VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

**BỘ MÔN AN TOÀN THÔNG TIN - CÔNG NGHỆ MẠNG**

**BÁO CÁO BÀI TẬP LỚN**

**Đề tài: Xây dựng công cụ tổng hợp tin tức**

**đa phương tiện**

|  |  |
| --- | --- |
| **Học viên thực hiện:** | **Hà Công Quốc Bảo - ANHTTT**  **Nông Phương Trang - ANHTTT** |
|  |  |
| **Giáo viên hướng dẫn** | **2// GV.TS.Phan Việt Anh** |

**Hà Nội - 2024**

# 1. Xây dựng kế hoạch và mục tiêu trinh sát

Mục đích của hoạt động OSINT: Theo dõi các đối tượng phản động (các hành động chống phá, định hướng dư luận, diễn biến hòa bình,... trên không gian mạng có thể gây bạo động, ảnh hưởng đến an ninh chính trị, trật tự an toàn xã hội) và các sự kiện đang được chú ý (tập trung vào các sự kiện liên quan đến liên quan đến tình hình chính trị, quân sự,...)

Mục tiêu của đề tài: xây dựng một công cụ Open Source Intelligence (OSINT) nhằm thực hiện nhiệm vụ tổng hợp thông tin trên Internet và các trang mạng xã hội về các sự kiện và các đối tượng trinh sát.

Phạm vi trinh sát: Các thông tin được đăng trên các nền tảng lớn như Google, Facebook và Youtube về những sự kiện được chỉ định và hoạt động của một số tổ chức phản động đang là mục tiêu trinh sát.

Nguồn dữ liệu: Các nền tảng Google, Facebook, YouTube và các trang báo điện tử liên quan đến đối tượng trinh sát hoặc đang theo dõi.

- ***Google***: Theo dõi, thu thập các thông tin, bài báo có thể tìm thấy qua công cụ tìm kiếm của Google (Liên quan đến đối tượng hoặc sự kiện đang trinh sát)

- ***Facebook***: Theo dõi, thu thập các thông tin từ nội dung các bài đăng trên Facebook (Các bài viết được đăng tải bởi các đối tượng hay có liên quan đến sự kiện đang trinh sát)

- ***Youtube***: Khai thác thông tin từ các video được đăng tải trên Youtube (Các video được đăng tải bởi các đối tượng hay có liên quan đến sự kiện đang trinh sát)

- ***Các trang báo điện tử*** (Việt Tân, BBC Tiếng Việt, RFA Tiếng Việt, Dân trí, VnExpress, ...): Khai thác thông tin từ các bài báo được đăng tải trên trang báo điện tử phải theo dõi hoặc các trang báo điện tử của thế lực thù địch, phản động.

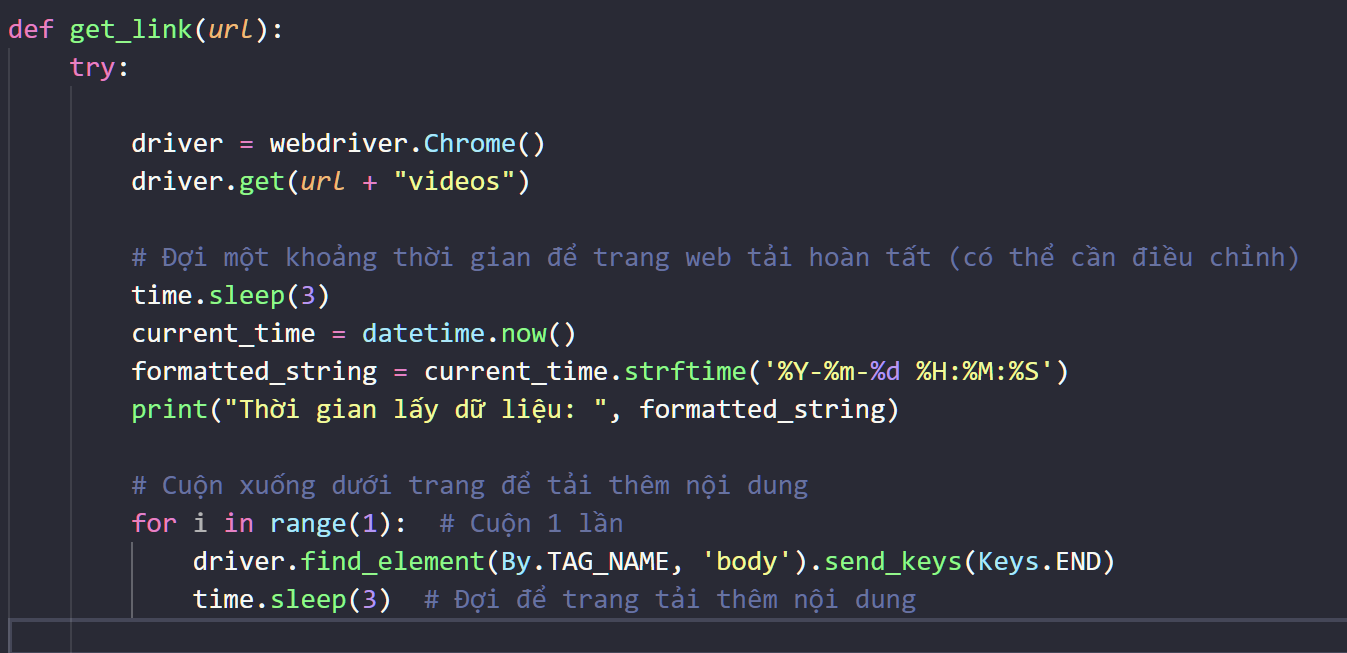
# 2. Thu thập dữ liệu

* Kỹ thuật thu thập dữ liệu:

- Ngôn ngữ lập trình: ***python***

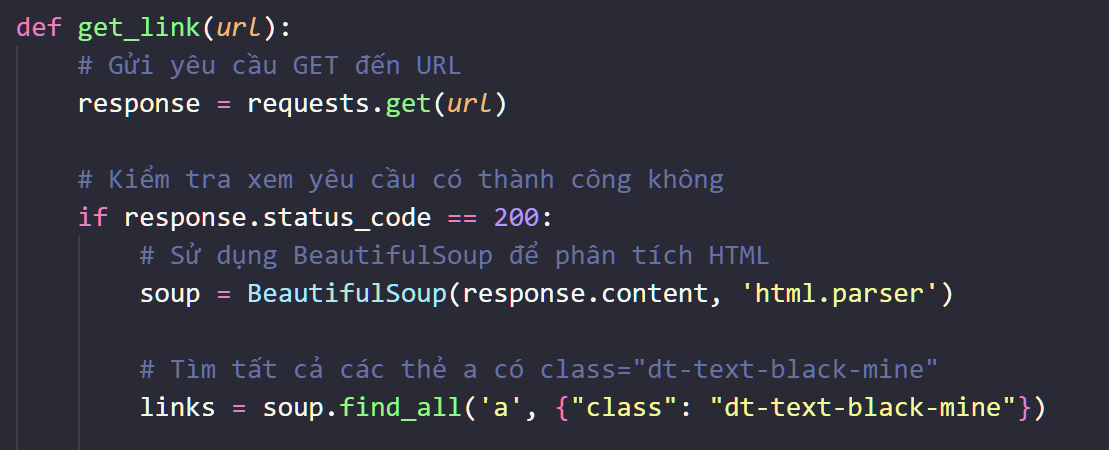
- Sử dụng công nghệ và công cụ có sẵn:

+ ***Selenium***: Tự động hóa trình duyệt web, tương tác với các trang web động



Hình . Hàm sử dụng thư viện Selenium để tự động điều khiển trình duyệt web

+ ***BeautifulSoup***: Phân tích cú pháp và trích xuất dữ liệu từ các tài liệu HTML và XML



Hình . Hàm sử dụng BeautifulSoup để phân tích HTML

+ ***Pytube***: Tải và trích xuất dữ liệu từ video trên YouTube

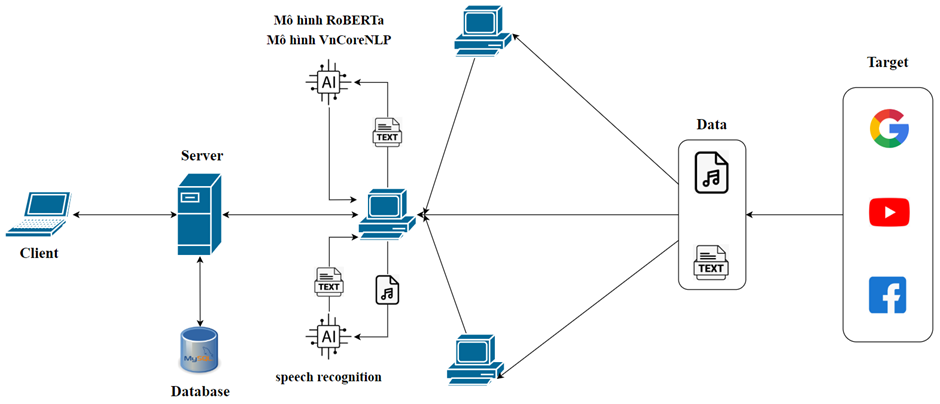


Hình . Hàm sử dụng Pytube trích xuất dữ liệu từ YouTube

* Lưu trữ dữ liệu: Dữ liệu được lưu trữ trong cơ sở dữ liệu ***MySQL*** để dễ dàng truy cập và xử lý sau này.

# 3. Xử lý và phân tích dữ liệu

**a. Mô hình xử lý dữ liệu**

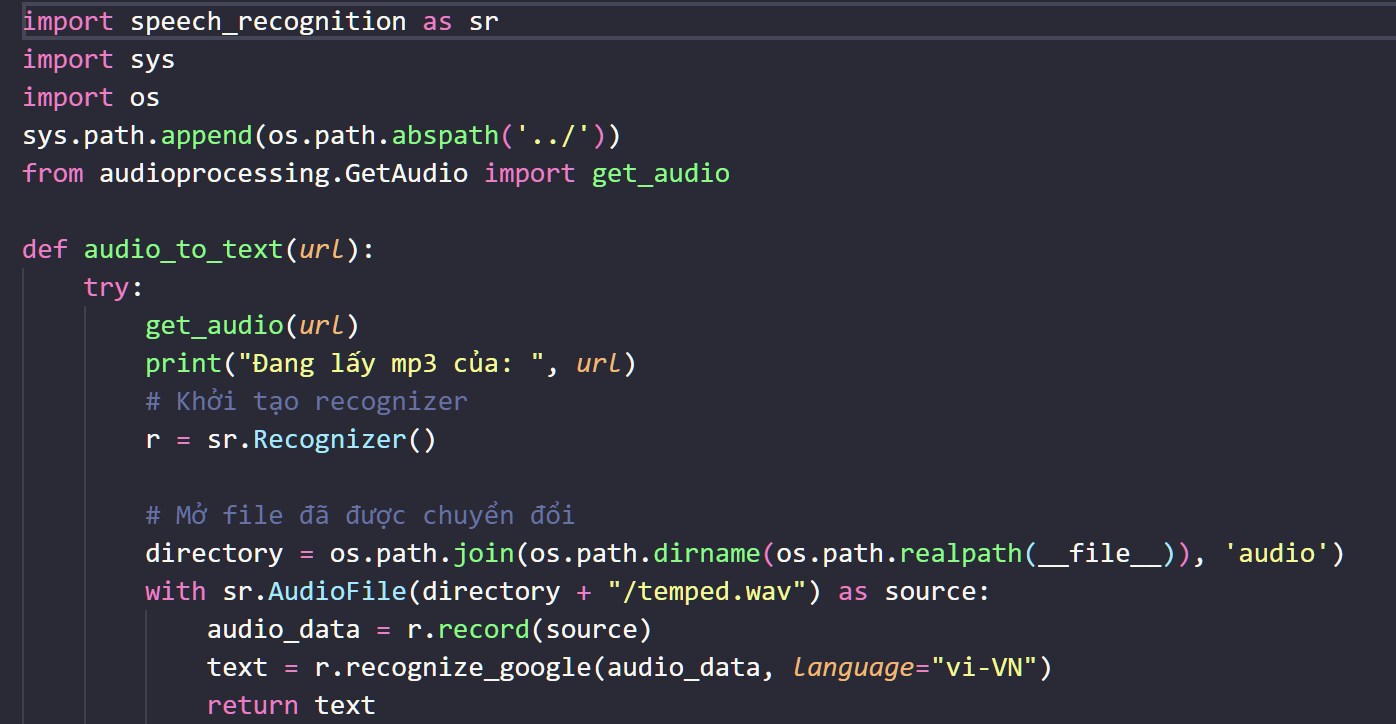


Hình . Mô hình thu thập và xử lý dữ liệu

**b. Phân tích dữ liệu**

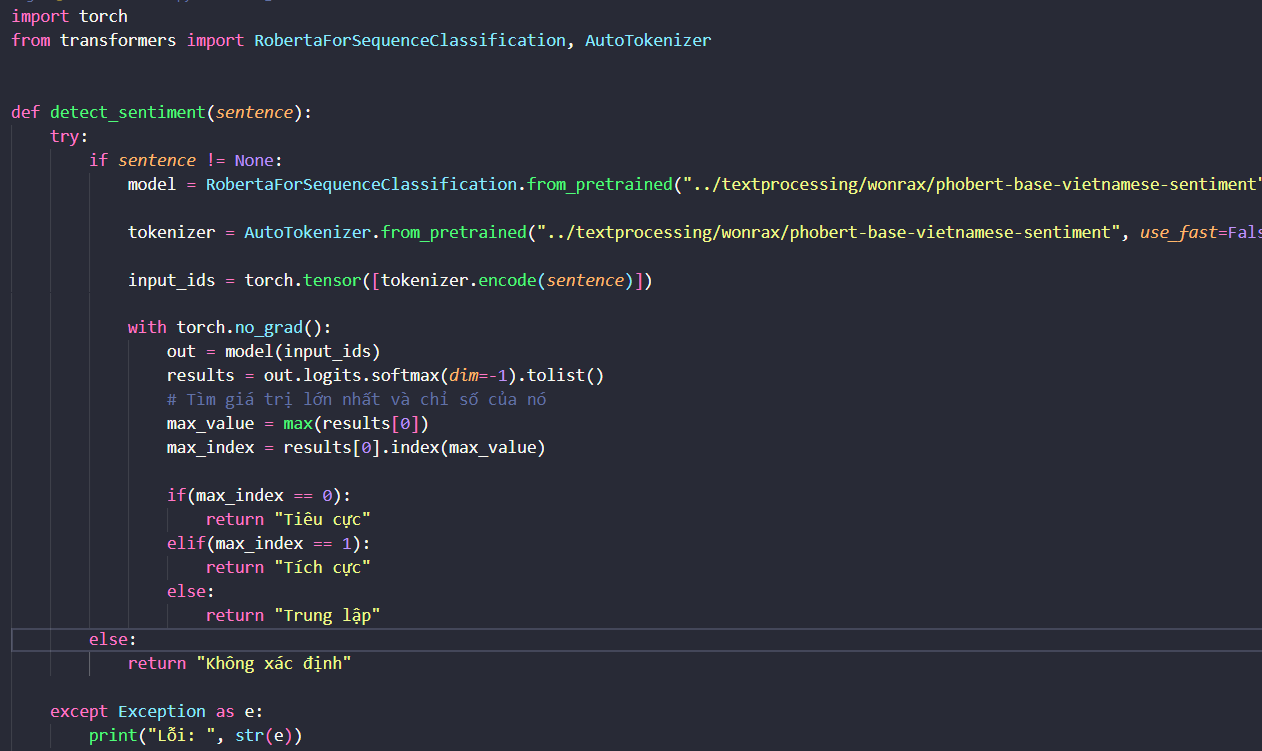
- Công cụ dùng để xử lý và phân tích dữ liệu:

+ Sử dụng ***Speech Recognition***: Chuyển đổi âm thanh thành văn bản



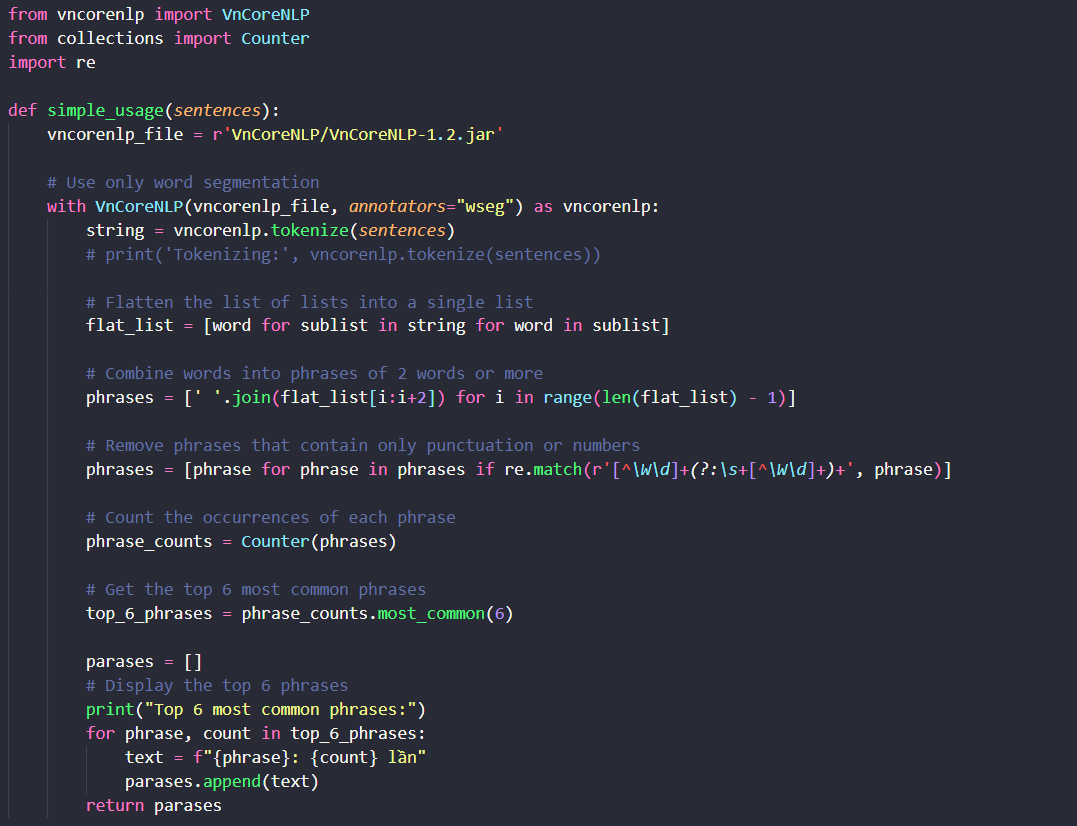
Hình . Hàm chuyển đổi âm thanh thành văn bản

+ Sử dụng mô hình ***RoBERTa***: Phân tích cảm xúc của một nội dung.



Hình . Hàm phân tích cảm xúc

+ Sử dụng mô hình ***VnCoreNLP***: Phân tích thành các cụm từ có nghĩa, từ đó nhận dạng chủ đề và số lần xuất hiện của mỗi cụm từ trong các bài viết.



Hình . Hàm phân tích nội dung và số lần xuất hiện

- Kỹ thuật xử lý và phân tích dữ liệu:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Nguồn** | **Dữ liệu thu được** | **Sử dụng**  **Speech Recognition** | **Sử dụng mô hình RoBERTa** | **Sử dụng mô hình VnCoreNLP** |
| Trang báo điện tử | Text |  | x | x |
| Youtube | Text |  | x | x |
| Audio | x | x | x |
| Facebook | Text |  | x | x |
| Google | Text |  | x | x |

- Tham chiếu đến các thông tin: Các thông tin được nhận diện cảm xúc là “Tiêu cực” thì người dùng có thể truy vấn đến và xem trực tiếp thông tin đó.

**c. Xây dựng hệ thống**

- Hệ thống được chia làm 2 phần: front-end và back-end.

+ ***Front-end***: Sử dụng html, css, NodeJS

+ ***Back-end***: Sử dụng python

- Hệ thống được lập lịch mỗi 6 tiếng sẽ gửi request đến back-end yêu cầu cập nhật thông tin của các đối tượng và các thông tin về sự kiện có trong cơ sở dữ liệu.

- Gồm 4 chức năng chính:

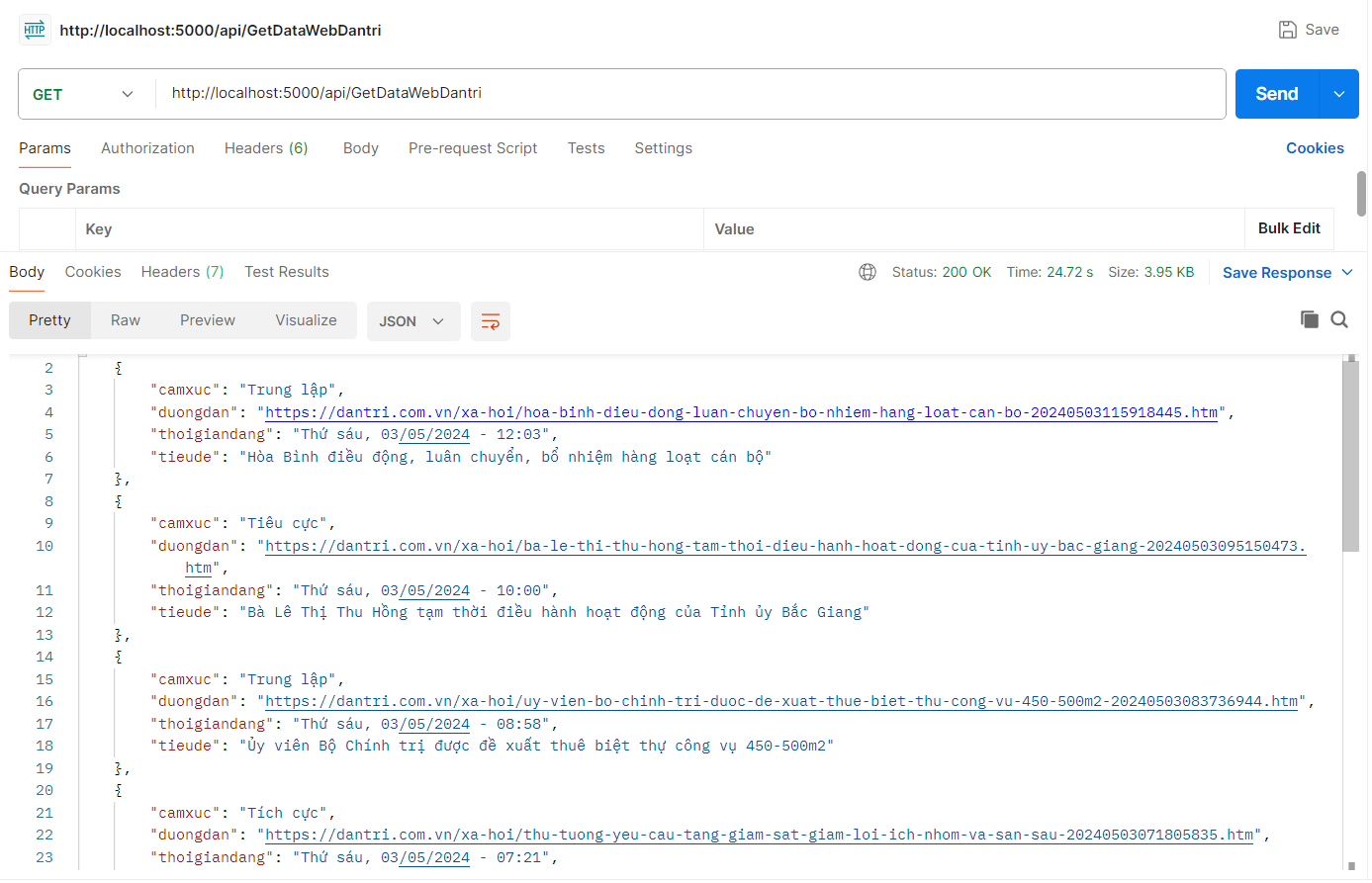
|  |  |  |
| --- | --- | --- |
| **STT** | **Chức năng** | **Mô tả chức năng** |
| 1 | Trinh sát thông tin đối tượng | Lấy và xử lý thông tin các bài báo mới trong web báo điện tử của đối tượng trinh sát đã được lưu trong cơ sở dữ liệu |
| Lấy và xử lý thông tin kênh Youtube, các video được đăng trên kênh Youtube của đối tượng trinh sát đã được lưu trong cơ sở dữ liệu |
| Lấy và xử lý thông tin trang, nhóm Facebook, các bài viết được đăng trên trang, nhóm Facebook của đối tượng trinh sát đã được lưu trong cơ sở dữ liệu |
| 2 | Trinh sát thông tin sự kiện | Lấy thông tin các sự kiện đang là xu hướng tại thời điểm hiện tại |
| Lấy và xử lý thông tin các sự kiện đang được theo dõi |
| 3 | Quản lý đối tượng | Thêm, sửa, xóa đối tượng. Một đối tượng bao gồm: Tên đối tượng, liên kết web, liên kết Youtube, liên kết Facebook |
| 4 | Quản lý sự kiện | Thêm, sửa, xóa sự kiện. Một sự kiện bao gồm tên sự kiện, số tin tích cực, số tin tiêu cực |

# 4. Báo cáo

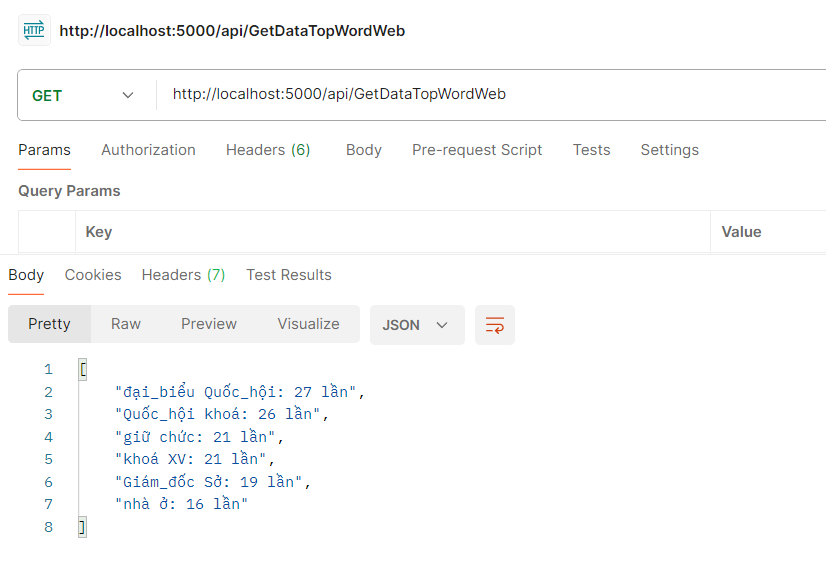
Kết quả thu nhận được của công cụ trinh sát đã xây dựng:

**a. Dữ liệu thu thập được**

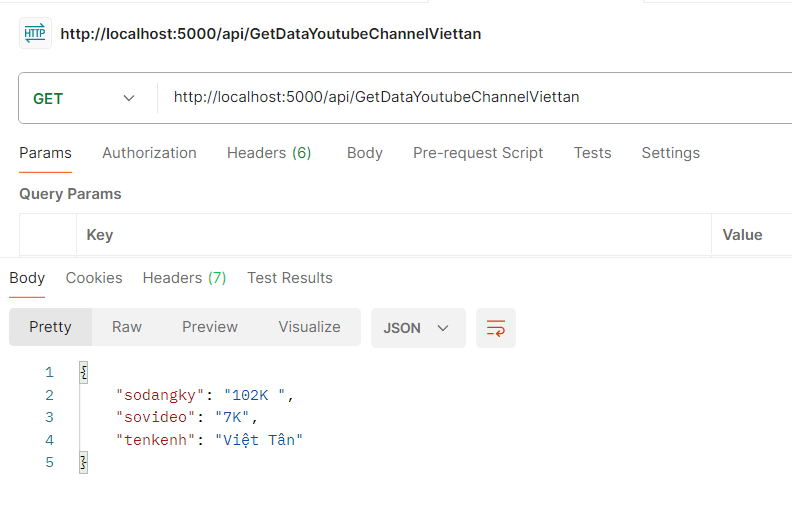
- Trinh sát thông tin đối tượng



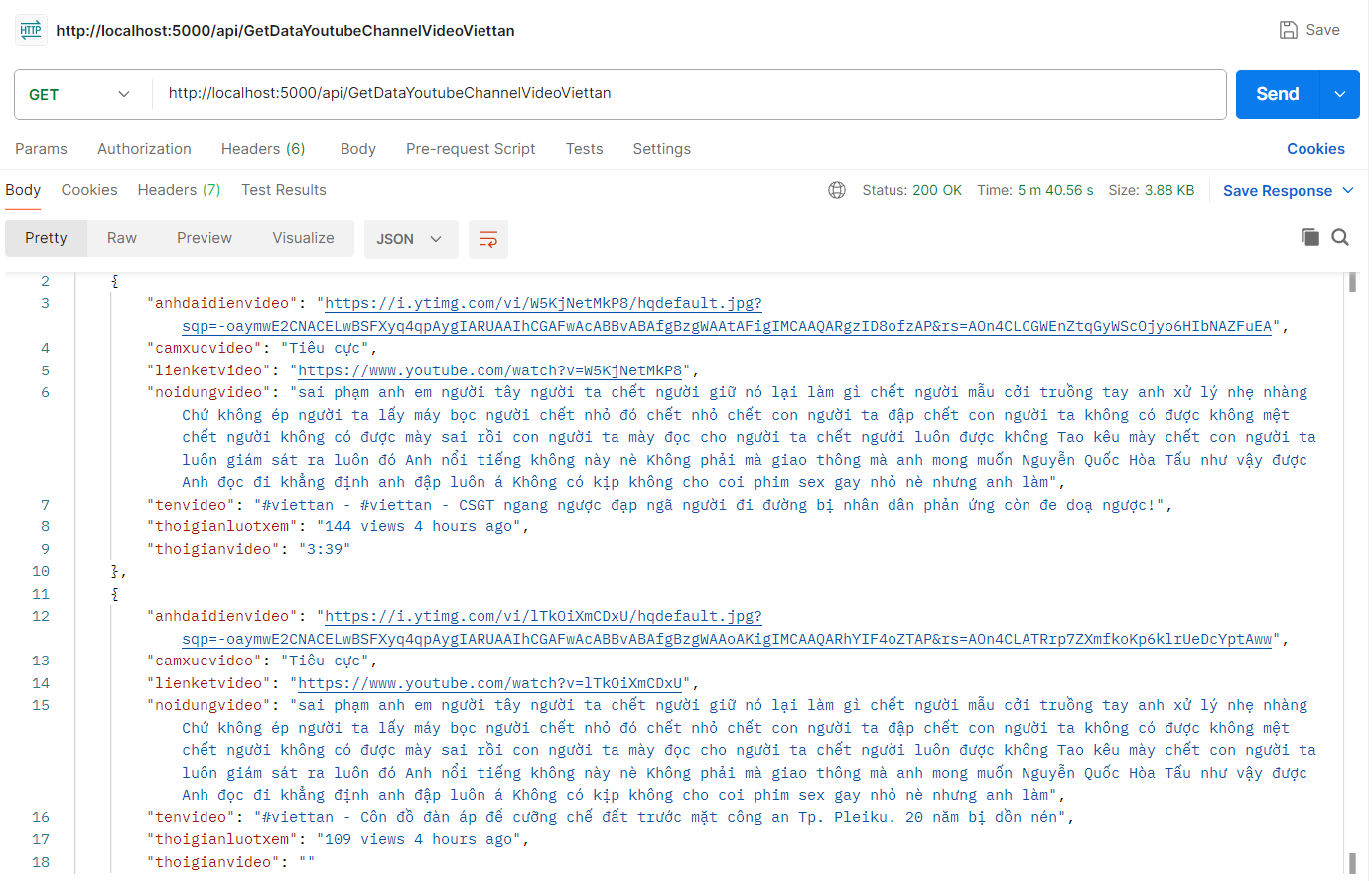
Hình . Thông tin từ trang web



Hình . Thống kê các từ khóa trong trang web



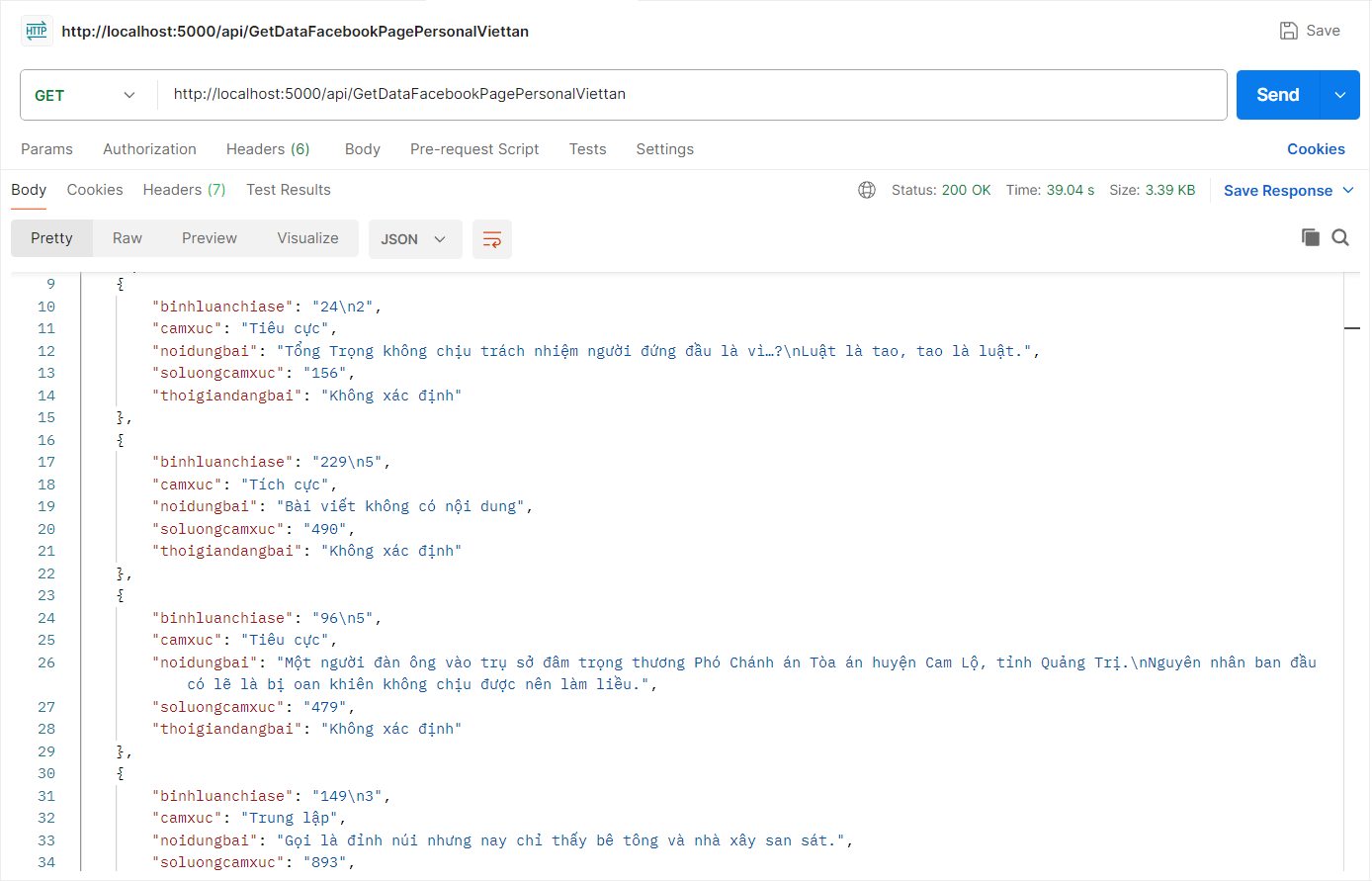
Hình . Thông tin kênh YouTube của đối tượng



Hình . Thông tin từ video của kênh YouTube

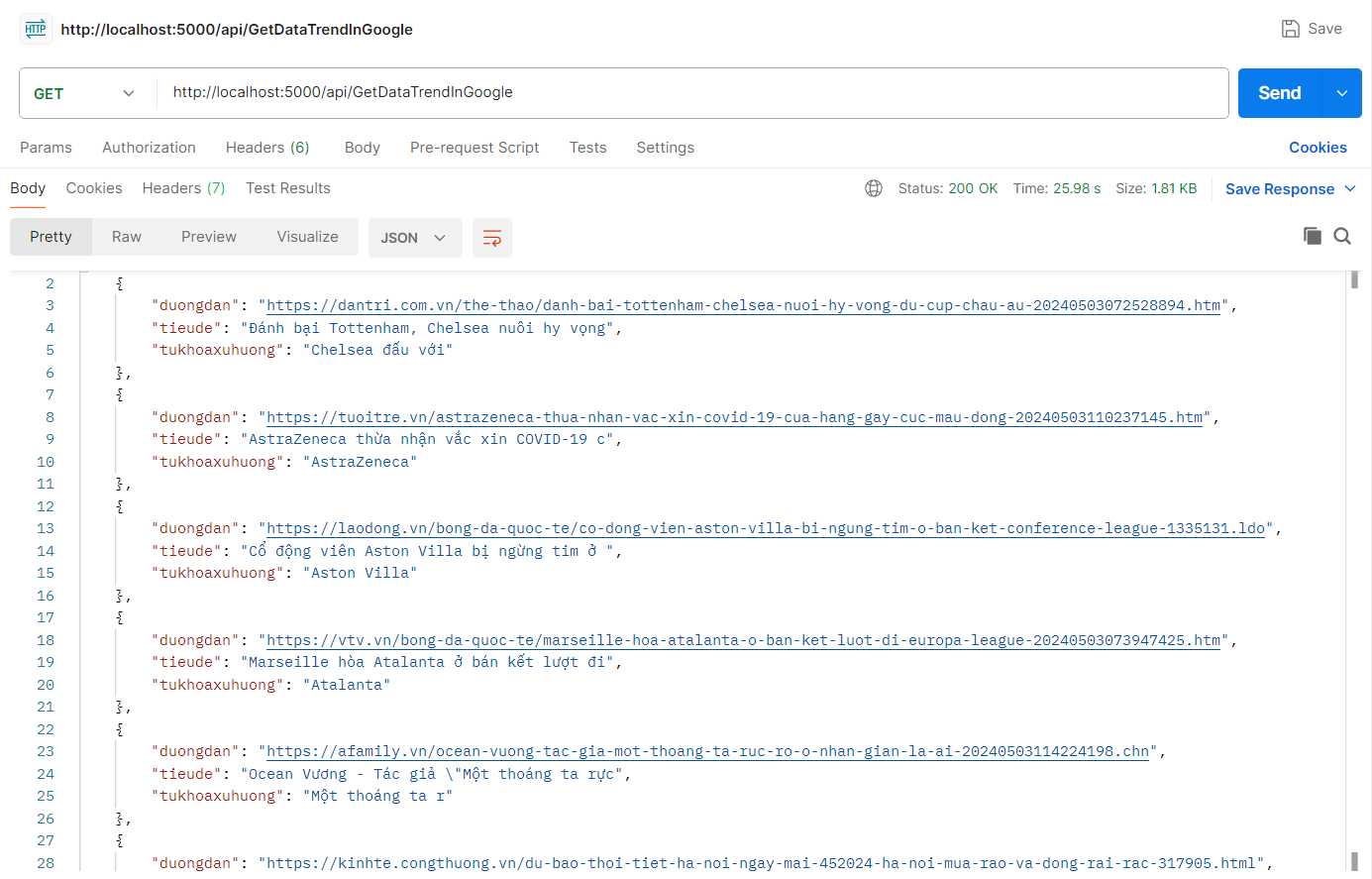


Hình . Thông tin trang Facebook đối tượng

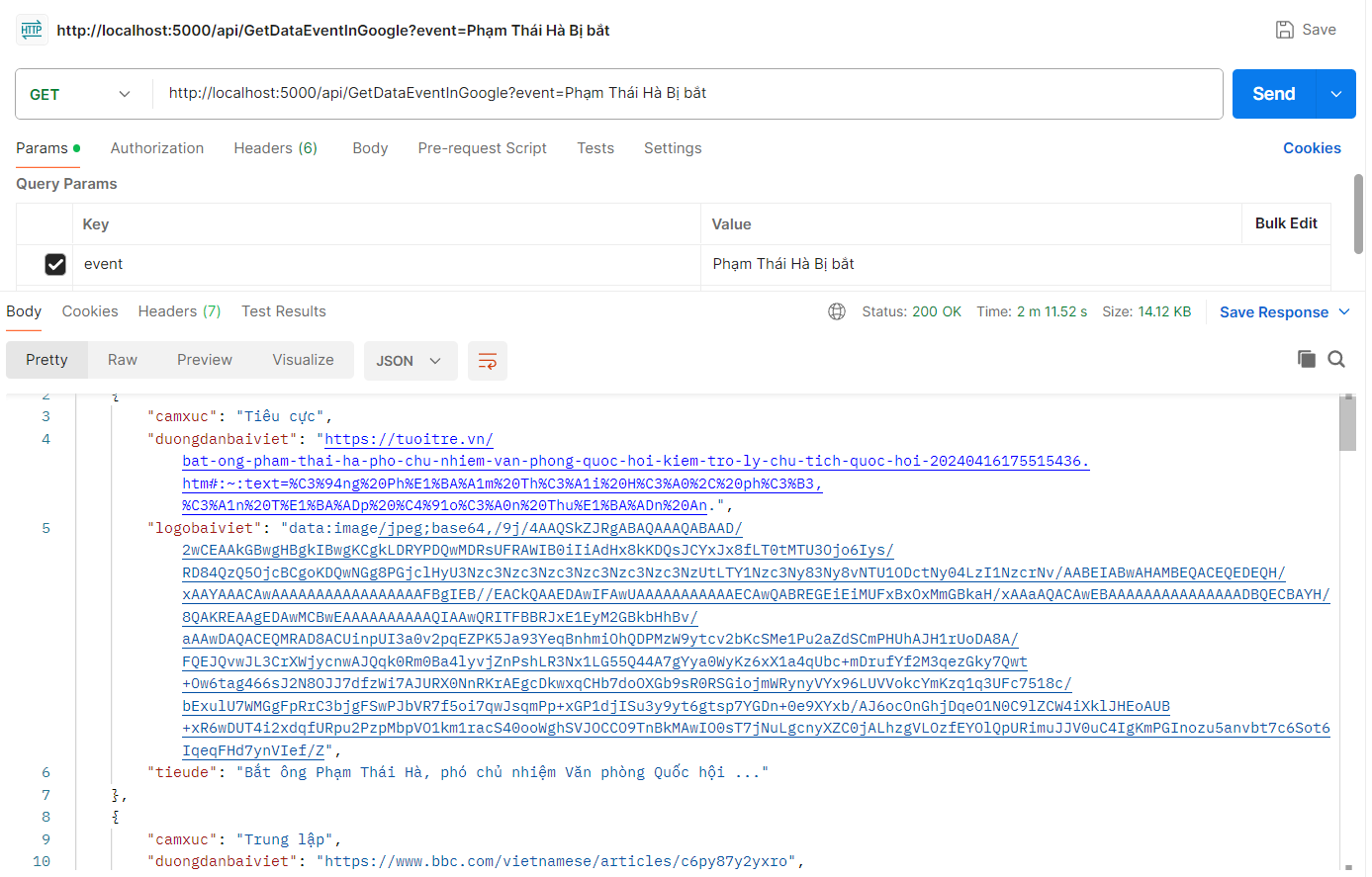


Hình . Thông tin từ bài đăng của trang Facebook đối tượng trinh sát

- Trinh sát thông tin sự kiện



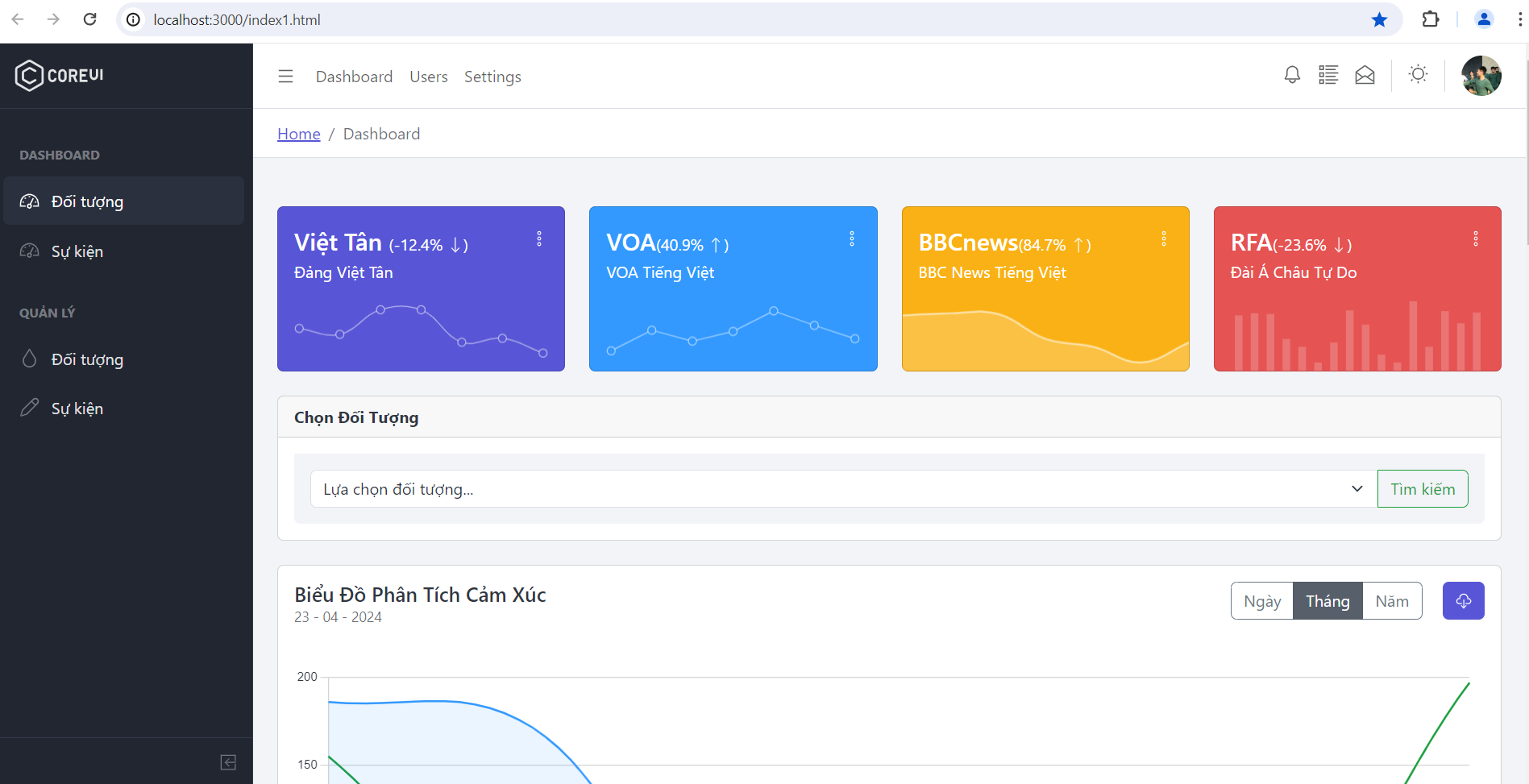
Hình . Thông tin nổi bật trên Google



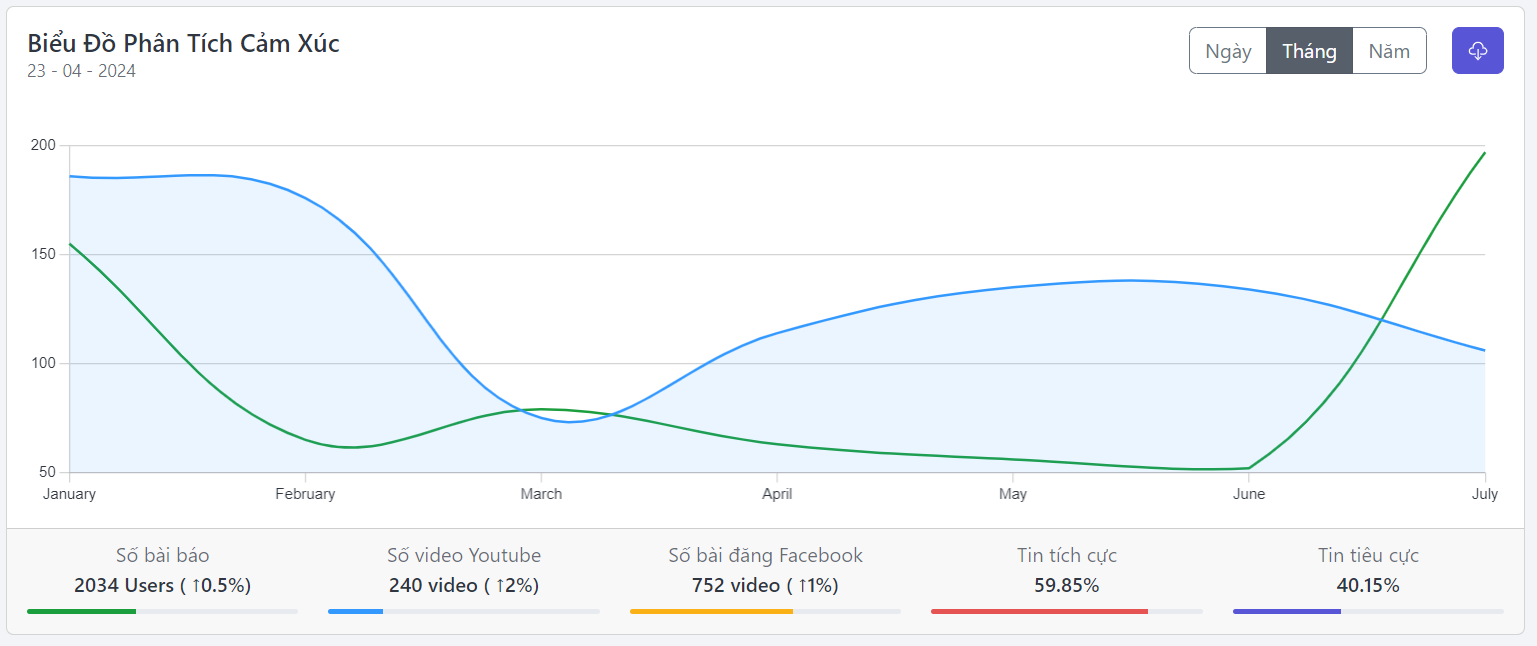
Hình . Thông tin liên quan đến sự kiện trinh sát (theo kết quả tìm kiếm Google)

**b. Giao diện chức năng**

- Trinh sát đối tượng

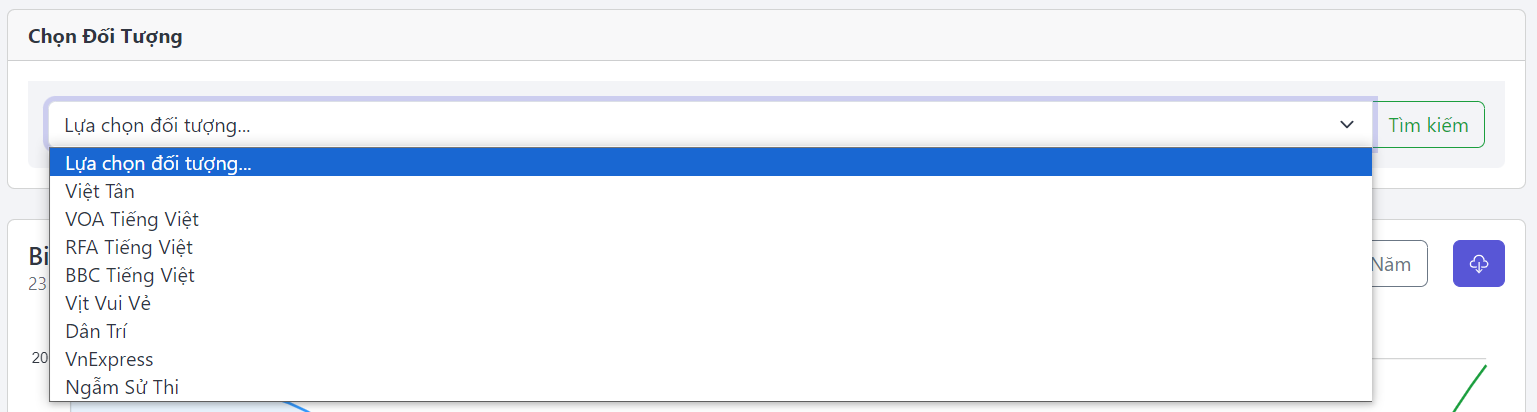


Hình . Dashboard trinh sát đối tượng

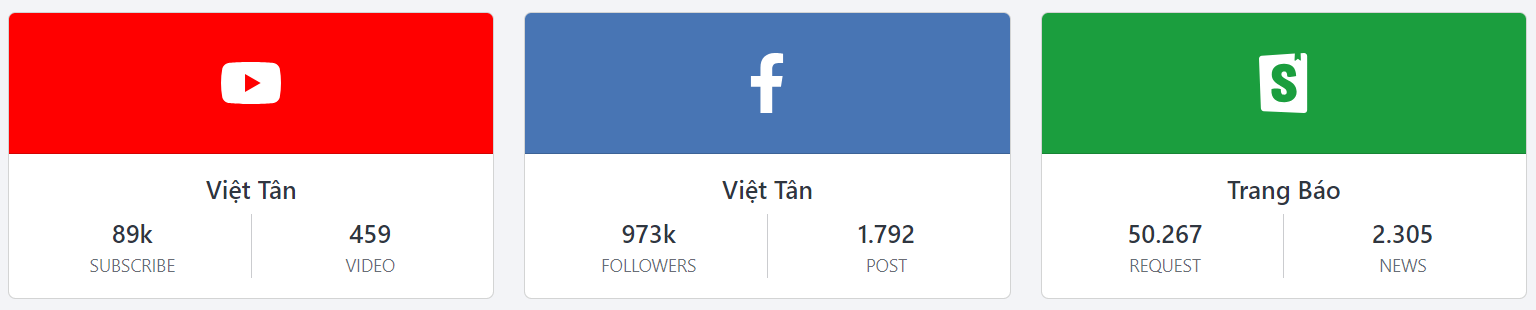


Hình . Biểu đồ phân tích cảm xúc

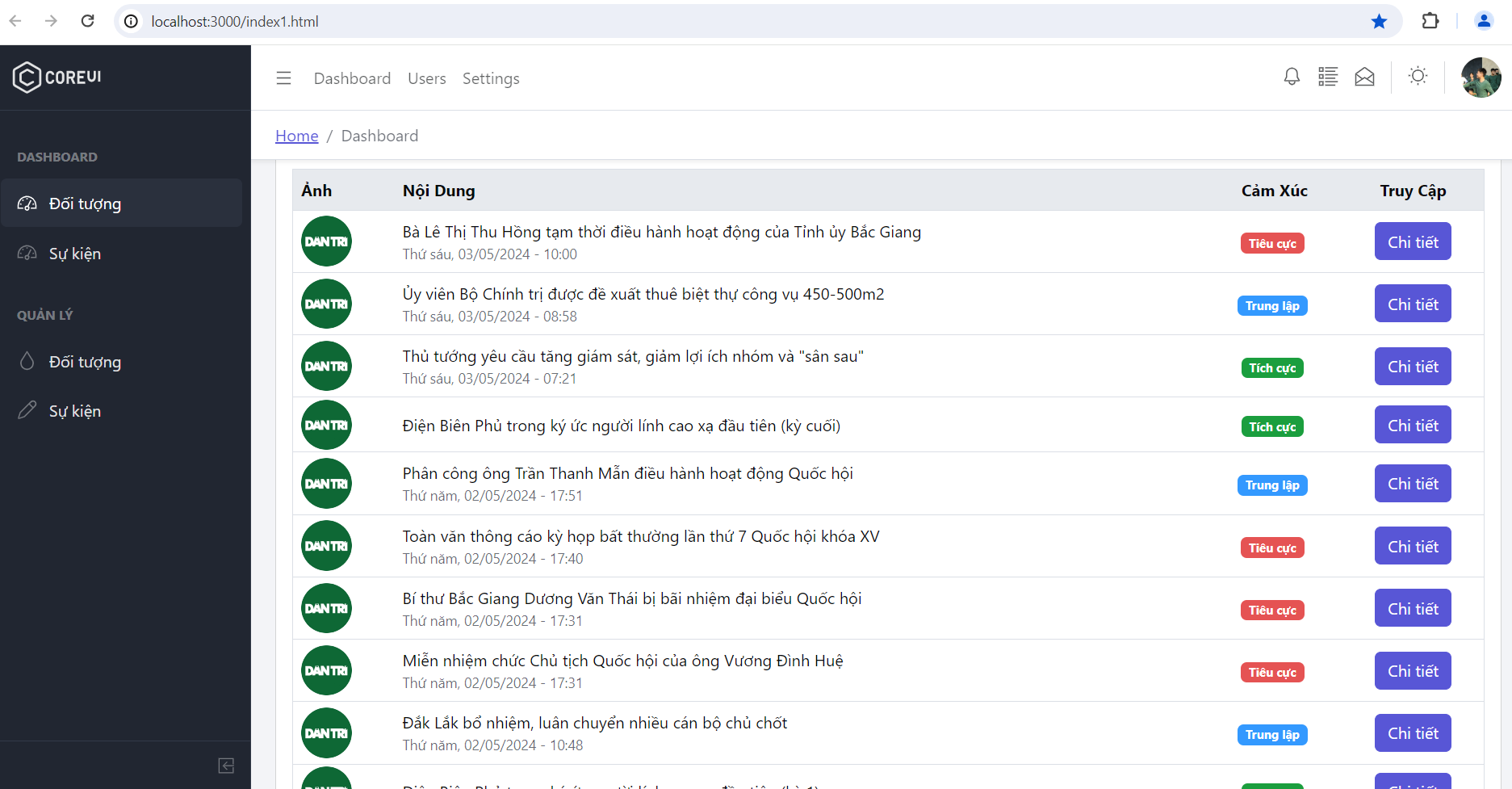
+ Thông tin theo từng đối tượng trinh sát



Hình . Lựa chọn đối tượng



Hình . Thống kê thông tin từ các nền tảng



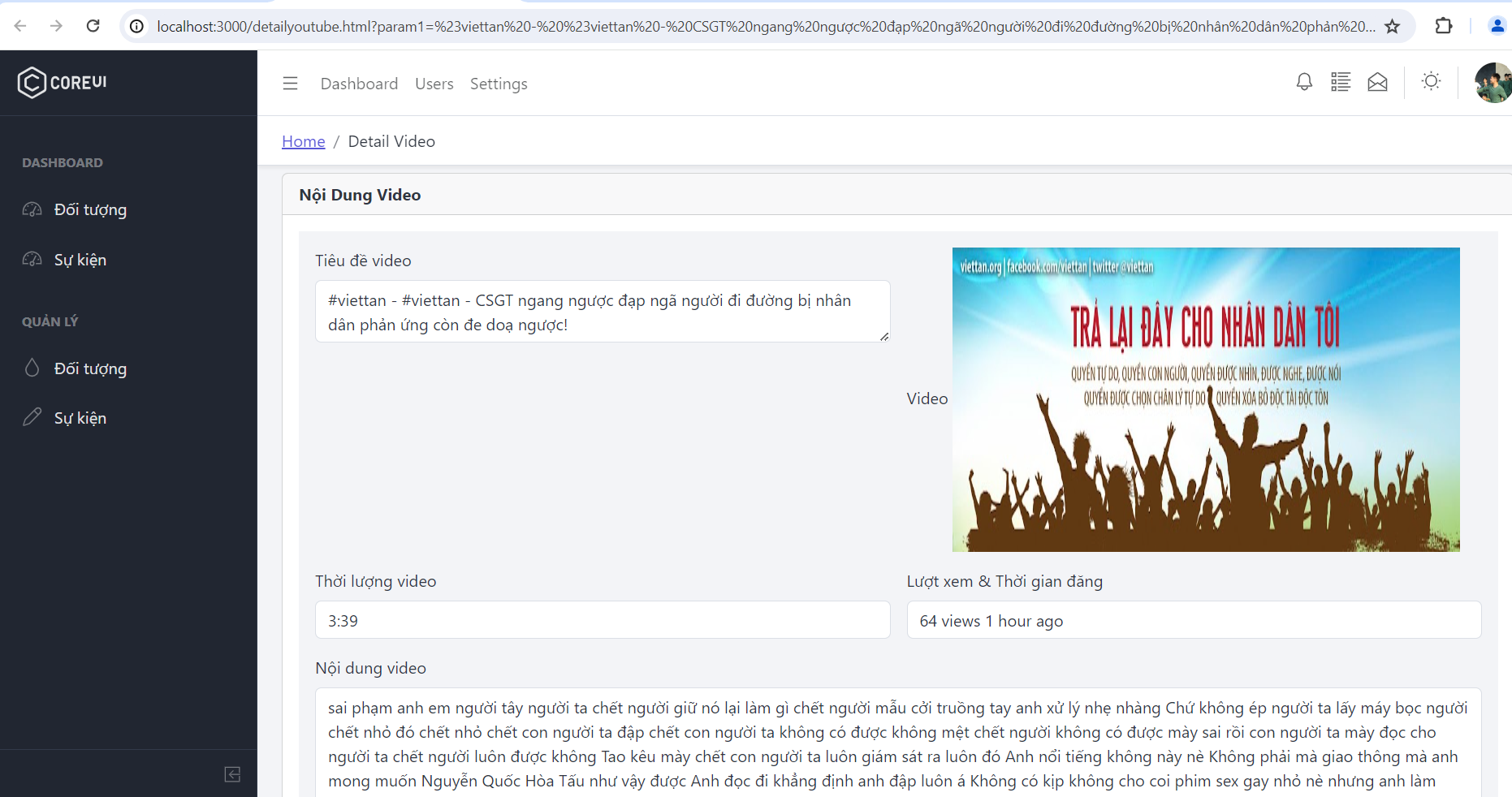
Hình . Bài báo mới của đối tượng



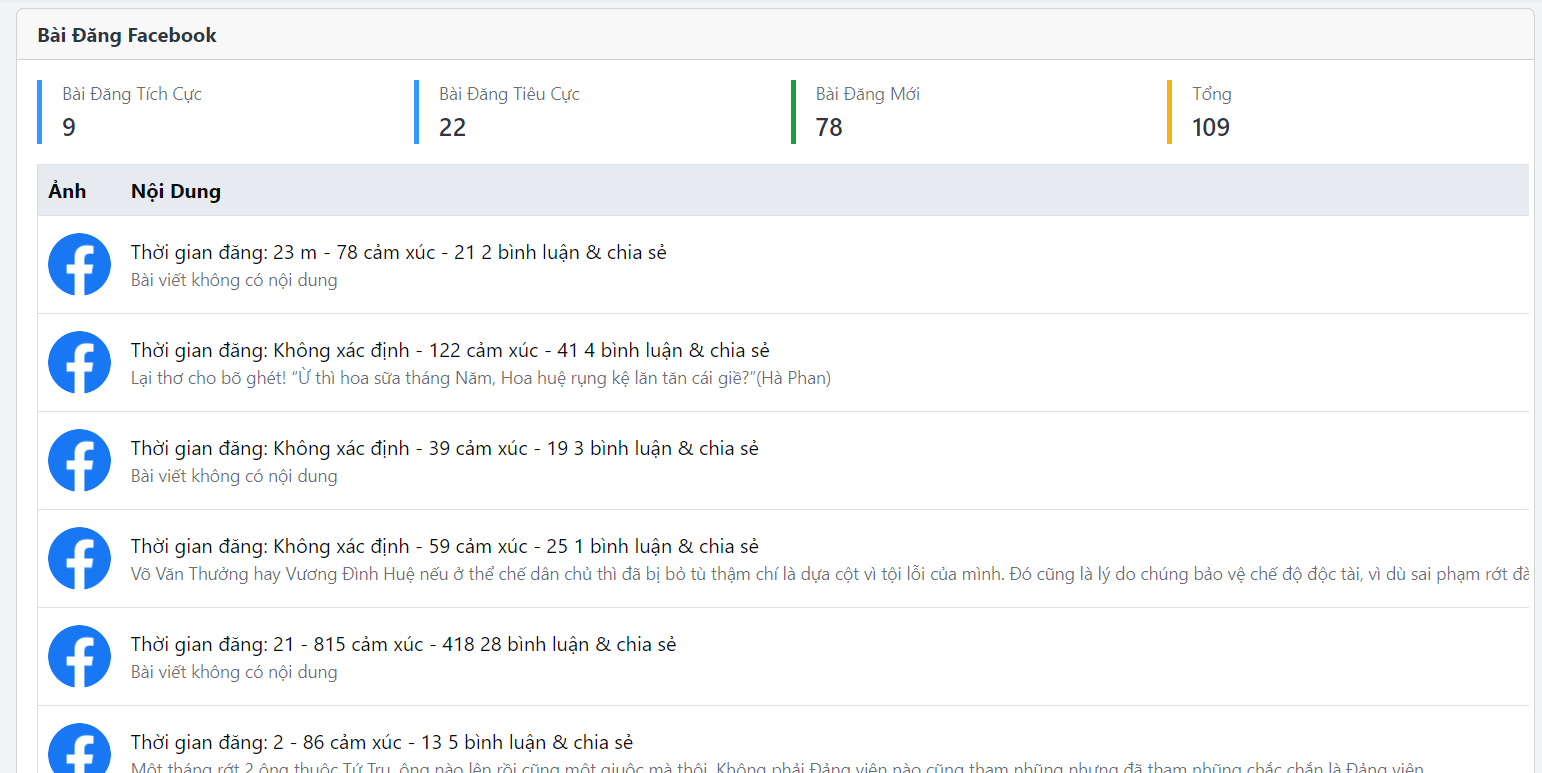
Hình . Top từ khóa trong các bài báo



Hình . Video mới của đối tượng

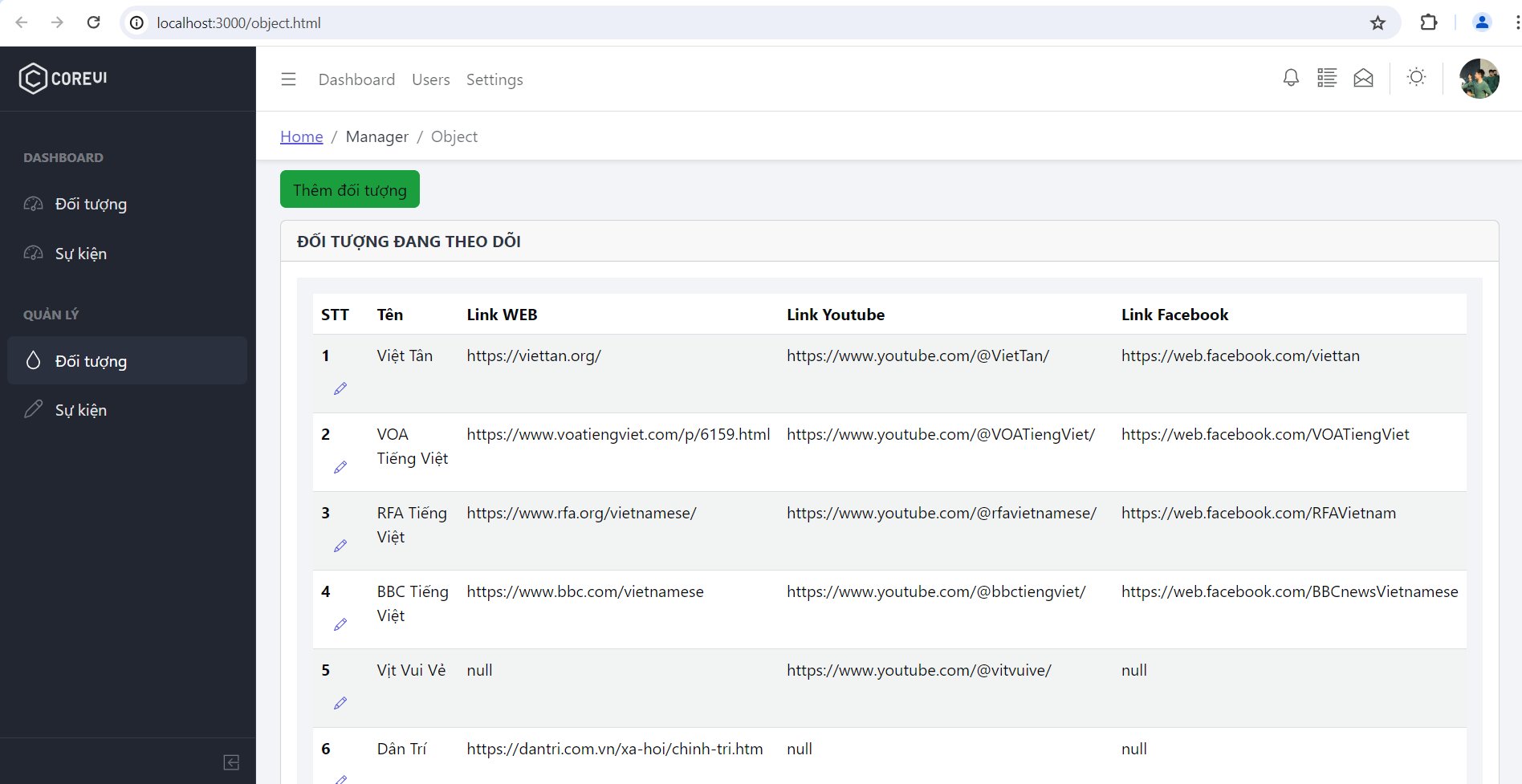


Hình . Chi tiết video từ Youtube



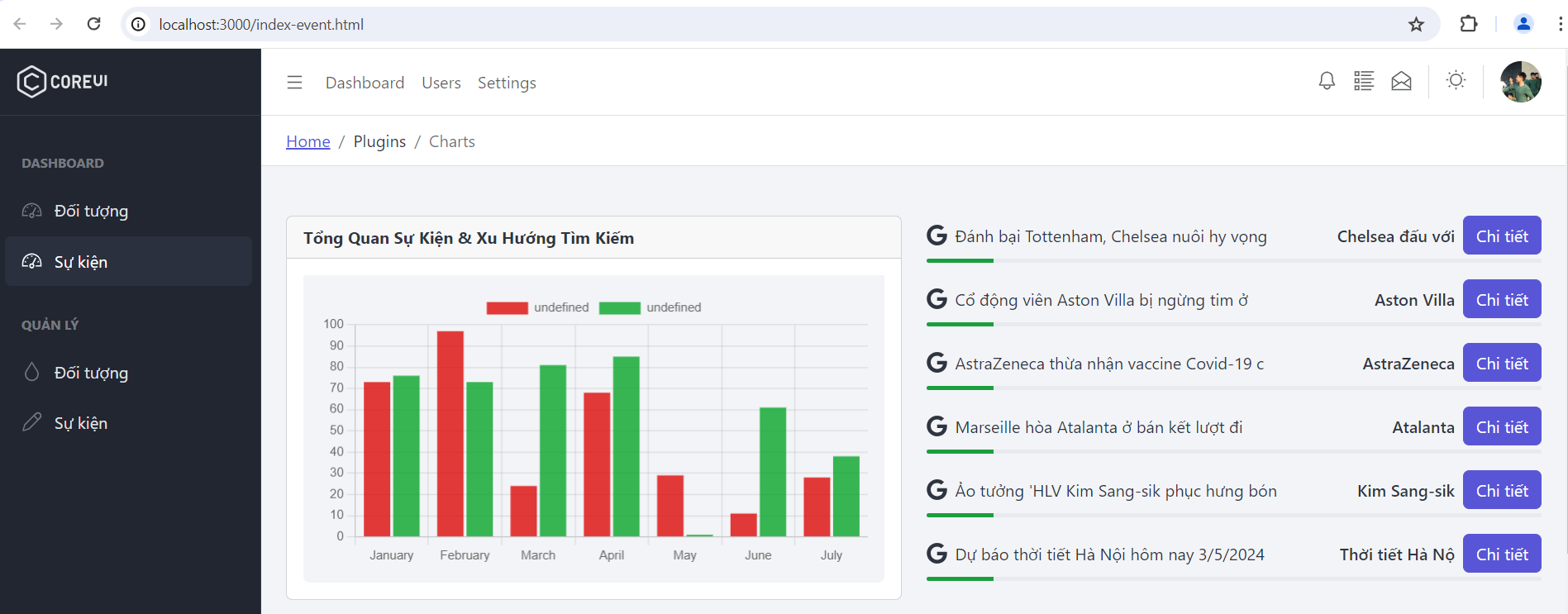
Hình . Các bài đăng mới trên trang Facebook của đối tượng

+ Quản lý đối tượng trinh sát



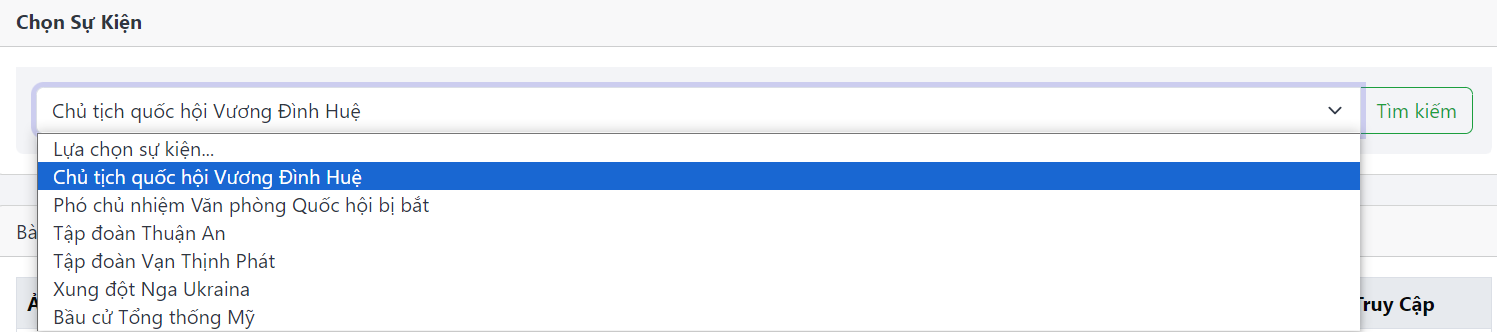
Hình . Giao diện quản lý đối tượng

- Trinh sát sự kiện

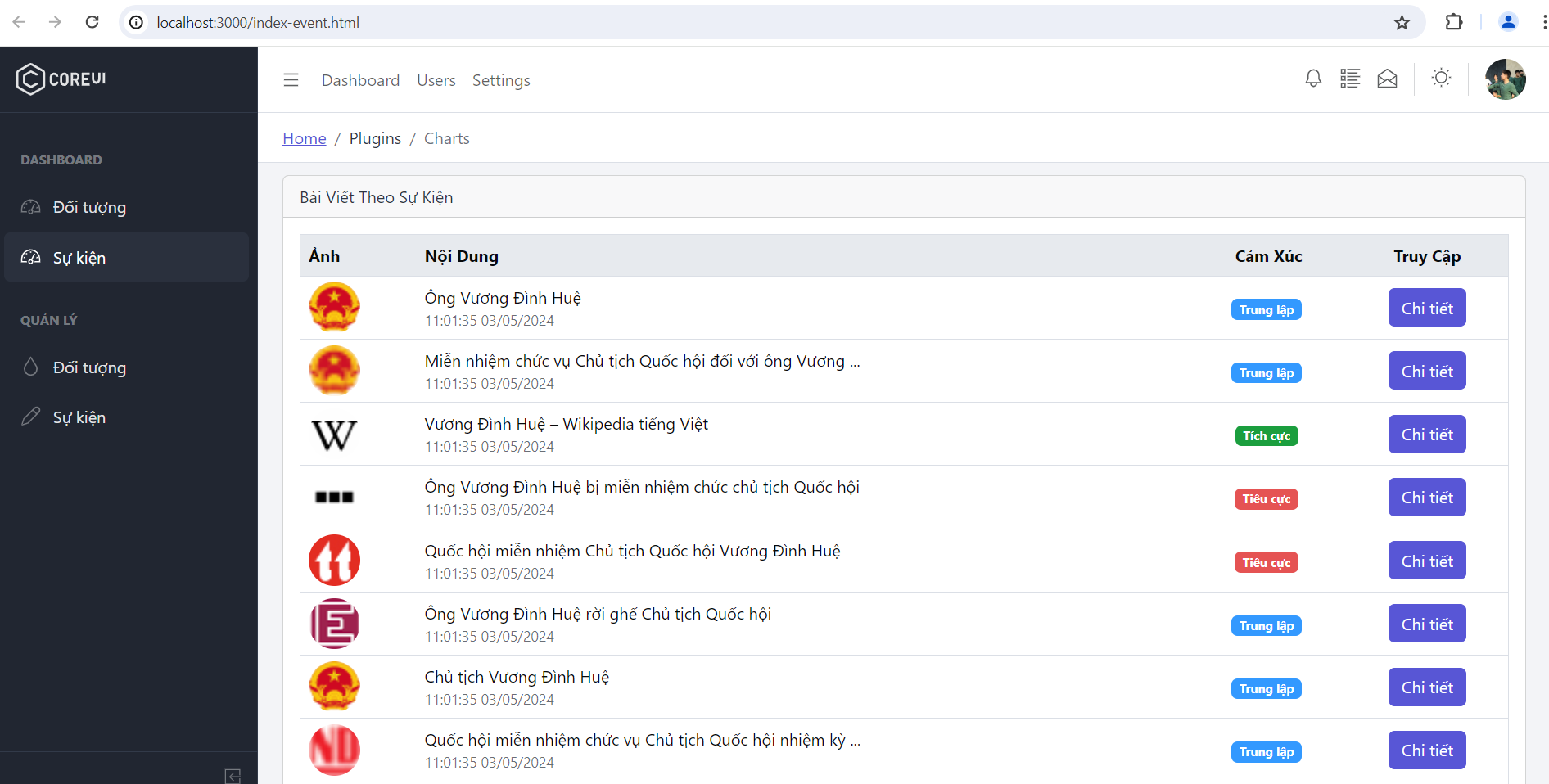


Hình . Dashboard trinh sát sự kiện

+ Thông tin của sự kiện trinh sát

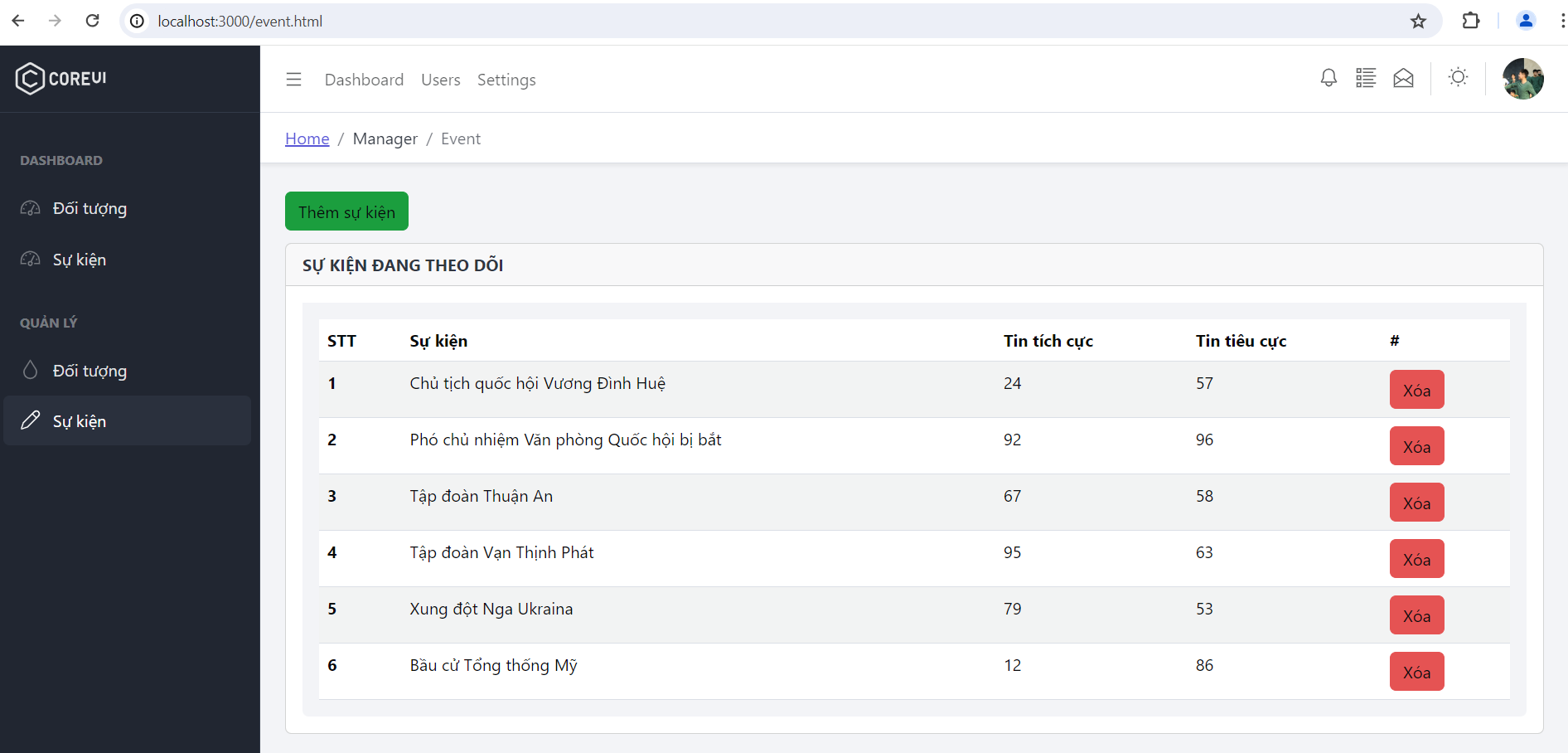


Hình . Lựa chọn sự kiện



Hình . Thông tin liên quan đến sự kiện

+ Quản lý sự kiện trinh sát



Hình . Giao diện quản lý sự kiện