**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](#_Toc110604427)

[1.1. Mục đích tài liệu 3](#_Toc110604428)

[1.2. Cấu trúc tài liệu 3](#_Toc110604429)

[1.3. Cấu trúc luồng xử lý 3](#_Toc110604430)

[A. Tổng quan các nguồn dữ liệu 5](#_Toc110604431)

[1. Dữ liệu về luật đất đai 5](#_Toc110604432)

[2. Dữ liệu về ô nhiễm môi trường 5](#_Toc110604433)

[3. Dữ liệu về ô nhiễm không khí 5](#_Toc110604434)

# 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

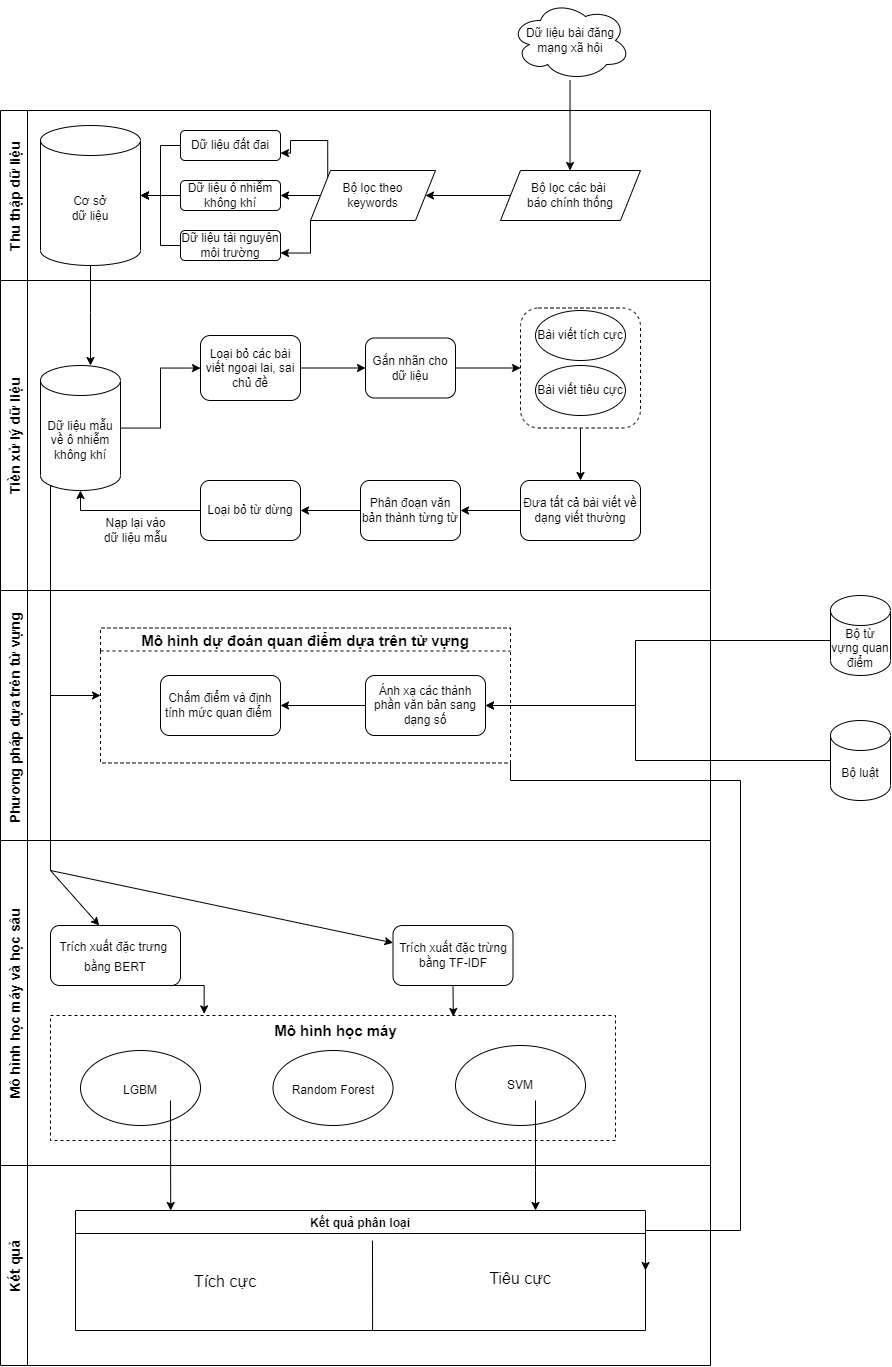
## Cấu trúc tài liệu

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:



Hình 1: Mô tả khái quá tchu trình xử lý

# Tổng quan các nguồn dữ liệu

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.

## Dữ liệu về luật đất đai

Tập dữ liệu về luật đất đai được tổng hợp từ 169.590 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, sử dụng đất,…

## Dữ liệu về ô nhiễm môi trường

Tập dữ liệu về ô nhiễm môi trường được tổng hợp từ 169.513 bài báo, sử dụng bộ lọc với các keyword: ô nhiễm, khói bụi, bụi mịn, bụi bẩn, vứt rác, nước thải, đen ngòm, xả rác, xả thải, chất thải,…

## Dữ liệu về ô nhiễm không khí

Tập dữ liệu này là kết quả thu được khi tinh chỉnh lại bộ lọc keyword của đề tài ô nhiễm môi trường khi các từ khóa chỉ chủ yếu xoay quanh các vấn đề về chất lượng không khí cũng như khí thải. Tập này chứa nội dung của 146.516 bài báo.

Từ yêu cầu của dự án, do cần thử nghiệm ở chủ đề nhỏ nhằm kiểm chứng khả năng triển khai và ứng dụng của giải pháp, tập dữ liệu chính được sử dụng trong quá trình xây dựng và huấn luyện mô hình sẽ là tập dữ liệu về ô nhiễm không khí.



Hình 2: Một số bản ghi mẫu từ tập dữ liệu ô nhiễm môi trường

Từ tập dữ liệu này, bước đầu sẽ chọn ra một tập con xấp xỉ 2000 bài báo, sau đó 2000 bài này sẽ được đọc, đánh giá thủ công và gắn nhãn “1” cho bài viết tích cực, “-1” cho bài viết tiêu cực. Đây là công đoạn chuẩn bị tập dữ liệu chuyên dụng để huấn luyện và kiểm thử khả năng vận hành cũng như độ chính xác của các mô hình được sử dụng về sau. Ngoài ra quá trình này còn lọc bớt đi được những bài viết có nội dung không liên quan đến chủ đề chúng ta đang xét đến (ô nhiễm không khí)