

**BỘ CÔNG THƯƠNG**  
**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP THỰC PHẨM TP. HCM**  
**KHOA CÔNG NGHỆ THÔNG TIN**

---



## **KHÓA LUẬN TỐT NGHIỆP**

### **HỆ THỐNG THU THẬP, PHÂN LOẠI, ĐÁNH GIÁ TIN TỨC LIÊN QUAN ĐẾN AN NINH QUỐC GIA – TRẬT TỰ AN TOÀN XÃ HỘI**

**GIÁO VIÊN HƯỚNG DẪN:**

ThS. Nguyễn Văn Tùng

ThS. Trần Đỗ Duy Quang

**SINH VIÊN THỰC HIỆN:**

Bùi Nguyễn Quang Hải                      2001170045 08DHTH2

Huỳnh Thanh Võ Hoàng Quân              2001170139 08DHTH1

TP. Hồ Chí Minh, tháng 01/2021

**BỘ CÔNG THƯƠNG**  
**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP THỰC PHẨM TP. HCM**  
**KHOA CÔNG NGHỆ THÔNG TIN**

---



## **KHÓA LUẬN TỐT NGHIỆP**

### **HỆ THỐNG THU THẬP, PHÂN LOẠI, ĐÁNH GIÁ TIN TỨC LIÊN QUAN ĐẾN AN NINH QUỐC GIA – TRẬT TỰ AN TOÀN XÃ HỘI**

**GIÁO VIÊN HƯỚNG DẪN:**

ThS. Nguyễn Văn Tùng

ThS. Trần Đỗ Duy Quang

**SINH VIÊN THỰC HIỆN:**

Bùi Nguyễn Quang Hải                      2001170045 08DHTH2

Huỳnh Thanh Võ Hoàng Quân              2001170139 08DHTH1

TP. Hồ Chí Minh, tháng 01/2021

## NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Tp.HCM, ngày tháng năm 2021

**Người nhận xét**

## NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Tp.HCM, ngày tháng năm 2021

**Người nhận xét**

## **LỜI CAM ĐOAN**

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi. Các số liệu, kết quả nêu trong đề án là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Tôi xin cam đoan rằng mọi sự giúp đỡ cho việc thực hiện khóa luận tốt nghiệp này đã được cảm ơn và các thông tin trích dẫn trong khóa luận tốt nghiệp đã được chỉ rõ nguồn gốc.

**Sinh viên thực hiện đề án**

Huỳnh Thanh Võ Hoàng Quân      Bùi Nguyễn Quang Hải

## LỜI CẢM ƠN

Trước tiên với tình cảm sâu sắc và chân thành nhất, cho phép chúng em được bày tỏ lòng biết ơn đến tất cả các cá nhân và tổ chức đã tạo điều kiện hỗ trợ, giúp đỡ chúng em trong toàn bộ quá trình tìm hiểu và nghiên cứu đề tài. Trong suốt thời gian từ khi bắt đầu học tập tại trường đến nay, chúng em đã nhận được rất nhiều sự quan tâm, giúp đỡ của quý thầy cô và bạn bè. Với lòng biết ơn sâu sắc nhất, chúng em xin gửi đến quý thầy cô ở khoa Công nghệ Thông tin đã truyền đạt vốn kiến thức quý báu cho chúng em trong suốt thời gian học tập tại trường. Đặc biệt là thầy Nguyễn Văn Tùng và anh Trần Đỗ Duy Quang đã hỗ trợ giúp đỡ chúng em trong quá trình thực hiện khóa luận.

Vì vốn kiến thức của bản thân còn hạn chế, trong quá trình thực hiện và hoàn thiện khóa luận tốt nghiệp này chúng em không tránh khỏi những sai sót, kính mong nhận được những ý kiến đóng góp từ thầy cô khoa Công nghệ Thông tin. Chúng em xin chân thành cảm ơn.

**Sinh viên thực hiện Đồ án**

Huỳnh Thanh Võ Hoàng Quân      Bùi Nguyễn Quang Hải

## TÓM TẮT

Hệ thống thu thập, phân loại, đánh giá tin tức liên quan đến An ninh quốc gia, An toàn trật tự xã hội là một hệ thống được xây dựng để hiện thực hóa vấn đề sau:

Dữ liệu đầu vào một tin tức dạng văn bản, kết quả trả về là nhãn của tin tức đó có thuộc về chủ đề An ninh quốc gia hoặc An toàn trật tự xã hội (gọi ngắn gọn là An ninh trật tự) hay không hay là thuộc một chủ đề khác, nếu tin tức đó thuộc vào một trong hai chủ đề trên thì mức độ của nó là tích cực hay tiêu cực?

Triển khai một công cụ có chức năng thu thập tin tức tự động. Tin tức sau khi được phân loại – đánh giá sẽ được lưu trữ trên cơ sở dữ liệu. Hệ thống quản lý tin tức được xây dựng trên website. Các thành phần trong hệ thống được viết dưới dạng mô-đun và liên kết với nhau thông qua WebAPI.

Hệ thống sau khi xây dựng sẽ được triển khai thực nghiệm ngoài thực tế với mục đích đóng góp giải pháp về việc sớm ghi nhận, phân loại các nguồn tin có ảnh hưởng xấu đến an nguy của Đất nước, An ninh trật tự xã hội.

# MỤC LỤC

<b>CHƯƠNG 1.TỔNG QUAN.....</b>	<b>1</b>
1.1. Thực trạng.....	1
1.2. Mục tiêu của đề tài.....	3
1.2.1. Đối tượng nghiên cứu.....	4
1.2.2. Phạm vi đề tài.....	5
1.3. Ý nghĩa khoa học và thực tiễn.....	5
1.3.1. Ý nghĩa khoa học.....	5
1.4. Đối tượng nghiên cứu và phạm vi đề tài.....	5
1.4.1. Thực tiễn.....	5
<b>CHƯƠNG 2.QUÁ TRÌNH TÌM HIỂU VÀ PHƯƠNG PHÁP NGHIÊN CỨU ..</b>	<b>6</b>
2.1. Bài toán thu thập tin tức tự động từ trang mạng xã hội Facebook.....	6
2.1.1. Thu thập tin tức tự động (quét dữ liệu tự động).....	6
2.1.2. Các phương pháp thu thập tin tức tự động.....	7
2.2. Bài toán phân loại, đánh giá tin tức.....	9
2.2.1. Khái niệm về phân loại văn bản.....	9
2.2.2. Thuật toán và các bộ thư viện hỗ trợ.....	10
2.2.3. Quy trình phân loại, đánh giá tin tức.....	12
2.3. Bài toán xây dựng hệ thống quản lý tin tức.....	16
2.3.1. Yêu cầu về chức năng của hệ thống.....	16
2.3.2. Lên kế hoạch xây dựng hệ thống.....	16
2.3.3. Công cụ triển khai.....	17
2.3.4. Ngôn ngữ triển khai.....	17
2.3.5. Mô hình và thư viện sử dụng.....	19
<b>CHƯƠNG 3.XÂY DỰNG VÀ TRIỂN KHAI HỆ THỐNG.....</b>	<b>22</b>
3.1. Cài đặt công cụ triển khai.....	22
3.1.1. Cài đặt Python.....	22
3.1.2. Cài đặt Pycharm.....	24
3.1.3. Cài đặt Postman.....	28
3.2. Cơ sở dữ liệu lưu trữ tin tức.....	29



3.2.1. Cơ sở dữ liệu .....	29
3.2.2. Mô tả cơ sở dữ liệu.....	30
3.3. Xây dựng tập huấn luyện.....	32
3.4. Huấn luyện mô hình phân loại, đánh giá tin tức.....	34
3.5. Công cụ thu thập, phân loại, đánh giá tin tức tự động.....	36
3.5.1. Thu thập tin tức tự động .....	36
3.5.2. Công cụ tách từ .....	40
3.5.3. Công cụ phân loại văn bản .....	43
3.6. WebAPI .....	44
3.6.1. Cấu trúc WebAPI .....	44
3.6.2. Kiểm tra API với Postman .....	48
3.7. Xây dựng Website quản lý, thống kê, báo cáo về tin tức .....	51
3.7.1. Giao diện đăng nhập.....	51
3.7.2. Giao diện trang chủ .....	51
3.7.3. Giao diện danh sách theo dõi .....	52
3.7.4. Giao diện danh sách bài viết .....	55
3.7.5. Giao diện thống kê, đánh giá.....	59
<b>CHƯƠNG 4.KẾT LUẬN.....</b>	<b>61</b>
4.1. Kết quả đạt được.....	61
4.2. Ưu điểm .....	61
4.3. Hạn chế .....	61
4.4. Hướng phát triển.....	62

## DANH MỤC CÁC TỪ VIẾT TẮT

Từ viết tắt	Tiếng Anh	Tiếng Việt
API	Application Programming Interface	Giao diện lập trình ứng dụng
CSS	Cascading Style Sheet language	Ngôn ngữ định kiểu theo tầng
HTML	Hypertext Markup Language	Ngôn ngữ đánh dấu siêu văn bản
HTTP	HyperText Transfer Protocol	Giao thức truyền tải siêu văn bản
ID	Identify	Mã định danh
MVC	Model – View – Controller	
URL	Uniform Resource Locator	Đường dẫn tham chiếu đến tài nguyên trên Internet.

## DANH MỤC CÁC BẢNG

Bảng 3.1 Account (Tài khoản hệ thống) .....	30
Bảng 3.2 FacebookType (Nguồn đăng tin trên Facebook) .....	30
Bảng 3.3 WatchList (Danh theo dõi) .....	31
Bảng 3.4 NewsLabel (Các nhãn tin tức) .....	31
Bảng 3.5 NewsLabel (Các nhãn tin tức) .....	31
Bảng 3.6 Post (Bài viết) .....	32
Bảng 3.7 Cấu hình máy tính.....	34
Bảng 3.8 Thống kê độ chính xác của từng nhãn .....	34
Bảng 3.9 Thống kê độ chính xác của từng nhãn .....	35

## DANH MỤC CÁC HÌNH ẢNH

Hình 2. 1 Sơ đồ huấn luyện.....	15
Hình 2. 2 Sơ đồ dự đoán .....	15
Hình 2. 3 Mô hình triển khai hệ thống.....	17
Hình 2. 4 Ví dụ thư viện Tkinter.....	19
Hình 2. 5 Mô hình MVC.....	20
Hình 3. 1 Cài đặt Python bước 1.....	22
Hình 3. 2 Cài đặt Python bước 2.....	22
Hình 3. 3 Cài đặt Python bước 3.....	23
Hình 3. 4 Cài đặt Python bước 4.....	23
Hình 3. 5 Cài đặt Python bước 5.....	24
Hình 3. 6 Cài đặt Pycharm bước 1 .....	24
Hình 3. 7 Cài đặt Pycharm bước 2 .....	25
Hình 3. 8 Cài đặt Pycharm bước 3 .....	25
Hình 3. 9 Cài đặt Pycharm bước 4 .....	26
Hình 3. 10 Cài đặt Pycharm bước 5.....	26
Hình 3. 11 Cài đặt Pycharm bước 6.....	27
Hình 3. 12 Cài đặt Pycharm bước 7 .....	27
Hình 3. 13 Quá trình cài đặt Postman bước 1 .....	28
Hình 3. 14 Quá trình cài đặt Postman bước 2 .....	28
Hình 3. 15 Quá trình cài đặt Postman bước 3 .....	29
Hình 3. 16 Sơ đồ diagram .....	30
Hình 3. 17 Giao diện chính của công cụ.....	36
Hình 3. 18 Khu vực nhập URL .....	36
Hình 3. 19 Danh sách theo dõi .....	37
Hình 3. 20 Tùy chọn loại trang và đăng nhập.....	37
Hình 3. 21 Giao diện đăng nhập.....	37
Hình 3. 22 Số lần cuộn trang.....	38

Hình 3. 23 Khu vực chức năng .....	38
Hình 3. 24 Xem nhật ký .....	38
Hình 3. 25 Khu vực thông báo .....	39
Hình 3. 26 Hộp thoại thông báo .....	39
Hình 3. 27 Giao diện thêm đối tượng.....	39
Hình 3. 28 Giao diện thông báo hủy .....	40
Hình 3. 29 Công cụ tách từ .....	40
Hình 3. 30 Khu vực nhận dữ liệu đầu vào .....	41
Hình 3. 31 Khu vực hiện thị kết quả .....	42
Hình 3. 32 Công cụ phân loại văn bản.....	43
Hình 3. 33 Kết quả phân loại .....	43
Hình 3. 34 Giao diện lưu bài viết.....	44
Hình 3. 35 Sơ đồ ánh xạ cơ sở dữ liệu .....	44
Hình 3. 36 Thao tác truy vấn với Postman.....	48
Hình 3. 37 Thao tác thêm với Postman.....	49
Hình 3. 38 Thao tác cập nhật với PostMan .....	50
Hình 3. 39 Thao tác xóa với Postman .....	50
Hình 3. 40 Giao diện đăng nhập.....	51
Hình 3. 41 Giao diện trang chủ .....	51
Hình 3. 42 Giao diện trang theo dõi.....	52
Hình 3. 43 Giao diện trang thêm đối tượng mới.....	53
Hình 3. 44 Giao diện thao tác trạng thái .....	53
Hình 3. 45 Chức năng lọc .....	54
Hình 3. 46 Chức năng tìm kiếm .....	54
Hình 3. 47 Chức năng sửa thông tin .....	55
Hình 3. 48 Trang quản lý bài viết .....	55
Hình 3. 49 Hiện thị số lượng.....	56
Hình 3. 50 Lọc bài viết.....	56
Hình 3. 51 Tìm kiếm bài viết .....	57

Hình 3. 52 Thông tin bài viết .....	57
Hình 3. 53 Xóa bài viết .....	58
Hình 3. 54 Xuất file excel .....	58
Hình 3. 55 Thông tin file excel .....	59
Hình 3. 56 Trang thống kê, đánh giá.....	59
Hình 3. 57 Khu vực thống kê chung .....	60
Hình 3. 58 Khu vực thống kê chi tiết .....	60

# CHƯƠNG 1. TỔNG QUAN

## 1.1. Thực trạng

Những năm gần đây, mạng xã hội đã có bước phát triển mạnh mẽ, tác động lớn đến đời sống xã hội ở hầu hết các quốc gia trên thế giới, trong đó có Việt Nam. Mạng xã hội cung cấp những tính năng đa dạng cho phép người dùng kết nối, chia sẻ, tiếp nhận thông tin một cách nhanh chóng, hiệu quả. Bên cạnh đó đây là nơi cung cấp tin tức, kiến thức về tất cả các lĩnh vực của đời sống xã hội. Qua đó giúp con người có thể nắm bắt được các xu thế của đời sống phục vụ cho công việc và cuộc sống của mình.

Mạng xã hội ra đời mang lại những lợi ích to lớn có thể kể đến như góp phần thúc đẩy quá trình hội nhập quốc tế trên lĩnh vực văn hóa của Việt Nam. Nhất là mạng xã hội xuyên quốc gia như Facebook, Youtube đã tạo ra những cơ hội, khả năng tiếp xúc, giao lưu văn hóa, hiểu biết lẫn nhau giữa dân tộc Việt Nam với các dân tộc khác trên thế giới. Thông qua nó, thế giới biết đến Việt Nam hơn như một dân tộc yêu chuộng hòa bình, tôn trọng công lý, năng động với một kho tàng các giá trị văn hóa phong phú, đầy bản sắc.

Tuy nhiên, bên cạnh mặt tích cực, mạng xã hội cũng đang tồn tại không ít những yếu tố tiêu cực, ảnh hưởng trực tiếp đến môi trường xã hội, lợi ích cộng đồng, An ninh quốc gia và An ninh trật tự. Điển hình có thể kể đến việc trong những năm qua các thế lực thù địch, phản động đã lập ra và sử dụng hàng ngàn trang mạng xã hội vào các hoạt động tuyên truyền, chống phá Đảng, Nhà nước. Nội dung tập trung về xuyên tạc, nói xấu chủ nghĩa Mác- Lênin, tư tưởng Hồ Chí Minh, vai trò lãnh đạo của Đảng, thường xuyên đăng tải những bài viết với lời lẽ chống đối chế độ, kích động thù địch núp bóng dưới các hình thức dân chủ, nhân quyền khiến cho một số bộ phận người dân bị ngộ độc về tin tức dẫn đến hình thành tâm lý phản kháng, tư tưởng bất mãn, chống đối, tiến tới kêu gọi biểu tình, bạo loạn lật đổ chế độ.

Sau khi được 87% đại biểu Quốc hội bỏ phiếu thông qua vào ngày 12/6/2018, kể từ ngày 1/1/2019, Luật An ninh mạng bắt đầu chính thức có hiệu lực. Việc ban hành Luật An ninh mạng đã góp phần phòng ngừa, đấu tranh làm thất bại các hoạt động sử dụng không gian mạng xâm phạm đến An ninh quốc gia, chống Nhà nước, tuyên truyền phá hoại tư tưởng, phá hoại khối đại đoàn kết dân tộc, kích động biểu tình, phá rối an ninh trên không gian mạng của các thế lực thù địch. Kể từ ngày Luật An ninh mạng được thông qua, nó đã trở thành cơ sở pháp lý để các cơ quan chức năng xử lý những trường hợp vi phạm pháp luật trên không gian mạng. Không ít các trường hợp vi phạm đã được cơ quan chức năng lập biên bản xử lý, thậm chí khởi tố hình sự đối với những trường hợp gây ảnh hưởng nghiêm trọng đến An ninh quốc gia.

Tuy nhiên, mạng xã hội là một môi trường không gian mạng tương đối lớn, mỗi ngày có hàng ngàn bài viết được đăng tải lên, việc kiểm soát tất cả nội dung gặp rất nhiều khó khăn. Chính vì vậy, đề tài này ra đời với mục tiêu đưa ra giải pháp hỗ trợ cho phía cơ quan chức năng có thể thuận tiện kiểm soát, phân tích, lưu trữ những bài viết nhằm phục vụ cho công tác theo dõi các đối tượng tình nghi trên không gian mạng để từ đó phục vụ công tác điều tra.



## 1.2. Mục tiêu của đề tài

Mục tiêu chính đặt ra là nghiên cứu các phương pháp về thu thập và phân tích dữ liệu dưới dạng văn bản từ đó hiện thực hệ thống về thu thập, phân loại và đánh giá tin tức trên trang mạng xã hội Facebook.

### ***Thu thập những tin tức trên Facebook dựa vào URL:***

- Nguồn thu thập từ Trang (Fanpage), Nhóm (Group), Trang cá nhân (Profile).
- Thu thập tự động 1 hoặc 1 danh sách các URL.
- Thu thập thủ công 1 bài viết bất kì.

### ***Phân loại tin tức:***

- Theo chủ đề An ninh quốc gia.
- Theo chủ đề An ninh trật tự.
- Chủ đề khác.

### ***Đánh giá tin tức dựa trên chủ đề***

- Đánh giá theo mức độ tích cực, tiêu cực đối với 2 chủ đề An ninh quốc gia, An ninh trật tự.
- Đánh giá mức độ bình thường đối với các chủ đề khác.

### ***Quản lý đối tượng đăng tin***

- Xem danh sách các đối tượng trong diện theo dõi.
- Xem thông tin chi tiết một đối tượng.
- Thêm một đối tượng mới vào diện theo dõi.
- Chỉnh sửa thông tin đối tượng.

### ***Quản lý tin tức***

- Xem danh sách các tin tức đã thu thập được.
- Xem nội dung chi tiết của một tin tức.
- Bộ lọc các tin tức dựa theo chủ đề, mức độ, thời gian đăng tin. Tìm kiếm.
- Xóa các tin tức không cần thiết cho việc lưu trữ.
- Xuất các tin tức ra file excel.

### ***Thống kê***

- Thống kê số lượng các đối tượng đang theo dõi.
- Thống kê tổng số lượng tin tức đã thu thập.
- Thống kê số lượng tin tức dựa theo mức độ.
- Thống kê xu hướng đăng tin trên từng đối tượng cụ thể.

#### **1.2.1. Đối tượng nghiên cứu**

Nghiên cứu về việc thu thập, phân loại, đánh giá tin tức trên trang mạng xã hội Facebook liên quan đến An ninh quốc gia, An ninh trật tự trong đó:

Chủ đề An ninh quốc gia là những tin tức, bài viết có xu hướng nói về Đảng, Nhà nước, chế độ, về độc lập chủ quyền lãnh thổ, về cán bộ cũng như các vị lãnh tụ. Vì vậy một tin tức được xem là tiêu cực khi có nội dung phản động, kích động thù địch, gây chia rẽ khối đại đoàn kết dân tộc, kích động bạo loạn, chống phá Nhà nước, nói xấu, bôi nhọ về cán bộ quản lý Nhà nước, về các vị lãnh tụ Đặc biệt là những nguồn tin từ các tổ chức phản động, các cá nhân bất mãn với chế độ đang ngày đêm gieo rắc những tư tưởng xấu đến quần chúng nhân dân trên các mạng xã hội. Ngược lại những tin được xem tích cực khi nó mang tính tuyên truyền, nêu cao vai trò lãnh đạo của Đảng và Nhà nước, những thành tựu mà Đảng đã đạt được trong quá trình xây dựng, phát triển đất nước, nêu cao tinh thần học tập và làm theo tấm gương đạo đức của Chủ tịch Hồ Chí Minh.

Về phần chủ đề An ninh trật tự, đây là một chủ đề nhỏ hơn nói về các vấn đề an ninh và trật tự xã hội ở trong nước bao gồm các vấn đề về tội phạm, các vụ án hình sự hoặc dân sự, các vấn đề về vi phạm pháp luật của các cơ quan, tổ chức, cá nhân. Những tin tức tích cực khi nói về các vụ án đã được giải quyết, đã bắt được nghi can hoặc khởi tố các đối tượng phạm tội. Những tin tức tiêu cực khi đề cập đến các vấn đề gây rối an ninh trật tự, ảnh hưởng đến tính mạng, sức khỏe của công dân, gây thiệt hại về kinh tế, tài sản, các vụ án không được giải quyết.

### 1.2.2. Phạm vi đề tài

Đề tài được thực hiện trong phạm vi giới hạn trên mạng xã hội Facebook. Sau khi hoàn tất quá trình nghiên cứu và xây dựng các mô-đun chức năng, hệ thống sẽ được triển khai chạy thử nghiệm trên máy chủ nội bộ của Công an tỉnh Bình Thuận.

## 1.3. Ý nghĩa khoa học và thực tiễn

### 1.3.1. Ý nghĩa khoa học

Với việc lĩnh vực công nghệ thông tin phát triển nhanh chóng như hiện và song song với đó là sự phát triển của cuộc cách mạng công nghiệp 4.0, việc tin học hóa các công việc trong đời sống đang là vấn đề được rất nhiều các cơ quan, tổ chức trên thế giới nói chung và tại Việt Nam nói riêng quan tâm đến. Hướng ứng xu thế đó, đề tài nghiên cứu này góp phần mở ra một giải pháp, hướng đi mới trong việc ứng dụng công nghệ thông tin vào trong hoạt động công tác của cơ quan chức năng về việc xử lý đối với các nguồn tin tức trên mạng xã hội Facebook.

## 1.4. Đối tượng nghiên cứu và phạm vi đề tài

### 1.4.1. Thực tiễn

Khi áp dụng vào thực tế, hệ thống sẽ hỗ trợ cơ quan chức năng trong việc nhanh chóng thu thập, phân loại được nguồn tin tức từ các đối tượng trong danh sách theo dõi, giúp giảm thiểu thời gian so với việc phải phân loại thủ công. Ngoài ra hệ thống còn cung cấp chức năng thống kê, lưu trữ giúp cho việc truy vết lại những nội dung tin tức trong trường hợp đối tượng cố tình gỡ bỏ bài viết đã đăng từ đó làm bằng chứng để xử lý vi phạm (nếu có). Qua đó, việc phát hiện sớm các nguồn tin tức tiêu cực phần nào giúp cho cơ quan chức năng ngăn chặn các hành vi vi phạm pháp luật có nguy cơ xảy ra, góp phần ổn định tình hình An ninh quốc gia, An ninh trật tự.

## **CHƯƠNG 2. QUÁ TRÌNH TÌM HIỂU VÀ PHƯƠNG PHÁP NGHIÊN CỨU**

Để phục vụ cho một bài toán lớn là việc xây dựng lên một hệ thống thu thập, phân tích và đánh giá tin tức liên quan đến chủ đề An ninh quốc gia - Trật tự an toàn xã hội. Chiến lược nghiên cứu đưa ra đó chính là chia nhỏ bài toán lớn ấy ra thành những bài toán nhỏ hơn để giải quyết vấn đề theo từng chức năng. Cụ thể được chia thành 3 bài toán con như sau:

- Bài toán thu thập tin tức tự động từ trang mạng xã hội Facebook
- Bài toán phân loại và đánh giá những tin tức đã thu thập được
- Xây dựng hệ thống có chức năng lưu trữ, quản lý, thống kê và báo cáo về những tin tức đã được phân loại.

Như vậy, việc phân chia một bài toán lớn thành các module (mô-đun) nhỏ hơn giúp cho việc nghiên cứu trở lên tường minh và cụ thể. Trong đó ta có thể thấy output (dữ liệu đầu ra) của bài toán con thứ nhất chính là input (dữ liệu đầu vào) của bài toán con thứ hai và tiếp tục như thế, dữ liệu của bài toán thứ hai sau khi được giải quyết sẽ trở thành dữ liệu để triển khai bài toán thứ ba. Cụ thể các phương pháp nghiên cứu sẽ được trình bày một cách chi tiết như sau:

### **2.1. Bài toán thu thập tin tức tự động từ trang mạng xã hội Facebook**

#### **2.1.1. Thu thập tin tức tự động (quét dữ liệu tự động)**

Hiện nay, nhu cầu về việc thu thập những tin tức, bài báo từ trên internet là rất nhiều. Nhu cầu này chủ yếu đến từ các doanh nghiệp, công ty và các tổ chức về tài chính. Mục đích sau khi thu thập dữ liệu là để phân tích về kinh doanh, nhu cầu của người dùng, phân tích về xu hướng trên thị trường. Nắm bắt được nhu cầu đó, rất nhiều những công cụ thu thập tin tức đã ra đời trong đó có thể kể đến như Cisco, Reputa. Nhưng hiện tại các công cụ như thế vẫn đang hoạt động chính ở trên các website của các trang báo điện tử và giá thành để thuê dịch vụ cũng tương đối lớn. Vậy thì thu thập tin tức tự động nghĩa là gì?

Thu thập tin tức tự động (quét dữ liệu tự động) là một thuật ngữ trong ngành Công nghệ thông tin, đây được xem là một quá trình khá quan trọng trong việc thu thập và lấy dữ liệu từ một trang web rồi phân tích dữ liệu lấy được và bóc tách thông tin dữ liệu theo yêu cầu đã được đặt ra. Có thể hiểu như thế này, để lấy thông tin từ một trang web nào đó thì phần quét dữ liệu sẽ cung cấp chức năng lấy được dữ liệu khi đang kết nối internet, sau đó tất cả dữ liệu thu thập được sẽ được lưu vào database một cách tự động mà không phải thông qua bất cứ một chi tiết nhập liệu nhập liệu nào cả - phương pháp này được coi là ưu điểm của giúp tiết kiệm thời gian cho người sử dụng. Nói một cách dễ hiểu, quét dữ liệu là phương pháp giúp chúng ta lấy thông tin khi biết đường link của website mà chúng ta muốn lấy thông tin, phần còn lại sẽ diễn ra một cách tự động.

### 2.1.2. Các phương pháp thu thập tin tức tự động

Đối với các kĩ thuật được sử dụng để thu thập dữ liệu tự động, có một vấn đề chung được đặt ra đó chính là các website chúng ta muốn lấy thông tin đều không có API kết nối trực tiếp để lấy dữ liệu từ database của website đó. Vì vậy cơ chế duy nhất đó chính là phân tích cấu trúc code, cấu trúc HTML của website đó để lấy được dữ liệu mong muốn.

Về phương pháp triển khai, mục tiêu đặt ra của đề tài là thu thập được tin tức được đăng trên các trang mạng xã hội (cụ thể là mạng xã hội Facebook). Để thu thập tin tức từ Facebook, chúng ta có 3 nguồn tin tức chính cần lấy về: Trang (Fanpage), Nhóm (Group) và Trang cá nhân (Profile). Trong 3 nguồn trên, việc thu thập tin tức từ các nhóm kín hoặc trang cá nhân bắt buộc phải thông qua bước đăng nhập vào Facebook.

Bằng cách sử dụng facebook-scraper (bộ thư viện hỗ trợ thu thập dữ liệu dựa trên ngôn ngữ lập trình Python). Việc thu thập tin tức trở nên thuận tiện hơn khi chúng ta có thể tận dụng được những hàm xây dựng sẵn có nhiệm vụ phân tích cấu trúc code HTML của trang web facebook để từ đó lấy về những dữ liệu mong muốn. Lấy một ví dụ cụ thể khi thu thập dữ liệu từ facebook cụ thể là một Trang

(Fanpage) chúng ta cần phải lấy về những thông tin sau:

- Link dẫn trực tiếp đến một bài đăng cụ thể của một trang. Cấu trúc có dạng như sau:  
`https://www.facebook.com/viettan/posts/10160993283250620`  
 trong đó `https://www.facebook.com` là tên miền của Facebook, `viettan` là id của trang Việt Tân.
- ID của bài viết, cụ thể là dãy số `10160993283250620` trên url
- Link dẫn trực tiếp đến đối tượng đã đăng bài viết
- Mốc thời gian cụ thể mà bài viết xuất hiện trên Facebook
- Mốc thời gian khi bài viết được thu thập về
- Nội dung của bài viết
- Hình ảnh đính kèm trên bài viết (nếu có)
- Tương tác của bài viết (thể hiện qua số lượt thích, bình luận, chia sẻ)

Tuy nhiên bộ thư viện này cũng có rất nhiều điểm hạn chế, điểm đầu tiên có thể kể đến là khả năng bóc tách cấu trúc trang web vẫn chưa thực sự tốt dẫn đến việc những dữ liệu thu thập về đôi khi sẽ bị thiếu hụt một vài trường dữ liệu. Ngoài ra điểm bất cập lớn chính là chưa hỗ trợ phân tích cấu trúc của Trang cá nhân, đó đó để lấy được dữ liệu từ nguồn này giải pháp đưa ra đó là sử dụng Selenium – một bộ công cụ kiểm thử có chức năng thao tác tự động trên trang web.

Áp dụng công cụ Selenium, việc thu thập tin tức trên Trang cá nhân diễn ra với quy trình tương tự như khi thu thập tin tức trên Trang và Nhóm. Khác nhau ở chỗ, thay vì sử dụng bộ thư viện facebook-scaper để hỗ trợ trong việc phân tích, bóc tách cấu trúc trang web thì nay công việc này được triển khai một cách thủ công. Quá trình triển khai một cách ngắn gọn như sau:

**Bước 1:** Trong công cụ thu thập dữ liệu, khởi động Selenium quét qua tất cả các bài viết của trang cá nhân cần thu thập tin tức.

**Bước 2:** Dựa vào cấu trúc code HTML của trang cá nhân trên Facebook, tìm ra những thẻ đặc trưng chứa những thông tin cần thu thập về như Url bài viết, Url

đối tượng đăng tin, nội dung bài viết, thời gian đăng tin, tương tác, v.v để trích xuất dữ liệu.

**Bước 3:** Dữ liệu sau khi trích xuất được lưu trữ để làm dữ liệu đầu vào cho mô-đun Phân loại – đánh giá tin tức.

### ***Ưu điểm***

- Lấy được thông tin từ trang web một cách nhanh chóng.
- Tiết kiệm thời gian, chi phí triển khai trong các dự án có nguồn dữ liệu từ các trang web.

### ***Hạn chế***

- Cấu trúc code HTML của trang web thường xuyên thay đổi (đặc biệt là Facebook) dẫn đến việc phải cấu hình lại phương pháp thu thập.
- Facebook được lập trình để phát hiện các hành vi thu thập dữ liệu trái phép trên hệ thống, gây khó khăn trong quá trình thực hiện.
- Không thể thu thập được trên những trang web có cấu trúc phức tạp, các thuộc tính trong thẻ HTML bị mã hóa.

## **2.2. Bài toán phân loại, đánh giá tin tức**

Tin tức chính là thông tin về các sự kiện hiện tại được cung cấp thông qua nhiều phương tiện khác nhau như: truyền miệng, in ấn, truyền hình, báo chí mà trong phạm vi nghiên cứu của đề tài, tin tức được thể hiện dưới dạng những văn bản ngắn, đoạn văn được đăng lên phương tiện truyền thông là mạng xã hội. Do đó để phân loại tin tức, phương án đề ra đó chính là áp dụng bài toán phân loại văn bản (Text classification) để giải quyết.

### **2.2.1. Khái niệm về phân loại văn bản**

Phân loại văn bản là một trong những bài toán phổ biến trong lĩnh vực xử lý ngôn ngữ tự nhiên (Nature language processing). Đây là bài toán thuộc nhóm học có giám sát trong học máy. Bài toán này yêu cầu đưa vào một tập huấn luyện đã được gán nhãn phân lớp sẵn (label). Mô hình sẽ học từ những dữ liệu đã được gán

nhân đó, sau đó dùng để dự đoán nhãn cho các dữ liệu mới mà mô hình chưa từng gặp. Lấy ví dụ, dữ liệu đầu vào là một tin tức đăng trên mạng xã hội Facebook, tin tức này sẽ được đưa vào mô hình học máy để dự đoán ra chủ đề (Chính trị, Thể thao, Âm nhạc v.v).

### 2.2.2. Thuật toán và các bộ thư viện hỗ trợ

#### *Thuật toán Naive Bayes*

Naive Bayes là một thuật toán dựa trên định lý Bayes về lý thuyết xác suất để đưa ra các dự đoán cũng như phân loại dữ liệu. Naive Bayes là một trong những thuật toán được ứng dụng rất nhiều trong các lĩnh vực học máy dùng để đưa các dự đoán chính xác nhất dựa trên một tập dữ liệu đã được thu thập với độ chính xác cao. Nó thuộc vào nhóm thuật toán học có hướng dẫn, tức là máy học từ các mẫu dữ liệu đã có.

Định lý Bayes được phát biểu như sau:

Định lý Bayes cho phép tính xác suất xảy ra của một sự kiện ngẫu nhiên A khi biết sự kiện liên quan B đã xảy ra. Xác suất này được ký hiệu là  $P(A|B)$ , và đọc là “xác suất của A nếu có B”. Đại lượng này được gọi xác suất có điều kiện hay xác suất hậu nghiệm vì nó được rút ra từ giá trị được cho của B hoặc phụ thuộc vào giá trị đó.

Gọi A, B là hai biến cố

Với  $P(B) > 0$ :

$$P(A | B) = \frac{P(AB)}{P(B)}$$

Suy ra:

$$P(AB) = P(A | B)P(B) = P(B | A)P(A)$$



Công thức Bayes:

$$\begin{aligned} P(B | A) &= \frac{P(AB)}{P(A)} = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A|B)P(B)}{P(AB)+P(A\bar{B})} \\ &= \frac{P(A|B)P(B)}{P(AB)+P(A\bar{B})} = \frac{P(A|B)P(B)}{P(A|B)P(B)+P(A|\bar{B})P(\bar{B})} \end{aligned}$$

Trong đó ta gọi A là một chứng cứ (evidence) (trong bài toán phân lớp A sẽ là một phần tử dữ liệu), B là một giả thiết nào để cho A thuộc về một lớp C nào đó. Trong bài toán phân lớp chúng ta muốn xác định giá trị  $P(B|A)$  là xác suất để giả thiết B là đúng với chứng cứ A thuộc vào lớp C với điều kiện ra đã biết các thông tin mô tả A.  $P(B|A)$  là một xác suất hậu nghiệm (posterior probability hay posteriori probability) của B với điều kiện A.

### ***Bộ thư viện Sklearn***

Sklearn là một bộ thư viện được xây dựng dành cho các thuật toán học máy được viết trên ngôn ngữ Python. Thư viện cung cấp một tập các công cụ xử lý các bài toán liên quan đến học máy và mô hình thống kê gồm: phân lớp, hồi quy, gom cụm. Một trong số những nhóm thuật toán nổi bật của bộ thư viện này là Supervised Models (học có giám sát) bao gồm mảng lớn các thuật toán học máy hiện nay trong đó thuật toán Naive Bayes.

### ***Bộ thư viện pyvi***

Pyvi là một bộ thư viện được xây dựng để hỗ trợ việc tiền xử lý văn bản trong lĩnh vực xử lý ngôn ngữ tự nhiên. Thư viện cung cấp các chức năng tiền xử lý ngôn ngữ tiếng Việt, nổi bật trong đó là chức năng tách từ tiếng Việt. Pyvi cung cấp một bộ từ điển tiếng Việt với số lượng lên đến 74 ngàn từ đảm bảo cho việc tách từ chính xác, đúng nghĩa.

### 2.2.3. Quy trình phân loại, đánh giá tin tức

Để giải quyết được bài toán phân loại chúng ta cần phải xây dựng một tập huấn luyện và mô hình huấn luyện, cụ thể như sau:

#### ***Xây dựng tập huấn luyện***

Dữ liệu là một trong những yếu tố quan trọng nhất trong việc huấn luyện một mô hình học máy. Trong quá trình xây dựng một hệ thống phân loại văn bản, các bước chuẩn bị và tiền xử lý dữ liệu có ảnh hưởng lớn đến độ chính xác của mô hình và quyết định tới thành bại của hệ thống.

Các bước để xây dựng tập huấn luyện:

#### **Bước 1: Xác định số lượng các chủ đề cần phân loại**

Đối với phạm vi đề tài, ngoài 2 chủ đề bắt buộc là An ninh quốc gia và An ninh trật tự thì hiện nay còn có các chủ đề phổ biến thường xuất hiện trên mạng xã hội Facebook. Do đó tập huấn luyện sẽ bao gồm thêm 10 chủ đề sau: Thể thao, Âm nhạc, Công nghệ, Thời trang, Giáo dục, Kinh doanh, Phim ảnh, Xe cộ, Ẩm thực, Sức khỏe. Tổng cộng là 12 chủ đề.

#### **Bước 2: Xác định số lượng nhãn dựa trên chủ đề**

Với 2 chủ đề An ninh quốc gia và An ninh trật tự, mỗi chủ đề bao gồm 2 mức độ tích cực và tiêu cực. Các chủ đề còn lại thuộc mức độ bình thường. Tổng cộng tập huấn luyện bao gồm 14 nhãn để phân loại.

#### **Bước 3: Tiền xử lý dữ liệu**

Từ những tin tức đã thu thập về trước đó, tiến hành tiền xử lý dữ liệu. Mục đích của việc này là để làm sạch dữ liệu trước khi đưa vào mô hình huấn luyện. Nguyên nhân là do dữ liệu khi thu thập về thường không theo một cấu trúc cụ thể và đôi khi sẽ có lẫn các code HTML của trang web, để nguyên như vậy mà đem đi xử lý sẽ rất khó khăn và độ chính xác trả về sẽ không cao. Đối với tiền xử lý văn bản, các công việc cần thực hiện bao gồm:

- **Làm sạch dữ liệu:** Loại bỏ các thẻ HTML trong đoạn văn.  
Ví dụ: <h1>Xin chào</h1> => Xin chào
- **Chuẩn hóa bộ gõ Unicode:** Hiện nay, có 2 loại mã Unicode đang được sử dụng phổ biến là Unicode tổ hợp và Unicode dựng sẵn. Điều này dẫn đến một vấn đề là 2 từ giống nhau nhưng về cấu trúc thì khác nhau. Đó đó hướng xử lý là chuyển hóa tất cả về một chuẩn Unicode dựng sẵn (phổ biến hơn Unicode tổ hợp).
- **Chuẩn hóa kiểu gõ dấu:** Kiểu gõ dấu là cách chúng ta sử dụng các phím trên bàn phím máy tính để tượng trưng cho các dấu và tạo thành một từ tiếng Việt. Hiện có 4 kiểu gõ dấu chữ tiếng Việt thông dụng là VIQR, VNI, Telex và Microsoft. Điều này dẫn đến cách dấu câu xuất hiện một cách không đồng nhất ví dụ như òa với oà. Hướng xử lý là chuyển hóa tất cả về một kiểu gõ dấu phổ biến nhất.
- **Tách từ tiếng việt:** Đơn vị từ trong tiếng Việt bao gồm từ đơn và từ ghép. Nên phương pháp đề ra là phải xử lý đoạn văn để cho mô hình học máy biết đâu là từ đơn, đâu là từ ghép. Nếu thiếu bước này, mô hình sẽ hiểu tất cả đề là từ đơn dẫn đến kết quả sai. Để tách từ, sử dụng hàm tách từ trong bộ thư viện pyvi. Nguyên lý hoạt động của việc tách từ là dựa vào bảng từ điển tiếng Việt, lần lượt ghép nối các từ đơn trong câu, nếu 2 từ được ghép là một từ ghép thì thay thế khoảng trắng giữa hai từ bằng dấu gạch dưới.  
Ví dụ: Học sinh học sinh học => Học\_sinh học sinh\_học
- **Chuẩn hóa về chữ in thường:** Việc đưa văn bản về dạng chữ in thường có tác dụng giảm số lượng đặc trưng (vì máy tính hiểu chữ hoa và chữ thường là 2 chữ khác nhau)
- **Loại bỏ các ký tự không cần thiết:** Ký tự thừa là các ký tự xuất hiện nhưng không có tác dụng cho việc phân loại văn bản. Việc loại bỏ các ký tự thừa giúp tăng tốc độ học và tốc độ xử lý.
- **Loại bỏ Stopword:** Đây là những từ xuất hiện nhiều trong tập huấn luyện, tuy nhiên lại không mang nhiều ý nghĩa. Ở tiếng Việt, stopwords là những từ như:

và, của, là, có, nhưng, còn, ... Hướng xử lý là tạo một bộ đếm quét qua hết tập huấn luyện, ghi nhận lại các từ xuất hiện nhiều lần và tiến hành xóa bỏ.

#### **Bước 4: Gắn nhãn cho dữ liệu đã qua tiền xử lý**

Ở bước này tiến hành gắn các nhãn cho các tin tức dựa theo nội dung. Việc xác định nội dung và gắn nhãn đúng góp phần nâng cao độ chính xác khi đưa vào mô hình huấn luyện.

Đối với chủ đề An ninh quốc gia: các từ khóa để xác định nội dung là: bò đỏ, bò vàng, nhân quyền, dân chủ, biểu tình, bạo loạn, đánh bom, khủng bố, phản động, việt tân, triều đại việt, tham nhũng, hối lộ, cộng sản, đa đảng, việt cộng, tàu cộng, hán nô, lãnh đạo, chính quyền, thiên đường xhcn, v.v.

Đối với chủ đề An ninh trật tự: các từ khóa để xác định nội dung là: trộm, cướp, giết người, án mạng, đánh nhau, gây rối, hành hung, làm nhục, hiếp dâm, lừa đảo, chiếm đoạt, v.v

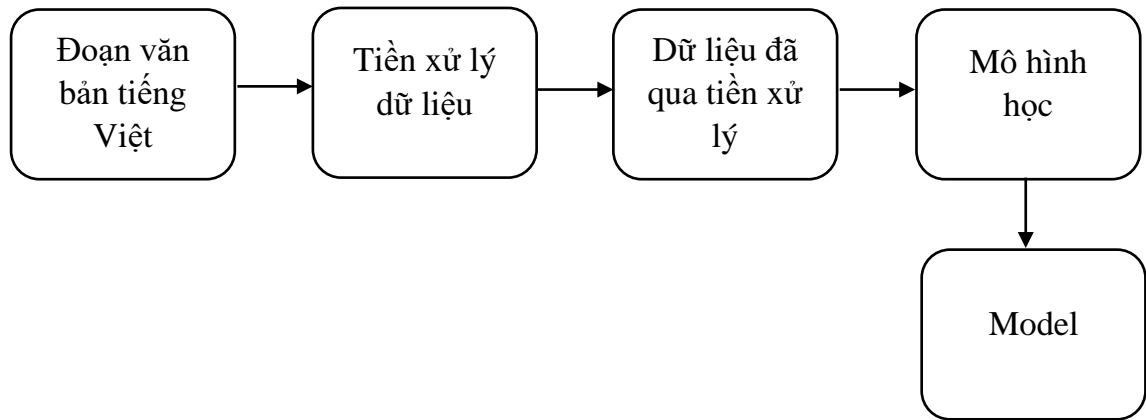
#### ***Xây dựng mô hình huấn luyện***

Việc triển khai mô hình huấn luyện phân loại văn bản bao gồm 2 giai đoạn:

##### **Giai đoạn 1: Huấn luyện**

Giai đoạn huấn luyện là giai đoạn mà mô hình sẽ học từ những dữ liệu đã được gắn nhãn trong tập huấn luyện đã được chuẩn bị trước. Dữ liệu văn bản sẽ được số hóa thông qua bộ trích xuất đặc trưng (feature extractor) để mỗi mẫu dữ liệu trong tập huấn luyện trở thành 1 véc-tơ nhiều chiều (đặc trưng). Thuật toán học máy sẽ học và tối ưu các tham số để đạt được kết quả tốt trên tập dữ liệu này. Nhãn của dữ liệu được dùng để đánh giá việc mô hình học có tốt không và dựa vào đó để tối ưu. Sau khi huấn luyện, kết quả thu được là một Model.

Sơ đồ huấn luyện được triển khai như sau:

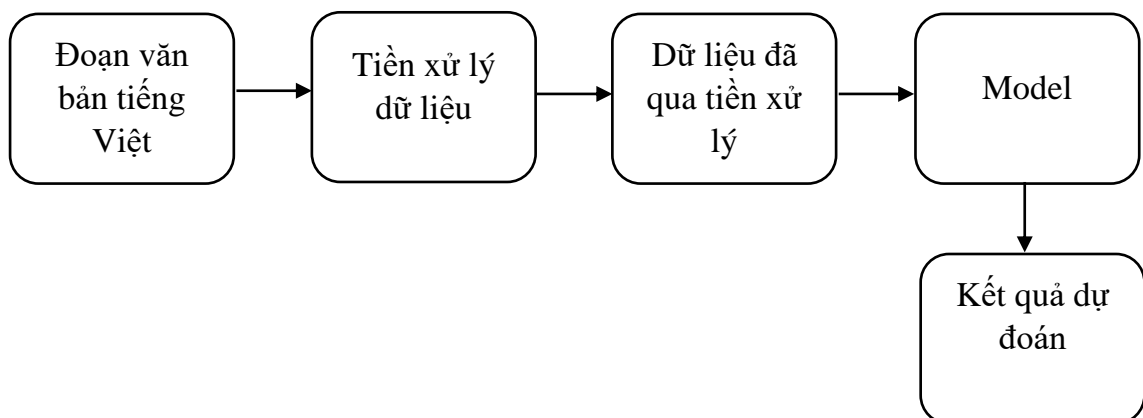


Hình 2. 1 Sơ đồ huấn luyện

### Giai đoạn 2: Dự đoán

Giai đoạn dự đoán (prediction) là giai đoạn đánh giá mô hình học máy sau khi đã được huấn luyện. Ở giai đoạn này, dữ liệu để đưa vào dự đoán cũng cần thực hiện các bước tiền xử lý và trích xuất đặc trưng. Mô hình học nhận đầu vào là đặc trưng đó để đưa vào model sau đó đưa ra kết quả dự đoán. Từ các kết quả dự đoán nhận chúng ta có thể rút ra được kết luận về độ chính xác của mô hình từ đó đưa ra quyết định có cần điều chỉnh tập dữ liệu hoặc tham số để huấn luyện lại mô hình hay không.

Sơ đồ đánh giá kết quả học được triển khai như sau:



Hình 2. 2 Sơ đồ dự đoán

## 2.3. Bài toán xây dựng hệ thống quản lý tin tức

### 2.3.1. Yêu cầu về chức năng của hệ thống

- Thu thập tự động nhưng bài viết trên mạng xã hội Facebook dựa vào Url (phạm vi bao gồm các Trang, Nhóm, Trang cá nhân)
- Phân loại tin tức theo chủ đề về An ninh quốc gia, An ninh trật tự
- Đánh giá các bài viết theo một số mức độ như tích cực, trung bình, tiêu cực
- Ghi nhận, lưu trữ, thống kê tin tức và gửi báo cáo.
- Lưu trữ, quản lý thông tin những đối tượng có hành vi, xu hướng đăng tin tiêu cực

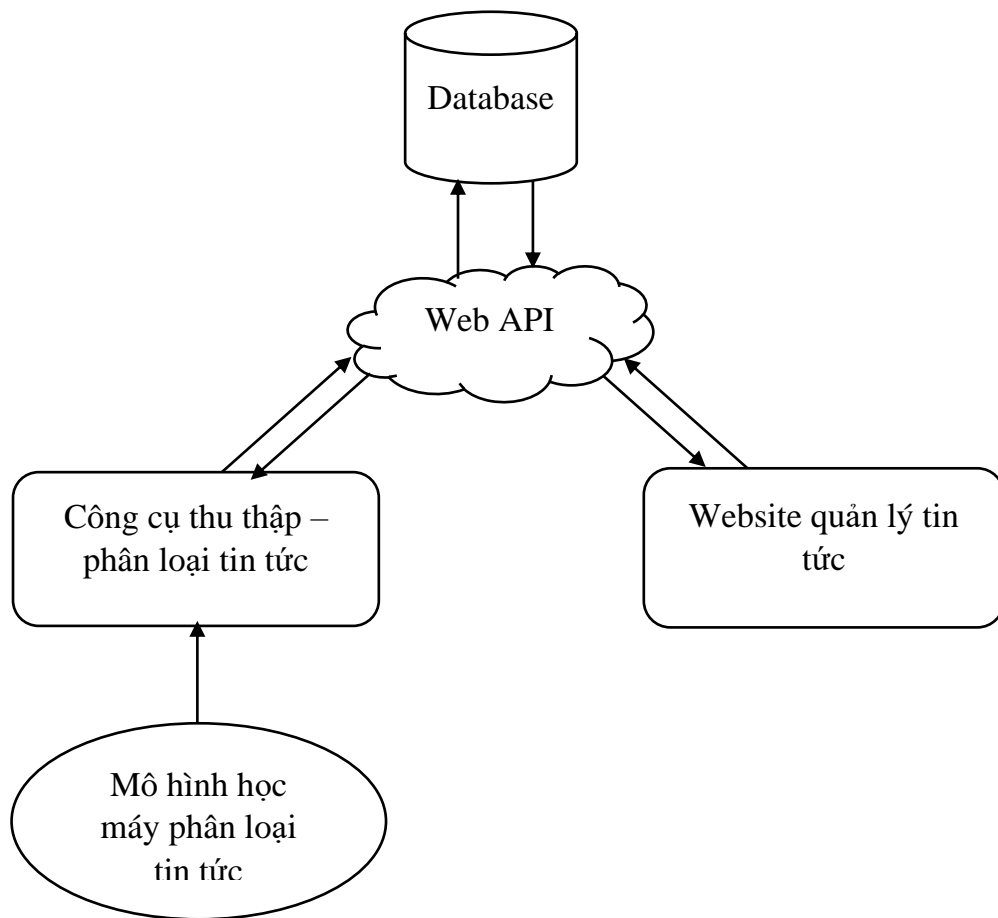
### 2.3.2. Lên kế hoạch xây dựng hệ thống

Dựa trên những yêu cầu về hệ thống, kế hoạch để triển khai như sau:

- Xây dựng cơ sở dữ liệu lưu trữ tin tức
- Huấn luyện mô hình phân loại – đánh giá tin tức
- Lập trình công cụ thu thập tin tức tự động từ mạng xã hội Facebook
- Xây dựng Web API
- Xây dựng Web-site quản lý, thống kê, báo cáo về tin tức

#### ***Phương án triển khai***

Với mục tiêu xây dựng hệ thống để triển khai thực tế. Phương án đề ra là viết ứng dụng, website hoạt động đa nền tảng, các thành phần được viết dưới dạng mô-đun và liên kết với nhau thông qua API.



Hình 2. 3 Mô hình triển khai hệ thống

### 2.3.3. Công cụ triển khai

- Visual studio 2019
- Visual studio code
- SQL Server 2012
- Pycharm IDE 2020.2.3
- Postman 7.36.1

### 2.3.4. Ngôn ngữ triển khai

#### *Python*

Python là một ngôn ngữ lập trình hướng đối tượng, đây là một ngôn ngữ bậc cao, mạnh mẽ, được tạo ra bởi Guido van Rossum. Python được thiết kế vào cuối những năm 1980 và được phát hành lần đầu tiên vào tháng 2 năm 1991. Không phải

chỉ Python mới hỗ trợ lập trình về học máy nhưng Python là một trong những ngôn ngữ tốt nhất cho lập trình học máy vì:

- Là một ngôn ngữ miễn phí, mã nguồn mở, có thể trực tiếp sửa đổi mã nguồn, sử dụng cho mục đích thương mại.
- Python có cú pháp rất đơn giản, rõ ràng. Nó dễ đọc, dễ viết hơn khi so sánh với những ngôn ngữ lập trình khác như C++, Java, C#.
- Python là một ngôn ngữ thông dịch do đó nó có ưu điểm về tốc độ thực thi. Khi chạy code Python, trình thông dịch sẽ tự động chuyển đổi code sang ngôn ngữ máy tính có thể hiểu mà không cần phải trải qua quá trình biên soạn code (compile) như một số ngôn ngữ lập trình khác.
- Các chương trình được viết bằng Python có thể được thực thi ở hầu hết các nền tảng phổ biến như Windows, macOS, Linux do đó nó có tính linh hoạt và khả năng mở rộng cao.
- Python có một cộng đồng hỗ trợ lớn và rất nhiều các bộ thư viện được xây dựng để phục vụ cho lĩnh vực trí tuệ nhân tạo, học máy. Vì vậy đây là điểm mấu chốt giúp Python trở thành ngôn ngữ số một hiện nay.

### ***C#***

C# (C Sharp) là một ngôn ngữ lập trình hướng đối tượng được phát triển bởi Microsoft. Đây được xem là một ngôn ngữ lập trình hiện đại, đơn giản, thuần về hướng đối tượng. C# được Microsoft phát triển dựa trên 2 ngôn ngữ huyền thoại đó là C++ và Java do đó nó được kế thừa những ưu điểm mạnh mẽ của 2 ngôn ngữ này. C# làm việc chủ yếu trên .NET Framework, ngôn ngữ lập trình này có khả năng tạo ra nhiều ứng dụng mạnh mẽ và an toàn cho nền tảng Windows bao gồm các ứng dụng desktop, dịch vụ web, ứng dụng di động và còn nhiều khả năng khác nữa.

### ***HTML/ CSS***

HTML là một ngôn ngữ đánh dấu siêu văn bản được dùng để tạo và cấu trúc các thành phần trong trang web. HTML không phải là một ngôn ngữ lập trình, đồng nghĩa với việc nó không thể tạo ra các chứng năng “động” được. Mục đích chính



của HTML là để bố cục và định dạng trang web.

CSS là một ngôn ngữ định kiểu theo tầng được dùng để tạo bố cục, trang trí, thiết lập màu nền, màu chữ, kích thước, ... cho trang web. Gọi CSS là ngôn ngữ định kiểu theo tầng vì mã CSS khi nhúng vào HTML sẽ được áp dụng theo nguyên tắc theo tầng (cascading). Điều này có nghĩa là nếu một đoạn code CSS được viết để định kiểu cho một phần tử HTML nào đó thì tất cả phần tử bên trong nó (phần tử con) cũng sẽ được kế thừa kiểu trang trí này.

### ***JavaScript***

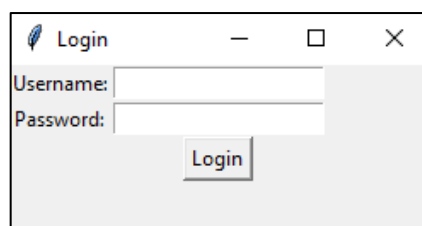
Javascript là một ngôn ngữ lập trình kịch bản cho phép triển khai những chức năng phức tạp trên trang web. Nó là một trong những ngôn ngữ phổ biến nhất thế giới trong suốt 20 năm qua. Cùng với HTML và CSS, Javascript có thể được xem là một trong 3 thành phần quan trọng cấu tạo nên một website.

#### **2.3.5. Mô hình và thư viện sử dụng**

##### ***Thư viện Tkinter***

Để lập trình một ứng dụng có giao diện người dùng (GUI) trên desktop thì thường các ngôn ngữ phổ biến được sử dụng là C# với winform, Wpf hoặc là Java Swing. Tuy nhiên trong phạm vi đề tài, để có thể nhúng mô-đun phân loại văn bản được viết bằng Python một cách thuận tiện lên giao diện ứng dụng thì phần giao diện cũng sẽ được viết bằng ngôn ngữ Python với thư viện Tkinter.

Tkinter là một thư viện được cài mặc định trên Python có chứa mô-đun Tk hỗ trợ cho việc lập trình GUI. Trong mô-đun Tk có chứa các thành phần hỗ trợ xây dựng giao diện cơ bản như Cửa sổ (Window), Khung (Frame), các thành phần con như Nút (Button), Nhãn (Label), Hộp văn bản (Textbox), ...



Hình 2. 4 Ví dụ thư viện Tkinter

### **Mô hình MVC**

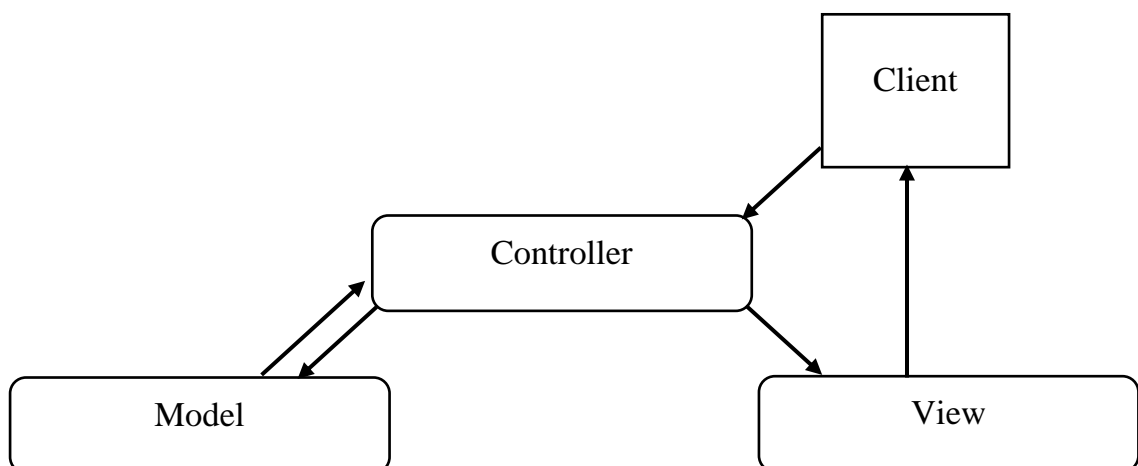
MVC là viết tắt của Model – View – Controller. Là một kiến trúc phần mềm hay mô hình được sử dụng trong thiết kế website. Nói cho dễ hiểu, nó là mô hình phân bố source code thành 3 phần, mỗi thành phần có một nhiệm vụ riêng biệt và độc lập với các thành phần khác.

**Controller:** Giữ nhiệm vụ nhận điều hướng các yêu cầu từ người dùng và gọi đúng những phương thức xử lý chúng... Chẳng hạn thành phần này sẽ nhận request từ url và form để thao tác trực tiếp với Model.

**Model:** Đây là thành phần chứa tất cả các nghiệp vụ logic, phương thức xử lý, truy xuất database, đối tượng mô tả dữ liệu như các Class, hàm xử lý...

**View:** Đảm nhận việc hiển thị thông tin, tương tác với người dùng, nơi chứa tất cả các đối tượng GUI như textbox, images... Hiểu một cách đơn giản, nó là tập hợp các form hoặc các file HTML.

Khi có một yêu cầu từ phía client gửi đến server, Bộ phận controller có nhiệm vụ nhận yêu cầu, xử lý yêu cầu đó. Và nếu cần, nó sẽ gọi đến phần model, vốn là bộ phận làm việc với Database. Sau khi xử lý xong, toàn bộ kết quả được đẩy về phần View. Tại View, sẽ gen ra mã Html tạo nên giao diện, và trả toàn bộ html về trình duyệt để hiển thị.



Hình 2. 6 Mô hình MVC

### ***Entity Framework***

Entity Framework là một thư viện ORM, một loại chương trình giúp ánh xạ qua lại giữa các object của chương trình và bản ghi/bảng của cơ sở dữ liệu quan hệ được Microsoft phát triển từ năm 2008. Entity Framework là công cụ làm việc với cơ sở dữ liệu được Microsoft khuyến nghị.

Trong phạm vi đề tài Entity Framework được sử dụng để xây dựng WebAPI phía server.

## CHƯƠNG 3. XÂY DỰNG VÀ TRIỂN KHAI HỆ THỐNG

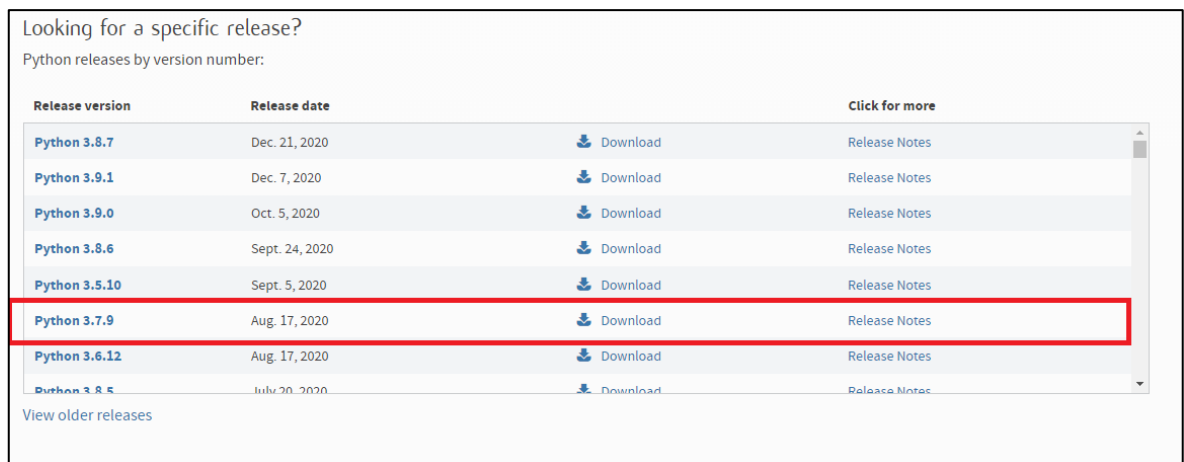
### 3.1. Cài đặt công cụ triển khai

#### 3.1.1. Cài đặt Python

Phiên bản cài đặt: 3.7.9

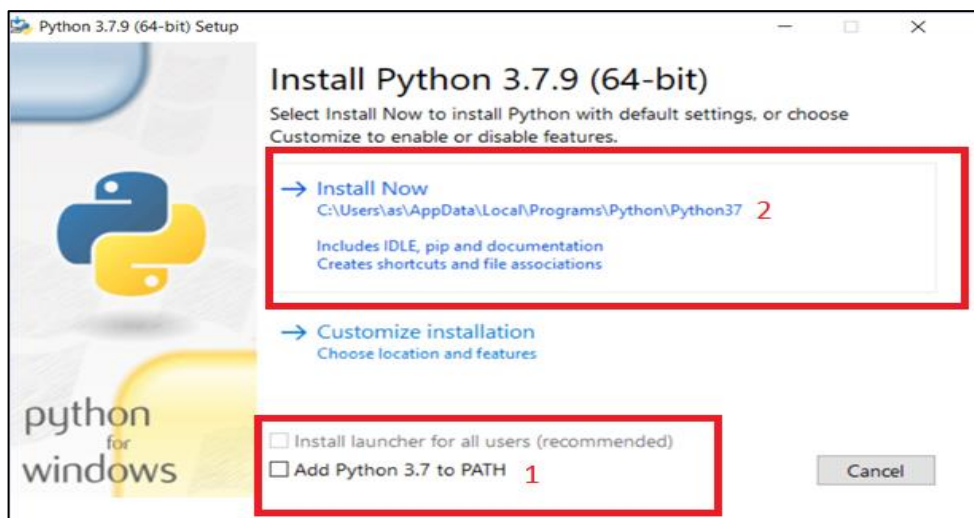
*Các bước cài đặt*

**Bước 1:** Tải file cài đặt tại địa chỉ <https://www.python.org/downloads/>

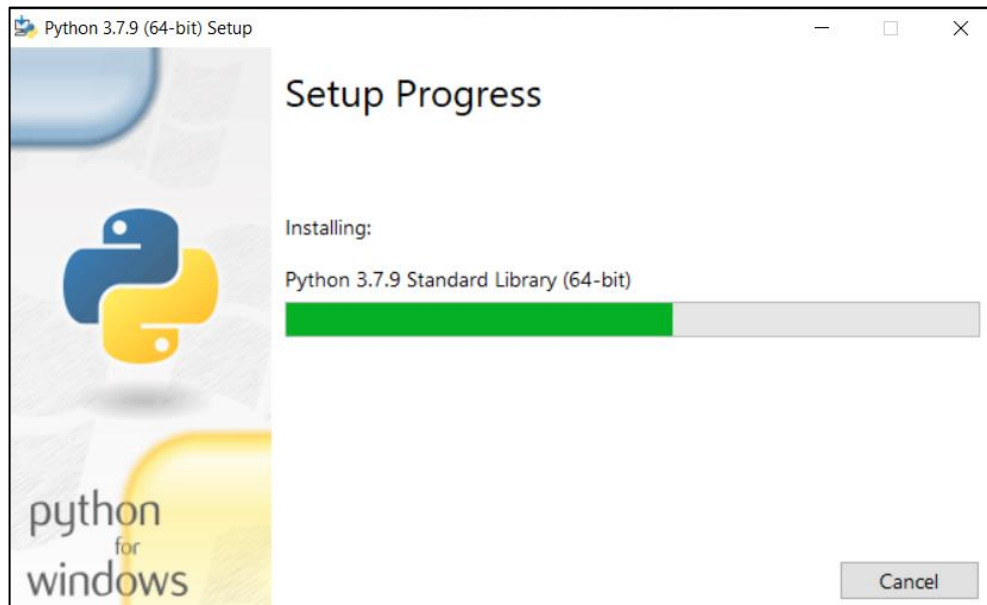


Hình 3. 1 Cài đặt Python bước 1

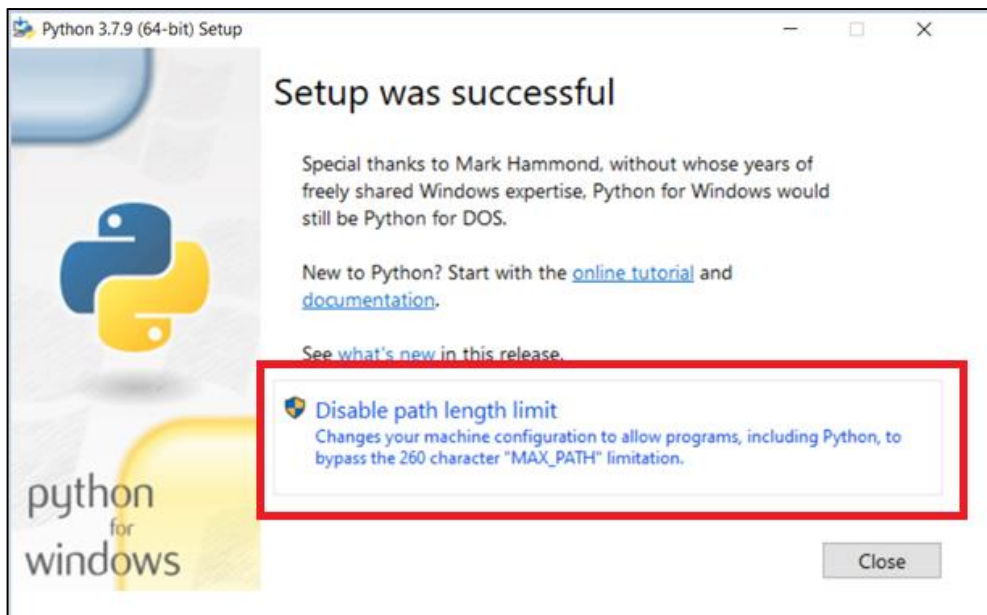
**Bước 2:** Chạy file cài đặt vừa tải về, chọn Add Python 3.7 to PATH sau đó chọn Install Now



Hình 3. 2 Cài đặt Python bước 2

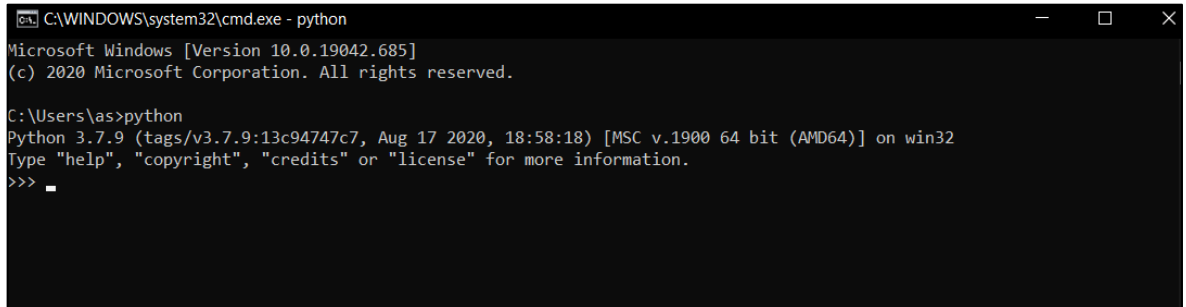
**Bước 3:** Quá trình cài đặt diễn ra tự động

Hình 3. 3 Cài đặt Python bước 3

**Bước 4:** Quá trình cài đặt hoàn tất, click chọn Disable path length limit

Hình 3. 4 Cài đặt Python bước 4

**Bước 5:** Kiểm tra python đã có trên máy hay chưa bằng cách mở cmd, gõ lệnh “python”. Nếu thông báo hiện như hình tức là quá trình cài đặt đã thành công.



```

C:\WINDOWS\system32\cmd.exe - python
Microsoft Windows [Version 10.0.19042.685]
(c) 2020 Microsoft Corporation. All rights reserved.

C:\Users\as>python
Python 3.7.9 (tags/v3.7.9:13c94747c7, Aug 17 2020, 18:58:18) [MSC v.1900 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>>
  
```

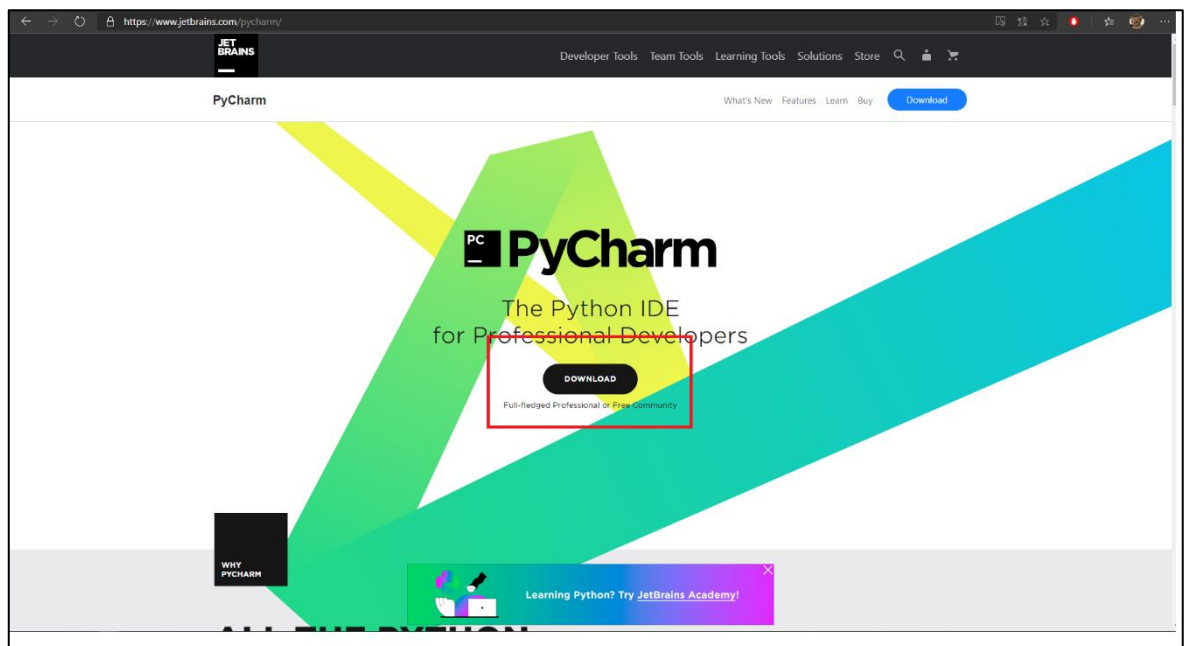
Hình 3. 5 Cài đặt Python bước 5

### 3.1.2. Cài đặt Pycharm

Phiên bản cài đặt: 2020.3.2

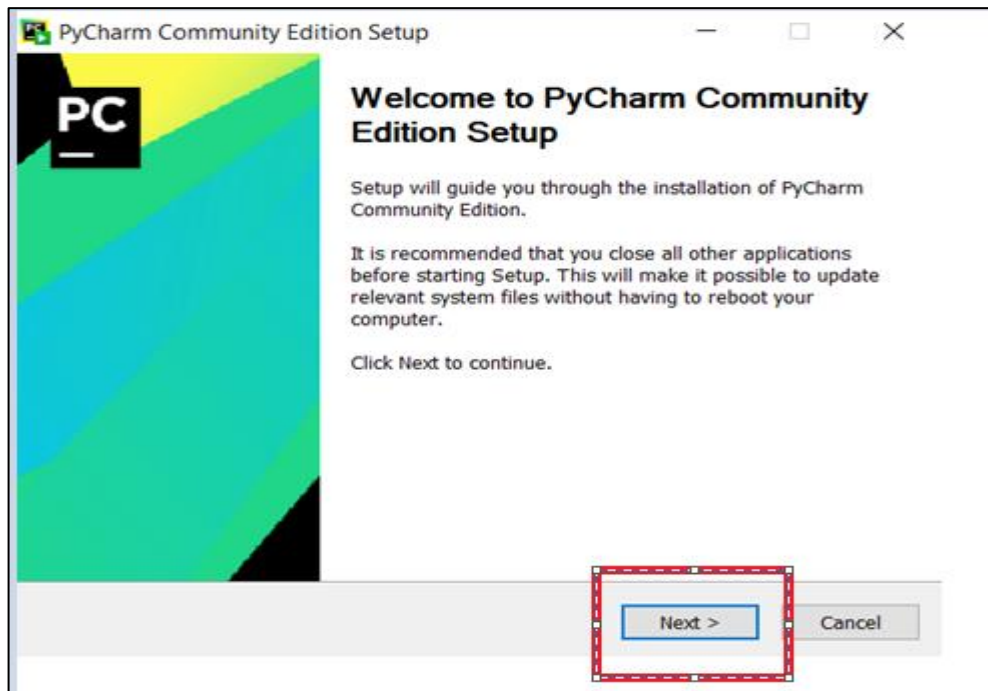
*Các bước cài đặt:*

**Bước 1:** Tải bộ công cụ cài đặt tại địa chỉ <https://www.jetbrains.com/pycharm/>



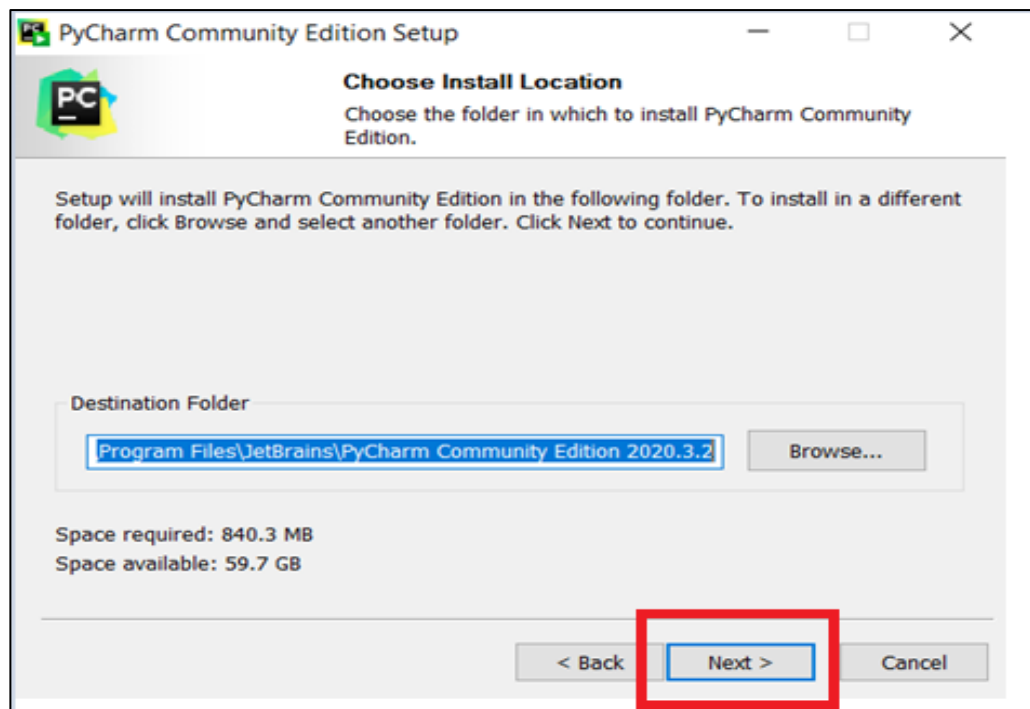
Hình 3. 6 Cài đặt Pycharm bước 1

**Bước 2:** Chạy file cài đặt vừa tải về, chọn Next



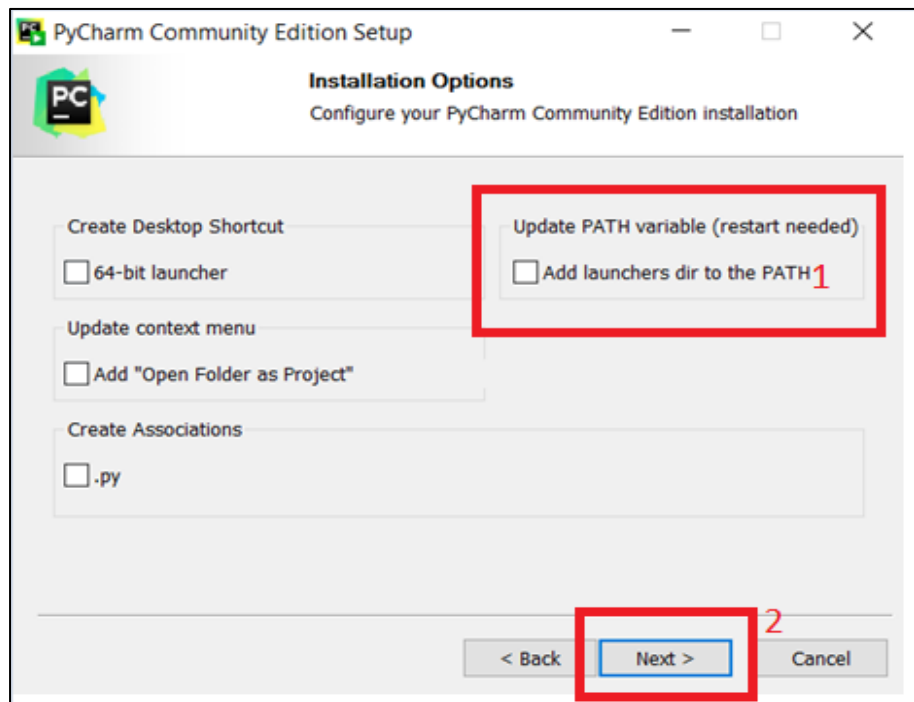
Hình 3. 7 Cài đặt Pycharm bước 2

**Bước 3:** Chọn đường dẫn cài đặt ứng dụng, chọn Next



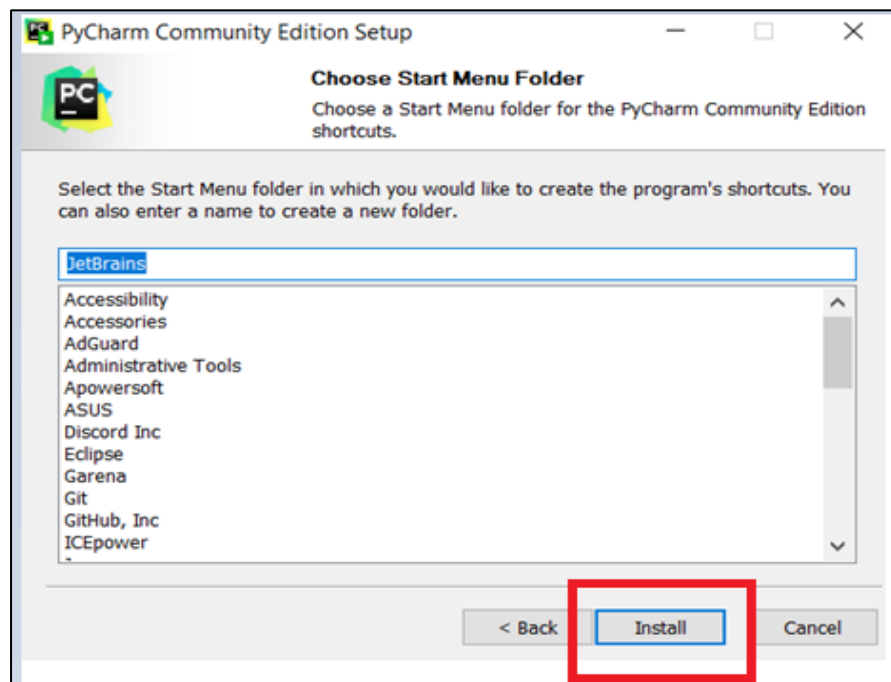
Hình 3. 8 Cài đặt Pycharm bước 3

**Bước 4:** Click vào tùy chọn Add launchers dir to the PATH sau đó chọn Next



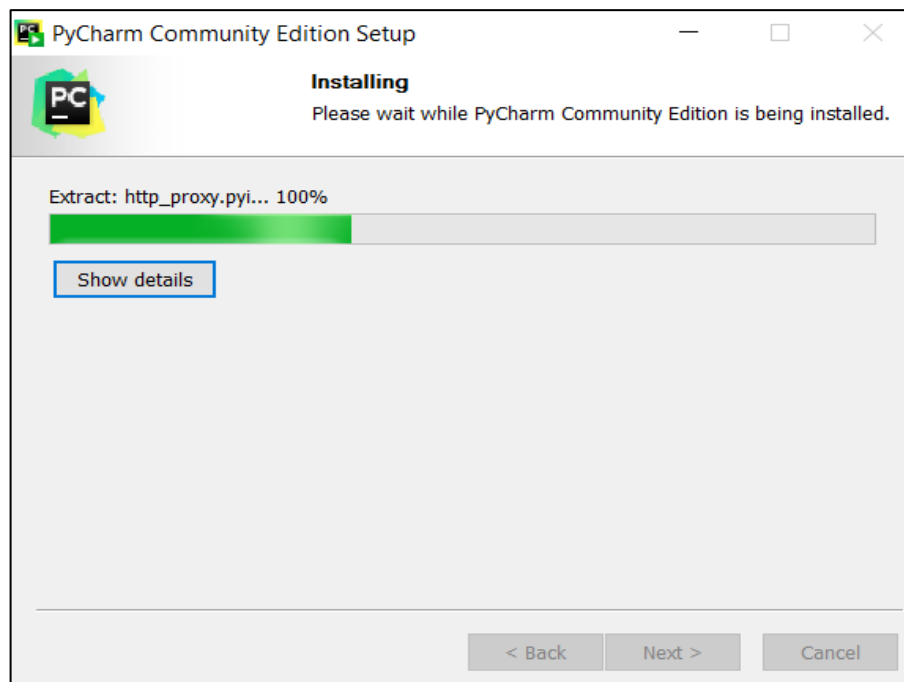
Hình 3. 9 Cài đặt Pycharm bước 4

**Bước 5:** Click Install

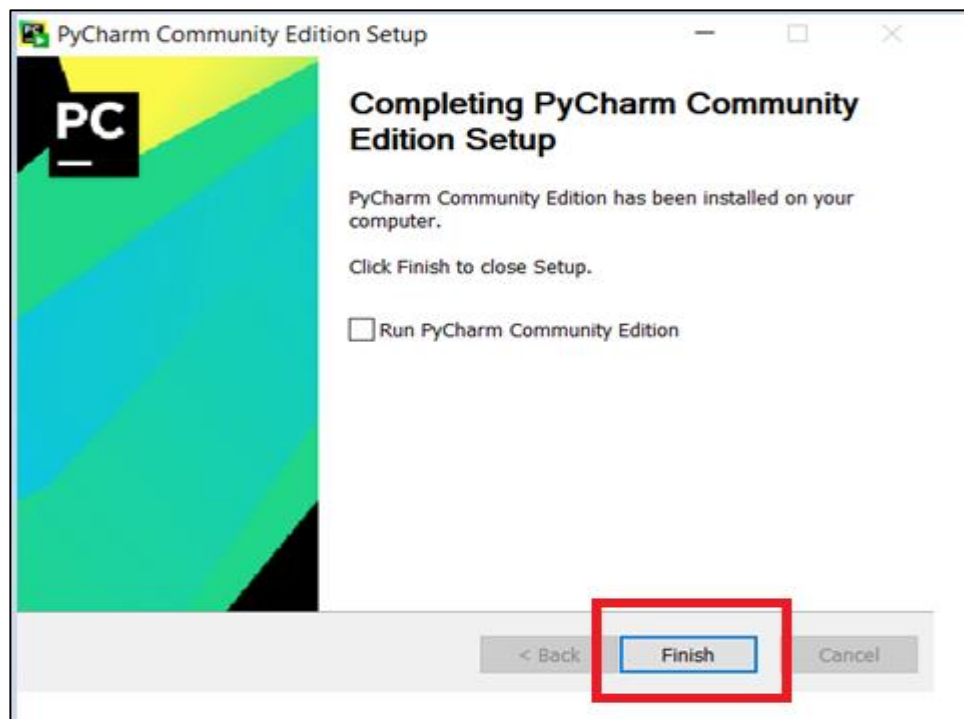


Hình 3. 10 Cài đặt Pycharm bước 5



**Bước 6:** Quá trình cài đặt được diễn ra tự động

Hình 3. 11 Cài đặt Pycharm bước 6

**Bước 7:** Quá trình cài đặt thành công, click Finish để hoàn tất quá trình cài đặt

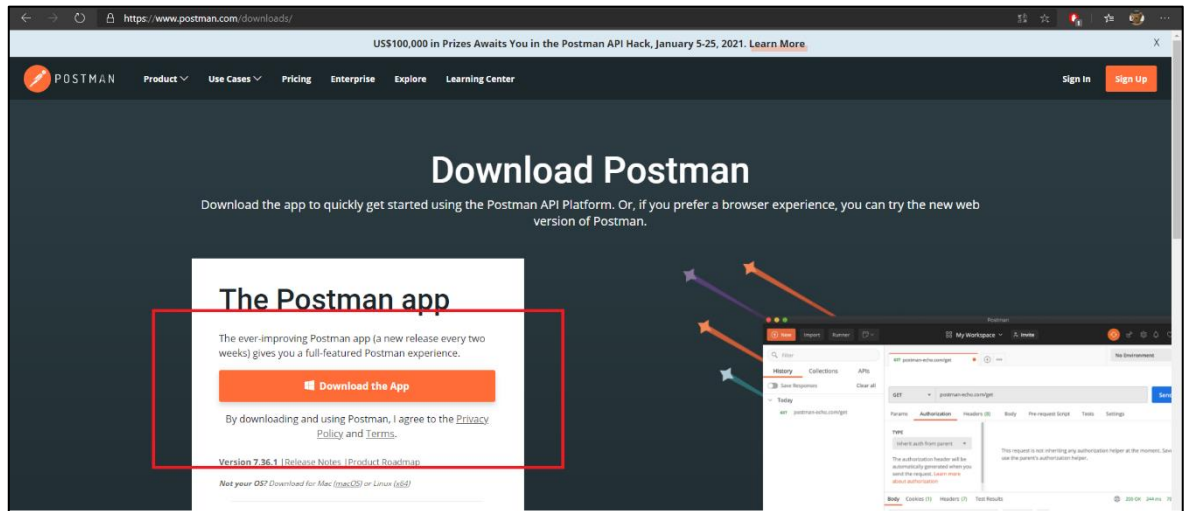
Hình 3. 12 Cài đặt Pycharm bước 7

### 3.1.3. Cài đặt Postman

Phiên bản cài đặt: 7.6.31

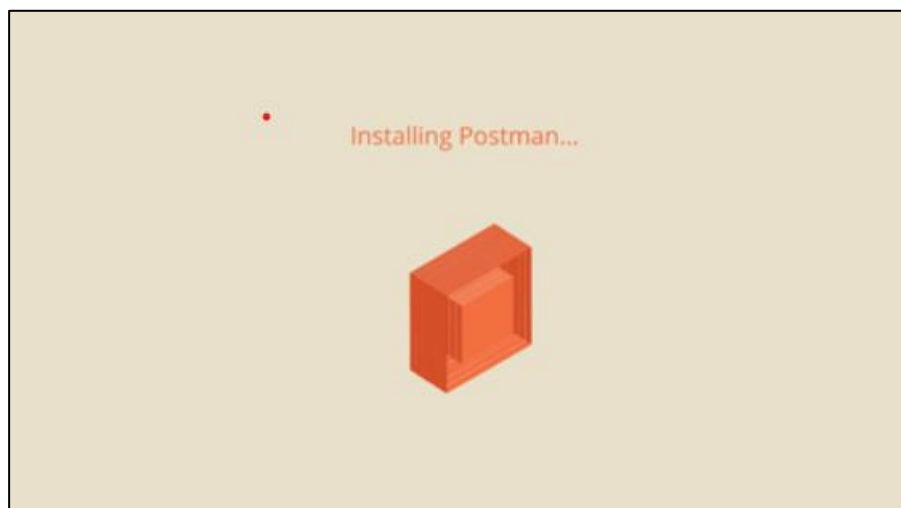
#### *Các bước cài đặt*

**Bước 1:** Tải bộ công cụ cài đặt tại địa chỉ <https://www.Postman.com/downloads/>



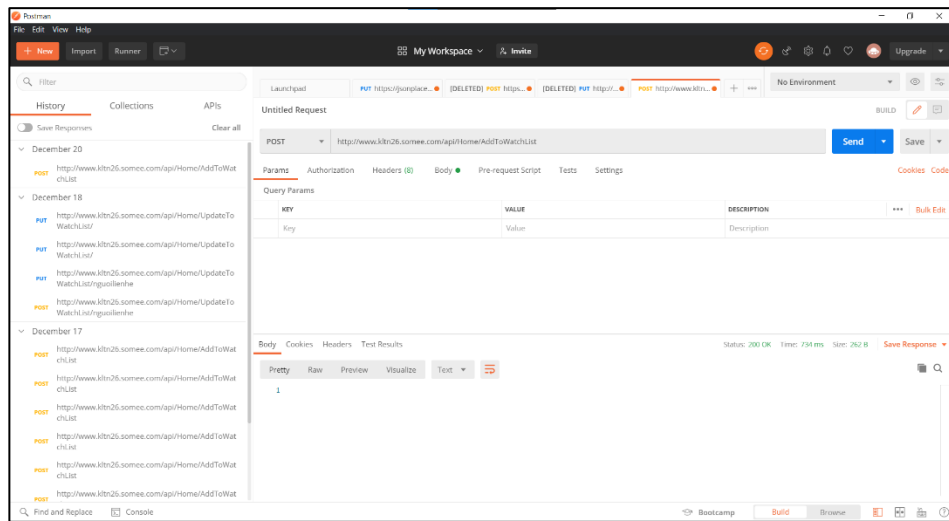
Hình 3. 13 Quá trình cài đặt Postman bước 1

**Bước 2:** Chạy file cài đặt vừa tải về, quá trình cài đặt diễn ra tự động



Hình 3. 14 Quá trình cài đặt Postman bước 2

### Bước 3: Quá trình cài đặt hoàn tất



Hình 3. 15 Quá trình cài đặt Postman bước 3

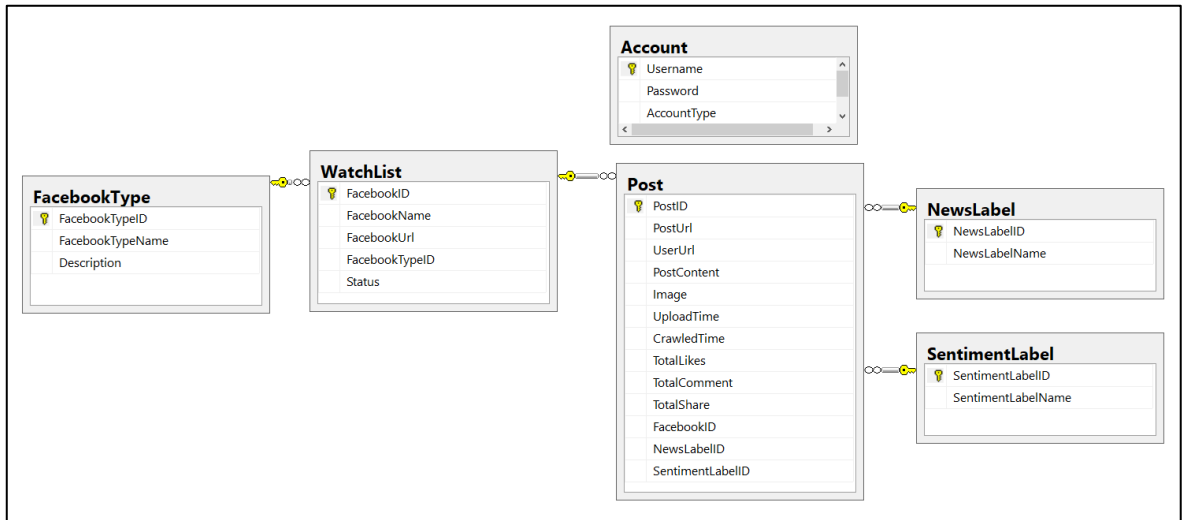
## 3.2. Cơ sở dữ liệu lưu trữ tin tức

### 3.2.1. Cơ sở dữ liệu

Gồm 6 bảng:

- Account: Thông tin tài khoản và mật khẩu đăng nhập vào hệ thống.
- FacebookType: Các nguồn đăng tin trên Facebook.
- WatchList: Danh sách theo dõi các đối tượng trên Facebook.
- Post: Thông tin chi tiết về một bài đăng.
- NewsLabel: Lưu trữ nhãn chủ đề của các bài đăng.
- SentimentsLabel: Lưu trữ nhãn mức độ của các bài đăng.

Sơ đồ diagram của cơ sở dữ liệu:



Hình 3. 16 Sơ đồ diagram

### 3.2.2. Mô tả cơ sở dữ liệu

#### Account (Tài khoản hệ thống)

Bảng 3.1 Account (Tài khoản hệ thống)

STT	Trường dữ liệu	Kiểu dữ liệu	Mô tả	Khóa
1	Username	Nvarchar (100)	Tên tài khoản	PK
2	Password	Nvarchar (200)	Mật khẩu	
3	AccountType	Int	Loại tài khoản	

#### FacebookType (Nguồn đăng tin trên Facebook)

Bảng 3.2 FacebookType (Nguồn đăng tin trên Facebook)

STT	Trường dữ liệu	Kiểu dữ liệu	Mô tả	Khóa
1	FacebookTypeID	Nvarchar (100)	Mã nguồn đăng tin	PK
2	FacebookTypeName	Nvarchar (200)	Tên nguồn đăng tin	
3	Description	Nvarchar (500)	Mô tả	

**WatchList (Danh theo dõi)**

Bảng 3.3 WatchList (Danh theo dõi)

STT	Trường dữ liệu	Kiểu dữ liệu	Mô tả	Khóa
1	FacebookID	Nvarchar (100)	Mã tài khoản Facebook	PK
2	FacebookName	Nvarchar (500)	Tên tài khoản Facebook	
3	FacebookUrl	Nvarchar (500)	Địa chỉ tài khoản Facebook	
4	FacebookTypeID	Nvarchar (100)	Mã nguồn đăng tin	FK
5	Status	Bit	Trạng thái	

**NewsLabel (Các nhãn tin tức)**

Bảng 3.4 NewsLabel (Các nhãn tin tức)

STT	Trường dữ liệu	Kiểu dữ liệu	Mô tả	Khóa
1	NewsLabelID	Nvarchar (100)	Mã nhãn tin tức	PK
2	NewsLabelName	Nvarchar (200)	Tên nhãn tin tức	

**SentimentLabel (Các nhãn mức độ)**

Bảng 3.5 SentimentLabel (Các nhãn mức độ)

STT	Trường dữ liệu	Kiểu dữ liệu	Mô tả	Khóa
1	SentimentLabelID	Nvarchar (100)	Mã nhãn mức độ	PK
2	SentimentLabelName	Nvarchar (200)	Tên nhãn mức độ	

**Post (Bài viết)**

Bảng 3.6 Post (Bài viết)

STT	Trường dữ liệu	Kiểu dữ liệu	Mô tả	Khóa
1	PostID	Nvarchar (100)	Mã bài viết	PK
2	PostUrl	Nvarchar (500)	Địa chỉ bài viết	
3	UserUrl	Nvarchar (500)	Địa chỉ người đăng	
4	PostContent	Nvarchar (MAX)	Nội dung bài viết	
5	Image	Nvarchar (MAX)	Hình ảnh	
6	UploadTime	Datetime	Thời gian đăng	
7	CrawledTime	Datetime	Thời gian thu thập	
8	TotalLikes	Int	Số lượt thích bài viết	
9	TotalComment	Int	Số bình luận bài viết	
10	TotalShare	INT	Số lượt chia sẻ bài viết	
11	FacebookID	Nvarchar (100)	Mã tài khoản	FK
12	NewsLabelID	Nvarchar (100)	Mã nhãn tin tức	FK
13	SentimentLabelID	Nvarchar (100)	Mã nhãn mức độ	FK

**3.3. Xây dựng tập huấn luyện**

Nguồn tin tức sử dụng để xây dựng tập huấn luyện được trích lọc từ các nguồn tin trên mạng xã hội Facebook, các nguồn báo như Zingnews.vn, Vietnamnet.vn, ... với khoảng hơn 200 bài viết cho mỗi chủ đề.

Trước khi đưa vào mô hình học phân loại, đánh giá tin tức, dữ liệu phải được qua bước tiền xử lý, cấu trúc đoạn code để tiền xử lý văn bản được miêu tả như sau:

```
def text_preprocess(text):
    remove_html(text) //Hàm loại bỏ code HTML khỏi văn bản
    convert_unicode(text) //Hàm chuẩn hóa bộ gõ unicode
    standardized_vietnamese_accent_sentence(document) // Hàm chuẩn hóa
    cách gõ dấu tiếng Việt
    tokenize(text) //Hàm tách từ tiếng Việt
    lower(text) //Hàm chuyển hóa văn bản về chữ thường
    re_sub(text) //Hàm loại bỏ các ký tự thừa
```

**Ví dụ:**

*Một đoạn văn bản trước khi tiền xử lý:*

Đại hội XIII Đảng Cộng sản Việt Nam - con đường đi đến những thành tựu mới. Là đội quân tiên phong của người dân Việt Nam - Đảng Cộng sản Việt Nam ngay từ những ngày đầu hoạt động cho đến nay luôn thể hiện là một tổ chức không ngừng sáng tạo, với những quyết sách táo bạo, không theo những tiêu chuẩn sáo mòn, được kết tinh bởi sự đoàn kết toàn dân tộc.

*Đoạn văn bản sau khi tiền xử lý:*

đại\_hội xiii đảng cộng\_sản việt\_nam con\_đường đi đến những thành\_tựu mới là đội\_quân tiên\_phong của người dân việt\_nam đảng cộng\_sản việt\_nam ngay từ những ngày đầu hoạt\_động cho đến nay luôn\_thể\_hiện là một tổ\_chức không\_ngừng sáng\_tạo với những quyết\_sách táo\_bạo không theo những tiêu\_chuẩn sáo\_mòn được kết\_tinh bởi sự đoàn\_kết toàn\_dân\_tộc

Công đoạn tiếp theo trong quá trình xây dựng tập huấn luyện là phân tích và gắn nhãn cho bài viết. Lấy ví dụ trên ta có từ khóa “đảng cộng\_sản” vậy nhãn cho đoạn văn trên là thuộc về chủ đề An ninh quốc gia. Tiếp theo ta xét tới các từ khóa thể hiện mức độ như “thành\_tự”, “tiên\_phong”, “sáng\_tạo”, “đoàn\_kết”. Đây là các từ khóa thể hiện mức độ tích cực cho nên bài viết sẽ được gắn nhãn tích cực. Cấu trúc của một đoạn văn sau khi được gắn nhãn như sau:

\_\_label\_\_[“nhãn chủ đề”][“nhãn mức độ”][“nội dung văn bản”]  
 \_\_label\_\_an\_ninh\_quốc\_gia\_tích\_cực đại\_hội xiii đảng cộng\_sản  
 việt\_nam...

### 3.4. Huấn luyện mô hình phân loại, đánh giá tin tức

Cấu hình máy tính được sử dụng để huấn luyện:

Bảng 3.7 Cấu hình máy tính dùng để huấn luyện mô hình

CPU	RAM	Ổ cứng	Hệ điều hành
Intel i5-7200U 2.5GHz	8 GB	SSD 128GB	Window 10 Home

Các kết quả thu được sau khi huấn luyện với thuật toán Naive Bayes và bộ dữ liệu đầu vào gần 3000 bài viết gồm:

- Thời gian huấn luyện: 0,87 giây
- Thống kê độ chính xác của từng nhãn:

Bảng 3.8 Thống kê độ chính xác của từng nhãn

Label	Precision	Recall	F1-score	Support
An ninh quốc gia (tiêu cực)	0.86	1.00	0.93	38
An ninh quốc gia (tích cực)	0.95	0.93	0.94	44
An ninh trật tự (tiêu cực)	0.73	0.69	0.71	32



An ninh trật tự (tích cực)	0.91	0.75	0.82	56
Công nghệ (bình thường)	0.91	0.97	0.94	40
Giáo dục (bình thường)	0.89	0.95	0.92	43
Kinh doanh (bình thường)	0.77	0.94	0.85	36
Phim ảnh (bình thường)	1.00	0.88	0.94	34
Sức khỏe (bình thường)	0.85	0.98	0.91	45
Thể thao (bình thường)	1.00	0.94	0.97	35
Thời trang (bình thường)	0.94	0.79	0.86	43
Xe (bình thường)	0.95	0.85	0.90	46
Âm nhạc (bình thường)	0.93	0.93	0.93	45
Ẩm thực (bình thường)	0.91	1.00	0.96	43

Bảng 3.9 Độ chính xác trung bình của mô hình

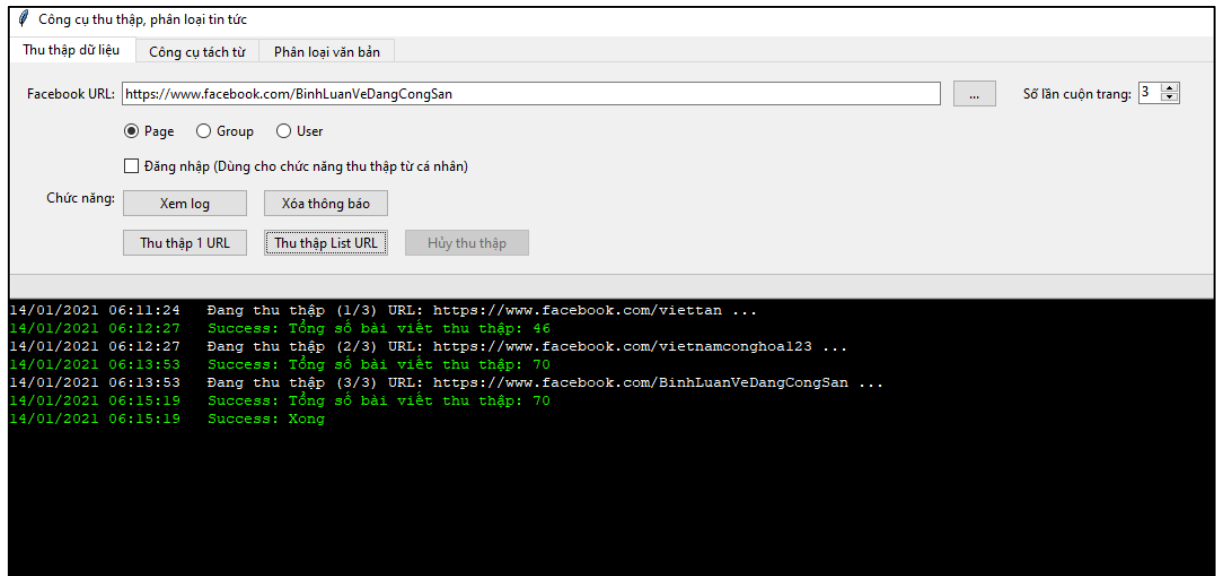
Accuracy			0.90	580
Macro avg	0.90	0.90	0.90	580
Weighted avg	0.90	0.90	0.90	580

Với bảng số liệu trên, độ chính xác trung bình của mô hình khi test trên tập huấn luyện là khoảng 90%. Tuy nhiên khi ta xét về độ chính xác của từng nhãn riêng biệt, nhãn An ninh trật tự với mức độ tiêu cực cho ra độ chính xác không như mong đợi (dưới 80%). Nguyên nhân có thể xuất phát từ tập huấn luyện chưa đủ tốt hoặc các đặc trưng của nhãn này chưa thật sự nổi bật để mô hình học có thể nhận diện rõ ràng. Để khắc phục vấn đề này, phương án là xây dựng lại một tập huấn luyện khác với nguồn dữ liệu đủ tốt và kết hợp với các mô hình học máy khác.

### 3.5. Công cụ thu thập, phân loại, đánh giá tin tức tự động

Phần công cụ được xây dựng để chạy trên môi trường desktop với 3 nhiệm vụ chính:

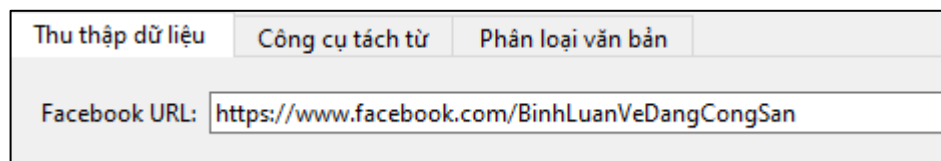
#### 3.5.1. Thu thập tin tức tự động



Hình 3. 17 Giao diện chính của công cụ

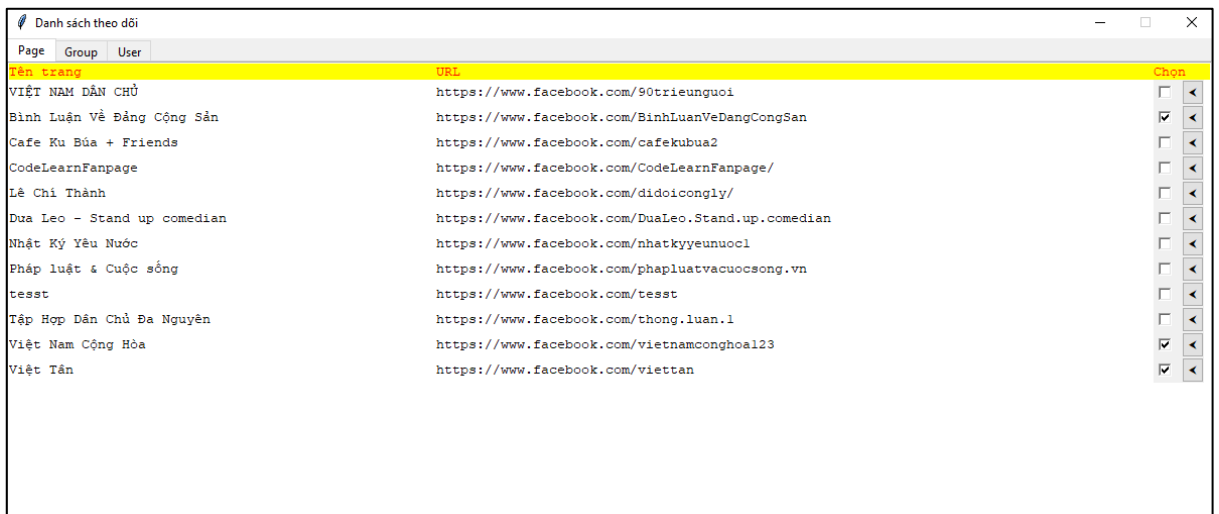
Mô tả: Giao diện có nhiệm vụ cung cấp các chức năng thu thập tin tức tự động bao gồm:

- Textbox cho phép nhập vào một URL Facebook.



Hình 3. 18 Khu vực nhập URL

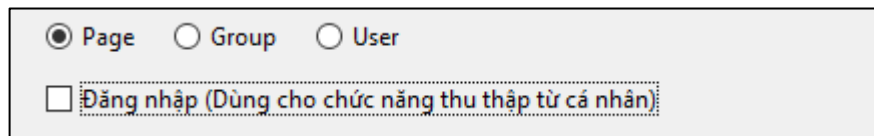
- Khi ấn vào nút mở rộng [...] một cửa sổ hiển thị danh sách các đối tượng đang theo dõi sẽ được hiện lên. Trong cửa sổ này có thể thực hiện các thao tác chọn 1 URL cụ thể hoặc thêm nhiều URL vào trong danh sách để cung cấp cho chức năng thu thập tin tức theo danh sách.



Tên trang	URL	Chọn
VIỆT NAM DÂN CHỦ	https://www.facebook.com/90trieunguoi	<input type="checkbox"/>
Bình Luận Về Đảng Cộng Sản	https://www.facebook.com/BinhLuanVeDangCongSan	<input checked="" type="checkbox"/>
Cafe Ku Búa + Friends	https://www.facebook.com/cafekubua2	<input type="checkbox"/>
CodeLearnFanpage	https://www.facebook.com/CodeLearnFanpage/	<input type="checkbox"/>
Lê Chí Thành	https://www.facebook.com/didoicongly/	<input type="checkbox"/>
Dua Leo - Stand up comedian	https://www.facebook.com/DuaLeo.Stand.up.comedian	<input type="checkbox"/>
Nhật Ký Yêu Nước	https://www.facebook.com/nhatkyyeuuoc1	<input type="checkbox"/>
Pháp luật & Cuộc sống	https://www.facebook.com/phapluatvacuocsong.vn	<input type="checkbox"/>
tesst	https://www.facebook.com/tesst	<input type="checkbox"/>
Tập Hợp Dân Chủ Đa Nguyên	https://www.facebook.com/thong.luan.1	<input type="checkbox"/>
Việt Nam Cộng Hòa	https://www.facebook.com/vietnamconghoa123	<input checked="" type="checkbox"/>
Việt Tân	https://www.facebook.com/viettan	<input checked="" type="checkbox"/>

Hình 3. 19 Danh sách theo dõi

- Khu vực Radiobutton để tùy chọn cho công cụ biết URL đã nhập thuộc từ nguồn tin nào trên Facebook (Trang, Nhóm, Trang cá nhân).

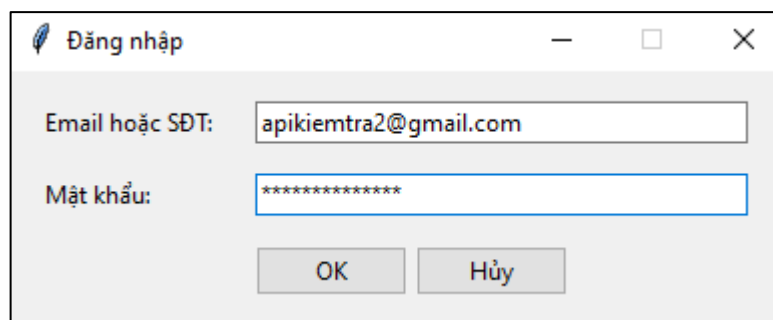


☒ Page
 ☐ Group
 ☐ User

☐ Đăng nhập (Dùng cho chức năng thu thập từ cá nhân)

Hình 3. 20 Tùy chọn loại trang và đăng nhập

- Checkbox đăng nhập là một tùy chọn bắt buộc khi lấy tin tức từ Trang cá nhân. Khi click chọn sẽ hiển thị hộp thoại đăng nhập.

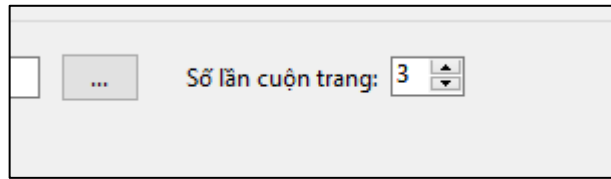


Email hoặc SĐT:

Mật khẩu:

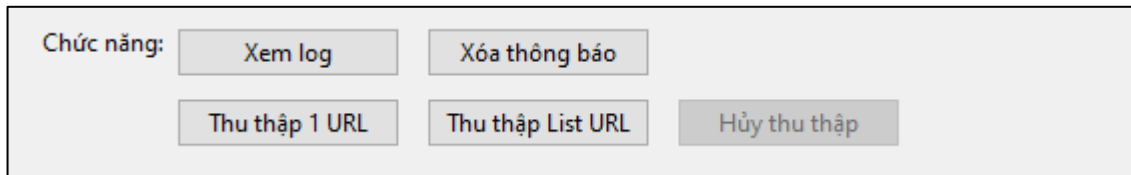
Hình 3. 21 Giao diện đăng nhập

- Tùy chọn số lần cuộn trang cho biết số lượng trang sẽ cuộn khi thao tác quét dữ liệu. Số lần cuộn trang càng lớn, lượng bài viết thu về càng nhiều nhưng tối đa là 20 lần để tránh trường hợp bị Facebook phát hiện là robot dẫn đến bị khóa tài khoản hoặc chặn truy cập.



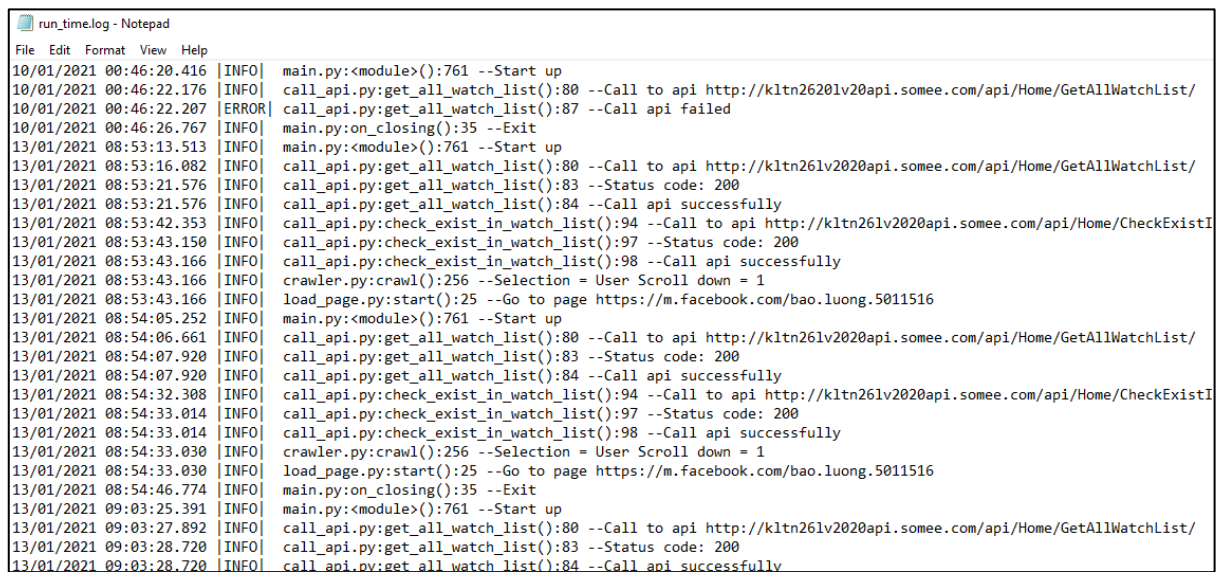
Hình 3. 22 Số lần cuộn trang

- Khu vực chức năng



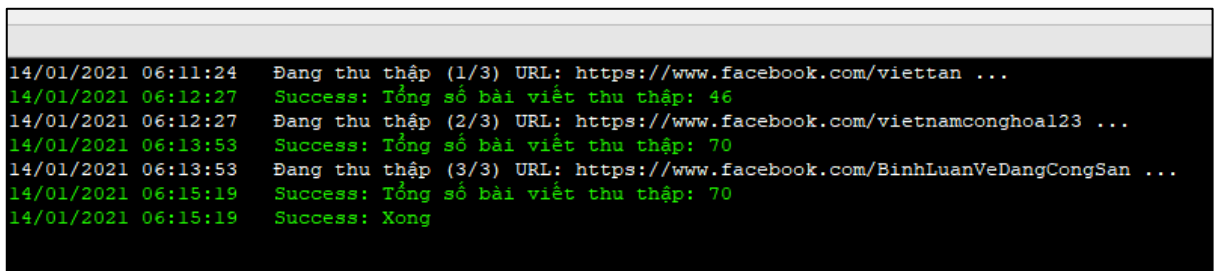
Hình 3. 23 Khu vực chức năng

- Chức năng Xem Log hiển thị nhật ký hoạt động của ứng dụng, mục đích để dễ dàng kiểm soát và sửa các lỗi phát sinh trong quá trình vận hành.



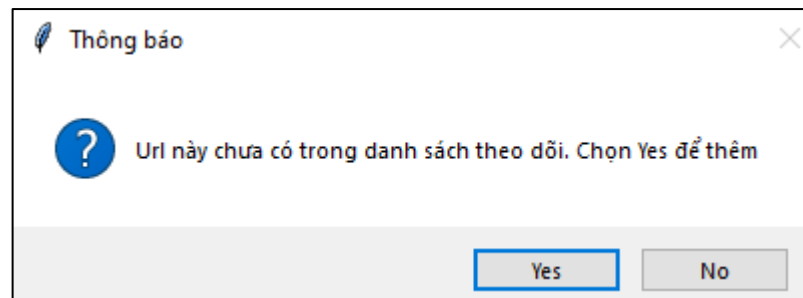
Hình 3. 24 Xem nhật ký

- Chức năng Xóa thông báo để xóa các dòng thông báo tiến trình hoạt động của hệ thống khi cần thiết.



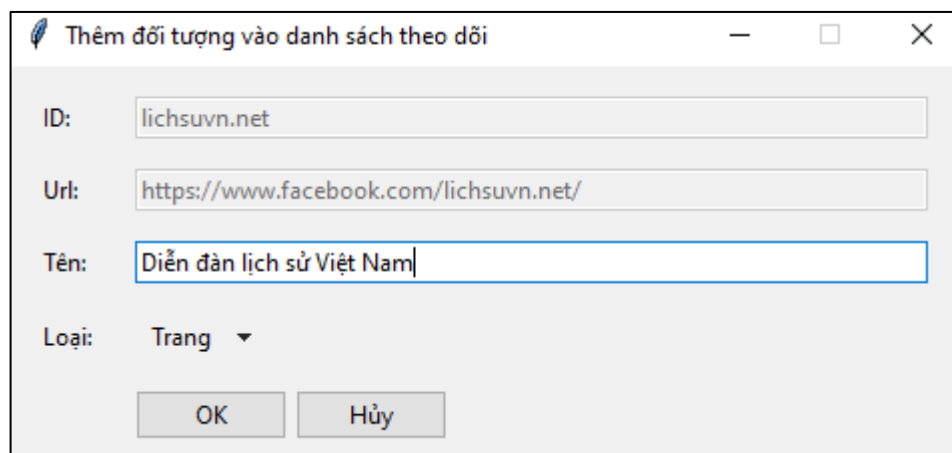
Hình 3. 25 Khu vực thông báo

- Chức năng Thu thập 1 URL: Tiến hành thu thập các bài viết theo URL đã nhập vào Textbox hoặc chọn trong cửa sổ danh sách theo dõi. Trong trường hợp URL nhập vào Textbox chưa có trong cơ sở dữ liệu, hộp thoại thông báo hiện ra yêu cầu thêm URL của đối tượng hiện tại vào trong danh sách theo dõi.



Hình 3. 26 Hộp thoại thông báo

- Nếu chọn “Yes” sẽ hiển thị giao diện cho phép thêm một đối tượng mới. Nếu chọn “No” thì sẽ hủy thao tác thu thập.



Hình 3. 27 Giao diện thêm đối tượng

- Chức năng Thu thập List URL: Tiến hành thu thập bài viết theo danh sách URL đã chọn trong danh sách theo dõi.
- Chức năng Hủy thu thập: Hủy quá trình thu thập tin tức.

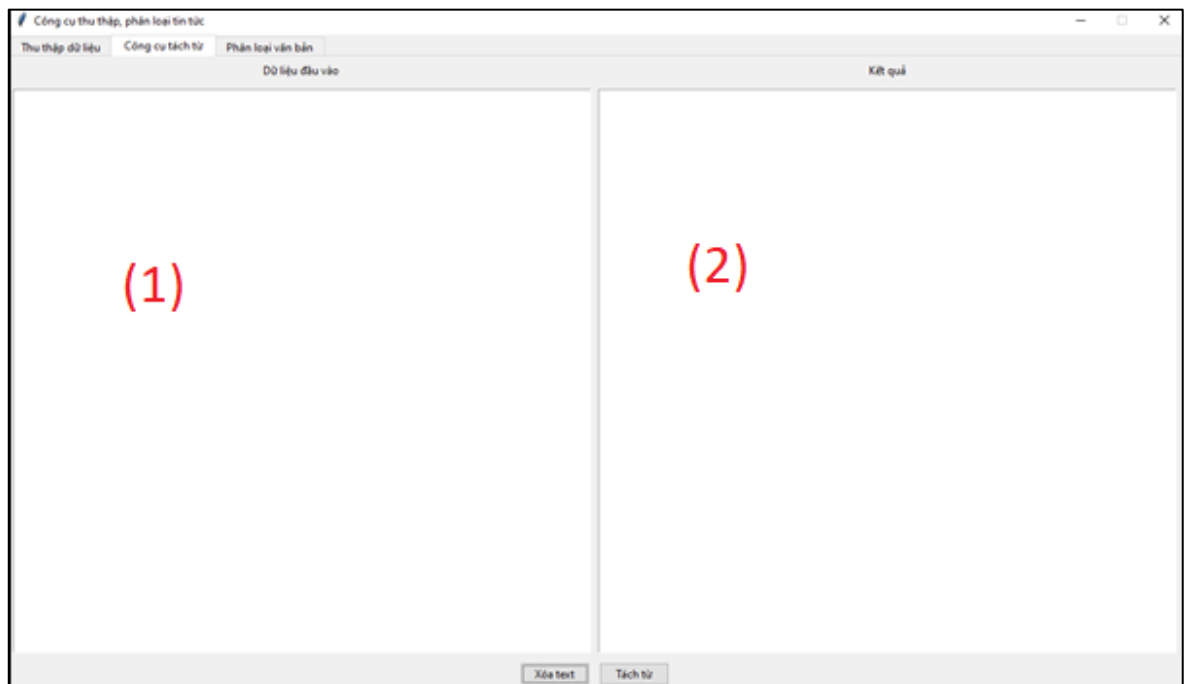
```

14/01/2021 06:55:12 Đang thu thập (1/4) URL: https://www.facebook.com/CodeLearnFanpage/ ...
14/01/2021 06:55:15 Đang hủy tiến trình thu thập dữ liệu...
14/01/2021 06:55:17 Warning: Hủy
14/01/2021 06:55:17 Đang thu thập (2/4) URL: https://www.facebook.com/didoicongly/ ...
14/01/2021 06:55:20 Warning: Hủy
14/01/2021 06:55:20 Đang thu thập (3/4) URL: https://www.facebook.com/DuaLeo.Stand.up.comedian ...
14/01/2021 06:55:21 Warning: Hủy
14/01/2021 06:55:21 Đang thu thập (4/4) URL: https://www.facebook.com/nhatkyyeunuocl ...
14/01/2021 06:55:23 Warning: Hủy
14/01/2021 06:55:23 Success: Xong

```

Hình 3. 28 Giao diện thông báo hủy

### 3.5.2. Công cụ tách từ



Hình 3. 29 Công cụ tách từ

Mô tả: Giao diện cung cấp chức năng tách từ hỗ trợ trong quá trình xây dựng tập huấn luyện gồm:

- Khu vực dữ liệu đầu vào nhận vào là một đoạn văn bản bình thường chưa thông qua các bước tiền xử lý dữ liệu.

Dữ liệu đầu vào
<p>Đại hội XIII Đảng Cộng sản Việt Nam - con đường đi đến những thành tựu mới. Là đội quân tiên phong của người dân Việt Nam - Đảng Cộng sản Việt Nam ngay từ những ngày đầu hoạt động cho đến nay luôn thể hiện là một tổ chức không ngừng sáng tạo, với những quyết sách táo bạo, không theo những tiêu chuẩn sáo mòn, được kết tinh bởi sự đoàn kết toàn dân tộc.</p> <p>Chiến lược này đã trở thành vũ khí lợi hại được Đảng Cộng sản Việt Nam vận dụng một cách tài tình, linh hoạt, đưa dân tộc Việt Nam tới mọi bến bờ chiến thắng trước tất cả các thế lực thực dân, đế quốc. Và thành tựu nổi bật nhất chính là Đại hội VI của Đảng Cộng sản Việt Nam, đã kịp thời đề ra chính sách "Đổi mới", đưa đất nước đến những thành quả đáng mừng như ngày nay.</p> <p>Nếu ở thập kỷ 70, 80 của thế kỷ trước, Liên Hợp Quốc liệt Việt Nam nằm trong số 25 nước nghèo nhất thế giới, thì chỉ sau đó một thập kỷ, vào giữa những năm 90, Việt Nam đã nhanh chóng gia nhập nhóm các nước đứng đầu thế giới về tốc độ tăng trưởng kinh tế.</p> <p>Việt Nam đã đạt được những thay đổi căn bản và toàn diện, nền kinh tế thoát khỏi khủng hoảng và đang vươn lên với tốc độ khá nhanh. Sự nghiệp công nghiệp hoá, hiện đại hoá, phát triển kinh tế thị trường định hướng xã hội chủ nghĩa đang tự tin tiến lên. Cuộc sống của người dân được cải thiện. Hệ thống chính trị và khối đại đoàn kết toàn dân tộc ngày càng được củng cố, ổn định. Quốc phòng và an ninh đáng tin cậy. Vị thế của Việt Nam trên trường quốc tế ngày càng vững chắc.</p>

Hình 3. 30 Khu vực nhận dữ liệu đầu vào

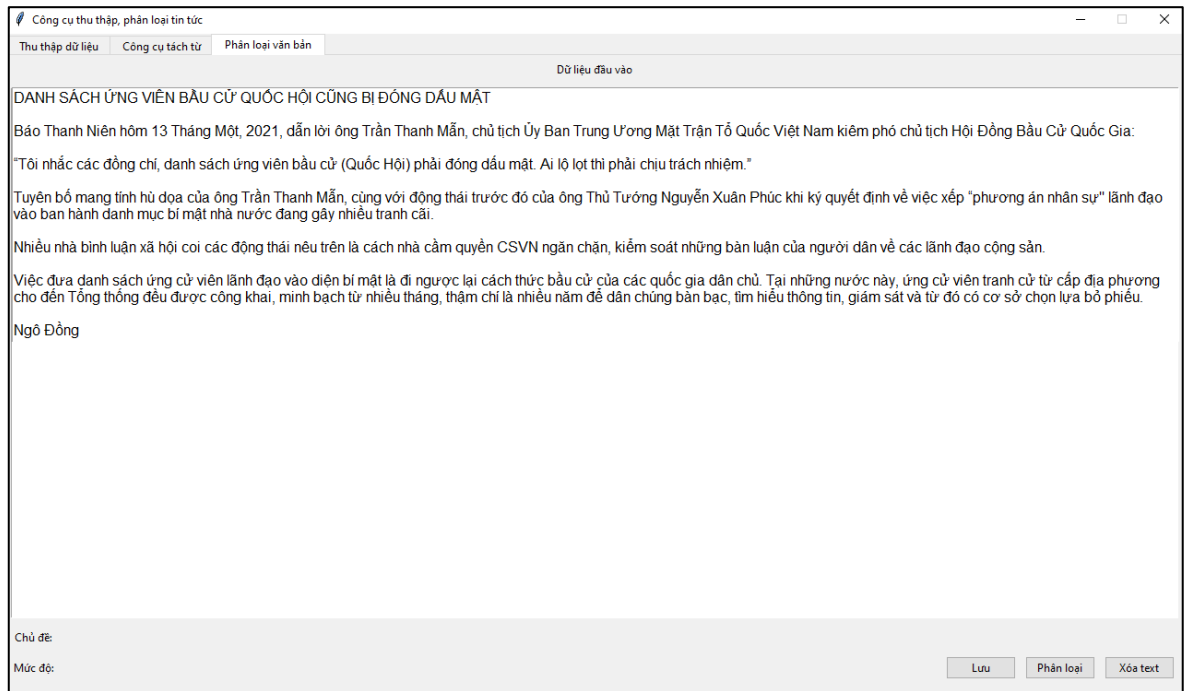
- Khu vực kết quả hiển thị đoạn văn bản sau khi đã được xử lý tách từ.

Kết quả
<p>đại_hội xiii đảng cộng_sản việt_nam con đường đi đến những thành_tựu mới là đội_quân tiên_phong của người dân việt_nam đảng cộng_sản việt_nam ngay từ những ngày đầu hoạt_động cho đến nay luôn_thể_hiện là một tổ_chức không ngừng sáng_tạo với những quyết_sách táo_bạo không theo những tiêu_chuẩn sáo_mòn được kết_tinh bởi sự đoàn_kết toàn_dân_tộc chiến_lược này đã trở_thành vũ_khí lợi_hại được đảng cộng_sản việt_nam vận_dụng một_cách tài_tình linh_hoạt đưa dân_tộc việt_nam tới mọi bến_bờ chiến_thắng trước tất_cả các thế_lực thực_dân đế_quốc và thành_tựu nổi_bật nhất chính là đại_hội_vi của đảng cộng_sản việt_nam đã kịp_thời đề_ra chính_sách đổi_mới đưa đất_nước đến những thành_quả đáng_mừng như ngày_nay nếu ở thập_kỷ 70 80 của thế_kỷ trước liên_hợp_quốc liệt_việt_nam nằm trong số 25 nước_nghèo nhất thế_giới thì chỉ sau đó một thập_kỷ vào giữa những năm 90 việt_nam đã nhanh_chóng gia_nhập nhóm các nước đứng_đầu thế_giới về tốc_độ tăng_trưởng kinh_tế việt_nam đã đạt được những thay_đổi căn_bản và toàn_diện nền kinh_tế thoát_khỏi khủng_hoảng và đang vươn_lên với tốc_độ khá nhanh sự_nghiệp công_nghiệp_hóa hiện_đại_hóa phát_triển kinh_tế thị_trường định_hướng xã_hội_chủ_nghĩa đang_tự_tin tiến_lên cuộc_sống của người dân được cải_thiện hệ_thống chính_trị và khối đại_đoàn_kết toàn_dân_tộc ngày_càng được củng_cố ổn_định quốc_phòng và an_ninh đáng_tin_cậy vị_thế của việt_nam trên trường quốc_tế ngày_càng vững_chắc</p>

Hình 3. 31 Khu vực hiển thị kết quả



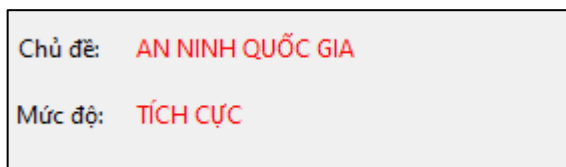
### 3.5.3. Công cụ phân loại văn bản



Hình 3. 32 Công cụ phân loại văn bản

Mô tả: Cung cấp chức năng phân loại văn bản thủ công theo chủ đề và mức độ. Dữ liệu đầu vào là một đoạn văn bản bất kỳ.

- Kết quả sau khi phân tích bao gồm chủ đề và mức độ của đoạn văn bản trên.



Hình 3. 33 Kết quả phân loại

- Ngoài ra sau khi phân loại và đánh giá, công cụ này cũng cung cấp thêm chức năng lưu trữ lại bài viết khi cần. Hộp thoại lưu bài viết bao gồm các thông tin về URL bài viết, URL người đăng tin, Thời gian đăng tin, Mức độ tương tác.

g tự tin tiến lên. Cuộc sống của người dân được cải thiện. Hệ thống chính trị và khối đại đoàn kết toàn dân tộc ngày càng vững chắc.

an ninh đáng tin cậy. Vị thế của Việt Nam trên trường quốc tế ngày càng vững chắc.

Lưu bài viết

URL bài viết:

URL người đăng:

Thời gian đăng:

0 giờ 0 phút

Ngày:

1/14/21

Tương tác:

Like:

Comment:

OK

Hủy

		January 2021						
	Mon	Tue	Wed	Thu	Fri	Sat	Sun	
53	28	29	30	31	1	2	3	
1	4	5	6	7	8	9	10	
2	11	12	13	14	15	16	17	
3	18	19	20	21	22	23	24	
4	25	26	27	28	29	30	31	
5	1	2	3	4	5	6	7	

Hình 3. 34 Giao diện lưu bài viết

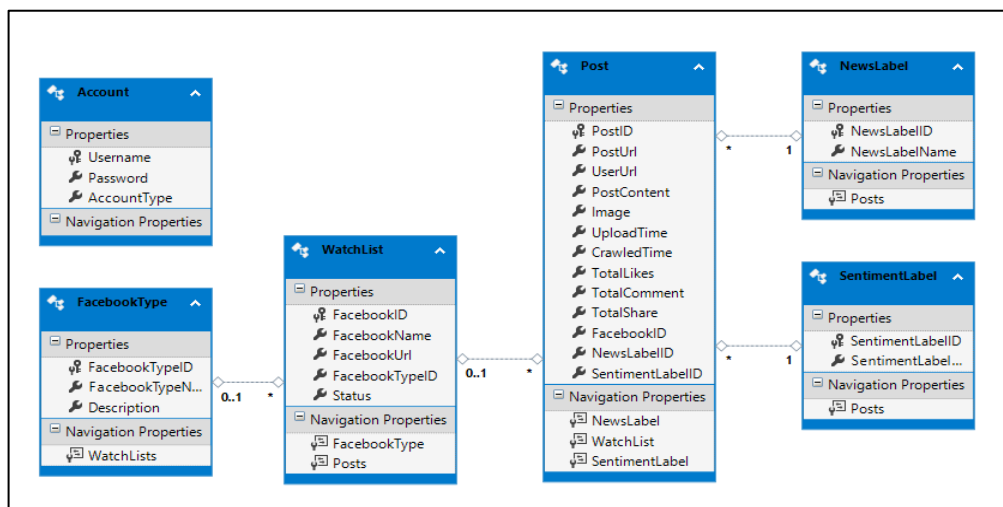
### 3.6. WebAPI

#### 3.6.1. Cấu trúc WebAPI

Hệ thống WebAPI có nhiệm vụ cầu nối giao tiếp giữa website quản lý và công cụ thu thập dữ liệu được cấu tạo bởi 3 thành phần chính:

##### *Data Access Layer (DAL)*

Lớp này có chức năng giao tiếp với hệ quản trị Cơ sở dữ liệu, thực hiện các thao tác truy vấn, thêm, xóa, sửa,... Ở lớp này, Entity Framework được sử dụng để ánh xạ các bảng trong Cơ sở dữ liệu thành các đối tượng (class).



Hình 3. 35 Sơ đồ ánh xạ cơ sở dữ liệu

### ***Business Logic Layer (BLL)***

Đây là lớp nhận nhiệm vụ xử lý một số nghiệp vụ chính như xác thực đăng nhập, lọc, tìm kiếm, thống kê dữ liệu theo điều kiện. Đây còn là nơi kiểm tra các ràng buộc, tính toàn vẹn và hợp lệ dữ liệu trước khi trả về kết quả hoặc hoặc truyền dữ liệu xuống lớp DAL.

### ***WebAPI***

Cung cấp các phương thức dùng để cho phép các ứng dụng khác nhau có thể giao tiếp, trao đổi dữ liệu qua lại. Dữ liệu được Web API trả lại có dạng JSON thông qua giao thức HTTP hoặc HTTPS. Các thao tác chính với WebAPI thường có 4 loại:

- GET: Lấy dữ liệu
- POST: Thêm dữ liệu mới
- PUT: Cập nhật dữ liệu
- DELETE: Xóa dữ liệu

Các phương thức được xây dựng trong hệ thống bao gồm:

Bảng 3.10 Các phương thức trong WebAPI

STT	Phương thức	Loại	Tham số	Dữ liệu trả về
1	GetAllWatchList()	GET	Không	Danh sách các đối tượng đang theo dõi.
2	FilterWatchList()	GET	Mã nguồn tin tức, trạng thái	Danh sách các đối tượng theo dõi được lọc theo nguồn đăng tin (Trang, Nhóm, Trang cá nhân), trạng thái (đang theo dõi, hủy theo dõi).

3	SearchWatchList()	GET	Từ khóa tìm kiếm	Danh sách các đối tượng theo dõi được tìm kiếm theo từ khóa.
4	GetWatchListItemByID()	GET	Id một đối tượng	Thông tin của một đối tượng cụ thể dựa trên Id.
5	CheckExistInWatchList()	GET	Id một đối tượng	True (đã tồn tại), False (chưa tồn tại)
6	AddToWatchList()	POST	Đối tượng cần thêm.	True (thêm thành công), False (thêm thất bại)
7	UpdateToWatchList()	PUT	Đối tượng cần sửa	True (sửa thành công), False (sửa thất bại)
8	Unfollow()	PUT	Id một đối tượng	True (hủy theo dõi thành công), False (hủy theo dõi thất bại)
9	Follow()	PUT	Id một đối tượng	True (theo dõi thành công), False (theo dõi thất bại)
10	GetAllPost()	GET	Không	Danh sách tất cả các bài viết
11	FilterPost()	GET	Tham số bộ lọc	Danh sách bài viết đã lọc theo dựa trên tham số
12	SearchPost()	GET	Tham số tìm kiếm	Danh sách bài viết đã tìm được dựa

				trên tham số tìm kiếm (từ khóa)
13	GetPostByID()	GET	Id bài viết	Thông tin chi tiết bài viết dựa trên Id
14	GetListPostByFacebookID()	GET	Id một đối tượng	Danh sách bài viết đã đăng của một đối tượng cụ thể
15	CheckExistPost()	GET	Id bài viết	True (bài viết đã tồn tại), False (bài viết chưa tồn tại)
16	AddNewPost()	POST	Thông tin bài viết	True (thêm bài viết thành công), False (thêm bài viết thất bại)
17	UpdatePost()	PUT	Thông tin bài viết	True (cập nhật bài viết thành công), False (cập nhật bài viết thất bại)
18	RemovePost()	DELETE	Id bài viết	True (xóa bài viết thành công), False (xóa bài viết thất bại)

### 3.6.2. Kiểm tra API với Postman

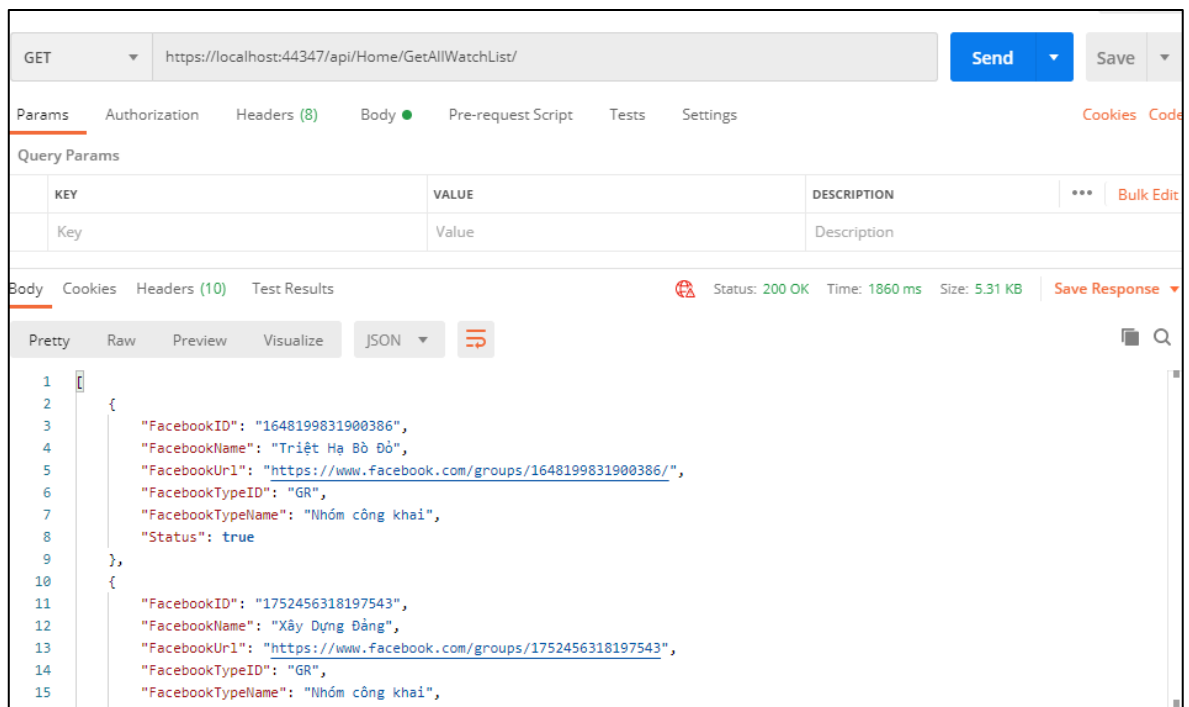
Sau khi xây dựng WebAPI, công đoạn kiểm tra nhằm đảm bảo webAPI hoạt động ổn định, chính xác. Nội dung kiểm tra bao gồm thực hiện các thao tác truy vấn, thêm, xóa, sửa.

#### *Thao tác truy vấn*

Thực hiện kiểm tra đối với tất cả các phương thức thuộc loại GET.

#### **Ví dụ:**

- API: api/Home/GetAllWatchList/
- Tham số truyền vào: Không
- Kết quả mong đợi: Danh sách đối tượng theo dõi.



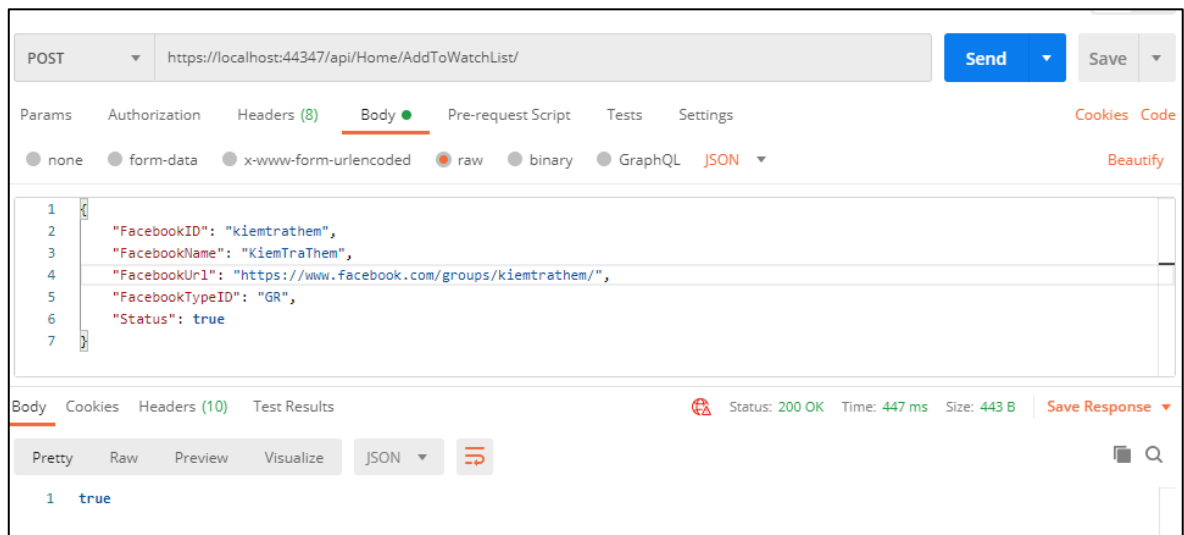
Hình 3. 36 Thao tác truy vấn với Postman

### ***Thao tác thêm***

Thực hiện kiểm tra đối với tất cả phương thức thuộc loại POST

#### **Ví dụ:**

- API: api/Home/AddToWatchList/
- Tham số truyền vào: Một đối tượng theo dõi bao gồm các trường (FacebookID, FacebookName, FacebookUrl, FacebookTypeID, Status)
- Kết quả mong đợi: True (thêm thành công) hoặc False (thêm thất bại)



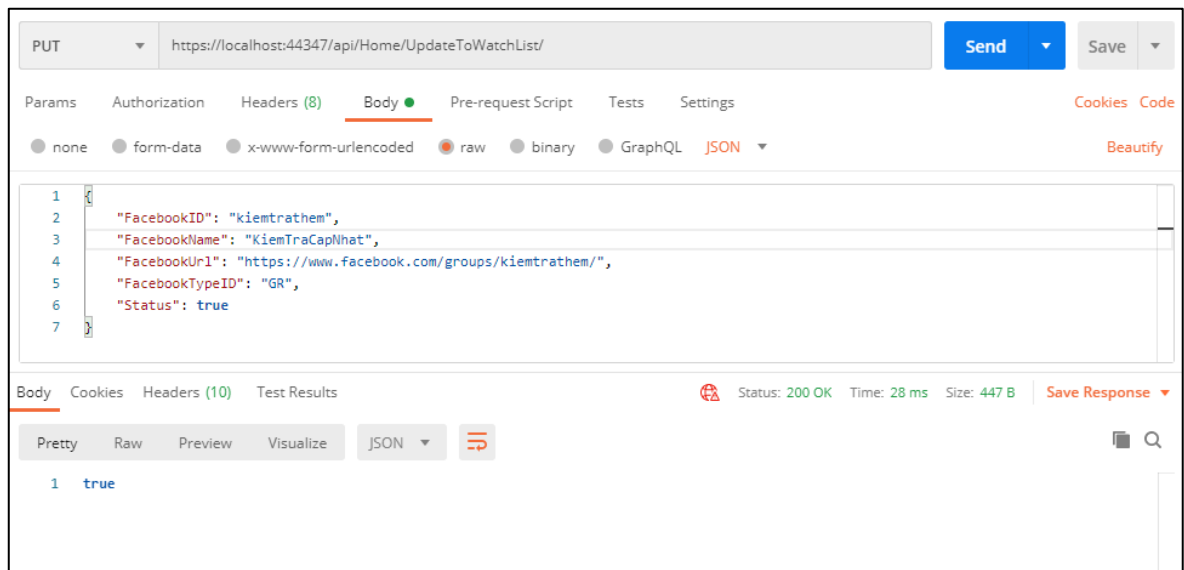
Hình 3. 37 Thao tác thêm với Postman

### ***Thao tác cập nhật***

Thực hiện kiểm tra đối với tất cả phương thức thuộc loại PUT

#### **Ví dụ:**

- API: api/Home/UpdateToWatchList/
- Tham số truyền vào: Một đối tượng theo dõi bao gồm các trường (FacebookID, FacebookName, FacebookUrl, FacebookTypeID, Status)
- Kết quả mong đợi: True (cập nhật thành công) hoặc False (cập nhật thất bại)



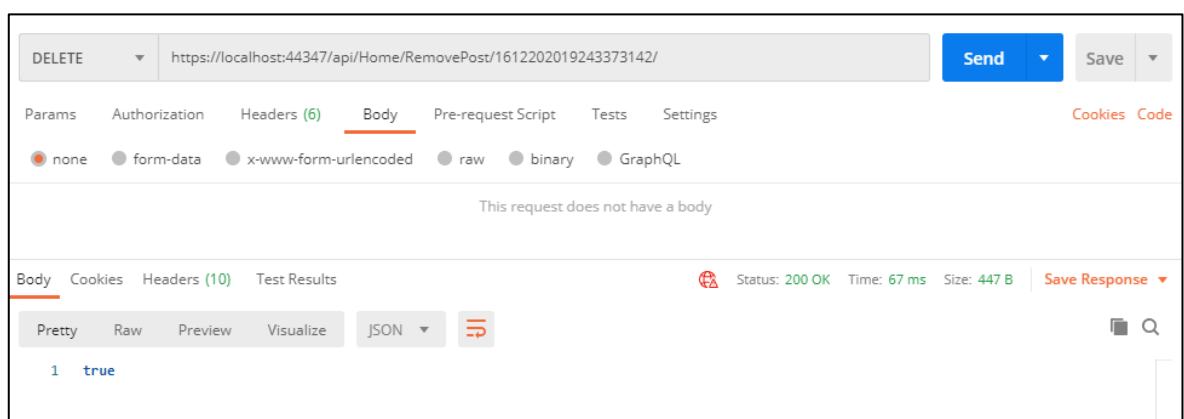
Hình 3. 38 Thao tác cập nhật với PostMan

### ***Thao tác xóa***

Thực hiện kiểm tra đối với tất cả phương thức thuộc loại DELETE

### **Ví dụ:**

- API: api/Home/RemovePost/
- Tham số truyền vào: ID của bài viết cần xóa
- Kết quả mong đợi: True (xóa thành công) hoặc False (xóa thất bại)



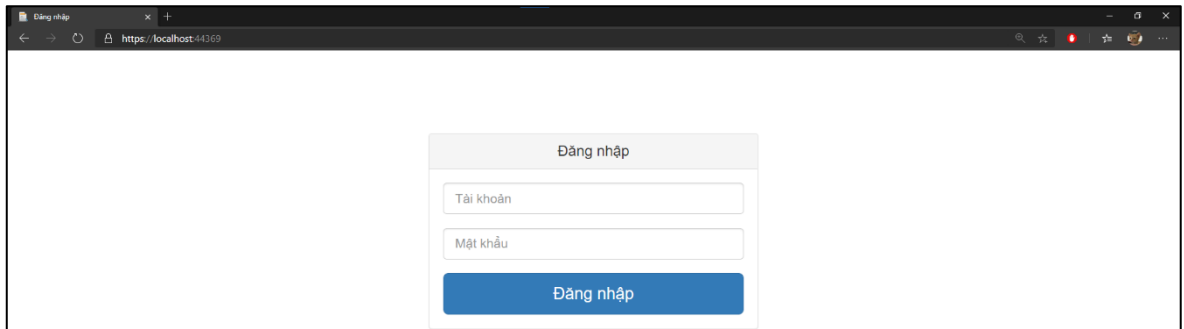
Hình 3. 39 Thao tác xóa với Postman



### 3.7. Xây dựng Website quản lý, thống kê, báo cáo về tin tức

#### 3.7.1. Giao diện đăng nhập

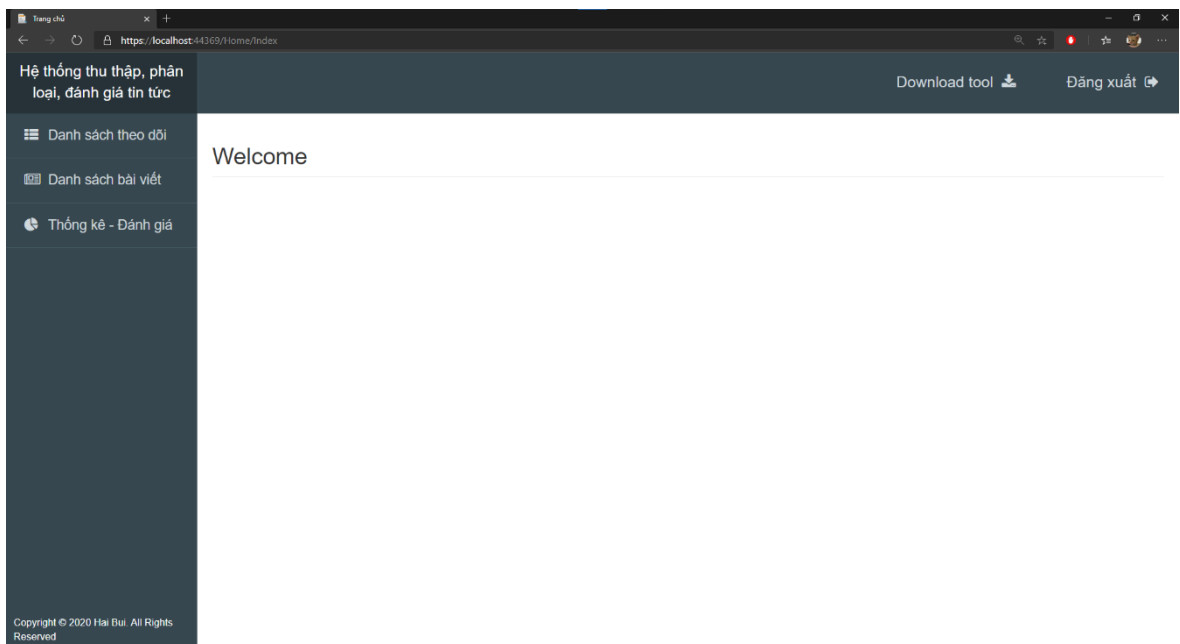
Giao diện cung cấp chức năng đăng nhập dùng để xác nhận người truy cập vào hệ thống.



Hình 3. 40 Giao diện đăng nhập

#### 3.7.2. Giao diện trang chủ

Hiện thị khi quá trình đăng nhập thành công. Phần layout bao gồm 2 khu vực. Khu vực bên trái hiển thị danh mục các chức năng của trang web bao gồm Danh sách theo dõi, Danh sách bài viết, Thống kê – Đánh giá. Khu vực phía trên cùng bên phải là 2 tùy chọn bao gồm Download tool và Đăng xuất.



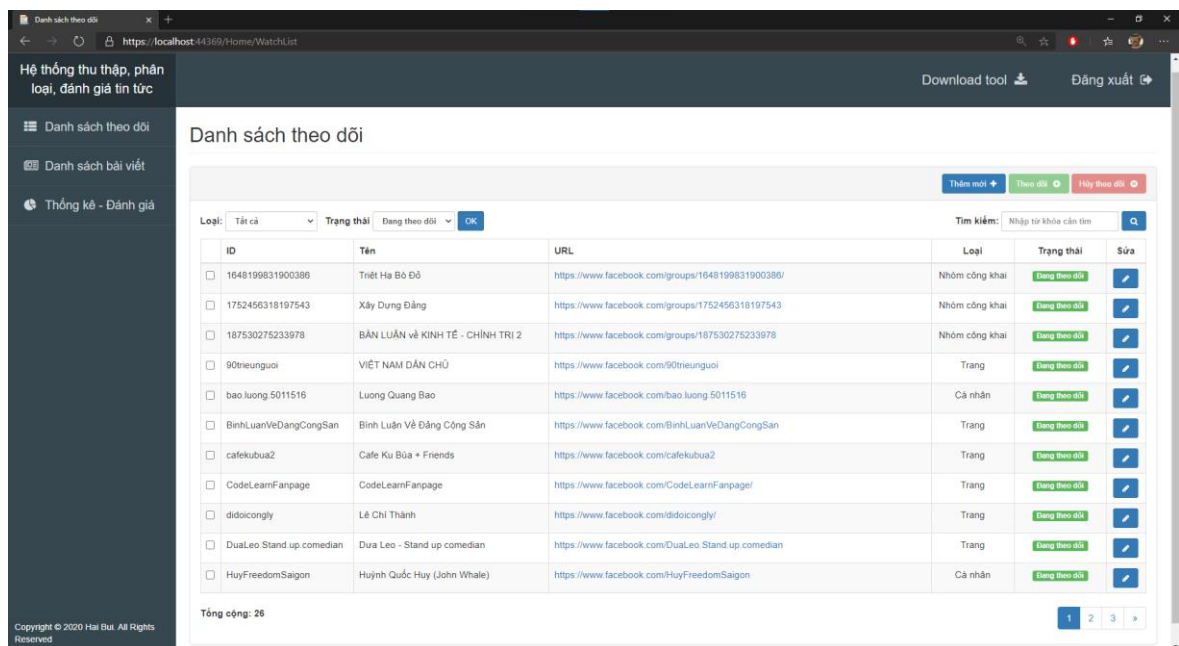
Hình 3. 41 Giao diện trang chủ

Đối với trường hợp chọn Download tool, trình duyệt sẽ điều hướng đến trang chứa công cụ thu thập, phân loại tin tức và tiến hành tải xuống.

Trường hợp chọn Đăng xuất, trình duyệt sẽ điều hướng về trang Đăng nhập.

### 3.7.3. Giao diện danh sách theo dõi

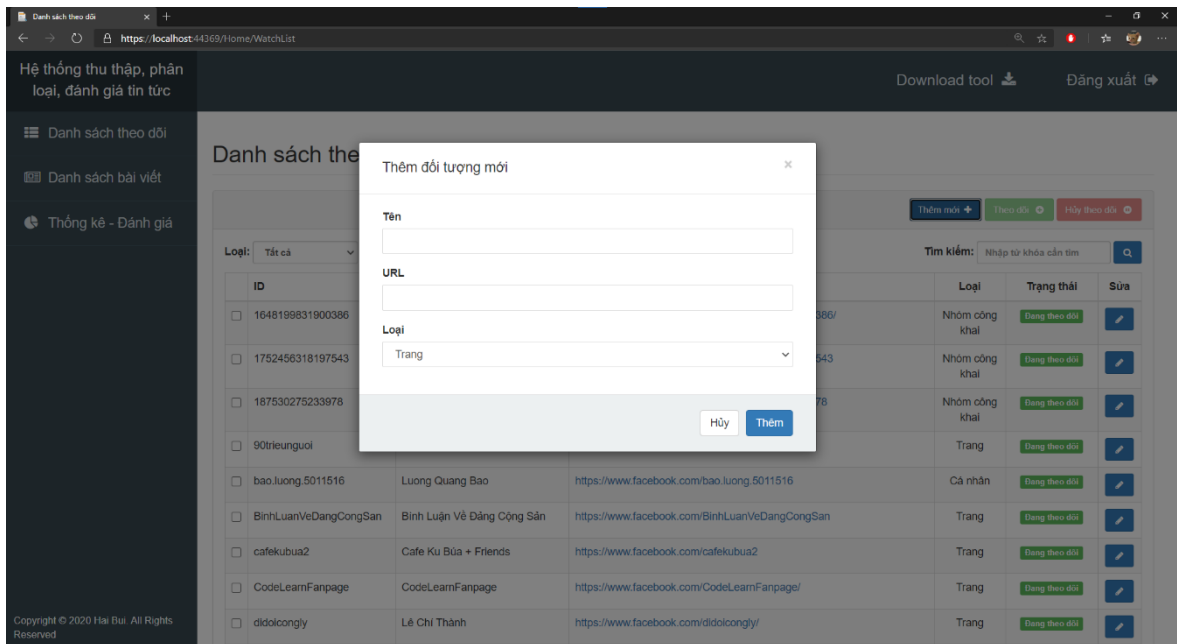
Giao diện cung cấp chức năng quản lý các đối tượng trong danh sách theo dõi. Hiện thị các thông tin về đối tượng bao gồm ID, tên, URL, loại, trạng thái.



Hình 3. 42 Giao diện trang theo dõi

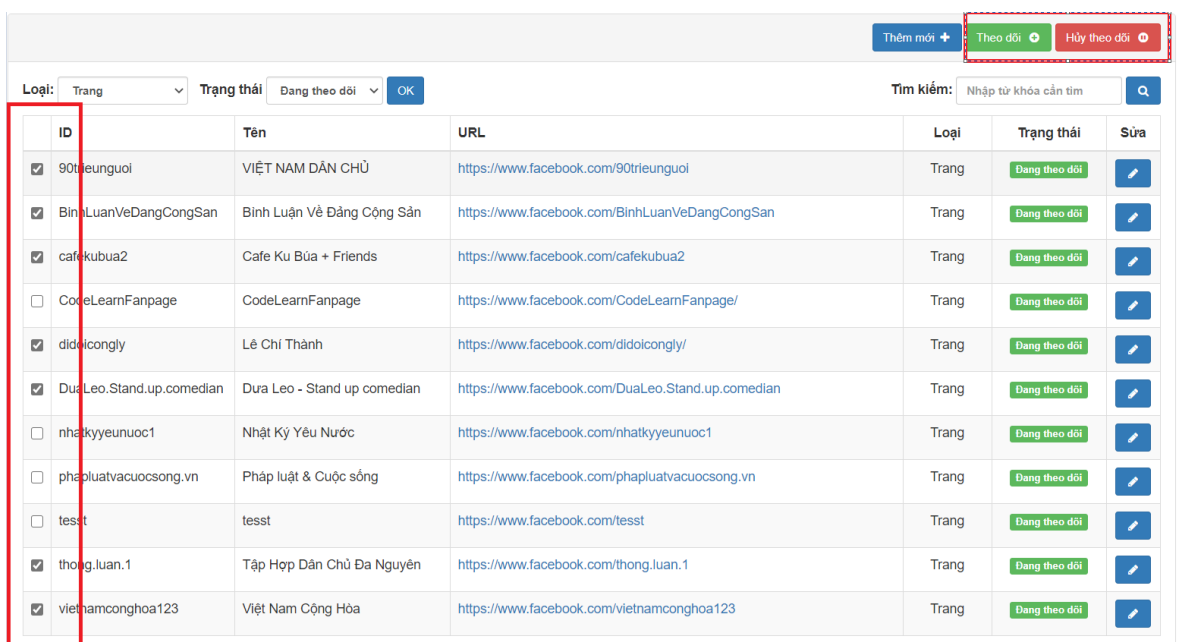
Các thành phần con:

- Nút Thêm mới mở ra một giao diện cho phép thêm một đối tượng vào trong danh sách theo dõi. Dữ liệu nhập vào bao gồm Tên, URL, lựa chọn loại là Trang, Nhóm, hoặc Trang cá nhân.



Hình 3. 43 Giao diện trang thêm đối tượng mới

- Nút Hủy theo dõi cho phép hủy theo dõi một hoặc nhiều đối tượng bằng cách click chọn vào các checkbox của đối tượng đó.
- Tương tự Nút theo dõi thực hiện việc theo dõi lại những đối tượng đã bị hủy theo dõi trước đó



Hình 3. 44 Giao diện thao tác trạng thái

- Các chức năng lọc dữ liệu hiển thị dựa trên Loại và Trạng thái của đối tượng

Loại: Trang

Trạng thái: Đang theo dõi

OK

ID	Tên	URL
<input type="checkbox"/> 90trieungui	VIỆT NAM DÂN CHỦ	<a href="https://www.facebook.com/90trieungui">https://www.facebook.com/90trieungui</a>
<input type="checkbox"/> BinhLuanVeDangCongSan	Bình Luận Về Đảng Cộng Sản	<a href="https://www.facebook.com/BinhLuanVeDangCongSan">https://www.facebook.com/BinhLuanVeDangCongSan</a>
<input type="checkbox"/> cafeclubua2	Cafe Ku Búa + Friends	<a href="https://www.facebook.com/cafeclubua2">https://www.facebook.com/cafeclubua2</a>
<input type="checkbox"/> CodeLearnFanpage	CodeLearnFanpage	<a href="https://www.facebook.com/CodeLearnFanpage/">https://www.facebook.com/CodeLearnFanpage/</a>
<input type="checkbox"/> didoicongly	Lê Chí Thành	<a href="https://www.facebook.com/didoicongly/">https://www.facebook.com/didoicongly/</a>
<input type="checkbox"/> DuaLeo.Stand.up.comedian	Dua Leo - Stand up comedian	<a href="https://www.facebook.com/DuaLeo.Stand.up.comedian">https://www.facebook.com/DuaLeo.Stand.up.comedian</a>
<input type="checkbox"/> nhathkyeunuc1	Nhật Ký Yêu Nước	<a href="https://www.facebook.com/nhathkyeunuc1">https://www.facebook.com/nhathkyeunuc1</a>
<input type="checkbox"/> phapluatvacuocsong.vn	Pháp luật & Cuộc sống	<a href="https://www.facebook.com/phapluatvacuocsong.vn">https://www.facebook.com/phapluatvacuocsong.vn</a>
<input type="checkbox"/> tesst	tesst	<a href="https://www.facebook.com/tesst">https://www.facebook.com/tesst</a>
<input type="checkbox"/> thong.luan.1	Tập Hợp Dân Chủ Đa Nguyên	<a href="https://www.facebook.com/thong.luan.1">https://www.facebook.com/thong.luan.1</a>
<input type="checkbox"/> vietnamconghoa123	Việt Nam Cộng Hòa	<a href="https://www.facebook.com/vietnamconghoa123">https://www.facebook.com/vietnamconghoa123</a>

Tìm kiếm: Nhập từ khóa cần tìm

Loại	Trạng thái	Sửa
Trang	Đang theo dõi	
Trang	Đang theo dõi	
Trang	Đang theo dõi	
Trang	Đang theo dõi	
Trang	Đang theo dõi	
Trang	Đang theo dõi	
Trang	Đang theo dõi	
Trang	Đang theo dõi	
Trang	Đang theo dõi	
Trang	Đang theo dõi	
Trang	Đang theo dõi	

Tổng cộng: 12

1

2

»

Hình 3. 45 Chức năng lọc

- Chức năng tìm kiếm giúp tìm nhanh các đối tượng dựa theo từ khóa.

Thêm mới +

Theo dõi

Hủy theo dõi

Loại: 

Tất cả

Trạng thái: 

Tất cả

OK

Tìm kiếm: 

viết tân


Q

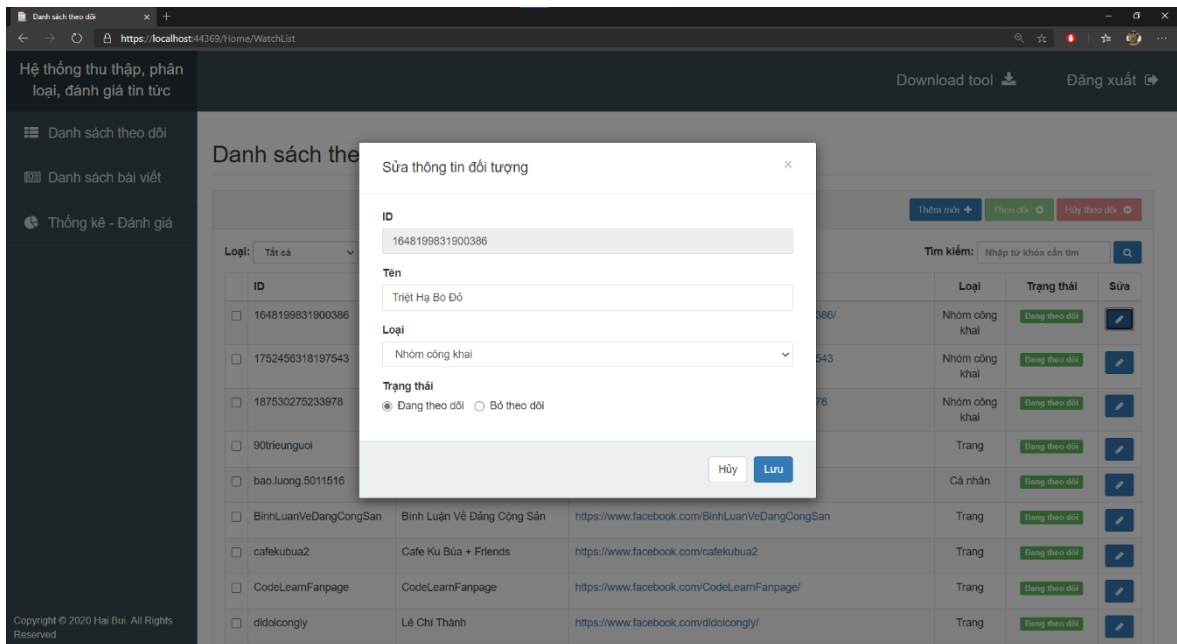
ID	Tên	URL	Loại	Trạng thái	Sửa
<input type="checkbox"/> viettan	Việt Tân	https://www.facebook.com/viettan	Trang	Đang theo dõi	

Tổng cộng: 1

1

Hình 3. 46 Chức năng tìm kiếm

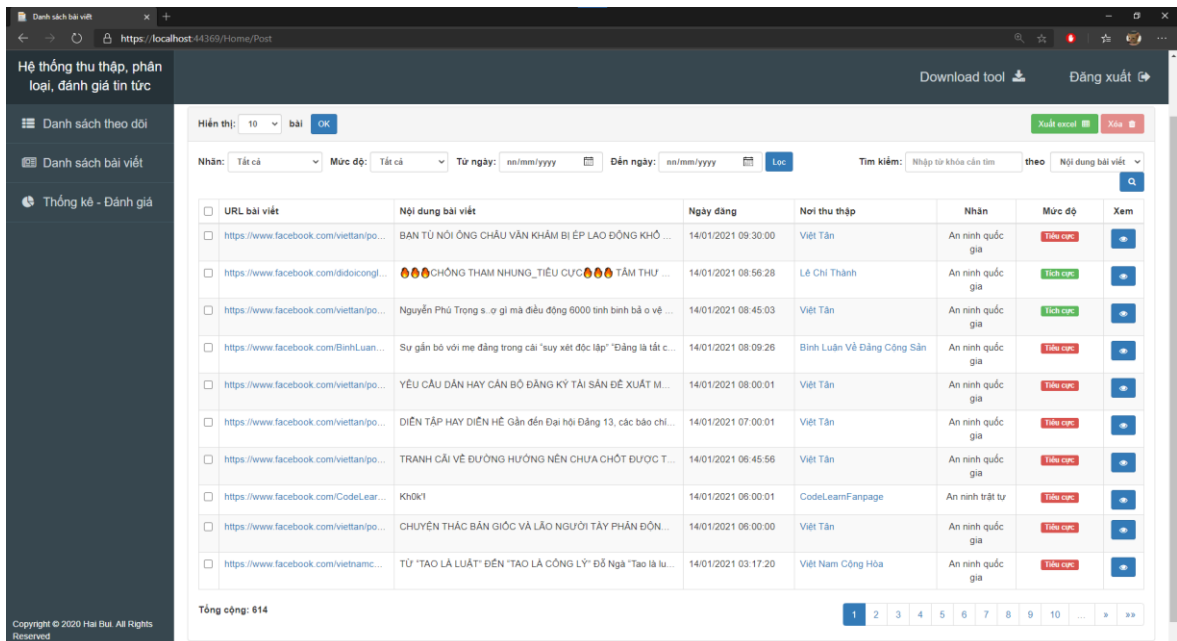
- Nút Sửa  hiển thị giao diện chứa thông tin chi tiết của đối tượng, cung cấp chức năng sửa Tên, Loại và Trạng thái của đối tượng.



Hình 3. 47 Chức năng sửa thông tin

### 3.7.4. Giao diện danh sách bài viết

Giao diện cung cấp chức năng quản lý các bài viết trong danh sách các bài viết đã thu thập. Hiển thị các thông tin về bài viết bao gồm URL, nội dung bài viết, ngày đăng, nơi thu thập bài viết, nhãn chủ đề, mức độ.



Hình 3. 48 Trang quản lý bài viết

Các thành phần con:

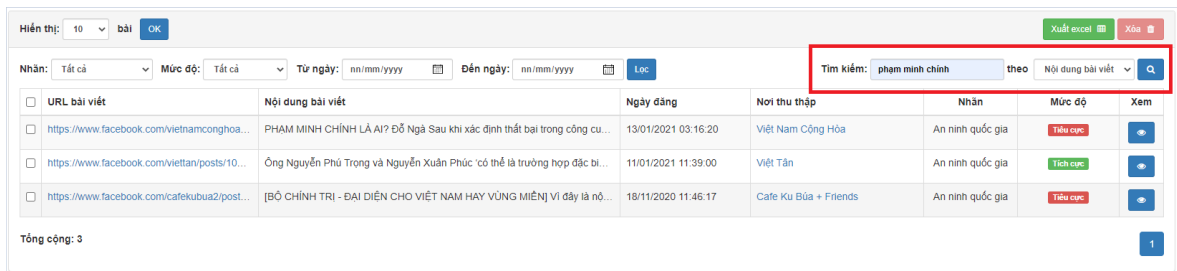
- Tùy chọn hiển thị số lượng bài viết sẽ xuất hiện khi phân trang.

Hình 3. 49 Hiển thị số lượng

- Chức năng lọc bài viết với các tùy chọn về nhãn chủ đề, mức độ, lọc các bài viết từ ngày đến ngày.

Hình 3. 50 Lọc bài viết


- Chức năng tìm kiếm bài viết dựa trên từ khóa, tìm kiếm dựa trên nơi thu thập bài viết đó.

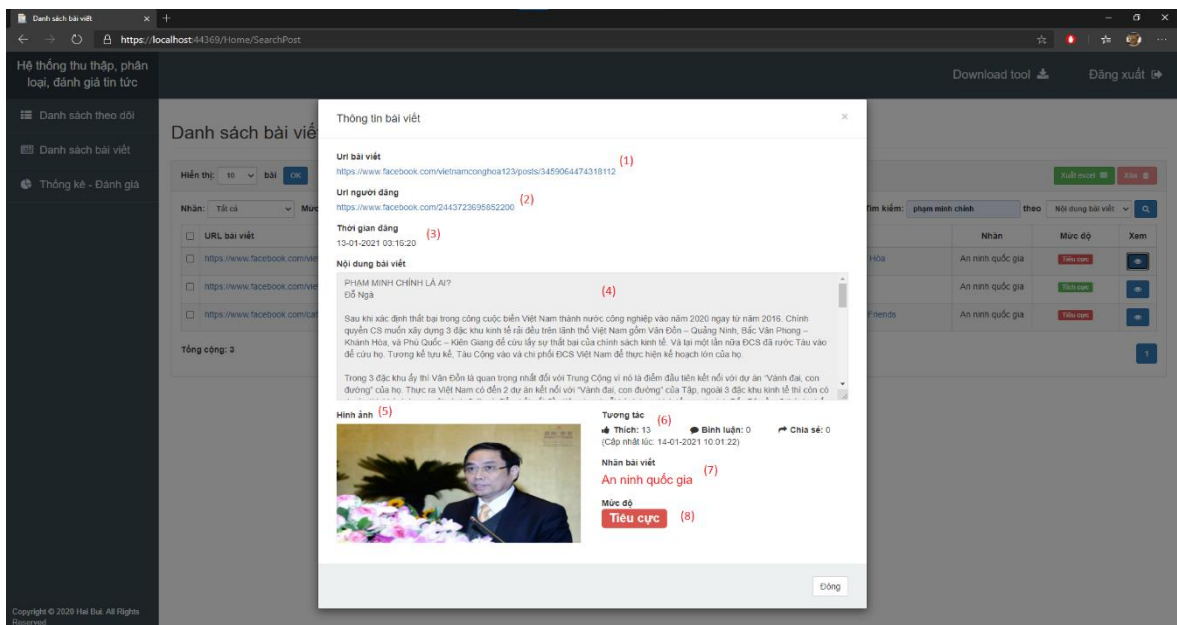


URL bài viết	Nội dung bài viết	Ngày đăng	Nơi thu thập	Nhãn	Mức độ	Xem
<a href="https://www.facebook.com/vietnamconghoa...">https://www.facebook.com/vietnamconghoa...</a>	PHẠM MINH CHÍNH LÀ AI? Đỗ Nga Sau khi xác định thất bại trong công cu...	13/01/2021 03:16:20	Việt Nam Cộng Hòa	An ninh quốc gia	Tiêu cực	
<a href="https://www.facebook.com/vietnam/posts/10...">https://www.facebook.com/vietnam/posts/10...</a>	Ông Nguyễn Phú Trọng và Nguyễn Xuân Phúc 'có thể là trường hợp đặc bi...	11/01/2021 11:39:00	Việt Tân	An ninh quốc gia	Tích cực	
<a href="https://www.facebook.com/cafehubua2/post...">https://www.facebook.com/cafehubua2/post...</a>	[BỘ CHÍNH TRỊ - ĐÀI DIỄN CHO VIỆT NAM HAY VÙNG MIỀN] Vì đây là nỏ...	18/11/2020 11:46:17	Cafe Ku Bua + Friends	An ninh quốc gia	Tiêu cực	

Tổng cộng: 3

Hình 3. 51 Tìm kiếm bài viết

- Nút xem  hiển thị giao diện chứa thông tin chi tiết của bài viết bao gồm:
  - (1) URL của bài viết, khi click chọn trình duyệt sẽ mở tab mới dẫn thẳng đến bài viết trên Facebook.
  - (2) URL người đăng, khi click chọn sẽ dẫn đến trang Facebook của đối tượng đăng tin.
  - (3) Thời gian mà đối tượng đăng tin.
  - (4) Nội dung chi tiết của bài viết.
  - (5) Hình ảnh trong bài viết (nếu có).
  - (6) Mức độ tương tác: số lượt thích, bình luận, chia sẻ bài viết.
  - (7) Nhãn chủ đề của bài viết.
  - (8) Nhãn mức độ của bài viết.




**Thông tin bài viết**

Url bài viết (1): <https://www.facebook.com/vietnamconghoa123/posts/3459064474318112>

Url người đăng (2): <https://www.facebook.com/2443723695652200>

Thời gian đăng (3): 13-01-2021 03:16:20

Nội dung bài viết (4): PHẠM MINH CHÍNH LÀ AI? Đỗ Nga

Hình ảnh (5): 

Tương tác (6): Thích: 13, Bình luận: 0, Chia sẻ: 0

Nhãn bài viết (7): An ninh quốc gia

Mức độ (8): Tiêu cực

Hình 3. 52 Thông tin bài viết

## - Chức năng xóa 1 hoặc nhiều bài viết

Hiển thị: 10 bài OK

Xuất excel Xóa

Nhân: Tất cả Mức độ: Tất cả Từ ngày: nn/mm/yyyy Đến ngày: nn/mm/yyyy Lọc

Tìm kiếm: Nhập từ khóa cần tìm theo Nội dung bài viết Q

<input type="checkbox"/>	URL bài viết	Nội dung bài viết	Ngày đăng	Nơi thu thập	Nhân	Mức độ	Xem
<input checked="" type="checkbox"/>	<a href="https://www.facebook.com/viettan/posts/10...">https://www.facebook.com/viettan/posts/10...</a>	FACEBOOK VÀ TWITTER BỊ CHẶN VÌ KIỂM DUYỆT THÔNG TIN CỦA KH...	13/01/2021 06:00:02	Việt Tân	Công nghệ	Bình thường	<a href="#">👁</a>
<input checked="" type="checkbox"/>	<a href="https://www.facebook.com/CodeLearnFanp...">https://www.facebook.com/CodeLearnFanp...</a>	Chữ "g" trong "Lập trình viên" tượng trưng cho sự "GIÁU CỎ" :)	13/01/2021 06:00:01	CodeLearnFanpage	Phim ảnh	Bình thường	<a href="#">👁</a>
<input type="checkbox"/>	<a href="https://www.facebook.com/viettan/posts/10...">https://www.facebook.com/viettan/posts/10...</a>	TRẦN LONG PHI - TÙ NHÂN LƯƠNG TÂM TRỀ NHẤT Lại thêm một cái tề...	13/01/2021 05:00:00	Việt Tân	An ninh quốc gia	Tiêu cực	<a href="#">👁</a>
<input type="checkbox"/>	<a href="https://www.facebook.com/viettan/posts/10...">https://www.facebook.com/viettan/posts/10...</a>	THÔNG DIỆP CỦA PHẢI ĐOÀN LIÊN MINH CHÂU ÂU Brussels, 12/01/202...	13/01/2021 04:33:25	Việt Tân	An ninh quốc gia	Tiêu cực	<a href="#">👁</a>
<input type="checkbox"/>	<a href="https://www.facebook.com/vietnamconghoa...">https://www.facebook.com/vietnamconghoa...</a>	PHẠM MINH CHÍNH LÃ AI? Đỗ Ngà Sau khi xác định thất bại trong công c...	13/01/2021 03:16:20	Việt Nam Cộng Hòa	An ninh quốc gia	Tiêu cực	<a href="#">👁</a>
<input type="checkbox"/>	<a href="https://www.facebook.com/vietnamconghoa...">https://www.facebook.com/vietnamconghoa...</a>	FB Ngạt Ngày 11/1/2021 Đài Loan phát hành hộ chiếu mới in rõ tên Trung H...	13/01/2021 03:15:32	Việt Nam Cộng Hòa	An ninh quốc gia	Tiêu cực	<a href="#">👁</a>
<input type="checkbox"/>	<a href="https://www.facebook.com/vietnamconghoa...">https://www.facebook.com/vietnamconghoa...</a>	Nhà báo Nguyễn Khanh HỒM NAY TẠI WASHINGTON D.C. 1- Mười một gi...	13/01/2021 03:14:42	Việt Nam Cộng Hòa	An ninh quốc gia	Tiêu cực	<a href="#">👁</a>
<input checked="" type="checkbox"/>	<a href="https://www.facebook.com/vietnamconghoa...">https://www.facebook.com/vietnamconghoa...</a>	Nguyễn Thị Bích Hậu Nhà Báo Giờ nhìn vào các trường đào tạo Y khoa của ...	13/01/2021 03:13:43	Việt Nam Cộng Hòa	Giáo dục	Bình thường	<a href="#">👁</a>
<input type="checkbox"/>	<a href="https://www.facebook.com/vietnamconghoa...">https://www.facebook.com/vietnamconghoa...</a>	FB Nguyễn Thủy Dương CANG BỊ LỬA??? Càng cường chế một mô nhà n...	13/01/2021 03:12:48	Việt Nam Cộng Hòa	An ninh quốc gia	Tiêu cực	<a href="#">👁</a>
<input checked="" type="checkbox"/>	<a href="https://www.facebook.com/BinhLuanVeDan...">https://www.facebook.com/BinhLuanVeDan...</a>	Tin cuối ngày: Thiệt hại về giá trị vốn hóa thị trường của Big Tech hôm 11/1 ...	12/01/2021 22:47:28	Bình Luận Về Đảng Cộng Sản	Kinh Doanh	Bình thường	<a href="#">👁</a>

Tổng cộng: 614

« 1 2 3 4 5 6 7 8 9 10 ... »

Hình 3. 53 Xóa bài viết

## - Chức năng xuất thông tin các bài viết ra file excel theo nội dung đã lọc hoặc tìm kiếm.

Hiển thị: 10 bài OK

Xuất excel Xóa

Nhân: Tất cả Mức độ: Tất cả Từ ngày: nn/mm/yyyy Đến ngày: nn/mm/yyyy Lọc

Tìm kiếm: Nhập từ khóa cần tìm theo Nội dung bài viết Q

<input type="checkbox"/>	URL bài viết	Nội dung bài viết	Ngày đăng	Nơi thu thập	Nhân	Mức độ	Xem
<input type="checkbox"/>	<a href="https://www.facebook.com/2740592832872...">https://www.facebook.com/2740592832872...</a>	Hôm 3/1/2021 có con mụ kia nói: ĐBQH phải gần dân. Ủy, vậy trước giờ ĐB...	13/01/2021 05:53:00	Luong Quang Bao	An ninh quốc gia	Tiêu cực	<a href="#">👁</a>
<input type="checkbox"/>	<a href="https://www.facebook.com/2740569322874...">https://www.facebook.com/2740569322874...</a>	Chúc mừng anh Trần nhật quang nùn đã nhận được giải thưởng cái bô văn...	13/01/2021 05:00:00	Luong Quang Bao	Âm nhạc	Bình thường	<a href="#">👁</a>
<input type="checkbox"/>	<a href="https://www.facebook.com/2740566662875...">https://www.facebook.com/2740566662875...</a>	Khi đỉnh cao trí tuệ đi kiểm tra an toàn vệ sinh thực phẩm thời đại 4.0 rúc rở...	13/01/2021 04:54:00	Luong Quang Bao	An ninh quốc gia	Tích cực	<a href="#">👁</a>
<input type="checkbox"/>	<a href="https://www.facebook.com/2740559382875...">https://www.facebook.com/2740559382875...</a>	CẤT BĂNG CẦU THĂNG THIÊN. Trời rét dưới mười độ Cái đy ở trong quần...	13/01/2021 04:39:00	Luong Quang Bao	An ninh quốc gia	Tiêu cực	<a href="#">👁</a>
<input type="checkbox"/>	<a href="https://www.facebook.com/2740550476210...">https://www.facebook.com/2740550476210...</a>	Có đứa em mới hỏi: hèn với giặc ác với dân là sao em không hiểu vậy anh. ...	13/01/2021 04:23:00	Luong Quang Bao	An ninh quốc gia	Tiêu cực	<a href="#">👁</a>
<input type="checkbox"/>	<a href="https://www.facebook.com/2740530972878...">https://www.facebook.com/2740530972878...</a>	Anh chị em sử dụng Facebook lưu ý: không nên viết tên cùng công ông cổ n...	13/01/2021 03:46:00	Luong Quang Bao	An ninh quốc gia	Tiêu cực	<a href="#">👁</a>
<input type="checkbox"/>	<a href="https://www.facebook.com/2740518702879...">https://www.facebook.com/2740518702879...</a>	Lựa chọn nhân sự lấy đức làm gốc. Hay là lựa chọn bọn ba đời bản cổ nồn...	13/01/2021 03:21:00	Luong Quang Bao	An ninh quốc gia	Tiêu cực	<a href="#">👁</a>
<input type="checkbox"/>	<a href="https://www.facebook.com/2740514792880...">https://www.facebook.com/2740514792880...</a>	Cả nhà cho em hỏi thiết: nhà nước của dân do dân và vì dân sao không cho...	13/01/2021 03:14:00	Luong Quang Bao	An ninh quốc gia	Tiêu cực	<a href="#">👁</a>
<input type="checkbox"/>	<a href="https://www.facebook.com/2740282812903...">https://www.facebook.com/2740282812903...</a>	AI thông thái trả lời em với: Phát ngôn viên bộ ngoại giao lên tivi ra rả tuyên ...	12/01/2021 20:07:00	Luong Quang Bao	An ninh quốc gia	Tiêu cực	<a href="#">👁</a>
<input type="checkbox"/>	<a href="https://www.facebook.com/2740272046237...">https://www.facebook.com/2740272046237...</a>	CẢM ƠN ẢNH MẮT VIỆT NAM ❤️❤️❤️ Tôi yêu tổ quốc Việt Nam ❤️❤️❤️ ...	12/01/2021 19:48:00	Luong Quang Bao	An ninh quốc gia	Tiêu cực	<a href="#">👁</a>

Tổng cộng: 28

1 2 3 »

Hình 3. 54 Xuất file excel

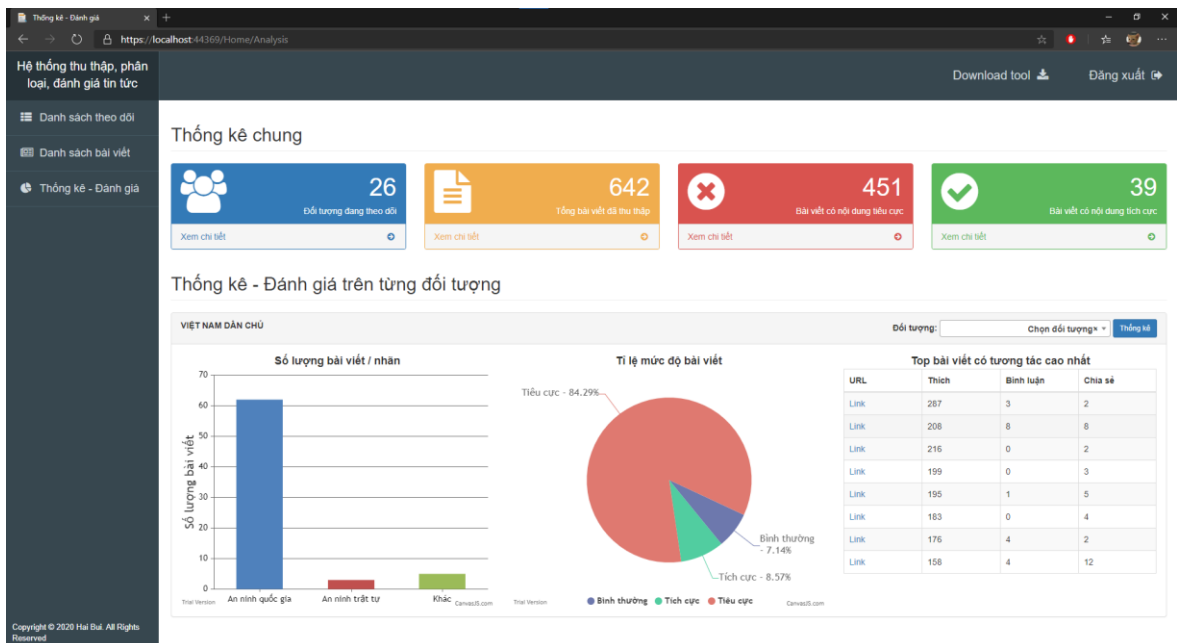


	A	B	C	D	E	F	G	H	I	J	K
	PostID	PostUrl	UserUrl	PostContent	Image	UploadTime	CrawledTime	Likes	Comments	Shares	Facebook
1	14012021173312XW19PN	https://www.facebook.com/2740592832872506/	https://www.facebook.com/100007654472749/	Hôm 3/1/2021 có con mụ kia nói: ĐBQH phải gần dân. Ủ, vậy trước giờ ĐBQH gần cái thứ gì vậy mụ ngu dân? Mụ không nói thì ai đầu biết mụ dân! Hôm 3/1/2021 có con mụ kia nói: ĐBQH phải gần dân. Ủ, vậy trước giờ ĐBQH gần cái thứ gì vậy mụ ngu dân? Mụ không nói thì ai đầu biết mụ dân!		1/13/2021 5:53	1/14/2021 17:34	38	1	0	bao.luong.50
2	14012021173312KV8FGV	https://www.facebook.com/2740569322874857/	https://www.facebook.com/100007654472749/	Chúc mừng anh Trần nhật quang nùn đã nhận được giải thưởng cái bô vàng. Vâng! xin chúc mừng anh.		1/13/2021 5:00	1/14/2021 17:34	77	8	2	bao.luong.50
3	14012021173312ZIV53E	https://www.facebook.com/2740566662	https://www.facebook.com/100007654472749/	Khi đỉnh cao trí tuệ đi kiểm tra an toàn vệ sinh thực phẩm thời đại 4.0 rục rở.		1/13/2021 4:54	1/14/2021 17:34	74	22	5	bao.luong.50
4				CẮT BĂNG CẦU THĂNG THIÊN. Trời rét dưới mười độ Cái ấy ở trong quần Bọc ba bốn lần vải Mả chỉ còn nửa phần. Thế mà các quan sản Đì cắt băng khánh thành Bát mẩy đưa con gái... More Lên cầu cho nó hành. Mẹ kiếp, rết thì rết Thống cầu là phải oai Phải hót gơ tha thướt Quần mỏng và áo dài. Cắt cái băng gì đấy Toàn quan to cấp trên Các em phải hấp dẫn Thi cầu kia mới bền. Tiên sư bố chúng nó Độc ác và khùng điên Chắc là xong cái lễ Mây châu này thăng thiên. Thờ của fb Ưng Hoàng		1/13/2021 4:39	1/14/2021 17:34	45	11	3	bao.luong.50
5	14012021173312479U8L	https://www.facebook.com/2740559382875851/	https://www.facebook.com/100007654472749/	Cá đưa em mới bắt: hàn uoi giặc ác uoi dân là							

Hình 3. 55 Thông tin file excel

### 3.7.5. Giao diện thống kê, đánh giá

Giao diện cung cấp chức năng thống kê, đánh giá giữa trên các đối tượng đang theo dõi và danh sách các bài viết đã thu thập được.

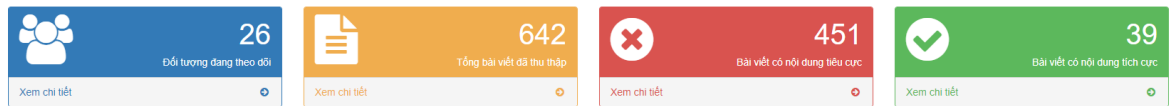


Hình 3. 56 Trang thống kê, đánh giá

Các thành phần con:

- Khu vực thống kê chung hiển thị thông kê về số lượng các đối tượng đang theo dõi, tổng số lượng các bài viết, số lượng bài viết có mức độ tích cực, tiêu cực.

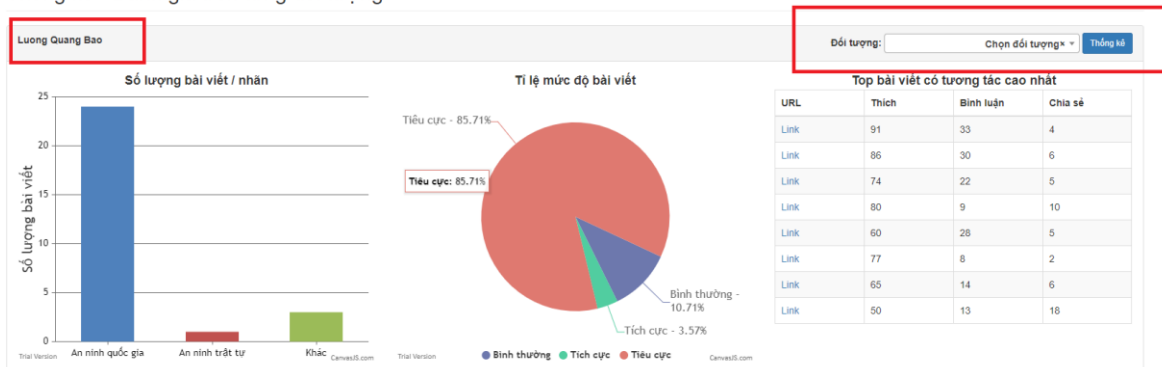
Thống kê chung



Hình 3. 57 Khu vực thống kê chung

- Khu vực thống kê chi tiết hiển thị biểu đồ đánh giá về xung hướng đăng tin của một đối tượng cụ thể qua các tiêu chí về số lượng bài viết/ nhân, tỉ lệ mức độ các bài viết mà đối tượng đó đã đăng. Ngoài ra còn có mục thống kê về top các bài viết có lượng tương tác cao nhằm nhận định độ ảnh hưởng của đối tượng đối với cộng đồng.

Thống kê - Đánh giá trên từng đối tượng



Hình 3. 58 Khu vực thống kê chi tiết

## CHƯƠNG 4. KẾT LUẬN

### 4.1. Kết quả đạt được

- Nghiên cứu các giải pháp về việc thu thập tin tức tự động, nguồn tin lấy từ mạng xã hội Facebook.
- Tìm hiểu về các cơ chế, phương pháp phân loại, đánh giá tin tức.
- Áp dụng thuật toán Naive Bayes và các bộ thư viện hỗ trợ xây dựng một công cụ có chức năng thu thập tin tức tự động, đồng thời phân loại, đánh giá tin tức đó theo hai chủ đề chính là An ninh quốc gia và An ninh trật tự.
- Áp dụng các kiến thức về lập trình web để xây dựng WebAPI có nhiệm vụ trung gian giao tiếp giữa các thành phần trong hệ thống.
- Xây dựng và triển khai thành công cơ sở dữ liệu và website quản lý tin tức có thể cài đặt và vận hành ổn định trên server thử nghiệm.

### 4.2. Ưu điểm

- Hệ thống được xây dựng mang tính thực tiễn cao, đáp ứng tương đối đầy đủ các yêu cầu đề ra.
- Hệ thống được thiết kế dưới dạng mô-đun, dễ dàng trong quá trình nâng cấp, bảo trì.
- Có tiềm năng trở thành công cụ hỗ trợ đắc lực cho cơ quan điều tra về lĩnh vực An ninh mạng.

### 4.3. Hạn chế

- Tập huấn luyện chưa nhận diện được chính xác hoàn toàn các nhãn chủ đề, mức độ. Kết quả dự đoán chỉ ở mức tương đối.
- Facebook thường xuyên cập nhật giao diện, rà soát các công cụ có tính năng quét tự động dẫn tới việc triển khai đôi lúc gặp nhiều khó khăn.
- Một số tính năng của hệ thống còn đang trong quá trình nghiên cứu và phát triển, chưa hoàn thiện 100% dẫn đến việc có thể xảy ra lỗi không mong muốn.

#### **4.4. Hướng phát triển**

- Tham gia triển khai thực nghiệm hệ thống trên máy chủ nội bộ của công an tỉnh Bình Thuận.
- Cải tiến các mô-đun nhằm cải thiện độ chính xác, tốc độ thực thi và tăng cường khả năng vận hành ổn định, lâu dài.
- Tiếp tục nghiên cứu, xây dựng tập huấn luyện và mô hình học máy phân loại với độ chính xác cao hơn.
- Mở rộng quy mô, phạm vi hệ thống ra các lĩnh vực khác như thu thập và phân loại tin tức và cảnh cáo về tin giả.

## TÀI LIỆU THAM KHẢO

### Tiếng Việt

- [1] Trịnh Tấn Đạt, Bài giảng máy học nâng cao (giáo trình nội bộ), Đại học Sài Gòn
- [2] Nhất Nghệ, Nguyễn Nghiệm, Giáo Trình ASP.NET MVC 5 from <<https://cuongquach.com/giao-trinh-asp-net-mvc-5-nhat-nghe-pdf.html>>

### Website

- [3] Nguyễn Văn Hiếu (2020), Phân loại văn bản tiếng Việt sử dụng Machine Learning from <<https://nguyenvanhieu.vn/phan-loai-van-ban-tieng-viet/>>
- [4] Microsoft, ASP.NET WebAPI, from<<https://docs.microsoft.com/en-us/aspnet/web-api/>>