

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

Báo cáo tiền độ dự án - SE363.Q11

Thực hiện:

Tăng Nhất¹ Lê Minh Nhật¹

GVHD: TS. Đỗ Trọng Hợp

Thầy Nguyễn Ngọc Quý

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

Tháng 12 năm 2025

Nội dung báo cáo

- 1 Tổng quan dự án
- 2 Dữ liệu & Thu thập
- 3 Phương pháp & Kiến trúc dự kiến
- 4 Khó khăn & Kế hoạch
- 5 Tài liệu tham khảo

Mục lục

- 1 Tổng quan dự án
- 2 Dữ liệu & Thu thập
- 3 Phương pháp & Kiến trúc dự kiến
- 4 Khó khăn & Kế hoạch
- 5 Tài liệu tham khảo

Mục tiêu đề tài

Xây dựng hệ thống **phát hiện xu hướng/sự kiện theo thời gian thực** từ dữ liệu mạng xã hội và báo chí, nhằm rút ngắn khoảng cách giữa *dữ liệu thô* và *thông tin có thể hành động*.

Hai nhóm đối tượng chính

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

Hướng tiếp cận: Kết hợp **tín hiệu đa nguồn** (Search–Social–News) và NLP để **gom topic, chấm điểm xu hướng** từ đó nắm bắt được tình trạng, phản ứng của mọi người về các xu hướng mới nhất.

Phân loại sự kiện (Taxonomy)

Hệ thống tập trung vào 7 loại sự kiện/xu hướng chính (phục vụ lọc nhiễu và gán nhãn đánh giá):

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.

Vì sao cần đa nguồn?

Search phản ánh *ý định*, Social phản ánh *độ lan truyền* nhưng nhiễu, News phản ánh *tính xác thực* nhưng có độ trễ.

Mục lục

- 1 Tổng quan dự án
- 2 Dữ liệu & Thu thập**
- 3 Phương pháp & Kiến trúc dự kiến
- 4 Khó khăn & Kế hoạch
- 5 Tài liệu tham khảo

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

- **Nguồn:** Các fanpage lớn / trang tin trên Facebook (ví dụ: Theanh28, Thông tin Chính Phủ, ...).
- **Định dạng:** JSON (nội dung, thời gian đăng, chỉ số tương tác).
- **Tín hiệu trend:** likes/comments/shares (có thể chuẩn hoá theo thời gian).
- **Trạng thái:** Đã hoàn thành crawl và đưa về schema thống nhất.

Mẫu dữ liệu JSON (Crawl từ Facebook)

```
{
  "page_name": "Theanh28",
  "published_time": "2025-12-15T01:54:10",
  "content": "Vào 100 đám cưới xin đồ ăn nuôi 120 mèo hoang...",
  "stats": { "likes": 3800, "comments": 99, "shares": 10 }
}
```

Nguồn dữ liệu 2: Tin tức (News)

- **Nguồn:** Báo Thanh Niên, Tuổi Trẻ, VnExpress, Vietnamnet, ...
- **Định dạng:** CSV (title, content, publish time, url, ...).
- **Vai trò:** Kiểm chứng sự kiện + tín hiệu **mật độ đưa tin** theo thời gian.
- **Trạng thái:** Đã **crawl** và hợp nhất vào pipeline chung.

Mẫu dữ liệu bài báo (News)

ID	URL	Tiêu đề	Nội dung	Thời gian
7287b...	vietnamnet.vn/...	Nhật Bản đau đầu với ngân sách	Ngân khố của chính phủ Nhật Bản đã nhận được 129 tỷ Yen từ các nguồn thu bất thường, gây tranh luận lớn trong nội bộ...	09/12/2025 10:37

Mục lục

- 1 Tổng quan dự án
- 2 Dữ liệu & Thu thập
- 3 Phương pháp & Kiến trúc dự kiến**
- 4 Khó khăn & Kế hoạch
- 5 Tài liệu tham khảo

Kiến trúc hệ thống dự kiến

Hệ thống chia làm 2 pha chính (phục vụ phát hiện xu hướng và vận hành thời gian thực):

Phase 1: Offline Discovery

- Thu thập dữ liệu lịch sử đa nguồn.
- Chuẩn hoá & tiền xử lý tiếng Việt.
- Vector hoá (embedding) và **clustering**.
- Tạo luật/gán nhãn bán giám sát để phân loại trend.

Phase 2: Online Real-time

- Thu thập định kỳ/streaming.
- Ghép bài viết vào topic + phát hiện topic mới.
- Chấm điểm xu hướng theo thời gian.
- Cảnh báo & lưu trữ cho dashboard.

Trọng tâm hiện tại

Phát hiện xu hướng (semantic matching + clustering + lọc nhiễu) là phần khó và tốn công nhất.

Pipeline NLP phát hiện xu hướng (đang phát triển)

Mục tiêu kỹ thuật

Gom các biểu đạt khác nhau (Facebook/News/Search) nhưng cùng một sự kiện thành **một topic thống nhất**.

- **Tiền xử lý:** chuẩn hoá Unicode, loại boilerplate, làm sạch ký tự/emoji.
- **Biểu diễn ngữ nghĩa:** dùng **embedding tiếng Việt** (PhoBERT / Vietnamese bi-encoder).
- **Tăng cường ngữ cảnh (2 hướng thử nghiệm):**
 - **NER (underthesea):** trích xuất thực thể để tăng độ khớp khi gom nhóm.
 - **Alias từ Google Trends:** mở rộng biên thể từ khoá để tăng recall khi khác cách diễn đạt.
- **Ghép & gom topic:** cosine similarity (matching) + UMAP (giảm chiều) + HDBSCAN (clustering).

Trend Scoring & Phân loại trend (theo đề xuất)

Với mỗi **topic**, hệ thống tổng hợp 3 tín hiệu:

- **Google Search Score (G)**: mức quan tâm/tăng trưởng tìm kiếm.
- **Facebook Engagement Score (F)**: mức tương tác được chuẩn hoá theo thời gian.
- **News Coverage Score (N)**: mật độ bài báo theo thời gian.

Unified Trend Score

$$T = w_G \cdot G + w_F \cdot F + w_N \cdot N$$

Trọng số có thể điều chỉnh theo mục tiêu (ưu tiên viral / xác thực / ý định tìm kiếm).

Phân loại xu hướng

Search-only / Social-only / News-only / Multi-source confirmed / Emerging / Fading.

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 **lặp hàng ngày** (thời tiết, giá vàng, xổ số, ...) xuất hiện dày đặc, tạo **topic giả** và gây nhiễu lớn.

Giải pháp điều chỉnh (đã/đang áp dụng)

- **Embedding tiếng Việt:** ưu tiên PhoBERT/Vietnamese bi-encoder để ghép ngữ nghĩa ổn định.
- **Bổ sung NER:** tăng độ chính xác khi gom nhóm + hỗ trợ lọc rác theo thực thể.
- **Bổ sung alias:** khai thác bộ từ khoá liên quan từ Google Trends để tăng khả năng match.

Kết quả thử nghiệm bước đầu

Những điểm đã đạt được

- Đã chạy được pipeline: **chuẩn hoá** → **embedding** → **matching** → **clustering**.
- Embedding tiếng Việt cho thấy khả năng gom các câu gần nghĩa tốt hơn so với so khớp từ khoá thuần.
- Alias từ Google Trends giúp tăng độ bao phủ khi cùng sự kiện nhưng khác cách gọi.

Vấn đề còn tồn tại

- **Cụm lớn bất thường** do nhiều (daily/spam) “nuốt” dữ liệu.
- **Overlap giữa cụm** khi nội dung na ná (bán hàng, câu view).
- Ngưỡng similarity nhạy: thay đổi nhỏ có thể làm lệch tỉ lệ matched/unmatched.

Hướng cải thiện ngay

Lọc nhiễu theo chu kỳ + từ khoá đặc trưng, tuning UMAP/HDBSCAN, và bổ sung đánh giá chất lượng cụm (kết hợp thủ công + chỉ số nội bộ).

Mục lục

- 1 Tổng quan dự án
- 2 Dữ liệu & Thu thập
- 3 Phương pháp & Kiến trúc dự kiến
- 4 Khó khăn & Kế hoạch**
- 5 Tài liệu tham khảo

Khó khăn hiện tại

1. Vấn đề Phân cụm (Clustering):

- Chất lượng phân cụm chưa cao; topic bị chồng lấn.
- Một số cụm quá lớn do “nuốt” dữ liệu nhiễu (daily/spam).

2. Xử lý nhiễu (Noise filtering):

- Chưa có cơ chế triệt để loại bỏ tin lặp (giá vàng, thời tiết, xổ số, ...).
- Cần kết hợp rule-based (chu kỳ/từ khoá) và thống kê theo thời gian.

3. Đồng bộ thời gian đa nguồn:

- Social nhanh, News có độ trễ, Search phản ánh ý định theo ngữ cảnh.
- Cần chuẩn hoá theo **time-binning** để chấm điểm công bằng.

Kế hoạch tiếp theo (Timeline)

Task	Mô tả	Trạng thái
Task 1	Chuẩn hoá schema + làm sạch cơ bản (Unicode/boilerplate)	Đang làm
Task 2	Thu thập dữ liệu (Crawler)	Hoàn thành
Task 3	Semantic matching (embedding) + alias/NER enrichment	Đang làm
Task 4	Tối ưu Clustering (UMAP + HDBSCAN) + đánh giá chất lượng cụm	TBD
Task 5	Trend scoring + phân loại trend đa nguồn	TBD
Task 6	Dashboard/báo cáo + tích hợp pipeline định kỳ/streaming	TBD

Bảng 1: Tiến độ thực hiện dự án

Mục lục

- 1 Tổng quan dự án
- 2 Dữ liệu & Thu thập
- 3 Phương pháp & Kiến trúc dự kiến
- 4 Khó khăn & Kế hoạch
- 5 Tài liệu tham khảo**

Tài liệu tham khảo I