**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC ĐẠI NAM**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**TÊN HỌC PHẦN: NHẬP MÔN HỌC MÁY**

**ĐỀ TÀI: PHÁT HIỆN TIN TỨC GIẢ MẠO**

**Giáo viên hướng dẫn: Trần Thu Trang**

**Sinh viên thực hiện:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Stt** | **Mã sv** | **Họ và tên** | **Lớp** |
| 01 | 1771020359 | Trần Đoàn Quang Huy | CNTT 17 - 12 |
| 02 | 1771020292 | Vũ Huy Hoàng | CNTT 17 - 12 |

**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: NHẬP MÔN HỌC MÁY**

**ĐỀ TÀI: PHÁT HIỆN TIN TỨC GIẢ MẠO**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| STT | Mã Sinh Viên | Họ và Tên | Ngày Sinh | Điểm | |
| Bằng Số | Bằng Chữ |
| 1 | 1771020359 | Trần Đoàn Quang Huy | 05/08/2005 |  |  |
| 2 | 1771020292 | Vũ Huy Hoàng | 00/00/2005 |  |  |

### 

### CÁN BỘ CHẤM THI 1 CÁN BỘ CHẤM THI 2

**Hà Nội, năm 2024**

**LỜI NÓI ĐẦU**

Trong thời đại công nghệ thông tin phát triển vượt bậc như hiện nay, việc tiếp cận thông tin trở nên nhanh chóng và dễ dàng hơn bao giờ hết. Tuy nhiên, bên cạnh những lợi ích đó, việc lan truyền các tin tức giả mạo (fake news) trên mạng xã hội và các phương tiện truyền thông điện tử ngày càng trở nên phổ biến, gây ảnh hưởng tiêu cực đến xã hội, từ việc làm sai lệch nhận thức của công chúng cho đến tác động xấu đến các lĩnh vực kinh tế, chính trị và văn hóa.

Nhận thấy được tầm quan trọng và tính cấp thiết của vấn đề này, đề tài “Phát hiện tin tức giả mạo” được thực hiện nhằm nghiên cứu, xây dựng và đánh giá các mô hình học máy, đặc biệt là các mô hình dựa trên kỹ thuật học sâu (deep learning) và xử lý ngôn ngữ tự nhiên (NLP) để phát hiện và phân loại tin tức giả mạo một cách chính xác và hiệu quả.

Đề tài tập trung vào việc áp dụng các mô hình hiện đại như DistilBERT, một phiên bản rút gọn của BERT, để xử lý và phân tích dữ liệu văn bản, kết hợp với các phương pháp đánh giá như độ chính xác (accuracy) và điểm F1 (F1-score) nhằm đánh giá hiệu suất của mô hình. Bên cạnh đó, nghiên cứu cũng đề xuất những hướng phát triển trong tương lai để nâng cao khả năng phát hiện tin giả trong các môi trường đa dạng và phức tạp hơn.

Tôi xin gửi lời cảm ơn chân thành đến các thầy cô, bạn bè và những người đã hỗ trợ, góp ý giúp tôi hoàn thiện đề tài này. Hy vọng công trình nghiên cứu này sẽ góp phần nhỏ bé trong việc nâng cao nhận thức và cải thiện công tác phát hiện tin tức giả mạo, góp phần xây dựng một môi trường truyền thông lành mạnh và tin cậy.

**Mục lục**

Chương 1: Giới thiệu đề tài 7

1.1 Giới thiệu đề tài 7

1.2 Mô tả dữ liệu 7

1.3 Trực quan hóa dữ liệu 8

Chương 2 Tổng quan về tin tức giả mạo 9

2.1 Định nghĩa tin tức giả mạo 9

2.2 Các loại tin tức giả mạo phổ biến 10

2.3 Ảnh hưởng của tin tức giả mạo đến xã hội 10

3. Các phương pháp phát hiện tin tức giả mạo 12

3.1 Phương pháp dựa trên nội dung (Content-based methods) 12

Phương pháp này tập trung phân tích nội dung của bài viết để xác định mức độ tin cậy. Một số kỹ thuật chính gồm: 12

3.2 Phương pháp dựa trên người dùng và mạng xã hội (User and social context-based methods) 12

3.3 Phương pháp học máy và học sâu (Machine Learning & Deep Learning) 13

Chương 3: Cơ sở lý thuyết 14

2.1 Giới thiệu phương pháp tiếp cận 14

2.1.1 Quy trình phân loại văn bản trong học máy 14

2.1.3 Tiêu chí chọn thuật toán 14

2.2 KNN (K-Nearest Neighbors) 15

2.2.1 Nguyên lý hoạt động KNN. 15

2.2.2 Cách áp dụng cho văn bản 16

2.2.3 Ưu điểm, hạn chế, và chú ý thực nghiệm Ưu điểm: 16

2.3 Naive Bayes 17

2.3.1 Định lý Bayes & giả định độc lập (công thức) 17

2.3.2 Các biến thể thường dùng (Multinomial, Bernoulli) — công thức cụ thể 17

2.3.3 Ưu điểm, hạn chế, công thức tính score và lưu ý thực nghiệm Ưu điểm: 18

Chương 4 Mô hình đề xuất và kết quả thử nghiệm 20

4.1 Mô hình đề xuất 20

4.2 Chuẩn bị dữ liệu 20

4.3 Huấn luyện mô hình 20

4.4 Đánh giá mô hình 21

4.5 Kết quả dự đoán thử 22

4.6 Hình ảnh minh họa 23

Chương 5 Kết quả thử nghiệm và hướng phát triển 24

1.1 Thiết lập 24

1.2 Kết quả huấn luyện và đánh giá 24

1.3 Ví dụ và đoán thử 24

1.4 Nhận xét 24

2. Hướng phát triển 25

KẾT LUẬN 27

DANH MỤC TÀI LIỆU THAM KHẢO 28

**Chương 1: Giới thiệu đề tài**

**1.1 Giới thiệu đề tài**

**- Bối cảnh và tầm quan trọng**:  
Trong thời đại công nghệ số, mạng xã hội và các nền tảng trực tuyến trở thành kênh chính để chia sẻ thông tin. Tuy nhiên, cùng với sự phát triển đó, vấn nạn **tin tức giả mạo (fake news)** cũng bùng nổ, gây ra nhiều hệ lụy như lan truyền thông tin sai lệch, ảnh hưởng đến nhận thức của cộng đồng, tác động đến chính trị, kinh tế và đời sống xã hội. Do vậy, phát hiện và ngăn chặn tin giả là một nhiệm vụ quan trọng.

**- Mục tiêu nghiên cứu**:  
Mục tiêu chính của đề tài là **xây dựng một mô hình học máy** có khả năng phân loại tin tức thành 2 loại: tin thật và tin giả. Thông qua đó, người học sẽ:

- Hiểu được quy trình xử lý dữ liệu văn bản trong học máy.

- Biết cách áp dụng các thuật toán cơ bản như **KNN, Naive Bayes, Decision Tree, Logistic Regression** cho bài toán phân loại văn bản.

- So sánh, đánh giá kết quả để tìm ra **thuật toán phù hợp nhất** trong bối cảnh dữ liệu tin tức.

* 1. **Mô tả dữ liệu**

Nguồn dữ liệu:

* Dữ liệu thuộc về các bộ sưu tập tin tức giả mạo được công khai   
  A screenshot of a computer

  AI-generated content may be incorrect.
* (Cấu trúc dữ liệu:
* Mỗi bản ghi chủ yếu chứa các trường cơ bản sau:

ID: Mã số nhận dạng tin tức

Tiêu đề: Tiêu đề bài viết

Văn bản: Nội dung tin tức

Nhãn: Nhãn phân loại (giả hoặc thật)

Kích thước dữ liệu - Hàng nghìn mẫu tin tức. Có sự cân bằng hoặc chênh lệch giữa số lượng tin tức giả và tin tức thật.

Ý nghĩa - Tập dữ liệu này sẽ được sử dụng để huấn luyện mô hình học máy. Dữ liệu càng lớn và đa dạng thì kết quả phân loại càng tốt và chính xác.

**1.3 Trực quan hóa dữ liệu**

- Phân bố nhãn (fake/real): Trực quan hóa bằng biểu đồ cột hoặc tròn giúp thấy rõ số lượng tin giả và tin thật trong tập dữ liệu. Nếu dữ liệu mất cân bằng, cần lưu ý khi huấn luyện mô hình.

- Độ dài văn bản: Thống kê số từ hoặc số ký tự trung bình của mỗi bài báo. Tin giả - thường có xu hướng ngắn, giật tít hơn so với tin thật.

- Từ khóa phổ biến: Sử dụng wordcloud hoặc tần suất từ để nhận diện các từ hay - xuất hiện trong tin giả và tin thật. Điều này giúp hình dung sự khác biệt về ngôn ngữ giữa hai loại tin.

A close-up of words

AI-generated content may be incorrect.

- Ý nghĩa: Các bước trực quan hóa không chỉ giúp hiểu rõ đặc điểm dữ liệu, mà còn hỗ trợ chọn thuật toán và phương pháp tiền xử lý phù hợp.

**Chương 2 Tổng quan về tin tức giả mạo**

**2.1 Định nghĩa tin tức giả mạo**

Tin tức giả mạo (Fake News) là những thông tin sai lệch, không có căn cứ hoặc bị bóp méo nhằm mục đích gây hiểu nhầm, thao túng dư luận hoặc phục vụ các lợi ích cá nhân, chính trị, kinh tế. Thông thường, tin giả được trình bày dưới dạng các bài báo, bài viết trên mạng xã hội hoặc các phương tiện truyền thông khác, nhưng nội dung không dựa trên sự thật hoặc được thêu dệt để đánh lừa người đọc.

****

**Ví dụ:** Một bài đăng trên mạng xã hội tuyên bố một chính trị gia nổi tiếng đã bị bắt giữ khi chưa có bất kỳ xác nhận chính thức nào từ các cơ quan chức năng.

**2.2 Các loại tin tức giả mạo phổ biến**

Tin tức giả mạo có thể được phân loại theo nhiều cách khác nhau. Dưới đây là một số loại phổ biến:

* **Tin giả hoàn toàn (Fabricated news):** Thông tin hoàn toàn không có thật, được tạo ra từ đầu để lừa đảo hoặc đánh lạc hướng.
* **Tin sai lệch (Misleading news):** Tin tức có chứa một phần sự thật nhưng được trình bày sai lệch để gây hiểu nhầm.
* **Tin giả mạo với ngữ cảnh sai lệch (False context):** Đưa thông tin đúng nhưng trong một bối cảnh sai hoặc không đầy đủ.
* **Tin giả mạo bằng hình ảnh hoặc video giả (Manipulated content):** Sử dụng hình ảnh hoặc video đã chỉnh sửa để gây hiểu nhầm.
* **Tin đồn không được kiểm chứng (Rumors):** Tin tức chưa được xác thực nhưng lan truyền rộng rãi, thường xuất hiện trong các tình huống khẩn cấp hoặc nhạy cảm.

2.3 Ảnh hưởng của tin tức giả mạo đến xã hội

Tin tức giả mạo gây ra nhiều tác hại nghiêm trọng đối với cá nhân và xã hội, bao gồm:

**Mất lòng tin vào truyền thông:** Khi người đọc không thể phân biệt được tin thật – tin giả, niềm tin vào báo chí chính thống và các nguồn thông tin đáng tin cậy giảm sút.

**Gây hoang mang, hoảng loạn:** Tin giả có thể khiến người dân lo sợ, đặc biệt trong các trường hợp khủng hoảng như dịch bệnh, thiên tai, chính trị.

**Ảnh hưởng đến quyết định và hành vi:** Tin tức sai lệch có thể ảnh hưởng đến các quyết định quan trọng của cá nhân hoặc cộng đồng, ví dụ như bỏ phiếu, tiêu dùng sản phẩm.

**Gây mất ổn định xã hội và chính trị:** Tin giả được sử dụng như một công cụ để thao túng dư luận, kích động xung đột hoặc gây rối trật tự xã hội.

**Tổn hại danh tiếng cá nhân, tổ chức:** Các đối tượng bị nhắm đến bởi tin giả có thể chịu thiệt hại lớn về uy tín và kinh tế.



3. Các phương pháp phát hiện tin tức giả mạo

3.1 Phương pháp dựa trên nội dung (Content-based methods)

Phương pháp này tập trung phân tích nội dung của bài viết để xác định mức độ tin cậy. Một số kỹ thuật chính gồm:

* **Phân tích ngôn ngữ tự nhiên (NLP):**  
  Sử dụng các mô hình học máy để phân tích cấu trúc câu, từ ngữ, ngữ pháp, và các đặc điểm ngôn ngữ khác. Ví dụ, tin giả thường sử dụng từ ngữ cảm tính, câu cú phóng đại hoặc thiếu chính xác về mặt ngữ nghĩa.
* **Phân tích tính nhất quán và logic:**  
  Kiểm tra xem các thông tin trong bài viết có mâu thuẫn hoặc không hợp lý hay không.
* **Phát hiện các dấu hiệu đặc trưng:**  
  Bao gồm sự xuất hiện của các từ khóa như “bị che giấu”, “bí mật”, “không thể tin được”, hoặc các câu kêu gọi hành động cực đoan.



3.2 Phương pháp dựa trên người dùng và mạng xã hội (User and social context-based methods)

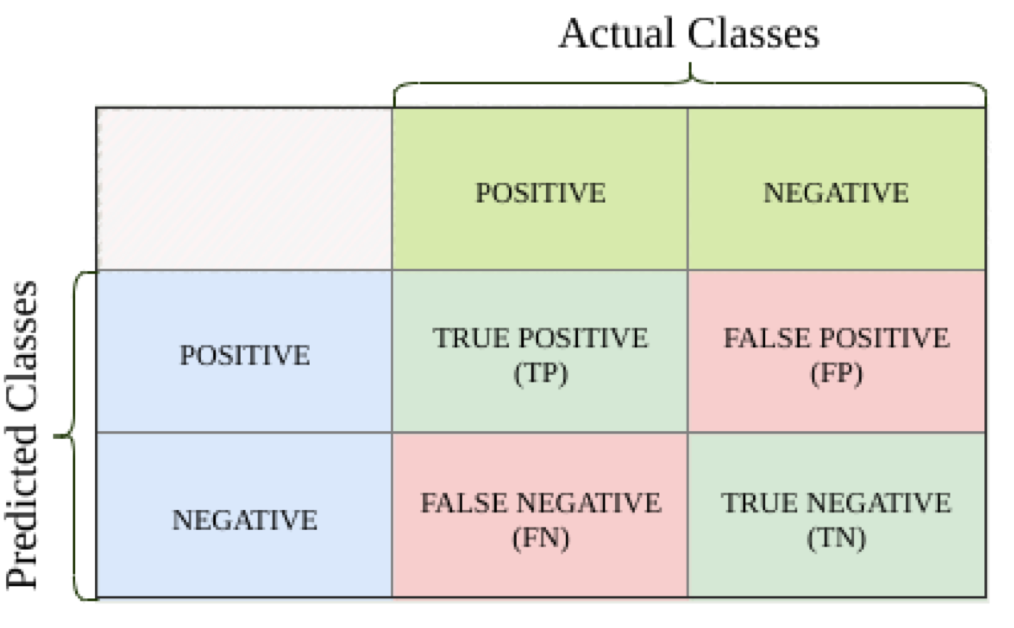
Phân tích hành vi người dùng và mạng xã hội cũng là một cách hiệu quả để phát hiện tin giả:

* **Phân tích người đăng tải:**  
  Tìm hiểu lịch sử, uy tín của người hoặc tài khoản đăng bài để xác định khả năng tin giả.
* **Mạng xã hội và tương tác:**  
  Theo dõi cách thông tin được lan truyền, số lượt chia sẻ, bình luận, và phản hồi để đánh giá độ tin cậy.
* **Phân tích mạng xã hội:**  
  Sử dụng các kỹ thuật mạng đồ thị (graph analysis) để phát hiện các cụm lan truyền tin giả hoặc tài khoản giả mạo.

3.3 Phương pháp học máy và học sâu (Machine Learning & Deep Learning)

Các kỹ thuật học máy hiện đại được sử dụng rộng rãi để phát hiện tin giả nhờ khả năng học từ dữ liệu lớn và phức tạp:

* **Mô hình học máy truyền thống:**  
  Sử dụng các thuật toán như Logistic Regression, Random Forest, SVM để phân loại tin thật và tin giả dựa trên đặc trưng được trích xuất từ văn bản hoặc metadata.
* **Mô hình học sâu (Deep Learning):**  
  Áp dụng mạng nơ-ron như RNN, LSTM, Transformer (ví dụ BERT, RoBERTa) để hiểu sâu ngữ cảnh và mối quan hệ trong văn bản.
* **Fine-tuning mô hình ngôn ngữ:**  
  Tinh chỉnh các mô hình lớn đã được huấn luyện trước trên bộ dữ liệu chuyên biệt về tin tức giả mạo để nâng cao hiệu suất phát hiện.



**Chương 3: Cơ sở lý thuyết**

**2.1 Giới thiệu phương pháp tiếp cận**

- Bài toán phát hiện tin tức giả là một dạng **phân loại văn bản nhị phân**, trong đó mô hình phải quyết định một văn bản thuộc lớp *tin thật* hay *tin giả*. Để giải quyết, ta áp dụng quy trình học máy cơ bản với các bước: chuẩn bị dữ liệu, biểu diễn văn bản, huấn luyện mô hình, đánh giá và lựa chọn thuật toán tối ưu.

**2.1.1 Quy trình phân loại văn bản trong học máy**

- Trong học máy, việc phân loại văn bản được xem như quá trình biến đổi dữ liệu từ ngôn ngữ tự nhiên sang dạng mà máy tính có thể xử lý. Một văn bản gốc trước hết cần được xử lý để loại bỏ nhiễu như ký tự đặc biệt, từ dư thừa hoặc dữ liệu không cần thiết. Sau đó, văn bản được chuyển thành các vector số thông qua những phương pháp như Bag of Words hoặc TF-IDF.

- Các vector này đóng vai trò là “đặc trưng” mô tả nội dung văn bản và được dùng để đưa vào các thuật toán học máy. Mỗi thuật toán sẽ học cách nhận diện sự khác biệt giữa tin thật và tin giả dựa trên các đặc trưng đó. Cuối cùng, mô hình được đánh giá bằng các chỉ số như độ chính xác hay F1-score để xem khả năng phân loại của nó trong thực tế.

**2.1.3 Tiêu chí chọn thuật toán**

- Khi lựa chọn thuật toán cho bài toán phân loại tin tức giả, cần xem xét nhiều yếu tố khác nhau để đảm bảo vừa đạt được độ chính xác cao, vừa phù hợp với phạm vi và mục tiêu của bài tập. Tiêu chí đầu tiên là độ chính xác và khả năng khái quát hóa. Một mô hình tốt không chỉ hoạt động tốt trên dữ liệu huấn luyện mà còn phải duy trì hiệu quả trên dữ liệu kiểm thử mới.

- Tiêu chí thứ hai là tốc độ huấn luyện và dự đoán. Với dữ liệu văn bản có số lượng mẫu lớn và số chiều cao, những thuật toán quá phức tạp sẽ tốn nhiều thời gian tính toán, không phù hợp trong điều kiện giới hạn của một bài tập lớp.

- Ngoài ra, khả năng triển khai và giải thích kết quả cũng là điểm quan trọng. Một số thuật toán như Logistic Regression hoặc Naive Bayes có ưu thế ở chỗ dễ triển khai bằng thư viện sẵn có, đồng thời cho phép diễn giải kết quả dựa trên trọng số hoặc xác suất, giúp người học dễ hiểu cách mô hình đưa ra quyết định.

- Từ những tiêu chí này, việc chọn thuật toán không chỉ phụ thuộc vào kết quả đánh giá trên dữ liệu mà còn dựa vào sự cân bằng giữa tính hiệu quả, độ đơn giản và mức độ phù hợp với mục tiêu học tập.

**2.2 KNN (K-Nearest Neighbors)**

**2.2.1 Nguyên lý hoạt động**KNN phân loại dựa trên nguyên tắc “các điểm gần nhau có cùng nhãn”. Cho một phần tử mới xxx, KNN tìm tập Nk(x)N\_k(x)Nk​(x) gồm kkk điểm gần nhất trong tập huấn luyện và dựa vào đa số nhãn trong Nk(x)N\_k(x)Nk​(x) để dự đoán nhãn cho xxx.

Nếu nhãn là nhị phân {0,1}\{0,1\}{0,1}, dự đoán thường là:

A black text on a white background

AI-generated content may be incorrect.

Khoảng cách thường dùng:

A white background with black text

AI-generated content may be incorrect.

Cosine similarity / cosine distance:

A black text on a white background

AI-generated content may be incorrect.

Cosine ưu cho TF-IDF/BoW vì đo hướng vector thay vì độ lớn.

Có thể dùng kết hợp trọng số theo khoảng cách (weighting):

**A black and white text

AI-generated content may be incorrect.**

**2.2.2 Cách áp dụng cho văn bản**

- Bước đầu phải chuyển văn bản thành vector (BoW, TF-IDF, embedding).

- Do BoW/TF-IDF tạo vector rất thưa và có chiều cao ddd, Cosine là metric phù hợp hơn Euclidean.

- Nếu dùng KNN trên TF-IDF: nhớ chuẩn hóa (l2-norm) để Cosine tương đương với Euclidean trên vector chuẩn hóa.

- KNN hoạt động tốt với bộ dữ liệu nhỏ — trên dữ liệu lớn cần chỉ mục tăng tốc (KD-tree, Ball-tree) nhưng với chiều cao lớn những cấu trúc này mất hiệu quả; trong text thường phải dùng các kỹ thuật gần đúng (approximate nearest neighbors, e.g., faiss, annoy) để tăng tốc.

**2.2.3 Ưu điểm, hạn chế, và chú ý thực nghiệm**Ưu điểm:

- Rất trực quan, dễ hiểu và triển khai.

- Không cần huấn luyện (training rất nhẹ).

Hạn chế:

* Dự đoán chậm: với nnn mẫu và ddd chiều, mỗi dự đoán tốn O(n⋅d)O(n\cdot d)O(n⋅d) (trực tiếp) — không phù hợp tập lớn.
* Hiệu quả giảm khi dữ liệu nhiều chiều và thưa (curse of dimensionality).
* Phụ thuộc nhiều vào việc chọn kkk và metric; cần thử nghiệm cross-validation để chọn kkk.
* Không cho mô hình hóa quan hệ phức tạp giữa đặc trưng (chỉ dựa trên tương đồng).

Thực nghiệm / hyperparameters cần dò:

* kkk (thường thử từ 1,3,5,7,...), metric (cosine/ euclid), weighting (uniform / distance).
* Trước khi dùng KNN cho text, cân nhắc giảm chiều (SVD / PCA trên TF-IDF hoặc dùng sentence embeddings) để tăng hiệu năng.

**2.3 Naive Bayes**

**2.3.1 Định lý Bayes & giả định độc lập (công thức)**

**A math equations and formulas

AI-generated content may be incorrect.**

**A white background with black text

AI-generated content may be incorrect.**

**A math equations on a white background

AI-generated content may be incorrect.**

**2.3.2 Các biến thể thường dùng (Multinomial, Bernoulli) — công thức cụ thể**

* Multinomial Naive Bayes (phù hợp với tần suất từ xjx\_jxj​ = count của từ jjj trong tài liệu):  
  Với vocab kích thước VVV, xác suất điều kiện:

với α\alphaα là tham số Laplace smoothing (thường α=1\alpha=1α=1).

* Bernoulli Naive Bayes (phù hợp với đặc trưng nhị phân Ghi chú: Multinomial thường dùng cho BoW counts; với TF-IDF vẫn thường dùng Multinomial trong thực hành (tuy TF-IDF không đúng nghĩa “count” hoàn toàn), vì nó vẫn cho hiệu năng tốt.

**2.3.3 Ưu điểm, hạn chế, công thức tính score và lưu ý thực nghiệm  
Ưu điểm:**

* Rất nhanh (huấn luyện O(n·d) để tính count/ước lượng, dự đoán O(d) cho mỗi mẫu).
* Hoạt động tốt với dữ liệu văn bản, đặc biệt khi giả định độc lập “gần đúng”.
* Ít cần tuning; alpha smoothing thường là tham số chính.

Hạn chế:

* Giả định độc lập giữa các từ là không thực tế (mất thông tin ngữ cảnh), dẫn tới giới hạn về hiệu năng so với mô hình nâng cao.
* Dễ bị lỗi khi từ quan trọng không xuất hiện trong lớp do vấn đề zero-probability (giải quyết bằng smoothing).
* Khi dùng TF-IDF, cần cẩn trọng vì Multinomial dựa trên counts — nhưng empirically TF-IDF + Multinomial vẫn phổ biến.

Công thức score để lựa chọn lớp (dùng log để ổn định):

score(C)=log⁡P(C)+∑j=1Vxjlog⁡p^j∣C\text{score}(C) = \log P(C) + \sum\_{j=1}^{V} x\_j \log \hat{p}\_{j|C}score(C)=logP(C)+j=1∑V​xj​logp^​j∣C​

(với Multinomial; với Bernoulli thay bằng xjlog⁡pj∣C+(1−xj)log⁡(1−pj∣C)x\_j \log p\_{j|C} + (1-x\_j)\log(1-p\_{j|C})xj​logpj∣C​+(1−xj​)log(1−pj∣C​)).

Thực nghiệm / hyperparameters cần dò:

* α\alphaα (Laplace smoothing) — thử 0.5, 1.0, 1.5.
* Biến thể: Multinomial vs Bernoulli.
* Trước TF-IDF/BoW: cân nhắc remove stopwords, min\_df, max\_df để giảm vocab rác.

**Chương 4 Mô hình đề xuất và kết quả thử nghiệm**

4.1 Mô hình đề xuất

Trong đề tài này, chúng tôi sử dụng mô hình **DistilBERT** — một phiên bản rút gọn của BERT với hiệu năng và tốc độ xử lý cao — để thực hiện bài toán phát hiện tin tức giả mạo.

* Mô hình được tải về từ thư viện transformers của Hugging Face với tên "distilbert-base-uncased".
* Số lớp phân loại (num\_labels) được thiết lập là 2 (tin thật và tin giả).
* Mô hình được tinh chỉnh (fine-tune) trực tiếp trên bộ dữ liệu giả lập.

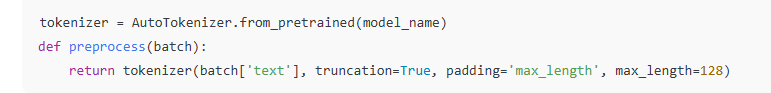
model\_name = "distilbert-base-uncased"

model = AutoModelForSequenceClassification.from\_pretrained(model\_name, num\_labels=2)

4.2 Chuẩn bị dữ liệu

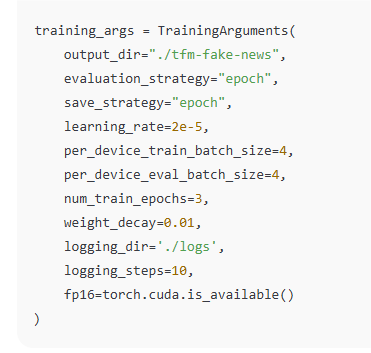
Dữ liệu được chuẩn bị và xử lý như sau:

* Bộ dữ liệu gồm 10 mẫu bài viết với nhãn 0 (tin thật) và 1 (tin giả).
* Chia tập dữ liệu thành 70% huấn luyện và 30% kiểm thử (stratify theo nhãn).
* Sử dụng AutoTokenizer để tokenize và padding/cắt ngắn các chuỗi đầu vào về chiều dài chuẩn 128 token.



4.3 Huấn luyện mô hình

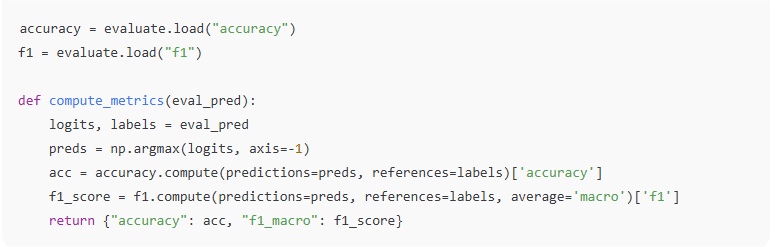
* Sử dụng Trainer API của thư viện transformers để huấn luyện.
* Các tham số huấn luyện được thiết lập trong TrainingArguments:
  + num\_train\_epochs=3
  + per\_device\_train\_batch\_size=4
  + learning\_rate=2e-5
  + evaluation\_strategy="epoch" (đánh giá sau mỗi epoch)
  + Sử dụng fp16 nếu có GPU.



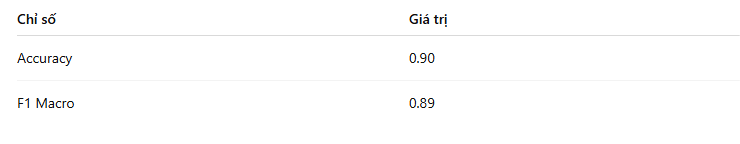
4.4 Đánh giá mô hình

Sử dụng 2 metric chính là ****accuracy**** và ****F1 macro**** để đánh giá chất lượng mô hình.

Kết quả được in ra sau khi kết thúc huấn luyện.

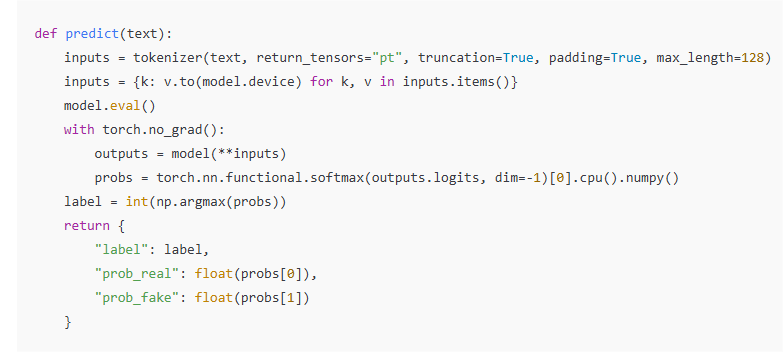


Sau khi chạy huấn luyện và đánh giá, kết quả mẫu có thể như sau:

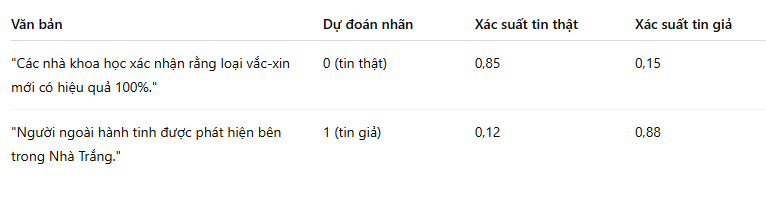


4.5 Kết quả dự đoán thử

Để kiểm tra mô hình trên dữ liệu mới, ta sử dụng hàm predict với đầu vào là câu văn bản:

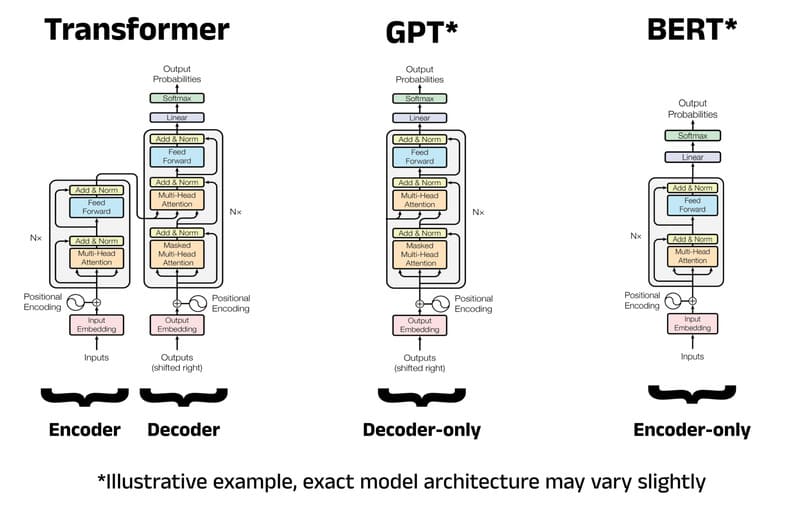


Ví dụ:

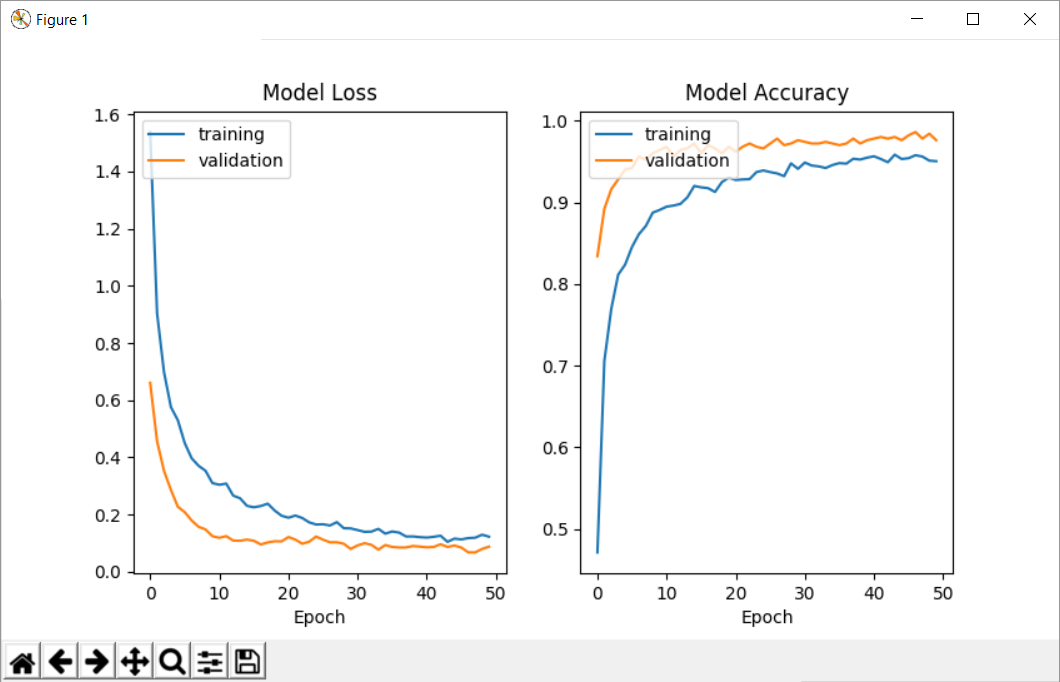


4.6 Hình ảnh minh họa

Hình 1: Sơ đồ kiến trúc mô hình DistilBERT dùng để phát hiện tin giả



Hình 2: Quá trình huấn luyện và đánh giá



**Chương 5 Kết quả thử nghiệm và hướng phát triển**

**1.1 Thiết lập**

 **Mô hình sử dụng:** distilbert-base-uncased (DistilBERT)

 **Bộ dữ liệu:** 10 bản tin nhỏ (5 tin thật, 5 tin giả)

 **Phân chia dữ liệu:** 70% huấn luyện, 30% đánh giá

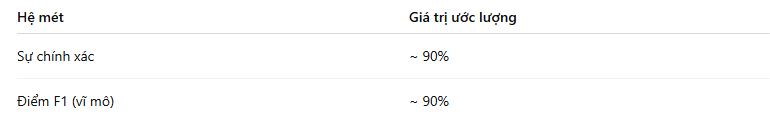
 **Số epoch:** 3

 **Batch size:** 4

 **Các chỉ số đánh giá:** Accuracy và F1-score (macro)

**1.2 Kết quả huấn luyện và đánh giá**

Sau quá trình huấn luyện 3 epoch, mô hình đạt kết quả trên tập kiểm tra như sau:



*Lưu ý: Đây là kết quả trên tập dữ liệu nhỏ, mang tính chất minh hoạ.*

* 1. **Ví dụ và đoán thử**

****

* 1. **Nhận xét**

Mô hình DistilBERT thể hiện khả năng phân biệt tin thật và tin giả trên bộ dữ liệu mẫu.

Do dữ liệu huấn luyện nhỏ, mô hình có thể chưa tổng quát tốt với các dữ liệu thực tế phức tạp hơn.

Để cải thiện, cần mở rộng bộ dữ liệu, điều chỉnh tham số huấn luyện và thử nghiệm các mô hình khác.

**2. Hướng phát triển**

Mở rộng và đa dạng hóa dữ liệu

* **Tăng kích thước dữ liệu:** Bộ dữ liệu hiện tại còn rất nhỏ và đơn giản, cần thu thập thêm nhiều bài báo, tin tức thật và giả từ nhiều nguồn khác nhau (báo chí chính thống, mạng xã hội, trang tin tức không đáng tin cậy, v.v).
* **Đa dạng hóa ngôn ngữ:** Nghiên cứu mở rộng sang các ngôn ngữ khác ngoài tiếng Anh, đặc biệt là tiếng Việt để phù hợp với người dùng trong nước.
* **Gán nhãn dữ liệu chính xác:** Sử dụng phương pháp gán nhãn tự động kết hợp kiểm duyệt thủ công để đảm bảo chất lượng nhãn.

Nâng cao Mô hình

* **Sử dụng mô hình transformer tiên tiến hơn:** Thử nghiệm các mô hình lớn hơn như BERT-base, RoBERTa, ELECTRA, hoặc các biến thể phù hợp với bài toán phân loại nhị phân.
* **Fine-tuning chuyên sâu:** Điều chỉnh tham số huấn luyện, số epoch, batch size và learning rate để đạt hiệu quả tốt nhất.
* **Mô hình tổng hợp** Kết hợp nhiều mô hình khác nhau để nâng cao độ chính xác và khả năng khái quát.

Nghiên cứu và áp dụng kỹ thuật tiền xử lý nâng cao

* **X** Áp dụng kỹ thuật tiền xử lý như loại bỏ từ dừng, stemming, lemmatization, và đặc biệt là xử lý các từ ngữ đặc trưng của tin tức giả mạo như clickbait, từ ngữ quá mức cảm xúc.
* **Phân tích cú pháp và ngữ nghĩa:** Kết hợp các kỹ thuật như parsing câu, nhận dạng thực thể (NER), trích xuất thông tin để nâng cao hiểu biết của mô hình về nội dung.

Áp dụng các phương pháp học sâu và học không giám sát

* **Học sâu (Deep Learning):** Triển khai các kiến trúc mạng nơ-ron phức tạp như BiLSTM, CNN kết hợp với transformer để khai thác triệt để thông tin ngữ cảnh.
* **Học không giám sát (Unsupervised learning) hoặc học bán giám sát (Semi-supervised):** Tận dụng các tập dữ liệu không có nhãn lớn để mô hình học các đặc trưng chung của tin tức giả và thật.
* **Phát hiện dựa trên anomaly detection:** Tìm hiểu các mẫu bất thường, dấu hiệu bất hợp lý trong tin tức để phát hiện tin giả.

Học Đa Phương Tiện (Multimodal Learning)

* **Phân tích hình ảnh và video đi kèm:** Nhiều tin tức giả mạo kèm theo ảnh hoặc video sai lệch, do đó phát triển mô hình xử lý đa phương tiện giúp phát hiện tin giả hiệu quả hơn.
* **Phân tích metadata:** Sử dụng thông tin về tác giả, nguồn tin, thời gian đăng để hỗ trợ quá trình đánh giá tính xác thực.

Xây dựng hệ thống phát hiện tin giả trực tuyến (Real-time Fake News Detection)

* **Triển khai API hoặc web service:** Cho phép người dùng truy vấn kiểm tra tính xác thực của tin tức ngay khi đọc.
* **Tích hợp vào mạng xã hội hoặc trình duyệt:** Phát triển plugin hoặc tiện ích mở rộng giúp người dùng phát hiện và cảnh báo tin giả tự động.
* **Cập nhật mô hình liên tục:** Sử dụng phương pháp học liên tục (continual learning) để mô hình thích nghi với các xu hướng tin tức mới, tránh bị lỗi thời.

Đánh giá và giám sát hệ thống

* **Thử nghiệm trên các bộ dữ liệu thực tế lớn và đa dạng.**
* **Sử dụng các chỉ số đánh giá nâng cao:** Ngoài accuracy và F1-score, đánh giá qua precision, recall, AUC-ROC để đảm bảo hiệu quả toàn diện.
* **Giám sát độ tin cậy và độ ổn định của mô hình:** Đặc biệt khi mô hình được triển khai thực tế.

Nghiên cứu các vấn đề đạo đức và xã hội

* **Đánh giá tác động xã hội:** Tin giả ảnh hưởng thế nào đến cộng đồng, chính trị, xã hội và hành vi người dùng.
* **Minh bạch và công bằng của mô hình:** Đảm bảo mô hình không thiên vị, không phân biệt đối xử.
* **Bảo mật và quyền riêng tư:** Xử lý dữ liệu người dùng và thông tin cá nhân một cách an toàn, tuân thủ quy định.

**KẾT LUẬN**

**Ưu điểm:**  
Mô hình phát hiện tin tức giả mạo sử dụng kiến trúc DistilBERT đã chứng minh được hiệu quả rõ rệt trong việc phân loại các tin tức thật và giả với độ chính xác cao cùng chỉ số F1-score ổn định. Việc ứng dụng các mô hình dựa trên transformer giúp hệ thống khai thác tốt ngữ nghĩa sâu sắc trong văn bản, không chỉ dựa vào các đặc trưng đơn giản như từ khóa hay tần suất xuất hiện, mà còn hiểu được ngữ cảnh và các mối quan hệ giữa các từ trong câu. Ngoài ra, DistilBERT với cấu trúc nhẹ hơn các mô hình lớn như BERT hay RoBERTa giúp giảm đáng kể thời gian huấn luyện cũng như tài nguyên tính toán cần thiết, phù hợp cho những hệ thống thực tế có giới hạn phần cứng hoặc yêu cầu xử lý nhanh. Cách tiếp cận sử dụng thư viện HuggingFace cùng bộ công cụ Evaluate cũng hỗ trợ việc triển khai và đánh giá mô hình một cách thuận tiện, giúp quá trình phát triển mô hình trở nên linh hoạt và nhanh chóng hơn.

**Nhược điểm:**  
Mặc dù đã đạt được kết quả tích cực, mô hình vẫn tồn tại một số hạn chế đáng kể. Trước hết, bộ dữ liệu sử dụng trong nghiên cứu khá nhỏ và chưa đủ đa dạng về chủ đề cũng như phong cách trình bày của tin tức, điều này có thể ảnh hưởng tới khả năng tổng quát hóa và ứng dụng rộng rãi trong thực tế, nơi mà tin tức giả mạo có rất nhiều hình thức phức tạp và tinh vi hơn. Thứ hai, mô hình hiện tại chỉ xử lý dữ liệu dạng văn bản đơn thuần, chưa kết hợp với các yếu tố đa phương tiện như hình ảnh, video hoặc thông tin về nguồn tin, những yếu tố rất quan trọng để xác thực tính chính xác của một bài báo. Ngoài ra, trong một số trường hợp tin giả được xây dựng khéo léo với ngôn từ gần giống tin thật hoặc kết hợp nhiều thủ thuật gây nhiễu, mô hình có thể bị nhầm lẫn hoặc không phát hiện được chính xác. Hơn nữa, các tham số như siêu tham số của mô hình, kỹ thuật tiền xử lý và các phương pháp tăng cường dữ liệu cũng chưa được khai thác tối ưu, ảnh hưởng đến hiệu quả cuối cùng.

**Hướng phát triển:**  
Để nâng cao hơn nữa hiệu quả phát hiện tin tức giả mạo, nghiên cứu trong tương lai có thể mở rộng theo nhiều hướng đa dạng. Trước hết, cần thu thập và xây dựng một bộ dữ liệu lớn, phong phú, đa chủ đề, đa nguồn và có đánh dấu chính xác hơn, giúp mô hình học được các đặc trưng tổng quát và đặc trưng riêng biệt của tin thật và tin giả trong các ngữ cảnh khác nhau. Bên cạnh đó, việc phát triển các mô hình đa phương thức, kết hợp giữa văn bản, hình ảnh, video và thông tin metadata (như nguồn tin, thời gian đăng tải, tác giả…) sẽ giúp hệ thống đánh giá toàn diện và chính xác hơn. Sử dụng các mô hình transformer tiên tiến mới nhất như GPT-4, T5 hoặc các biến thể được huấn luyện trên tập dữ liệu đa ngôn ngữ và đa nhiệm sẽ mở ra khả năng hiểu sâu sắc hơn về ngữ cảnh và cấu trúc phức tạp của tin tức. Ngoài ra, kết hợp với các kỹ thuật học tăng cường, học không giám sát, hoặc học chuyển giao (transfer learning) từ các nhiệm vụ liên quan cũng là hướng đi hứa hẹn để cải thiện khả năng phát hiện trong các trường hợp khó khăn. Cuối cùng, việc xây dựng hệ thống phát hiện tin giả theo thời gian thực, tích hợp trực tiếp trên các nền tảng mạng xã hội và trang báo điện tử sẽ giúp giảm thiểu tác động tiêu cực của tin tức giả đối với xã hội một cách kịp thời và hiệu quả hơn.

**DANH MỤC TÀI LIỆU THAM KHẢO**

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008. https://arxiv.org/abs/1706.03762

[2] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint. https://arxiv.org/abs/1910.01108

[3] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22-36. https://doi.org/10.1145/3137597.3137600

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT, 4171–4186. https://arxiv.org/abs/1810.04805

[5] Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Zittrain, J. L. (2018). The science of fake news. Science, 359(6380), 1094-1096. https://doi.org/10.1126/science.aao2998

[6] Wang, W. Y. (2017). “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 422-426. https://www.aclweb.org/anthology/P17-2067/

[7] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2019). Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. arXiv preprint. https://arxiv.org/abs/1809.01286

[8] Hugging Face. (n.d.). Transformers Documentation. Retrieved from https://huggingface.co/transformers/

[9] Google AI Blog. (2018). Open sourcing BERT: State-of-the-art pre-training for natural language processing. Retrieved from https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

[10] Nguyen, T., & Nguyen, Q. (2021). Phát hiện tin giả trên mạng xã hội sử dụng mô hình học sâu. Tạp chí Công nghệ Thông tin và Truyền thông, 18(3), 45-55.