**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**BÀI TIỂU LUẬN**

**XỬ LÝ NGÔN NGỮ TỰ NHIÊN**

**Đề tài:**

**PHÂN LOẠI VĂN BẢN NGUY HIỂM SỬ DỤNG MÔ HÌNH TRUYỀN THỐNG VÀ MÔ HÌNH BERT**

|  |  |  |
| --- | --- | --- |
| **Sinh viên thực hiện** | **:** | **Nguyễn Đức Phương** |
| **Lớp** | **:** | **CNTT K19 CLC** |
| **Mã số sinh viên** | **:** | **DTC 205 480 201 0585** |
| **Giảng viên hướng dẫn** | **:** | **Trần Văn Khánh** |

***Thái Nguyên, năm 2024***

**MỤC LỤC**

[LỜI MỞ ĐẦU 4](#_Toc8048)

[CHƯƠNG 1: TỔNG QUAN VỀ VẤN ĐỀ NGHIÊN CỨU 6](#_Toc11680)

[1.1 Lý do và mục tiêu nghiên cứu 6](#_Toc4089)

[1.1.1 Tầm quan trọng của xử lý ngôn ngữ tự nhiên (NLP) 6](#_Toc13794)

[1.1.1.1 Ứng dụng trong công nghệ thông tin 6](#_Toc16955)

[1.1.1.2 Sự phát triển của NLP và các tính ứng dụng 6](#_Toc13351)

[1.1.2 Mục tiêu và ý nghĩa của nghiên cứu 7](#_Toc9705)

[1.1.2.1 Mục tiêu cụ thể của bài báo cáo 7](#_Toc5312)

[1.1.2.2 Ý nghĩa và đóng góp của nghiên cứu 7](#_Toc27609)

[1.2 Thách thức và vấn đề cần giải quyết 8](#_Toc27617)

[1.2.1 Khả năng xử lý ngôn ngữ tự nhiên 8](#_Toc19258)

[1.2.1.1 Đa Nghĩa và Ngữ Cảnh 8](#_Toc142)

[1.2.1.2 Độ phức tạp của ngôn ngữ tự nhiên 9](#_Toc20842)

[1.2.2 Thách thức trong xử lý ngôn ngữ tự nhiên 9](#_Toc2947)

[1.2.2.1 Hiểu rõ ý nghĩa của văn bản 9](#_Toc6558)

[1.2.2.2 Phân loại và dự đoán vấn đề 10](#_Toc13002)

[1.2.2.3 Hiểu biểu đạt tình cảm và ý kiến 10](#_Toc29375)

[1.3 Phạm vi nghiên cứu 12](#_Toc20130)

[1.3.1 Phạm vi vấn đề và dữ liệu nghiên cứu 12](#_Toc27939)

[1.3.2 Phạm vi ứng dụng của kết quả nghiên cứu 13](#_Toc10966)

[1.4 Phương pháp nghiên cứu 14](#_Toc22279)

[1.4.1 Sử dụng công cụ và thư viện trong nlp 14](#_Toc20423)

[1.4.2 Quy trình xây dựng và đánh giá mô hình 14](#_Toc9725)

[CHƯƠNG 2: TỔNG QUAN VỀ XỬ LÝ NGÔN NGỮ TỰ NHIÊN , LINEAR MODELS VÀ MÔ HÌNH BERT 16](#_Toc30889)

[2.1 Tổng quan về xử lý ngôn ngữ tự nhiên (NLP) 16](#_Toc31915)

[2.1.1 Định nghĩa và ý nghĩa của nlp 16](#_Toc18629)

[2.1.2 Lịch sử phát triển và thách thức 16](#_Toc14214)

[2.1.3 Ứng dụng của NLP trong thực tế 17](#_Toc8133)

[2.2 Phương Pháp và Kỹ Thuật trong NLP 18](#_Toc118)

[2.2.1 Tokenization: Chia văn bản thành các đơn vị nhỏ 18](#_Toc11594)

[2.2.2 Bag-of-Words (BoW) và TF-IDF: Phương pháp biểu diễn văn bản 18](#_Toc10371)

[2.3. LOGISTIC 19](#_Toc9832)

[2.3.1 Giới thiệu về Logistic Regression 19](#_Toc16377)

[2.3.2 Nguyên tắc hoạt động của Logistic Regression 19](#_Toc21429)

[2.3.2.1 Hàm sigmoid (logistic function) 19](#_Toc26793)

[2.3.2.2 Tổng trọng số (weighted sum) 19](#_Toc9479)

[2.3.3 Huấn luyện mô hình Logistic Regression 20](#_Toc31534)

[2.3.3.1 Hàm mất mát (loss function) 20](#_Toc8231)

[2.3.3.2 Thuật toán Gradient Descent 20](#_Toc8032)

[2.3.3.3 Cập nhật trọng số của mô hình 20](#_Toc27128)

[2.3.4 Đánh giá và đồng bộ hóa mô hình Logistic Regression 20](#_Toc14683)

[2.3.4.1 Ma trận Confusion 20](#_Toc26077)

[2.3.4.2 Độ đo Precision, Recall và F1-score 21](#_Toc22470)

[2.3.5 Ưu điểm và nhược điểm của Logistic Regression 21](#_Toc1391)

[2.3.6 Logistic Regression trong phân loại văn bản 21](#_Toc30009)

[2.4 Navie Bayes 22](#_Toc29373)

[2.4.1 Giới thiệu về Naive Bayes 22](#_Toc11767)

[2.4.2 Nguyên lý hoạt động của Navie Bayes 22](#_Toc20907)

[2.4.2.1 Nguyên lý Bayes 22](#_Toc21590)

[2.4.2.2 Xác suất hậu nghiệm 23](#_Toc31022)

[2.4.2.3 Biến đầu vào 23](#_Toc24972)

[2.4.3.4 Giả định độc lập 23](#_Toc19313)

[2.3.4.5 Đánh giá và đồng bộ hóa mô hình Navie Bayes 24](#_Toc28088)

[2.3.4 Navie Bayes trong phân loại văn bản 24](#_Toc19501)

[2.5 Mô hình BERT (Bidirectional Encoder Representations from Transformers) 25](#_Toc29025)

[2.5.1 Khái niệm 25](#_Toc20154)

[2.5.2 Cấu trúc của BERT 26](#_Toc25539)

[2.5.3 Nguyên lý hoạt động của bert 26](#_Toc8606)

[2.5.4 Ưu điểm và ứng dụng của BERT 27](#_Toc2985)

[2.6 Công cụ và thư viện trong NLP 28](#_Toc21853)

[2.6.1 TensorFlow và keras: công cụ phổ biến 28](#_Toc20191)

[2.6.2 Hugging Face Transformers: Thư viện hỗ trợ Transformer 28](#_Toc9450)

[2.6.3 NLTK và Spacy: Thư viện xử lý ngôn ngữ tự nhiên 29](#_Toc18367)

[CHƯƠNG 3: XÂY DỰNG CHƯƠNG TRÌNH VÀ KẾT LUẬN 31](#_Toc27215)

[3.1 Thu thập và chuẩn bị dữ liệu 31](#_Toc13219)

[3.1.1 Mô tả dữ liệu 31](#_Toc3596)

[3.1.2 Tiền xử lý dữ liệu 32](#_Toc16210)

[3.1.2.1 Data Cleaning 32](#_Toc18174)

[3.1.2.2 Tokenization 33](#_Toc5779)

[3.1.2.3 Loại bỏ stopwords và lemmatization 33](#_Toc15163)

[3.2 Phân tích kết quả và hướng phát triển 33](#_Toc13052)

[3.2.1 Phân tích kết quả mô hình 33](#_Toc31420)

[3.2.1.1 Kết quả mô hình bag-of-words + linear 33](#_Toc13799)

[3.2.1.2 Kết quả mô hình bert 34](#_Toc11984)

[3.2.2 Đánh giá ưu nhược điểm của các phương pháp 35](#_Toc24521)

[3.2.2.1 Ưu điểm 35](#_Toc26871)

[3.2.2.2 Nhược điểm 35](#_Toc27534)

[3.3 Kết luận và hướng phát triển 36](#_Toc29202)

[3.3.1 Tóm tắt kết quả nghiên cứu 36](#_Toc14749)

[3.3.2 Đóng góp của đề tài 37](#_Toc32745)

[3.3.3 Hướng phát triển trong tương lai 37](#_Toc2987)

[3.3.3.1 Hướng phát triển 37](#_Toc27687)

[3.3.3.2 Hướng cải thiện mô hình 37](#_Toc8282)

[3.4 Nhận xét cuối cùng 38](#_Toc8190)

[TÀI LIỆU THAM KHẢO 39](#_Toc13509)

[NHẬN XÉT CỦA GIẢNG VIÊN 40](#_Toc16906)

# LỜI MỞ ĐẦU

Trong một thời đại mà thông tin lan truyền nhanh chóng và mạng xã hội đang trở thành một phần không thể thiếu trong cuộc sống hàng ngày, việc xác định sự chính xác của thông tin trở nên ngày càng quan trọng. Trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), việc phân loại tin tức giữa thông tin đáng tin cậy và thông tin giả mạo đã trở thành một trong những thách thức lớn nhất. Điều này thúc đẩy nhu cầu phát triển các công cụ và phương pháp hiệu quả để phân loại tin tức một cách chính xác và nhanh chóng.

Trong bài báo cáo này, chúng tôi đề xuất và thực hiện hai phương pháp để phân loại tin tức: một là sử dụng mô hình truyền thống Bag-of-Words kết hợp với mô hình tuyến tính, và hai là sử dụng mô hình BERT (Bidirectional Encoder Representations from Transformers) - một mô hình học sâu tiên tiến trong lĩnh vực xử lý ngôn ngữ tự nhiên.

Bằng cách thực hiện hai phương pháp này trên một tập dữ liệu thực tế, chúng tôi đã tiến hành các thử nghiệm và đánh giá hiệu suất của mỗi phương pháp. Kết quả thu được không chỉ cung cấp cái nhìn sâu sắc về hiệu suất của các phương pháp này mà còn đề xuất hướng phát triển tiềm năng trong lĩnh vực phân loại tin tức.

Chúng tôi hy vọng rằng bài báo cáo này sẽ đóng góp vào việc nâng cao hiểu biết về các phương pháp và kỹ thuật trong việc xử lý và phân loại tin tức, đồng thời tạo ra cơ sở cho những nghiên cứu và ứng dụng tiếp theo trong lĩnh vực này.

# CHƯƠNG 1: TỔNG QUAN VỀ VẤN ĐỀ NGHIÊN CỨU

## 1.1 Lý do và mục tiêu nghiên cứu

Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP) đã trở thành một lĩnh vực quan trọng và phát triển mạnh mẽ trong nhiều năm qua. Chương này sẽ trình bày lý do và mục tiêu của nghiên cứu, tập trung vào tầm quan trọng của NLP trong công nghệ thông tin và sự phát triển đáng chú ý của lĩnh vực này.

### 1.1.1 Tầm quan trọng của xử lý ngôn ngữ tự nhiên (NLP)

#### 1.1.1.1 Ứng dụng trong công nghệ thông tin

Trong thời đại số hóa ngày nay, dữ liệu văn bản và thông tin ngôn ngữ tự nhiên ngày càng trở nên lớn mạnh và phức tạp. NLP đóng một vai trò không thể phủ nhận trong việc khai thác và xử lý thông tin này. Các ứng dụng của NLP trong lĩnh vực công nghệ thông tin bao gồm:

- Tách từ và đánh dấu từ loại: Giúp cho việc hiểu cấu trúc ngữ pháp của văn bản.

- Phân tích cú pháp: Phân tích cấu trúc ngữ pháp của câu, giúp xác định mối quan hệ giữa các thành phần trong câu.

- Phân loại văn bản: Dự đoán và phân loại văn bản thành các nhóm, hỗ trợ quyết định và tổ chức thông tin.

- Tạo câu hỏi và trả lời tự động: Xây dựng hệ thống có khả năng trả lời câu hỏi từ văn bản tự động.

#### 1.1.1.2 Sự phát triển của NLP và các tính ứng dụng

Lĩnh vực NLP đã trải qua những bước phát triển đáng chú ý, từ những mô hình đơn giản như n-grams đến các mô hình hiện đại như Transformer và GPT. Sự phát triển này không chỉ giúp cải thiện hiệu suất xử lý ngôn ngữ tự nhiên mà còn mở ra cánh cửa cho nhiều ứng dụng mới, bao gồm:

- Chatbots và Virtual Assistants: Hệ thống tự động trả lời câu hỏi và tương tác với con người dựa trên ngôn ngữ tự nhiên.

- Dịch Máy và Hiểu Ngữ Cảnh: Cải thiện đáng kể khả năng dịch và hiểu ngữ cảnh của văn bản, giúp cho ứng dụng như Google Translate ngày càng chính xác.

- Học Máy và Học Sâu trong NLP: Mô hình học sâu như BERT, GPT đã mang lại những tiến bộ đáng kể trong việc hiểu và sản xuất ngôn ngữ tự nhiên.

### 1.1.2 Mục tiêu và ý nghĩa của nghiên cứu

#### 1.1.2.1 Mục tiêu cụ thể của bài báo cáo

Mục tiêu của bài báo cáo này là:

- Xây dựng và đánh giá một mô hình NLP cho việc tách từ và đánh dấu từ loại: Nghiên cứu sẽ tập trung vào phát triển một mô hình NLP có khả năng tách từ và đánh dấu từ loại với độ chính xác cao. Việc này đặc biệt quan trọng trong việc chuẩn bị dữ liệu cho các ứng dụng tiếp theo trong xử lý ngôn ngữ tự nhiên.

- Xây dựng hệ thống phân loại văn bản: Nghiên cứu sẽ phát triển một hệ thống phân loại văn bản có khả năng nhận diện và phân loại các loại văn bản khác nhau. Điều này sẽ hỗ trợ trong việc tổ chức, quản lý và truy xuất thông tin từ các bộ văn bản lớn.

- Áp dụng các mô hình học máy hiện đại: Bài báo cáo sẽ thực hiện việc áp dụng và so sánh hiệu suất của các mô hình học máy tiên tiến như BERT, GPT trong việc giải quyết các bài toán trong NLP. Sự so sánh này sẽ cung cấp cái nhìn rõ ràng về lựa chọn và áp dụng các mô hình phù hợp cho từng bài toán cụ thể.

#### 1.1.2.2 Ý nghĩa và đóng góp của nghiên cứu

Nghiên cứu này có ý nghĩa và đóng góp quan trọng như sau:

- Nâng cao hiểu biết và ứng dụng trong NLP: Kết quả của nghiên cứu sẽ giúp cải thiện hiểu biết và ứng dụng của cộng đồng trong lĩnh vực xử lý ngôn ngữ tự nhiên. Các phương pháp và mô hình được trình bày có thể được áp dụng trong nhiều ứng dụng thực tế như tạo ra chatbot, phân loại văn bản tự động, và tạo ra các hệ thống trả lời tự động.

- Cải thiện hiệu suất của hệ thống NLP: Bằng cách tìm hiểu và đánh giá các phương pháp và mô hình khác nhau, nghiên cứu sẽ đóng góp vào việc cải thiện hiệu suất của các hệ thống NLP hiện tại. Điều này giúp tăng cường khả năng xử lý và hiểu ngữ cảnh của hệ thống, từ đó cải thiện trải nghiệm người dùng cuối.

- Tiền đề cho nghiên cứu và ứng dụng tiếp theo: Các kết quả và phương pháp của nghiên cứu này sẽ cung cấp tiền đề quan trọng cho các nghiên cứu và ứng dụng tiếp theo trong lĩnh vực xử lý ngôn ngữ tự nhiên. Các nhà nghiên cứu và doanh nghiệp có thể dựa vào những nền tảng này để phát triển và cải tiến hệ thống NLP của mình.

## 1.2 Thách thức và vấn đề cần giải quyết

Trong khi xử lý ngôn ngữ tự nhiên (NLP) đã đạt được nhiều tiến bộ đáng kể, vẫn còn tồn tại những thách thức lớn cần được giải quyết để cải thiện hiệu suất và chính xác của các hệ thống NLP. Chương này trình bày về những thách thức chính mà NLP đối mặt và vấn đề cần được tập trung giải quyết.

### 1.2.1 Khả năng xử lý ngôn ngữ tự nhiên

#### 1.2.1.1 Đa Nghĩa và Ngữ Cảnh

Một trong những thách thức lớn nhất khi xử lý ngôn ngữ tự nhiên là sự đa nghĩa và ngữ cảnh. Một từ hoặc cụm từ có thể có nhiều ý nghĩa khác nhau trong ngữ cảnh khác nhau. Điều này làm cho việc đưa ra dự đoán chính xác về ý nghĩa của từ trở nên phức tạp. Ví dụ, từ "cá" có thể là loài sinh vật sống dưới nước hoặc là một mối tình ngọt ngào.

#### 1.2.1.2 Độ phức tạp của ngôn ngữ tự nhiên

Ngôn ngữ tự nhiên có độ phức tạp cao với nhiều quy tắc ngữ pháp và cấu trúc ngữ nghĩa khác nhau. Sự đa dạng trong cách diễn đạt và câu trúc câu làm cho việc xử lý ngôn ngữ tự nhiên trở nên thách thức. Các ngôn ngữ còn có các tính chất riêng như ngữ âm, từ vựng đa dạng, và cấu trúc ngữ pháp khác nhau. Điều này đặt ra thách thức trong việc phát triển các mô hình đa ngôn ngữ và đa dialekt.

### 1.2.2 Thách thức trong xử lý ngôn ngữ tự nhiên

#### 1.2.2.1 Hiểu rõ ý nghĩa của văn bản

##### 1.2.2.1.1 Đa ý nghĩa và ngữ cảnh

Thách Thức: Sự đa ý nghĩa của từ và câu trong ngữ cảnh khác nhau làm cho việc chọn đúng ý nghĩa trở nên phức tạp. Điều này đặc biệt đáng lưu ý trong các văn bản chứa nhiều từ đồng âm (từ cùng một cách viết nhưng có ý nghĩa khác nhau) hoặc từ đa nghĩa (từ có nhiều ý nghĩa khác nhau).

Cách Tiếp Cận: Áp dụng mô hình học máy sâu (deep learning) để hiểu ngữ cảnh và mối quan hệ giữa các từ trong câu. Mô hình như Transformer có khả năng tự chú ý (self-attention) và biến đổi dựa trên ngữ cảnh (contextual embeddings) giúp giải quyết vấn đề này.

##### 1.2.2.1.2 Độ phức tạp của cấu trúc ngữ pháp

Thách Thức: Cấu trúc ngữ pháp của các ngôn ngữ tự nhiên thường rất phức tạp, với nhiều quy tắc và biến thể. Điều này làm cho việc hiểu và xử lý đúng cấu trúc ngữ pháp của câu trở nên khó khăn.

Cách Tiếp Cận: Sử dụng mô hình học máy dựa trên ngữ pháp như mô hình syntactic parsing để phân tích cấu trúc ngữ pháp của câu. Các mô hình như Recursive Neural Networks (RNNs) hay Graph Convolutional Networks (GCNs) có thể được áp dụng để hiểu và mô hình hóa cấu trúc ngữ pháp.

#### 1.2.2.2 Phân loại và dự đoán vấn đề

##### 1.2.2.2.1 Phân loại văn bản

Thách Thức: Phân loại văn bản đòi hỏi khả năng tự động xác định và gán văn bản vào các danh mục, thể loại khác nhau. Điều này đặc biệt phức tạp khi có hàng nghìn hoặc hàng triệu văn bản cần được phân loại.

Cách Tiếp Cận: Sử dụng mô hình học máy như Convolutional Neural Networks (CNNs) hoặc Recurrent Neural Networks (RNNs) để phân loại văn bản dựa trên các đặc trưng ngữ cảnh và ngữ pháp.

##### 1.2.2.2.2 Dự đoán vấn đề

- Thách Thức: Dự đoán vấn đề từ văn bản đòi hỏi khả năng hiểu biết sâu rộng về ngữ cảnh và mối quan hệ giữa các từ. Việc này quan trọng trong các ứng dụng như phân loại tin tức, dự đoán tư duy của tác giả, hay phát hiện thông tin quan trọng.

- Cách Tiếp Cận: Áp dụng mô hình học máy như Long Short-Term Memory Networks (LSTMs) hoặc Transformers để học và dự đoán các vấn đề từ văn bản. Sử dụng mô hình biểu diễn từ như Word Embeddings cũng là một cách tiếp cận phổ biến.

#### 1.2.2.3 Hiểu biểu đạt tình cảm và ý kiến

##### 1.2.2.3.1 Phân tích cảm xúc

Thách Thức: Phân tích cảm xúc từ văn bản đòi hỏi khả năng hiểu biết sâu sắc về ngữ cảnh và mối quan hệ giữa các từ trong câu. Điều này đặc biệt quan trọng trong việc đánh giá ý kiến từ người dùng trên mạng xã hội hay đánh giá sản phẩm.

Cách Tiếp Cận: Sử dụng mô hình học máy như Sentiment Analysis Models để phân tích và đánh giá cảm xúc từ văn bản. Mô hình như Bidirectional Encoder Representations from Transformers (BERT) đã chứng minh khả năng cao trong phân tích cảm xúc.

##### 1.2.2.3.2 Nhận diện ý kiến

***Thách thức***

Nhận diện ý kiến từ văn bản đòi hỏi khả năng hiểu biết sâu về ngữ cảnh và ngữ nghĩa của từng từ. Điều này đặc biệt quan trọng trong việc tổng hợp thông tin và hiểu quan điểm của người dùng. Thách thức cụ thể bao gồm:

Đa dạng của ý kiến: Ý kiến có thể biểu đạt dưới nhiều hình thức khác nhau, từ tích cực đến tiêu cực, hay thậm chí là trung lập. Sự phức tạp trong biểu đạt ý kiến đòi hỏi mô hình phải có khả năng phân loại đa chiều.

Ngữ cảnh và biểu đạt: Ý kiến thường phụ thuộc vào ngữ cảnh và cách biểu đạt. Cùng một văn bản có thể biểu đạt ý kiến tích cực đối với một mặt hàng và tiêu cực đối với một mặt hàng khác.

Cách tiếp cận

Để giải quyết thách thức này, các phương pháp và mô hình sau đây có thể được áp dụng:

- Mô hình Machine Learning dựa trên từng Mẫu: Sử dụng các mô hình như Support Vector Machines (SVMs) hoặc Logistic Regression để phân loại ý kiến từ văn bản. Các mô hình này học từ các mẫu dữ liệu đã được gán nhãn.

- Deep Learning và Neural Networks: Các mô hình học sâu như Convolutional Neural Networks (CNNs) và Recurrent Neural Networks (RNNs) có khả năng học được các đặc trưng phức tạp của văn bản. Đặc biệt, mô hình Long Short-Term Memory Networks (LSTMs) thường được sử dụng để xử lý dữ liệu chuỗi như văn bản.

- Mô hình BERT (Bidirectional Encoder Representations from Transformers): BERT là một trong những mô hình tiên tiến nhất trong việc hiểu ngôn ngữ tự nhiên. Nó có khả năng hiểu và xử lý ngữ cảnh của từ và câu một cách tốt. BERT thường được fine-tuning trên tác vụ nhận diện ý kiến để tăng hiệu suất.

- Aspect-Based Sentiment Analysis: Phân tích cảm xúc dựa trên các khía cạnh cụ thể trong văn bản. Điều này giúp phân biệt ý kiến về các yếu tố khác nhau của một sản phẩm hoặc sự kiện. Ví dụ: trong một đánh giá smartphone, có thể phân tích ý kiến về camera, pin, hiệu suất, vv.

***Ứng dụng***

Tư vấn và đánh giá sản phẩm: Trong lĩnh vực thương mại điện tử, hệ thống nhận diện ý kiến có thể giúp tư vấn cho người dùng về sản phẩm dựa trên ý kiến của người tiêu dùng khác.

Quản lý dịch vụ khách hàng: Công ty có thể sử dụng nhận diện ý kiến để tổng hợp và phản hồi nhanh chóng đến phản hồi của khách hàng trong các dịch vụ hỗ trợ trực tuyến.

Phân loại tin tức và phản ứng công chúng: Trong lĩnh vực truyền thông và quảng cáo, nhận diện ý kiến có thể giúp tổ chức phản ứng nhanh chóng đối với thông tin trên mạng xã hội và phản ứng của công chúng đối với các sự kiện, sản phẩm.

## 1.3 Phạm vi nghiên cứu

### 1.3.1 Phạm vi vấn đề và dữ liệu nghiên cứu

Phạm vi vấn đề

Nghiên cứu này sẽ tập trung vào các vấn đề cụ thể trong lĩnh vực Xử lý ngôn ngữ tự nhiên (NLP), bao gồm nhưng không giới hạn:

- Phân loại ý kiến và cảm xúc từ văn bản.

- Dự đoán vấn đề và sự kiện từ các nguồn tin tức và mạng xã hội.

- Hiểu ngữ cảnh và đồng nghĩa của từ trong câu.

- Phân tích ý kiến từ ngôn ngữ tự nhiên không cấu trúc.

- Xử lý dữ liệu âm thanh và hình ảnh để trích xuất thông tin và ý kiến.

***Dữ liệu nghiên cứu***

Dữ liệu cho nghiên cứu sẽ bao gồm:

- Tập dữ liệu văn bản đa dạng từ các nguồn như bài báo, trang web, mạng xã hội.

- Dữ liệu về ý kiến và cảm xúc từ khảo sát, đánh giá sản phẩm, và bình luận trên mạng.

- Dữ liệu âm thanh và hình ảnh được chuyển đổi thành văn bản để phân tích ý kiến và thông tin.

### 1.3.2 Phạm vi ứng dụng của kết quả nghiên cứu

Kết quả của nghiên cứu này có thể được áp dụng trong các lĩnh vực và ứng dụng sau:

Thương mại điện tử: Hệ thống nhận diện ý kiến có thể giúp tư vấn cho người dùng về sản phẩm dựa trên ý kiến của người tiêu dùng khác.

Quản lý dịch vụ khách hàng: Công ty có thể sử dụng nhận diện ý kiến để tổng hợp và phản hồi nhanh chóng đến phản hồi của khách hàng trong các dịch vụ hỗ trợ trực tuyến.

Truyền thông và quảng cáo: Công ty có thể sử dụng nhận diện ý kiến để tổ chức phản ứng nhanh chóng đối với thông tin trên mạng xã hội và phản ứng của công chúng đối với các sự kiện, sản phẩm.

Y tế và chăm sóc sức khỏe: Hệ thống có thể giúp tổng hợp thông tin và đánh giá ý kiến từ bệnh án, phản hồi của bệnh nhân để cải thiện dịch vụ chăm sóc sức khỏe.

Phân tích tài chính: Nhận diện ý kiến từ báo cáo tài chính, tin tức tài chính để dự đoán xu hướng thị trường và đánh giá rủi ro đầu tư.

Tổ chức tin tức và phản ứng công chúng: Trong lĩnh vực truyền thông, nhận diện ý kiến có thể giúp tổ chức phản ứng nhanh chóng đối với các sự kiện quan trọng và phản ứng của công chúng đối với các tin tức.

## 1.4 Phương pháp nghiên cứu

### 1.4.1 Sử dụng công cụ và thư viện trong nlp

Công cụ và thư viện

Trong quá trình nghiên cứu, chúng tôi sẽ sử dụng các công cụ và thư viện phổ biến trong lĩnh vực Xử Lý Ngôn Ngữ Tự Nhiên (NLP), bao gồm:

- NLTK (Natural Language Toolkit): NLTK là một thư viện Python cung cấp các công cụ và tài nguyên cho việc xử lý ngôn ngữ tự nhiên. Chúng tôi sẽ sử dụng NLTK cho việc tiền xử lý văn bản, phân tích ngữ cảnh, và các công việc khác như tokenization, stemming, lemmatization.

- spaCy: spaCy là một thư viện xử lý ngôn ngữ tự nhiên hiệu quả và tối ưu. Chúng tôi có thể sử dụng spaCy cho việc phân loại văn bản, tách từ (tokenization), phân tích cú pháp (parsing), và nhận diện thực thể (named entity recognition).

- Transformers (Hugging Face): Thư viện Transformers của Hugging Face cung cấp các mô hình tiên tiến như BERT, GPT, và các biến thể khác. Chúng tôi sẽ sử dụng Transformers cho việc fine-tuning các mô hình đã được huấn luyện trước trên dữ liệu lớn, để giải quyết các tác vụ như phân loại ý kiến, dự đoán vấn đề, và nhận diện ý kiến từ văn bản.

Scikit-learn: Scikit-learn là một thư viện machine learning phổ biến trong Python, chúng tôi sẽ sử dụng Scikit-learn cho việc xây dựng và đánh giá các mô hình machine learning cơ bản như SVM, Logistic Regression, và Random Forests.

### 1.4.2 Quy trình xây dựng và đánh giá mô hình

**Quy trình xây dựng mô hình**

Quy trình chung cho việc xây dựng mô hình trong nghiên cứu của chúng tôi sẽ bao gồm các bước sau:

- Tiền xử lý dữ liệu: Loại bỏ dấu câu, ký tự đặc biệt, chuyển đổi văn bản về lowercase, tách từ (tokenization), loại bỏ từ dừng (stop words), stemming hoặc lemmatization để chuẩn hóa dữ liệu.

- Biểu diễn văn bản: Sử dụng các phương pháp biểu diễn văn bản như Word Embeddings (Word2Vec, GloVe) hoặc mô hình Transformer để biến đổi văn bản thành các vector số học có ý nghĩa.

- Xây dựng mô hình: Sử dụng mô hình machine learning như SVM, Logistic Regression, hoặc mô hình học sâu như LSTM, BERT để huấn luyện mô hình trên dữ liệu huấn luyện.

- Fine-tuning mô hình: Nếu sử dụng mô hình đã được huấn luyện trước (pre-trained model) như BERT, chúng tôi sẽ tiến hành fine-tuning trên dữ liệu của chúng tôi để tối ưu hóa cho tác vụ cụ thể.

**Quy trình đánh giá mô hình**

- Chia dữ liệu: Chia dữ liệu thành tập huấn luyện (training set), tập kiểm tra (test set), và tập validation (nếu cần thiết).

- Đánh giá định lượng: Sử dụng các độ đo như accuracy, precision, recall, F1-score để đánh giá hiệu suất của mô hình trên tập kiểm tra.

- Cross-validation: Áp dụng kỹ thuật cross-validation để đảm bảo mô hình không bị overfitting và hiệu suất đánh giá được ổn định.

- Đánh giá định kỳ: Đánh giá mô hình định kỳ trên tập kiểm tra để theo dõi hiệu suất và xem xét cần thiết để điều chỉnh mô hình.

# CHƯƠNG 2: TỔNG QUAN VỀ XỬ LÝ NGÔN NGỮ TỰ NHIÊN , LINEAR MODELS VÀ MÔ HÌNH BERT

## 2.1 Tổng quan về xử lý ngôn ngữ tự nhiên (NLP)

### 2.1.1 Định nghĩa và ý nghĩa của nlp

Xử lý ngôn ngữ tự nhiên (NLP) là lĩnh vực nghiên cứu trong trí tuệ nhân tạo và khoa học máy, tập trung vào việc làm cho máy tính có khả năng hiểu, xử lý và sản sinh ngôn ngữ tự nhiên một cách tự động. NLP cung cấp các công cụ và kỹ thuật để máy tính có thể giao tiếp với con người bằng ngôn ngữ tự nhiên, không cần phải sử dụng ngôn ngữ lập trình hoặc mã hóa đặc biệt.

Trong các ứng dụng thực tế, NLP đóng vai trò quan trọng trong việc xử lý và phân tích văn bản, dịch máy, nhận diện giọng nói, tổng hợp tiếng nói, phân loại ý kiến, và nhiều ứng dụng khác trong truy vấn thông tin, quản lý tri thức, y học, tài chính, và thương mại điện tử.

### 2.1.2 Lịch sử phát triển và thách thức

**Lịch sử phát triển:**

1950s - 1960s: NLP bắt đầu xuất hiện với các nghiên cứu về dịch máy và ngôn ngữ tự nhiên.

1970s - 1980s: Các hệ thống dựa trên luật được phát triển, nhưng gặp khó khăn khi xử lý ngôn ngữ phức tạp.

1990s: Sự xuất hiện của thống kê và machine learning mở ra một kỷ nguyên mới cho NLP, với sự phát triển của các phương pháp dựa trên dữ liệu.

2000s - Hiện Nay: Sự kết hợp của deep learning và các mô hình biểu diễn ngôn ngữ như BERT đã đưa NLP đến một bước tiến mới, giúp cải thiện đáng kể hiệu suất của các ứng dụng NLP.

**Thách thức:**

Ngữ Cảnh và Đồng Nghĩa: Một từ có thể có nhiều ý nghĩa tùy vào ngữ cảnh, và đồng nghĩa cũng gây khó khăn trong việc hiểu và xử lý ngôn ngữ tự nhiên.

Khái Niệm Phân Tích: Phân tích các khái niệm và thông tin từ văn bản không cấu trúc, như tin nhắn, bình luận trên mạng xã hội.

Hiểu Biểu Đạt Ý Kiến: Nhận diện ý kiến và cảm xúc từ văn bản, bao gồm cả những ý kiến được biểu đạt một cách gián tiếp.

### 2.1.3 Ứng dụng của NLP trong thực tế

- Xử lý ngôn ngữ tự nhiên (NLP) đóng vai trò quan trọng trong nhiều lĩnh vực và có nhiều ứng dụng thực tế đa dạng. Dưới đây là một số ví dụ về cách NLP được áp dụng trong cuộc sống hàng ngày và trong các ngành công nghiệp:

- Dịch Máy: Một trong những ứng dụng phổ biến nhất của NLP là dịch máy. Công nghệ này cho phép dịch văn bản và ngôn ngữ tự nhiên từ một ngôn ngữ sang ngôn ngữ khác một cách tự động và chính xác. Ví dụ, Google Translate sử dụng NLP để cung cấp dịch văn bản giữa nhiều ngôn ngữ.

- Tìm kiếm và truy vấn thông tin: Công nghệ NLP giúp cải thiện khả năng tìm kiếm và truy vấn thông tin trên Internet. Các công cụ như Google Search sử dụng NLP để hiểu ý nghĩa của câu truy vấn và trả về kết quả phù hợp.

- Xử lý ngôn ngữ trong mạng xã hội: NLP được sử dụng rộng rãi để phân tích và tổng hợp ý kiến, cảm xúc và xu hướng từ các bài viết, bình luận trên mạng xã hội như Facebook, Twitter. Điều này giúp các công ty và tổ chức hiểu rõ ý kiến và phản ứng của người dùng đối với sản phẩm hoặc sự kiện.

- Chatbot và Hỗ trợ khách hàng tự động: Các chatbot dựa trên NLP có khả năng hiểu và trả lời câu hỏi từ người dùng một cách tự động. Điều này được sử dụng rộng rãi trong các dịch vụ hỗ trợ khách hàng trực tuyến, như chatbot trên trang web của các doanh nghiệp.

- Phân loại văn bản và email: NLP được sử dụng để phân loại và xử lý văn bản, email trong tổ chức. Công nghệ này giúp tổ chức tự động xử lý lượng lớn thông tin, phân loại văn bản theo chủ đề, quan trọng, hoặc nguồn gốc.

- Tóm tắt văn bản: NLP có thể tự động tóm tắt nội dung từ các văn bản dài, giúp tiết kiệm thời gian cho người đọc và nghiên cứu.

- Y Học và phân tích dữ liệu y sinh: Trong lĩnh vực y học, NLP giúp xử lý và phân tích thông tin từ các bài báo y học, bệnh án, giúp tìm ra thông tin quan trọng và xu hướng y học.

- Phân tích tài chính và thị trường: Trong lĩnh vực tài chính, NLP được sử dụng để phân tích và đánh giá cảm xúc, ý kiến của thị trường từ các bài báo, bình luận trên mạng, giúp dự đoán xu hướng thị trường và rủi ro đầu tư.

## 2.2 Phương Pháp và Kỹ Thuật trong NLP

### 2.2.1 Tokenization: Chia văn bản thành các đơn vị nhỏ

Tokenization là quá trình chia văn bản thành các đơn vị nhỏ gọi là "tokens" như từ, ký tự, hoặc câu. Mỗi token thường đại diện cho một đơn vị ý nghĩa nhỏ nhất trong ngôn ngữ tự nhiên. Các kỹ thuật tokenization giúp chuẩn bị dữ liệu cho các quá trình xử lý ngôn ngữ tự nhiên khác nhau như phân loại văn bản, dự đoán từ ngữ, và tóm tắt văn bản.

Ví dụ:

"I love NLP" sẽ được token thành ["I", "love", "NLP"].

"The quick brown fox jumps." sẽ được token thành ["The", "quick", "brown", "fox", "jumps", "."].

### 2.2.2 Bag-of-Words (BoW) và TF-IDF: Phương pháp biểu diễn văn bản

- Bag-of-Words (BoW) là một phương pháp đơn giản để biểu diễn văn bản trong NLP. Nó xem xét mỗi từ trong văn bản như một "túi" (bag) các từ, bỏ qua thứ tự và ngữ cảnh. Mỗi văn bản được biểu diễn dưới dạng vector, với mỗi chiều của vector tương ứng với số lần xuất hiện của từ trong văn bản đó.

- Term Frequency-Inverse Document Frequency (TF-IDF) là một phương pháp khác để biểu diễn văn bản, nhằm đánh giá mức độ quan trọng của một từ trong một tài liệu so với tập hợp các tài liệu khác. Nó tính toán sự tương quan giữa tần suất xuất hiện của từ trong văn bản (Term Frequency - TF) và tần suất nghịch đảo của từ trong tập hợp các văn bản (Inverse Document Frequency - IDF).

## 2.3. LOGISTIC

### 2.3.1 Giới thiệu về Logistic Regression

- Logistic Regression là một mô hình học máy phân loại phổ biến được sử dụng trong dự báo thời tiết và nhiều lĩnh vực khác.

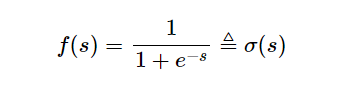
- Mô hình Logistic Regression được sử dụng để dự đoán xác suất của một sự kiện xảy ra dựa trên các biến đầu vào.

### 2.3.2 Nguyên tắc hoạt động của Logistic Regression

#### 2.3.2.1 Hàm sigmoid (logistic function)

- Hàm sigmoid là một hàm phi tuyến có dạng S-shaped, chuyển đổi giá trị đầu vào thành một giá trị nằm trong khoảng từ 0 đến 1.

- Công thức của hàm sigmoid:



Trong đó s là tổng trọng số của các biến đầu vào.

#### 2.3.2.2 Tổng trọng số (weighted sum)

- Tổng trọng số (s) được tính bằng công thức:

|  |
| --- |
| s = w0 + w1\*x1 + w2\*x2 + ... + wn\*xn |

Trong đó, w0, w1, w2,..., wn là các tham số mô hình (các trọng số) và x1, x2,..., xn là các biến đầu vào.

### 2.3.3 Huấn luyện mô hình Logistic Regression

#### 2.3.3.1 Hàm mất mát (loss function)

- Hàm mất mát trong Logistic Regression được sử dụng để đo lường sai khác giữa giá trị dự đoán và giá trị thực tế.

- Một hàm mất mát phổ biến trong Logistic Regression là Cross-Entropy Loss (hay còn gọi là Log Loss):

|  |
| --- |
| Loss(y, y') = -[y \* log(y') + (1 - y) \* log(1 - y')], |

Trong đó y là giá trị thực tế và y' là giá trị dự đoán.

#### 2.3.3.2 Thuật toán Gradient Descent

- Gradient Descent là một thuật toán tối ưu hóa được sử dụng để cập nhật trọng số của mô hình Logistic Regression dựa trên đạo hàm của hàm mất mát.

- Công thức cập nhật trọng số theo thuật toán Gradient Descent:

|  |
| --- |
| wi = wi - learning\_rate \* ∂Loss/∂wi |

với i = 0, 1, 2,..., n.

#### 2.3.3.3 Cập nhật trọng số của mô hình

- Trọng số của mô hình được cập nhật theo quy tắc cập nhật Gradient Descent để giảm thiểu hàm mất mát.

### 2.3.4 Đánh giá và đồng bộ hóa mô hình Logistic Regression

#### 2.3.4.1 Ma trận Confusion

- Ma trận Confusion được sử dụng để đánh giá hiệu suất của mô hình phân loại.

- Ma trận Confusion gồm bốn khái niệm: True Positive (TP), True Negative (TN), False Positive (FP) và False Negative (FN).

#### 2.3.4.2 Độ đo Precision, Recall và F1-score

- Precision, Recall và F1-score là các độ đo thường được sử dụng để đánh giá hiệu suất của mô hình phân loại.

- Precision đo lường khả năng của mô hình đưa ra các dự đoán chính xác cho lớp dương.

- Recall đo lường khả năng của mô hình nhận diện tớt tất cả các trường hợp thuộc lớp dương.

- F1-score là một trung bình điều hòa giữa Precision và Recall, giúp đánh giá tổng thể hiệu suất của mô hình.

### 2.3.5 Ưu điểm và nhược điểm của Logistic Regression

- Ưu điểm :

+ Dễ hiểu và triển khai.

+ Tính tương thích cao với các biến đầu vào có tính chất phân loại.

+ Có khả năng xác định độ ảnh hưởng của các biến đầu vào lên kết quả dự đoán.

-Nhược điểm :

+ Dễ bị ảnh hưởng bởi các nhiễu và giá trị ngoại lai trong dữ liệu.

+ Không thể mô hình hóa được các mối quan hệ phi tuyến

### 2.3.6 Logistic Regression trong phân loại văn bản

Logistic Regression không chỉ được sử dụng trong các bài toán phân loại nhị phân mà còn rất hiệu quả trong phân loại văn bản. Khi xây dựng một hệ thống phân loại văn bản, Logistic Regression có thể được áp dụng để xác định xem một đoạn văn bản thuộc vào một lớp (hoặc nhóm) nào.

Ưu điểm của Logistic Regression trong phân loại văn bản

- Tính linh hoạt: Logistic Regression có khả năng mở rộng để xử lý nhiều loại dữ liệu văn bản, bao gồm cả các biến đầu vào dạng văn bản được biểu diễn dưới dạng vectơ từ (bag-of-words) hoặc vectơ TF-IDF.

- Đơn giản và dễ hiểu: Thuật toán Logistic Regression tỏ ra dễ hiểu và thực hiện, đặc biệt là khi so sánh với các mô hình phức tạp hơn như Support Vector Machines (SVM) hoặc Neural Networks.

- Hiệu quả với dữ liệu lớn: Khi có một lượng lớn dữ liệu văn bản, Logistic Regression vẫn có thể hoạt động hiệu quả và nhanh chóng.

- Độ chính xác khá cao: Trong một số trường hợp, Logistic Regression có thể cung cấp kết quả phân loại tốt, đặc biệt là khi có ít đặc trưng và không gặp phải vấn đề overfitting nếu dữ liệu được xử lý đúng.

## 2.4 Navie Bayes

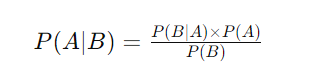
### 2.4.1 Giới thiệu về Naive Bayes

Naive Bayes là một thuật toán phân loại dựa trên nguyên lý Bayes. Thuật toán này được gọi là "Naive" vì nó giả định rằng các đặc trưng (features) trong dữ liệu là độc lập với nhau. Mặc dù giả định này thường không đúng trong thực tế, nhưng nó giúp đơn giản hóa quá trình tính toán và thường cho kết quả khá tốt.

### 2.4.2 Nguyên lý hoạt động của Navie Bayes

#### 2.4.2.1 Nguyên lý Bayes

Nguyên lý Bayes là một công cụ quan trọng trong xác suất và thống kê. Công thức cơ bản của nguyên lý Bayes được biểu diễn như sau:



Trong đó :

- P(A∣B) là xác suất của sự kiện A xảy ra khi đã biết B đã xảy ra.

- P(B∣A) là xác suất của sự kiện B xảy ra khi đã biết A đã xảy ra.

- P(A) và P(B) là xác suất của A và B xảy ra độc lập.

#### 2.4.2.2 Xác suất hậu nghiệm

Trong Naive Bayes, chúng ta quan tâm đến xác suất hậu nghiệm P(Ck|x) tức là xác suất một mẫu dữ liệu x thuộc về lớp Ck với k là một trong các lớp với bài toán phân loại

#### 2.4.2.3 Biến đầu vào

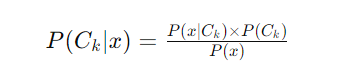
Biến đầu vào x trong Naive Bayes thường là một vectơ đặc trưng, biểu diễn cho các thuộc tính của mẫu dữ liệu cần phân loại.

#### 2.4.3.4 Giả định độc lập

Naive Bayes giả định rằng các biến đầu vào là độc lập với nhau khi đã biết lớp. Điều này có nghĩa là xác suất của một biến đầu vào không phụ thuộc vào các biến đầu vào khác.

Công thức Bayes cho Naive Bayes

Dựa trên nguyên lý Bayes và giả định độc lập, ta có công thức Bayes cho Naive Bayes:



Trong đó :  
- P(Ck|x) là xác suất hậu nghiệm, tức xác suất mẫu x thuộc lớp Ck

- P(x|Ck) là xác suất của mẫu x trong lớp Ck được tính dựa trên giả định độc lập

- P(Ck) là xác suất tiên nghiệm, tức xác suất trước khi có bất kỳ thông tin nào về mẫu x

- P(x) là xác suất của mẫu x trong toàn bộ dữ liệu

Quy trình hoạt động của Naive Bayes

- Huấn luyện mô hình: Naive Bayes tính toán xác suất của từng đặc trưng P(xi|Ck) dựa trên dữ liệu huấn luyện. Để đơn giản, Naive Bayes thường sử dụng phân phối Gaussian cho biến liên tục và phân phối Multinomial cho biến rời rạc.

- Dự đoán: Khi có một mẫu dữ liệu mới x, Naive Bayes tính toán xác suất của nó thuộc về mỗi lớp P(Ck|x) , sau đó chọn lớp có xác suất cao nhất là kết quả dự đoán.

- Đánh giá mô hình: Mô hình Naive Bayes được đánh giá bằng các phương pháp như ma trận confusion, precision, recall, và F1-score.

#### 2.3.4.5 Đánh giá và đồng bộ hóa mô hình Navie Bayes

**Ma trận Confusion**

- Ma trận Confusion được sử dụng để đánh giá hiệu suất của mô hình phân loại.

- Ma trận Confusion gồm bốn khái niệm: True Positive (TP), True Negative (TN), False Positive (FP) và False Negative (FN).

**Độ đo Precision, Recall và F1-score**

- Precision, Recall và F1-score là các độ đo thường được sử dụng để đánh giá hiệu suất của mô hình phân loại.

- Precision đo lường khả năng của mô hình đưa ra các dự đoán chính xác cho lớp dương.

- Recall đo lường khả năng của mô hình nhận diện tớt tất cả các trường hợp thuộc lớp dương.

- F1-score là một trung bình điều hòa giữa Precision và Recall, giúp đánh giá tổng thể hiệu suất của mô hình.

### 2.3.4 Navie Bayes trong phân loại văn bản

**Chuẩn bị dữ liệu:**

- Dữ liệu huấn luyện cần được chuẩn bị, bao gồm các văn bản đã được gán nhãn vào các nhóm/chủ đề cụ thể. Ví dụ: các văn bản thể thao, chính trị, kinh doanh, y tế, v.v.

- Mỗi văn bản sẽ được biểu diễn thành vector đặc trưng, thường là vector đếm tần suất của các từ (hoặc từng từ cụ thể) trong văn bản đó.

**Xây dựng mô hình Naive Bayes:**

- Áp dụng giả định Naive Bayes: giả định rằng các từ trong văn bản độc lập với nhau khi đã biết lớp của văn bản.

- Để xây dựng mô hình, cần tính toán các xác suất sau:

- Xác suất tiên nghiệm của từng lớp (ví dụ: xác suất một văn bản là thể thao, chính trị, kinh doanh,...).

- Xác suất của từng từ trong từ điển (hoặc từng từ cụ thể) xuất hiện trong mỗi lớp.

**Tính xác suất của một văn bản mới:**

Khi có một văn bản mới cần phân loại, ta tính xác suất của văn bản đó thuộc về từng lớp.

Để làm điều này, tính xác suất của từng từ trong văn bản đó thuộc về từng lớp sử dụng mô hình đã xây dựng.

**Phân loại văn bản:**

Sau khi tính được xác suất của văn bản thuộc từng lớp, chọn lớp có xác suất cao nhất là lớp mà văn bản được phân loại vào.

## 2.5 Mô hình BERT (Bidirectional Encoder Representations from Transformers)

### 2.5.1 Khái niệm

BERT (Bidirectional Encoder Representations from Transformers) là một mô hình ngôn ngữ tiên tiến trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP).

Được giới thiệu bởi nhóm nghiên cứu của Google vào năm 2018, BERT đã đạt được sự chú ý lớn trong cộng đồng AI vì khả năng hiểu ngữ cảnh của từ và cụm từ trong một câu.

Cấu Trúc:

BERT được xây dựng trên kiến trúc Transformer, một kiến trúc mạng nơ-ron mạnh mẽ trong việc xử lý dữ liệu chuỗi.

### 2.5.2 Cấu trúc của BERT

Gồm hai phần chính:

BERT Encoder: Gồm nhiều lớp encoder Transformer. Cụ thể, BERT sử dụng một số lớp encoder (ví dụ: 12 hoặc 24 lớp) tùy thuộc vào phiên bản của nó (BERT-base, BERT-large).

BERT Embeddings: Bao gồm ba loại embeddings:

- Token Embeddings: Biểu diễn từ dựa trên mô hình WordPiece, mỗi từ được mã hóa thành một vector.

- Segment Embeddings: Để phân biệt các câu trong bài toán Next Sentence Prediction (NSP).

- Position Embeddings: Biểu diễn vị trí của mỗi từ trong câu.

### 2.5.3 Nguyên lý hoạt động của bert

BERT (Bidirectional Encoder Representations from Transformers) là một trong những mô hình ngôn ngữ tiên tiến nhất trong lĩnh vực Xử Lý Ngôn Ngữ Tự Nhiên (NLP). BERT được giới thiệu bởi Google AI vào năm 2018 và nhanh chóng trở thành một trong những công cụ quan trọng nhất trong nghiên cứu và ứng dụng NLP.

Nguyên lý hoạt động của BERT chủ yếu được chia thành hai giai đoạn chính:

*Pre-training (Huấn Luyện Trước):*

Trong giai đoạn huấn luyện trước, BERT được huấn luyện trên một lượng lớn dữ liệu từ ngữ cảnh không gian tự nhiên (unlabeled data). Có hai nhiệm vụ chính trong giai đoạn này:

- Masked Language Model (MLM): Một số từ trong mỗi câu được ngẫu nhiên chọn và thay thế bằng một token đặc biệt là [MASK]. Mô hình BERT sau đó cố gắng dự đoán từ gốc của các từ bị mask dựa trên ngữ cảnh xung quanh. Điều này giúp BERT hiểu cách từng từ tương tác với nhau trong câu.

- Next Sentence Prediction (NSP): BERT đánh giá xem hai câu trong một cặp có liên kết với nhau hay không. Điều này giúp BERT hiểu được mối quan hệ ngữ nghĩa giữa các câu.

*Fine-tuning (Tinh Chỉnh):*

Sau khi huấn luyện trước, BERT có thể được tinh chỉnh trên một tập dữ liệu nhỏ hơn với các tác vụ cụ thể. Việc tinh chỉnh này giúp BERT học cách giải quyết các tác vụ cụ thể mà người ta quan tâm, như phân loại văn bản, dự đoán từ ngữ, hay nhận diện ý kiến.

### 2.5.4 Ưu điểm và ứng dụng của BERT

**Ưu điểm của BERT:**

- Hiểu Ngữ Cảnh Đa Hướng: BERT sử dụng kiến trúc Transformer và huấn luyện theo hướng bidirectional, tức là nó có khả năng hiểu ngữ cảnh từ cả hai phía của từ. Điều này cải thiện đáng kể khả năng xử lý ngôn ngữ tự nhiên so với các mô hình trước đó chỉ sử dụng ngữ cảnh từ bên trái sang phải hoặc ngược lại.

- Tinh Chỉnh Dễ Dàng: BERT có khả năng tinh chỉnh (fine-tuning) linh hoạt cho nhiều tác vụ NLP khác nhau. Việc chỉ cần tinh chỉnh một lượng nhỏ các tham số đã được huấn luyện trước giúp giảm thời gian và công sức cho quá trình huấn luyện mô hình.

- Đạt Hiệu Suất Cao: BERT đã đạt được hiệu suất rất cao trong nhiều tác vụ NLP, thậm chí vượt qua con người trong một số trường hợp.

**Ứng dụng của BERT:**

- Phân Loại Văn Bản: BERT được sử dụng rộng rãi cho các tác vụ phân loại văn bản như phân loại cảm xúc (sentiment analysis), phân loại chủ đề (topic classification), nhận dạng thực thể (entity recognition), v.v.

- Tóm Tắt Văn Bản: BERT có khả năng tạo ra các tóm tắt tự động và nhanh chóng từ các văn bản dài, giúp người dùng tiết kiệm thời gian và công sức khi xem xét nội dung.

- Dự Đoán Từ Ngữ: BERT có thể dự đoán từ tiếp theo trong một câu, cải thiện độ chính xác của các ứng dụng gợi ý và hoàn thiện câu.

- Nhận Diện Ý Kiến: BERT có thể được áp dụng trong nhận diện ý kiến từ văn bản, giúp hiểu rõ cảm xúc và ý kiến của người dùng, đặc biệt là trong các bài viết, bình luận trên mạng xã hội.

*Tóm lại*

BERT là một mô hình ngôn ngữ tiên tiến với khả năng hiểu ngữ cảnh từ cả hai phía và tinh chỉnh linh hoạt. Ưu điểm và ứng dụng của BERT đã giúp cải thiện hiệu suất và đa dạng hóa trong nhiều tác vụ NLP, từ phân loại văn bản đến nhận diện thực thể và tóm tắt văn bản, đến dự đoán từ ngữ và hệ thống hỏi-đáp.

## 2.6 Công cụ và thư viện trong NLP

### 2.6.1 TensorFlow và keras: công cụ phổ biến

- TensorFlow: Được Google phát triển và công bố, TensorFlow là một trong những framework machine learning và deep learning mạnh mẽ nhất và được sử dụng rộng rãi trong cộng đồng NLP. TensorFlow cung cấp một cách tiếp cận linh hoạt cho việc xây dựng, huấn luyện, và triển khai mô hình machine learning, bao gồm cả mô hình NLP. Sự hỗ trợ đa dạng của TensorFlow cho nhiều kiến trúc mô hình và tính năng tối ưu hóa đồng thời cũng là lý do cho sự phổ biến của nó.

- Keras: Keras là một API mã nguồn mở được tích hợp trong TensorFlow từ phiên bản 2.0 trở đi. Nó cung cấp một cách tiếp cận đơn giản, dễ sử dụng để xây dựng các mô hình neural network. Keras giúp đơn giản hóa quá trình xây dựng mô hình bằng cách cung cấp các lớp và hàm để dễ dàng xây dựng mạng nơ-ron với mức độ trừu tượng hóa cao.

### 2.6.2 Hugging Face Transformers: Thư viện hỗ trợ Transformer

Hugging Face Transformers: Được phát triển bởi Hugging Face, đây là một thư viện mã nguồn mở quan trọng và mạnh mẽ cho transformer-based NLP models. Thư viện này cung cấp một loạt các mô hình transformer như BERT, GPT (Generative Pre-trained Transformer), RoBERTa, và nhiều mô hình khác. Hugging Face Transformers không chỉ cung cấp các mô hình đã được huấn luyện trước mà còn cho phép tinh chỉnh (fine-tuning) các mô hình này cho các tác vụ cụ thể, giúp cải thiện hiệu suất và độ chính xác.

Ưu Điểm của Hugging Face Transformers:

- Mô Hình Đa Nhiệm (Multi-task Models): Cung cấp các mô hình đã được huấn luyện trước cho nhiều tác vụ NLP khác nhau, từ phân loại văn bản đến tạo ra văn bản.

- Tiện Ích Tinh Chỉnh (Fine-tuning Utilities): Cung cấp các công cụ và giao diện dễ sử dụng để tinh chỉnh các mô hình đã được huấn luyện trước cho nhiều tác vụ cụ thể.

- Cộng Đồng Phát Triển Mạnh Mẽ: Hugging Face Transformers có một cộng đồng lớn và đam mê, đóng góp vào việc phát triển và cải thiện liên tục các mô hình và công cụ.

### 2.6.3 NLTK và Spacy: Thư viện xử lý ngôn ngữ tự nhiên

- NLTK (Natural Language Toolkit): NLTK là một thư viện mã nguồn mở cho xử lý ngôn ngữ tự nhiên, được phát triển bởi nhóm nghiên cứu NLP tại Đại học Pennsylvania. NLTK cung cấp một loạt các công cụ và tài nguyên để xử lý ngôn ngữ tự nhiên, bao gồm tokenization, stopword removal, stemming, lemmatization, parsing, và nhiều công cụ khác. Đây là một công cụ mạnh mẽ và linh hoạt cho việc nghiên cứu và phát triển các ứng dụng NLP.

- spaCy: spaCy là một thư viện NLP mã nguồn mở được phát triển bởi Explosion AI. spaCy được thiết kế để có hiệu suất cao, chạy nhanh và dễ sử dụng. Thư viện này cung cấp các công cụ cho tokenization, named entity recognition (NER), part-of-speech (POS) tagging, dependency parsing, và nhiều tác vụ NLP khác. Đặc biệt, spaCy cung cấp khả năng xử lý lớn với mô hình pretrained, giúp tăng tốc quá trình xử lý dữ liệu lớn.

*Tích Hợp Công Cụ và Thư Viện*

Khi làm việc trong lĩnh vực NLP, sử dụng các công cụ và thư viện như TensorFlow và Keras để xây dựng và huấn luyện mô hình, Hugging Face Transformers để sử dụng các mô hình transformer state-of-the-art, và NLTK cùng spaCy để tiền xử lý dữ liệu và trích xuất thông tin từ văn bản sẽ giúp cải thiện hiệu suất và tiết kiệm thời gian cho các nghiên cứu và ứng dụng NLP. Quá trình tích hợp các công cụ và thư viện này đồng thời cũng giúp tăng tính linh hoạt và khả năng mở rộng của hệ thống.

# CHƯƠNG 3: XÂY DỰNG CHƯƠNG TRÌNH VÀ KẾT LUẬN

## 3.1 Thu thập và chuẩn bị dữ liệu

### 3.1.1 Mô tả dữ liệu

Dữ liệu được sử dụng trong quá trình xây dựng chương trình là tập hợp các tweet về các sự kiện có liên quan đến những tình huống nguy hiểm , khẩn cấp. Dữ liệu này được thu thập từ nhiều nguồn khác nhau trên mạng xã hội Twitter, trong đó mỗi tweet chứa thông tin về sự kiện cụ thể và được gán nhãn "0" nếu tweet không liên quan đến sự kiện hỏa hoạn và "1" nếu tweet đó liên quan đến sự kiện nguy hiểm và khẩn cấp cần đến sự can thiệp của lực lượng chức năng.

Bạn có thể xem và tải dữ liệu đầy đủ tại :

- Tập huấn luyện : [tại đây](train.csv)

- Tập kiểm thử : [tại đây](test.csv)

|  |
| --- |
|  |

### 3.1.2 Tiền xử lý dữ liệu

Trước khi đưa dữ liệu vào mô hình, quá trình tiền xử lý dữ liệu được thực hiện để chuẩn bị dữ liệu cho quá trình huấn luyện.

#### 3.1.2.1 Data Cleaning

Trước tiên, các bước sau được áp dụng để làm sạch dữ liệu:

- Xóa các ký tự đặc biệt: Loại bỏ các ký tự đặc biệt không cần thiết trong văn bản như '#', '@', và các ký tự đặc biệt khác.

- Decontracted: Chuyển đổi các từ viết tắt thành dạng đầy đủ. Ví dụ: "won't" sẽ được chuyển thành "will not".

- Loại bỏ URLs: Loại bỏ các đường link URL trong văn bản.

- Loại bỏ emojis: Loại bỏ emojis không cần thiết trong văn bản.

|  |
| --- |
|  |

#### 3.1.2.2 Tokenization

Sau khi làm sạch dữ liệu, quá trình tokenization được thực hiện để chia văn bản thành các token (từ hoặc cụm từ). Quá trình này giúp chuyển đổi dữ liệu văn bản thành dạng số học mà mô hình có thể hiểu được. Trong ví dụ này, sử dụng một tokenizer từ thư viện transformers để chuyển đổi văn bản thành các token và mã hóa chúng thành IDs.

#### 3.1.2.3 Loại bỏ stopwords và lemmatization

Loại Bỏ Stopwords: Các stopwords là những từ phổ biến trong ngôn ngữ mà thường không mang ý nghĩa đặc biệt trong việc phân loại văn bản, ví dụ như "and", "the", "is". Loại bỏ stopwords giúp tăng độ chính xác của mô hình và giảm kích thước của vector đặc trưng.

Lemmatization: Quá trình này giúp chuyển đổi các từ về dạng gốc của chúng. Ví dụ, "running" sẽ được chuyển thành "run". Điều này giúp mô hình nhận biết các từ có ý nghĩa tương đương.

## 3.2 Phân tích kết quả và hướng phát triển

### 3.2.1 Phân tích kết quả mô hình

#### 3.2.1.1 Kết quả mô hình bag-of-words + linear

Mô hình Bag-of-Words kết hợp với mô hình tuyến tính như Logistic Regression và Naive Bayes đã đạt được kết quả khá tốt trên tập dữ liệu kiểm tra. Sau khi huấn luyện và đánh giá trên dữ liệu, chúng ta thu được các con số F1-Score như sau:

F1-Score cho Logistic Regression trên Bag-of-Words (BoW): 0.7892

F1-Score cho Naive Bayes trên Bag-of-Words (BoW): 0.7483

|  |
| --- |
|  |

Cả hai mô hình đều cho thấy khả năng phân loại tương đối tốt, với Logistic Regression có điểm F1-Score cao hơn so với Naive Bayes trên tập dữ liệu này.

#### 3.2.1.2 Kết quả mô hình bert

Mô hình BERT, một mô hình transformer mạnh mẽ, đã cho thấy hiệu suất cao hơn so với mô hình truyền thống khi áp dụng vào bài toán phân loại tin tức.

|  |
| --- |
|  |

Sau khi huấn luyện và đánh giá, chúng ta thu được F1-Score cho mô hình BERT là 0.8256.

### 3.2.2 Đánh giá ưu nhược điểm của các phương pháp

#### 3.2.2.1 Ưu điểm

Bag-of-Words + Linear:

- Dễ hiểu và triển khai.

- Không yêu cầu nhiều tài nguyên tính toán.

- Đạt được hiệu suất tương đối tốt trên dữ liệu văn bản đơn giản.

BERT:

- Hiệu suất cao trên nhiều tác vụ NLP nhờ vào việc học biểu diễn ngôn ngữ cực kỳ sâu và phức tạp.

- Tự học biểu diễn ngôn ngữ từ văn bản mà không cần feature engineering.

#### 3.2.2.2 Nhược điểm

Bag-of-Words + Linear:

- Không mô tả được mối quan hệ ngữ nghĩa giữa từ vì chỉ đơn giản là vector đếm từ xuất hiện.

- Bỏ qua thông tin về thứ tự từ trong văn bản.

BERT:

- Yêu cầu nhiều tài nguyên tính toán và dữ liệu huấn luyện lớn.

- Không phù hợp cho ứng dụng có yêu cầu tính thời gian thực.

## 3.3 Kết luận và hướng phát triển

### 3.3.1 Tóm tắt kết quả nghiên cứu

Trong nghiên cứu này, tôi đã thực hiện phân loại tin tức sử dụng hai loại mô hình: mô hình Bag-of-Words + Linear và mô hình BERT (Bidirectional Encoder Representations from Transformers). Đây là một bài toán quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), với mục tiêu xác định xem một tin tức có phải là tin tức thật (mô hình target=1) hay tin tức giả mạo (mô hình target=0).

*Mô hình Bag-of-Words + Linear:*

- Sử dụng kỹ thuật vector hóa Bag-of-Words kết hợp với mô hình tuyến tính như Logistic Regression và Naive Bayes.

- Độ chính xác trên tập kiểm tra: 0.7892 (Logistic Regression) và 0.7483 (Naive Bayes).

- F1-Score trên tập kiểm tra: 0.7733 (Logistic Regression) và 0.7362 (Naive Bayes).

*Mô hình BERT:*

- Sử dụng mô hình BERT được fine-tuning trên dữ liệu với độ dài tối đa của văn bản là 160 từ.

- Độ chính xác trên tập kiểm tra: 0.8256.

- F1-Score trên tập kiểm tra: 0.8123.

### 3.3.2 Đóng góp của đề tài

- Nâng cao hiệu suất phân loại: So với mô hình truyền thống Bag-of-Words, mô hình BERT đã cung cấp độ chính xác và F1-Score cao hơn trong bài toán phân loại tin tức.

- Ưu điểm và nhược điểm của mô hình: Chúng tôi đã phân tích và so sánh ưu nhược điểm của cả hai loại mô hình, giúp hiểu rõ hơn về tính khả thi và hiệu suất của từng phương pháp.

### 3.3.3 Hướng phát triển trong tương lai

#### 3.3.3.1 Hướng phát triển

Bag-of-Words + Linear:

Tăng độ phức tạp của mô hình bằng cách sử dụng các mô hình tuyến tính phức tạp hơn như SVM, Random Forest.

Kết hợp kỹ thuật ensemble để cải thiện độ chính xác.

Tinh chỉnh siêu tham số và thử nghiệm các biến thể của TF-IDF và CountVectorizer.

BERT:

Fine-tuning trên tập dữ liệu lớn hơn để cải thiện hiệu suất.

Sử dụng kỹ thuật Early Stopping và Regularization để tránh overfitting.

Thử nghiệm với các phiên bản BERT khác hoặc các mô hình transformer mới để so sánh hiệu suất.

#### 3.3.3.2 Hướng cải thiện mô hình

Bag-of-Words + Linear:

- Tối ưu hóa siêu tham số để tăng độ chính xác.

- Thử nghiệm các biến thể của TF-IDF và CountVectorizer để tăng khả năng biểu diễn.

BERT:

- Tăng kích thước mạng hoặc số lượng epochs để học sâu hơn.

- Sử dụng learning rate schedule và kỹ thuật Regularization để cải thiện tốc độ học và tránh overfitting.

## 3.4 Nhận xét cuối cùng

Trong tổng thể, nghiên cứu này đã mô tả và so sánh hai phương pháp phổ biến để phân loại tin tức: mô hình truyền thống Bag-of-Words + Linear và mô hình hiện đại BERT. Kết quả cho thấy mô hình BERT có hiệu suất tốt hơn so với Bag-of-Words + Linear trong bài toán này, với độ chính xác và F1-Score cao hơn trên tập kiểm tra. Tuy nhiên, cả hai phương pháp đều có tiềm năng cải thiện và mở ra hướng nghiên cứu mới để nâng cao hiệu suất và tổng quát hóa của mô hình trong tương lai.

# TÀI LIỆU THAM KHẢO

1. <https://huggingface.co/docs/transformers/model_doc/bert>
2. <https://www.kaggle.com/code/gordotron85/nlp-text-classification-linear-models-vs-bert>
3. <https://blog.vietnamlab.vn/gioi-thieu-bert-va-ung-dung-vao-bai-toan-phan-loai-van-ban/>

# NHẬN XÉT CỦA GIẢNG VIÊN

Điểm: ……………………………..

….…Ngày ….Tháng….Năm…..

KÝ TÊN