*ĐHQG TP. Hồ Chí Minh - Trường ĐH Công Nghệ Thông Tin*

**-**oOo**-**

**Logo

Description automatically generated**

**ĐỒ ÁN CUỐI KỲ**

**PHÂN TÍCH CẢM XÚC BẰNG CÁCH SỬ DỤNG DỮ LIỆU ĐÁNH GIÁ SẢN PHẨM SHOPEE**

Môn học: Tư duy tính toán – CS117.M21.KHCL

Giảng viên: Ngô Đức Thành

Sinh viên thực hiện: Hoàng Gia Huy – 19521607

Đoàn Tấn Phát – 20520269

Bùi Thị Bích Hậu -19521483

*TP. Hồ Chí Minh, 11 tháng 06 năm 2022*

**Mục lục**

[**I. Giới thiệu:** 3](#_Toc106031801)

[**1.1. Tổng quan:** 3](#_Toc106031802)

[**1.2. Nhiệm vụ của bài toán:** 3](#_Toc106031803)

[**1.3. Sơ đồ Hierarchy:** 3](#_Toc106031804)

[**1.4. Graphic Organizer:** 3](#_Toc106031805)

[**1.4. FlowChart:** 11](#_Toc106031806)

[**II. Bài toán:** 12](#_Toc106031807)

[**2.1. Xác định input và output của bài toán:** 12](#_Toc106031808)

[**2.2. Chuẩn bị dữ liệu (Prepare Dataset):** 12](#_Toc106031809)

[**2.3. Tiền xử lý dữ liệu (Data Preprocessing):** 13](#_Toc106031810)

[**2.4. Xây dựng và Huấn luyện model (Choosing and Training model):** 16](#_Toc106031811)

[**2.5. Kết quả:** 18](#_Toc106031812)

[**III. Ứng dụng demo:** 22](#_Toc106031813)

[**IV. Bảng phân công:** 23](#_Toc106031814)

[**V. Tài liệu tham khảo:** 23](#_Toc106031815)

# **I. Giới thiệu:**

## **1.1. Tổng quan:**

- Ngày nay, với sự phát triển vượt bậc của khoa học và công nghệ, đặc biệt là sự bùng nổ của Internet với các phương tiện truyền thông xã hội, thương mại điện tử,... cho phép mọi người không những chia sẻ thông tin trên Internet mà còn thể hiện thái độ, quan điểm của mình đối với các sản phẩm, dịch vụ và các vấn đề xã hội khác trong cuộc sống.

**-** Việc thu thập và xem xét các thông tin phản hồi đó của khách hàng là một cách để giúp cho các doanh nghiệp hiểu được điểm mạnh, điểm yếu trong sản phẩm, dịch vụ của mình; đồng thời nhanh chóng nắm bắt được tâm lý và nhu cầu khách hàng để mang đến cho họ sản phẩm và dịch vụ hoàn hảo nhất.

## **1.2. Nhiệm vụ của bài toán:**

- Với nhu cầu thị trường và sự phát triển của ngành công nghệ thông tin hiện nay, việc xây dựng một mô hình tự động đánh giá và phân loại các câu bình luận, phê bình của người tiêu dùng đóng vai trò rất quan trọng, nhằm giúp:

* Người dùng sử dụng nó có thể tìm kiếm, tham khảo trước khi đưa ra quyết định sử dụng một sản phẩm hay dịch vụ nào đó.
* Các nhà cung cấp dịch vụ cũng có thể sử dụng những nguồn thông tin này để đánh giá về sản phẩm của mình, từ đó có thể đưa ra những cải tiến phù hợp hơn với người dùng, mang lại lợi nhuận cao hơn, tránh các rủi ro đáng tiếc xảy ra.

## **1.3. Sơ đồ Hierarchy:**

Graphical user interface, application, Teams, PowerPoint

Description automatically generated

**Cấu trúc giải quyết cho bài toán:**

Để giải quyết bài toán nhóm, đã phân chia bài toán thành cấu trúc gồm 3 phần chính:

* Data
* Training
* Demo

Với 3 phần trên nhóm thấy rằng có thể áp dụng các pattern của việc giải quyết một bài toán máy học (cụ thể là bài toán phân loại). Sau khi phân chia bài toán thành 3 vấn đề nhỏ hơn, chúng em sẽ thực hiện giải quyết vấn đề bằng các bước tương tự khi giải quyết một vấn đề máy học (các công đoạn để giải quyết một bài toán máy học điển hình)

Ngoài ra ở mỗi bước thực hiện, sẽ có một số bước nhỏ kèm theo (rõ hơn ở phần Graphic Organizer). Các vấn đề ở node trên tuy chưa là một vấn đề quen đơn giản nhưng ta có thể sử dụng các pattern recognition về bài toán text classification để giải quyết (các phần ở các node lá là các công đoạn được xem là cơ bản trong các bài toán có sử dụng máy học).

## **1.4. Graphic Organizer:**

**Iteration 1:**Diagram

Description automatically generated

**Iteration 2:**Text

Description automatically generated**- Iteration 3:**Text

Description automatically generated

**Iteration 4:**Text

Description automatically generated

**Iteration 5:**Text

Description automatically generated with medium confidence

**Iteration 6:**Text

Description automatically generated

**Iteration 7:**Text

Description automatically generated

## **1.4. FlowChart:**

Diagram

Description automatically generated

# **II. Bài toán:**

**2.1. Xác định input và output của bài toán:**

- Input:

* Một input duy nhất dưới định dạng text.
* Input đó là một hay nhiều câu comments/reviews ngắn của một người về một sản phẩm/dịch vụ trong ngôn ngữ tiếng Việt có dấu, được nhập trên cùng một hàng.
* Input không bao gồm các từ ngữ tiếng Anh, các con số và các kí tự đặc biệt như dấu câu và emojis.
* Input chỉ có thể chứa tối đa 200 từ.( vì quá 200 từ thường là spam)

- Output:

* Một output duy nhất cũng dưới định dạng dạng text.
* Output chỉ có thể là 1 trong 3 giá trị: -1, 0, 1.

*“-1”*  - Negative: Thể hiện rằng comment/reviews của người dùng mang hàm ý phê bình, chê bai, góp ý về sản phẩm/dịch vụ.

*“0”* - Neutral: Thể hiện rằng comment/reviews của người dùng mang hàm ý trung tính. Output ở dạng này thường xuất hiện khi:

+ Người dùng không nêu rõ ý kiến của mình ở input.

+ Người dùng vừa thể hiện tích tiêu cực và tích cực ở 2 phần khác nhau của input.

+ Comments/Reviews của người dùng khó phân tích hoặc không liên quan đến sản phẩm/dịch vụ. (đa phần là các tin nhắn spam, hoặc là các đường link mà không phải là bình luận đánh giá,…).

*“1”* - Positive: Thể hiện rằng comment/reviews của người dùng mang hàm ý ủng hộ, đánh giá cáo về sản phẩm/dịch vụ.

**2.2. Chuẩn bị dữ liệu (Prepare Dataset):**

**- Crawl Data:** Đầu tiên, dữ liệu sẽ được thu thập bằng cách crawl data từ các câu bình luận, đánh giá của khách hàng về sản phẩm/dịch vụ ở trang bán hàng online shopee sử dụng thư viện BeautifulSoup – 1 package của Python dùng để phân tích cú pháp các tài liệu HTML và XML.

➔ Kết quả, nhóm em thu về được ~11000 bình luận (file **FinalData.csv**) với các thông tin (‘label, ‘text’) và được 1 thành viên trong nhóm dán nhãn thủ công.

**- Phân phối dữ liệu:**

**Chart, bar chart

Description automatically generated**

1 7897

0 2150

-1 1457

**- Link Colab:** [**https://colab.research.google.com/drive/1iHeTCzlso1NaPKKQT3K59Lamhh-VkqGU?usp=sharing**](https://colab.research.google.com/drive/1iHeTCzlso1NaPKKQT3K59Lamhh-VkqGU?usp=sharing)

**2.3. 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.

Text

Description automatically generated with low confidence

- Đưa về kiểu chữ thường:

Text

Description automatically generated with low confidence

- Loại bỏ các con số: 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 ‘’.

Text

Description automatically generated with medium confidence

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

Calendar

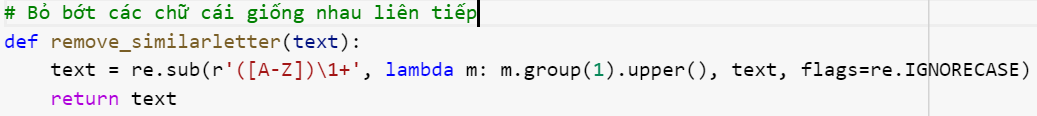
Description automatically generated with low confidence

- Loại bỏ khoảng trắng thừa: 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 “ ”.

Text

Description automatically generated with low confidence

- Loại bỏ các chữ cái giống nhau liên tiếp:



- Tách từ tiếng Việt sử dụng thư viện underthesea có sẵn:

A picture containing text

Description automatically generated

- Loại bỏ các stopwords tiếng Việt sử dụng bộ stopwords có sẵn:

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.

Table

Description automatically generated

Graphical user interface, text, application

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.

Text, letter

Description automatically generated

Text

Description automatically generated

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

Text

Description automatically generated

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.

Calendar

Description automatically generated

w"Shop giao hàng nhanh chóng. Hàng đúng đủ y hình. Chất lượng tốt. Sẽ ủng hộ shop dài dài. Các bạn yên tâm mua hàng nhé"

"

**2.4. 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 70/30.



- Đánh giá model:

A picture containing text

Description automatically generated

* 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.

|  |  |  |  |
| --- | --- | --- | --- |
| -1 | **True Negative** | False Neutral | False Positive |
| 0 | False Negative | **True Neutral** | False Positive |
| 1 | False Negative | False Neutral | **True Positive** |
|  | -1 | 0  Predicted label | 1 |

True label

- Precision\_1: là tỉ lệ số điểm True Negative trong số những điểm được phân loại là Negative.

Precision\_1 = TNeg/(TNeg + FNeg + FNeg)

- Precision\_2: là tỉ lệ số điểm True Neutral trong số những điểm được phân loại là Neutral.

Precision\_2 = TNeu/(TNeu + FNeu + FNeu)  
- Precision\_3: là tỉ lệ số điểm True Positive trong số những điểm được phân loại là Positive.

Precision\_3 = TPo/(TPo + FPo + FPo)

- Recall\_1: là tỉ lệ số điểm True Negative trong số những điểm thực sự là Negative.

Recal\_1 = TNeg/(TNeg + FNeu + FPo)

- Recall\_2: là tỉ lệ số điểm True Neutral trong số những điểm thực sự là Neutral.

Recal\_2 = TNeu/(TNeu + FNeg + FPo)

- Recall\_3: là tỉ lệ số điểm True Positive trong số những điểm thực sự là Positive.

Recal\_3 = TPo/(TPo + FNeu + FNeg)

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

F1-Score\_1 = 2 x (Precision\_1 x Recall\_1)/(Precision\_1 + Recall\_1)

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

F1-Score\_2 = 2 x (Precision\_2 x Recall\_2)/(Precision\_2 + Recall\_2)

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

F1-Score\_3 = 2 x (Precision\_3 x Recall\_3)/(Precision\_3 + Recall\_3)

➔ F1-Score = (F1-Score\_1 + F1-Score\_2 + F1-Score\_3)/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.

Text

Description automatically generated

## **2.5. Kết quả:**

**-** Mô hình SVM:

Text

Description automatically generated

Chart

Description automatically generatedTable

Description automatically generated

**-** Mô hình Multinominal Naïve Bayes:

Text

Description automatically generated with medium confidence

Chart

Description automatically generated

Table

Description automatically generated

**-** Mô hình Logistic Regression:

Text

Description automatically generated

Chart

Description automatically generated

A screenshot of a computer

Description automatically generated with low confidence

**-** Mô hình Random Forest:

Text

Description automatically generated

Chart, waterfall chart

Description automatically generated

A screenshot of a computer

Description automatically generated with low confidence

**-** Mô hình Decision Tree:

Text

Description automatically generated

Chart

Description automatically generated

Table

Description automatically generated

**Bảng so sánh kết quả chạy của các mô hình:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Kết quả** | **SVM** | **Naïve Bayes** | **Logistic Regression** | **Random Forest** | **Decision Tree** |
| Recall | 0.70 | 0.53 | 0.70 | 0.68 | 0.66 |
| Precision | 0.76 | 0.76 | 0.77 | 0.73 | 0.67 |
| F1\_score | 0.73 | 0.58 | 0.73 | 0.70 | 0.67 |
| Accuracy | 0.83 | 0.76 | 0.83 | 0.81 | 0.78 |

# **III. Ứng dụng demo:**

**- Giao diện thông thường của Web:**

**A picture containing graphical user interface

Description automatically generated**

**- Một số bình luận mẫu và kết quả trả về:**

**A picture containing graphical user interface

Description automatically generated**

**Graphical user interface

Description automatically generated**

**Graphical user interface

Description automatically generated**

**IV. Bảng phân công:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Người Nhận** | **MSSV** | **Công việc được phân công** | **Đánh giá tỉ lệ hoàn thành** |
| Hoàng Gia Huy | 19521607 | Thực hiện crawl data, thu thập data, label các nhãn, tổng hợp data hoàn chỉnh, xây dựng demo, tiền xử lý data. | 100% |
| Bùi Thị Bích Hậu | 19521483 | Label data, format báo cáo, debug code,làm poster. | 100% |
| Đoàn Tấn Phát | 20520269 | Thực hiện đánh giá, viết báo cáo, training model, tiền xử lý data. | 100% |

**V. Tài liệu tham khảo:**

1. https://machinelearningcoban.com/2017/08/31/evaluation/#truefalsepositivenegative

2. <https://www.digitalocean.com/community/tutorials/how-to-scrape-webpages-with-beautiful-soup-and-python-3>

3. <https://blog.vietnamlab.vn/2019/08/04/xay-dung-1-model-machine-learningdon-gian-de-giai-quyet-bai-toan-phan-loai-sac-thai-binh-luan-trong-tiengviet/>

4. <https://codetudau.com/bag-of-words-tf-idf-xu-ly-ngon-ngu-tunhien/index.html>

5.https://scikitlearn.org/stable/modules/generated/sklearn.feature\_extraction.text.TfidfVectorizer.html

6. <https://towardsdatascience.com/how-sklearns-tf-idf-is-different-from-thestandard-tf-idf-275fa582e73d>

7. https://machinelearningmastery.com/precision-recall-and-f-measure-forimbalancedclassification/?fbclid=IwAR3NT5Ui6YmU4i8lDTCt9sTekHNjyWg4- vn4HSto8aZg5OP5yVhxHupyVpc#:~:text=Once%20precision%20and%20r ecall%20have

8. <https://realpython.com/python-gui-tkinter/>

9. <https://docs.python.org/3/library/re.html>

10. <https://github.com/undertheseanlp/word_tokenize>

11. Code được tham khảo từ nhóm:

<https://github.com/Long-1234kfgkl/CS114.K21/blob/master/BaoCaoCuoiKy_CS114.K21/Main.ipynb>

<https://viblo.asia/p/phan-tich-phan-hoi-khach-hang-hieu-qua-voi-machine-learningvietnamese-sentiment-analysis-Eb85opXOK2G>

[Working With Text Data — scikit-learn 0.24.2 documentation](https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html)

[CS114.K21/Text\_Classification.ipynb at master · ThuanPhong0126/CS114.K21 (github.com)](https://github.com/ThuanPhong0126/CS114.K21/blob/master/Project/Text_Classification.ipynb)