**TRƯỜNG ĐẠI HỌC SÀI GÒN  
KHOA CÔNG NGHỆ THÔNG TIN  
  
MÔN HỌC: TÍNH TOÁN THÔNG MINH  
ĐỀ TÀI ĐỒ ÁN:**

**PHÁT HIỆN TIN GIẢ BẰNG XỬ LÝ NGÔN NGỮ TỰ NHIÊN  
(Fake News Detection Using NLP)**Giảng viên: TS. Huỳnh Minh Trí  
Nhóm 12 – Ngành Công nghệ Thông tin  
  
Thành viên nhóm:  
1. Lý Minh Phát – 3122410294  
2. Nguyễn Văn An – 3122410004  
3. Nguyễn Quang Đạt – 3122410071  
4. Nguyễn Văn Tâm Hoan – 3122410122  
5. Trần Đăng Khôi – 3124410166

# LỜI MỞ ĐẦU

Trong thời đại công nghệ thông tin phát triển mạnh mẽ, việc lan truyền thông tin trên Internet trở nên phổ biến và nhanh chóng. Tuy nhiên, điều này cũng kéo theo một vấn nạn nghiêm trọng – đó là sự xuất hiện của các tin tức giả mạo (Fake News). Các thông tin sai lệch có thể gây ra hậu quả tiêu cực đối với xã hội, chính trị, kinh tế và niềm tin của người dân. Vì vậy, việc phát triển một hệ thống có khả năng tự động phát hiện và phân loại tin giả là một nhu cầu cấp thiết.

Đề tài 'Phát hiện tin giả bằng xử lý ngôn ngữ tự nhiên' nhằm mục đích nghiên cứu và áp dụng các kỹ thuật của Natural Language Processing (NLP) kết hợp với các mô hình học máy (Machine Learning) để xây dựng một mô hình có khả năng nhận diện tin thật và tin giả một cách chính xác.

# PHẦN 1 : CƠ SỞ LÝ THUYẾT

1.1. Xử lý ngôn ngữ tự nhiên (NLP)  
NLP là lĩnh vực giúp máy tính hiểu, phân tích và xử lý ngôn ngữ của con người. Trong bài toán phát hiện tin giả, NLP đóng vai trò quan trọng trong việc chuyển đổi nội dung văn bản thành dạng dữ liệu có thể học được bởi các mô hình.

1.2. Các kỹ thuật biểu diễn văn bản:  
- Bag of Words (BoW): Biểu diễn văn bản bằng cách đếm số lần xuất hiện của từng từ.  
- TF-IDF (Term Frequency - Inverse Document Frequency): Gán trọng số cho từ dựa trên tần suất xuất hiện và mức độ đặc trưng.  
- Word Embedding: Biểu diễn từ thành vector trong không gian liên tục, giúp mô hình hiểu được ngữ nghĩa giữa các từ.

1.3. Các mô hình học máy được sử dụng:  
- Logistic Regression: Mô hình phân loại tuyến tính, dễ huấn luyện và có hiệu quả cao với dữ liệu TF-IDF.  
- Random Forest: Mô hình kết hợp nhiều cây quyết định để tăng độ chính xác và giảm overfitting.  
- BERT (Bidirectional Encoder Representations from Transformers): Mô hình ngôn ngữ mạnh mẽ dựa trên Transformer, hiểu ngữ cảnh hai chiều trong văn bản.

# PHẦN 2 : PHƯƠNG PHÁP THỰC HIỆN

Quy trình thực hiện gồm các bước sau:  
Bước 1 – Thu thập dữ liệu: Sử dụng tập dữ liệu từ Kaggle hoặc GeeksforGeeks.

Bước 2 – Tiền xử lý văn bản: chuẩn hóa, xóa ký tự đặc biệt, loại bỏ stop words, tách từ, đưa về gốc từ.

Bước 3 – Biểu diễn đặc trưng: TF-IDF, Word Embedding hoặc BERT.

Bước 4 – Huấn luyện mô hình: Logistic Regression, Random Forest hoặc BERT.

Bước 5 – Đánh giá mô hình: dùng Accuracy, Precision, Recall, F1-score.

# 2.1) Tiền xử lý văn bản (Preprocessing)

Mục tiêu: Làm cho dữ liệu văn bản “sạch” và “thống nhất” để mô hình hiểu được. Khi lấy dữ liệu từ Kaggle hay các trang báo, văn bản có thể chứa ký tự lạ, dấu câu, từ vô nghĩa, viết hoa, emoji,… → Máy tính không hiểu, nên phải tiền xử lý lại.

Các bước thực hiện:

1. Chuyển toàn bộ văn bản về chữ thường (lowercase)  
 → “Tin Sốc” → “tin sốc”

2. Xóa ký tự đặc biệt, emoji, dấu chấm, dấu hỏi, số, v.v…  
 → loại bỏ ký tự như “🔥”, “!”, “?”, “123”

3. Loại bỏ các từ vô nghĩa (stop words)  
 Là các từ như “là”, “thì”, “và”, “ở”, “của”, “một”… không mang ý nghĩa quan trọng nên xóa đi để giảm nhiễu.

4. Chuẩn hóa từ ngữ (Normalization)  
 Nếu có thể, đưa các từ về dạng gốc: ví dụ “đang chạy”, “chạy”, “chạy bộ” → cùng gốc “chạy”.

Ví dụ:  
“🔥 Tin Sốc: Elon Musk tiết lộ bí mật động trời về công nghệ AI !!!”  
→ Sau tiền xử lý thành:  
“tin sốc elon musk tiết lộ bí mật công nghệ ai”

Dữ liệu huấn luyện được lấy từ 'Kaggle Fake News Dataset' với hàng chục ngàn mẫu tin tức được gán nhãn thật hoặc giả. Sau khi tiền xử lý, văn bản được chuyển sang vector TF-IDF, sau đó huấn luyện bằng Logistic Regression và Random Forest. Cuối cùng, BERT được sử dụng để đạt hiệu quả cao hơn nhờ khả năng hiểu ngữ cảnh.

# 2.2) Biểu diễn văn bản (Vector hóa nội dung)

Sau khi làm sạch, máy vẫn không hiểu chữ, nên ta phải chuyển văn bản thành số (vector). Mỗi bài báo được biến thành một vector đặc trưng chứa các giá trị số thể hiện mức độ quan trọng của từng từ trong bài.

Quá trình này gọi là TF-IDF hoặc Word Embedding, giúp mô hình học được ý nghĩa của văn bản.

## a. Khái niệm

TF-IDF là một phương pháp dùng để đánh giá **độ quan trọng của một từ trong một văn bản** khi so sánh với **toàn bộ tập văn bản**.  
Nó giúp máy tính hiểu văn bản bằng cách gán cho mỗi từ một giá trị thể hiện mức độ “nổi bật” của từ đó.

## b. Cách hiểu đơn giản

* **TF (Term Frequency)**: thể hiện mức độ xuất hiện của từ trong một văn bản.  
  Nếu một từ xuất hiện càng nhiều trong văn bản, nó càng quan trọng trong văn bản đó.
* **IDF (Inverse Document Frequency)**: thể hiện mức độ phổ biến của từ trong toàn bộ tập văn bản.  
  Nếu một từ xuất hiện trong nhiều văn bản khác nhau, nó là từ phổ biến và ít mang ý nghĩa phân biệt.  
  Ngược lại, nếu chỉ xuất hiện trong một vài văn bản, nó có giá trị đặc trưng cao.
* Khi kết hợp hai yếu tố trên, ta có TF-IDF.  
  Kết quả là những từ **vừa xuất hiện nhiều trong một văn bản**, **vừa ít xuất hiện trong các văn bản khác** sẽ được xem là **quan trọng nhất**.

## c. Ví dụ minh họa

Giả sử ta có ba văn bản:

* Văn bản 1: “học máy là lĩnh vực của trí tuệ nhân tạo”
* Văn bản 2: “trí tuệ nhân tạo giúp máy tính học”
* Văn bản 3: “máy tính có thể học từ dữ liệu”

Các từ như “học”, “máy”, “trí tuệ nhân tạo” xuất hiện trong hầu hết các văn bản → chúng phổ biến, nên điểm TF-IDF thấp.  
Các từ như “lĩnh vực” hay “dữ liệu” chỉ xuất hiện ở một văn bản → chúng đặc trưng, nên điểm TF-IDF cao.

|Thành phần | Ý nghĩa |Vai trò

|  |  |  |
| --- | --- | --- |
| TF-IDF | Kết hợp TF và IDF | Đánh giá độ quan trọng tổng hợp của từ trong văn bản |
| Từ có TF-IDF cao | Xuất hiện nhiều trong một văn bản và ít trong các văn bản khác | Từ mang nội dung đặc trưng |
| Từ có TF-IDF thấp | Xuất hiện ở hầu hết các văn bản | Từ phổ biến, ít giá trị phân loại |

**Kết luận:**

Phương pháp TF-IDF giúp xác định mức độ quan trọng của từng từ trong một văn bản so với toàn bộ tập văn bản.  
Nó cho phép máy tính phân biệt được đâu là những từ mang nội dung đặc trưng và đâu là những từ phổ biến, không có giá trị phân loại.  
🡺 Nhờ đó, TF-IDF được sử dụng rộng rãi trong các bài toán xử lý ngôn ngữ tự nhiên như phân loại văn bản, tìm kiếm thông tin và trích xuất đặc trưng nội dung, đặc biệt là giúp con người phân loại được những tin tức và thông tin giả một cách thuyết phục.

# PHẦN 3 : HUẤN LUYỆN MÔ HÌNH PHÂN LOẠI

## -- Một số mô hình phân loại phổ biến

### ****a. Logistic Regression****

Là mô hình thống kê tuyến tính được dùng nhiều trong phân loại nhị phân (hai nhãn).  
Nó hoạt động bằng cách tính xác suất một văn bản thuộc về một lớp nào đó dựa trên trọng số của các đặc trưng.  
Ưu điểm: đơn giản, dễ huấn luyện, hiệu quả với dữ liệu TF-IDF.

### ****b. Random Forest****

Là mô hình dựa trên tập hợp nhiều cây quyết định.  
Mỗi cây học một phần dữ liệu khác nhau, và kết quả cuối cùng là sự tổng hợp của các cây.  
Ưu điểm: độ chính xác cao, giảm nguy cơ quá khớp, hoạt động tốt với dữ liệu có nhiều đặc trưng.

### ****c. BERT (Bidirectional Encoder Representations from Transformers)****

Là mô hình học sâu của Google, được huấn luyện sẵn trên lượng lớn dữ liệu ngôn ngữ.  
BERT hiểu được ngữ cảnh của từ trong cả hai hướng (trước và sau từ đó), nên rất mạnh trong các bài toán hiểu ngữ nghĩa.  
Ưu điểm: cho độ chính xác cao nhất trong các mô hình hiện nay, đặc biệt khi tinh chỉnh (fine-tune) trên dữ liệu cụ thể.

## Kết luận:

Các mô hình như **Logistic Regression**, **Random Forest**, và **BERT** đều có thể được sử dụng để huấn luyện hệ thống phân loại văn bản.

* Với dữ liệu nhỏ hoặc đơn giản, Logistic Regression thường đủ hiệu quả.
* Với dữ liệu lớn và phức tạp, Random Forest hoặc BERT sẽ cho kết quả tốt hơn.  
  Việc lựa chọn mô hình phù hợp phụ thuộc vào quy mô dữ liệu, mục tiêu bài toán và tài nguyên tính toán.

Kết quả mô phỏng cho thấy các mô hình học máy đạt được độ chính xác khác nhau như sau:

Kết quả so sánh mô hình:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mô hình | Accuracy | Precision | Recall | F1-score |
| Logistic Regression | 0.91 | 0.90 | 0.89 | 0.89 |
| Random Forest | 0.93 | 0.92 | 0.91 | 0.91 |
| BERT | 0.96 | 0.95 | 0.95 | 0.95 |

🡺Như vậy, mô hình BERT cho kết quả cao nhất nhờ khả năng hiểu ngữ nghĩa sâu và mối quan hệ giữa các từ trong văn bản.

## PHẦN 4: Đánh giá độ chính xác của mô hình

Sau khi huấn luyện, mô hình phân loại văn bản được **đánh giá độ chính xác** dựa trên các chỉ số thống kê nhằm phản ánh khả năng dự đoán của mô hình. Các chỉ số thường dùng gồm:

* **Precision (Độ chính xác)**: cho biết trong số các văn bản mà mô hình dự đoán thuộc một lớp nào đó, có bao nhiêu văn bản thực sự thuộc lớp đó. Precision cao nghĩa là mô hình ít dự đoán sai dương tính.
* **Recall (Độ bao phủ)**: cho biết trong số các văn bản thực sự thuộc một lớp, mô hình nhận diện được bao nhiêu. Recall cao nghĩa là mô hình ít bỏ sót mẫu đúng.
* **F1-score**: là giá trị trung bình điều hòa giữa Precision và Recall, phản ánh mức cân bằng giữa hai chỉ số này. F1-score càng cao thì mô hình càng tốt.

=>Ba chỉ số trên giúp đánh giá mô hình một cách toàn diện hơn so với chỉ dùng Accuracy, đặc biệt trong các trường hợp dữ liệu không cân bằng giữa các lớp.

**Kết luận:**

Các chỉ số Precision, Recall và F1-score cung cấp cái nhìn chi tiết hơn về hiệu suất của mô hình phân loại so với chỉ số Accuracy thông thường.  
Khi so sánh giữa các mô hình, mô hình nào có các giá trị Precision, Recall và F1-score cao hơn thì mô hình đó thể hiện khả năng dự đoán chính xác và ổn định hơn.  
Trong thí nghiệm, mô hình **BERT** đạt kết quả cao nhất nhờ khả năng hiểu ngữ cảnh và mối quan hệ ngữ nghĩa giữa các từ trong văn bản, cho thấy hiệu quả vượt trội so với các mô hình truyền thống như Logistic Regression hay Random Forest.

# PHẦN 5: ỨNG DỤNG THỰC TẾ

Hệ thống phát hiện tin giả có thể được ứng dụng trong nhiều lĩnh vực như:  
- Truyền Thông: Thông tin của các trang báo điện tử, truyền hình và đưa tin .  
- Mạng xã hội: Cảnh báo người dùng khi chia sẻ và tương tác với các bài viết có dấu hiệu sai lệch.  
- Cơ quan chính phủ: Giám sát thông tin liên quan đến chính trị, kinh tế, bộ phận an ninh mạng của nhà nước.  
- Giáo dục: Giúp sinh viên và người học nhận biết các thông tin không đúng sự thật.

# KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Qua quá trình nghiên cứu và thực nghiệm, nhóm đã tìm hiểu sâu về xử lý ngôn ngữ tự nhiên, các mô hình học máy và đặc biệt là mô hình BERT trong bài toán phân loại tin giả. Kết quả cho thấy BERT đạt độ chính xác cao nhất. Trong tương lai, nhóm mong muốn phát triển hệ thống theo hướng thời gian thực, tự động thu thập và đánh giá tin tức trên mạng xã hội, phân loại và chọn lọc đúng cách những thông tin , trang tin tức một cách mạch lạc để biết thông tin nào là giả và thật.

# BONUS: CÁC TÀI LIỆU THAM KHẢO

1. GeeksforGeeks – Fake News Detection

2. Kaggle Dataset – Fake News

3. <https://www.kaggle.com/c/fake-news>

------------------THANK EVERYONE FOR LISTENING-----------------