|  |  |  |
| --- | --- | --- |
| A blue and yellow logo  Description automatically generated with low confidence | **TRƯỜNG ĐẠI HỌC BÁCH KHOA**  **KHOA CÔNG NGHỆ THÔNG TIN** |  |

TIỂU LUẬN CUỐI KỲ

HỌC PHẦN: KHOA HỌC DỮ LIỆU

**NHẬN DẠNG TIÊU ĐỀ BÀI BÁO**

|  |  |  |
| --- | --- | --- |
| HỌ VÀ TÊN SINH VIÊN | LỚP HỌC PHẦN | ĐIỂM BẢO VỆ |
| Lê Hữu Hưng | 20TCLC-KHDL |  |
| Hoàng Nguyên Bách | 20TCLC-KHDL |  |
| Đinh Huy Hoàng | 20TCLC-KHDL |  |

**GIẢNG VIÊN GIẢNG DẠY:**

**TS. Ninh Khánh Duy**

**Đà Nẵng, 05/2023**

**TÓM TẮT**

Trong tiểu luận này, chúng tôi tập trung vào giải quyết vấn đề nhận dạng tiêu đề bài báo từ văn bản. Vấn đề này là một trong những thách thức quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên, do tiêu đề bài báo thường chứa thông tin quan trọng và đóng vai trò quan trọng trong việc thu hút sự chú ý của người đọc.

Phương pháp giải quyết của chúng tôi là sử dụng PhoBert, một mô hình nhúng từ vựng tiếng Việt dựa trên kiến trúc Transformer. PhoBert cho phép chúng tôi biểu diễn các từ trong tiêu đề bài báo dưới dạng các vectơ số thực, giúp mô hình có khả năng hiểu được ngữ nghĩa và tương quan giữa các từ trong tiêu đề.

Kết quả đạt được khi xây dựng mô hình nhận dạng tiêu đề bài báo là khả năng nhận dạng và phân loại chính xác các tiêu đề bài báo từ văn bản. Mô hình của chúng tôi đã được huấn luyện trên một tập dữ liệu đa dạng và đạt được độ chính xác cao trong việc nhận dạng và phân loại các tiêu đề bài báo. Điều này có thể hỗ trợ trong việc tổ chức và phân loại thông tin từ các bài báo, giúp người dùng tiết kiệm thời gian và nỗ lực trong việc tìm hiểu và tìm kiếm thông tin cần thiết từ các tiêu đề bài báo.

**BẢNG PHÂN CÔNG NHIỆM VỤ**

|  |  |  |
| --- | --- | --- |
| Sinh viên thực hiện | Các nhiệm vụ | Tự đánh giá theo 3 mức  (Đã hoàn thành/Chưa hoàn thành/Không triển khai) |
| Lê Hữu Hưng | * Xử lý dữ liệu * Trích xuất đặc trưng | * Đã hoàn thành * Đã hoàn thành |
| Hoàng Nguyên Bách | * Crawl data | * Đã hoàn thành |
| Đinh Huy Hoàng | * Mô hình hóa dữ liệu | * Đã hoàn thành |

**MỤC LỤC**

[1. Giới thiệu 5](#_Toc134827351)

[1.1 Lý do chọn đề tài 5](#_Toc134827352)

[1.2 Giới thiệu bài toán 5](#_Toc134827353)

[1.2.1 Mục tiêu 6](#_Toc134827354)

[1.2.2 Giải pháp 6](#_Toc134827355)

[2. Thu thập và mô tả dữ liệu 5](#_Toc134827356)

[2.1 Thu thập dữ liệu 5](#_Toc134827357)

[2.2 Mô tả dữ liệu 5](#_Toc134827358)

[3. Trích xuất đặc trưng 5](#_Toc134827359)

[3.1 Lý chọn đặc trưng 5](#_Toc134827360)

[3.2 Làm sạch dữ liệu 5](#_Toc134827361)

[3.3 Chuẩn hóa dữ liệu 5](#_Toc134827362)

[3.4 Giảm chiều dữ liệu 5](#_Toc134827363)

[3.5 Trực quan hóa dữ liệu 5](#_Toc134827364)

[4. Mô hình hóa dữ liệu 5](#_Toc134827365)

[4.1. Phát biểu bài toán 5](#_Toc134827366)

[4.2. Phân Chia dữ liệu 5](#_Toc134827367)

[4.3. Thuận toán 5](#_Toc134827368)

[4.3.1. Support Vector Machine 6](#_Toc134827369)

[4.3.2. Random Forest Classifier 6](#_Toc134827370)

[4.4. Đánh giá thuật toán 5](#_Toc134827371)

[4.4.1. Accuracy 6](#_Toc134827372)

[4.4.2. F1 Score 6](#_Toc134827373)

[4.4.3. Recall 6](#_Toc134827374)

[4.4.4. Ma trận nhầm lẫn 6](#_Toc134827375)

[5. Kết luận 5](#_Toc134827376)

[5.1 Kết quả đạt được 6](#_Toc134827377)

[5.2 Giới hạn của đề tài 6](#_Toc134827378)

[6. Tài liệu tham khảo 5](#_Toc134827379)

# Giới thiệu

## Lý do chọn đề tài

Trong thế giới thông tin kỹ thuật số ngày nay, hàng ngàn bài báo được xuất bản hàng ngày trên các nền tảng trực tuyến. Việc xác định và phân loại tiêu đề bài báo có thể giúp chúng ta nắm bắt thông tin nhanh chóng và tiết kiệm thời gian. Điều này có thể áp dụng trong nhiều lĩnh vực, bao gồm việc tìm kiếm thông tin, theo dõi xu hướng, và phân tích dữ liệu.

Việc sử dụng mô hình nhận dạng tiêu đề bài báo có thể giúp tự động hóa quy trình công việc liên quan đến xử lý và phân loại thông tin. Thay vì phải đọc từng tiêu đề bài báo một cách thủ công, ta có thể áp dụng mô hình để tự động phân loại và tổ chức thông tin dựa trên tiêu đề. Mô hình nhận dạng tiêu đề bài báo có thể được ứng dụng trong nhiều lĩnh vực thực tế, bao gồm phân loại tin tức, theo dõi sự kiện, quảng cáo và tiếp thị.

Chính vì vậy, nhóm chúng tôi thực hiện đề tài ‘Nhận dạng tiêu đề bài báo’ để có thể phân loại các bài báo theo các tiêu đề một cách tự động.

## Giới thiệu bài toán

### Mục tiêu

Xây dựng mô hình phân loại thể loại của các bài báo.

* Đầu vào là của một bài báo bất kì.
* Đầu ra trả về là chủ đề chính của bài báo(bài báo đó thuộc lĩnh vực nào).

### Giải pháp

Phương pháp giải quyết của chúng tôi là sử dụng PhoBert, một mô hình nhúng từ vựng tiếng Việt dựa trên kiến trúc Transformer. PhoBert cho phép chúng tôi biểu diễn các từ trong tiêu đề bài báo dưới dạng các vectơ số thực, giúp mô hình có khả năng hiểu được ngữ nghĩa và tương quan giữa các từ trong tiêu đề.

Từ đó có thể xây dựng mô hình nhận dạng tiêu đề bài báo có độ chính xác từ các tiêu đề bài báo văn bản.

# Thu thập và mô tả dữ liệu

## Thu thập dữ liệu

Là quá trình thu thập các dữ liệu liệu cần thiết dùng để thực hiện dự án. Việc thu thập dữ liệu phục vụ cho xử lý ngôn ngữ tự nhiên có thể được thực hiện trên nhiều hình thức khác nhau. Phổ biến trong đó là thu thập dữ liệu từ các website.

Trang website nhóm chúng tôi thực hiện việc thu thập dữ liệu là: vtv.vn



Mỗi bài báo trên website sẽ gồm có những nội dung như sau:

* Tiêu đề.
* Ngày viết bài.
* Chủ đề.
* Mô tả ngắn gọn.

Từ những nội dung trên website nhóm chúng tôi quyết định sẽ thu thập nội dung để thực hiện việc huấn luyện mô hình là:

* Tiêu đề.
* Chủ đề.

## Mô tả dữ liệu

Từ những dữ liệu trên nhóm đã thu thập thành công được hai tập dữ liệu.

* Small data: gồm 1374 mẫu.
* Big data: gồm 10981 mẫu.

Số lượng nhãn(tiêu đề) của từng tập dữ liệu bao gồm:

* Small data

|  |  |
| --- | --- |
| **Tiêu đề** | **Số lượng** |
| kinh-te | 2326 |
| truyen-hinh | 2164 |
| chinh-tri | 1856 |
| van-hoa-giai-tri | 1564 |
| doi-song | 1456 |
| xa-hoi | 1005 |
| cong-nghe | 548 |

* Big data

|  |  |
| --- | --- |
| **Tiêu đề** | **Số lượng** |
| kinh-te | 566 |
| truyen-hinh | 213 |
| chinh-tri | 189 |
| van-hoa-giai-tri | 156 |
| doi-song | 113 |
| xa-hoi | 102 |
| cong-nghe | 36 |

# Trích xuất đặc trưng

## Chọn đặc trưng

## Làm sạch dữ liệu

## Chuẩn hóa dữ liệu

## Giảm chiều dữ liệu

## Trực quan hóa dữ liệu

# Mô hình hóa dữ liệu

## Phát biểu bài toán

Ádasd

## Phân Chia dữ liệu

## Thuận toán

### Support Vector Machine

### Random Forest Classifier

## Đánh giá thuật toán

### Accuracy

### F1 Score

### Recall

### Ma trận nhầm lẫn

# Kết luận

## Kết quả đạt được

## Giới hạn của đề tài

# Tài liệu tham khảo