**CỘNG HÒA XẪ HỘI CHỦ NGHĨA VIỆT NAM  
Độc lập – Tự do – Hạnh phúc  
-------------------------------**

Báo cáo kết quả thực hiện

Dự án: phân tích quan điểm

Hà Nội – 2022

Mục lục

[A. Tổng quan dự án 3](#_Toc110957192)

[1.1. Mục đích tài liệu 3](#_Toc110957193)

[1.2. Cấu trúc tài liệu 3](#_Toc110957194)

[1.3. Cấu trúc luồng xử lý 3](#_Toc110957195)

[B. Tổng quan dữ liệu 5](#_Toc110957196)

[1. Khái quát 5](#_Toc110957197)

[2. Các tập dữ liệu 6](#_Toc110957198)

[2.1. Dữ liệu về luật đất đai 6](#_Toc110957199)

[2.2. Dữ liệu về ô nhiễm môi trường 6](#_Toc110957200)

[2.3. Dữ liệu về ô nhiễm không khí 6](#_Toc110957201)

[3. Định nghĩa và mô tả nhãn 8](#_Toc110957202)

[C. Mô hình dự đoán và kết quả 10](#_Toc110957203)

[1. Mô hình dự đoán quan điểm dựa trên từ vựng (LBA) 10](#_Toc110957204)

[2. TF-IDF kết hợp với một số mô hình học máy 11](#_Toc110957205)

[3. PhoBERT kết hợp với một số mô hình học máy 11](#_Toc110957206)

[4. Mô hình học sâu Maximum Entropy(MaxEnt) 12](#_Toc110957207)

# Tổng quan dự án

## Mục đích tài liệu

Tài liệu này nhằm tóm tắt, thu gọn lại quá trình thực hiện cũng như đưa ra kết quả của dự án ứng dụng học máy/ học sâu vào phân loại sắc thái cảm xúc (tích cực hay tiêu cực) của những bài báo và nguồn tin chính thống về các chủ để nhất định.

Bài toán và yêu cầu:

* Thu thập thông tin từ các nguồn báo chính thống.
* Đánh giá phân loại sắc thái cảm xúc.
* Các chủ đề xoay quanh bao gồm: Tài nguyên môi trường, môi trường, đất đai,..
* Thử nghiệm từ chủ đề nhỏ trước nhằm đánh giá khả năng giải pháp và khả năng triển khai

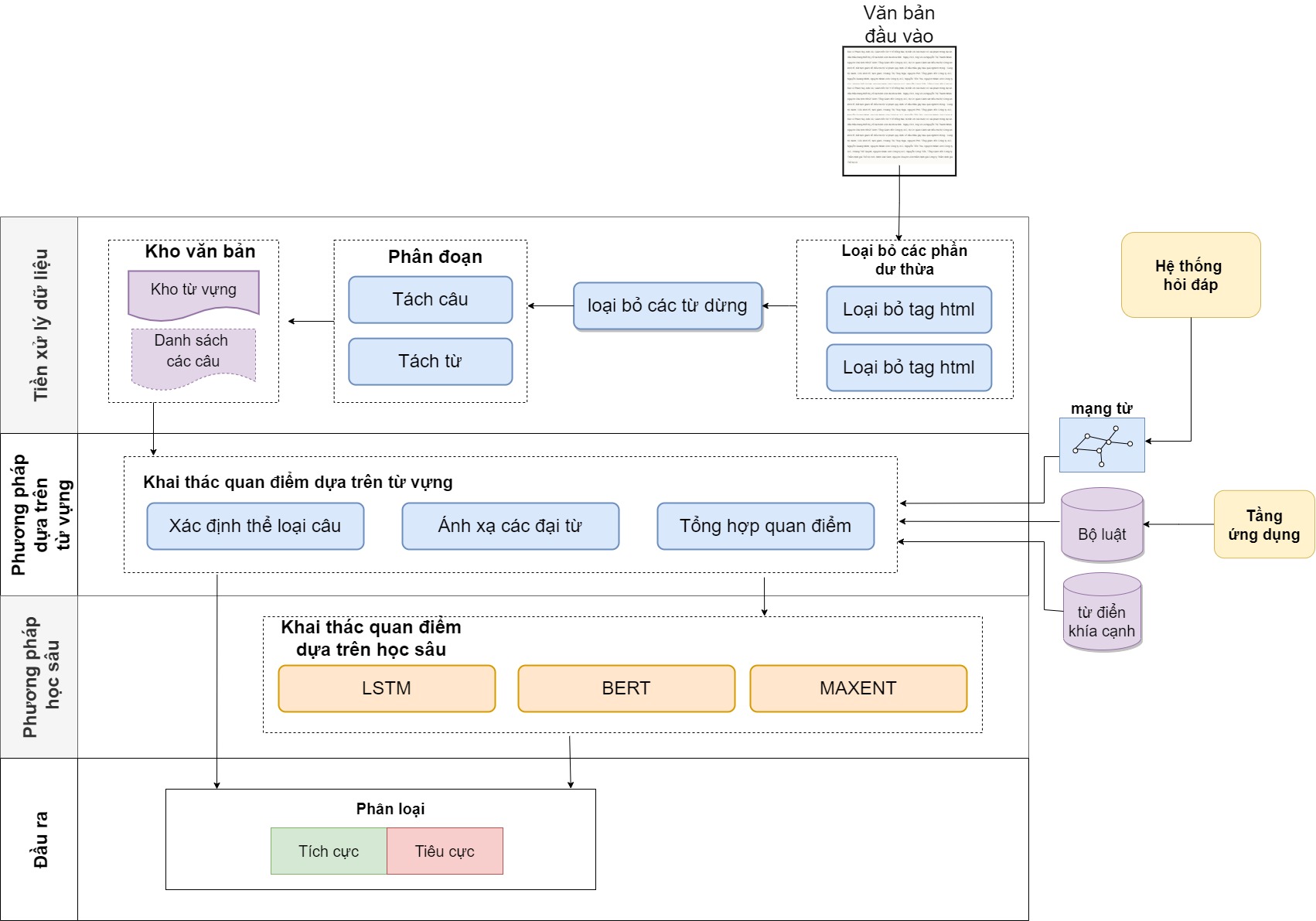
## Cấu trúc tài liệu

Báo cáo bao gồm 2 phần chính:

* *Tổng quan và mô tả dữ liệu*: khái quát nguồn dữ liệu và mô tả các tập dữ liệu đã thu thập
* *Các mô hình dự đoán được sử dụng*: mô tả các mô hình được sử dụng từ xử lí dữ liệu đến kết quả đạt được

## Cấu trúc luồng xử lý

Cấu trúc quá trình từ khi thu thập dữ liệu cho đến khi cho ra kết quả cuối cùng gồm nhiều công đoạn, cụ thể được tóm gọn lại dưới đây:



Mô tả khái quát chu trình xử lý

# Tổng quan dữ liệu

## Khái quát

Dữ liệu được thu thập và sử dụng trong dự án đều được lấy từ hàng trăm nguồn báo chính thống trong nước như Vnexpress, Báo điện tử VTV, VOVTV,... được cập nhật liên tục từng ngày, từng giờ, nhằm bám sát thực tế bối cảnh thực tại.

Qua việc sử dụng các bộ lọc thể loại cũng như sử dụng các bộ từ khóa đặc trưng, các bộ dữ liệu theo chủ đề luật đất đai và ô nhiễm môi trường lần lượt được thu về. Ở mỗi bài báo, thực hiện thu về dữ liệu với các trường sau:

|  |  |
| --- | --- |
| **Nhãn** | **Precision** |
| -1 | 0.62 |
| 1 | 0.26 |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
| Nhãn | Precision |

### 2.1. Dữ liệu về luật đất đai

Tập dữ liệu về luật đất đai được tổng hợp từ khoảng 20.000 bài báo, là kết quả nhận được khi luồng tổng hợp dữ liệu đã được đi qua bộ lọc với các keyword như: đất đai, sổ đỏ, …

### 2.2. Dữ liệu về ô nhiễm môi trường

Tập dữ liệu về ô nhiễm môi trường được tổng hợp từ khoảng 16.000 bài báo, sử dụng bộ lọc với các keyword: ô nhiễm, khói bụi, bụi mịn, bụi bẩn, vứt rác, nước thải, đen ngòm, xả thải, …

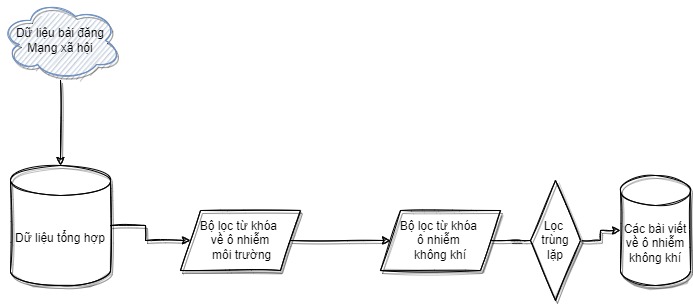
|  |  |
| --- | --- |
| **Nhãn** | **Precision** |
| -1 | 0.62 |
| 1 | 0.26 |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
| Nhãn | Precision |
| -1 | 0.91 |
| 1 | 0.85 |
|  |  |
|  |  |

### 2.3. Dữ liệu về ô nhiễm không khí

Do yêu cầu sử dụng dữ liệu về ô nhiễm không khí để thử nghiệm đánh giá giải pháp và khả năng triển khai, đây sẽ là tập dữ liệu được đi qua nhiều phễu lọc nhất nhằm cho ra dữ liệu đảm bảo chất lượng.

Tập dữ liệu về ô nhiễm không khí là sản phẩm khi tinh chỉnh lại bộ từ khóa cho bộ từ của về ô nhiễm môi trường, trong đó bổ sung thêm một số từ đặc trưng cho chủ đề như: chất lượng không khí, mùi hôi thối, CO2, CO, SO2, PM2.5, NH3, Sox, đốt rơm đốt rạ,…

Sau khi thu được dữ liệu từ các bộ lọc từ khóa, tiếp tục lọc lại dữ liệu nhằm loại bỏ đi các bài viết trùng lặp về nội dung (có thể do 2 hay nhiều nguồn đăng lại một bài, hoặc do trùng lặp do thời gian lấy dữ liệu thay đổi).



Tóm gọn quá trình thu thập dữ liệu về ô nhiễm không khí

Sau khi dữ liệu được đi qua các phễu lọc kể trên, ta thu được tập dữ liệu về ô nhiễm không khí với khoảng trên 2000 bản ghi.



Một số bản ghi mẫu từ tập dữ liệu ô nhiễm không khí

Tập dữ liệu này sẽ là tập dữ liệu chính sử dụng cho huấn luyện và kiểm thử hiệu quả và tính khả thi của các mô hình dự đoán.

## Định nghĩa và mô tả nhãn

Mục tiêu của các mô hình dự đoán là phân loại các bài viết về 2 nhãn:

|  |  |  |
| --- | --- | --- |
| **Nhãn** | **Ý nghĩa** | **Ví dụ** |
| 1 | Bài viết tích cực | - Bài viết về cắt giảm khí thải  - Bài viết về các công tác xử lý ô nhiễm |
| -1 | Bài viết tiêu cực | - Bài về xả thải, cháy rừng gây ô nhiễm  - Bài viết về hậu quả ô nhiễm ảnh hưởng đến đời sống |

# Mô hình dự đoán và kết quả

## Mô hình dự đoán quan điểm dựa trên từ vựng (LBA)

**LBA (Lexicon Based Aproach)** là một phương pháp chấm điểm điểm cảm xúc của đoạn văn bản dựa trên điểm cảm xúc của các từ vựng và cụm từ trong đoạn văn bản đó.

Ban đầu mô hình được triển khai với bộ Vietnamese SentiWordNet, kết quả cho ra được kết quả khá thấp với độ chuẩn xác (accuracy) chỉ khoảng 30%.

|  |  |  |  |
| --- | --- | --- | --- |
| **Nhãn** | **Precision** | **Recall** | **F-score** |
| -1 | 0.62 | 0.11 | 0.18 |
| 1 | 0.26 | 0.83 | 0.4 |

*Bảng mô tả kết quả LBA với bộ từ Vietnamese SentiWordNet*

Bộ Vietnamese SentiWordNet tuy đã được nghiên cứu và phát triển bởi các chuyên gia nhưng lại cho ra kết quả không tốt vì

* Các bài báo đa phần được viết theo cách rất khách quan, mang ít sắc thái cảm xúc nên điều quyết định đến sự tích cực hay tiêu cực của bài báo chỉ dựa vào thông tin của nó.
* Các bài báo gồm nhiều từ vựng thuộc chủ đề ô nhiễm không khí không chứa trong bộ Vietnamese SentiWordNet

Vì kết quả LBA với bộ Vietnamese SentiWordNet không tốt nên nhóm triển khai đã tiến hành phát triển 1 bộ từ riêng dành cho chủ đề ô nhiễm không khí chứa khoảng 127 từ vựng và cải thiện được mô hình với độ chuẩn xác 55%.

Kết quả cho ra:

|  |  |  |  |
| --- | --- | --- | --- |
| **Nhãn** | **Precision** | **Recall** | **F-score** |
| -1 | 0.9 | 0.31 | 0.46 |
| 1 | 0.45 | 0.94 | 0.61 |

Tuy nhiên, kết quả này vẫn chưa đạt yêu cầu đề ra nên nhóm triển khai đã chuyển hướng sang sử dụng học máy và học sâu để giải quyết bài toán này.

## TF-IDF kết hợp với một số mô hình học máy

**TF-IDF (Term Frequency – Inverse Document Frequency):** là 1 kĩ thuật sử dụng trong khai phá dữ liệu văn bản. Trọng số này được sử dụng để đánh giá tầm quan trọng của một từ trong một văn bản. Giá trị cao thể hiện độ quan trọng cao và nó phụ thuộc vào số lần từ xuất hiện trong văn bản nhưng bù lại bởi tần suất của từ đó trong tập dữ liệu. Một vài biến thể của TF-IDF thường được sử dụng trong các hệ thống tìm kiếm như một công cụ chính để đánh giá và sắp xếp văn bản dựa vào truy vấn của người dùng. TF-IDF cũng được sử dụng để lọc những từ stopwords trong các bài toán như tóm tắt văn bản và phân loại văn bản.

Kết quả sử dụng TF-IDF cùng với một số mô hình học máy:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Mô hình** | **Nhãn** | **Precision** | **Recall** | **F-score** |
| SVM(accuracy: 86.17%) | -1 | 0.85 | 0.81 | 0.83 |
| 1 | 0.87 | 0.9 | 0.88 |
| LightGBM(accuracy: 81.91%) | -1 | 0.78 | 0.79 | 0.78 |
| 1 | 0.85 | 0.84 | 0.84 |
| Random Forest(accuracy: 85.64%) | -1 | 0.84 | 0.81 | 0.82 |
| 1 | 0.87 | 0.89 | 0.88 |

## PhoBERT kết hợp với một số mô hình học máy

Mô hình huấn luyện trước **PhoBert** là một mô hình học sâu xử lý ngôn ngữ hiện đại bậc nhất hiện nay dành cho ngôn ngữ Tiếng Việt. Đặc điểm nổi bật của phoBert:

Là mô hình quy mô lớn cho việc xử lý đơn ngôn ngữ Tiếng Việt đầu tiên. Hướng tiếp cận huấn luyện trước của PhoBert được dựa trên RoBERTa, tối ưu hóa quy trình huấn luyện trước của BERT để có hiệu suất mạnh mẽ hơn.

Kết quả cho ra sau khi kết hợp PhoBert với một số mô hình:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Mô hình** | **Nhãn** | **Precision** | **Recall** | **F-score** |
| SVM(accuracy: 78.19%) | -1 | 0.74 | 0.69 | 0.72 |
| 1 | 0.81 | 0.84 | 0.82 |
| LightGBM(accuracy: 87.23%) | -1 | 0.86 | 0.81 | 0.84 |
| 1 | 0.88 | 0.91 | 0.9 |
| Random Forest(accuracy: 87.23%) | -1 | 0.89 | 0.77 | 0.86 |
| 1 | 0.86 | 0.94 | 0.9 |

## Mô hình học sâu Maximum Entropy(MaxEnt)

Phương pháp phân loại Maximum Entropy (MaxEnt) là một bộ phân loại văn bản phổ biến, bằng cách tham số hóa mô hình để đạt được entropy phân loại tối đa, với ràng buộc rằng xác suất thu được trên dữ liệu huấn luyện với mô hình bằng với phân phối thực.

Kết quả sử dụng mô hình MaxEnt (Accuracy: 86.7%):

|  |  |  |  |
| --- | --- | --- | --- |
| **Nhãn** | **Precision** | **Recall** | **F-score** |
| -1 | 0.91 | 0.76 | 0.83 |
| 1 | 0.85 | 0.95 | 0.89 |