Phát hiện Xu hướng và Phân loại cảm xúc 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.

Kế hoạch

Task	Mô tả	Trạng thái
Task 1	Thu thập dữ liệu đa nguồn (Crawler / Apify / News)	**Đã hoàn thành**
Task 2	Chuẩn hoá schema + làm sạch dữ liệu (Unicode, nội dung mẫu lặp)	Đang làm
Task 3	Semantic matching (embedding) + tăng cường ngữ cảnh (NER / alias)	Đang thử nghiệm
Task 4	Tối ưu phân cụm (UMAP + HDBSCAN) + đánh giá chất lượng cụm	Đang tìm giải pháp tối ưu hơn
Task 5	Trend scoring để xác định xu hướng	Đang thử nghiệm
Task 6	Đặt tên trend và tóm tắt thông tin cốt lõi cho mỗi cụm	TBD
Task 7	Phân loại cảm xúc, tâm lý cộng đồng theo từng trend	TBD
Task 8	Dashboard/báo cáo + tích hợp pipeline định kỳ/streaming	TBD

Mục lục

Nguồn dữ liệu 1: Mạng xã hội Facebook

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
}
```

Lợi ích

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

Pipeline NLP phát hiện xu hướng (đang phát triển)

Mục tiêu kỹ thuật

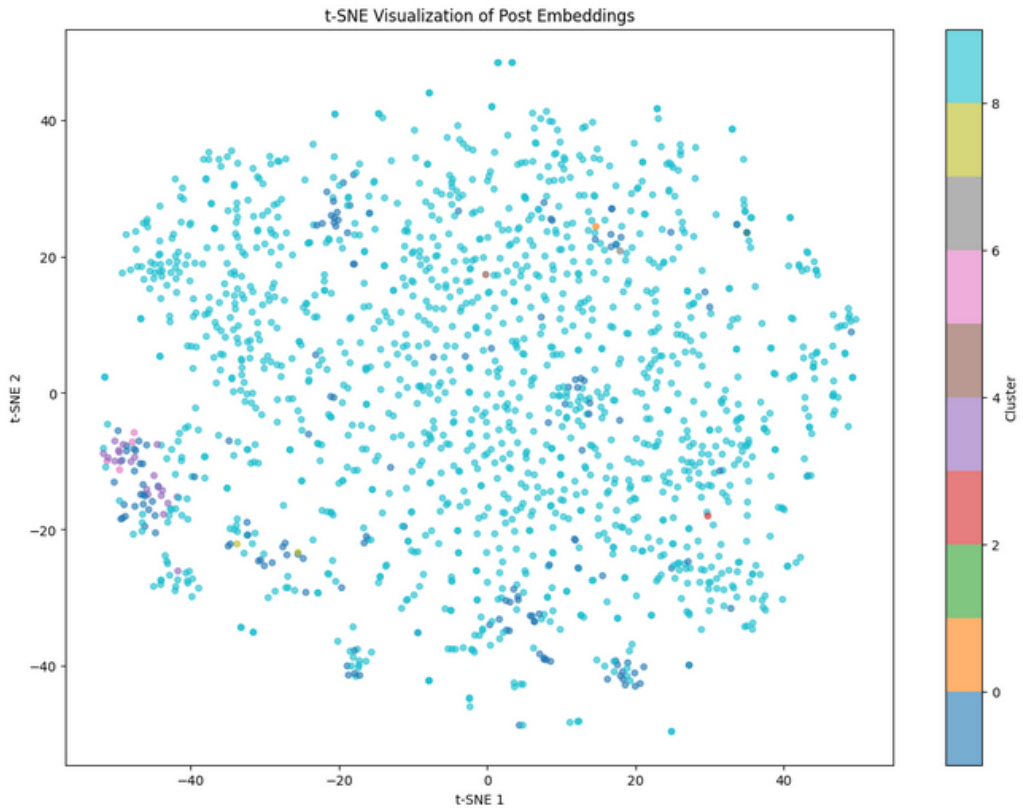
Gom các biểu đạt khác nhau (Facebook/News/Search) nhưng cùng một sự kiện thành **một topic thống nhất**.

- **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 (phân cụm).

So sánh trực quan: phân cụm với NER

Thiết lập

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

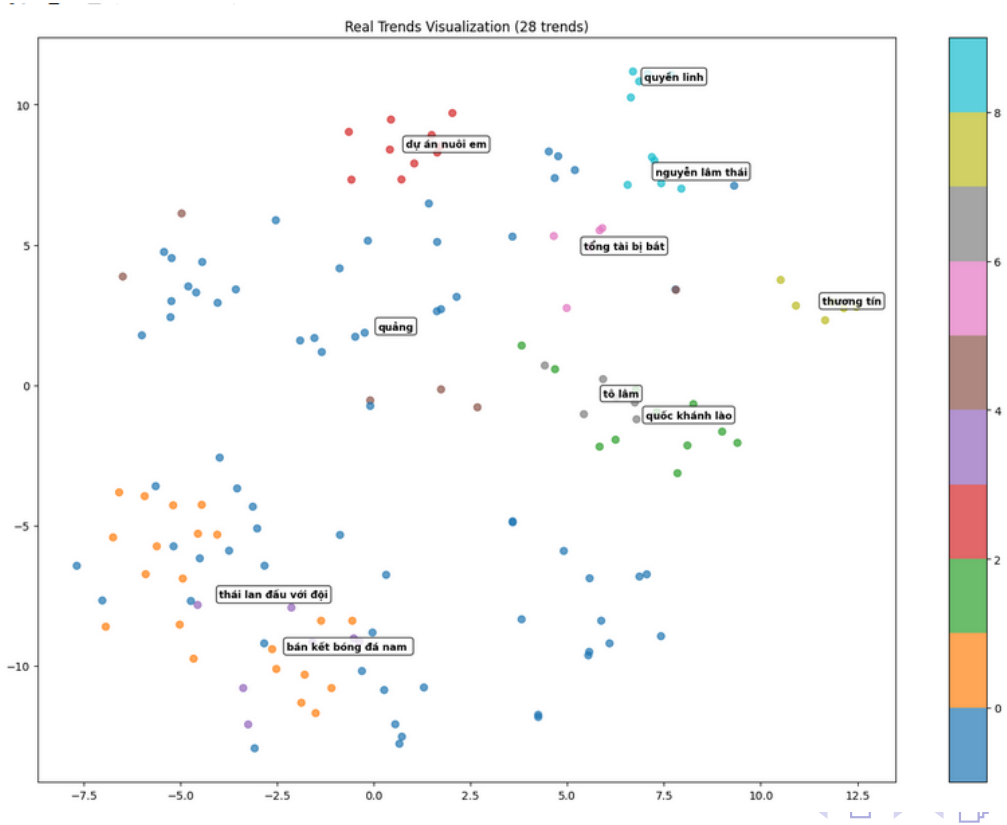


Hình 1: Phân bố cụm khi tăng cường ngữ cảnh bằng Named Entity Recognition

So sánh trực quan: phân cụm với Alias từ Google Trends

Thiết lập

Embedding tiếng Việt (PhoBERT / bi-encoder) + Cosine Matching (Alias normalization từ Google Trends) $\rightarrow$ UMAP + HDBSCAN.



Trend Scoring & Phân loại trend (ý tưởng)

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.

Kết quả thử nghiệm bước đầu

Những điểm đã đạt được

- Đã chạy được pipeline: **chuẩn hoá** $\rightarrow$ **embedding** $\rightarrow$ **matching** $\rightarrow$ **phân cụm**.
- 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).
- **Nếu phụ thuộc quá vào từ khóa google trend, sẽ kém hiệu quả với các trend mới hoàn toàn.**

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 (phân cụm):**
   - 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 có cơ chế đồng bộ công bằng.

Mục lục

Tài liệu tham khảo I