# Đề tài: Xây dựng công cụ tổng hợp tin tức đa phương tiện

Học viên thực hiện: Hà Công Quốc Bảo Lớp: ANHTTT

Nông Phương Trang Lớp ANHTTT

# 1. Mô tả bài toán

Công cụ này có nhiệm vụ: thường xuyên cập nhật, phân loại thông tin trên google, youtube, facebook.

Hệ thống gồm 3 chức năng chính:

- Thu thập dữ liệu văn bản:

+ Google, Facebook: Lấy thông tin dạng văn bản.

+ Youtube: Download audio, chuyển đổi âm thanh thành văn bản.

- Phân tích văn bản để tổng hợp thông tin theo đối tượng:

+ Thu thập thông tin của đối tượng bằng các mục tiêu nền tảng: google, youtube, facebook đã có trong cơ sở dữ liệu

+ Phân cụm văn bản theo đối tượng: Tìm ra chủ đề trong mỗi bài viết của đối tượng

+ Phân tích thông tin của đối tượng: Tóm tắt thông tin thu thập được, phân tích sắc thái.

- Phân tích văn bản để tổng hợp thông tin theo sự kiện:

+ Phân cụm văn bản theo sự kiện: Các bản tin liên quan đến 1 sự kiện được gán vào cùng một nhóm.

+ Phân tích thông tin trong sự kiện: Tóm tắt bản tin, phân tích sắc thái.

# 2. Các thành phần của hệ thống

- Nguồn tin: Google, Youtube, Facebook

- Thư viện sử dụng: BeautifulSoup, selenium, pytube

- Lưu trữ: MySQL

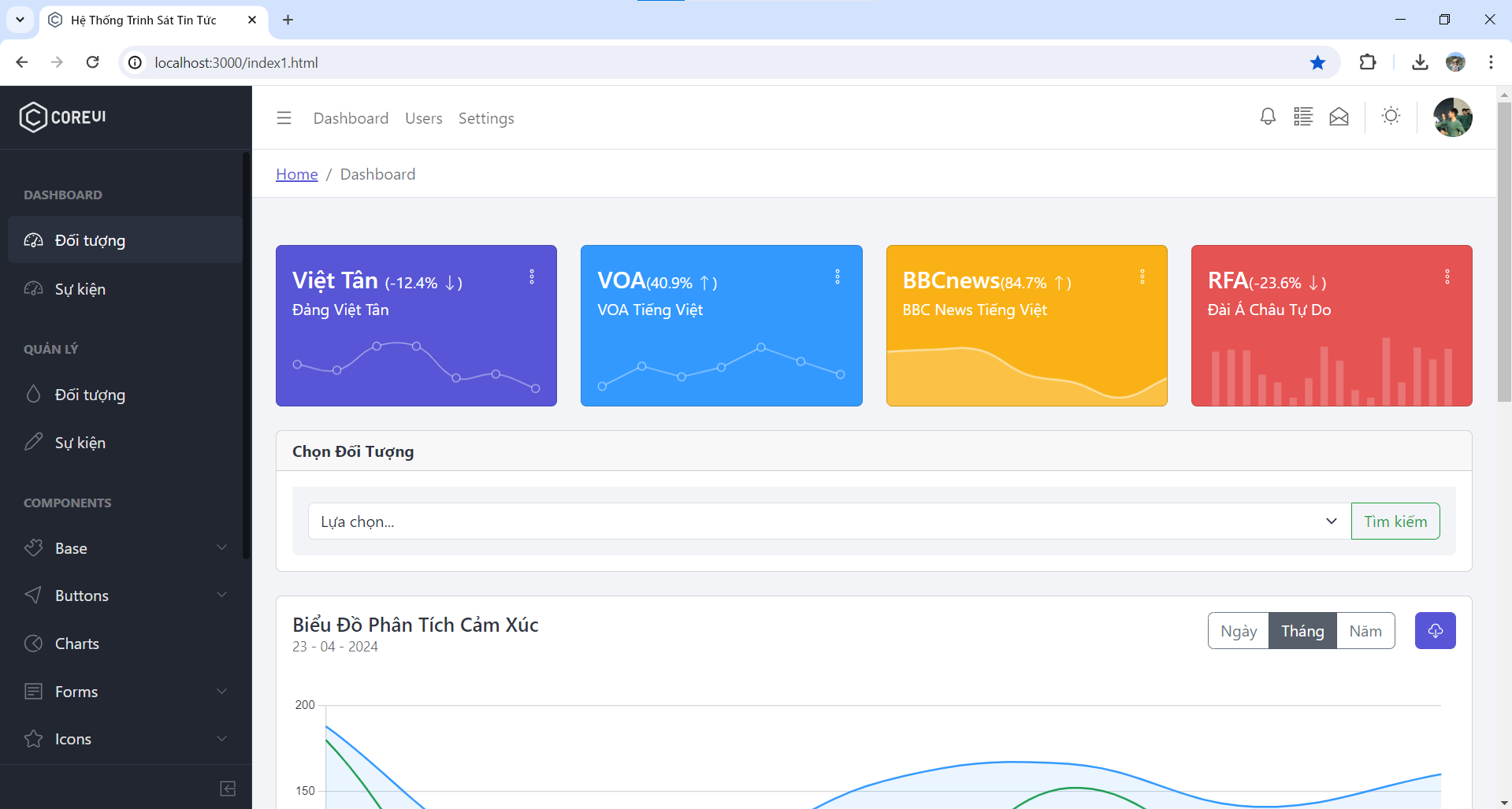
- Xử lý:

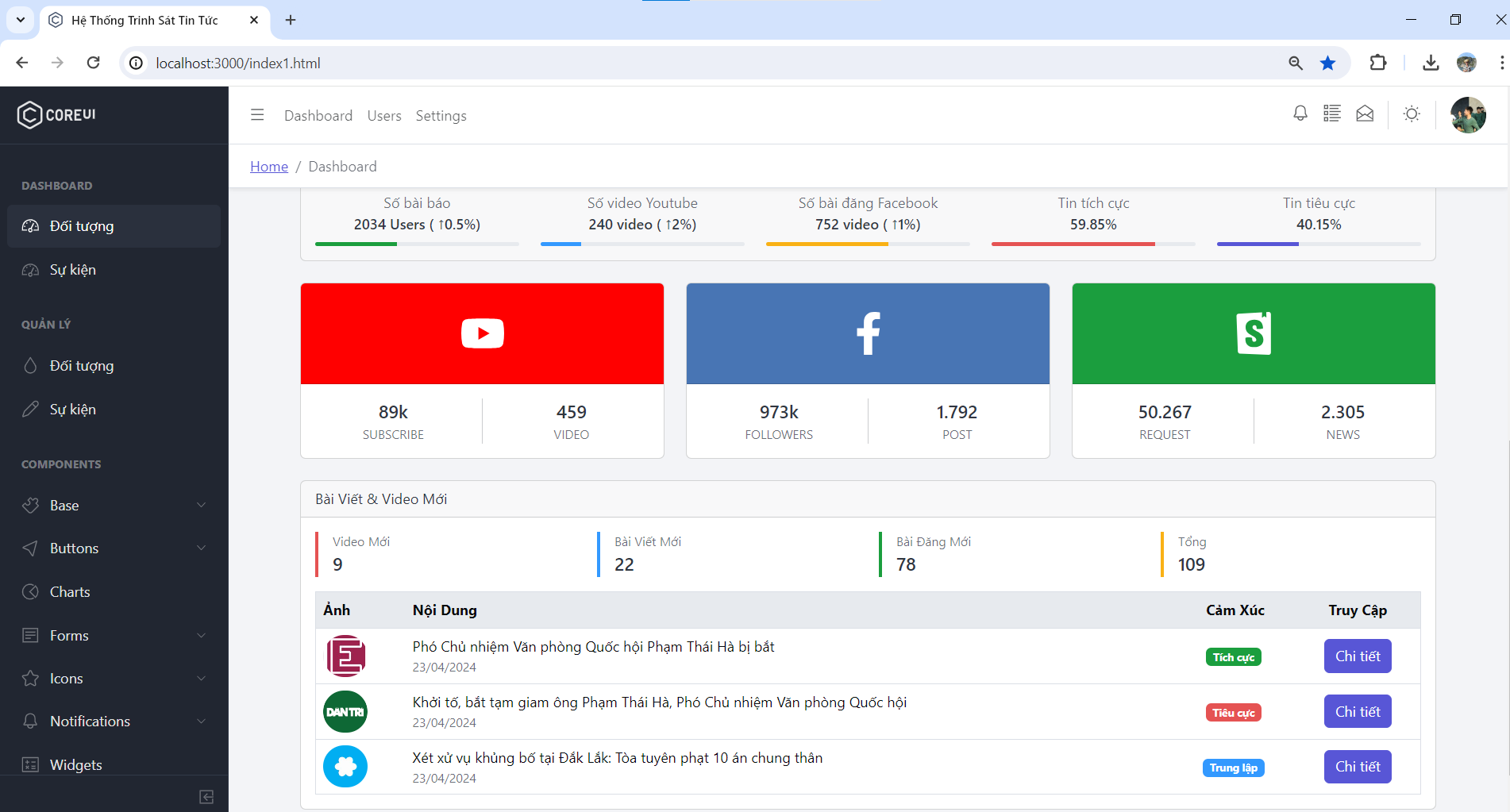
+ Nhận dạng âm thanh: Thư viện speech recognition

+ Phân tích cảm xúc: Mô hình RoBERTa

+ Nhận dạng chủ đề: Mô hình VnCoreNLP

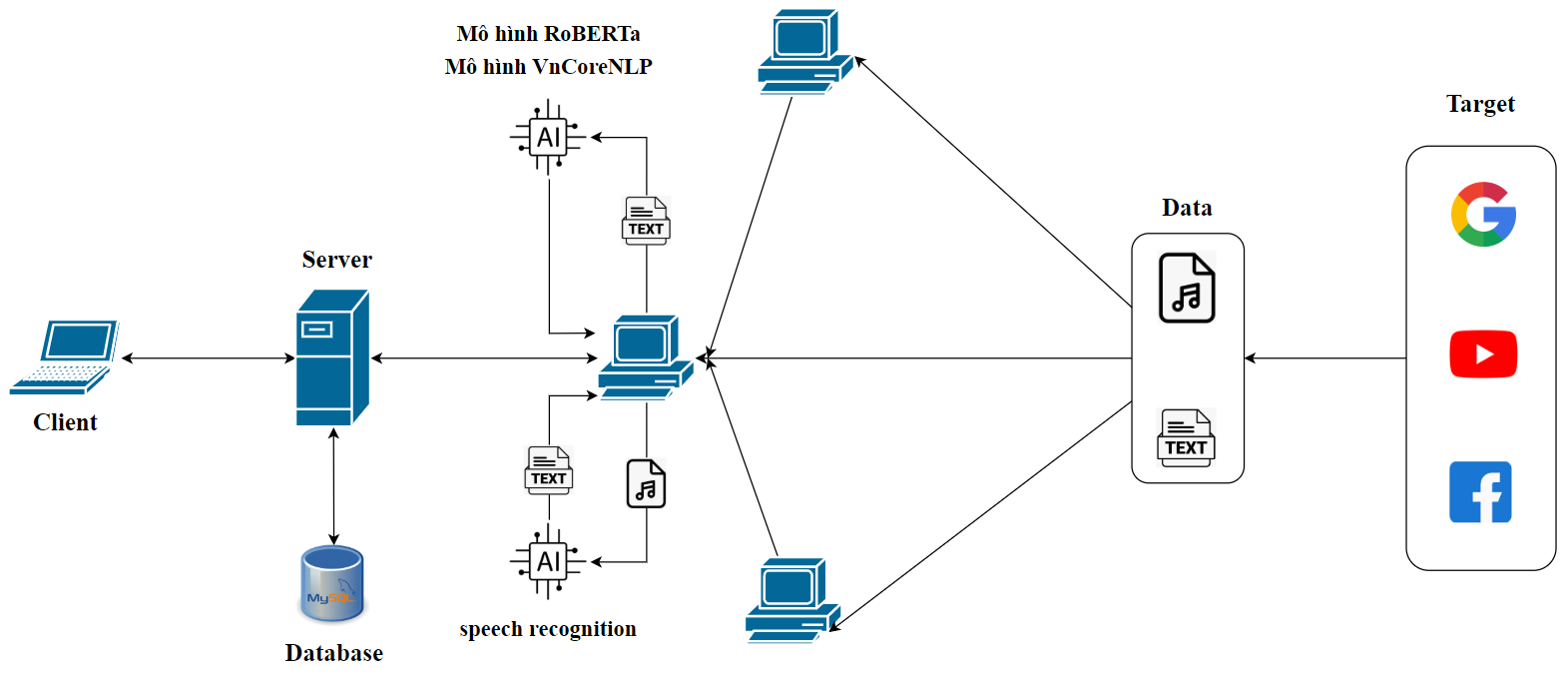
- Dashboard:





# 3. Xây dựng công cụ

**-** Mô hình triển khai hệ thống:



- Giải pháp thu thập dữ liệu: Thu thập đủ tất cả các dữ liệu từ các nguồn Google, Youtube, Facebook.

- Giải pháp xử lý thông tin để đạt được mục tiêu mô tả bài toán:

+ Thu thập đủ thông tin từ các nguồn dữ liệu của đối tượng.

+ Sử dụng thư viện BeautifulSoup, Selenium: Lấy thông tin dạng văn bản.

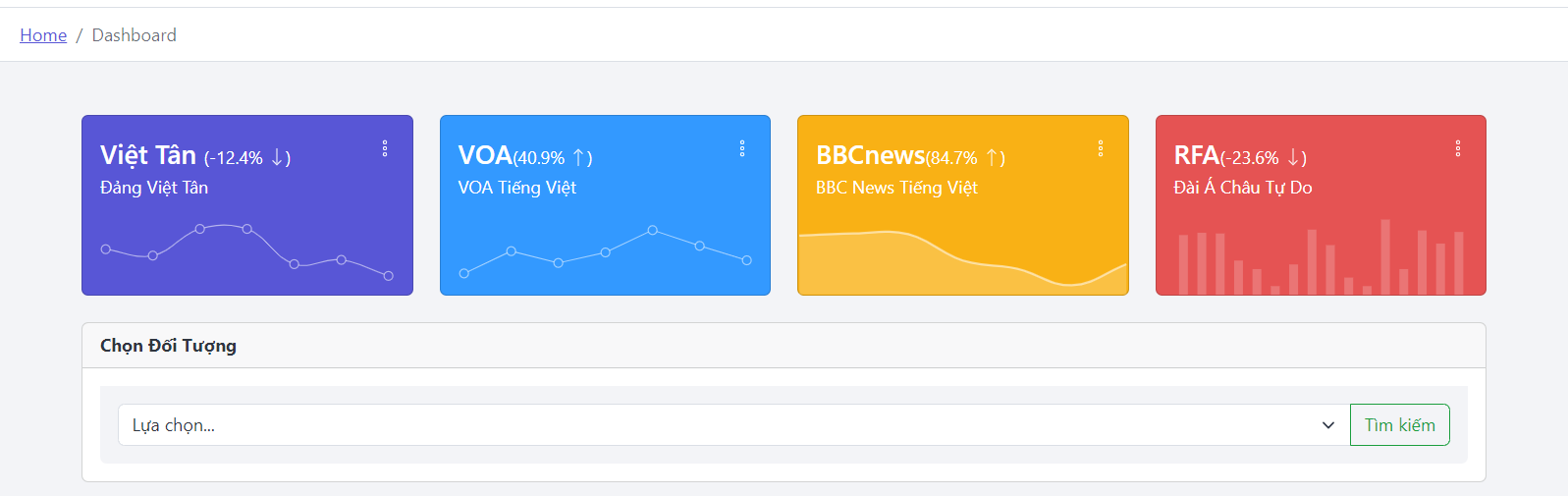
+ Sử dụng thư viện Pytube, Speech Recognition: Tải xuống thông tin dạng audio chuyển đổi thành văn bản.

+ Sử dụng mô hình RoBERTa xử lý bài toán phân tích cảm xúc.

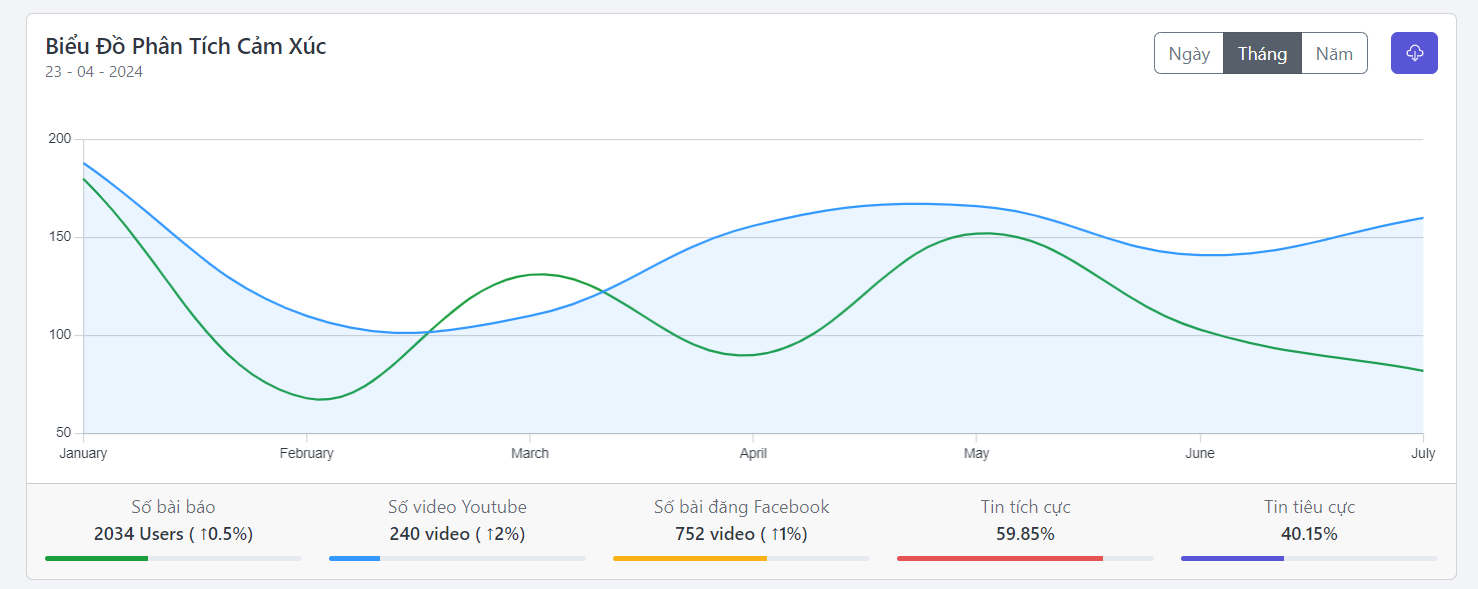
+ Sử dụng mô hình VnCoreNLP xử lý bài toán nhận dạng thực thể.

- Xây dựng giao diện, thống kê, trực quan hóa kết quả:

* *Chức năng trinh sát thông tin theo đối tượng*



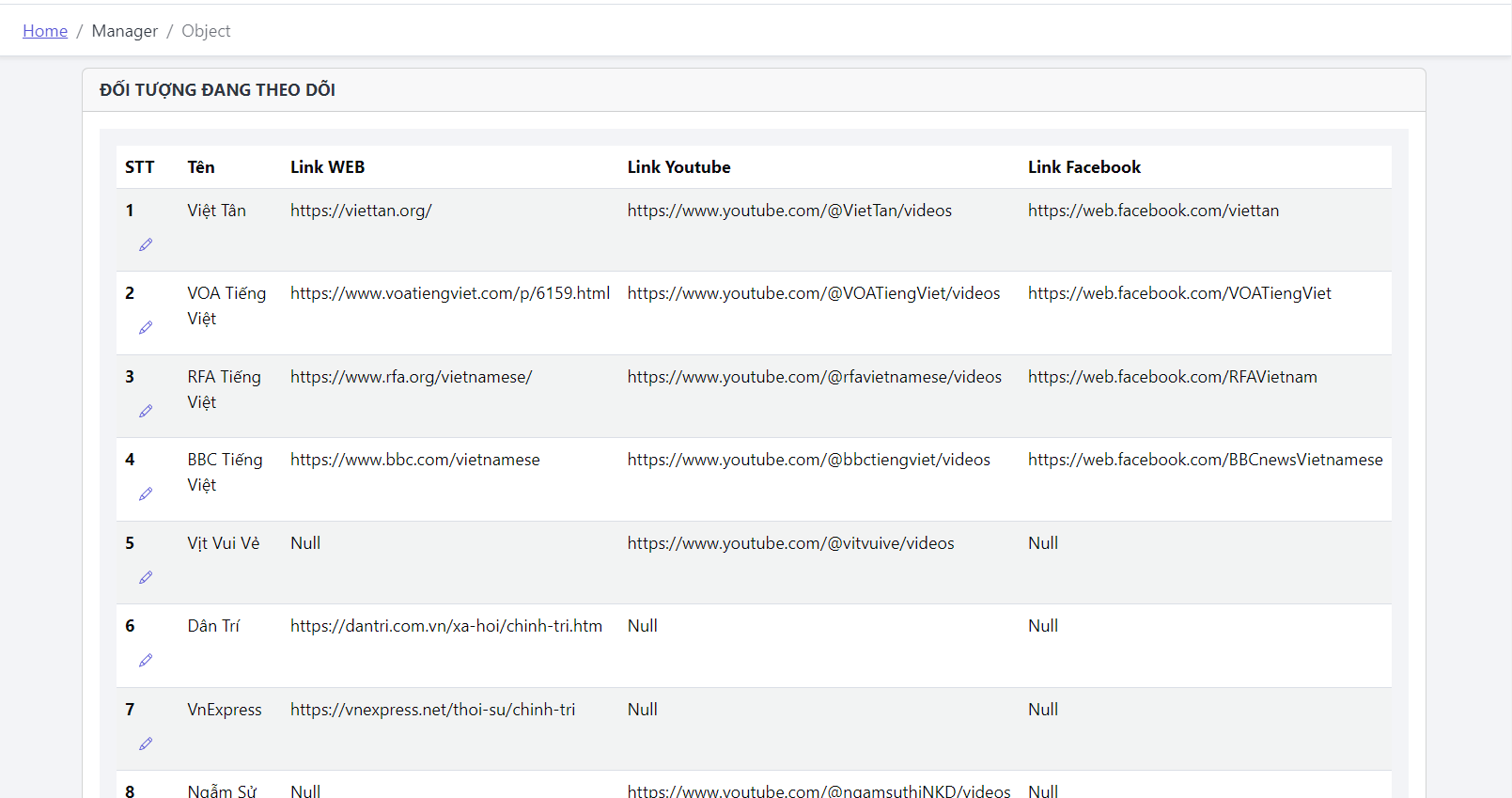
*Giao diện: Dashboard theo dõi danh sách đối tượng*



*Giao diện: Biểu đồ phân tích cảm xúc của 1 đối tượng*



*Giao diện: Tổng quan và các bài đăng mới của đối tượng*

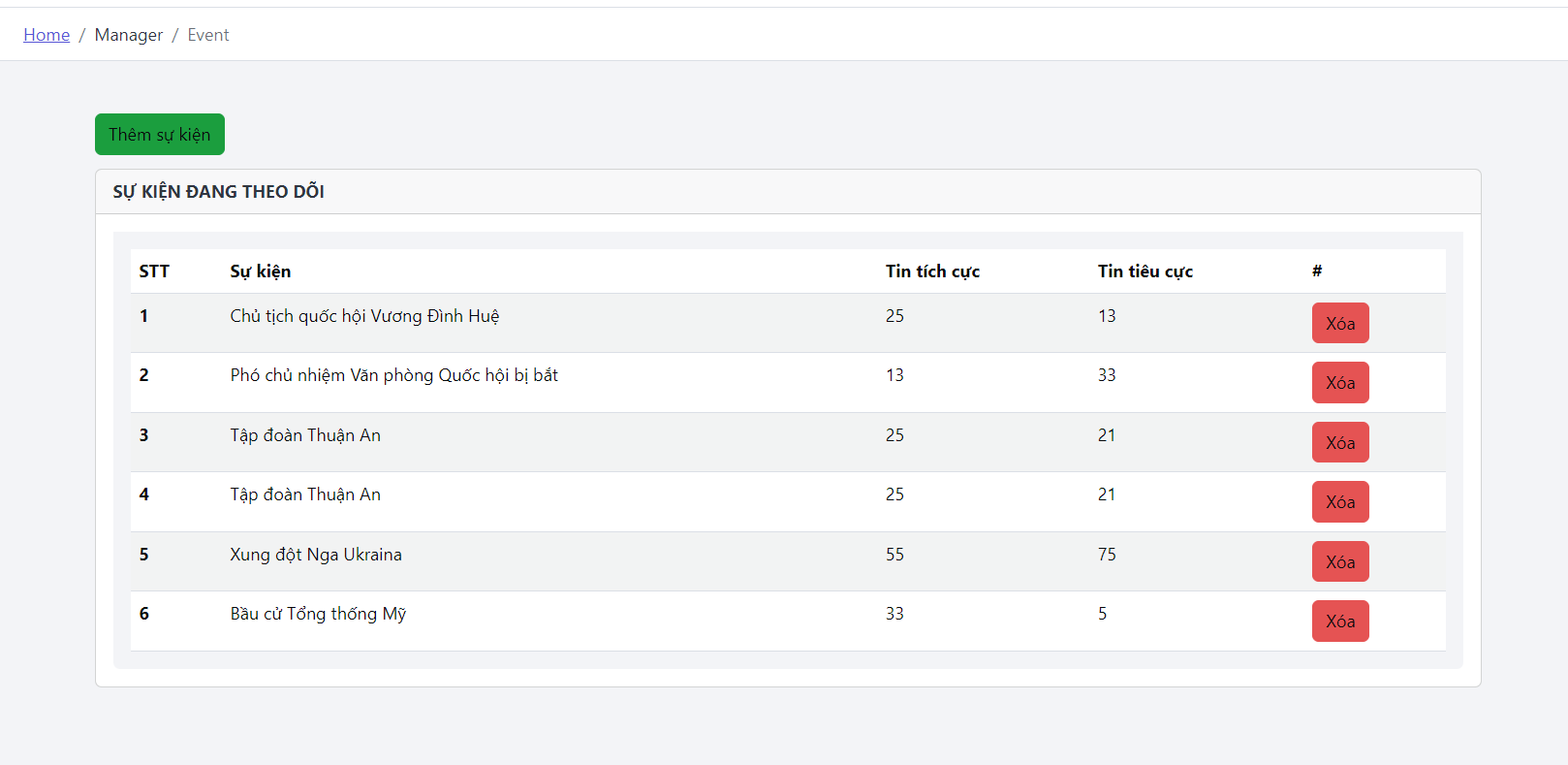


*Giao diện: Quản lý đối tượng*

* *Chức năng trinh sát thông tin theo sự kiện*



*Giao diện: Tổng quan về Sự kiện & Xu hướng tìm kiếm*



*Giao diện: Quản lý sự kiện đang theo dõi*