**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**

**KHOA CÔNG NGHỆ THÔNG TIN**





**TIỂU LUẬN HỌC MÁY VÀ ỨNG DỤNG**

**ĐỀ TÀI: Hệ thống Phát hiện Bắt Nạt Trực Tuyến: Học Máy và Phân tích Văn Bản**

**Giảng viên hướng dẫn: Th.S BÙI TIẾN ĐỨC**

**Sinh viên thực hiện: PHAN THANH TÂM**

**MSSV: 2100008769**

**Chuyên ngành: Khoa học dữ liệu**

**Môn học: Học máy và ứng dụng**

**Khóa: 2021**

**Tp.HCM, 10 tháng 06 năm 2024**

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**

**KHOA CÔNG NGHỆ THÔNG TIN**





**TIỂU LUẬN KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG**

**ĐỀ TÀI: Hệ thống Phát hiện Bắt Nạt Trực Tuyến: Học Máy và Phân tích Văn Bản**

**Giảng viên hướng dẫn:** **Th.S BÙI TIẾN ĐỨC**

**Sinh viên thực hiện: PHAN THANH TÂM**

**MSSV: 2100008769**

**Chuyên ngành: Khoa học dữ liệu**

**Môn học: Học máy và ứng dụng**

**Khóa: 2021**

**Tp.HCM, 10 tháng 06 năm 2024**

# LỜI CẢM ƠN

Em xin bày tỏ lòng biết ơn chân thành và sâu sắc tới tất cả các **Thầy Cô tại Khoa Công Nghệ Thông Tin, trường Đại học Nguyễn Tất Thành** đã dành thời gian và kiến thức quý báu để chia sẻ với chúng em trong suốt quá trình học tập và nghiên cứu tại trường. Sự cống hiến và tâm huyết của các Thầy Cô đã giúp chúng em trưởng thành và phát triển không chỉ về mặt chuyên môn mà còn về kỹ năng thực tiễn.

Đặc biệt, em xin gửi lời cảm ơn sâu sắc và trân trọng nhất tới **Thầy Th.S Bùi Tiến Đức**. Thầy đã ân cần hướng dẫn và truyền đạt cho em những kiến thức và kỹ năng cần thiết để hoàn thành dự án Hệ thống Phát hiện Bắt Nạt Trực Tuyến: Học Máy và Phân tích Văn Bản. Dưới sự chỉ dẫn của Thầy, em không chỉ nắm vững kiến thức về học máy và phân tích văn bản, mà còn học được cách tiền xử lý dữ liệu, trực quan hóa dữ liệu và ứng dụng học máy vào dữ liệu một cách hiệu quả.

Nhờ vào sự dạy bảo và hỗ trợ tận tình của Thầy, em đã phát triển được khả năng nghiên cứu và phân tích một cách tự tin và sâu sắc hơn. Lời cảm ơn này cũng xin được gửi đến tất cả các Thầy Cô đã tạo điều kiện tốt nhất cho sự phát triển của chúng em, từ việc tài trợ tài liệu, hỗ trợ kỹ thuật cho đến sự khích lệ và định hướng đúng đắn trong quá trình nghiên cứu.

Xin chân thành cảm ơn và kính chúc quý Thầy Cô ngày càng gặt hái nhiều thành công và phát triển vững mạnh trong sứ mệnh cao quý của mình.

Em xin chân thành cảm ơn.

TP.HCM, ngày 10 tháng 06 năm 2024

# LỜI MỞ ĐẦU

Trong thời đại số hóa hiện nay, khi mạng xã hội và các nền tảng trực tuyến trở thành một phần thiết yếu của cuộc sống hàng ngày, vấn đề bảo vệ sự an toàn trên mạng đã trở thành ưu tiên hàng đầu. Một trong những thách thức nghiêm trọng hiện nay là vấn đề bắt nạt trực tuyến, hay cyberbullying, một hiện tượng ngày càng phổ biến và ảnh hưởng sâu rộng đến cả sức khỏe tâm lý và chất lượng cuộc sống của nhiều người, đặc biệt là giới trẻ.

Báo cáo này tập trung vào việc phát triển hệ thống phát hiện bắt nạt trực tuyến bằng cách áp dụng các kỹ thuật học máy và phân tích văn bản. Em đã tiến hành một nghiên cứu sâu rộng để khám phá các phương pháp và thuật toán học máy có thể giúp nhận diện và phân loại các hành vi bắt nạt trên nền tảng mạng xã hội và các diễn đàn trực tuyến. Bằng cách sử dụng phân tích văn bản, em đã khai thác các đặc điểm ngữ nghĩa và ngữ cảnh của các thông điệp để cải thiện khả năng phát hiện và ứng phó với các hành vi bắt nạt.

Mục tiêu của báo cáo là cung cấp một cái nhìn toàn diện về cách các công nghệ học máy và phân tích văn bản có thể được ứng dụng để cải thiện việc phát hiện và ngăn chặn bắt nạt trực tuyến, từ đó góp phần xây dựng một môi trường mạng an toàn và lành mạnh hơn. Em kỳ vọng rằng nghiên cứu này sẽ đóng góp vào các giải pháp hiệu quả cho vấn đề bắt nạt trực tuyến và là nền tảng cho các nghiên cứu và ứng dụng tiếp theo trong lĩnh vực này.

Em xin gửi lời cảm ơn chân thành đến các **Thầy Cô tại Khoa Công Nghệ Thông Tin, trường Đại học Nguyễn Tất Thành**, và đặc biệt là **Thầy Th.S Bùi Tiến Đức**, người đã cung cấp sự hỗ trợ quý báu và hướng dẫn tận tình trong suốt quá trình thực hiện nghiên cứu. Sự hỗ trợ của các thầy cô đã đóng vai trò quan trọng trong việc hoàn thành dự án này và là nguồn động viên lớn lao cho em.

# NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

Điểm đồ án:

*TPHCM, Ngày …… tháng …… năm*

**Giáo viên hướng dẫn**

(Ký tên, đóng dấu)

|  |  |
| --- | --- |
| TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH  **TRUNG TÂM KHẢO THÍ** | **KỲ THI KẾT THÚC HỌC PHẦN**  **HỌC KỲ 3 – NĂM HỌC 20232024** |

# PHIẾU CHẤM THI TIỂU LUẬN/ĐỒ ÁN

Môn thi: **Học máy và ứng dụng**

Lớp học phần: **21DTH2C**

Sinh viên thực hiện : Phan Thanh Tâm Mssv:2100008769

Ngày thi:04/09/2024

Phòng thi:L.512

Đề tài tiểu luận/báo cáo của sinh viên : Hệ thống Phát hiện Bắt Nạt Trực Tuyến: Học Máy và Phân tích Văn Bản

Phần đánh giá của giảng viên (căn cứ trên thang rubrics của môn học):

|  |  |  |  |
| --- | --- | --- | --- |
| **Tiêu chí (theo CĐR HP)** | **Đánh giá của GV** | **Điểm tối đa** | **Điểm đạt được** |
| Cấu trúc của báo cáo |  | 2 |  |
| Nội dung |  |  |  |
| * Các nội dung thành phần |  | 2 |  |
| * Lập luận |  | 4 |  |
| * Kết luận |  | 1 |  |
| Trình bày |  | 1 |  |
| **TỔNG ĐIỂM** |  | 10 |  |

|  |  |
| --- | --- |
|  | **Giảng viên chấm thi**  *(ký, ghi rõ họ tên)* |

Th.S Bùi Tiến Đức

MỤC LỤC

[- LỜI CẢM ƠN 3](#_Toc175692805)

[- LỜI MỞ ĐẦU 4](#_Toc175692806)

[- NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN 5](#_Toc175692807)

[- PHIẾU CHẤM THI TIỂU LUẬN/ĐỒ ÁN 6](#_Toc175692808)

[MỤC LỤC 7](#_Toc175692809)

[DANH MỤC HÌNH 8](#_Toc175692810)

[DANH MỤC CHỮ VIẾT TẮT 9](#_Toc175692811)

[Chương 1. Giới thiệu 1](#_Toc175692812)

[1.1. Lý Do Chọn Đề Tài 1](#_Toc175692813)

[1.2. Mục Tiêu Và Nhiệm Vụ Nghiên Cứu 1](#_Toc175692814)

[1.2.1. Đối tượng và phạm vi nghiên cứu 2](#_Toc175692815)

[1.2.2. Phương pháp nghiên cứu 2](#_Toc175692816)

[1.3. Ý nghĩa khoa học và thực tiễn 2](#_Toc175692817)

[Chương 2. Cơ sở lý thuyết 3](#_Toc175692818)

[2.1. Giới thiệu về Bình Luận 3](#_Toc175692819)

[2.2. Công nghệ phần mềm, công cụ phân tích được sử dụng 4](#_Toc175692820)

[2.3. Biểu đồ Tròn (Pie Chart) và Biểu đồ Nhiệt (Heatmap) 5](#_Toc175692821)

[2.4. Mô hình học máy 6](#_Toc175692822)

[Chương 3. Thực Nghiệm 10](#_Toc175692823)

[3.1. Phân tích dữ liệu hiện tại 10](#_Toc175692824)

[3.2. Kiểm tra và tiền sử lý dữ liệu 11](#_Toc175692825)

[3.3. Triển khai trực quan hóa dữ liệu. 12](#_Toc175692826)

[3.4. Phát triển và đánh giá mô hình 15](#_Toc175692827)

[3.4.1. Xây dựng mô hình 15](#_Toc175692828)

[3.4.2. Đánh giá mô hình, dự đoán và phân tích 17](#_Toc175692829)

[Chương 4. Kết Luận 18](#_Toc175692830)

[Tài Liệu Tham Khảo 20](#_Toc175692831)

# DANH MỤC HÌNH

[Hình 1: Naive Bayes. 6](#_Toc175692940)

[Hình 2: Xác suất posterior. 6](#_Toc175692941)

[Hình 3: Xác suất của đặc trưng (P(X | C\_k)). 7](#_Toc175692942)

[Hình 4: Xác suất của đặc trưng (P(x\_i | C\_k)) trong mô hình Multinomial. 7](#_Toc175692943)

[Hình 5: Support Vector Machine (SVM). 8](#_Toc175692944)

[Hình 6: Siêu phẳng phân chia. 8](#_Toc175692945)

[Hình 7: Hàm mất mát (Loss Function): SVM sử dụng hàm mất mát hinge loss. 8](#_Toc175692946)

[Hình 8: Tối ưu hóa: SVM. 9](#_Toc175692947)

[Hình 9: Random Forest. 9](#_Toc175692948)

[Hình 10: Tiêu chí phân chia. 10](#_Toc175692949)

[Hình 11: Dữ liệu ban đầu. 10](#_Toc175692950)

[Hình 12: Hàm tiền sử lý văn bản. 11](#_Toc175692951)

[Hình 13: Xử lý giá trị thiếu (NaN) và tiền xử lý văn bản. 11](#_Toc175692952)

[Hình 14: Chuyển đổi văn bản thành dạng số. 11](#_Toc175692953)

[Hình 15: Mã hóa nhãn và chia dữ liệu. 12](#_Toc175692954)

[Hình 16: In dữ liệu sau tiền sử lý. 12](#_Toc175692955)

[Hình 17: Dữ liệu sau khi tiền sử lý. 12](#_Toc175692956)

[Hình 18: Code minh họa biểu đồ tròn(pie chart). 13](#_Toc175692957)

[Hình 19: Biểu đồ tròn(pie chart). 13](#_Toc175692958)

[Hình 20: Code minh họa Biểu đồ nhiệt (heatmap). 14](#_Toc175692959)

[Hình 21: Biểu đồ nhiệt (heatmap). 14](#_Toc175692960)

[Hình 22: Xây dựng mô hình học máy. 15](#_Toc175692961)

[Hình 23: Kết quả mô hình. 16](#_Toc175692962)

[Hình 24: Đánh giá mô hình. 17](#_Toc175692963)

[Hình 25: Dự đoán các văn bản mới. 17](#_Toc175692964)

# DANH MỤC CHỮ VIẾT TẮT

(SVM) Support Vector Machine.

(MultinomialNB) Naive Bayes.

(SVC) Support Vector Classification.

(NaN) Giá trị thiếu.

(TFIDF) Term Frequency-Inverse Document Frequency.

# Chương 1. Giới thiệu

## 1.1. Lý Do Chọn Đề Tài

Việc lựa chọn đề tài "Phát triển Hệ thống Phát hiện Bắt nạt Trực tuyến bằng Kỹ thuật Học Máy và Phân tích Văn bản" được dựa trên các lý do:

Tính chất quan trọng và cấp bách của vấn đề: Bắt nạt trực tuyến đã trở thành một vấn đề nghiêm trọng, ảnh hưởng sâu rộng đến sức khỏe tâm lý và chất lượng cuộc sống của nhiều người, đặc biệt là giới trẻ. Việc phát triển hệ thống phát hiện và ngăn chặn hành vi này không chỉ góp phần bảo vệ người dùng trên mạng mà còn nâng cao sự an toàn và lành mạnh của môi trường trực tuyến.

Thách thức và cơ hội trong lĩnh vực công nghệ: Ngành công nghệ thông tin, đặc biệt là trong lĩnh vực học máy và phân tích văn bản, đang phát triển nhanh chóng. Xây dựng hệ thống phát hiện bắt nạt trực tuyến không chỉ là một thách thức kỹ thuật mà còn mở ra cơ hội ứng dụng công nghệ tiên tiến để giải quyết các vấn đề xã hội cấp thiết. Đề tài này cung cấp cái nhìn về cách công nghệ có thể được áp dụng để giải quyết những vấn đề thực tiễn.

Áp dụng kiến thức học được: Dự án này yêu cầu sử dụng các kỹ thuật học máy và phân tích văn bản để nhận diện và phân loại các hành vi bắt nạt. Việc thực hiện nghiên cứu này giúp củng cố và mở rộng kỹ năng trong lĩnh vực học máy và xử lý ngôn ngữ tự nhiên, đồng thời tạo cơ hội cho sinh viên hoặc nhà nghiên cứu phát triển kỹ năng phân tích và giải quyết vấn đề trong một lĩnh vực đang phát triển nhanh chóng.

Ứng dụng thực tế và giá trị xã hội: Kết quả của nghiên cứu có thể được áp dụng trong thực tế để xây dựng các công cụ và giải pháp giúp nhận diện và ngăn chặn bắt nạt trực tuyến hiệu quả hơn. Điều này không chỉ mang lại giá trị cho các tổ chức và nền tảng trực tuyến mà còn đóng góp vào việc xây dựng một cộng đồng mạng an toàn và tích cực hơn cho mọi người.

## 1.2. Mục Tiêu Và Nhiệm Vụ Nghiên Cứu

Mục tiêu:

Phát triển Hệ thống Phát hiện Bắt nạt Trực tuyến: Xây dựng một hệ thống sử dụng kỹ thuật học máy và phân tích văn bản để tự động nhận diện và phân loại các hành vi bắt nạt trực tuyến trên nền tảng mạng xã hội và diễn đàn trực tuyến.

Cải thiện Chính xác và Hiệu quả: Tối ưu hóa khả năng của hệ thống trong việc phát hiện và phân loại hành vi bắt nạt bằng cách sử dụng các mô hình học máy tiên tiến và khai thác các đặc điểm ngữ nghĩa và ngữ cảnh của văn bản.

Cung cấp Giải pháp Thực tiễn: Đưa ra các khuyến nghị và giải pháp cụ thể nhằm cải thiện các công cụ và chính sách hiện tại liên quan đến việc phòng chống và ứng phó với bắt nạt trực tuyến.

Nhiệm vụ:

Tiền xử lý Dữ liệu:

Chuyển đổi văn bản thành chữ thường.

Loại bỏ ký tự đặc biệt và placeholder.

Xử lý giá trị thiếu.

Trực Quan Hóa Dữ Liệu:

Biểu đồ tròn.

Biểu đồ nhiệt (Heatmap).

Học Máy:

Naive Bayes: Phương pháp phân loại dựa trên xác suất, thường được sử dụng cho văn bản.

SVM (Support Vector Machine): Phương pháp phân loại mạnh mẽ với khả năng tìm ranh giới phân lớp tốt nhất.

Random Forest: Mô hình ensemble sử dụng nhiều cây quyết định để cải thiện độ chính xác và giảm overfitting.

Huấn luyện và đánh giá mô hình.

Dự đoán với mô hình tốt nhất: Sử dụng mô hình học máy phù hợp nhất để dự đoán nhãn cho các bình luận mới và kiểm tra hiệu suất của mô hình trên dữ liệu chưa thấy trước đó.

### 1.2.1. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu:

Người dùng mạng xã hội và diễn đàn trực tuyến.

Các nhà nghiên cứu và sinh viên: Những người quan tâm đến nghiên cứu về lĩnh vực khoa học dữ liệu, thống kê, hoặc học máy và phân tích văn bản, và muốn áp dụng kiến thức của họ vào việc việc nghiên cứu và phát triển các giải pháp công nghệ để phát hiện và ngăn chặn hành vi bắt nạt trực tuyến.

Phạm vi nghiên cứu:

Bài báo cáo nghiên cứu sâu sắc và phân tích nội dung và hình thức bắt nạt trực tuyến, kỹ thuật và công nghệ học máy, đánh giá và so sánh hiệu suất, ứng dụng và giải pháp thực tiễn.

### 1.2.2. Phương pháp nghiên cứu

Tìm kiếm tài liệu trên internet có liên quan đến nội dung nghiên cứu

Thu thập tài liệu liên quan đến nghiệp vụ quản lý kiểm tra.

Phân Tích và Đánh Giá Các Phương Pháp và Kỹ Thuật: Nghiên cứu các phương pháp thống kê và học máy, phân tích dữ liệu và dự đoán

Phát triển và triển khai hệ thống: Xây dựng mô hình phát hiện, thực hiện kiểm tra và đánh giá.

## 1.3. Ý nghĩa khoa học và thực tiễn

**Ý nghĩa khoa học:**

Nâng cao Hiểu Biết về Bắt Nạt Trực Tuyến: Nghiên cứu này cung cấp cái nhìn sâu sắc về các hình thức và mẫu hành vi bắt nạt trực tuyến thông qua việc áp dụng các kỹ thuật học máy và phân tích văn bản. Điều này giúp mở rộng hiểu biết về cách các hành vi này diễn ra và cách chúng có thể được nhận diện và phân loại một cách hiệu quả.

Thúc Đẩy Nghiên Cứu Trong Lĩnh Vực An Ninh Mạng và Học Máy: Bằng việc áp dụng các kỹ thuật học máy và phân tích văn bản vào việc phát hiện bắt nạt trực tuyến, nghiên cứu này không chỉ đóng góp vào sự phát triển của các công nghệ phát hiện và phòng chống bắt nạt mà còn thúc đẩy nghiên cứu trong các lĩnh vực học máy và xử lý ngôn ngữ tự nhiên.

Xây Dựng Cơ Sở Cho Các Nghiên Cứu Tiếp Theo: Kết quả của nghiên cứu này có thể trở thành nền tảng cho các nghiên cứu tiếp theo trong việc cải tiến hệ thống phát hiện hành vi xâm hại trực tuyến, cũng như phát triển các phương pháp mới trong việc xử lý và phân tích nội dung mạng xã hội.

**Ý nghĩa thực tiễn:**

Cải Thiện Công Cụ và Chính Sách An Ninh Mạng: Các tổ chức và nền tảng trực tuyến có thể sử dụng kết quả nghiên cứu để cải thiện các công cụ và chính sách quản lý nội dung của họ, nhằm phát hiện và ứng phó hiệu quả hơn với các hành vi bắt nạt trực tuyến. Điều này giúp xây dựng một môi trường trực tuyến an toàn hơn cho người dùng.

Hỗ Trợ Quản Lý và Giám Sát Nội Dung: Các nhà quản lý và quản trị viên nền tảng mạng xã hội có thể sử dụng thông tin từ nghiên cứu này để phát triển và áp dụng các phương pháp hiệu quả hơn trong việc giám sát và kiểm soát nội dung, từ đó giảm thiểu tình trạng bắt nạt và bảo vệ người dùng.

Hỗ Trợ Cá Nhân và Cộng Đồng: Các cá nhân và cộng đồng mạng có thể sử dụng các công cụ phát hiện và giải pháp được phát triển từ nghiên cứu này để bảo vệ mình khỏi hành vi bắt nạt trực tuyến. Điều này cung cấp các công cụ hữu ích cho việc nhận diện và báo cáo hành vi xâm hại, giúp nâng cao nhận thức và khả năng phòng chống bắt nạt trong cộng đồng.

# Chương 2. Cơ sở lý thuyết

## 2.1. Giới thiệu về Bình Luận

**Bình luận (comment)** trên mạng xã hội, website, hay các báo điện tử, được dùng để bạn đọc chia sẻ cảm xúc, bày tỏ ý kiến, trao đổi, phản biện, tranh luận về một vấn đề nào đó.

Nhưng trên thực tế, còn có nhiều loại “comment bẩn” (hay là rác bình luận) khác mà ai cũng từng gặp và đang xuất hiện ngày càng nhiều. Dễ nhận biết nhất là những comment tục tĩu, bậy bạ, vô văn hóa, văng tục, chửi thề trên mạng, với những lời lẽ dữ dằn có thể gây sốc và choáng cho người tiếp nhận. Dạng này xuất hiện khá nhiều trên MXH, nhưng với các diễn đàn online có “bộ lọc” tốt, có bộ phận chuyên trách quản lý, biên tập, duyệt đăng các bình luận, thì hầu hết bình luận thô tục kiểu này bị xóa bỏ vì không phù hợp.

Lại có những người không phải “bình luận” mà chỉ muốn “chửi cho sướng miệng”. Đó là chưa biết đúng –sai, thực – hư thế nào, đã “tay nhanh hơn não” comment chửi bới vô tội vạ, kết án, bôi nhọ, xúc phạm danh dự cá nhân, tổ chức khác. Bình luận hay tranh luận đều cần có văn hóa và tôn trọng sự khác biệt, nhưng nhiều người chỉ cần ý kiến khác biệt với mình là… chửi. Có người còn lợi dụng một vụ việc để công kích chính quyền, xuyên tạc một chủ trương, chính sách đúng đắn của Đảng Nhà nước.

Tinh vi hơn, khó nhận biết hơn và nguy hại hơn, một dạng “comment bẩn” nữa là cố tình phát tán các tin giả, bịa đặt lên mạng (hoặc thêm thắt tình tiết vào các vụ án, vụ việc), rồi tạo ra các bình luận để lôi kéo, kích động đám đông mạng hùa theo “ném đá”, nhằm vu khống, xúc phạm một tổ chứccá nhân nào đó. Loại này rất nguy hiểm có thể gây bất ổn xã hội, nên chính quyền và công an thường khuyến cáo người dân cần tỉnh táo, thận trọng trước khi bình luận, chia sẻ một thông tin nhạy cảm, chưa được kiểm chứng, dẫn tới hậu quả đáng tiếc.

## 2.2. Công nghệ phần mềm, công cụ phân tích được sử dụng

Ngôn ngữ lập trình python:

Python là một ngôn ngữ lập trình bậc cao được sử dụng rộng rãi trong nhiều lĩnh vực khác nhau như phát triển web, phần mềm, tự động hóa các tác vụ và phân tích dữ liệu. Python là một ngôn ngữ lập trình đa mục đích, có khả năng phát triển nhiều loại chương trình khác nhau mà không bị giới hạn bởi một vấn đề cụ thể nào.

**Thư viện được sử dụng ở đồ án:**

**Pandas:** Được sử dụng để thao tác và phân tích dữ liệu dạng bảng. Pandas cung cấp các cấu trúc dữ liệu như DataFrame và Series giúp việc xử lý dữ liệu trở nên dễ dàng và hiệu quả.

**import pandas as pd**

**Matplotlib:** Thư viện cơ bản và mạnh mẽ để tạo ra các biểu đồ, đồ thị trong Python. Matplotlib cho phép tạo ra các biểu đồ từ đơn giản đến phức tạp. Được sử dụng để vẽ biểu đồ tròn (pie charts) và biểu đồ nhiệt (heatmap).

**import matplotlib.pyplot as plt**

**Seaborn:** Một thư viện được xây dựng trên Matplotlib, cung cấp giao diện trực quan và dễ sử dụng hơn cho việc tạo ra các biểu đồ thống kê. Được sử dụng để vẽ biểu đồ nhiệt (heatmap) với phong cách và màu sắc đẹp mắt.

**import seaborn as sns**

**scikitlearn:** Scikitlearn: Được sử dụng cho tiền xử lý văn bản (TFIDF), mã hóa nhãn (LabelEncoder), chia dữ liệu (train\_test\_split), xây dựng và đánh giá các mô hình học máy (Naive Bayes, SVM, Random Forest).

**from sklearn.feature\_extraction.text import TfidfVectorizer**

**from sklearn.preprocessing import LabelEncoder**

**from sklearn.model\_selection import train\_test\_split**

**from sklearn.naive\_bayes import MultinomialNB**

**from sklearn.metrics import classification\_report**

**from sklearn.svm import SVC**

**from sklearn.ensemble import RandomForestClassifier**

Tiền xử lý văn bản: Sử dụng TfidfVectorizer từ sklearn.feature\_extraction.text.

Mã hóa nhãn: Sử dụng LabelEncoder từ sklearn.preprocessing.

Chia dữ liệu: Sử dụng train\_test\_split từ sklearn.model\_selection.

Xây dựng và đánh giá mô hình: Sử dụng các mô hình học máy từ sklearn.naive\_bayes, sklearn.svm, và sklearn.ensemble, cùng với các công cụ đánh giá từ sklearn.metrics.

## 2.3. Biểu đồ Tròn (Pie Chart) và Biểu đồ Nhiệt (Heatmap)

**Biểu Đồ Tròn (Pie Chart)**

Định Nghĩa:

Biểu đồ tròn là một loại biểu đồ được sử dụng để thể hiện tỷ lệ phần trăm của các phần trong tổng thể. Biểu đồ này có hình dạng giống như một chiếc bánh tròn và được chia thành các phần tương ứng với tỷ lệ phần trăm của từng thành phần.

Quá Trình Xây Dựng:

Xác định các danh mục và giá trị liên quan.

Chuyển đổi các giá trị thành tỷ lệ phần trăm của tổng giá trị.

Tạo hình tròn và chia thành các phần với kích thước tương ứng với tỷ lệ phần trăm.

Thêm nhãn cho từng phần và cung cấp thông tin về tỷ lệ phần trăm của từng phần.

Ưu Điểm:

Dễ dàng trực quan hóa tỷ lệ phần trăm của các phần trong tổng thể.

Hiệu quả với số lượng danh mục hạn chế.

Nhược Điểm:

Khó so sánh các phần với nhau khi số lượng danh mục nhiều.

Không hiệu quả với dữ liệu lớn hoặc dữ liệu có sự thay đổi mạnh mẽ.

**Biểu Đồ Nhiệt (Heatmap)**

Định Nghĩa:

Biểu đồ nhiệt là một loại biểu đồ sử dụng màu sắc để thể hiện giá trị của dữ liệu trong một ma trận. Mỗi ô trong ma trận được tô màu theo giá trị của nó, giúp người xem nhanh chóng nhận biết các mẫu và xu hướng.

Quá Trình Xây Dựng:

Xác định các biến số cần so sánh và thu thập dữ liệu.

Tạo ma trận với các hàng và cột tương ứng với các biến số và giá trị.

Gán các giá trị dữ liệu vào ma trận và tô màu các ô dựa trên giá trị của chúng, thường sử dụng một bảng màu gradient.

Sử dụng phần mềm hoặc công cụ trực quan hóa để tạo heatmap từ ma trận dữ liệu.

Ưu Điểm:

Hiệu quả trong việc phân tích dữ liệu đa chiều và nhận diện các mẫu và xu hướng.

Có thể hiển thị dữ liệu lớn và phức tạp một cách rõ ràng.

Nhược Điểm:

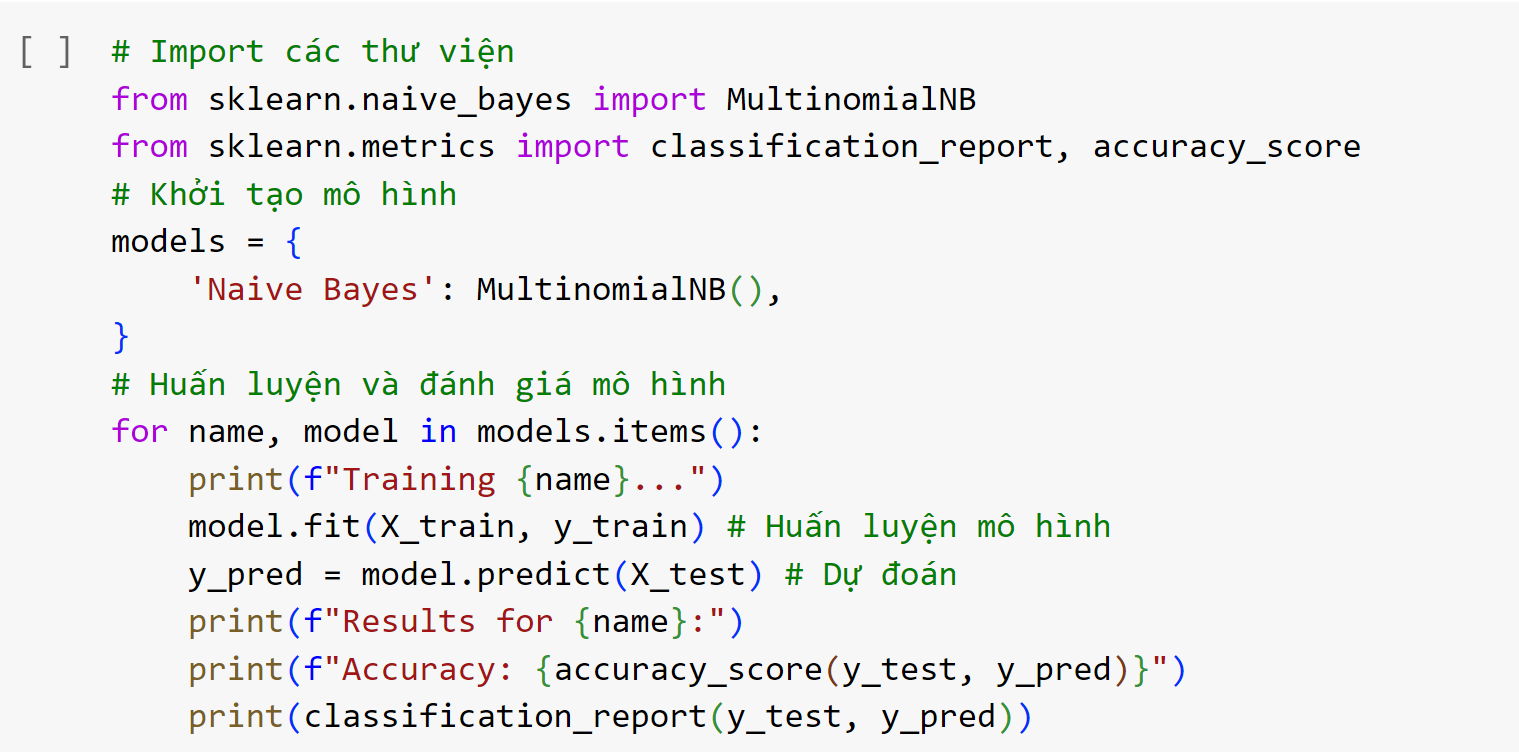
Có thể khó đọc nếu ma trận quá lớn hoặc màu sắc không được chọn phù hợp.

Không thích hợp cho dữ liệu có số lượng rất nhỏ hoặc đơn giản.

## 2.4. Mô hình học máy

**Naive Bayes:**

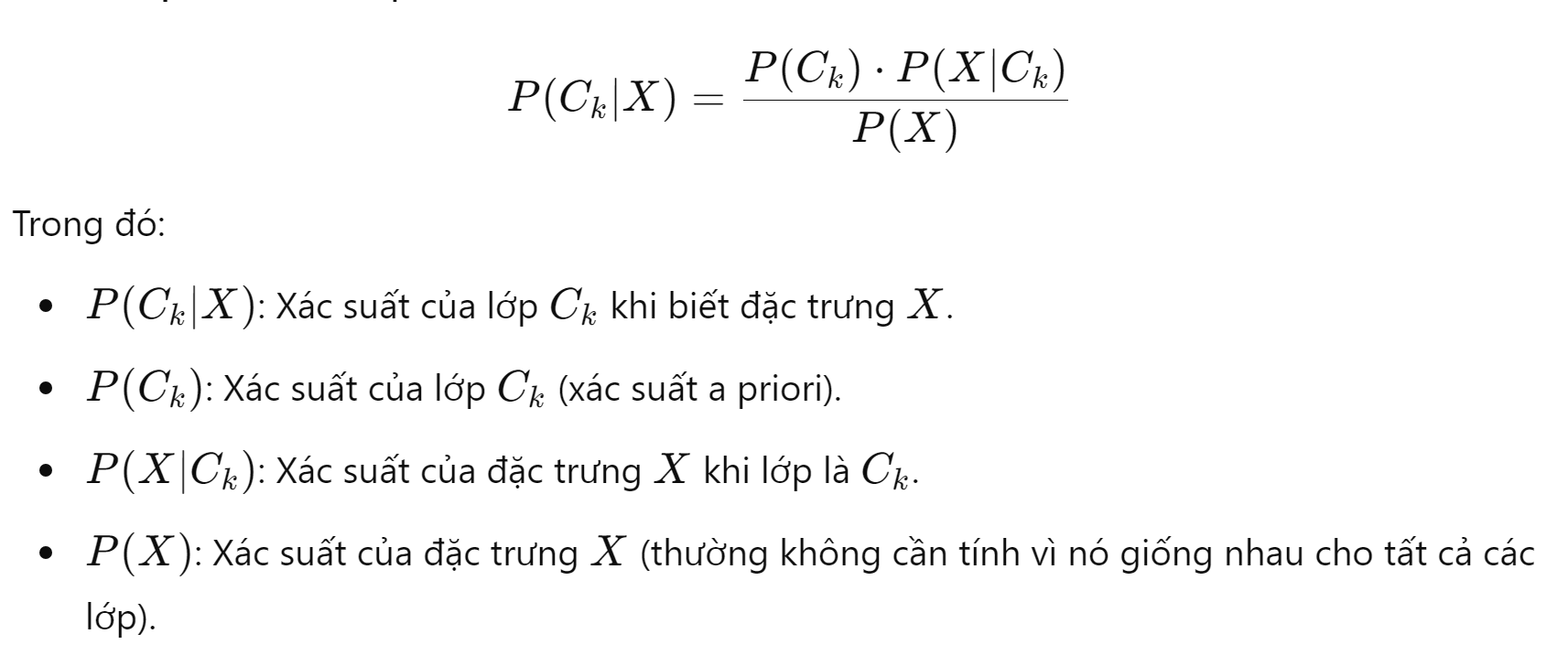
Naive Bayes là một mô hình phân loại dựa trên định lý Bayes với giả định rằng các đặc trưng (từ khóa trong văn bản) là độc lập với nhau. MultinomialNB là biến thể của Naive Bayes được sử dụng cho dữ liệu phân loại nhiều lớp và thường được áp dụng cho bài toán phân loại văn bản.

Code:   


Hình 1: Naive Bayes.

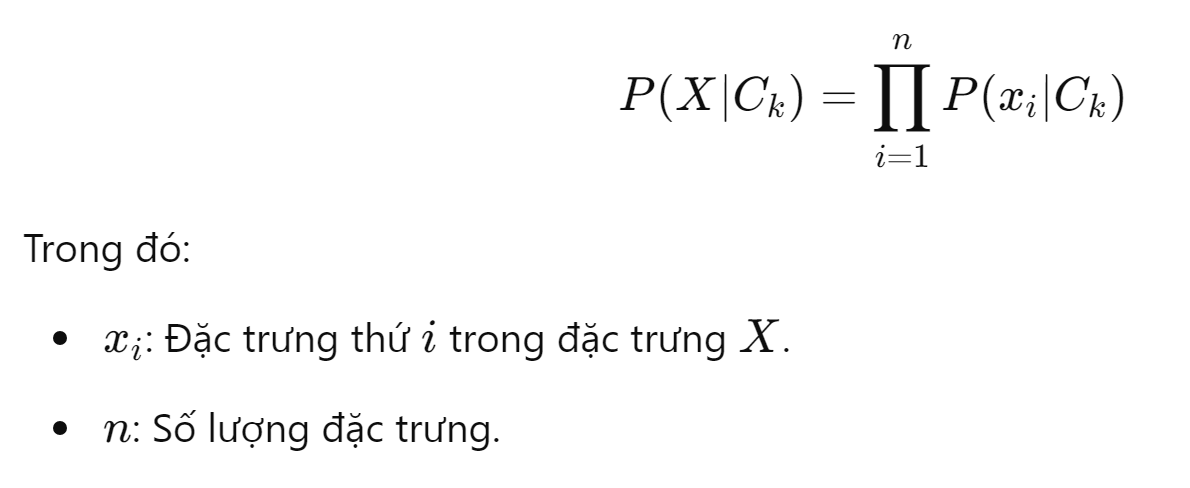
Công thức tính: Naive Bayes dựa trên định lý Bayes và giả định độc lập giữa các đặc trưng.

Xác suất posterior (P(C|X)):



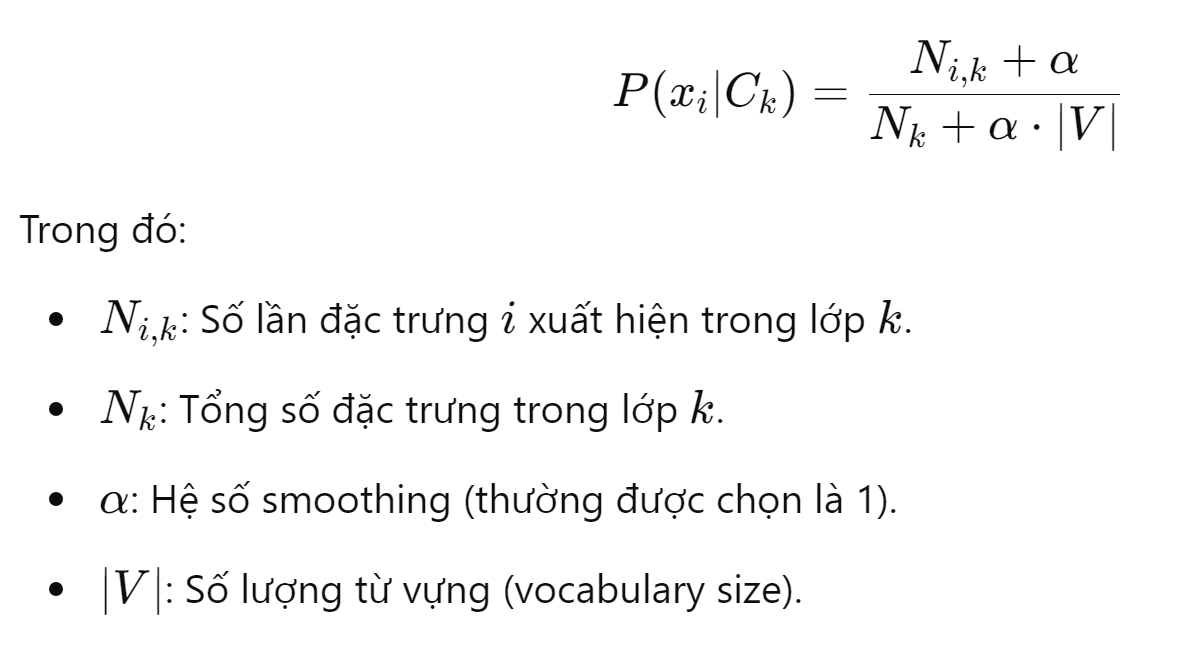
Hình 2: Xác suất posterior.

Xác suất của đặc trưng (P(X | C\_k)):



Hình 3: Xác suất của đặc trưng (P(X | C\_k)).

Xác suất của đặc trưng (P(x\_i | C\_k)) trong mô hình Multinomial:

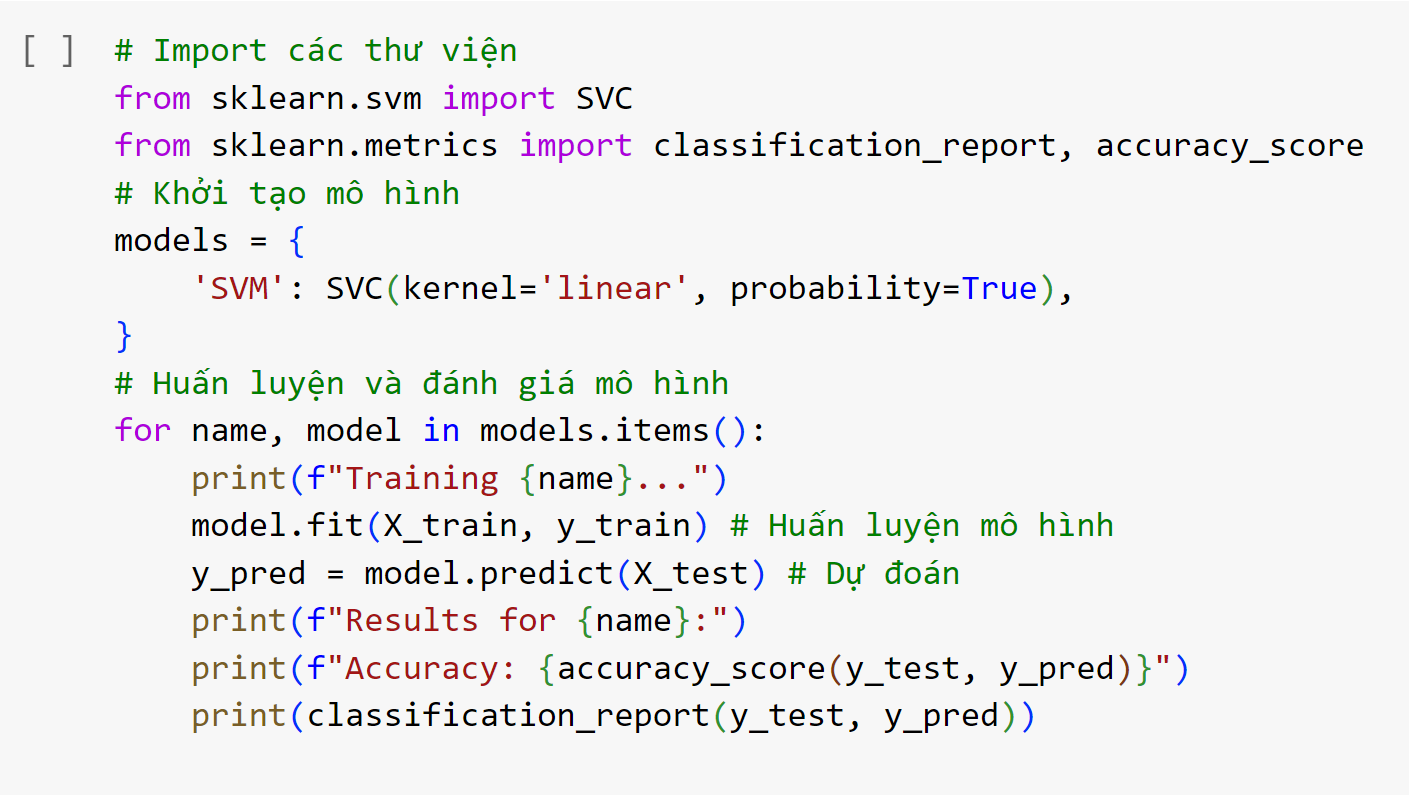


Hình 4: Xác suất của đặc trưng (P(x\_i | C\_k)) trong mô hình Multinomial.

**Support Vector Machine (SVM):**

SVM là một mô hình phân loại mạnh mẽ, tìm kiếm siêu phẳng (hyperplane) tối ưu để phân chia các lớp trong không gian đặc trưng. SVC (Support Vector Classification) với kernel tuyến tính (linear) là một biến thể của SVM thường được sử dụng cho các bài toán phân loại.

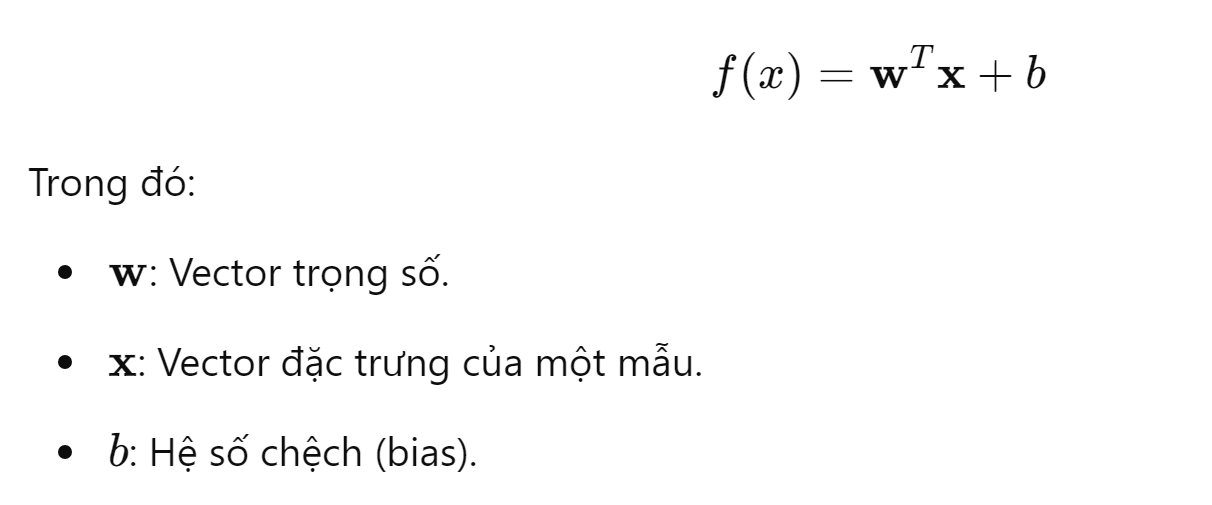
Code:



Hình 5: Support Vector Machine (SVM).

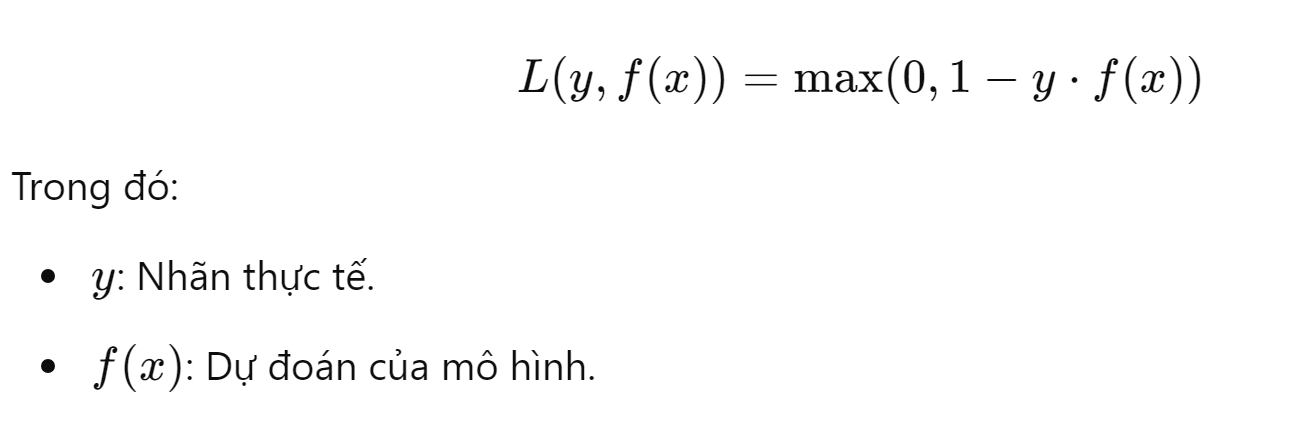
Công thức tính: SVM tìm siêu phẳng tối ưu để phân chia các lớp. Đối với SVM với kernel tuyến tính.

Siêu phẳng phân chia:



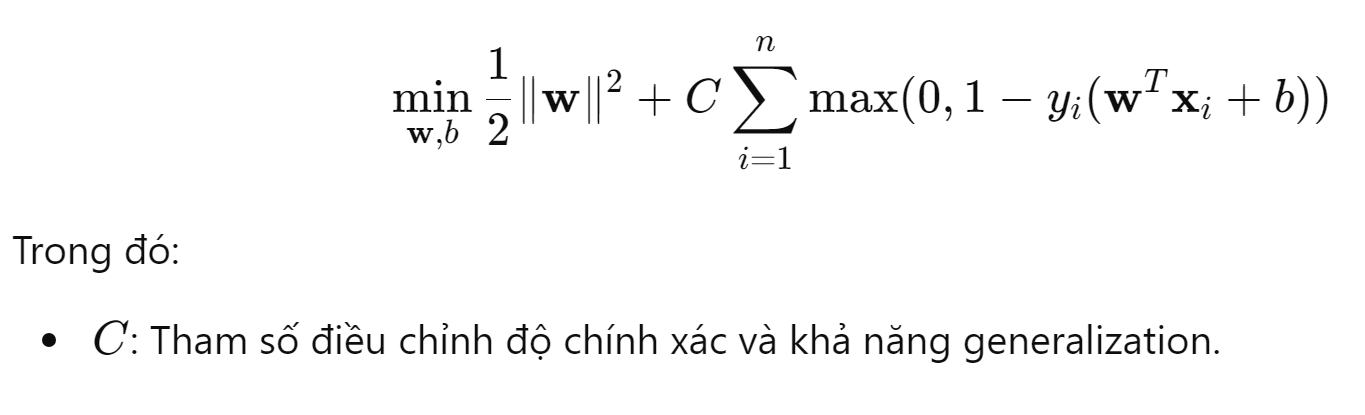
Hình 6: Siêu phẳng phân chia.

Hàm mất mát (Loss Function): SVM sử dụng hàm mất mát hinge loss:



Hình 7: Hàm mất mát (Loss Function): SVM sử dụng hàm mất mát hinge loss.

Tối ưu hóa: SVM tối ưu hóa hàm mất mát và thêm điều kiện ràng buộc để đảm bảo siêu phẳng phân chia đúng.

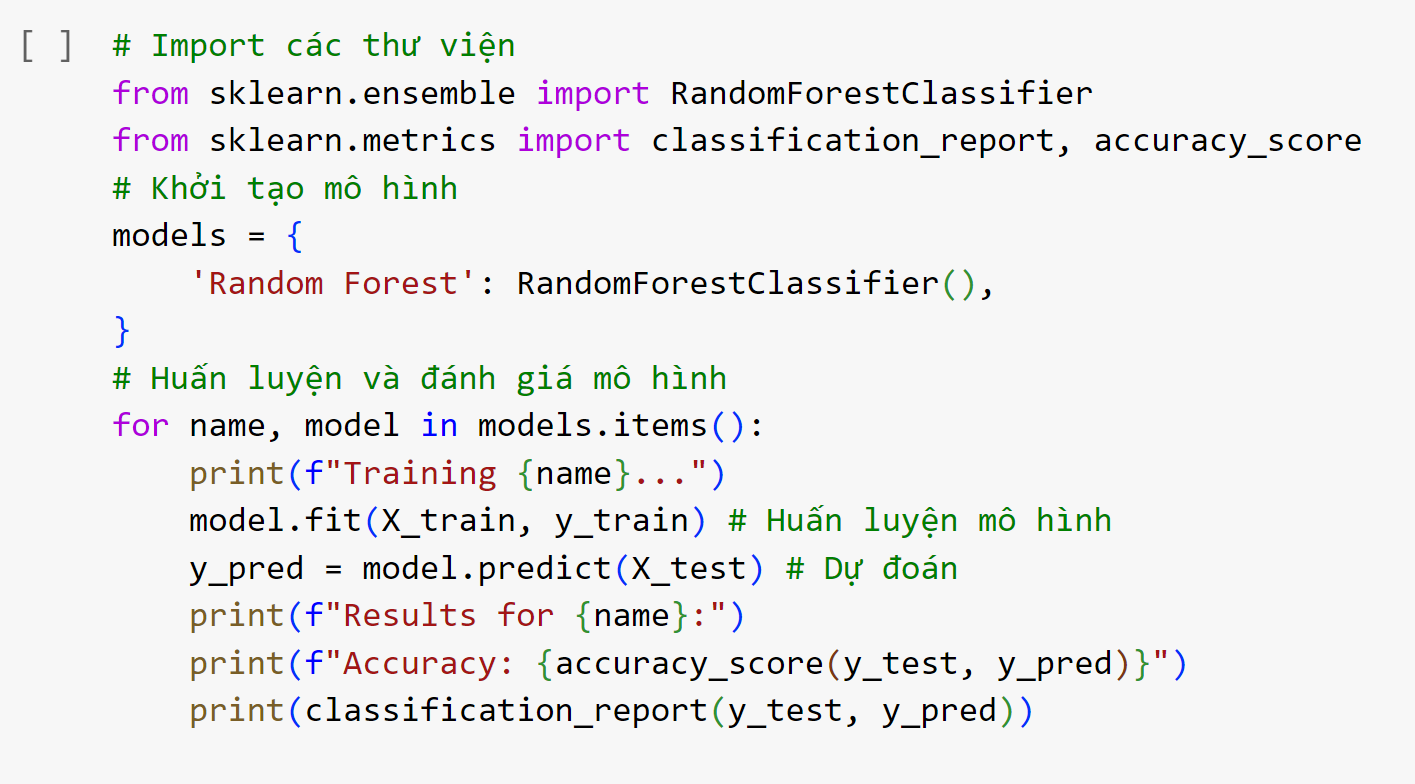


Hình 8: Tối ưu hóa: SVM.

**Random Forest:**

Random Forest là một mô hình phân loại dựa trên việc kết hợp nhiều cây quyết định để đưa ra dự đoán. Mô hình này thường hoạt động tốt với dữ liệu có nhiều đặc trưng và giúp giảm nguy cơ overfitting.

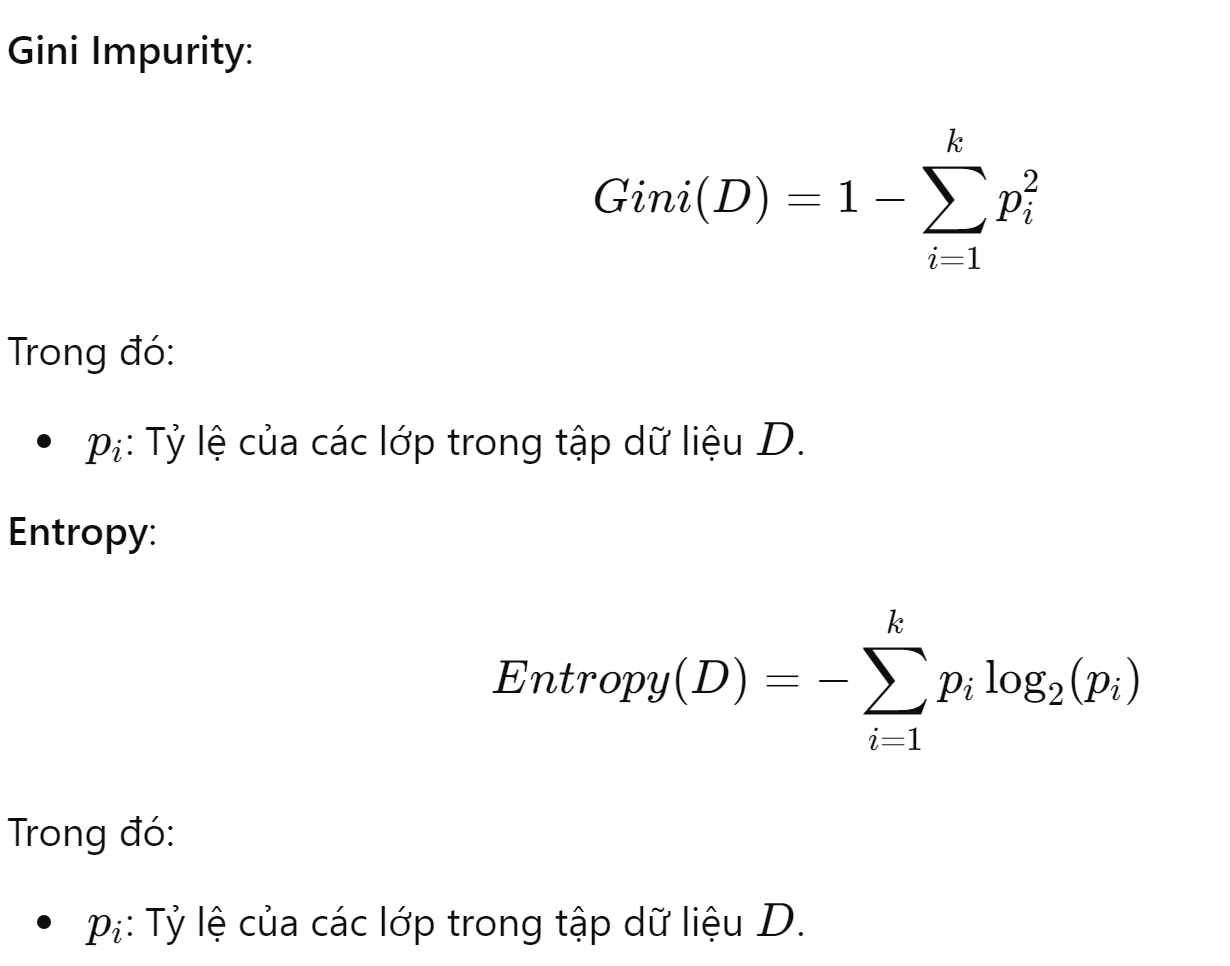
Code:



Hình 9: Random Forest.

Công thức tính : Random Forest là một tập hợp các cây quyết định (decision trees), và quyết định cuối cùng được đưa ra dựa trên đa số phiếu từ các cây.

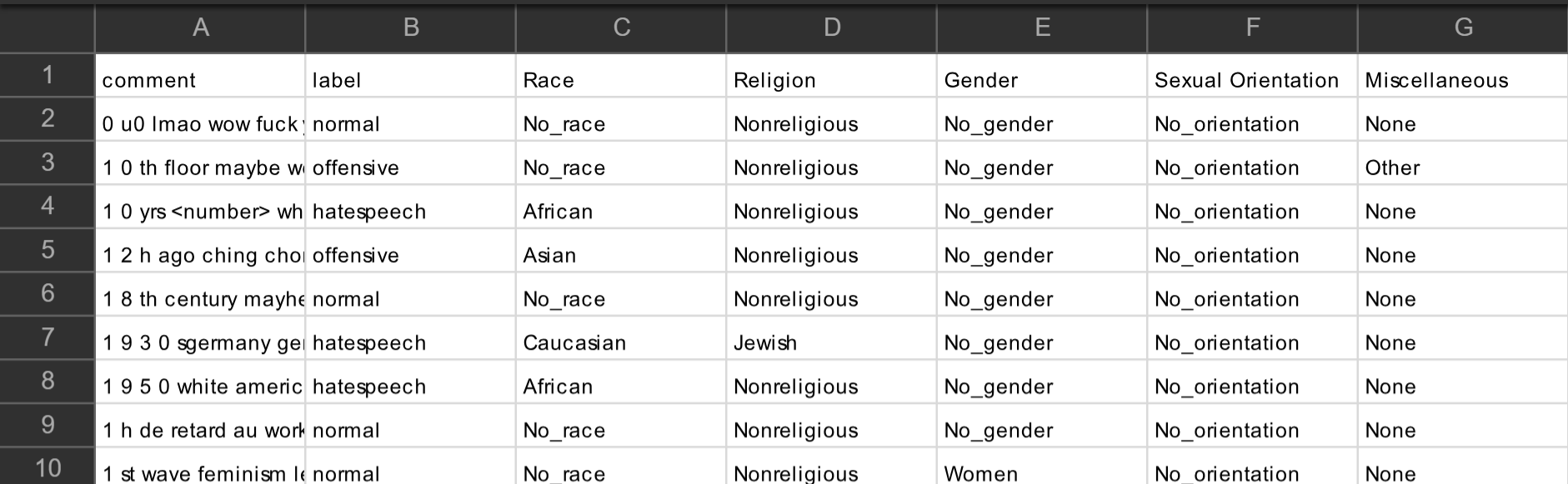
Tiêu chí phân chia:



Hình 10: Tiêu chí phân chia.

# Chương 3. Thực Nghiệm

## 3.1. Phân tích dữ liệu hiện tại



Hình 11: Dữ liệu ban đầu.

Dữ liệu gồm:

Có 20109 hàng

Và có 7 cột thuộc tính khác nhau

Được lấy từ trang web https://www.kaggle.com/

**Cột comment**: Chứa các bình luận hay ghi chú từ người dùng hoặc người cung cấp dữ liệu.

**Cột label**:

normal: Là các mục dữ liệu được phân loại là bình thường nội dung không chứa bất kỳ ngôn từ hoặc hành vi tiêu cực nào.

offensive: Là các mục dữ liệu chứa ngôn từ hoặc hành vi có thể gây xúc phạm, nhưng không nghiêm trọng đến mức gọi là "hatespeech."

hatespeech: Là các mục dữ liệu chứa ngôn từ hoặc hành vi thù địch, phân biệt chủng tộc, hoặc ngôn ngữ kích động hận thù đối với cá nhân hoặc nhóm người nào đó.

**Cột Race**: Chứa thông tin về chủng tộc của cá nhân.

**Cột Religion**: Là thông tin về tôn giáo mà cá nhân theo hoặc tin theo.

**Cột Gender**: Chứa thông tin về giới tính của cá nhân.

**Cột Sexual Orientation**: Liên quan đến xu hướng tình dục của cá nhân.

**Cột Miscellaneous**: Chứa thông tin bổ sung.

Qua đó ta có thể xây dựng một hệ thống để phát hiện và phân loại ngôn từ xúc phạm bằng cách kết hợp các thông tin từ các cột trên.

## 3.2. Kiểm tra và tiền sử lý dữ liệu

Chuyển đổi chữ thường, thay thế chuỗi, và loại bỏ các ký tự không mong muốn bằng các mẫu phù hợp sử dụng các phép toán cơ bản trên chuỗi và các biểu thức chính quy (regular expressions):



Hình 12: Hàm tiền sử lý văn bản.

Xử lý giá trị thiếu (NaN) và tiền xử lý văn bản.

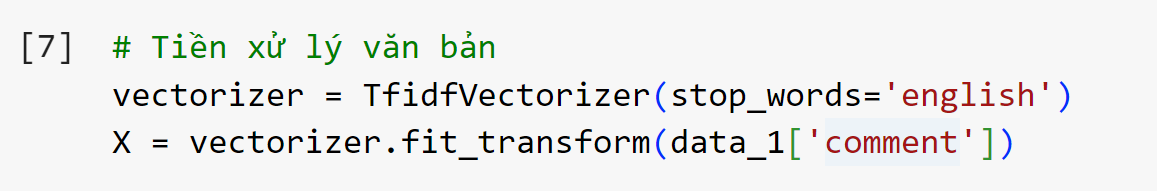


Hình 13: Xử lý giá trị thiếu (NaN) và tiền xử lý văn bản.

Chuyển đổi văn bản thành dạng số mà mô hình học máy có thể sử dụng:

TfidfVectorizer giúp chuyển đổi văn bản thành dạng số bằng cách sử dụng Tfidf để đo lường tầm quan trọng của từ trong tài liệu và toàn bộ tập dữ liệu.

Tham số stop\_words='english' loại bỏ các từ dừng để giảm thiểu ảnh hưởng của những từ không mang nhiều thông tin.

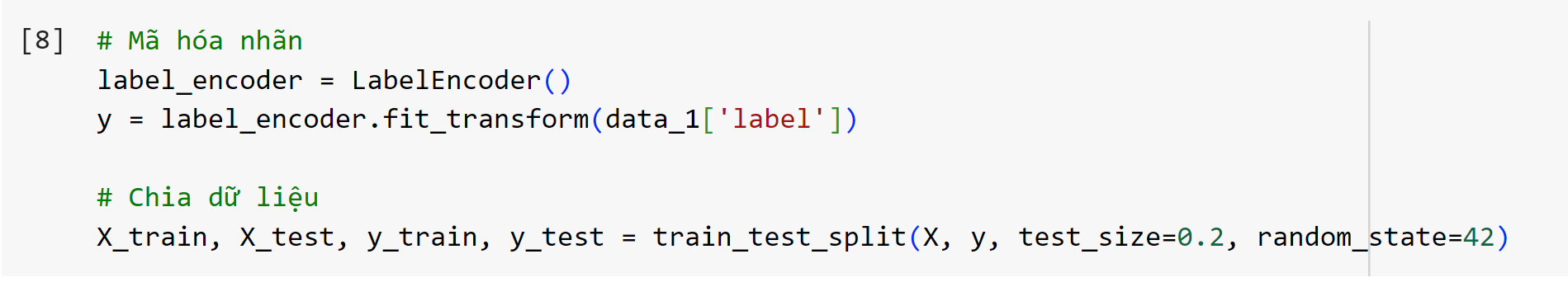
fit\_transform() tạo ra ma trận Tfidf từ dữ liệu văn bản, với các giá trị đại diện cho mức độ quan trọng của các từ trong các văn bản.

Hình 14: Chuyển đổi văn bản thành dạng số.

Mã hóa nhãn và chia dữ liệu:

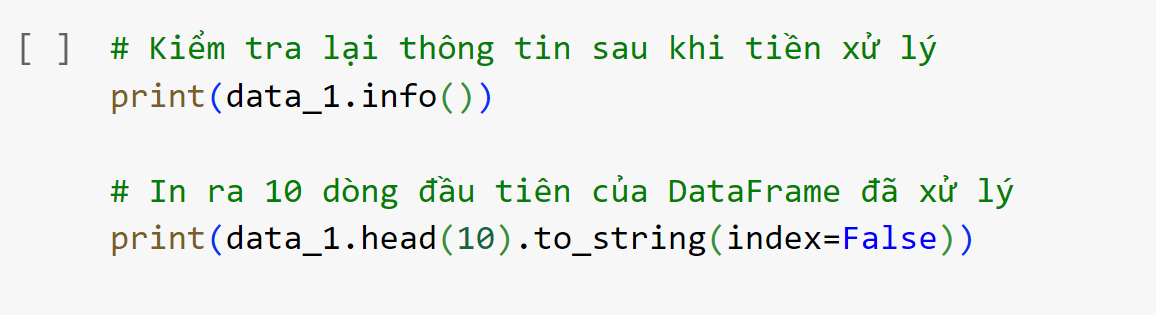
X\_train và y\_train chứa dữ liệu và nhãn tương ứng cho tập huấn luyện.

X\_test và y\_test chứa dữ liệu và nhãn tương ứng cho tập kiểm tra.

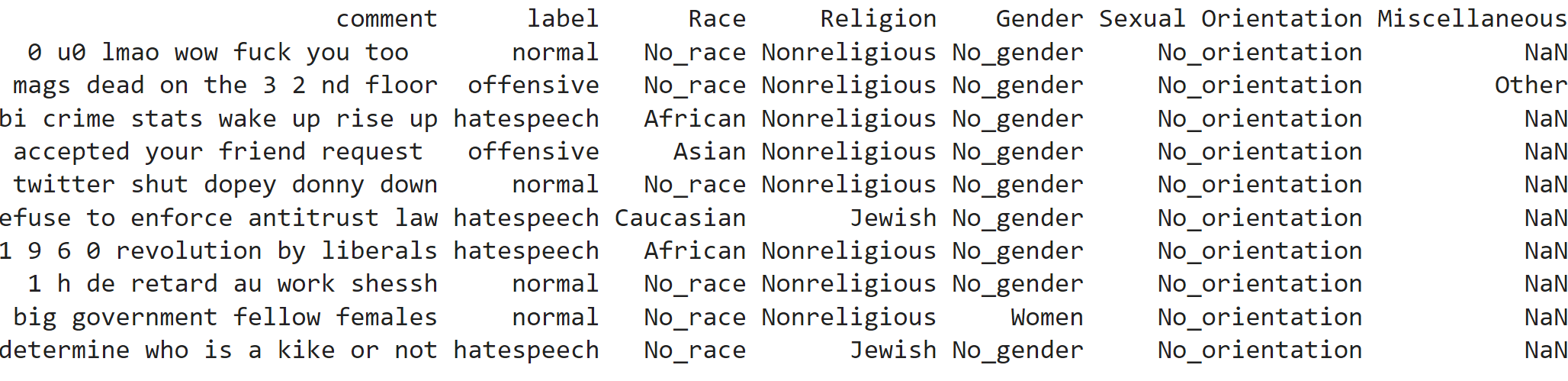


Hình 15: Mã hóa nhãn và chia dữ liệu.

In dữ liệu sau khi đã tiền sử lý.



Hình 16: In dữ liệu sau tiền sử lý.



Hình 17: Dữ liệu sau khi tiền sử lý.

## 3.3. Triển khai trực quan hóa dữ liệu.

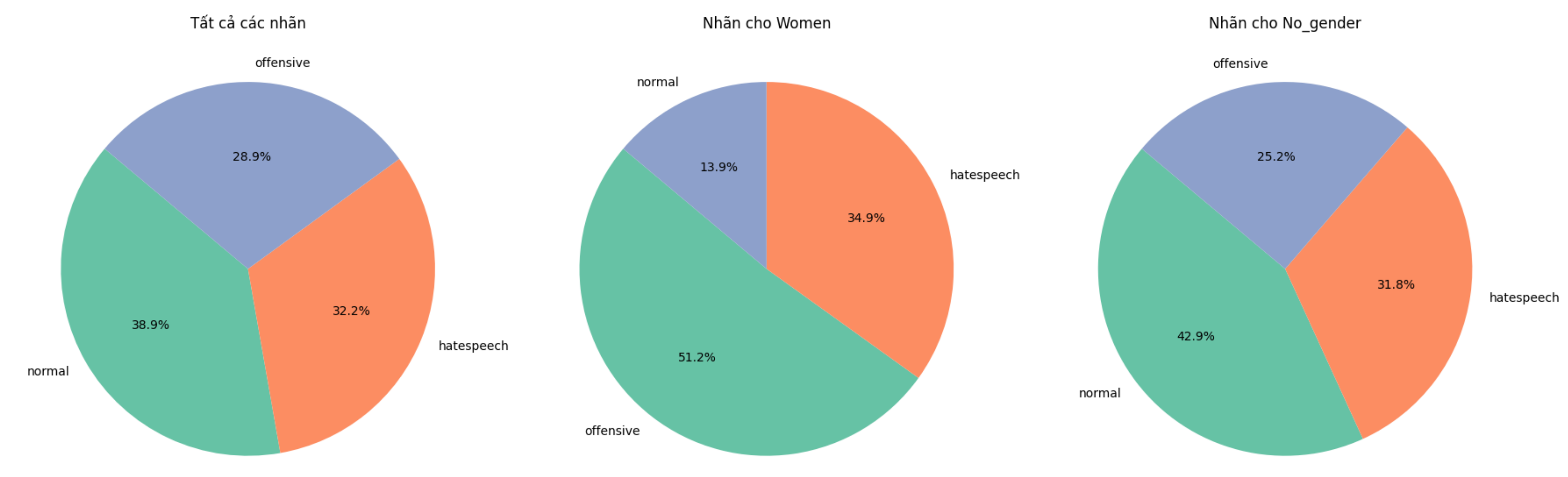
Sử dụng **biểu đồ tròn(pie chart)** để phân tích và trực quan hóa các loại nhãn trong dữ liệu dựa trên phân nhóm giới tính.

Code:



Hình 18: Code minh họa biểu đồ tròn(pie chart).

Sử dụng cột '**label**' để đếm số lượng nhãn và cột '**Gender**' để phân chia dữ liệu theo giới tính, từ đó vẽ ba biểu đồ tròn để thể hiện tỷ lệ phần trăm của từng nhãn trong ba nhóm khác nhau.



Hình 19: Biểu đồ tròn(pie chart).

Biểu đồ trực quan cho thấy được Nhãn cho Women có 51.2% về offensive là rất cao cho thấy có nghĩa là hơn một nửa của các nhãn cho nhóm phụ nữ được phân loại là offensive (xúc phạm) có thể là do phụ nữ xuất phát từ nhiều yếu tố phức tạp.

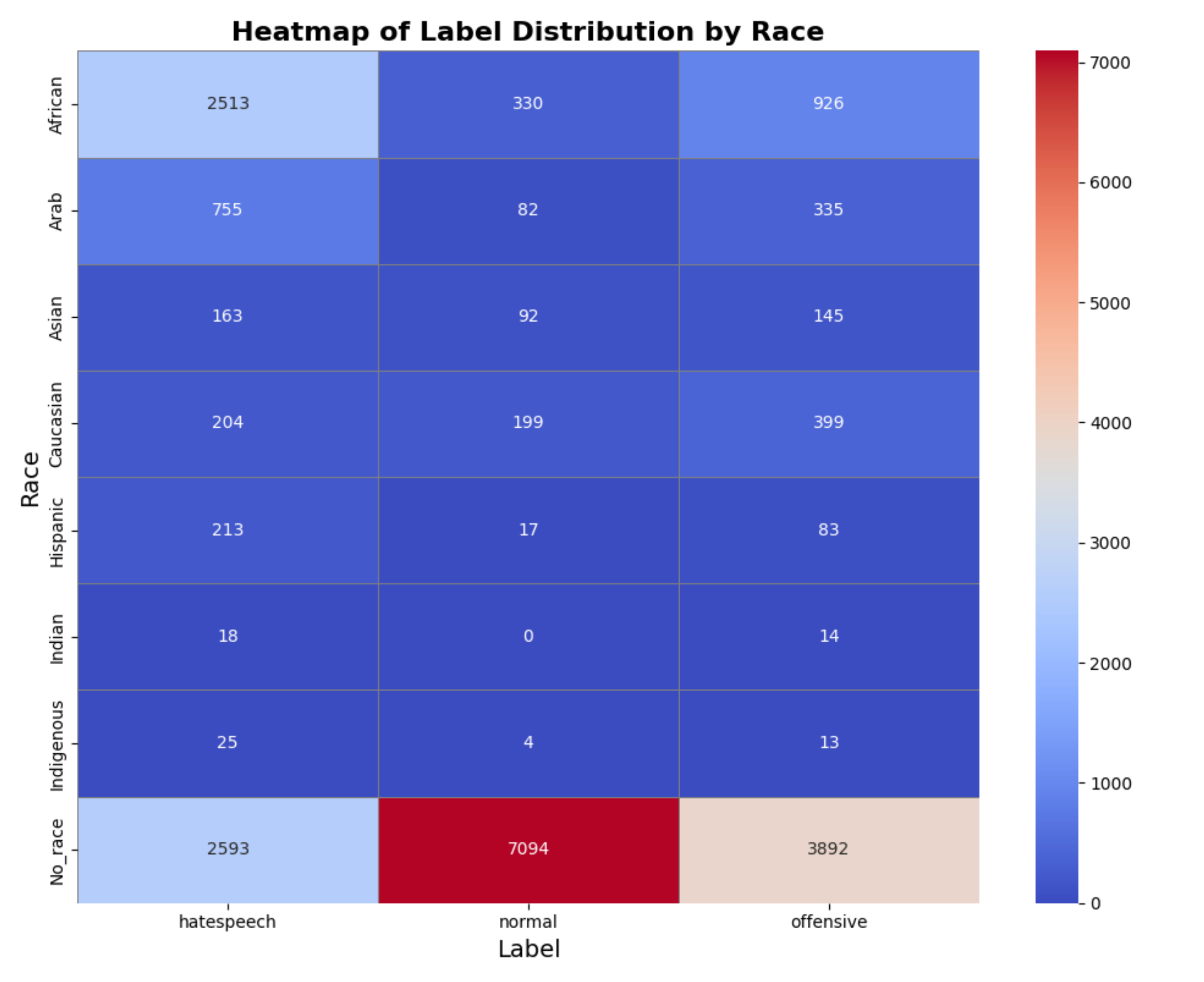
Sử dụng **Biểu đồ nhiệt (heatmap)** phân phối của các loại nhãn (labels) theo các nhóm chủng tộc (races)

Code:



Hình 20: Code minh họa Biểu đồ nhiệt (heatmap).

Biểu đồ này giúp phát hiện các mẫu hoặc xu hướng trong phân phối nhãn dựa trên chủng tộc, điều này có thể cung cấp thông tin hữu ích cho phân tích dữ liệu, các giá trị số trong ô giúp biết chính xác số lượng nhãn trong từng nhóm chủng tộc:



Hình 21: Biểu đồ nhiệt (heatmap).

Biểu đồ trên cho ta thấy:

**African:** Có số lượng lớn nhất đối với nhãn hatespeech (2513) và offensive (926). Điều này cho thấy rằng trong nhóm người có chủng tộc là **African**, tần suất của các bình luận hoặc nội dung bị phân loại là hatespeech và offensive là cao hơn so với các nhóm chủng tộc khác.

**Indian**: Số lượng nhãn hatespeech thấp nhất (18). Điều này cho thấy rằng trong nhóm người có chủng tộc là **Indian**, sự xuất hiện của nội dung được phân loại là hatespeech là ít hơn nhiều so với các nhóm khác.

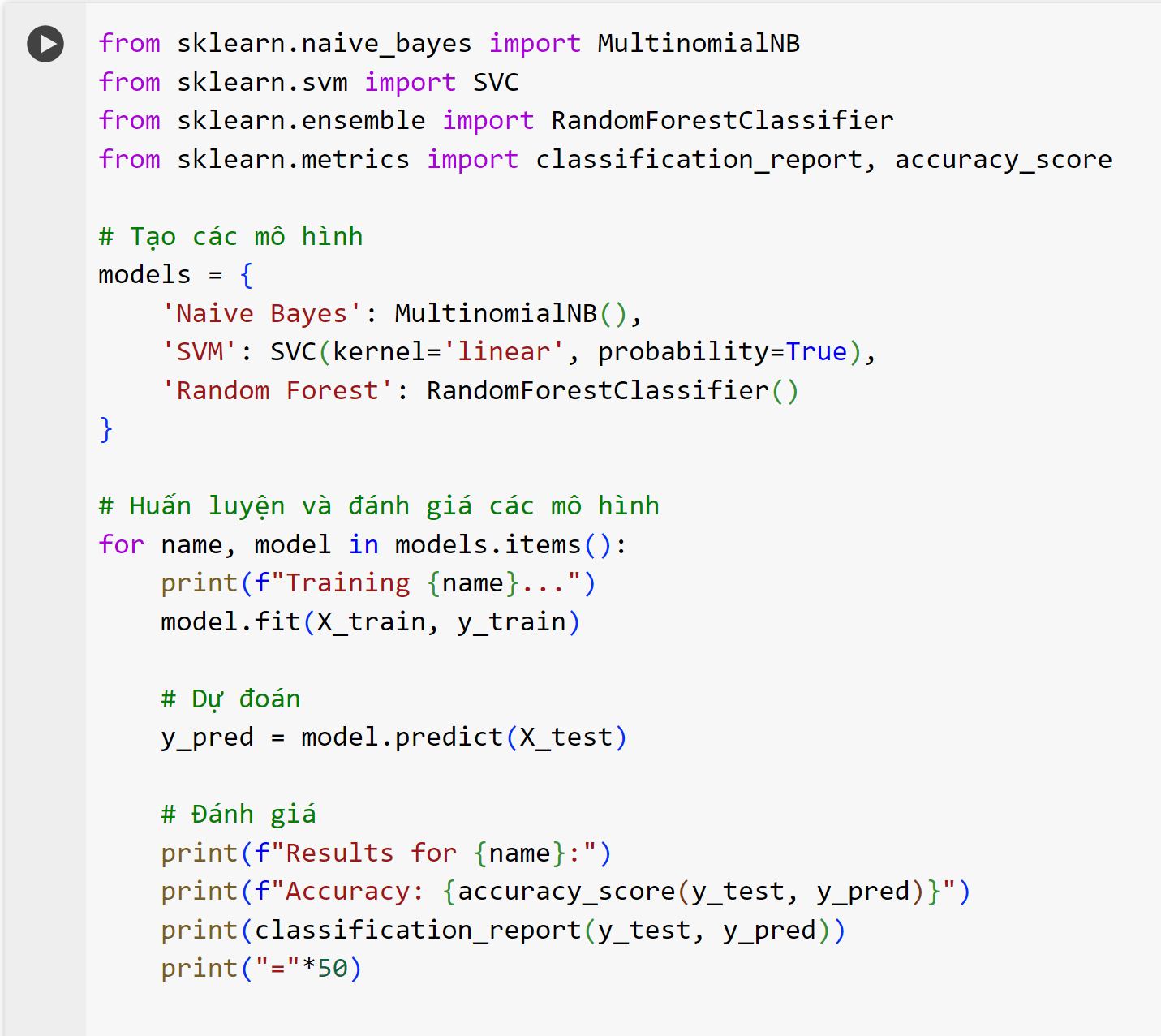
**Indigenous**: Số lượng nhãn offensive thấp nhất (13). Điều này cho thấy rằng nhóm người có chủng tộc là **Indigenous** có ít bình luận hoặc nội dung bị phân loại là offensive.

## 3.4. Phát triển và đánh giá mô hình

### 3.4.1. Xây dựng mô hình

Khởi tạo các mô hình học máy **Naive Bayes (MultinomialNB), Support Vector Machine (SVM), Random Forest**:

Code:



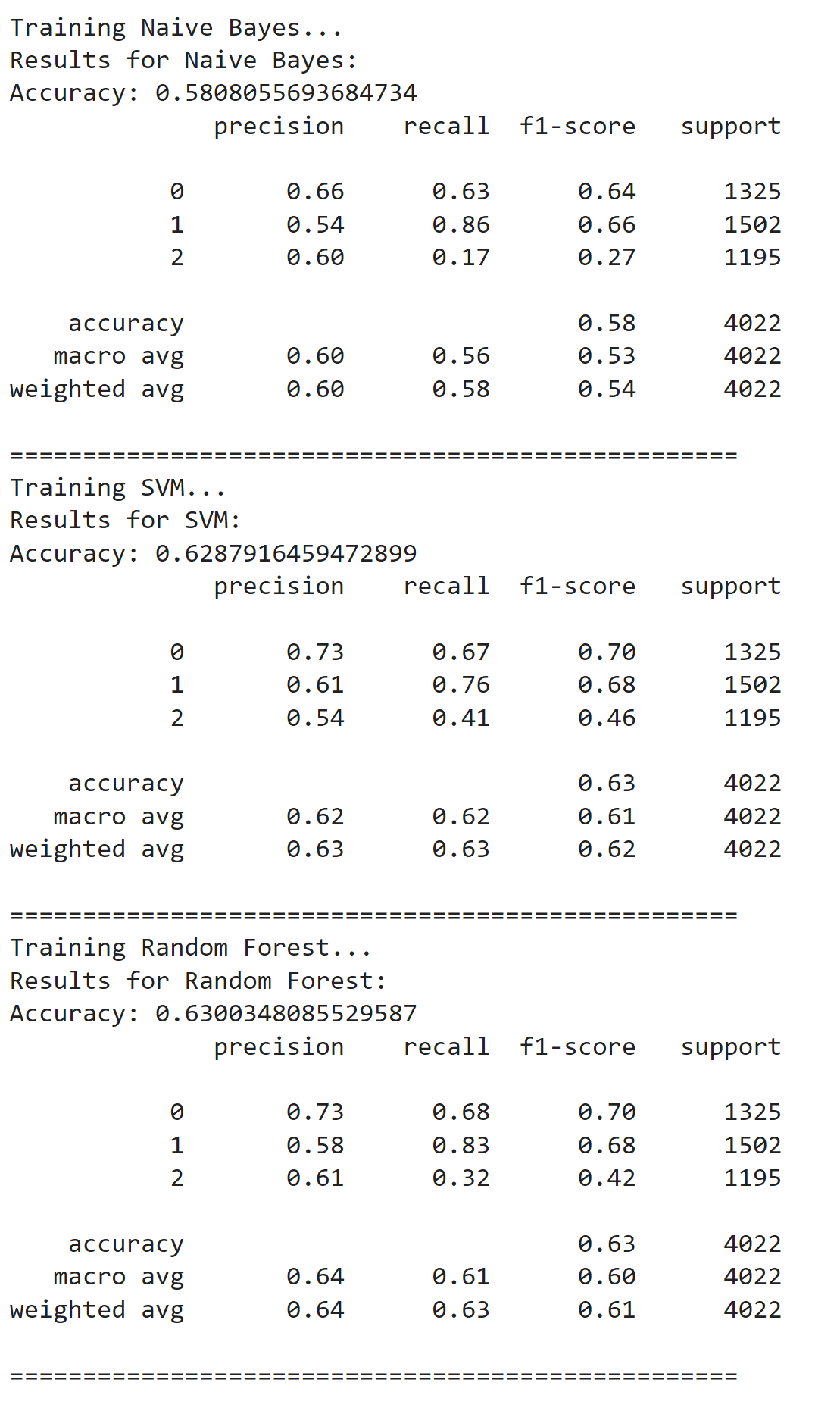
Hình 22: Xây dựng mô hình học máy.

Naive Bayes (MultinomialNB): tính xác suất của một lớp dựa trên tần suất xuất hiện của các đặc trưng trong lớp đó.

Support Vector Machine (SVC): Mô hình sử dụng hàm mục tiêu là việc tối đa hóa khoảng cách (hoặc margin) giữa các lớp. Đối với kernel 'linear', mặt phẳng phân cách được định nghĩa bởi hàm số tuyến tính trong không gian đặc trưng.

Random Forest (RandomForestClassifier): Xây dựng nhiều cây quyết định dự đoán của mỗi cây được thu thập. Dự đoán cuối cùng được tính bằng cách tổng hợp (hoặc bỏ phiếu đa số) kết quả từ tất cả các cây.

Kết quả:



Hình 23: Kết quả mô hình.

Nhận xét:

Mô hình Naive Bayes có độ chính xác tương đối thấp (58.08%).

Mô hình SVM có độ chính xác tốt hơn so với Naive Bayes (62.88%). Đối với lớp 0 và lớp 1, SVM cho thấy sự cân bằng tốt giữa precision và recall, dẫn đến F1Score cao hơn.

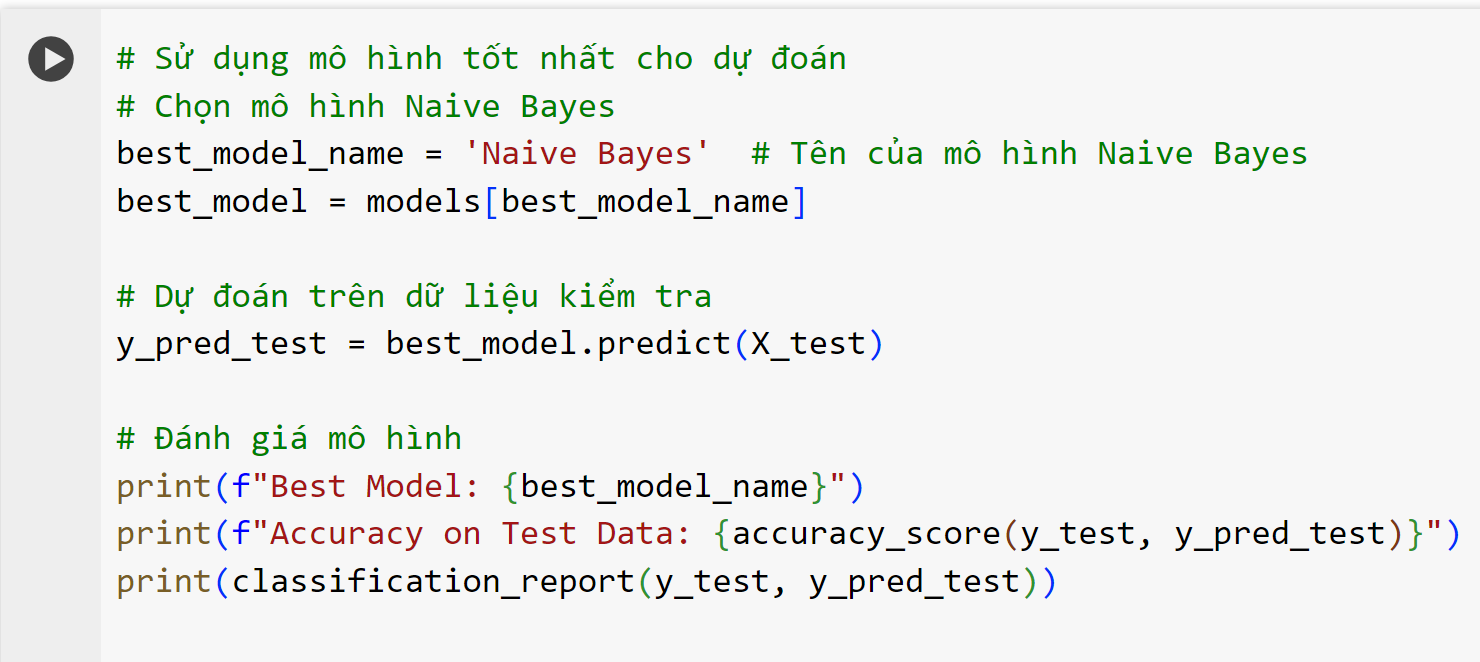
Random Forest đạt độ chính xác cao nhất trong ba mô hình (63.00%). Mô hình cho thấy sự cân bằng tốt giữa các lớp với precision và recall, đặc biệt là đối với lớp 0 và lớp 1.

### 3.4.2. Đánh giá mô hình, dự đoán và phân tích

Đánh giá mô hình.

Mục tiêu chính là phát hiện hate speech một cách chính xác nên chọn mô hình Naive Bayes sẽ giải thích ở bên dưới

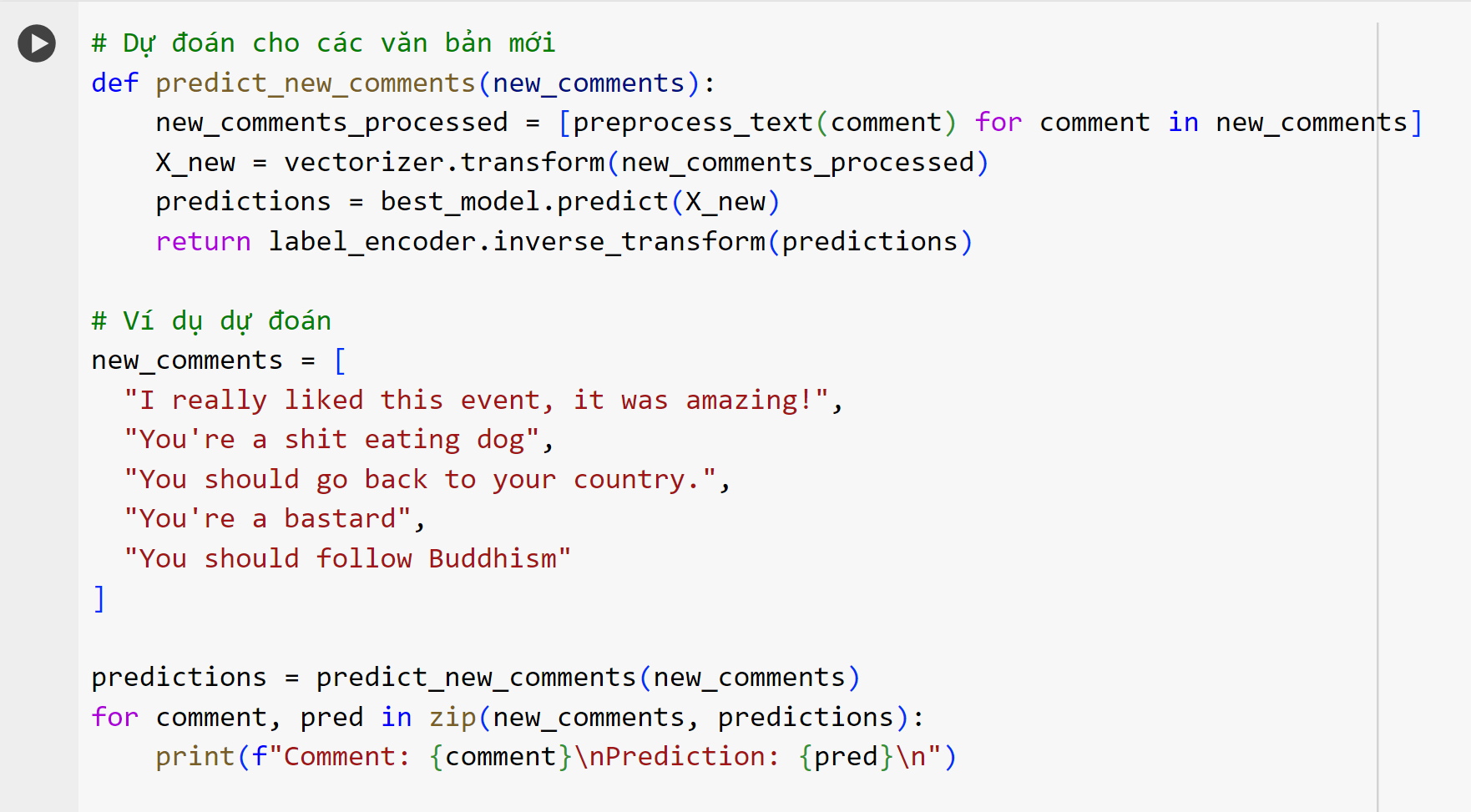
Code:



Hình 24: Đánh giá mô hình.

Dự đoán.

Code:



Hình 25: Dự đoán các văn bản mới.

So sánh Dự đoán.

Dự đoán của Random Forest:

"You're a shit eating dog": normal (dự đoán không chính xác; đúng là hate speech)

"You should go back to your country.": normal (dự đoán không chính xác; đúng là hate speech)

"You're a bastard": normal (dự đoán không chính xác; đúng là hate speech)

Dự đoán của SVM:

"You're a shit eating dog": normal (dự đoán không chính xác; đúng là hate speech)

"You should go back to your country.": hate speech (dự đoán chính xác)

"You're a bastard": normal (dự đoán không chính xác; đúng là hate speech)

Dự đoán của Naive Bayes:

"You're a shit eating dog": hate speech (dự đoán chính xác)

"You're a bastard": hate speech (dự đoán chính xác)

Phân Tích

Random Forest có vẻ đưa ra nhiều dự đoán không chính xác, đặc biệt là trong các câu có nội dung tiêu cực (hate speech).

SVM có một số dự đoán chính xác hơn so với Random Forest, nhưng vẫn còn một số lỗi trong việc phân loại các câu tiêu cực.

Naive Bayes thể hiện khả năng phân loại các câu tiêu cực (hate speech) tốt hơn, với cả hai dự đoán về các câu tiêu cực đều chính xác.

Vì thế **Naive Bayes** có vẻ là mô hình phù hợp nhất trong trường hợp này, bởi vì nó đã phân loại chính xác các câu hate speech. Mặc dù Naive Bayes có thể không có hiệu suất tổng thể tốt nhất theo các chỉ số thống kê như accuracy hay f1score, nhưng trong bối cảnh của bài toán cụ thể (phân loại hate speech), khả năng phân loại đúng các trường hợp hate speech là quan trọng. Do đó, mục tiêu chính là phát hiện hate speech một cách chính xác, thì Naive Bayes có vẻ là sự lựa chọn tốt nhất trong ba mô hình đã thử nghiệm.

# Chương 4. Kết Luận

Dựa trên phân tích và đánh giá các mô hình học máy để phát hiện và phân loại ngôn từ xúc phạm, em đã thực hiện các bước sau:

Tiền xử lý Dữ liệu:

Em đã chuyển đổi văn bản thành dạng số bằng cách sử dụng `TfidfVectorizer`, loại bỏ các từ dừng và áp dụng các phương pháp xử lý văn bản cơ bản như chuyển đổi chữ thường, thay thế chuỗi và loại bỏ ký tự không mong muốn.

Dữ liệu được chia thành các tập huấn luyện và kiểm tra để chuẩn bị cho việc đào tạo và đánh giá mô hình.

Trực Quan Hóa Dữ liệu:

Các biểu đồ tròn và biểu đồ nhiệt được sử dụng để phân tích phân phối nhãn theo giới tính và chủng tộc. Kết quả cho thấy tỷ lệ nhãn "offensive" cao hơn trong nhóm phụ nữ và có sự phân phối không đồng đều của các loại nhãn giữa các nhóm chủng tộc khác nhau.

Đánh Giá Mô Hình:

Naive Bayes: Đạt độ chính xác thấp nhất (58.08%) nhưng thể hiện khả năng phân loại hate speech tốt, điều này quan trọng trong bối cảnh của bài toán. Mặc dù không có hiệu suất tổng thể cao nhất, nhưng mô hình này chính xác hơn trong việc phân loại các trường hợp hate speech.

SVM: Cung cấp độ chính xác tốt hơn so với Naive Bayes (62.88%) và cho thấy sự cân bằng tốt giữa precision và recall, với điểm F1 cao hơn.

Random Forest: Đạt độ chính xác cao nhất (63.00%) và có sự cân bằng tốt giữa các lớp, đặc biệt là lớp 0 và lớp 1. Tuy nhiên, mô hình này gặp khó khăn trong việc phân loại chính xác các câu tiêu cực.

Ưu tiên lựa chọn:

Trong bối cảnh của bài toán phân loại hate speech, mặc dù Naive Bayes không có hiệu suất tổng thể tốt nhất, nó lại là mô hình chính xác hơn trong việc phân loại các câu hate speech. Do đó, Naive Bayes là lựa chọn phù hợp nhất cho nhiệm vụ này, bởi khả năng phát hiện đúng các trường hợp hate speech là ưu tiên hàng đầu.

Mặc dù có một số nhược điểm và không hoàn hảo trong tất cả các khía cạnh, Naive Bayes được xác định là mô hình tốt nhất cho việc phát hiện hate speech trong tình huống hiện tại. Các mô hình khác như SVM và Random Forest cũng có giá trị, nhưng Naive Bayes có thể đáp ứng tốt hơn mục tiêu chính của bài toán này.

# Tài Liệu Tham Khảo

[1].https://viblo.asia/p/tienxulidulieuvanbanvoinltkAz45b0LgZxY

[2].https://machinelearningcoban.com/2017/08/08/nbc/

[3].https://codelearn.io/sharing/top30librariespackages4beginnerp1

[4].<https://tuoitre.vn/thenaolaxucphamnguoikhactrenmangxahoi20171022101531938.htm>

[5].https://baobinhthuan.com.vn/binhluanvabinhloan110460.html

Link source code:

<https://colab.research.google.com/drive/1kCewoIEGujPel9cHYqbkNHR9m03Nukce#scrollTo=4pFiejQfYSw9>

Link dataset:

[https://drive.google.com/drive/u/0/folders/13HXQKhSLE6jlgO0B88sB6FaCTnnN5i9N](https://www.kaggle.com/datasets/chopper53/data-engineer-salary-in-2024)