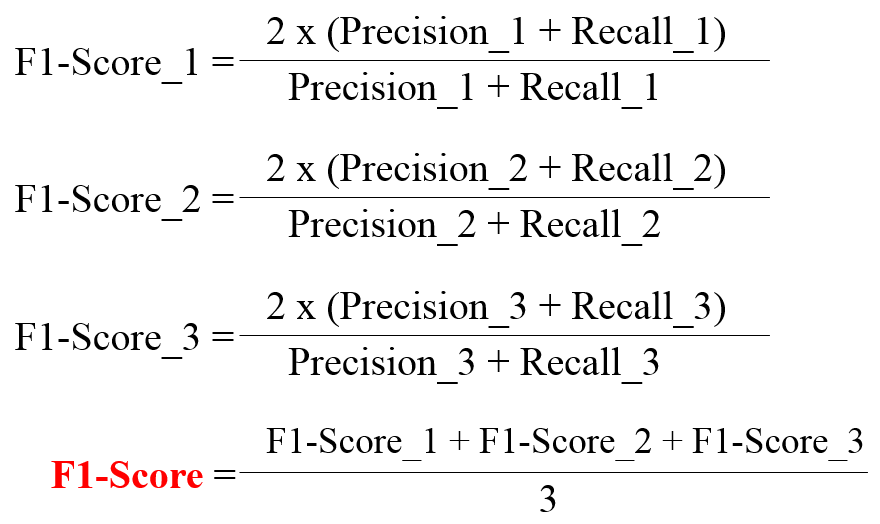
* Precision\_1: là tỉ lệ số điểm True Negative trong số những điểm được phân loại là Negative.
* Precision\_2: là tỉ lệ số điểm True Neutral trong số những điểm được phân loại là Neutral.
* Precision\_3: là tỉ lệ số điểm True Positive trong số những điểm được phân loại là Positive.



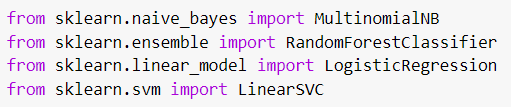
* + Recall\_1: là tỉ lệ số điểm True Negative trong số những điểm thực sự là Negative.
  + Recall\_2: là tỉ lệ số điểm True Neutral trong số những điểm thực sự là Neutral.
  + Recall\_3: là tỉ lệ số điểm True Positive trong số những điểm thực sự là Positive.



* F1-Score\_1 là hàm harmonic mean của Precision\_1 và Recall\_1.
* F1-Score\_2 là hàm harmonic mean của Precision\_2 và Recall\_2.
* F1-Score\_3 là hàm harmonic mean của Precision\_3 và Recall\_3.



* Tiến hành thử nghiệm với các model khác nhau để tìm được một model tốt nhất, phù hợp nhất cho đồ án lần này thông qua *F1 Score* trong quá trình training model.
* Chúng em nhắm tới 4 model:
* SVC.
* Multinomial Naïve Bayes.
* Logistic Regression.
* Random Forest.



* Kết quả thu về được:

\* Link colab: <https://colab.research.google.com/drive/1BfdrT8tM_aTOLOyD9v6PaEhh5bsrcRpq?usp=sharing&fbclid=IwAR3S2uA5G9e6tA0PJcqDjX6J02yUQf05DikwE-XehcliDH6JHCjknZvAcsQ#scrollTo=PsOD6b2EqC3Y>

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

* **Nhận xét:**

*F1 Score* của model SVC ~ 0.7447

*F1 Score* của model Multinomial Naïve Bayes ~ 0.7414

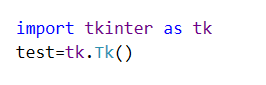
*F1 Score* của model Logistic Regression ~ 0.7304

*F1 Score* của model Random Forest ~ 0.6489

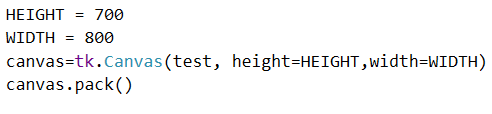
* + *F1 Score* của model SVC đạt giá trị cao nhất, chọn model này xây dựng ứng dụng.

## **2.4 Thiết kế giao diện cho người dùng**

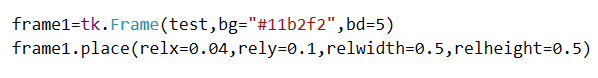
* Tiến hành import Tkinter – một package trong python có chứa module Tk hỗ trợ cho việc lập trình GUI:



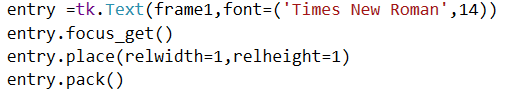
* .title(“tên giao diện”) : đặt tên cho giao diện lập trình.
* Để thiết kế một giao diện hoàn thiện, ta sử dụng đến các khái niệm canvas, frame, label, text và button,… ngay trong thư viện tkinter:
* Canvas: dùng để vẽ các hình học không gian nhất định (hình chữ nhật, hình vuông, hình tròn,…) cho giao diện:
  + - .Canvas (parameter, height= “chiều cao mong muốn”, width= “chiều rộng mong muốn”).
    - .pack() để hiển thị lên giao diện.
    - Ta tạo một hình chữ nhật với chiều cao = 700 và chiều rộng = 800 làm một khung chính cho toàn bộ giao diện.



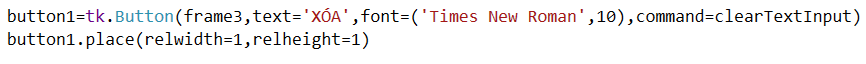
* Frame: dùng để tạo khung hỗ trợ thuận tiện cho việc đưa các label, entry, text, button,… lên giao diện dễ dàng:
  + - .Frame(parameter, bg=”màu background mong muốn”).
    - Hàm .place: vị trí mong muốn trên giao diện (tính theo tọa độ).
    - .place(relx=”hoành độ điểm xuất phát”, rely=”tung độ điểm xuất phát”, relwidth=”chiều rộng mong muốn ”,relheight=”chiều cao mong muốn ”.



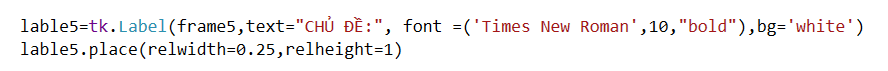
* Text: tạo một Text trên giao diện:
  + - .Text(parameter, font=(‘Tên front’, cỡ chữ)).



* Button: tạo một nút bấm trên giao diện:
  + - .Button(parameter, text=”tên nút”, font=(“tên font”,cỡ chữ), command=”một hàm ta truyền vào để thực hiện sau khi thao tác trên nút bấm này”).



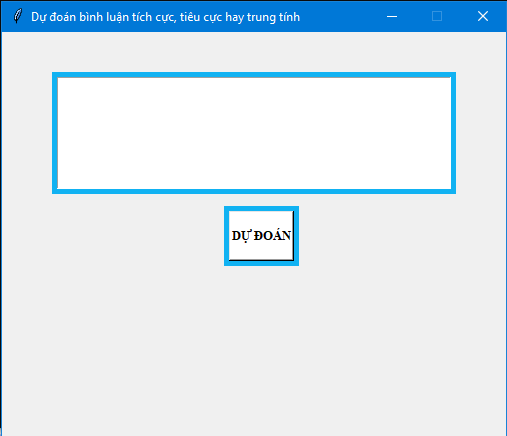
* Label: tạo một nhãn trên giao diện (thường được dùng để hiển thị kết quả):
  + - .Lable(parameter,text=”tên nhãn”, font=(“tên font”,cỡ chữ), bg=” màu background mong muốn).



* <Tên giao diện>.mainloop():

Dừng giao diện lại xem kết quả.



****

*Hình ảnh giao diện thiết kế được*

## **2.5 Kết quả thực nghiệm**

Thử nghiệm với 5 bình luận đánh giá sản phẩm bất kì với model SVC được chọn.

\*Các bình luận dưới đây được lấy từ trang web tiki.vn.

**Bình luận 1**: Đầu tiên phải nói về khâu giao hàng, mình đặt hàng từ tối hôm trước thế là sang chiều hôm sau là đã giao tới rồi giao rất nhanh( giao hàng tiêu chuẩn ). Hàng nguyên seal full box kèm phụ kiện như là cáp sạc nhanh Vooc, dây sạc và ốp lưng bằng nhựa. Trải nghiệm ban đầu về máy phải nói là máy rất mượt, cấu hình chơi game mạnh và đặc biệt là pin rất khoẻ. Do có sạc nhanh nên thời gian sạc cũng nhanh, pin vừa khoẻ mà sạc lại nhanh thì còn gì bằng :)). Máy bắt wifi tốt và màn hình 6.3 inch mang lại trải nghiệm tốt khi xem phim. Đối với ai chỉ cần chụp hình thông thường selfie các kiểu thì sẽ thấy máy ảnh của realme 3 pro thật sự tốt, với giá 3tr750 thì đây là một chiếc điện thoại phải nói là thật sự xứng đáng để bỏ tiền mua. Dưới đây là một số tấm hình mình chụp bằng Realme 3 Pro

-Sản phẩm: Realme 3 Pro (128GB/6GB)

-Khách hàng: Hoàng Tấn Đức

-Kết quả:



**Bình luận 2**: Đây là lần thứ hai mình mua Iphone trên Tiki. Về sản phẩm thì mình tin tưởng Tiki, sản phẩm chính hãng Apple. Mã VN/A. Mình đang dùng thử, mới được 1 tuần. Hi vọng sản phẩm sẽ không có lỗi gì. Nhưng mình khá thất vọng vì việc giao hàng lần này. Thời gian giao hàng quá lâu. Nhân viên còn định gửi hàng cho cái quầy tạp hóa dưới tòa nhà mình ở, cầm hộ khi mình bảo chờ mình 10' (Vì giao hàng không có đúng ngày hẹn ban đầu nên mình phải vòng từ cty về)'.

-Sản phẩm: Iphone 7 Plus 128GB

-Khách hàng: Phạm Bá Hiếu

-Kết quả:



**Bình luận 3**: Mình mua cho mẹ dùng. Ưu điểm: cầm nhẹ, màn hình lớn, giao hàng nhanh, chụp ảnh ok. Nhược: cầm ko được chắc tay cho lắm, một hộp chỉ có điện thoại và dây sạc, không có tai nghe. Màn hình và ốp bám vân tay. Sau khi mua xong thì đọc review là sử dụng chip cũ snap450 cùi :)). Giá không phù hợp với cấu hình. Mua trả góp nên mình không huỷ được đơn hàng :(

-Sản phẩm: Samsung Galaxy A11

-Khách hàng: Dohuyen

-Kết quả:

A screenshot of a cell phone

Description automatically generated

**Bình luận 4**: Điện thoại phù hợp giá tiền thôi, mua cho bà để gọi điện với con cháu mà camera xấu tệ :)) , pin tuột nhanh, cấu hình tạm ổn, mình thấy mọi người hỏi chơi game được ko thì mình tải thử Free Fire cũg tạm hơi lag nhẹ nhưng vẫn chơi đc. mặt sau dễ bám vân tay,loa hơi nhỏ,nói chung tốt trog tầm giá, ngoài camara thì ko nói gì đc. Giao hàng nhanh.

-Sản phẩm: Vsmart Star 3

-Khách hàng: Nguyễn Thị Tâm Như

-Kết quả:

A screenshot of a cell phone

Description automatically generated

**Bình luận 5**: máy trâu pin, chơi game mượt không bị giậy lag, bắt mạng rất nhanh, chụp hình đẹp nói chung là rất ok

-Sản phẩm: OPPO A5s

-Khách hàng: Lê Hữu Quốc Minh

-Kết quả:

A screenshot of a cell phone

Description automatically generated

# Tổng kết

## **3.1 Những điểm hạn chế**

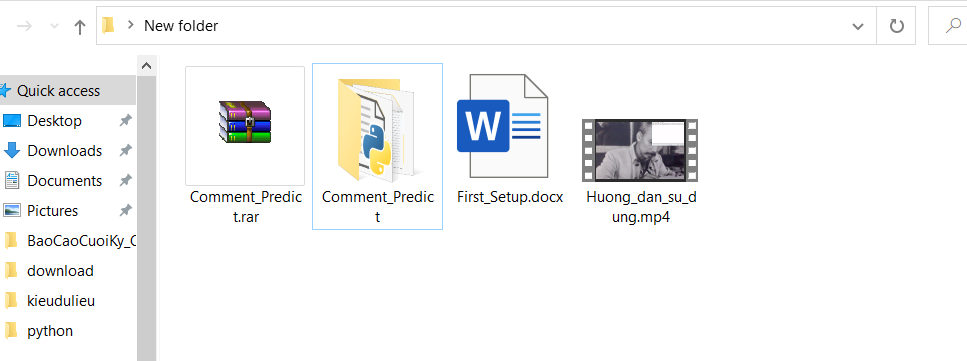
* Model SVC khá nhanh, nhẹ nhàng, dùng tốt trong những trường hợp bình luận thông thường nhưng vẫn còn phân tích sai trong những trường hợp bình luận khó và phức tạp về ý nghĩa.
* Các nhãn trong bộ data có thể gán chưa được chuẩn và việc xuất hiện thêm nhiều từ mới có thể gây nhầm lẫn trong việc phân loại các bình luận.
* Việc chuẩn hóa vẫn chưa xử lí được các trường hợp bình luận lỗi font, sai chính tả, thiếu chữ,… (VD: ah, mu, mụa, giạo hàng, sạt pin, kiêủ,…) gây nhiễu trong quá trình huấn luyện.
* Mặc dù việc dùng TF-IDF để trích xuất vector đặc trưng đã có để ý đến thứ tự sắp xếp các từ trong câu nhưng TF-IDF vẫn chủ yếu là để đánh trọng số những từ quan trọng là chính nên có thể sai trong những trường hợp bình luận kiểu như: “không gây thất vọng” bị xếp vào nhóm tiêu cực.
* Giải pháp:

Sử dụng word embedding (Word2vec) và LSTM để cải thiện.

Thu thập thêm data.

## **3.2 Hướng dẫn sử dụng source code**

**Bước 1**: Tải file Comment\_Predict.rar được cung cấp về máy và giải nén.



File Comment\_Predict: chứa toàn bộ source code của đồ án lần này.

File First\_Setup.docx: hướng dẫn install các thư viện cần thiết trước khi chạy chương trình.

File Huong\_dan\_su\_dung.mp4: Demo hướng dẫn sử dụng.

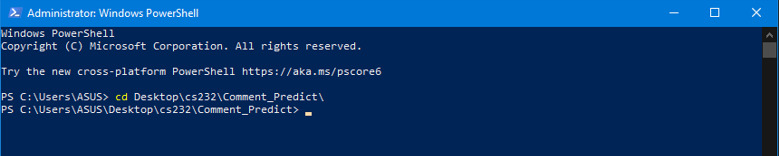
**Bước 2**: Để chạy chương trình ta tiến hành bật CMD hoặc PowerShell.

Dùng lệnh *cd* để truy cập đến đường dẫn nơi lưu file thực thi, file thực thi ở đây là Predict.py

A screenshot of a social media post

Description automatically generated

*Đường dẫn nơi lưu file thực thi (Predict.py)*



**Bước 3**: (Nếu không phải lần đầu thực thi bỏ qua bước này)

**\*Yêu cầu**: Python phiên bản 3.7

Dùng lệnh sau để cài đặt các thư viện cần thiết:

pip install -r requirements.txt

hoặc pip3 install -r requirements.txt

A screenshot of a cell phone

Description automatically generated

**Bước 4**: Nhập lệnh *python* <tênfile>.py để chạy chương trình.

python Predict.py

A screenshot of a cell phone

Description automatically generated

# Tài liệu tham khảo

1. <https://machinelearningcoban.com/2017/08/31/evaluation/#-truefalse-positivenegative>
2. <https://www.digitalocean.com/community/tutorials/how-to-scrape-web-pages-with-beautiful-soup-and-python-3>
3. <https://blog.vietnamlab.vn/2019/08/04/xay-dung-1-model-machine-learning-don-gian-de-giai-quyet-bai-toan-phan-loai-sac-thai-binh-luan-trong-tieng-viet/>
4. <https://codetudau.com/bag-of-words-tf-idf-xu-ly-ngon-ngu-tu-nhien/index.html>
5. <https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html>
6. <https://towardsdatascience.com/how-sklearns-tf-idf-is-different-from-the-standard-tf-idf-275fa582e73d>
7. <https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/?fbclid=IwAR3NT5Ui6YmU4i8lDTCt9sTekHNjyWg4-vn4HSto8aZg5OP5yVhxHupyVpc#:~:text=Once%20precision%20and%20recall%20have>
8. <https://realpython.com/python-gui-tkinter/>
9. <https://docs.python.org/3/library/re.html>
10. <https://github.com/undertheseanlp/word_tokenize>

# Bảng phân công

|  |  |  |
| --- | --- | --- |
| **Mã số sinh viên** | **Họ và tên** | **Công việc được giao** |
| 18521060 | Trịnh Hưng Long | Thực hiện crawl data từ trang web, tinh chỉnh lại các nhãn và tổng hợp bộ data hoàn chỉnh, chỉnh sửa file báo cáo và slide thuyết trình. |
| 18521062 | Hà Văn Luân | Thực hiện các bước tiền xử lí data, tinh chỉnh lại các nhãn thích hợp cho bộ dataset bằng tay, thu thập các từ cần chuẩn hóa trong bộ dataset. |
| 18521274 | Lỡ Đình Phương | Thực hiện đánh giá, training model, thiết kế giao diện ứng dụng và tinh chỉnh lại các nhãn thích hợp cho bộ dataset bằng tay. |