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[2]
Nguyễn Ngọc Quí[2]

[1]Khoa Khoa học Máy tính
Trường Đại học Công nghệ Thông tin
[2]Khoa Khoa học và Kỹ thuật thông tin
Trường Đại học Công nghệ Thông tin

Ngày 16 Tháng 12 năm 2025

Nội dung báo cáo

Mục lục

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.

Ví dụ cụ thể: Theo dõi phản ứng xã hội theo thời gian thực

Ví dụ: Dự kiến đánh thuế hộ kinh doanh thu nhập 200 triệu/năm

Thông tin dự kiến đánh thuế được lan truyền nhanh trên mạng xã hội, kèm theo nhiều cách diễn đạt khác nhau và cảm xúc trái chiều.

- Mục tiêu hệ thống:
  - **Tóm tắt sự việc:** từ thông tin ban đầu $\rightarrow$ lan truyền $\rightarrow$ phản hồi chính thức.
  - **Cảm xúc chủ đạo của người dân:** lo lắng, bức xúc, hoang mang.
  - **Mức độ lan rộng:** thảo luận tăng mạnh trong thời gian ngắn.

Ý nghĩa

Giúp cơ quan quản lý **nắm bắt phản ứng xã hội sớm**, chủ động truyền thông để tránh hiểu sai và giảm bức xúc dư luận.

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

Nguồn dữ liệu 1: Mạng xã hội (Facebook)

- **Nguồn:** Fanpage lớn / trang tin Facebook (Theanh28, Dân trí, . . .).
- **Định dạng:** JSON (nội dung + thời gian đăng + tương tác).
- **Tín hiệu trend:** likes / comments / shares.
- **Trạng thái:** Đã crawl và đưa về schema thống nhất.

Hạn chế

Crawl thủ công **không luôn lấy được giờ/phút chính xác** với bài đăng cũ $\rightarrow$ ảnh hưởng đánh giá xu hướng theo thời gian.

Schema hiện tại (rút gọn)

```
{
  "page_name": "Theanh28",
  "published_time": "2025-12-15T01:54:10",
  "content": "...",
  "stats": { "likes": 3800, "comments": 99, "shares": 10 }
}
```

Nguồn dữ liệu 1: Facebook – Định hướng nâng chất lượng (Apify)

Mục tiêu

Sử dụng **Apify** để crawl Facebook với **timestamp chính xác** (giờ/phút/giây) cho cả bài cũ.

- Lấy được **time ISO + epoch timestamp**.
- Giữ thông tin postId, url, pageName, link bài báo.
- Đầy đủ tương tác: likes / comments / shares.

Schema Apify (các trường chính)

```
{
  "pageName": "baodantridientu",
  "postId": "1191254559782876",
  "time": "2025-12-15T09:41:09.000Z",
  "timestamp": 1765791669,
  "text": "...",
  "likes": 9, "comments": 1, "shares": 1
}
```

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

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

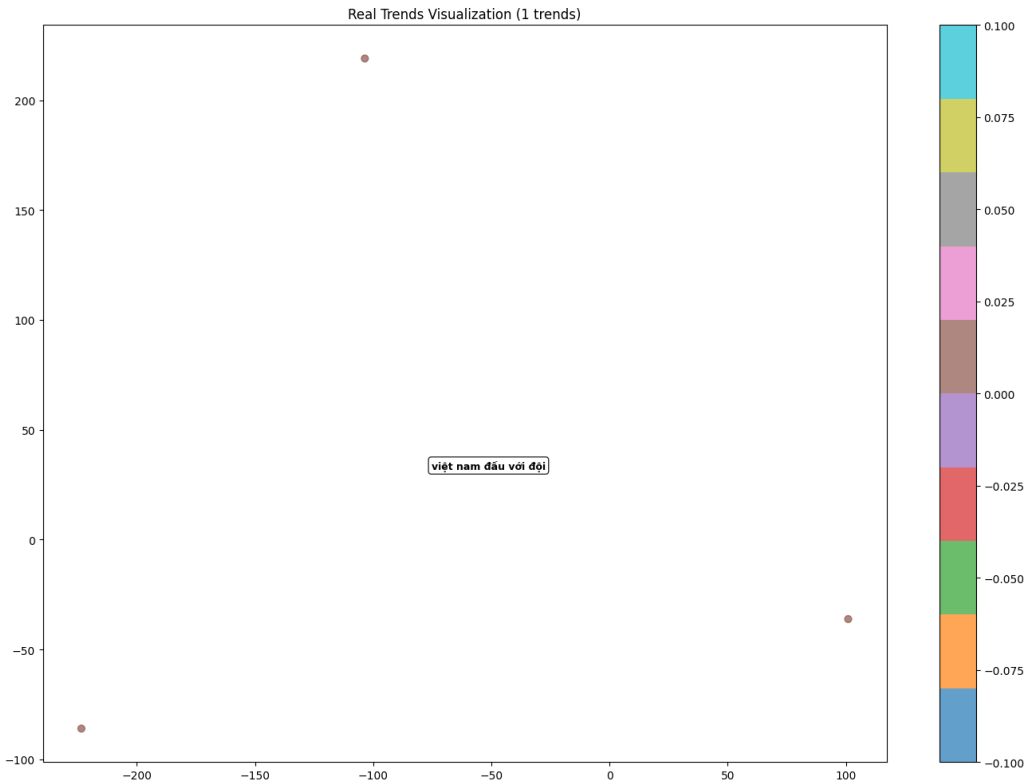
Gom các biểu đạt khác nhau (Facebook/News/Search) nhưng cùng một sự kiện thành **một topic thống nhất**.

- **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 biến thể từ khoá để tăng recall khi khác cách diễn đạt.
- **Ghép & gom topic:** cosine similarity (matching) + UMAP (giảm chiều) + HDBSCAN (clustering).

So sánh trực quan: Clustering với NER

Thiết lập

Embedding tiếng Việt (PhoBERT / bi-encoder) + **NER enrichment (underthesea)** → UMAP + HDBSCAN.

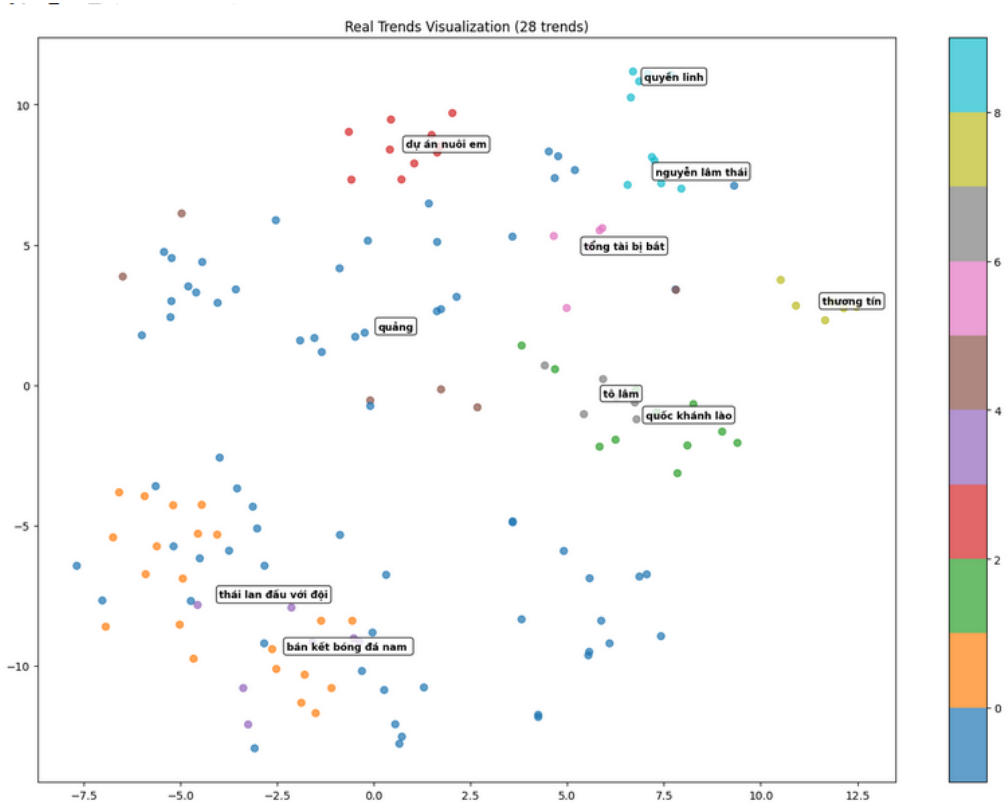


Hình 1: Phân bố cụm khi tăng cường ngữ cảnh bằng Named Entity Recognition

So sánh trực quan: Clustering với Alias từ Google Trends

Thiết lập

Embedding tiếng Việt (PhoBERT / bi-encoder) + **Alias normalization từ Google Trends** → UMAP + HDBSCAN.



Hình 2: Phân bố cụm khi mở rộng biến thể từ khoá từ Google Trends

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** $\rightarrow$ **embedding** $\rightarrow$ **matching** $\rightarrow$ **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 nhiễu (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

Khó khăn hiện tại

Trong thực tế, việc phát hiện một xu hướng "thật sự" là bài toán khó, do các thảo luận trên mạng xã hội thường bùng phát nhanh, nhiễu nhiều, và không phải lúc nào mức độ lan truyền cũng phản ánh đúng tầm quan trọng của sự kiện.

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)	Đã Hoàn thành các bộ Crawl
Task 3	Semantic matching (embedding) + alias/NER enrichment	Đang thử nghiệm
Task 4	Tối ưu Clustering (UMAP + HDBSCAN) + đánh giá chất lượng cụm	Đang tìm giải pháp tối ưu hơn
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

Tài liệu tham khảo I