Big Data Course

Capstone Project   
Final Report

For students (instructor review required)

ⓒ2023 SAMSUNG. All rights reserved.

Samsung Electronics Corporate Citizenship Office holds the copyright of this document.

This document is a literary property protected by copyright law so reprint and reproduction without permission are prohibited.

To use this document other than the curriculum of Samsung Innovation Campus, you must receive written consent from copyright holder.

| Phân tích dữ liệu tài chính và thương mại điện tử |
| --- |

14/08/24

**OU - HCM - BigData02 – Nhóm 2**

Lê Minh Kiệt

Võ Trần Yến Như

Dương Ngọc Minh Huy

Trần Hữu Hậu

Trần Trọng Nhân

Nguyễn Thanh Nam

Nguyễn Thanh Nở

**Content**

1. Introduction

1.1. Background Information

1.2. Motivation and Objective

1.3. Members and Role Assignments

1.4. Schedule and Milestones

2. Project Execution

2.1. Simulated Scenario Description

2.2. Datasets Selection and Description

2.3. Data Ingestion Pipeline

2.4. Data Transformation Processing

2.5. Data Query and Insight

3. Results

3.1. Data Ingestion Scripts and Code

3.2. Data Transformation Scripts and Code

3.3. Description and Sample of Transformed Datasets

3.4. Data Visualization of Query Results

4. Projected Impact

4.1. Accomplishments and Benefits

4.2. Future Improvements

5. Team Member Review and Comment

6. Instructor Review and Comment

1. Introduction

**1.1. Background Information**

Bình luận và đánh giá của khách hàng là một phần quan trọng trong thương mại điện tử, giúp những khách hàng cũng như người bán có thể đánh giá chung về chất lượng sản phẩm và mức độ hài lòng của khách hàng sau khi mua hàng.

Các bình luận có thể được chia làm 3 loại:

* **Hài lòng**:
  + **Chất lượng sản phẩm:** Khách hàng bày tỏ sự hài lòng về chất lượng của sản phẩm, bao gồm cả hiệu suất, thiết kế, và độ bền. Những bình luận như "Sản phẩm tuyệt vời", "Hoạt động rất tốt", hay "Hoàn toàn hài lòng với chất lượng" là điển hình của nhóm này.
  + **Trải nghiệm mua hàng:** Quy trình mua hàng diễn ra suôn sẻ, từ việc đặt hàng, thanh toán đến giao hàng. Khách hàng thường khen ngợi về thời gian giao hàng nhanh, đóng gói cẩn thận, và sản phẩm đến tay trong tình trạng hoàn hảo. Các nhận xét như "Giao hàng rất nhanh", "Mua hàng dễ dàng" là đặc trưng cho nhãn này.
  + **Dịch vụ khách hàng:** Khách hàng nhận được sự hỗ trợ tận tình từ dịch vụ khách hàng, với phản hồi nhanh chóng và giải quyết các vấn đề một cách thỏa đáng. Các bình luận như "Dịch vụ khách hàng tuyệt vời", "Rất hài lòng với sự hỗ trợ" thể hiện sự hài lòng về dịch vụ.
  + **Cảm nhận chung:** Khách hàng có xu hướng thể hiện sự hài lòng chung với sản phẩm và dịch vụ, thể hiện bằng những bình luận tích cực như "Rất hài lòng", "Sẽ tiếp tục mua hàng", "Sản phẩm đáng giá".
* **Bình thường:**
  + **Chất lượng sản phẩm:** Sản phẩm có thể đạt được mong đợi ở mức cơ bản nhưng không nổi bật. Các nhận xét có thể bao gồm những cụm từ như "Chất lượng ổn", "Không có gì đặc biệt nhưng chấp nhận được", "Sản phẩm đúng như mô tả nhưng không ấn tượng".
  + **Trải nghiệm mua hàng:** Quy trình mua hàng diễn ra bình thường, không gặp sự cố lớn nhưng cũng không có điểm nổi bật nào để khen ngợi. Khách hàng có thể cảm thấy rằng các yếu tố như giao hàng, đóng gói và hỗ trợ khách hàng đều ở mức chấp nhận được nhưng không có gì xuất sắc.
  + **Dịch vụ khách hàng:** Dịch vụ khách hàng được đánh giá là trung bình, với các phản hồi như "Hỗ trợ bình thường", "Dịch vụ ở mức chấp nhận được", hoặc "Không có vấn đề gì lớn nhưng cũng không ấn tượng".
  + **Cảm nhận chung:** Khách hàng thể hiện sự trung lập, không có cảm xúc mạnh mẽ về sản phẩm hoặc dịch vụ. Những bình luận như "Cũng được", "Không tệ nhưng cũng không quá tốt", hoặc "Tạm ổn" thường xuất hiện trong nhãn này.
* **Không hài lòng:**
  + **Chất lượng sản phẩm:** khách hàng không hài lòng với sản phẩm, có thể do các vấn đề như sản phẩm không hoạt động đúng cách, chất lượng kém, hoặc không giống như mô tả. Các bình luận tiêu cực như "Sản phẩm tệ", "Không như mong đợi", "Chất lượng rất kém" là phổ biến trong nhóm này.
  + **Trải nghiệm mua hàng:** Trải nghiệm mua hàng không đạt yêu cầu, có thể do giao hàng chậm trễ, sản phẩm bị hỏng khi đến tay, hoặc các vấn đề khác trong quá trình mua sắm. Những phản hồi như "Giao hàng rất chậm", "Đóng gói kém" là dấu hiệu của sự không hài lòng.
  + **Dịch vụ khách hàng:** Khách hàng có thể không nhận được sự hỗ trợ thỏa đáng, dẫn đến sự thất vọng. Những bình luận như "Dịch vụ khách hàng rất kém", "Không nhận được sự hỗ trợ như mong đợi" thể hiện sự không hài lòng với dịch vụ.
  + **Cảm nhận chung:** Khách hàng có cảm xúc tiêu cực mạnh mẽ về trải nghiệm tổng thể của họ, có thể dẫn đến những bình luận như "Không bao giờ mua lại", "Rất thất vọng", hoặc "Trải nghiệm tệ hại".

Sự đánh giá của khách hàng có vai trò quyết định sự thành công của sản phẩm trong thương mại điện tử, nếu một sản phẩm có đánh giá tích cực, nhận được sự hài lòng của khách hàng thì sản phẩm đó sẽ có khả năng thành công cao hơn rõ rệt so với những sản phẩm có đánh giá tiêu cực và sự hài lòng của khách hàng ở mức thấp. Báo cáo này nhằm đánh giá các bình luận của khách hàng cho từng sản phẩm nhằm giúp người dùng cũng như khách hàng có cái nhìn chính xác nhất về chất lượng cũng như mức độ hài lòng của khách hàng đối với sản phẩm.

**1.2. Motivation and Objective**

* Trong thị trường thương mại điện tử ngày càng mở rộng, việc đánh giá đúng chất lượng sản phẩm và mức độ hài lòng của người dùng là cực kỳ quan trọng. Bình luận của khách hàng đóng vai trò như một nguồn thông tin đáng tin cậy, giúp người tiêu dùng đưa ra quyết định mua sắm, đồng thời cung cấp cho nhà bán hàng thông tin cần thiết để cải thiện sản phẩm và dịch vụ của họ. Tuy nhiên, việc đánh giá sai chất lượng sản phẩm hoặc mức độ hài lòng của người dùng có thể gây ra nhiều tác hại nghiêm trọng cho cả người mua và người bán.
  + **Tác hại đối với người tiêu dùng:** 
    - **Lựa chọn sản phẩm không phù hợp:** Khi chất lượng sản phẩm được đánh giá sai, người tiêu dùng có thể bị dẫn đến việc mua những sản phẩm không đáp ứng được nhu cầu hoặc mong đợi của họ. Điều này không chỉ gây mất thời gian và tiền bạc mà còn có thể làm giảm niềm tin của họ vào các nền tảng thương mại điện tử.
    - **Mất niềm tin vào đánh giá trực tuyến:** Nếu các đánh giá và bình luận bị phân loại sai, người tiêu dùng có thể cảm thấy khó tin tưởng vào các đánh giá trực tuyến, khiến họ gặp khó khăn trong việc đưa ra quyết định mua hàng trong tương lai. Điều này có thể dẫn đến việc họ chuyển sang mua sắm tại các cửa hàng truyền thống hoặc trên các nền tảng khác.
  + **Tác hại đối với nhà bán hàng**
    - **Mất khách hàng tiềm năng:** Khi một sản phẩm tốt bị đánh giá sai là kém chất lượng, nhà bán hàng có thể mất đi nhiều khách hàng tiềm năng. Ngược lại, nếu một sản phẩm kém được đánh giá cao một cách không chính xác, sẽ có nhiều khách hàng thất vọng sau khi mua, dẫn đến tăng tỷ lệ trả hàng và khiếu nại.
    - **Giảm uy tín và doanh thu:** Việc có quá nhiều đánh giá không chính xác có thể làm giảm uy tín của nhà bán hàng trên nền tảng thương mại điện tử. Uy tín giảm dẫn đến doanh thu giảm sút do người tiêu dùng có xu hướng tránh xa những nhà bán hàng hoặc sản phẩm có đánh giá không rõ ràng hoặc mâu thuẫn.
    - **Quản lý sản phẩm kém hiệu quả:** Nếu nhà bán hàng không nhận được phản hồi chính xác về sản phẩm, họ sẽ gặp khó khăn trong việc cải thiện và quản lý sản phẩm một cách hiệu quả. Điều này có thể dẫn đến việc không giải quyết kịp thời các vấn đề của sản phẩm, khiến chất lượng dịch vụ tổng thể bị ảnh hưởng.

Vì những lý do trên, việc phát triển một hệ thống tự động và chính xác để phân tích và đánh giá bình luận của khách hàng là cần thiết. Hệ thống này sẽ giúp giảm thiểu các sai sót trong việc đánh giá chất lượng sản phẩm và mức độ hài lòng của người dùng, từ đó tạo ra môi trường mua sắm trực tuyến tin cậy và hiệu quả hơn. Hơn nữa, hệ thống còn hỗ trợ nhà bán hàng trong việc duy trì và nâng cao chất lượng sản phẩm, từ đó tăng cường khả năng cạnh tranh trên thị trường.

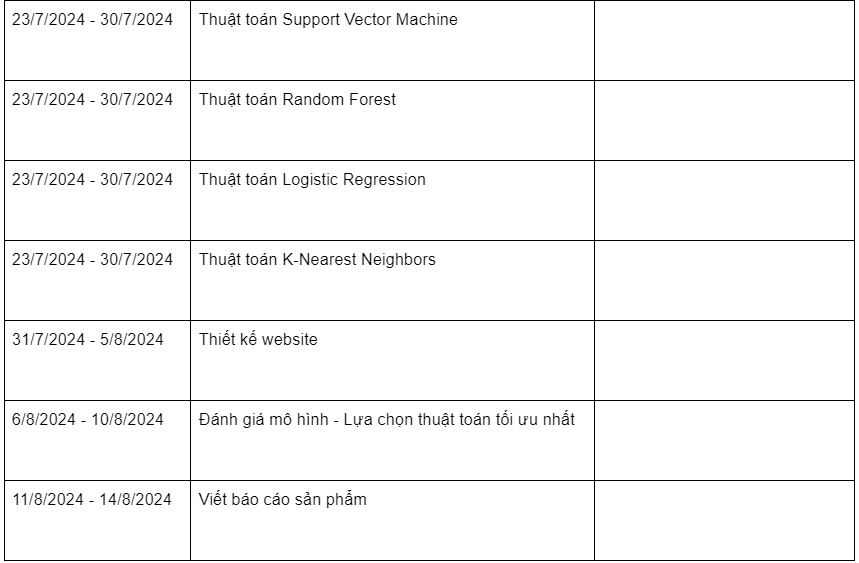
* Mục tiêu chính của dự án là xây dựng một hệ thống dự đoán có khả năng phân loại sản phẩm thương mại điện tử thành ba nhóm “tốt”, “trung bình” và “chưa tốt” dựa trên bình luận của người mua. Để đạt được điều này, dự án sẽ áp dụng các thuật toán học máy như Random Forest, Logistic Regression, Support Vector Machine và K-Nearest Neighbors kết hợp với kỹ thuật xử lý ngôn ngữ tự nhiên để phân tích nội dung bình luận. Ngôn ngữ lập trình Python sẽ được sử dụng để triển khai các mô hình và thực hiện quá trình xử lý dữ liệu. Hệ thống này nhằm cung cấp một công cụ đáng tin cậy cho cả người tiêu dùng lẫn người bán trong việc đánh giá chất lượng sản phẩm dựa trên phản hồi thực tế từ người dùng.

**1.3. Members and Role Assignments**

* Lê Minh Kiệt: Thu thập dữ liệu – Thuật toán Logistic – Xử lý ngôn ngữ tự nhiên – Thiết kế web – Làm sạch dữ liệu
* Võ Trần Yến Như: Thu thập dữ liệu – Làm sạch dữ liệu – Xử lý ngôn ngữ tự nhiên
* Dương Ngọc Minh Huy: Thu thập dữ liệu – Thuật toán SVM – Xử lý ngôn ngữ tự nhiên
* Trần Hữu Hậu: Thu thập dữ liệu – Thuật toán Random Forest – Xử lý ngôn ngữ tự nhiên
* Nguyễn Thanh Nam: Thu thập dữ liệu – Làm sạch dữ liệu – Thiết kế web – Xử lý ngôn ngữ tự nhiên
* Trần Trọng Nhân: Thu thập dữ liệu – Vector hóa dữ liệu – Thuật toán KNN – Xử lý ngôn ngữ tự nhiên
* Nguyễn Thanh Nở: Thu thập dữ liệu – Thiết kế web – Xử lý ngôn ngữ tự nhiên

**1.4. Schedule and Milestones**

A screenshot of a white sheet

Description automatically generated

2. Project Execution

**2.1. Simulated Scenario Description**

Trong bối cảnh phát triển mạnh mẽ của thương mại điện tử, người tiêu dùng phải đối mặt với tình trạng quá tải thông tin khi mua sắm trực tuyến. Họ thường xuyên phải đọc qua hàng loạt nhận xét để đánh giá chất lượng sản phẩm trước khi đưa ra quyết định mua hàng. Điều này không chỉ tốn thời gian mà còn gây khó khăn khi phải lọc qua các nhận xét có giá trị thấp hoặc không liên quan.

Dự án của chúng em mô phỏng một tình huống thực tế, nơi người tiêu dùng có thể sử dụng một hệ thống thông minh để tự động phân tích nhận xét từ các trang web bán hàng. Hệ thống này sẽ xác định chất lượng của sản phẩm dựa trên nhận xét của người dùng, giúp người tiêu dùng đưa ra quyết định nhanh.

Tập dữ liệu sử dụng trong dự án này bao gồm các đánh giá sản phẩm từ một trang thương mại điện tử là Lazada. Mỗi dòng dữ liệu chứa thông tin về đánh giá của khách hàng, bao gồm:

Comment: Bình luận cảm xúc của khách hàng.

StarCount: Điểm số đánh giá sản phẩm (trong thang điểm từ 1 đến 5 sao)

Label: Được đánh giá dựa trên số sao (trong thang điểm từ 1 đến 3 tương ứng Chưa Tốt – Trung Bình – Tốt)

**2.2. Datasets Selection and Description**

Tập dữ liệu hơn 55.000 đánh giá, được phân chia cho các nhãn sản phẩm. Các đánh giá có chiều dài từ 10 đến 100 từ, bao gồm cả các từ ngữ tích cực và tiêu cực, phản ánh cảm nhận của khách hàng về sản phẩm. Dữ liệu sử dụng trong dự án được thu thập từ các trang thương mại điện tử lớn tại Việt Nam, nơi có nhiều nhận xét từ người tiêu dùng về các sản phẩm khác nhau. Chúng em tập trung vào các sản phẩm phổ biến như điện thoại di động, thiết bị gia dụng và mỹ phẩm.

**2.3. Data Ingestion Pipeline**

* Sử dụng thư viện pandas, chúng em sẽ đọc dữ liệu từ file XLSX chứa các đánh giá sản phẩm.
* Dữ liệu được thu thập thông qua quá trình scraping sử dụng công cụ Selenium.
* Scraping: Sử dụng Selenium để tự động thu thập dữ liệu từ các trang web thương mại điện tử.
* Lưu trữ tạm thời: Dữ liệu sau khi scraping được lưu trữ tạm thời trong định dạng XLSX để dễ dàng xử lý tiếp.
* Nhập dữ liệu vào cơ sở dữ liệu: Dữ liệu sau đó được nhập vào một cơ sở dữ liệu lớn, nơi chúng em có thể thực hiện các bước xử lý tiếp theo.

**2.4. Data Transformation Processing**

Sau khi dữ liệu được thu thập và nhập vào hệ thống, quá trình xử lý dữ liệu được thực hiện để chuẩn hóa và biến đổi dữ liệu thành dạng phù hợp cho phân tích:

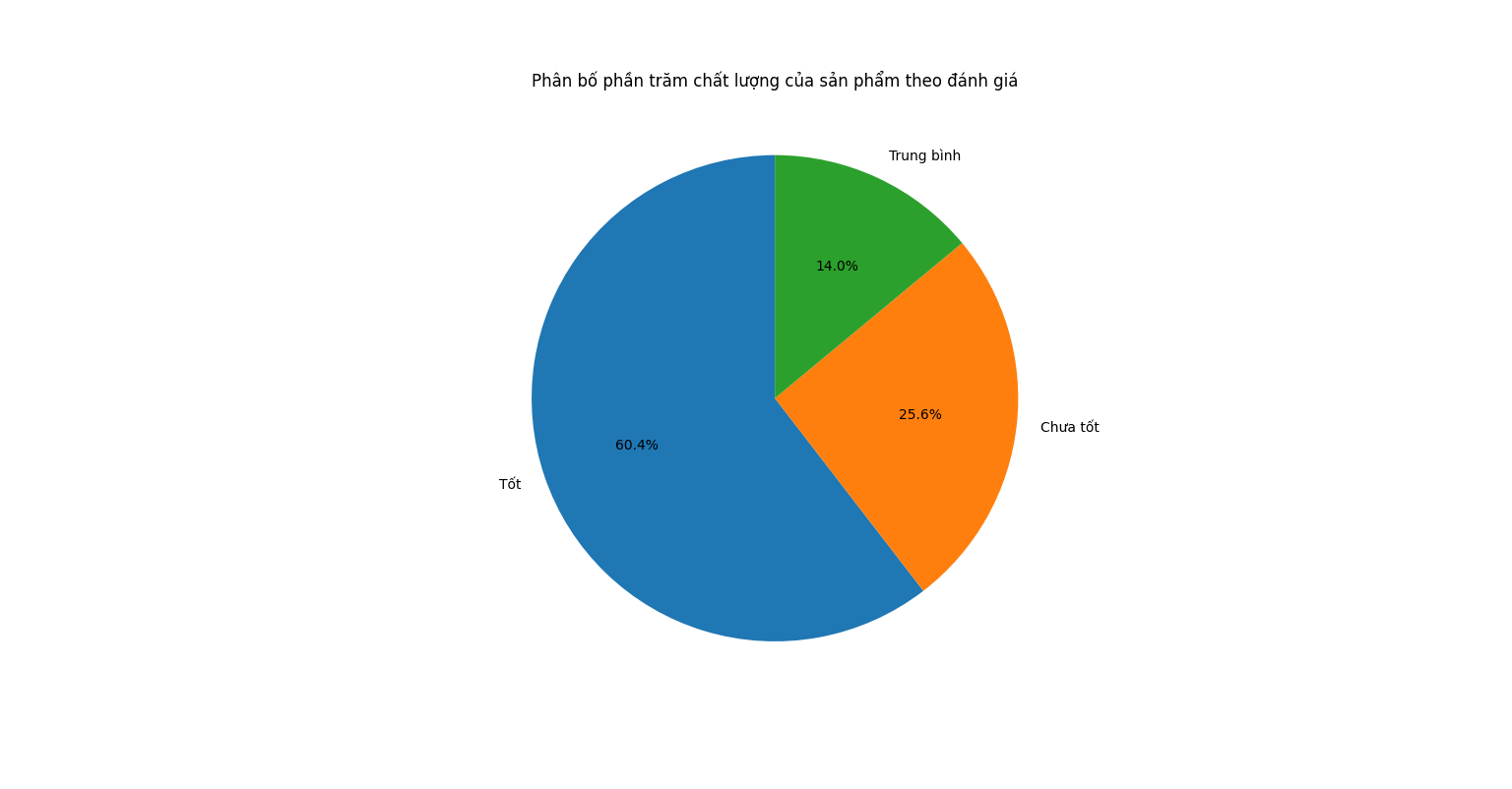
* Xử lý lỗi chính tả và viết tắt: Do dữ liệu nhận xét chủ yếu bằng tiếng Việt, chúng em phải xử lý các lỗi chính tả và các từ viết tắt phổ biến để đảm bảo tính nhất quán.
* Loại bỏ ký tự đặc biệt.
* Loại bỏ từ dừng (stop words).
* Loại bỏ các nhận xét không liên quan: Những nhận xét ngắn hoặc không cung cấp thông tin hữu ích sẽ bị loại bỏ.
* Vector hóa dữ liệu: Sử dụng TfidfVectorizer (TF-IDF) để biến đổi các nhận xét thành dạng vector số, phục vụ cho việc huấn luyện mô hình học máy.

**2.5. Data Query and Insight**

Sau khi dữ liệu đã được xử lý và biến đổi, chúng em thực hiện các truy vấn để rút ra những insight từ dữ liệu:

* Phân tích chất lượng sản phẩm: Dựa trên mô hình RandomForestClassifier, chúng em xác định tỷ lệ các sản phẩm có nhận xét Tốt - Trung bình - Chưa tốt.
* Xu hướng mua sắm của người tiêu dùng: Phân tích những sản phẩm nào nhận được nhiều nhận xét tích cực nhất, đánh giá cao từ trải nghiệm, giúp xác định xu hướng mua sắm.
* Nhận diện vấn đề phổ biến: Thông qua phân tích nội dung bình luận, chúng em có thể xác định các vấn đề thường gặp mà người tiêu dùng phản ánh, như chất lượng sản phẩm kém, thời gian giao hàng lâu, v.v.

Mô hình được đánh giá trên tập kiểm tra bằng các chỉ số phổ biến như độ chính xác, F1-score, độ nhạy (recall) và độ đặc hiệu (precision). Kết quả cho thấy mô hình đạt độ chính xác **92%** với F1-score trung bình là 0.901 (sử dụng average='weighted' do dữ liệu không cân bằng), cho thấy mô hình hoạt động tốt trong việc phân loại các sản phẩm dựa trên đánh giá của khách hàng.



Mô hình này có thể được tích hợp vào hệ thống quản lý sản phẩm của các doanh nghiệp thương mại điện tử để tự động phân loại các sản phẩm theo chất lượng dựa trên đánh giá khách hàng. Điều này không chỉ giúp tiết kiệm thời gian mà còn cải thiện chất lượng dịch vụ thông qua phản hồi nhanh chóng và chính xác từ khách hàng.

## **2.6 Building Models to Predict Product Quality Based on Customer Comments**

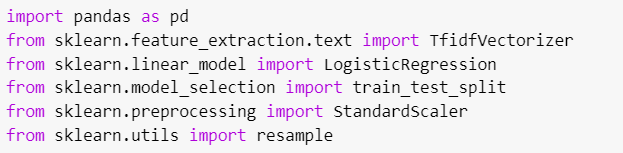
**1. Logistic Regression:**

**1.1. Introduction to Logistic Regression:**

Hồi quy logistic là một kỹ thuật phân tích dữ liệu sử dụng toán học để tìm ra mối quan hệ giữa hai yếu tố dữ liệu. Sau đó, kỹ thuật này sử dụng mối quan hệ đã tìm được để dự đoán giá trị của những yếu tố đó dựa trên yếu tố còn lại. Dự đoán thường cho ra một số kết quả hữu hạn, như có hoặc không.

**1.2. Building a Logistic Regression model.**

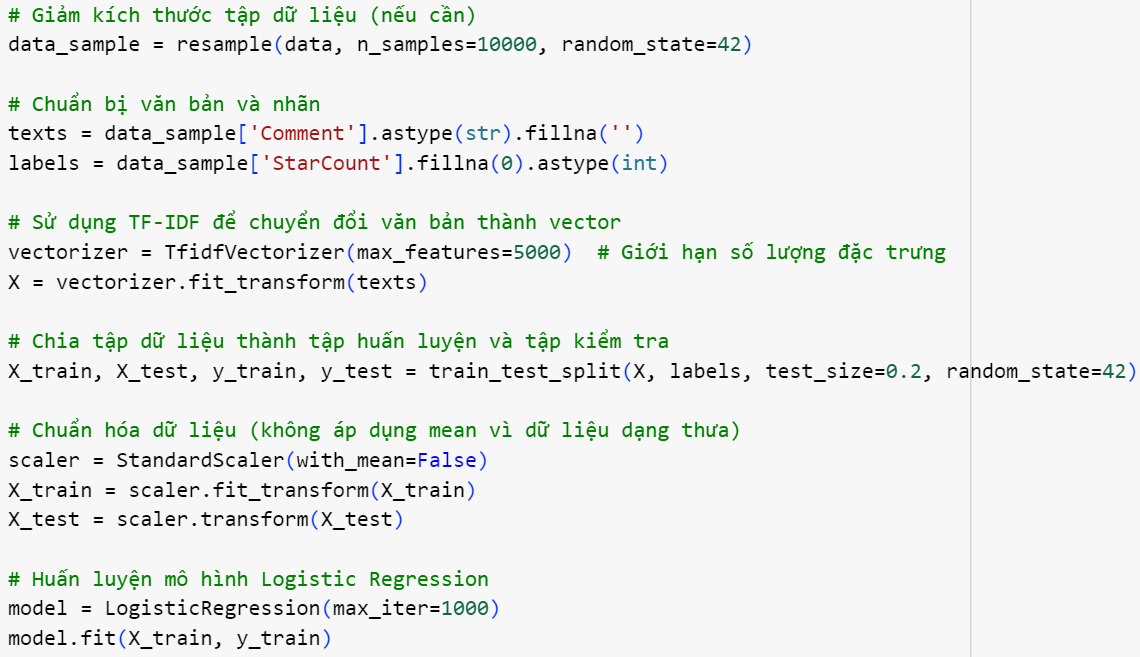
* **Import the necessary libraries to build the model**



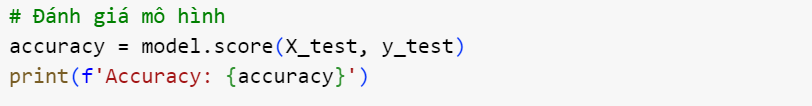
* **Read Datasets**



* **Build model**



* **Model evaluation**



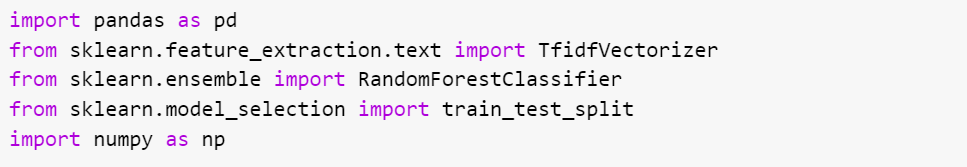
**2. Random Forest:**

**2.1. Introduction to Random Forest:**

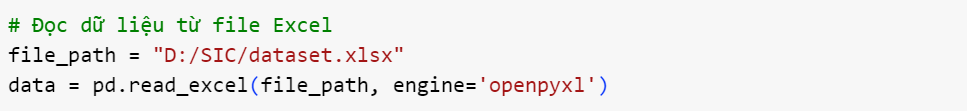
Random là ngẫu nhiên, Forest là rừng, nên ở thuật toán Random Forest mình sẽ xây dựng nhiều cây quyết định bằng thuật toán Decision Tree, tuy nhiên mỗi cây quyết định sẽ khác nhau (có yếu tố random). Sau đó kết quả dự đoán được tổng hợp từ các cây quyết định.

**2.2. Building a Random Forest model.**

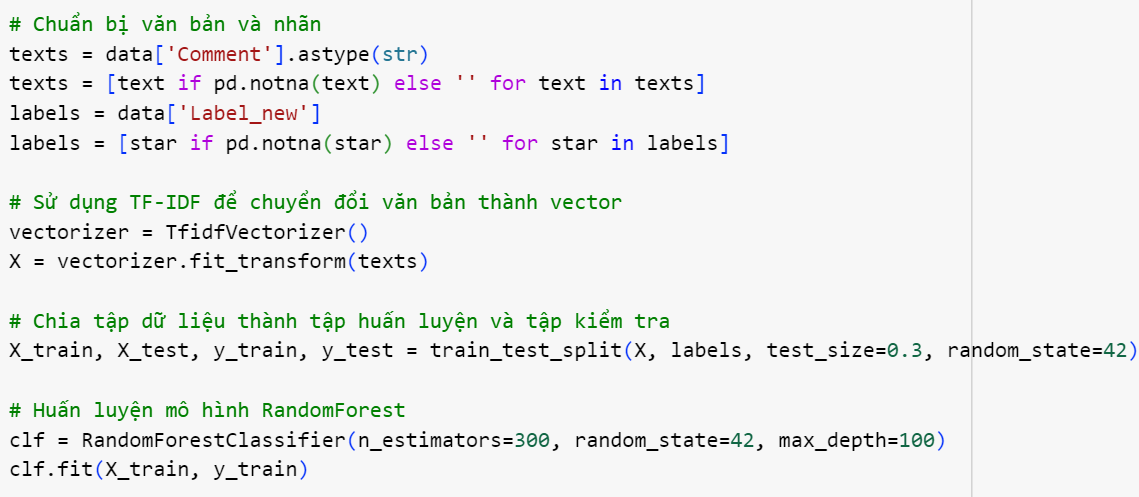
* **Import the necessary libraries to build the model**

****

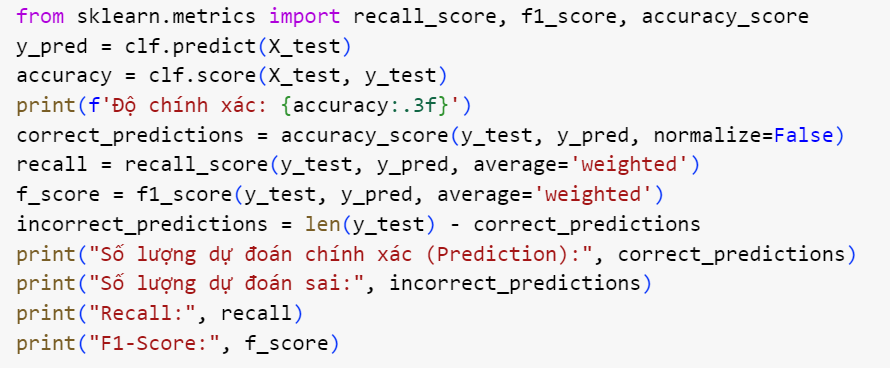
* **Read Datasets**

****

* **Build model**

****

* **Model evaluation**

****

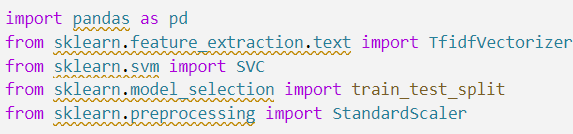
**3. Support Vector Machine (SVM):**

**3.1. Introduction to SVM:**

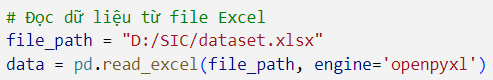
SVM là một thuật toán giám sát, nó có thể sử dụng cho cả việc phân loại hoặc đệ quy. Tuy nhiên nó được sử dụng chủ yếu cho việc phân loại. Trong thuật toán này, chúng ta vẽ đồi thị dữ liệu là các điểm trong n chiều ( ở đây n là số lượng các tính năng bạn có) với giá trị của mỗi tính năng sẽ là một phần liên kết. Sau đó chúng ta thực hiện tìm "đường bay" (hyper-plane) phân chia các lớp. Hyper-plane nó chỉ hiểu đơn giản là 1 đường thẳng có thể phân chia các lớp ra thành hai phần riêng biệt.

**3.2. Building a SVM model.**

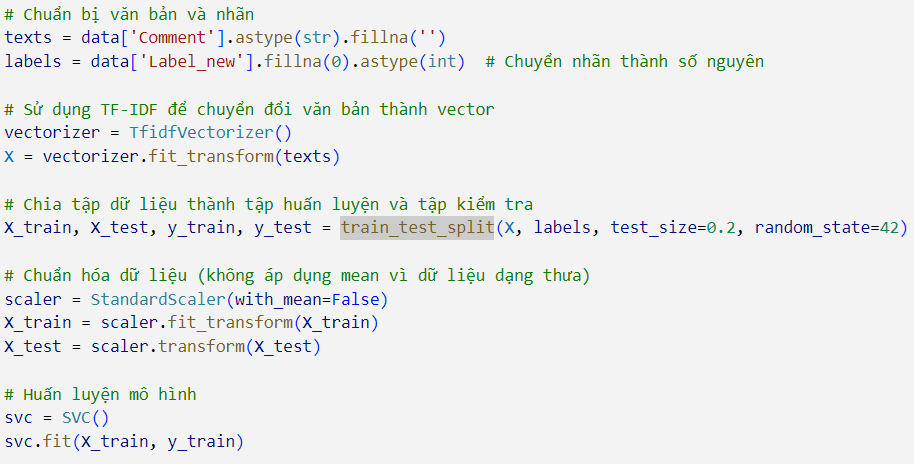
* **Import the necessary libraries to build the model**

****

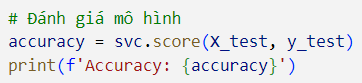
* **Read Datasets**

****

* **Build model**

****

* **Model evaluation**

****

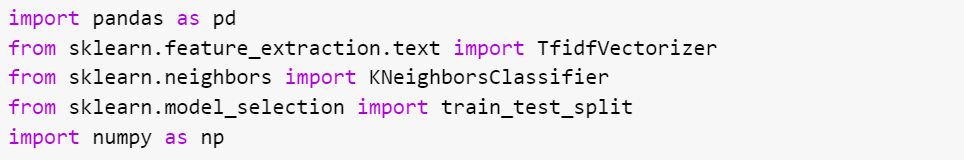
**4. K-nearest neighbors (KNN):**

**4.1. Introduction to KNN:**

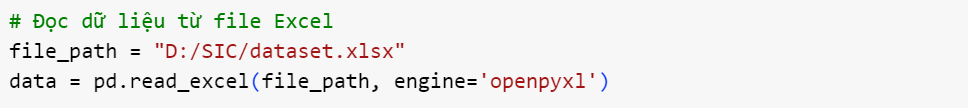
K-nearest neighbor là một trong những thuật toán supervised-learning đơn giản nhất (mà hiệu quả trong một vài trường hợp) trong Machine Learning. Khi training, thuật toán này không học một điều gì từ dữ liệu training (đây cũng là lý do thuật toán này được xếp vào loại lazy learning), mọi tính toán được thực hiện khi nó cần dự đoán kết quả của dữ liệu mới. K-nearest neighbor có thể áp dụng được vào cả hai loại của bài toán Supervised learning là Classification và Regression. KNN còn được gọi là một thuật toán Instance-based hay Memory-based learning.

**4.2. Building a KNN model.**

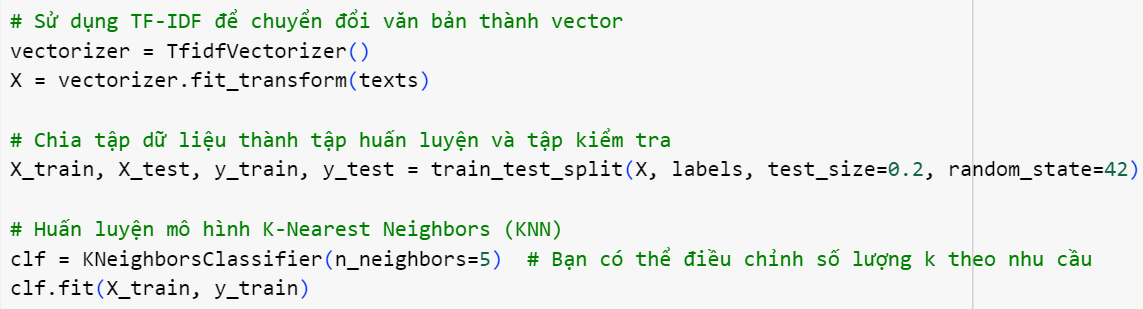
* **Import the necessary libraries to build the model**

****

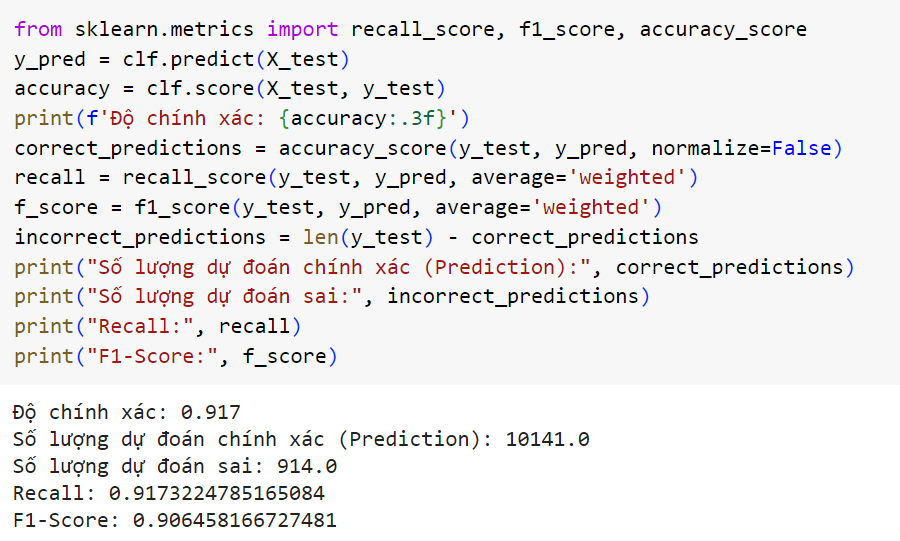
* **Read Datasets**

****

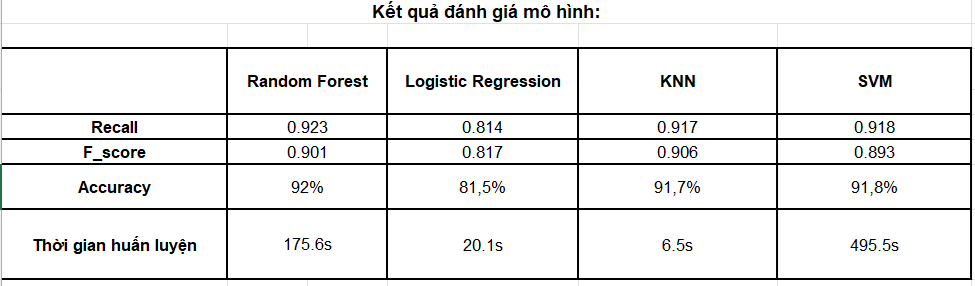
* **Build model**

****

* **Model evaluation**

****

**5. Comparison between algorithms**

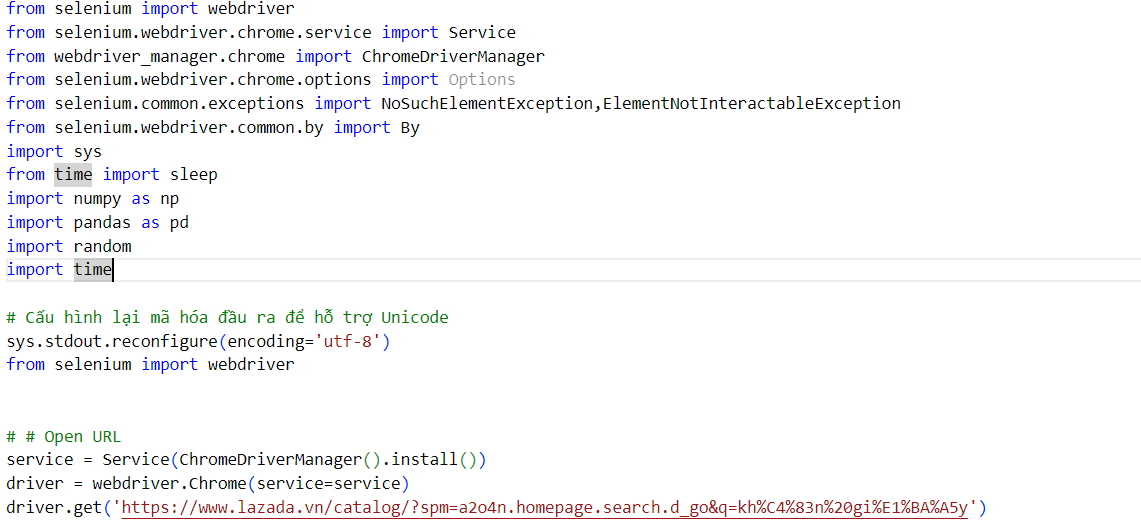


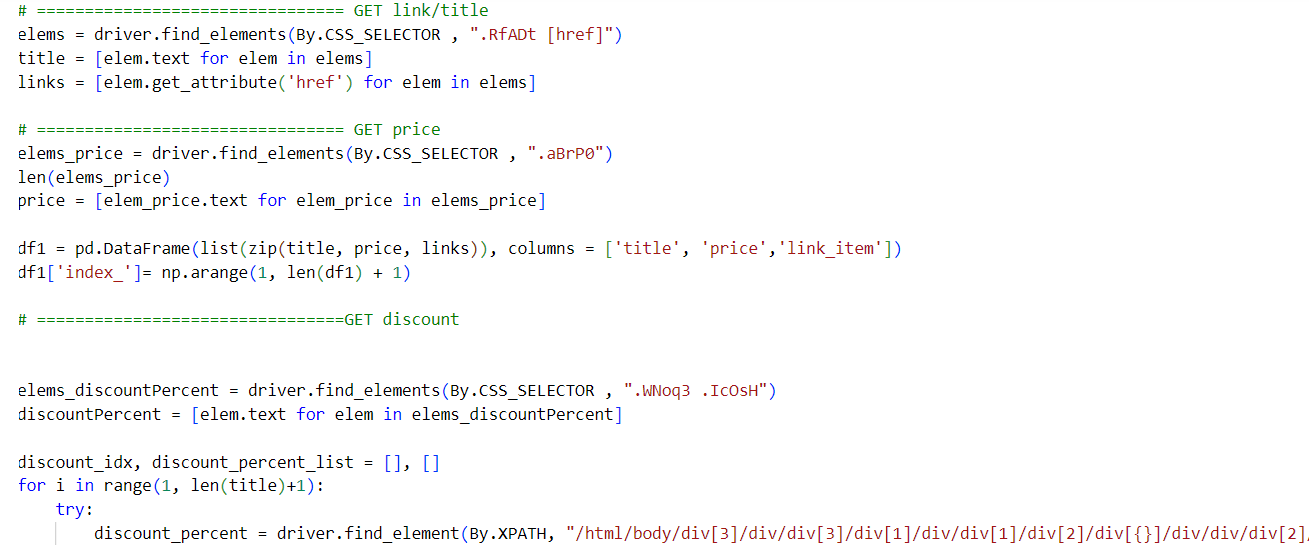
3. Results

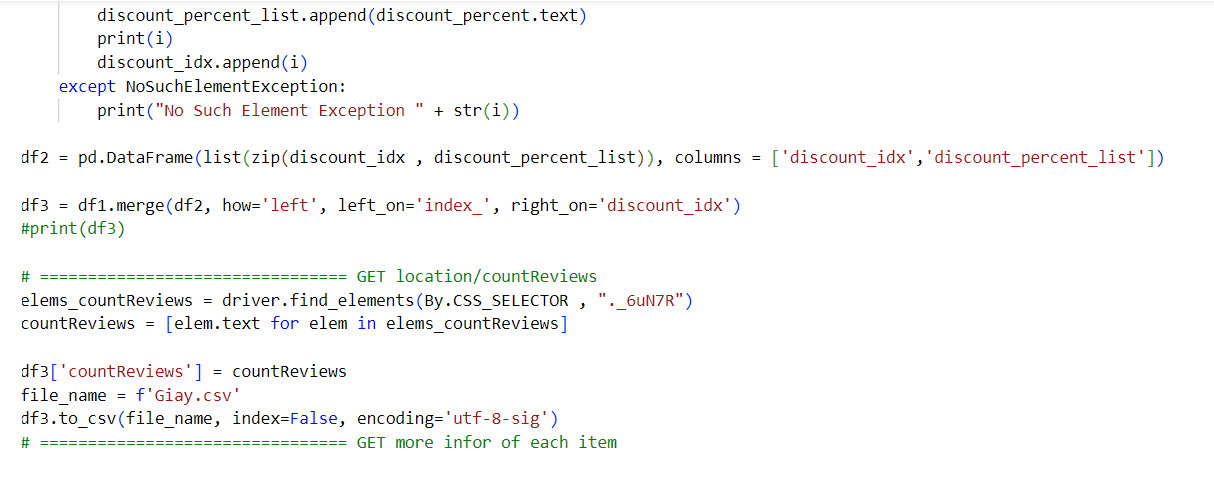
**3.1. Data Ingestion Scripts and Code**

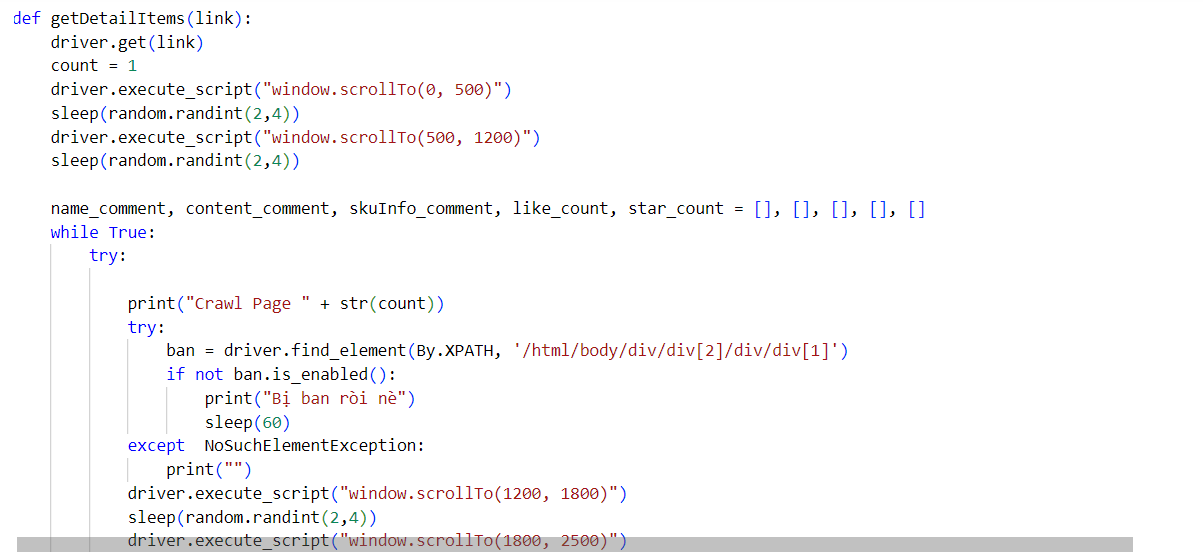
Trong phần này, chúng em mô tả quy trình nạp dữ liệu, bao gồm việc thu thập và nhập dữ liệu từ sàn thương mại điện tử Lazada. Đối với dự án này, dữ liệu chủ yếu là các bình luận của khách hàng bằng tiếng Việt, được thu thập từ nền tảng thương mại điện tử Lazada. Vì dữ liệu này không có cấu trúc rõ ràng nên cần được tiền xử lý đáng kể khi sử dụng cho phân tích.

**\*Lấy dữ liệu từ Lazada**

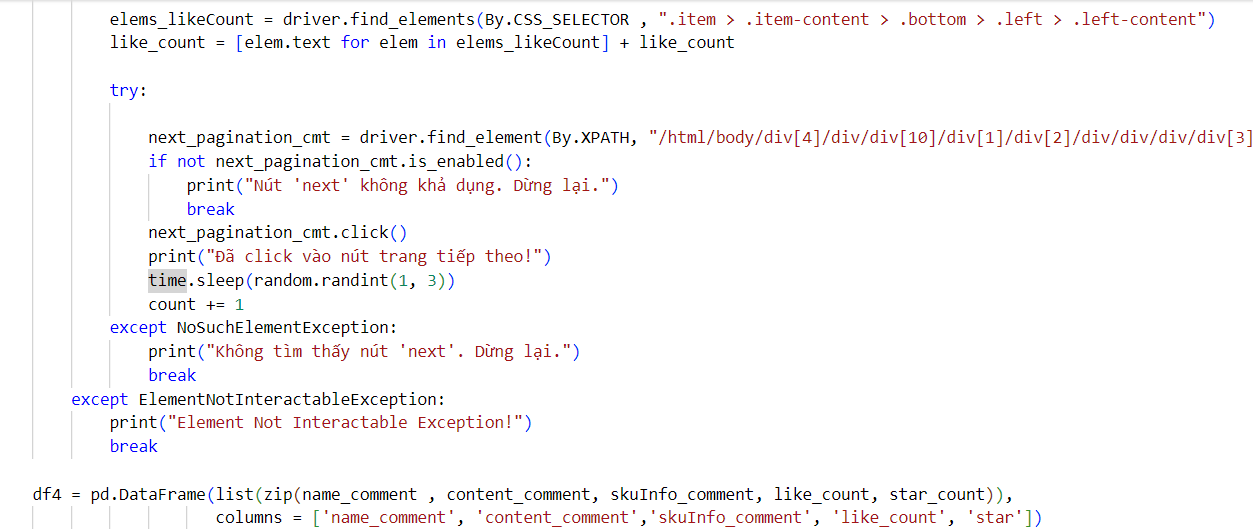
****

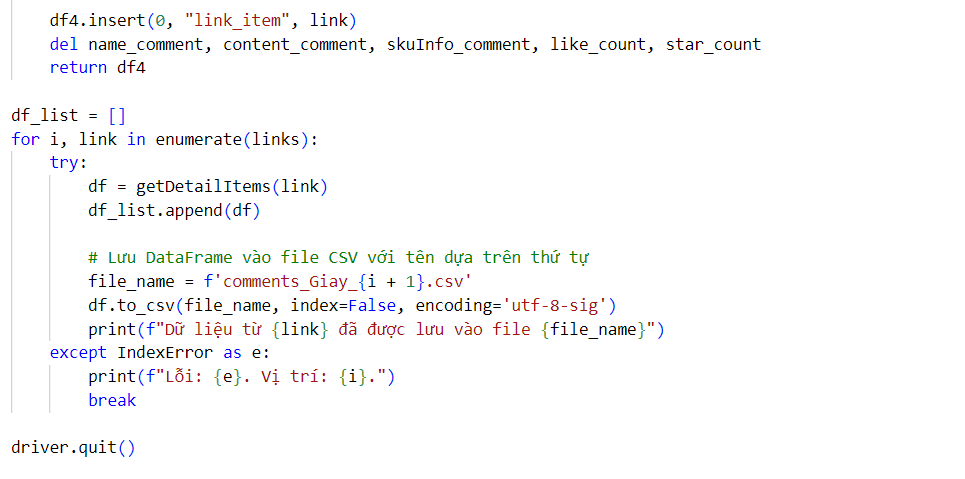








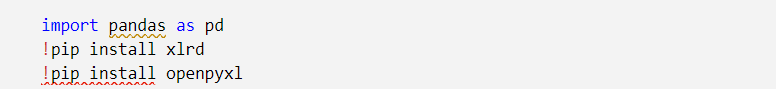


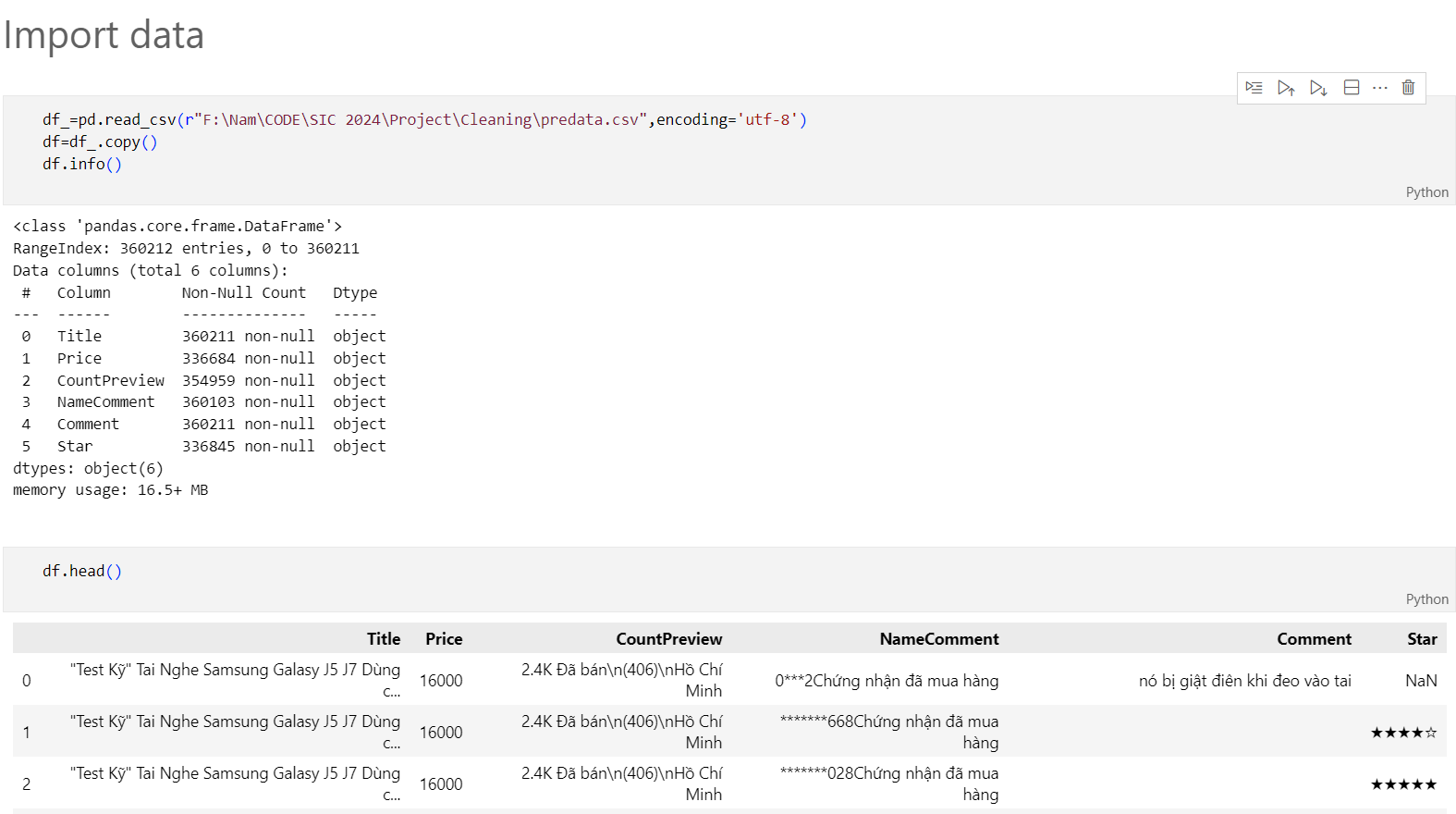


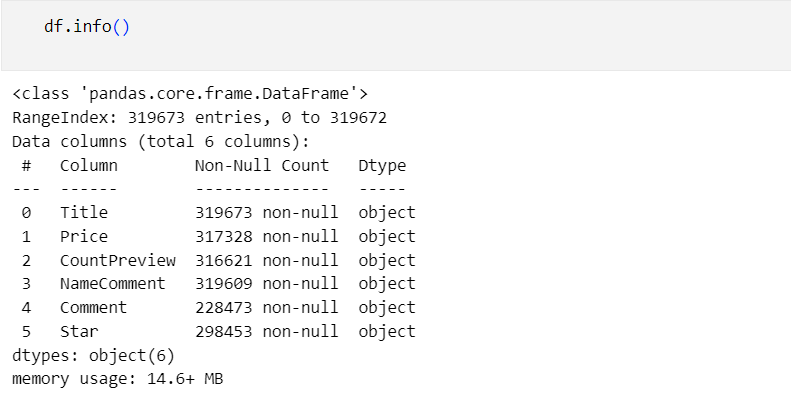
**3.2. Data Transformation Scripts and Code**

Dữ liệu sau khi được thu thập cần được biến đổi để phù hợp với mô hình học máy. Quy trình biến đổi dữ liệu bao gồm xử lý lỗi chính tả, chuẩn hóa từ ngữ, loại bỏ các từ không có ý nghĩa (stop words), và vector hóa dữ liệu. Phần xử lý lỗi chính tả, chuẩn hóa từ hay loại bỏ các từ không ý nghĩa được làm một cách thủ công. Do các công cụ đã không còn hỗ trợ hoặc đã ngừng cập nhật dữ liệu rất nhiều năm về trước.

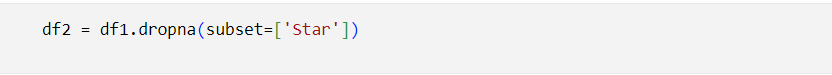
**\* Làm sạch dữ liệu**

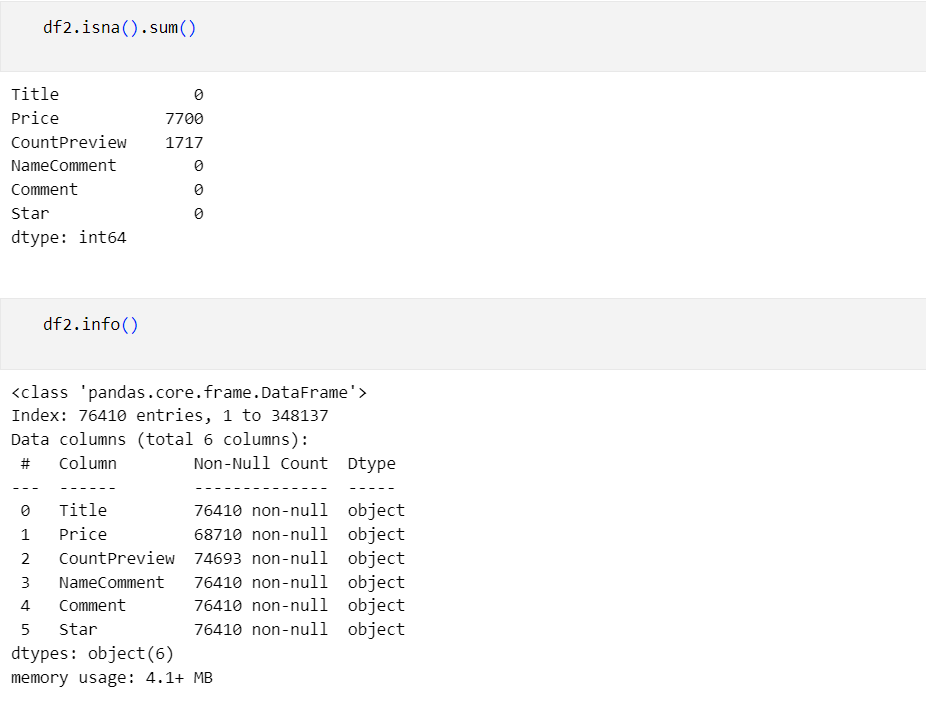
****

****

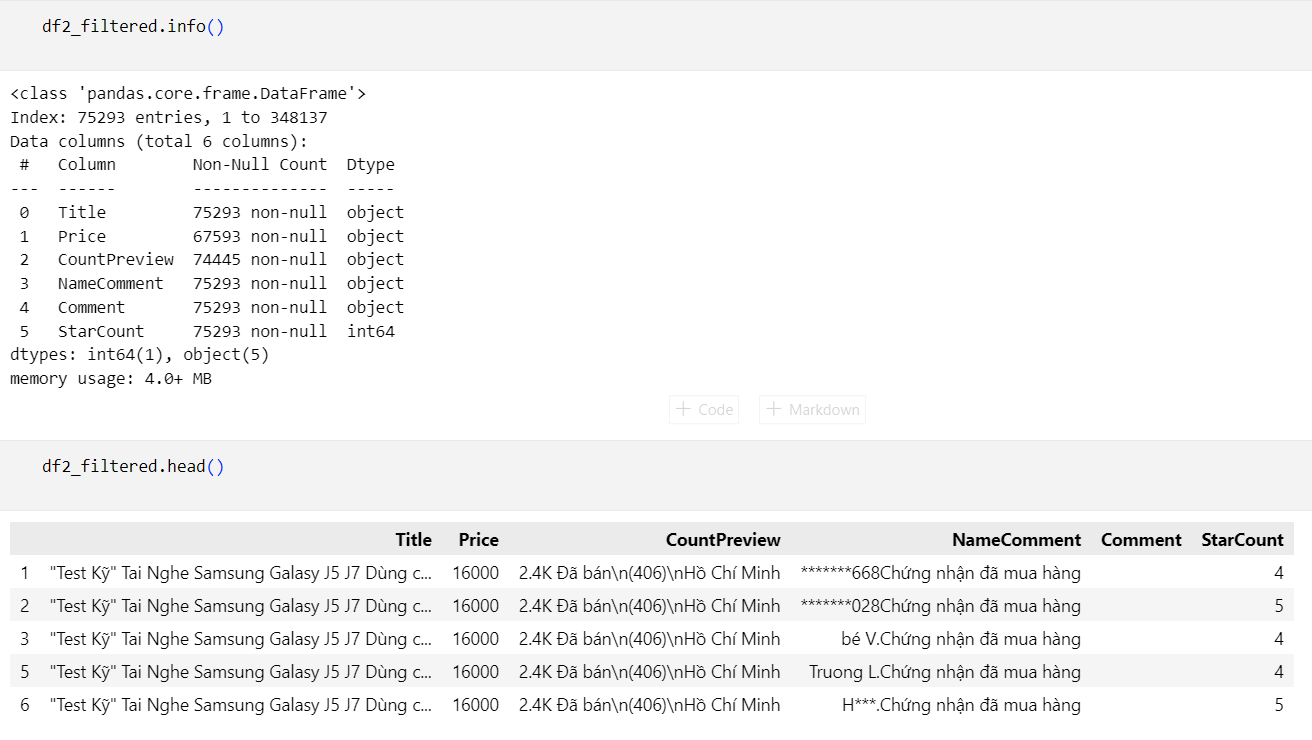
****

****



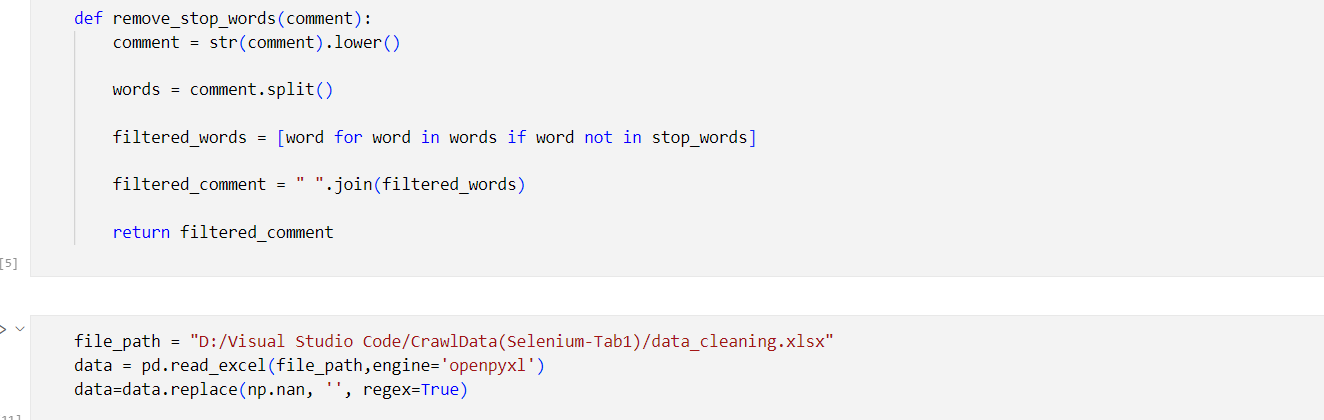
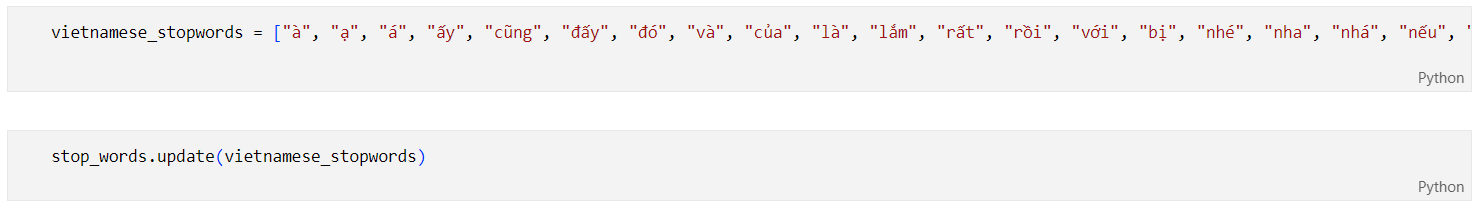
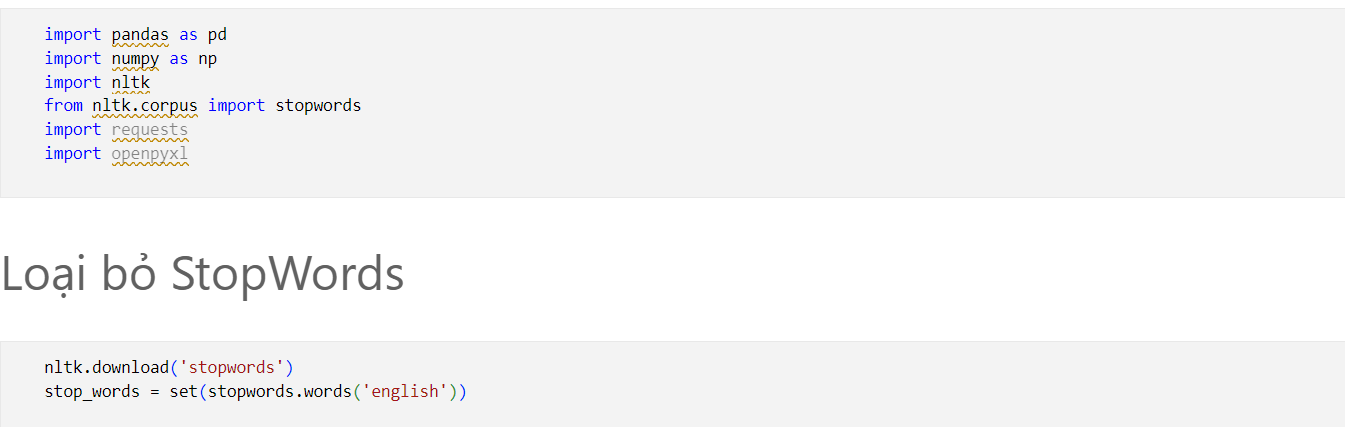






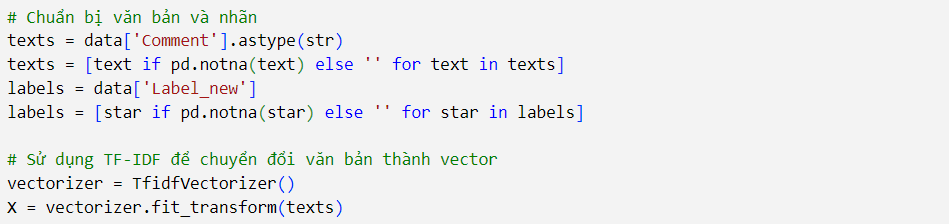


**\*Loại bỏ Stop-Word và gán nhãn Comment**



**\*Vector hóa dữ liệu**

Chúng em sử dụng TF-IDF để vector hóa dữ liệu vì nó giúp xác định các từ quan trọng trong bình luận bằng cách giảm thiểu ảnh hưởng của các từ thông dụng, đồng thời tăng trọng số cho những từ mang ý nghĩa đặc biệt trong ngữ cảnh cụ thể. Giúp cải thiện độ chính xác của mô hình khi phân tích và phân loại bình luận.



**3.3. Description and Sample of Transformed Datasets**

Cột Comment bên trái là cột do người dùng nhập, dữ liệu này được lấy từ những sản phẩm trên trang thương mại điện tử Lazada. Cột Filtered\_Comment bên phải là cột đã được lọc bỏ đi những ký tự đặc biệt, những từ không mang ý nghĩa và không ảnh hưởng đến ý nghĩa của câu. Ngoài ra, những hàng có bình luận trùng lắp cũng sẽ bị loại khỏi dataset. Những chữ cái viết hoa hoặc những khoảng trắng lớn cũng sẽ được xử lý về dạng chữ cái thường, mỗi từ chỉ cách nhau một khoảng trắng. Ban đầu, thu thập dữ liệu từ Lazada khoảng 320.000 dòng, sau khi thực hiện bước loại bỏ giá trị trùng lắp và bỏ đi những dòng bình luận trống thì dataset còn 65.000 dòng.

A close-up of a document

Description automatically generated



4. Projected Impact

**4.1. Accomplishments and Benefits**

**Accomplishments**

* Chúng em đã thu thập và xử lý thành công một lượng lớn dữ liệu từ các bình luận sản phẩm bằng tiếng Việt, một nhiệm vụ không hề dễ dàng do đặc thù ngôn ngữ và sự đa dạng trong cách diễn đạt của người dùng. Việc xử lý đã giúp nhóm nâng cao kỹ năng xử lý dữ liệu thực tế, đồng thời làm phong phú thêm kho dữ liệu để phân tích.
* Dự án đã giúp chúng em có thêm những kiến thức mới, bước ra khỏi vùng an toàn của bản thân. Dự án là nền tảng để chúng em đạt thêm nhiều thành tựu từ các khóa học khác hoặc trong cuộc sống

**Benefits**

* Một trong những lợi ích lớn nhất của dự án là khả năng giúp người tiêu dùng đưa ra quyết định mua sắm thông minh hơn. Thay vì phải đọc qua hàng ngàn bình luận, hệ thống của chúng em sử dụng các thuật toán machine learning tiên tiến như SVM, Random Forest, và Logistic Regression để phân loại và đánh giá chất lượng sản phẩm dựa trên những bình luận thực tế. Điều này không chỉ tiết kiệm thời gian cho người tiêu dùng mà còn tăng cường độ tin cậy trong việc lựa chọn sản phẩm phù hợp với nhu cầu.
* Ngoài ra, dự án còn đem lại nhiều lợi ích khác cho các nhà bán lẻ và nhà sản xuất. Thông qua việc phân tích và hiểu rõ hơn về phản hồi của khách hàng, họ có thể điều chỉnh chiến lược kinh doanh, cải thiện chất lượng sản phẩm, và tối ưu hóa dịch vụ chăm sóc khách hàng. Khả năng xử lý Big Data của dự án cũng mở ra cơ hội để các doanh nghiệp áp dụng những phân tích sâu hơn, phát hiện xu hướng tiêu dùng và dự đoán nhu cầu thị trường một cách chính xác hơn.

**4.2. Future Improvements**

* Chúng em mong muốn cải tiến chương trình hơn là có thể tự động thu thập được tất cả bình luận của sản phẩm khi người dùng nhập liên kết của sản phẩm, được hợp tác với các sàn thương mại điện tử sẽ giúp chúng em thực hiện điều này một cách dễ dàng hơn.
* Bên cạnh đó, việc xử lý dữ liệu tiếng Việt, đặc biệt là các bình luận có chứa lỗi chính tả, viết tắt, hoặc tiếng lóng, vẫn còn là một thách thức lớn. Hiện tại, phần lớn quá trình này vẫn yêu cầu xử lý thủ công trong dự án lần này vì hầu như các công cụ đã ngừng hỗ trợ hoặc cập nhật ở nhiều năm trước. Do đó, một trong những cải tiến chính trong tương lai là phát triển và tích hợp các công cụ xử lý ngôn ngữ tự nhiên (NLP) mạnh mẽ hơn, giúp tự động hóa quá trình này một cách hiệu quả hơn. Điều này không chỉ giúp nâng cao độ chính xác của mô hình mà còn giảm bớt thời gian và công sức xử lý dữ liệu.
* Trong tương lai, chúng em sẽ thử nghiệm và triển khai thêm các mô hình machine learning tiên tiến khác như deep learning, đặc biệt là mạng neuron sâu (Deep Neural Networks), để cải thiện khả năng phân tích và dự đoán. Việc sử dụng các mô hình phức tạp hơn có thể giúp chúng em khai thác sâu hơn những thông tin ẩn chứa trong dữ liệu, từ đó đưa ra những dự đoán chính xác hơn.
* Hiện tại, hệ thống của chúng em chủ yếu phân tích các bình luận về sản phẩm bằng tiếng Việt. Trong tương lai, chúng em sẽ hướng đến việc mở rộng khả năng xử lý và phân tích bình luận bằng nhiều ngôn ngữ khác nhau, giúp dự án có tính ứng dụng toàn cầu

5. Team Member Review and Comment 

| <ATTACH A TEAM PICTURE HERE> |
| --- |

| NAME | REVIEW and COMMENT |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

6. Instructor Review and Comment

| CATEGORY | SCORE | REVIEW and COMMENT |
| --- | --- | --- |
| IDEA | \_\_/10 |  |
| APPLICATION | \_\_/30 |  |
| RESULT | \_\_/30 |  |
| PROJECT MANAGEMENT | \_\_/10 |  |
| PRESENTATION & REPORT | \_\_/20 |  |
| TOTAL | \_\_/100 |  |