Phát hiện Xu hướng và Phân tích Tâm lý theo Thời gian thực

Báo cáo tiến độ dự án - SE363.Q11

Thực hiện:
Tăng Nhất[1] Lê Minh Nhựt[1]
GVHD: TS. Đỗ Trọng Hợp[1]

[1]Khoa Khoa học Máy tính
Trường Đại học Công nghệ Thông tin

Tháng 12 năm 2025

Nội dung báo cáo

Mục lục

Mục tiêu đề tài

Xây dựng hệ thống **Real-time Event Detection** trên mạng xã hội nhằm thu hẹp khoảng cách giữa dữ liệu thô và thông tin chi tiết có thể hành động.

Hai nhóm đối tượng chính

- **Chính phủ/An toàn công cộng:** Phát hiện sớm các rủi ro xã hội, biểu tình, tin giả (Social Risk).
- **Doanh nghiệp/Marketing:** Nắm bắt nhanh các xu hướng tiêu dùng, viral trends (Market Opportunity).

Phương pháp tiếp cận: Kết hợp Học bán giám sát (Semi-Supervised Learning) để xử lý dữ liệu stream thiếu nhãn.

Phân loại sự kiện (Taxonomy)

Hệ thống tập trung vào 7 loại sự kiện chính:

1. **Social Controversy:** Tranh cãi xã hội, bê bối (Ưu tiên cao cho CP).
2. **Civil Unrest:** Biểu tình, đình công.
3. **Natural Disaster:** Thiên tai, lũ lụt, môi trường.
4. **Public Safety:** Tai nạn, cháy nổ.
5. **Politics & Policy:** Bầu cử, phản ứng chính sách.
6. **Viral Lifestyle:** Xu hướng ăn uống, thời trang (Ưu tiên cho Marketing).
7. **Entertainment:** Giải trí, văn hóa đại chúng.

Mục lục

Nguồn dữ liệu 1: Mạng xã hội (Social Media)

- **Nguồn:** Các Fanpage lớn (Ví dụ: Theanh28, Beatvn...)
- **Định dạng:** JSON (Content, Media, Stats).
- **Trạng thái:** Đã hoàn thành crawl.

Mẫu dữ liệu JSON (Crawl từ Facebook)


```
{
  "page_name": "Theanh28",
  "published_time": "2025-12-15T01:54:10",
  "content": "Vào 100 đám c■´■i xin đồ̀ăn nuôi 120 mèo hoang...",
  "stats": { "likes": 3800, "comments": 99 }
}
```

Nguồn dữ liệu 2: Tin tức (News)

- **Nguồn:** Báo Thanh Niên, Tuổi Trẻ, VnExpress, Vietnamnet...
- **Định dạng:** CSV (Title, Description, Content, Author).
- **Mục đích:** Dùng làm dữ liệu nền (baseline) hoặc kiểm chứng thông tin.

Mẫu dữ liệu CSV

```
article_id, url, title, content
"7287b...", "vietnamnet.vn/...", "Nhật Bản đau đầu...",
"Ngân khố của chính phủ Nhật Bản đãnhận đ■■c 129 tỷ Yen..."
```

Mục lục

Kiến trúc hệ thống dự kiến

Hệ thống chia làm 2 pha chính:

Phase 1: Offline Discovery

- Crawl dữ liệu quá khứ.
- Tiền xử lý & Vector hóa.
- Phân cụm (Clustering).
- Gán nhãn bán giám sát (Label Propagation).

Phase 2: Online Real-time

- Streaming Ingestion (Kafka).
- Inference thời gian thực.
- Cảnh báo & Lưu trữ.

Thay đổi & Cập nhật mô hình

Sau quá trình thử nghiệm (08/12/2025), nhóm đã thực hiện các điều chỉnh:

Vấn đề gặp phải

Dữ liệu "Daily basis"quá nhiều (Dự báo thời tiết hàng ngày, giá vàng, xổ số...) gây nhiễu lớn.

Giải pháp điều chỉnh

- **Thay đổi Model:** Chuyển từ mini-LLM sang **PhoBERT** để xử lý tiếng Việt tốt hơn.
- **Bổ sung NER:** Trích xuất thực thể để lọc nội dung rác và gom nhóm chính xác hơn.

Mục lục

Khó khăn hiện tại

1. **Vấn đề Phân cụm (Clustering):**
   - Chất lượng phân cụm chưa cao.
   - Cụm bị chồng lấn bởi bài viết tương tự (bán hàng, spam) nhưng không cùng sự kiện.
2. **Xử lý nhiễu (Noise filtering):**
   - Chưa có cơ chế triệt để loại bỏ tin lặp hàng ngày (giá vàng, thời tiết) khỏi luồng sự kiện nóng.

Kế hoạch tiếp theo (Timeline)

Task	Mô tả	Trạng thái
Task 1	Phân tích yêu cầu & Thiết kế	Đang làm
Task 2	Thu thập dữ liệu (Crawler)	**Hoàn thành**
Task 3	Nghiên cứu NLP (PhoBERT/NER)	TBD
Task 4	Cải thiện thuật toán Clustering	TBD
Task 5	Huấn luyện mô hình phân loại	TBD
Task 6	Tích hợp Pipeline Streaming	TBD

Bảng 1: Tiến độ thực hiện dự án

Mục lục

Tài liệu tham khảo I