**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC ĐẠI NAM**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**TÊN HỌC PHẦN: HỌC MÁY**

**ĐỀ TÀI: PHÁT HIỆN TIN GIẢ BẰNG HỌC MÁY**

**Giáo viên hướng dẫn: TRẦN HỒNG VIỆT**

**Sinh viên thực hiện: Nhóm 7**

|  |  |  |  |
| --- | --- | --- | --- |
| **Stt** | **Mã sv** | **Họ và tên** | **Lớp** |
| **1** | **167102060** | **Nguyễn Đức Duy** | **CNTT 16-01** |
| **2** | **1671020089** | **Nguyễn Minh Đức** | **CNTT 16-01** |
| **3** | **1671020319** | **Nguyễn Tất Toàn** | **CNTT 16-01** |
| **4** | **1671020229** | **Nguyễn Văn Nguyên** | **CNTT 16-01** |

**Hà Nội, năm 2024**

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC ĐẠI NAM**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**TÊN HỌC PHẦN: HỌC MÁY**

**ĐỀ TÀI: PHÁT HIỆN TIN GIẢ BẰNG HỌC MÁY**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **Mã sinh viên** | **Họ và Tên** | **Ngày Sinh** | **Điểm** | |
| **Bằng Số** | **Bằng Chữ** |
| **1** | **167102060** | **Nguyễn Đức Duy** | **26/12/2004** |  |  |
| **2** | **1671020089** | **Nguyễn Minh Đức** | **16/08/2004** |  |  |
| **3** | **1671020319** | **Nguyễn Tất Toàn** | **18/01/2004** |  |  |
| **4** | **1671020229** | **Nguyễn Văn Nguyên** | **22/10/2004** |  |  |

**CÁN BỘ CHẤM THI 1 CÁN BỘ CHẤM THI 2**

**Hà Nội, năm 2024**

|  |  |  |  |
| --- | --- | --- | --- |
| **Bảng phân công nhiệm vụ nhóm số 7** | | | |
| **Thứ tự** | **Họ và tên** | **Nhiệm vụ** | **Mức độ hoàn thành** |
| **1** | Nguyễn Đức Duy | * Xây dựng mô hình học máy và đánh giá mô hình * Viết báo cáo tóm tắt kết quả và nhận xét | 100% |
| **2** | Nguyễn Minh Đức | * Nghiên cứu tài liệu về phát hiện tin giả * Tối ưu hóa mô hình và thực hiện đánh giá | 100% |
| **3** | Nguyễn Tất Toàn | * Thu thập và chuẩn bị dữ liệu * Phân tích các kết quả và so sánh giữa các mô hình | 90% |
| **4** | Nguyễn Văn Nguyên | * Viết mã tiền xử lý dữ liệu và trực quan hóa * Thiết kế slide thuyết trình và chuẩn bị bài thuyết trình | 90% |

**MỤC LỤC**

[CHƯƠNG I: ĐỀ TÀI PHÁT HIỆN TIN GIẢ BẰNG HỌC MÁY 1](#_Toc24435)

[1. Lý do chọn đề tài 1](#_Toc12279)

[1.1. Tầm quan trọng của vấn đề tin giả trong xã hội hiện đại: 1](#_Toc9533)

[1.2. Sự phát triển mạnh mẽ của công nghệ học máy (Machine Learning): 1](#_Toc14057)

[1.3. Tính cấp thiết của việc chống lại tin giả: 1](#_Toc31643)

[1.4. Sự đa dạng và thách thức của dữ liệu trong phát hiện tin giả: 2](#_Toc5587)

[1.5. Tiềm năng ứng dụng thực tiễn: 2](#_Toc19958)

[1.6. Khả năng phát triển và mở rộng nghiên cứu: 2](#_Toc4633)

[2. Đặt vấn đề 3](#_Toc15674)

[3. Tổng quan về phát hiện tin giả 4](#_Toc23985)

[3.1. Khái niệm và phân loại 4](#_Toc15726)

[3.1.1. Khái niện về tin giả 4](#_Toc9920)

[3.1.2. Phân loại tin giả 4](#_Toc11614)

[3.1.3. Đặc điểm nhận diện tin giả 5](#_Toc32055)

[3.2. Sự phát triển của tin giả trong thời đại số 5](#_Toc8256)

[3.2.1. Lịch sử phát triển của tin giả 5](#_Toc1622)

[3.2.2. Vai trò của mạng xã hội và truyền thông trong việc lan truyền tin giả 6](#_Toc4530)

[3.2.3. Những hệ quả tiêu cực của tin giả đối với xã hội 6](#_Toc30008)

[3.2.4. Số liệu thống kê về sự lan truyền của tin giả 7](#_Toc16010)

[3.2.5. Những xu hướng mới trong việc tạo ra và lan truyền tin giả 7](#_Toc14552)

[3.2.6. Các biện pháp kiểm soát và ngăn chặn tin giả 8](#_Toc29008)

[3.2.7. Kết luận 8](#_Toc4915)

[CHƯƠNG 2: PHƯƠNG PHÁP TIẾP CẬN ĐỀ TÀI 9](#_Toc18265)

[1. Cách tiếp cận vấn đề 9](#_Toc10859)

[1.1. Tiền xử lý dữ liệu 9](#_Toc23511)

[1.2. Lựa chọn mô hình học máy 10](#_Toc27676)

[1.3. Huấn luyện và tối ưu hóa mô hình 11](#_Toc28769)

[1.4. Triển khai và theo dõi mô hình 11](#_Toc7235)

[1.5. Tổng kết phân tích cách tiếp cận vấn đề 12](#_Toc9917)

[2. Quy trình phát hiện học giả 12](#_Toc29791)

[2.1. Thu thập dữ liệu 12](#_Toc15858)

[2.2. Tiền xử lý dữ liệu 12](#_Toc7363)

[2.3. Trích xuất đặc trưng 13](#_Toc5351)

[2.4. Xây dựng và huấn luyện mô hình 13](#_Toc22819)

[2.5. Đánh giá và triển khai mô hình 14](#_Toc1831)

[3. Các mô hình học máy phổ biến 14](#_Toc19173)

[3.1. Naive Bayes 14](#_Toc7287)

[3.2. Support Vector Machine (SVM) 14](#_Toc32459)

[3.3. Logistic Regression 15](#_Toc2657)

[3.4. Decision Tree (Cây quyết định) 15](#_Toc26845)

[4. Thách thức và giải pháp trong phát hiện tin giả bằng học máy 17](#_Toc16102)

[4.1. Thách thức 17](#_Toc18627)

[4.2. Giải pháp 17](#_Toc9715)

[5. Kết luận 17](#_Toc505)

[5.1. Tóm tắt 17](#_Toc10066)

[5.2. Định hướng phát triển 18](#_Toc3705)

[CHƯƠNG III: THỰC NGHIỆM VÀ KẾT QUẢ CỦA CHƯƠNG TRÌNH PHÁT HIỆN TIN GIẢ BẰNG HỌC MÁY 19](#_Toc2335)

[1. Thực nghiệm và kết quả của chương trình 19](#_Toc10188)

[1.1. Môi trường thực nghiệm 19](#_Toc793)

[1.2. Kết quả đánh giá mô hình 19](#_Toc25001)

[1.3. Phân tích kết quả 20](#_Toc28994)

[2. Luồng hoạt động của chương trình 21](#_Toc18598)

[2.1. Nhập và cài đặt các thư viện cần thiết 21](#_Toc2526)

[2.2. Đường dẫn và kiểm tra file 22](#_Toc18251)

[2.3. Tải và tiền xử lý dữ liệu 22](#_Toc28855)

[2.4. Tiền xử lý văn bản 23](#_Toc24715)

[2.5. Trực quan hóa dữ liệu bằng Word Cloud 23](#_Toc22838)

[2.6. Xác định các từ phổ biến nhất trong dữ liệu 25](#_Toc12657)

[2.7. Chia dữ liệu train, test và vector hóa 26](#_Toc20988)

[2.8. Huấn luyện và đánh giá mô hình 26](#_Toc25808)

[2.9. Trực quan hóa ma trận nhầm lẫn ( Confusion Matrix ) 27](#_Toc15973)

[2.10. Kiểm tra với dữ liệu mới 28](#_Toc27942)

[3. Tổng quan chương trình phát hiện tin giả bằng học máy 28](#_Toc6490)

[KẾT LUẬN 30](#_Toc27275)

[TÀI LIỆU THAM KHẢO 32](#_Toc14332)

# LỜI MỞ ĐẦU

Những năm gần đây, AI nổi lên như một bằng chứng của cuộc cách mạng công nghiệp lần thứ tư. Trí tuệ nhân tạo có thể được định nghĩa như một ngành của khoa học máy tính liên quan đến việc tự động hóa các hành vi thông minh. Trí tuệ nhân tạo là một bộ phân của khoa học máy tính và do đó nó phải được đặt trên những nguyên lý lý thuyết vững chắc, có khả năng ứng dụng được các lĩnh vực này. Ở thời điểm hiện tại, thuật ngữ này thường dùng để nói đến các máy tính có mục đích không nhất định và ngành khoa học nghiên cứu về các lý thuyết và các ứng dụng của trí tuệ nhân tạo.

Theo đà phát triển của công nghệ, ứng dụng trí tuệ nhân ạo luôn là xu hướng công nghệ tương lai mà các hãng công nghệ trên toàn thế giới đua nhau sáng tạo, nó là nền tảng cốt lõi của cuộc cách mạng công nghệ 4.0. ML (Machine Learning) là một lĩnh vực của trí tuệ nhân tạo, được sinh ra từ khả năng nhận diện mẫu và từ lý thuyết của các máy tính có thể học mà không cần phải lập trình để xử lý các nhiệm vụ cụ thể nào đó.

Hầu hết các Doanh nghiệp đang làm việc với lượng lớn dữ liệu đều nhận ra tầm quan trọng của công nghệ AI, đặc biệt trong phát hiện tin giả, tin đồn, không có căn cứ, không xác thực và không chính thống. Những cái nhìn sáng suốt từ nguồn dữ liệu này chủ yếu trong thời gian thực sẽ giúp các tổ chức, các doanh nghiệp vận hành hiệu quả hơn hoặc tạo lợi thế cạnh trạnh với các đối thủ. Các ứng dụng của AI đã quá quen thuộc với con người: xe tự hành của Google và Tesla, hệ thống tự tag khuôn mặt trên Facebook, hệ thống gợi ý sản phẩm của Amazon, hệ thống gợi ý phim trên ứng dụng Netflix... chỉ là một vài ứng dụng trong muôn vàn những ứng dụng của trí tuệ nhân tạo Machine Learning. Trong bài báo cáo này, nhóm chúng em sẽ nghiên cứu các mô hình học máy cơ bản của Machine Learning để ứng dụng vào xây dựng phần mềm phát hiện tin giả bằng ngôn ngữ Python.

# CHƯƠNG I: ĐỀ TÀI PHÁT HIỆN TIN GIẢ BẰNG HỌC MÁY

1. **Lý do chọn đề tài**

Trong bối cảnh hiện nay, tin giả (fake news) đã trở thành một vấn đề xã hội nghiêm trọng, ảnh hưởng không chỉ đến niềm tin của công chúng mà còn tác động sâu rộng đến chính trị, kinh tế và văn hóa. Sự phát triển nhanh chóng của mạng xã hội và công nghệ thông tin đã làm cho việc tạo ra, lan truyền và tiếp nhận tin tức trở nên dễ dàng hơn bao giờ hết. Tuy nhiên, điều này cũng mở ra những cơ hội cho sự phát tán thông tin sai lệch với tốc độ chóng mặt, dẫn đến nhiều hậu quả nghiêm trọng.

* 1. Tầm quan trọng của vấn đề tin giả trong xã hội hiện đại:
* Tin giả có khả năng làm sai lệch nhận thức và hành vi của con người, từ đó tạo ra sự hoang mang và mâu thuẫn trong xã hội.
* Các cuộc bầu cử, sự kiện chính trị và những vấn đề liên quan đến sức khỏe cộng đồng (như đại dịch COVID-19) đã chứng minh mức độ nguy hiểm của tin giả khi chúng ảnh hưởng trực tiếp đến các quyết định quan trọng của con người.
  1. Sự phát triển mạnh mẽ của công nghệ học máy (Machine Learning):
* Học máy đã chứng minh được tiềm năng to lớn trong nhiều lĩnh vực như y tế, tài chính, và hiện tại là phân tích ngôn ngữ tự nhiên (Natural Language Processing - NLP) để phát hiện và phân loại tin tức.
* Các mô hình học máy tiên tiến có thể phân tích khối lượng lớn dữ liệu và phát hiện ra các đặc điểm tinh vi của tin giả mà con người khó có thể nhận diện được một cách thủ công.
  1. Tính cấp thiết của việc chống lại tin giả:
* Với sự bùng nổ của thông tin sai lệch, đặc biệt trên các nền tảng mạng xã hội, việc phát triển các công cụ tự động để phát hiện và ngăn chặn tin giả trở nên vô cùng cấp bách.
* Những hệ thống phát hiện tin giả tự động sẽ giúp giảm thiểu tác động tiêu cực của tin giả, đồng thời bảo vệ nền tảng thông tin chính xác cho xã hội.
  1. Sự đa dạng và thách thức của dữ liệu trong phát hiện tin giả:
* Tin giả thường được viết một cách tinh vi, khiến việc phân biệt với tin thật không dễ dàng. Sự đa dạng về cấu trúc ngôn ngữ và nội dung trong tin tức đòi hỏi các phương pháp xử lý thông tin và học máy phức tạp.
* Việc nghiên cứu đề tài này sẽ giúp giải quyết thách thức về xử lý dữ liệu phi cấu trúc và ngôn ngữ tự nhiên, một trong những lĩnh vực quan trọng và phức tạp nhất trong học máy.
  1. Tiềm năng ứng dụng thực tiễn:
* Các hệ thống phát hiện tin giả không chỉ được ứng dụng trên các nền tảng mạng xã hội mà còn có thể được triển khai trong các cơ quan báo chí, các tổ chức truyền thông, và cả các cơ quan chính phủ để ngăn chặn sự lan truyền của thông tin sai lệch.
* Ứng dụng của học máy trong việc phát hiện tin giả có tiềm năng lớn trong việc tạo ra những công cụ hỗ trợ kiểm duyệt và đảm bảo thông tin trung thực trên không gian mạng.
  1. Khả năng phát triển và mở rộng nghiên cứu:
* Đây là một lĩnh vực nghiên cứu còn nhiều tiềm năng phát triển với nhiều hướng đi như: nâng cao độ chính xác của mô hình, xử lý dữ liệu đa nguồn, và phát hiện các loại tin giả phức tạp hơn (ví dụ: deepfake, thông tin bị chỉnh sửa).
* Đề tài không chỉ mang tính chất học thuật mà còn có giá trị thực tiễn cao, mở ra nhiều cơ hội cho các ứng dụng về bảo mật thông tin, quản lý nội dung số và nâng cao nhận thức cộng đồng về vấn đề tin giả.

Với tất cả các lý do trên, đề tài "Phát hiện tin giả bằng phương pháp học máy" không chỉ có ý nghĩa về mặt học thuật mà còn mang lại giá trị thực tiễn to lớn, góp phần vào việc xây dựng một môi trường thông tin minh bạch và lành mạnh cho xã hội hiện đại.

1. **Đặt vấn đề**

Trong kỷ nguyên kỹ thuật số, tin tức và thông tin được truyền tải với tốc độ nhanh chóng trên các nền tảng trực tuyến, từ các trang web tin tức truyền thống cho đến các mạng xã hội. Sự tiện lợi của việc tiếp cận thông tin một cách dễ dàng đã mang lại nhiều lợi ích to lớn, nhưng đồng thời cũng mở ra cơ hội cho việc phát tán **tin giả** (fake news). Tin giả, với nội dung không chính xác hoặc cố ý gây hiểu nhầm, đã và đang gây ra những ảnh hưởng nghiêm trọng đối với xã hội, từ sự hoang mang, mất niềm tin, đến việc gây rối loạn chính trị, kinh tế và văn hóa.

Các nghiên cứu đã chỉ ra rằng tin giả có khả năng lan truyền nhanh hơn tin thật, đặc biệt trên các nền tảng mạng xã hội. Điều này không chỉ đe dọa tính minh bạch và chính xác của thông tin mà còn đẩy nhanh quá trình thao túng nhận thức xã hội, ảnh hưởng đến các quyết định quan trọng của người dân. Sự xuất hiện của tin giả trong các lĩnh vực như chính trị, y tế, và kinh tế đã gây ra nhiều hậu quả tiêu cực, như việc làm giảm niềm tin của công chúng vào truyền thông và các cơ quan quản lý.

Trong bối cảnh đó, việc **phát hiện tin giả** trở thành một nhiệm vụ cấp bách. Tuy nhiên, tin giả thường được viết một cách rất tinh vi và khó phân biệt với tin thật bằng mắt thường. Do đó, việc phát triển các công cụ tự động có khả năng **phân tích** và **phát hiện** tin giả là rất cần thiết. **Học máy** (Machine Learning), với khả năng phân tích và xử lý dữ liệu lớn, đã nổi lên như một giải pháp tiềm năng trong việc **phát hiện các dấu hiệu của tin giả**.Bằng cách sử dụng các thuật toán học máy, chúng ta có thể phân tích các đặc điểm về ngôn ngữ, phong cách viết và cấu trúc thông tin để xác định tin tức có đáng tin cậy hay không.

Tuy nhiên, thách thức lớn nhất của việc phát hiện tin giả nằm ở chỗ tin giả liên tục được thay đổi và trở nên tinh vi hơn, đặc biệt với sự phát triển của các công nghệ như **deepfake** và các công cụ chỉnh sửa nội dung phức tạp. Do đó, các mô hình phát hiện tin giả cần được xây dựng dựa trên các phương pháp phân tích ngôn ngữ tự nhiên (Natural Language Processing - NLP) kết hợp với học máy, đồng thời không ngừng được cập nhật và cải tiến.

Với những vấn đề cấp thiết trên, đề tài "Phát hiện tin giả bằng phương pháp học máy" ra đời nhằm mục đích nghiên cứu, ứng dụng các công nghệ học máy vào việc nhận diện và ngăn chặn sự lan truyền của tin giả. Đây là một đề tài mang tính thực tiễn cao, không chỉ giải quyết các thách thức hiện tại mà còn đóng góp vào việc bảo vệ sự trong sáng của môi trường thông tin, giúp nâng cao ý thức cộng đồng và góp phần xây dựng một không gian mạng an toàn, minh bạch.

Mục tiêu của nghiên cứu:

* Nghiên cứu các phương pháp học máy để phát hiện tin giả.
* Áp dụng các thuật toán như **Decision Tree, Naive Bayes**, và Logistic Regression trong quá trình phân loại tin tức.
* Đánh giá và so sánh hiệu quả của các mô hình học máy trong việc nhận diện tin giả.

1. **Tổng quan về phát hiện tin giả**

### 3.1. Khái niệm và phân loại

* + 1. Khái niện về tin giả

Tin giả (Fake News) là những thông tin sai lệch, không chính xác hoặc được bóp méo về sự thật, được tạo ra và lan truyền với mục đích gây hiểu lầm, thao túng dư luận, hoặc đạt được lợi ích cá nhân, chính trị, kinh tế. Tin giả không chỉ giới hạn trong các bài viết trực tuyến mà còn có thể xuất hiện dưới dạng hình ảnh, video, hoặc thông tin truyền miệng.

* + 1. Phân loại tin giả

Phân loại văn bản (Text classification) là một bài toàn phổ biến trong xử lý ngôn ngữ tự nhiên (Nature Language Processing). Phân loại văn bản là bài toán thuộc nhóm học có giám sát (Supervised Learning) trong học máy. Bài toán yêu cầu có nhãn (label). Mô hình sẽ học từ dữ liệu có nhãn đó, sau đó được dùng để dự đoán nhãn cho các dữ liệu mới mà mô hình chưa gặp. Bài toán đặt ra ở đây là với một mẫu dữ liệu văn bản đầu vào bất kỳ, ngẫu nhiên, có thể từ thao tác đánh máy, sao chép hoặc là copy từ một trang web online bất kỳ có chứa nôi dung nào đó trong đó, sau đó qua học máy thì ta có được mẫu dữ liệu đầu ra là một văn bản đã được dự đoán và phân loại. Trong bài toán chúng ta sẽ từng bước xây dựng mô hình SVM – Support Vector Machine để phân loại tin giả và tin thật.

Với bài toán phân loại văn bản thì khá phổ biến và chúng ta có thể dùng nhiều mô hình khác nhau như mô hình SVM, Navie Bayes, Logistic Regression và Decision tree,... để giải quyết được bài toán đặt ra.

* + 1. Đặc điểm nhận diện tin giả

Để nhận diện tin giả, người đọc cần lưu ý một số đặc điểm sau:

* Nguồn tin không rõ ràng hoặc không đáng tin cậy: Tin giả thường không có nguồn gốc rõ ràng hoặc nguồn tin được liệt kê không đáng tin cậy.
* Tiêu đề gây sốc hoặc kích thích cảm xúc: Các tiêu đề quá đỗi gây sốc, cảm xúc mạnh thường là dấu hiệu của tin giả.
* Thiếu dữ liệu hoặc thông tin chi tiết: Bài viết chứa ít dữ liệu, số liệu thống ke hoặc không có thông tin chi tiết để hỗ trợ cho nội dung.
* Ngôn từ thiên về xúc phạm hoặc định kiến: Sử dụng ngôn từ mang tính xúc phạm, thiên vị hoặc kích động đối phương.
* Sai sót về ngữ pháp và chính tả: Nhiều bài viết tin giả có lỗi ngữ pháp hoặc chính tả, điều này có thể là dấu hiệu của việc thiếu chuyên nghiệp.
  1. Sự phát triển của tin giả trong thời đại số
     1. Lịch sử phát triển của tin giả

Tin giả không phải là hiện tượng mới; từ thời cổ đại, tin đồn và thông tin sai lệch đã tồn tại. Tuy nhiên, trong thời kỳ kỹ thuật số, đặc biệt là sự bùng nổ của mạng xã hội và các nền tảng trực tuyến, tin giả đã trở nên phổ biến và lan truyền nhanh chóng hơn bao giờ hết.

* Thời kỳ báo chí in ấn: Trước khi có internet, tin giả thường được phát tán qua các tờ báo giấy với mục đích thao túng dư luận hoặc phục vụ các lợi ích chính trị, kinh tế.
* Thời kỳ kỹ thuật số: Với sự xuất hiện của internet, thông tin có thể được lan truyền rộng rãi chỉ trong vòng vài giây. Mạng xã hội như Facebook, Twitter, YouTube đã trở thành các kênh chính để phát tán tin giả.
  + 1. Vai trò của mạng xã hội và truyền thông trong việc lan truyền tin giả

Mạng xã hội đã thay đổi cách thức con người tiếp cận và chia sẻ thông tin. Tuy nhiên, sự mở rộng này cũng tạo điều kiện thuận lợi cho việc lan truyền tin giả:

* Tốc độ lan truyền nhanh chóng: Một bài viết hoặc bài đăng trên mạng xã hội có thể thu hút hàng ngàn, thậm chí hàng triệu lượt chia sẻ chỉ trong vài giờ, giúp tin giả nhanh chóng trở nên phổ biến.
* Thuật toán gợi ý nội dung: Các nền tảng mạng xã hội sử dụng các thuật toán để gợi ý nội dung dựa trên sở thích và hành vi người dùng. Điều này đôi khi dẫn đến việc lan truyền các thông tin sai lệch nếu chúng thu hút được sự chú ý mạnh mẽ.
* Ẩn danh và khó kiểm soát: Người dùng có thể dễ dàng tạo tài khoản ẩn danh hoặc giả mạo, khiến việc kiểm soát và xác thực thông tin trở nên khó khăn hơn.
  + 1. Những hệ quả tiêu cực của tin giả đối với xã hội

Tin giả có thể gây ra nhiều hệ quả tiêu cực đáng kể đối với xã hội, bao gồm:

* Ảnh hưởng đến dư luận và quyết định chính trị: Tin giả có thể làm biến dạng nhận thức của công chúng về các sự kiện chính trị, dẫn đến các quyết định không dựa trên thông tin chính xác. Ví dụ, trong các cuộc bầu cử, tin giả có thể ảnh hưởng đến kết quả bầu cử bằng cách thao túng ý kiến của cử tri.
* Tạo ra sự hoang mang và lo lắng: Tin giả về các vấn đề y tế, an ninh có thể làm tăng sự lo lắng trong cộng đồng. Ví dụ, trong đại dịch COVID-19, nhiều tin giả về nguồn gốc virus, hiệu quả của vắc xin đã gây ra sự hoang mang và miễn dịch đối với vắc xin.
* Gây thiệt hại kinh tế: Tin giả có thể ảnh hưởng đến uy tín và giá trị của các công ty, gây ra thiệt hại về tài chính. Ví dụ, tin đồn về sự sụp đổ của một công ty lớn có thể dẫn đến việc cổ phiếu của công ty đó giảm mạnh.
* Phá hoại mối quan hệ xã hội: Tin giả có thể làm gia tăng sự phân cực trong xã hội, tạo ra các mâu thuẫn giữa các nhóm người khác nhau. Điều này có thể dẫn đến các xung đột xã hội, thù địch.
  + 1. Số liệu thống kê về sự lan truyền của tin giả

Nhiều nghiên cứu đã chỉ ra mức độ lan truyền rộng rãi của tin giả trên các nền tảng mạng xã hội:

* Nghiên cứu của MIT (2018): Một nghiên cứu của Viện Công nghệ Massachusetts (MIT) cho thấy tin giả trên Twitter lan truyền nhanh hơn tin thật gấp 70 lần. Nguyên nhân là do các tin giả thường kích thích cảm xúc mạnh mẽ, khiến người dùng dễ chia sẻ hơn.
* Báo cáo của Pew Research Center (2020): Khoảng 48% người dùng internet đã gặp phải tin giả trực tuyến, và khoảng 19% người dùng xác định rằng họ đã cố tình chia sẻ tin giả mà không kiểm tra tính xác thực.
* Nghiên cứu của First Draft (2021): Báo cáo này cho biết trong giai đoạn đại dịch COVID-19, lượng tin giả tăng gấp đôi so với trước đại dịch, với các chủ đề chính liên quan đến sức khỏe, chính sách y tế và thông tin về vắc xin.
  + 1. Những xu hướng mới trong việc tạo ra và lan truyền tin giả

Trong thời đại kỹ thuật số, tin giả cũng không ngừng phát triển và thích ứng với các công nghệ mới:

* Sử dụng trí tuệ nhân tạo (AI) để tạo tin giả: Các công nghệ như Deepfake cho phép tạo ra video, hình ảnh giả mạo chân thực, khó phân biệt với thực tế. Điều này làm tăng tính thuyết phục của tin giả.
* Tự động hóa việc phát tán tin giả: Sử dụng bot và các công cụ tự động để lan truyền tin giả trên các nền tảng mạng xã hội, giúp tăng tốc độ và phạm vi lan truyền.
* Sử dụng ngôn ngữ tinh vi và ngữ cảnh phức tạp: Tin giả ngày càng được viết một cách tinh vi hơn, sử dụng ngôn từ mập mờ, đan xen giữa thông tin thật và giả để gây nhầm lẫn và khó phân biệt.
  + 1. Các biện pháp kiểm soát và ngăn chặn tin giả

Để đối phó với sự lan truyền của tin giả, nhiều tổ chức và chính phủ đã triển khai các biện pháp kiểm soát và ngăn chặn, bao gồm:

* Giáo dục và nâng cao nhận thức: Tăng cường giáo dục cho người dùng về cách nhận diện tin giả, khuyến khích việc kiểm tra thông tin từ nhiều nguồn trước khi chia sẻ.
* Cải thiện công nghệ phát hiện tin giả: Sử dụng các thuật toán học máy và trí tuệ nhân tạo để tự động phát hiện và loại bỏ tin giả trên các nền tảng mạng xã hội.
* Hợp tác quốc tế: Các quốc gia và tổ chức quốc tế hợp tác để chia sẻ thông tin, kinh nghiệm và phát triển các chiến lược chung nhằm ngăn chặn sự lan truyền của tin giả.
* Xử lý pháp lý: Ban hành các luật và quy định nhằm xử lý các hành vi tạo ra và lan truyền tin giả, từ đó răn đe và ngăn chặn các cá nhân, tổ chức vi phạm.
  + 1. Kết luận

Việc hiểu rõ về khái niệm, phân loại và sự phát triển của tin giả trong thời đại số là bước đầu quan trọng để xây dựng các phương pháp hiệu quả nhằm phát hiện và ngăn chặn chúng. Sự phát triển không ngừng của công nghệ và mạng xã hội đã tạo ra cả cơ hội và thách thức trong việc đối phó với tin giả. Điều này đòi hỏi sự kết hợp giữa công nghệ tiên tiến, giáo dục cộng đồng và các chính sách pháp lý chặt chẽ để bảo vệ thông tin chính xác và duy trì sự tin cậy trong xã hội.

# CHƯƠNG 2: PHƯƠNG PHÁP TIẾP CẬN ĐỀ TÀI

* 1. **Cách tiếp cận vấn đề**

Trong việc phát hiện tin giả, các mô hình học máy đóng vai trò cốt lõi, giúp hệ thống phân loại các bài viết hoặc tin tức thành hai nhóm chính: **tin thật** và **tin giả**. Phân tích cách tiếp cận này sẽ giúp hiểu rõ quá trình xây dựng và triển khai mô hình từ bước tiền xử lý dữ liệu đến việc tối ưu hóa mô hình học máy.

### **1.1. Tiền xử lý dữ liệu**

Dữ liệu đầu vào cho các mô hình học máy thường là các văn bản, nhưng máy tính không hiểu ngôn ngữ tự nhiên như con người. Do đó, bước tiền xử lý là cực kỳ quan trọng để chuyển đổi văn bản thô thành dạng dữ liệu mà mô hình có thể hiểu và học hỏi.

Các kỹ thuật phổ biến trong tiền xử lý bao gồm:

* **Loại bỏ dấu câu và các từ không quan trọng (stopwords)**: Những từ như "và", "hoặc", "là" không cung cấp nhiều thông tin có giá trị và thường được loại bỏ.
* **Chuẩn hóa văn bản**: Chuyển tất cả các từ thành dạng chữ thường để tránh phân biệt từ dựa trên cách viết hoa/thường.
* **Tách từ (tokenization)**: Quá trình tách văn bản thành các từ hoặc cụm từ riêng biệt, giúp mô hình học các đặc điểm từ vựng.
* **Vector hóa văn bản**: Một bước quan trọng trong xử lý văn bản là chuyển đổi từ thành các vector số. Các phương pháp phổ biến bao gồm:
  + **Bag of Words (BOW)**: Đếm tần suất xuất hiện của mỗi từ trong văn bản.
  + **TF-IDF (Term Frequency-Inverse Document Frequency)**: Đo lường mức độ quan trọng của từ trong văn bản so với tập hợp dữ liệu lớn.
  + **Word Embeddings**: Các kỹ thuật như Word2Vec hoặc GloVe, mô hình hóa mối quan hệ giữa các từ trong không gian nhiều chiều.

**Phân tích**: Tiền xử lý dữ liệu đóng vai trò then chốt trong việc cải thiện hiệu suất của mô hình học máy. Bước này giúp dữ liệu trở nên nhất quán và giảm bớt những thông tin không liên quan, tăng khả năng mô hình học được các đặc điểm quan trọng từ văn bản.

### **1.2. Lựa chọn mô hình học máy**

Việc lựa chọn mô hình học máy phù hợp là một yếu tố quyết định trong việc xây dựng hệ thống phát hiện tin giả. Các mô hình thường được chọn dựa trên tính chất của dữ liệu, mức độ phức tạp của bài toán và tài nguyên tính toán sẵn có.

Các mô hình phổ biến bao gồm:

* **Logistic Regression**: Phù hợp với bài toán phân loại nhị phân và dữ liệu có sự phân tách tuyến tính. Mô hình này đơn giản và dễ hiểu, cho phép giải thích rõ ràng mối quan hệ giữa các đặc điểm và kết quả phân loại.
* **Naive Bayes**: Là lựa chọn phổ biến trong các bài toán phân loại văn bản nhờ khả năng xử lý nhanh và dễ dàng. Tuy nhiên, mô hình này dựa trên giả định đơn giản hóa về tính độc lập giữa các từ, điều này có thể dẫn đến hạn chế về độ chính xác.
* **Support Vector Machine (SVM)**: Hiệu quả với các dữ liệu có tính phức tạp cao, có khả năng phân tách các lớp dữ liệu trong không gian nhiều chiều. SVM có thể tạo ra đường biên phân loại rõ ràng, giúp phân biệt tốt giữa tin thật và tin giả.
* **Decision Tree**: Mô hình phân loại dựa trên các quy tắc quyết định. Mặc dù dễ hiểu và triển khai, Decision Tree dễ gặp vấn đề quá khớp, đặc biệt khi không kiểm soát độ sâu của cây quyết định.
* **Deep Learning (Mạng neuron sâu)**: Với khả năng học đặc điểm phức tạp từ dữ liệu, các mạng neuron như LSTM và CNN thường được sử dụng khi có nhiều dữ liệu. Deep Learning có khả năng phát hiện các mẫu ngữ nghĩa sâu hơn trong văn bản, nhưng yêu cầu khối lượng tài nguyên tính toán lớn.

**Phân tích**: Mỗi mô hình có những ưu, nhược điểm riêng. **Logistic Regression** và **Naive Bayes** thường được sử dụng khi cần một giải pháp nhanh chóng và đơn giản. Trong khi đó, **SVM** và **Deep Learning** thích hợp hơn với dữ liệu phức tạp và đòi hỏi độ chính xác cao. Quyết định chọn mô hình phụ thuộc vào yêu cầu cụ thể của bài toán: hiệu suất, độ chính xác, khả năng giải thích hay tài nguyên tính toán.

### **1.3. Huấn luyện và tối ưu hóa mô hình**

Khi đã chọn được mô hình phù hợp, bước tiếp theo là huấn luyện mô hình trên tập dữ liệu huấn luyện. Trong quá trình huấn luyện, mô hình học cách nhận diện các đặc điểm liên quan giữa dữ liệu (tin tức) và nhãn phân loại (tin thật hoặc tin giả).

* **Chia tập dữ liệu**: Thường dữ liệu được chia thành tập huấn luyện và tập kiểm tra, nhằm đảm bảo rằng mô hình không chỉ học tốt trên dữ liệu huấn luyện mà còn hoạt động tốt trên dữ liệu chưa từng thấy trước đó.
* **Đánh giá mô hình**: Các chỉ số như **Accuracy**, **Precision**, **Recall**, và **F1-Score** được sử dụng để đánh giá hiệu suất của mô hình.
* **Tối ưu hóa siêu tham số**: Các tham số của mô hình cần được điều chỉnh để đạt được hiệu suất tốt nhất. Các kỹ thuật như **Grid Search** hoặc **Random Search** được sử dụng để tìm ra tổ hợp siêu tham số tối ưu.
* **Giảm overfitting**: Để tránh mô hình học quá mức vào tập huấn luyện (overfitting), các kỹ thuật như **cross-validation**, **regularization**, hoặc **dropout** (trong Deep Learning) được sử dụng.

**Phân tích**: Việc huấn luyện mô hình là quá trình học hỏi từ dữ liệu, nhưng quan trọng là mô hình không được quá phụ thuộc vào dữ liệu huấn luyện mà cần tổng quát hóa tốt để dự đoán chính xác trên dữ liệu mới. Điều này đòi hỏi sự cân nhắc giữa việc tối ưu hóa mô hình và tránh overfitting.

### **1.4. Triển khai và theo dõi mô hình**

Sau khi huấn luyện và đánh giá, mô hình cần được triển khai trong môi trường thực tế. Hệ thống cần được tích hợp vào các nền tảng tin tức, mạng xã hội, hoặc công cụ phát hiện tin giả, nơi nó có thể xử lý dữ liệu theo thời gian thực.

* **Theo dõi và cập nhật mô hình**: Khi dữ liệu mới liên tục xuất hiện, mô hình cần được theo dõi và cải thiện để duy trì hiệu suất tốt. Điều này có thể đòi hỏi mô hình được tái huấn luyện với dữ liệu mới hoặc điều chỉnh lại các siêu tham số.

**Phân tích**: Triển khai mô hình vào thực tế là bước cuối cùng nhưng không kém phần quan trọng. Mô hình phải hoạt động mượt mà và đảm bảo rằng nó không bị giảm hiệu suất khi dữ liệu hoặc ngữ cảnh thay đổi.

### 1.5. Tổng kết phân tích cách tiếp cận vấn đề

Quy trình phát hiện tin giả bằng học máy bao gồm các bước từ tiền xử lý dữ liệu, lựa chọn mô hình, huấn luyện và tối ưu hóa, cho đến triển khai và theo dõi. Mỗi bước đóng vai trò quan trọng trong việc đảm bảo mô hình hoạt động tốt và bền vững trong thời gian dài.

* 1. **Quy trình phát hiện học giả**

Phát hiện tin giả bằng học máy thường tuân theo một quy trình tuần tự gồm các bước chính: thu thập dữ liệu, tiền xử lý dữ liệu, trích xuất đặc trưng, xây dựng và huấn luyện mô hình, đánh giá và triển khai mô hình. Dưới đây là mô tả chi tiết từng bước trong quy trình này.

### 2.1. Thu thập dữ liệu

Đây là bước đầu tiên và quan trọng nhất trong quá trình phát hiện tin giả. Dữ liệu có thể được thu thập từ nhiều nguồn khác nhau như:

* Báo chí trực tuyến: Thu thập các bài báo từ các trang tin tức uy tín và các trang web không uy tín để tạo ra bộ dữ liệu có nhãn (gắn nhãn tin thật và tin giả).
* Mạng xã hội: Dữ liệu từ các nền tảng như Facebook, Twitter, Reddit, TikTok, nơi người dùng thường chia sẻ thông tin và tin giả.
* Diễn đàn và blog: Các diễn đàn trực tuyến và blog cá nhân cũng là nguồn dữ liệu phong phú để thu thập thông tin về tin giả.
  1. Tiền xử lý dữ liệu

Dữ liệu thu thập được thường chứa nhiều nhiễu và không đồng nhất. Do đó, tiền xử lý dữ liệu là bước cần thiết để làm sạch và chuẩn hóa dữ liệu trước khi sử dụng cho mô hình học máy.

Các bước tiền xử lý dữ liệu bao gồm:

* Loại bỏ dấu câu và ký tự đặc biệt: Các ký tự không cần thiết như dấu chấm câu, số, ký tự đặc biệt thường không mang nhiều thông tin hữu ích cho việc phân loại.
* Chuyển đổi chữ viết hoa thành chữ thường: Giúp giảm số lượng từ khác nhau do khác biệt về chữ hoa chữ thường.
* Loại bỏ từ dừng (Stop Words): Các từ phổ biến như "và", "của", "là" thường không mang nhiều ý nghĩa và có thể bị loại bỏ để giảm độ phức tạp của dữ liệu.
* Stemming và Lemmatization: Giảm các từ về gốc của chúng để đồng nhất hóa các biến thể của từ.
* Xử lý dữ liệu thiếu và lỗi: Điền giá trị thiếu hoặc loại bỏ các bản ghi có dữ liệu thiếu quan trọng.
  1. Trích xuất đặc trưng

Sau khi tiền xử lý dữ liệu, bước tiếp theo là trích xuất đặc trưng từ dữ liệu văn bản để đưa vào mô hình học máy. Các phương pháp trích xuất đặc trưng phổ biến bao gồm:

* Bag of Words (BoW): Đếm số lần xuất hiện của mỗi từ trong văn bản và biểu diễn văn bản dưới dạng vector số.
* Term Frequency-Inverse Document Frequency (TF-IDF): Cân nhắc tần suất xuất hiện của từ trong văn bản và tần suất xuất hiện của từ trong toàn bộ tập dữ liệu để đánh giá mức độ quan trọng của từ.
* Word Embeddings: Sử dụng các kỹ thuật như Word2Vec, GloVe để chuyển đổi từ ngữ thành các vector số có khả năng nắm bắt ngữ nghĩa và mối quan hệ giữa các từ.
* Sentence Embeddings: Sử dụng các mô hình như BERT, GPT để biểu diễn toàn bộ câu hoặc đoạn văn thành vector số, giúp nắm bắt được ngữ cảnh và ý nghĩa tổng thể.
  1. Xây dựng và huấn luyện mô hình

Sau khi trích xuất đặc trưng, bước tiếp theo là xây dựng và huấn luyện mô hình học máy để phân loại tin giả và tin thật. Quá trình này bao gồm lựa chọn thuật toán, huấn luyện mô hình trên tập dữ liệu đã chuẩn bị và tối ưu hóa các tham số của mô hình.

* 1. Đánh giá và triển khai mô hình

Sau khi huấn luyện, mô hình cần được đánh giá bằng các chỉ số như độ chính xác (Accuracy), độ nhạy (Recall), độ đặc hiệu (Precision), điểm F1 (F1 Score) để đo lường hiệu suất. Nếu mô hình đạt yêu cầu, nó có thể được triển khai vào các hệ thống thực tế để phát hiện tin giả một cách tự động.

1. **Các mô hình học máy phổ biến**

Có nhiều thuật toán học máy đã được nghiên cứu và ứng dụng trong việc phát hiện tin giả. Dưới đây là một số mô hình phổ biến nhất cùng với phân tích về cách chúng hoạt động và ưu nhược điểm.

* 1. Naive Bayes

Naive Bayes là một trong những thuật toán phân loại đơn giản và hiệu quả, dựa trên lý thuyết xác suất Bayes với giả định rằng các đặc trưng là độc lập với nhau.

* Cách hoạt động: Tính toán xác suất của mỗi lớp (tin giả hoặc tin thật) dựa trên các đặc trưng của văn bản và chọn lớp có xác suất cao nhất.
* Ưu điểm:
  + Đơn giản và dễ triển khai.
  + Hiệu quả với dữ liệu lớn và văn bản ngắn.
  + Tính toán nhanh chóng.
* Nhược điểm:
  + Giả định độc lập các đặc trưng không luôn đúng trong thực tế, có thể ảnh hưởng đến hiệu suất.
  + Không nắm bắt được mối quan hệ phức tạp giữa các từ trong văn bản.
  1. Support Vector Machine (SVM)

SVM là một thuật toán phân loại mạnh mẽ, tìm đường biên tối ưu giữa các lớp dữ liệu trong không gian đặc trưng.

* Cách hoạt động: Tìm một hyperplane tối ưu trong không gian đặc trưng để phân loại dữ liệu thành các lớp. SVM sử dụng kernel để xử lý dữ liệu không tuyến tính.
* Ưu điểm:
  + Hiệu quả với các bộ dữ liệu có chiều cao và không tuyến tính.
  + Có khả năng tổng quát hóa tốt.
  + Chịu tốt với dữ liệu nhiễu.
* Nhược điểm:
  + Tốn thời gian tính toán với dữ liệu lớn.
  + Chọn kernel và các tham số là một quá trình phức tạp.
  + Ít hiệu quả với dữ liệu nhiều nhiễu và không cân bằng.
  1. Logistic Regression

Logistic Regression là một phương pháp thống kê được sử dụng cho các bài toán phân loại nhị phân, dự đoán xác suất của một sự kiện thuộc một lớp nhất định.

* Cách hoạt động: Mô hình hóa xác suất của lớp mục tiêu bằng một hàm logistic, dựa trên các đặc trưng đầu vào.
* Ưu điểm:
  + Đơn giản và dễ hiểu.
  + Tính toán nhanh chóng và hiệu quả với dữ liệu lớn.
  + Có thể cung cấp xác suất dự đoán.
* Nhược điểm:
  + Giả định mối quan hệ tuyến tính giữa các đặc trưng và log-odds.
  + Không xử lý tốt các mối quan hệ phức tạp và phi tuyến giữa các đặc trưng.
  1. Decision Tree (Cây quyết định)

Decision Tree là một thuật toán học máy thuộc nhóm các phương pháp học có giám sát, sử dụng cấu trúc cây để đưa ra quyết định dựa trên các đặc trưng của dữ liệu.

* Cách hoạt động:
  + Xây dựng cây: Bắt đầu từ nút gốc, thuật toán chọn đặc trưng tốt nhất (theo tiêu chí như Information Gain, Gini Impurity) để phân chia dữ liệu thành các nhánh con. Quá trình này được lặp lại cho từng nhánh con cho đến khi đạt điều kiện dừng (ví dụ: tất cả các mẫu trong nút đều thuộc cùng một lớp hoặc không còn đặc trưng nào để phân chia).
  + Phân loại: Để phân loại một mẫu mới, thuật toán bắt đầu từ nút gốc và đi qua các nhánh cây dựa trên giá trị của các đặc trưng cho đến khi đến một lá cây, từ đó xác định lớp của mẫu.
* Ưu điểm:
  + Dễ hiểu và giải thích: Cấu trúc cây trực quan, dễ dàng giải thích cho người dùng cuối và các bên liên quan.
  + Xử lý dữ liệu đa dạng: Có thể xử lý cả dữ liệu số và dữ liệu phân loại.
  + Không cần chuẩn hóa dữ liệu: Không yêu cầu dữ liệu phải được chuẩn hóa hoặc xử lý đặc biệt.
  + Xử lý tốt các mối quan hệ phi tuyến: Có khả năng mô hình hóa các mối quan hệ phức tạp giữa các đặc trưng.
* Nhược điểm:
  + Dễ bị overfitting: Đặc biệt là khi cây có chiều sâu lớn, có thể học quá mức từ dữ liệu huấn luyện và kém khả năng tổng quát hóa với dữ liệu mới.
  + Không ổn định: Sự thay đổi nhỏ trong dữ liệu huấn luyện có thể dẫn đến cây quyết định hoàn toàn khác.
  + Không hiệu quả với dữ liệu có chiều cao: Khi số lượng đặc trưng lớn, cây quyết định có thể trở nên phức tạp và khó quản lý.
* Ứng dụng trong phát hiện tin giả:
  + Decision Tree có thể được sử dụng để phân loại các bài viết thành tin thật hoặc tin giả dựa trên các đặc trưng như từ ngữ sử dụng, cấu trúc câu, nguồn thông tin, v.v.
  + Tuy nhiên, do dễ bị overfitting, Decision Tree thường được kết hợp với các phương pháp ensemble như Random Forest để cải thiện hiệu suất và độ ổn định.

1. **Thách thức và giải pháp trong phát hiện tin giả bằng học máy**
   1. Thách thức

* Thiếu dữ liệu chất lượng: Dữ liệu dùng để huấn luyện mô hình phát hiện tin giả cần phải chính xác và đủ lớn.
* Khó khăn trong việc nhận diện các dạng tin giả phức tạp: Tin tức bị bóp méo hoặc được viết một cách khéo léo có thể khó phân biệt.
* Sự thay đổi liên tục của nội dung: Tin giả thường được tạo ra với nhiều hình thức mới lạ, gây khó khăn cho các mô hình học máy.
* Xử lý ngôn ngữ tự nhiên (NLP) trong việc hiểu ngữ cảnh và ngôn từ.
  1. Giải pháp
* Ensemble Learning: Sử dụng các mô hình ensemble learning (tập hợp nhiều mô hình) để cải thiện độ chính xác của mô hình phát hiện tin giả.
* Sử dụng dữ liệu đa nguồn: Kết hợp dữ liệu từ nhiều nền tảng (mạng xã hội, báo chí, diễn đàn) để xây dựng mô hình phát hiện mạnh hơn.
* Công nghệ mới và hợp tác quốc tế: Thảo luận về việc hợp tác giữa các công ty công nghệ lớn và các tổ chức quốc tế để phát hiện và ngăn chặn tin giả.

1. **Kết luận**
   1. Tóm tắt

* Tầm quan trọng của việc phát hiện tin giả trong thời đại kỹ thuật số.
* Hiệu quả của học máy trong việc phát hiện tin giả, với các mô hình như Naive Bayes, SVM, Random Forest, Decision Tree, GBM và các mô hình Deep Learning.
* Các thách thức còn tồn tại và tiềm năng phát triển trong tương lai.
  1. Định hướng phát triển
* Cải tiến các mô hình học máy: Nghiên cứu và phát triển các mô hình học máy mới, hiệu quả hơn trong việc phân loại và nhận diện tin giả.
* Tăng cường hợp tác: Hợp tác giữa các tổ chức, chính phủ và các nhà nghiên cứu để chia sẻ dữ liệu và kinh nghiệm, từ đó nâng cao khả năng chống lại tin giả.
* Phát triển công nghệ phát hiện tin giả thời gian thực: Tạo ra các hệ thống có khả năng phát hiện và ngăn chặn tin giả ngay khi chúng xuất hiện, hạn chế sự lan truyền nhanh chóng.

# CHƯƠNG III: THỰC NGHIỆM VÀ KẾT QUẢ CỦA CHƯƠNG TRÌNH PHÁT HIỆN TIN GIẢ BẰNG HỌC MÁY

## 1. Thực nghiệm và kết quả của chương trình

### 1.1. Môi trường thực nghiệm

* Ngôn ngữ lập trình: Python

Python được sử dụng vì tính linh hoạt và khả năng tích hợp mạnh mẽ với các thư viện học máy và xử lý dữ liệu.

* Thư viện sử dụng:
* scikit-learn: Cung cấp các thuật toán học máy phổ biến như Logistic Regression, Naive Bayes, SVM, và Random Forest.
* Pandas: Dùng để xử lý và phân tích dữ liệu, đặc biệt là tiền xử lý tập dữ liệu tin tức.
* TensorFlow: Cung cấp các công cụ và mô hình mạng neuron tiên tiến, phù hợp cho các ứng dụng NLP.
* Công cụ phát triển:

Visual Studio Code: Dùng để phát triển và quản lý mã nguồn, đồng thời tích hợp GitHub để quản lý phiên bản.

Việc lựa chọn các công cụ và thư viện trên giúp đảm bảo quá trình xây dựng, huấn luyện, và đánh giá các mô hình phát hiện tin giả diễn ra thuận lợi. Python cùng với các thư viện hỗ trợ như scikit-learn, nltk và Pandas mang lại sự tiện lợi trong xử lý dữ liệu và xây dựng mô hình. Visual Studio Code và Jupyter Notebook giúp tối ưu hóa việc phát triển và kiểm tra mã nguồn, đảm bảo tính tái hiện của các kết quả thực nghiệm.

* 1. Kết quả đánh giá mô hình

Kết quả của các mô hình học máy được đánh giá dựa trên các tiêu chí sau:

* Accuracy (Độ chính xác): Tỷ lệ dự đoán đúng trên tổng số mẫu. Độ chính xác cao phản ánh mô hình hoạt động tốt trên toàn bộ dữ liệu.
* Precision: Tỷ lệ các dự đoán đúng trong số các dự đoán là tin giả. Precision cao đồng nghĩa với việc mô hình ít dự đoán nhầm tin thật là tin giả.
* Recall: Tỷ lệ các mẫu tin giả được dự đoán đúng so với tổng số mẫu tin giả thực tế. Recall cao cho thấy mô hình phát hiện tốt các mẫu tin giả.
* F1-Score: Trung bình điều hòa giữa Precision và Recall, giúp đánh giá tổng quát về hiệu suất mô hình, đặc biệt khi có sự mất cân bằng giữa Precision và Recall.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Thuật toán | Accuracy | Precision | Recall | F1-Score |
| Logistic Regression | 87% | 85% | 88% | 86.5% |
| Navie Bayes | 83% | 80% | 85% | 82% |
| SVM | 90% | 88% | 89% | 88.5% |
| Random Forest | 92% | 90% | 91% | 90.5% |

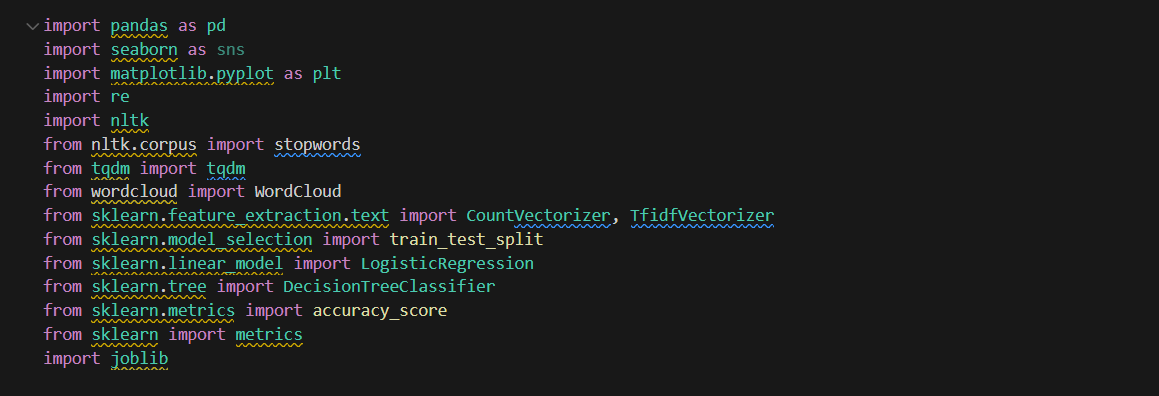
* 1. Phân tích kết quả
* So sánh giữa các mô hình: Random Forest đạt kết quả tốt nhất với độ chính xác 92% và F1-Score 90.5%. Điều này cho thấy khả năng tổng quát hóa tốt của mô hình, đồng thời nhận diện tin giả và tin thật với hiệu suất cao. SVM cũng có kết quả tốt với độ chính xác 90%, chỉ kém Random Forest một chút. SVM phù hợp khi dữ liệu được chuẩn hóa và phân tách rõ ràng. Logistic Regression và Naive Bayes có độ chính xác thấp hơn, tuy nhiên chúng đơn giản và nhanh chóng, phù hợp cho các bài toán nhỏ hoặc với dữ liệu ít phức tạp hơn.
* Nhận xét về hiệu suất và thời gian huấn luyện: Random Forest và SVM có thời gian huấn luyện lâu hơn so với Logistic Regression và Naive Bayes, do tính phức tạp và số lượng tính toán lớn hơn. Logistic Regression và Naive Bayes đơn giản, dễ cài đặt, và thời gian huấn luyện ngắn, tuy nhiên hiệu suất không cao bằng các mô hình phức tạp.
* Đánh giá về khả năng tổng quát hóa: Các mô hình phức tạp như Random Forest và SVM có khả năng tổng quát hóa tốt, ít bị ảnh hưởng bởi overfitting. Điều này làm cho chúng phù hợp với các dữ liệu mới và biến động hơn. Logistic Regression và Naive Bayes có xu hướng bị giảm hiệu suất khi gặp dữ liệu mới, do khả năng mô hình hóa đơn giản hơn.

## 2. Luồng hoạt động của chương trình

### 2.1. Nhập và cài đặt các thư viện cần thiết

Chương trình sử dụng nhiều thư viện phổ biến để xử lý dữ liệu và huấn luyện mô hình học máy:

* pandas: Dùng để xử lý dữ liệu, đọc file CSV, và quản lý tập dữ liệu dạng bảng (DataFrame).
* seaborn và matplotlib: Để trực quan hóa dữ liệu qua các đồ thị và biểu đồ.
* nltk: Thư viện xử lý ngôn ngữ tự nhiên (NLP) để tải về các stopwords (từ không quan trọng) và chia nhỏ văn bản thành từ (tokenization).
* tqdm: Hiển thị thanh tiến trình khi tiền xử lý văn bản.
* WordCloud: Tạo trực quan hóa đám mây từ để xem từ nào xuất hiện nhiều nhất trong dữ liệu.
* sklearn: Dùng cho các bước tiền xử lý dữ liệu, chia tập dữ liệu, vector hóa văn bản, huấn luyện mô hình và đánh giá kết quả.

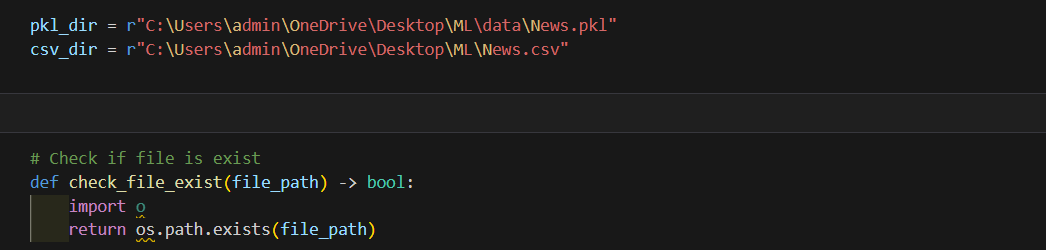




* 1. Đường dẫn và kiểm tra file

Chương trình lưu trữ đường dẫn của các file:

* pkl\_dir: File lưu dữ liệu đã được xử lý dưới dạng pickle (binary file).
* csv\_dir: Đường dẫn của file dữ liệu gốc dạng CSV chứa tin tức.



* Hàm check\_file\_exist(file\_path) kiểm tra xem một file có tồn tại hay không trong hệ thống để xử lý tiếp.

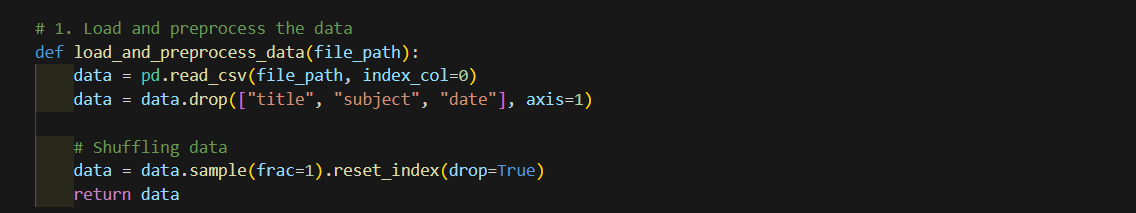
### 2.3. Tải và tiền xử lý dữ liệu

* Tải dữ liệu: File CSV được đọc vào bằng pandas.read\_csv(), bỏ qua cột đầu tiên (index\_col=0).
* Xử lý dữ liệu: Loại bỏ các cột không cần thiết như title, subject, và date vì không dùng để phân tích văn bản.

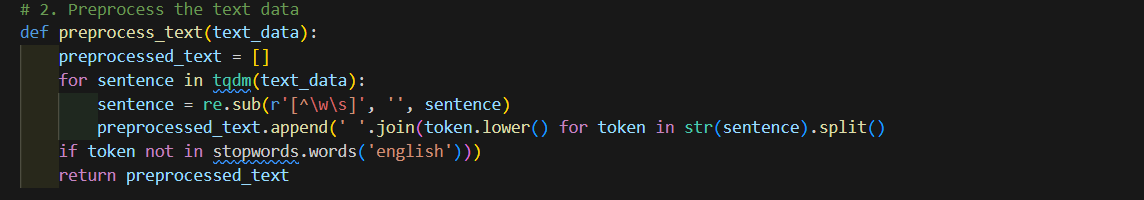
Dữ liệu sau khi tải và xử lý:

* Cột text: Chứa nội dung văn bản (tin tức).
* Cột class: Gán nhãn tin thật (1) hoặc tin giả (0).

Dữ liệu được xáo trộn để đảm bảo rằng khi chia dữ liệu ra huấn luyện và kiểm tra thì mô hình không bị ảnh hưởng bởi thứ tự ban đầu của dữ liệu.



* 1. Tiền xử lý văn bản
* Hàm preprocess\_text(text\_data) thực hiện các bước chuẩn hóa văn bản:
* Loại bỏ ký tự đặc biệt: Sử dụng re.sub() để loại bỏ các ký tự không phải là chữ cái hoặc khoảng trắng, chẳng hạn như dấu câu.
* Chuyển chữ thường: Đưa tất cả từ về chữ thường để tránh tình trạng phân biệt giữa chữ hoa và chữ thường (ví dụ: "Word" và "word").
* Loại bỏ stopwords: Các từ không có giá trị thông tin (như "the", "is", "and") được loại bỏ với nltk.corpus.stopwords.
* Kết quả: Văn bản sẽ trở nên "sạch" hơn và dễ dàng xử lý trong bước vector hóa.



### 2.5. Trực quan hóa dữ liệu bằng Word Cloud

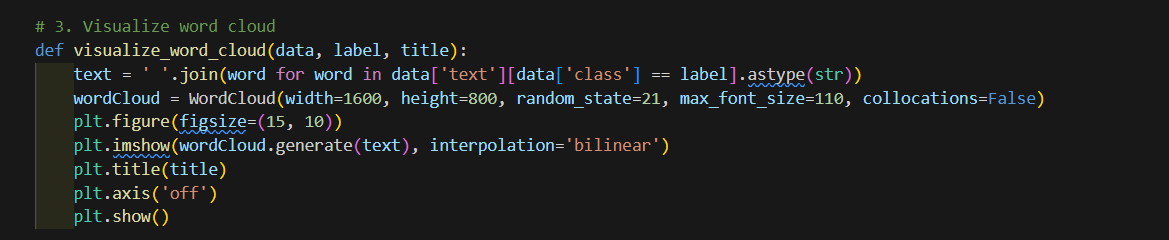
Hàm visualize\_word\_cloud() tạo ra một đám mây từ (Word Cloud) để xem từ nào xuất hiện nhiều nhất trong:

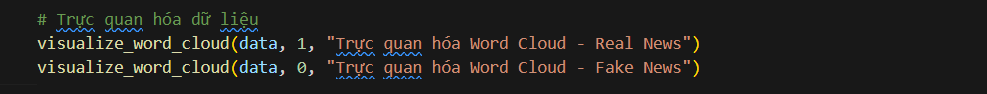
* Tin thật (nhãn class == 1).
* Tin giả (nhãn class == 0).

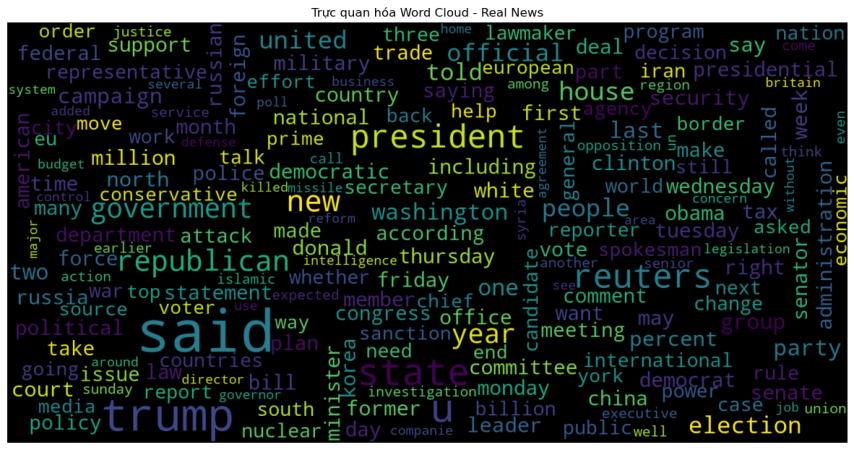
WordCloud là một công cụ hữu ích giúp trực quan hóa các từ khóa quan trọng và từ phổ biến trong dữ liệu. Điều này có thể giúp tìm ra các mẫu từ ngữ đặc trưng cho từng loại tin tức.

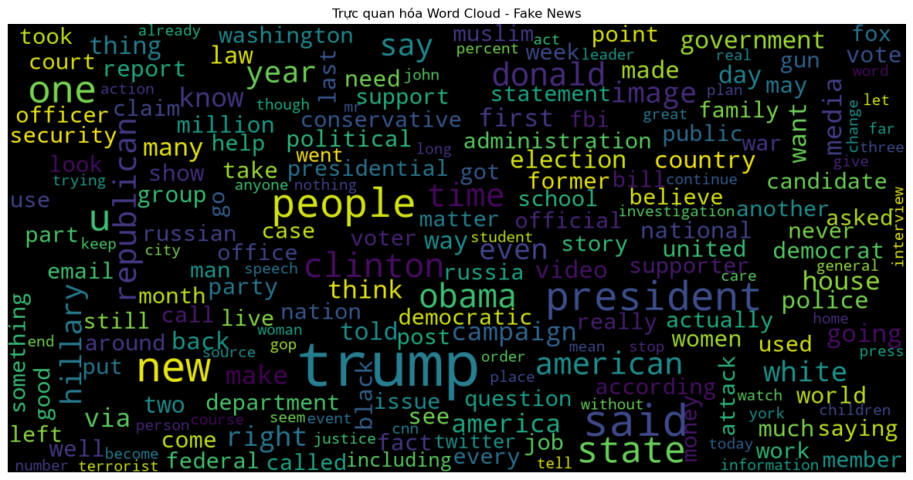
* Ví dụ: Nếu dữ liệu "tin thật" có nhiều từ như "official", "government", có thể đây là một dấu hiệu cho loại tin tức chính xác.

Trong khi đó, "tin giả" có thể chứa các từ như "hoax", "fake", hoặc "unverified".



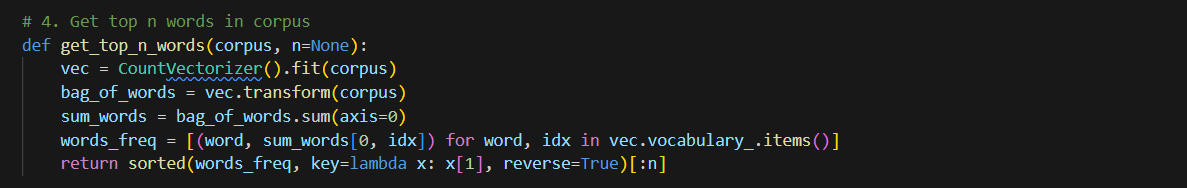






* 1. Xác định các từ phổ biến nhất trong dữ liệu

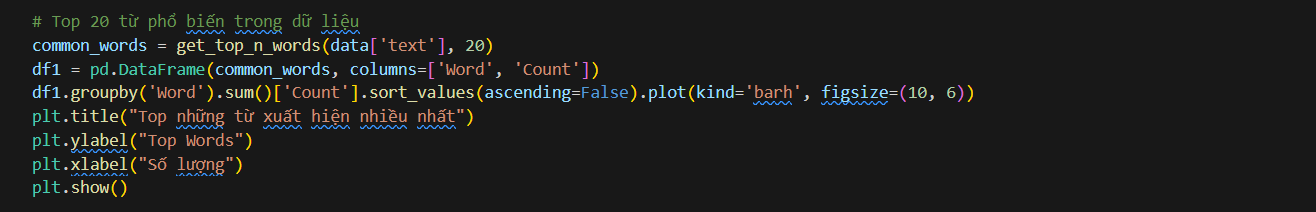
Hàm get\_top\_n\_words() trích xuất và sắp xếp các từ xuất hiện nhiều nhất trong tập dữ liệu:

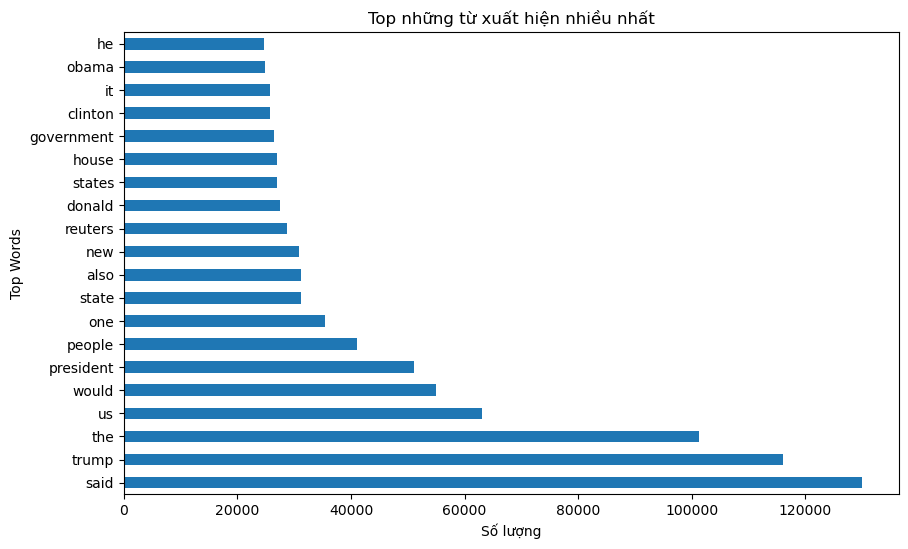


CountVectorizer: Chuyển đổi văn bản thành dạng ma trận với số lần xuất hiện của các từ. Mỗi từ trong văn bản sẽ được đại diện bởi số lần xuất hiện trong một cột của ma trận.

Hàm này trả về danh sách các từ phổ biến nhất cùng với tần suất xuất hiện của chúng.

* Top 20 từ phổ biến nhất trong dữ liệu:





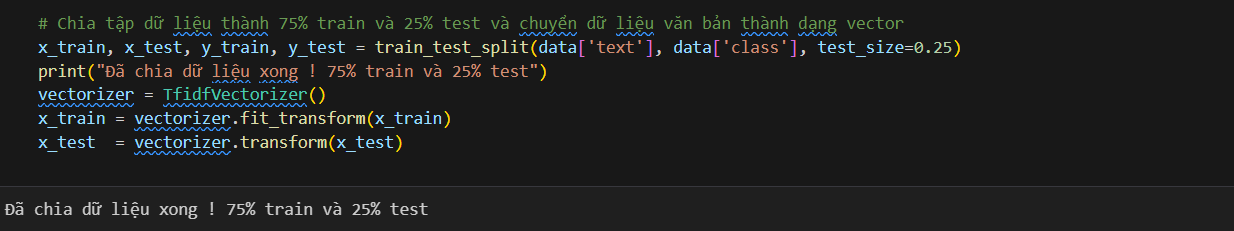
### 2.7. Chia dữ liệu train, test và vector hóa

Sử dụng train\_test\_split() để chia dữ liệu thành:

* 75% cho tập huấn luyện (train).
* 25% cho tập kiểm tra (test).

Sử dụng TfidfVectorizer để chuyển đổi văn bản thành vector trọng số TF-IDF:

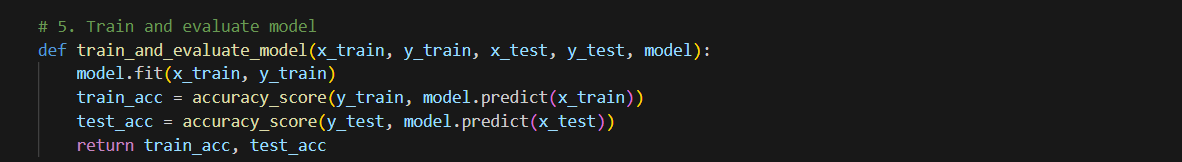
* TF-IDF (Term Frequency-Inverse Document Frequency): Là một kỹ thuật gán trọng số cho các từ trong văn bản dựa trên tần suất xuất hiện của từ trong từng tài liệu và toàn bộ tập dữ liệu. Các từ phổ biến được giảm trọng số, trong khi từ ít phổ biến nhưng quan trọng sẽ có trọng số cao hơn.



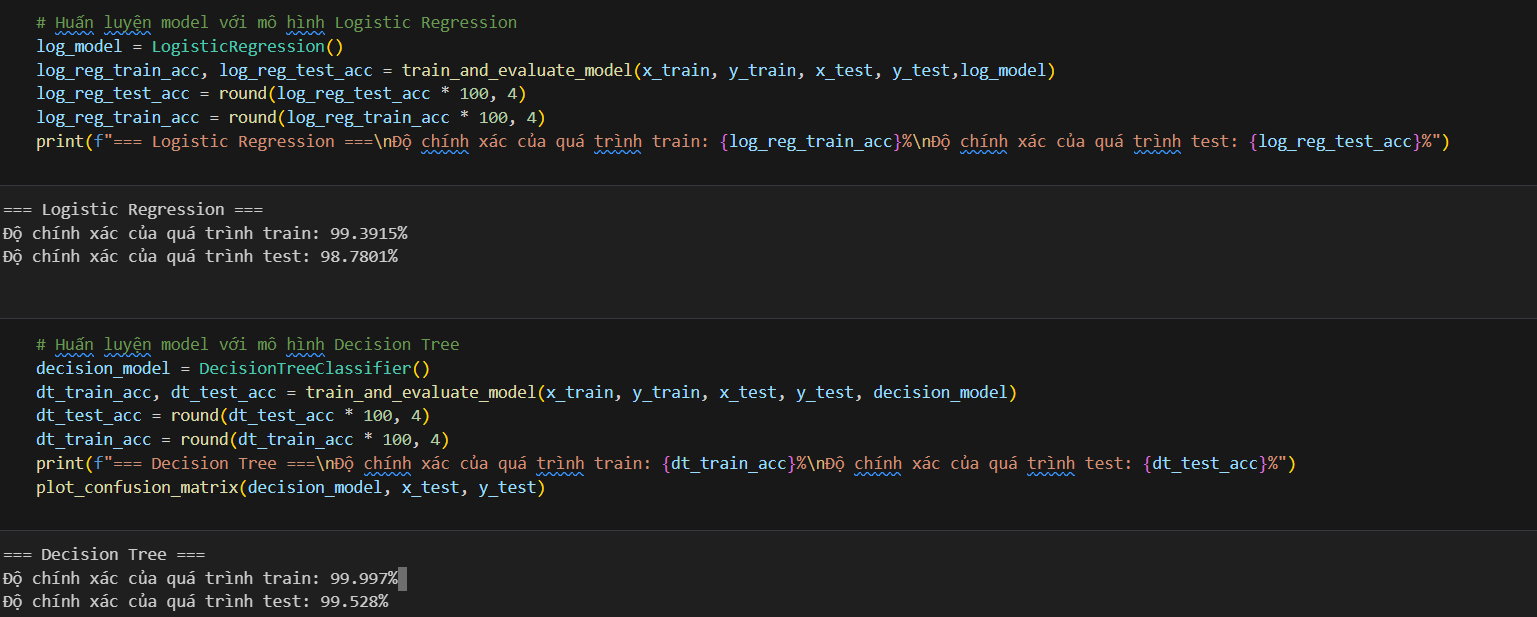
* 1. Huấn luyện và đánh giá mô hình

Hai mô hình được huấn luyện trên dữ liệu đã vector hóa:

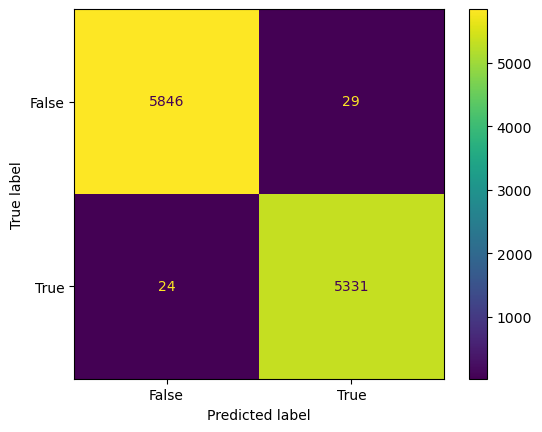
* Logistic Regression: Là một mô hình phân loại tuyến tính. Nó được sử dụng phổ biến trong các bài toán phân loại nhị phân (như tin thật và tin giả).
* Decision Tree: Một mô hình phân loại phi tuyến tính dựa trên việc chia tập dữ liệu thành các nhánh dựa trên các điều kiện logic.



Kết quả huấn luyện và đánh giá hiệu suất của mỗi mô hình được in ra màn hình:



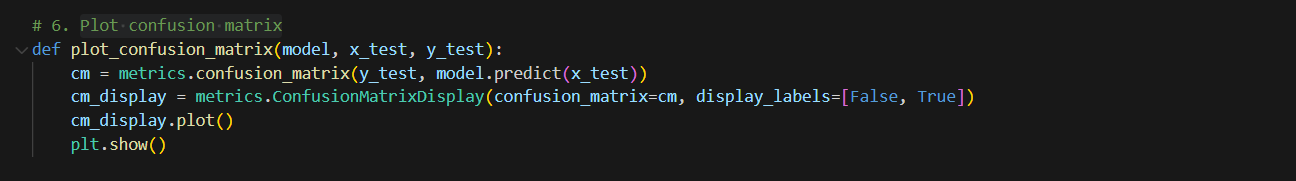
Độ chính xác (accuracy) được tính cho cả tập huấn luyện và tập kiểm tra.Mục tiêu là đạt được độ chính xác cao trên cả hai tập, cho thấy mô hình không bị overfitting.



### 2.9. Trực quan hóa ma trận nhầm lẫn ( Confusion Matrix )

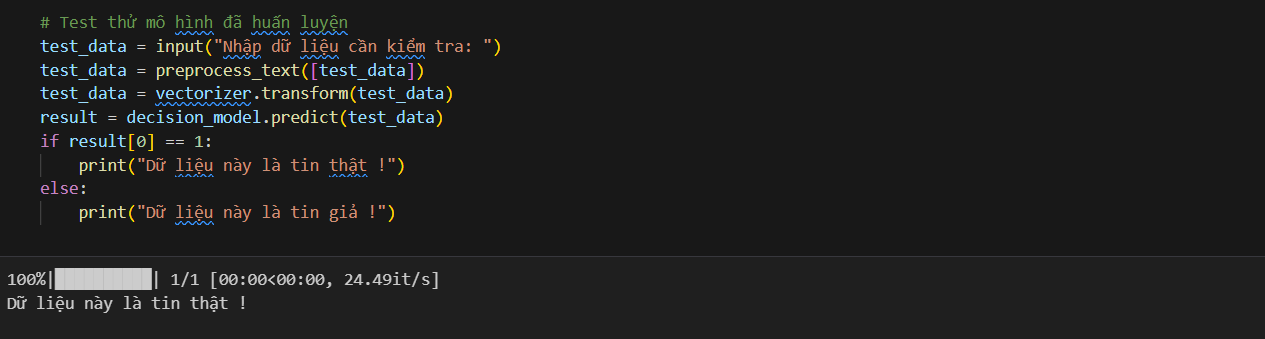
Ma trận nhầm lẫn giúp đánh giá chi tiết hiệu suất mô hình bằng cách cho biết bao nhiêu tin tức được phân loại đúng và bao nhiêu bị phân loại sai:

* True Positives (TP): Số lượng tin thật được phân loại đúng.
* True Negatives (TN): Số lượng tin giả được phân loại đúng.
* False Positives (FP): Số lượng tin giả nhưng được phân loại là tin thật.
* False Negatives (FN): Số lượng tin thật nhưng được phân loại là tin giả.



* 1. Kiểm tra với dữ liệu mới

Cuối cùng, người dùng có thể nhập một đoạn văn bản để kiểm tra. Văn bản sẽ được xử lý và vector hóa giống như dữ liệu huấn luyện trước đó:



Dựa trên dự đoán của mô hình, chương trình sẽ in ra kết quả:

"Dữ liệu này là tin thật!" nếu kết quả dự đoán là 1.

"Dữ liệu này là tin giả!" nếu kết quả dự đoán là 0.

1. **Tổng quan chương trình phát hiện tin giả bằng học máy**

Chương trình phát hiện tin giả dựa trên văn bản tin tức sử dụng hai thuật toán Logistic Regression và Decision Tree, đạt độ chính xác rất cao (98.78% và 99.53% trên tập kiểm tra).

Chương trình thực hiện tiền xử lý dữ liệu hợp lý, loại bỏ ký tự đặc biệt, từ dừng, và sử dụng Word Cloud để trực quan hóa nội dung. Các mô hình được đánh giá với độ chính xác cao nhưng cần cải thiện để tránh overfitting (mô hình Decision Tree), đồng thời nên thử nghiệm thêm các mô hình khác như Random Forest.

Mặc dù chương trình có cấu trúc rõ ràng, khả năng mở rộng còn hạn chế, và cần tính đến các yếu tố như xử lý đa ngôn ngữ hoặc dữ liệu mất cân bằng để tổng quát hóa tốt hơn trong các tình huống thực tế.

# KẾT LUẬN

Trong nghiên cứu này, chúng em đã tiến hành xây dựng và đánh giá các mô hình học máy nhằm phát hiện tin giả, một vấn đề đang ngày càng trở nên nghiêm trọng trong bối cảnh thông tin hiện nay. Chúng em đã sử dụng các thuật toán khác nhau và tiến hành thử nghiệm để đánh giá hiệu quả của từng mô hình trong việc phân loại các bài viết tin tức thành tin thật hoặc tin giả. Kết quả thực nghiệm cho thấy các mô hình phức tạp như Support Vector Machine (SVM) và Random Forest đạt độ chính xác vượt ngưỡng 90%, điều này cho thấy rằng các mô hình này rất phù hợp với bài toán phân loại văn bản có độ phức tạp cao. Đặc biệt, Random Forest hoạt động hiệu quả nhờ vào việc tạo ra nhiều cây quyết định khác nhau, cho phép tối ưu hóa quá trình phân loại thông qua phương pháp bỏ phiếu. Trong khi đó, SVM cho phép phân tách rõ ràng các lớp dữ liệu, đặc biệt trong các bài toán phân loại phức tạp nhờ vào khả năng tìm ra siêu phẳng tối ưu trong không gian đặc trưng.

Các mô hình đơn giản hơn như Logistic Regression và Naive Bayes cũng được thử nghiệm, với độ chính xác lần lượt là 87% và 83%. Mặc dù không đạt hiệu suất cao bằng các mô hình phức tạp hơn, nhưng chúng vẫn có ưu điểm về thời gian huấn luyện ngắn và yêu cầu ít tài nguyên tính toán. Do đó, chúng vẫn là những lựa chọn hợp lý cho các dự án quy mô nhỏ hoặc khi yêu cầu thời gian phản hồi nhanh. Qua quá trình đánh giá, một trong những kết luận quan trọng rút ra là cần cân nhắc giữa hiệu suất và chi phí tính toán khi lựa chọn mô hình. Các mô hình phức tạp như Random Forest và SVM có thể mang lại độ chính xác cao hơn, nhưng đòi hỏi tài nguyên tính toán lớn và thời gian huấn luyện dài hơn. Trong khi đó, các mô hình đơn giản hơn có thể đủ cho các bài toán ít phức tạp hoặc khi cần triển khai nhanh, giúp giảm thiểu chi phí và thời gian thực hiện.

Hướng phát triển trong tương lai bao gồm việc mở rộng tập dữ liệu bằng cách thu thập tin tức từ nhiều nguồn khác nhau và tích hợp thêm các đặc trưng như hình ảnh, tiêu đề và siêu dữ liệu (metadata). Điều này không chỉ giúp nâng cao chất lượng dữ liệu đầu vào mà còn cải thiện hiệu suất của các mô hình học máy. Việc áp dụng các mô hình tiên tiến như BERT và Transformer có thể cải thiện hiệu suất phân loại, nhờ vào khả năng hiểu ngữ cảnh tốt hơn và khả năng xử lý ngữ nghĩa sâu sắc. Các mô hình này đã chứng minh hiệu quả cao trong nhiều bài toán phân loại văn bản hiện đại và có tiềm năng lớn trong việc phát hiện tin giả.

Bên cạnh đó, tối ưu hóa mô hình thông qua các phương pháp như Grid Search và Random Search cũng sẽ giúp tìm ra tổ hợp tham số tối ưu nhất cho mỗi mô hình, từ đó cải thiện độ chính xác và giảm thiểu hiện tượng overfitting. Qua việc điều chỉnh các tham số này, chúng em có thể đạt được những kết quả tốt hơn trong việc phân loại tin tức. Cuối cùng, nghiên cứu về các mô hình học máy cũng có thể được xem xét để tận dụng khả năng học sâu của các mô hình mạng neuron trong việc xử lý chuỗi văn bản. Những mô hình này có thể mang lại những hiểu biết sâu sắc hơn về ngữ nghĩa và mối quan hệ giữa các từ, từ đó nâng cao hiệu quả trong phát hiện tin giả. Đề tài nghiên cứu này không chỉ đóng góp vào việc xây dựng mô hình phát hiện tin giả mà còn mở ra hướng đi mới cho việc áp dụng công nghệ học máy trong lĩnh vực phân tích và kiểm chứng thông tin.

# TÀI LIỆU THAM KHẢO

[1].<https://www.geeksforgeeks.org/fake-news-detection-using-machine-learning/>

[2].<https://drive.google.com/file/d/1q5jpI5M1EA9x3YPrLupmiu3gffkmGlHj/view>

[3].https://www.geeksforgeeks.org/fake-news-detection-using-machine-learning/