Nhóm: AIT2006-1-2.4

Đỗ Hoàng Long - 24022389
Phạm Gia Bảo - 24022267
Nguyễn Sỹ Quyền - 24022437

Năm dữ liệu: 2011
Ngày nộp: 19/12/2025
Commit hash: 0c78b600ddd393c869c3a4f76f0983ba6f702fc5

1. Giới thiệu & tổng quan dữ liệu

1.1. Nguồn gốc và Mục tiêu

Dự án này phân tích bộ dữ liệu từ **IEEE VAST Challenge 2011 - Mini Challenge 1**. Mục tiêu là truy vết đại dịch tại thành phố giả lập Vastopolis thông qua việc kết hợp dữ liệu microblogs và thời tiết, từ đó xác định nguồn gốc, sự lây lan và tác động của dịch bệnh.

1.2. Phạm vi dữ liệu (Scope)

Nhóm 2.4 chịu trách nhiệm phân tích dữ liệu trong khoảng thời gian:

- **Từ ngày:** 12/05/2011
- **Đến ngày:** 15/05/2011

1.3. Đặc điểm chính của dữ liệu

Dữ liệu đầu vào bao gồm:

- **Microblogs:** Chứa nội dung văn bản (text), thời gian (created_at), vị trí (lat, long), id
- **Weather:** Dữ liệu thời tiết theo chu kỳ, bao gồm hướng gió và tốc độ gió.
- **Map:** Ảnh bản đồ Vastopolis dùng để đối chiếu tọa độ địa lý.

2. Làm sạch & chuẩn hóa

2.1. Quy tắc Đảm bảo Chất lượng (QA Rules)

Áp dụng các quy tắc sau để kiểm tra dữ liệu:

- Gắn cờ các dòng thiếu nội dung (text), thiếu vị trí hoặc thiếu thông tin thời tiết.
- Gắn cờ các tọa độ nằm ngoài phạm vi bản đồ Vastopolis hoặc tốc độ gió âm.
- Kiểm tra định dạng thời gian (timestamp).
- Kiểm tra trùng lặp trong danh sách từ khóa.

Bảng dữ liệu	Tiêu chí kiểm tra (Flag)	Số lượng vi phạm	Tổng số dòng	Tỷ lệ (%)
Microblogs	Sai định dạng thời gian	0	182,422	0.0%
Microblogs	Thiếu nội dung (missing_text)	36	182,422	0.02%
Microblogs	Thiếu vị trí (missing_location)	0	182,422	0.0%
Microblogs	Vị trí sai (invalid_location)	0	182,422	0.0%
Weather	Lỗi dữ liệu chung	0	4	0.0%
Keywords	Từ khóa trùng lặp	1	1,302	0.08%

Chất lượng dữ liệu khá tốt. 36 dòng thiếu text trong Microblogs được giữ lại (gắn cờ) để không làm mất thông tin thời gian/vị trí. 1 từ khóa trùng lặp trong bảng Keywords đã được xử lý.

2.2. Chuẩn hoá dữ liệu

- **Thời gian:** Chuyển đổi sang định dạng **`datetime`** chuẩn (YYYY-MM-DD HH:MM:SS).
- **Văn bản:** Chuyển về chữ thường (lowercase), loại bỏ khoảng trắng thừa, xử lý encoding (UTF-8/Latin-1).
- **Danh mục:** Chuẩn hóa cột **`weather`** và **`wind_direction`**

2.3. Kết quả

- **Input:** microblog.csv, weather.csv, keyword.csv (thư mục **raw/**)
- **Output:** microblog_clean.csv, weather_clean.csv, keyword_cleaned.csv (thư mục **processed/**)

3. Tổng hợp dữ liệu

3.1. Kết nối dữ liệu (Join)

Dữ liệu thời tiết được nội suy (upsample) theo giờ và kết nối với microblogs dựa trên trường thời gian.

- **Input:** microblog_clean.csv, weather_clean.csv, keyword_cleaned.csv (thư mục **processed/**).
- **Output:** Bảng tổng hợp chứa thông tin thời tiết cho từng tweet.(không lưu file)

3.2. Phân loại từ khóa

Sử dụng danh sách **`keywords_cleaned.csv`** để phân loại thành 2 nhóm:

- **Symptom Keywords:** flu, fever, stomach, ...
- **Other Keywords:** ...

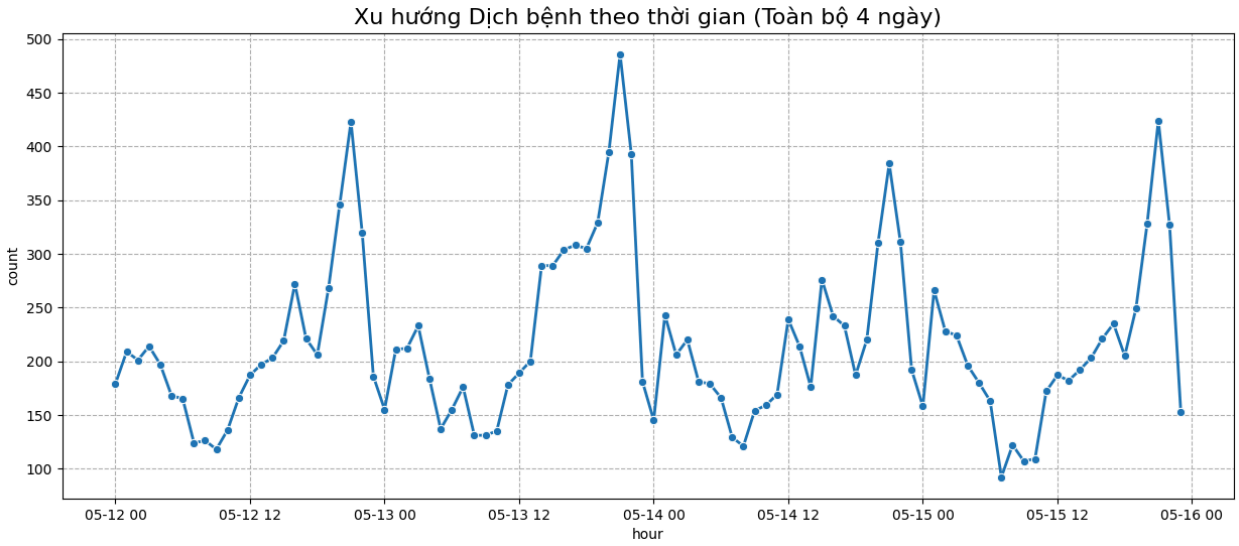
3.3. Bảng tổng hợp kết quả

Dữ liệu đã xử lý được lưu tại thư mục **`processed/`** với định dạng:

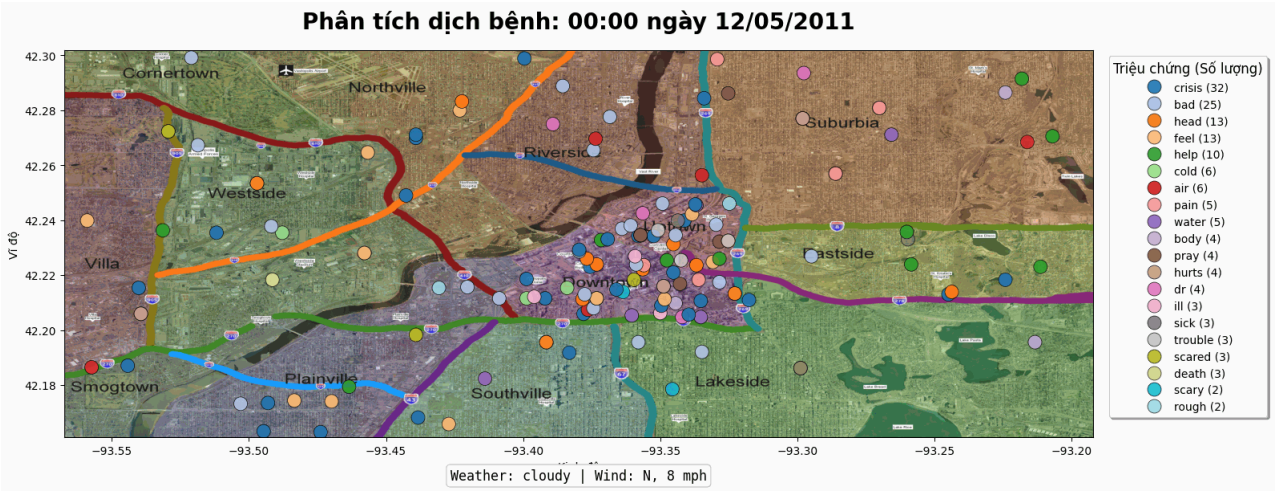
- **`stat_hourly_HH_DD_MM.csv`:** Thống kê số lượng keyword theo giờ.
- **`keyword_location_mapping_hourly_HH_DD_MM.csv`**: Ánh xạ vị trí.
- **keyword.csv:** tổng hợp các keyword liên quan đến dịch bệnh

4. Trực quan hóa

4.1. Biểu đồ xu hướng

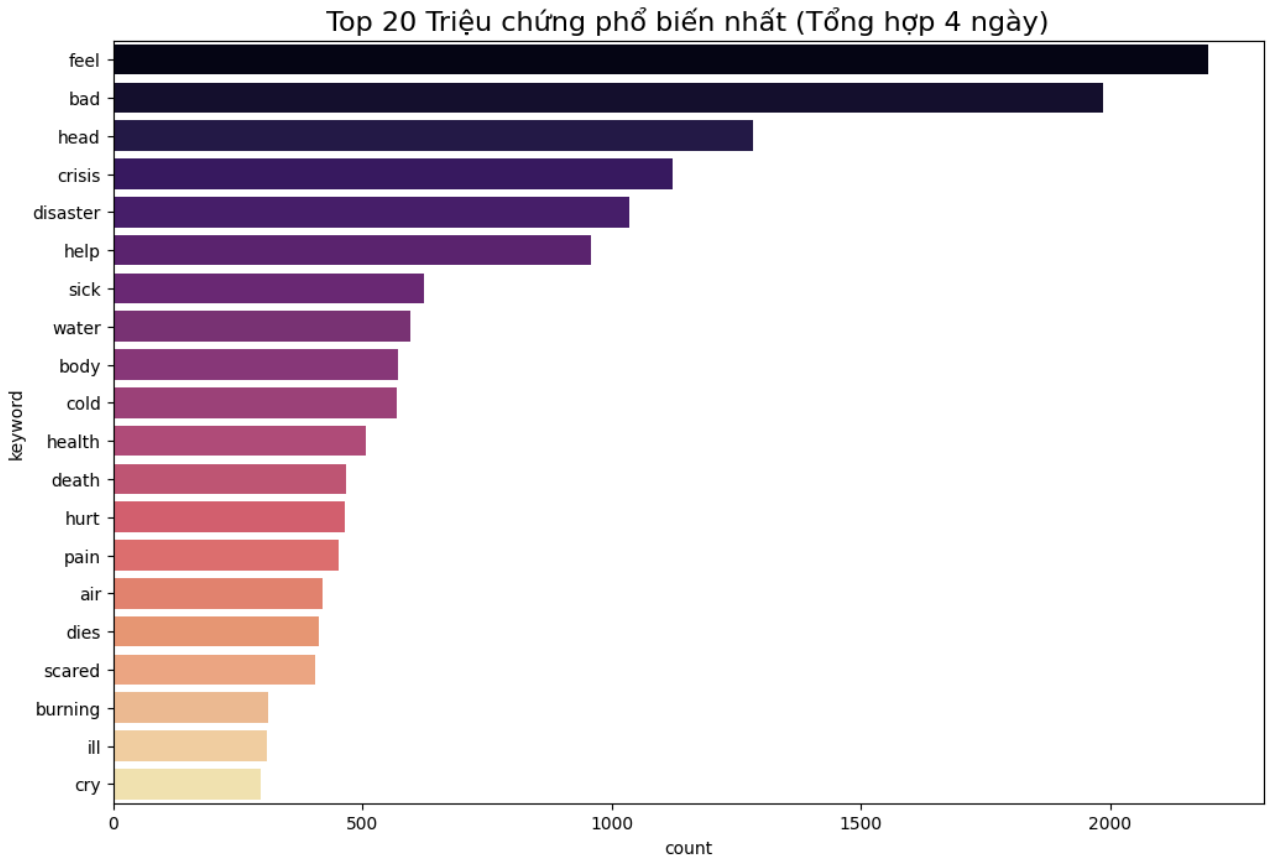


4.2. Bản đồ phân bố dịch bệnh theo giờ (GIS Map)



4.3. Wordcloud

4.4 Biểu đồ cột



5. Các phân tích

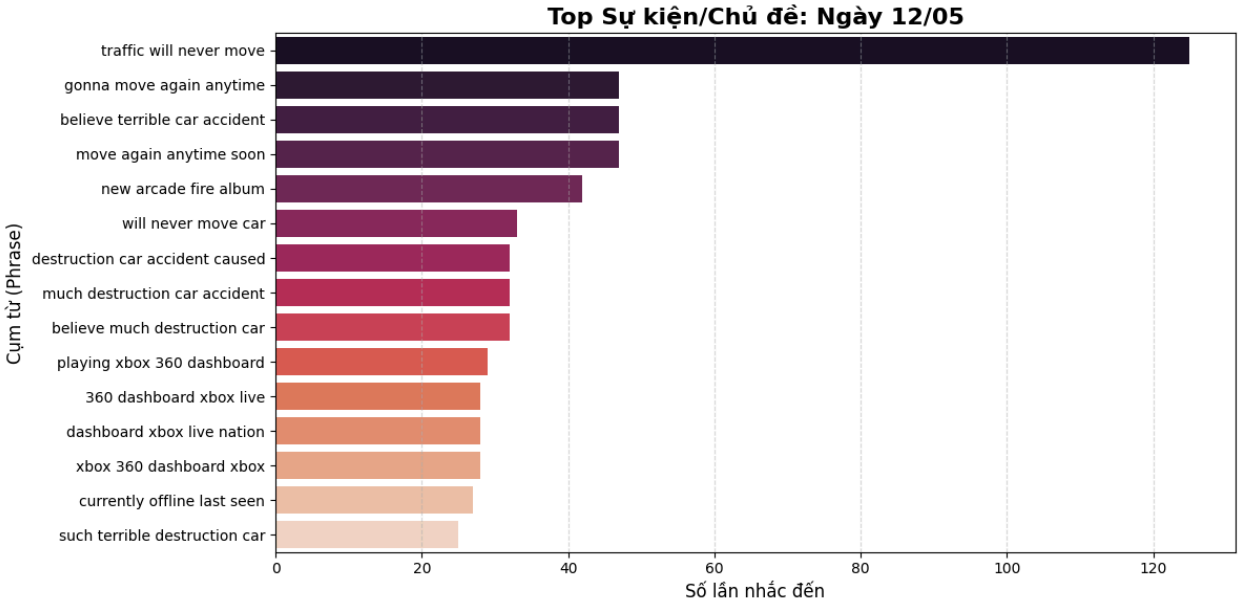
I. PHÂN TÍCH CÁC SỰ KIỆN CHÍNH THEO THỜI GIAN

1.Ngày 12/05

Dữ liệu cho thấy thành phố hoạt động bình thường.

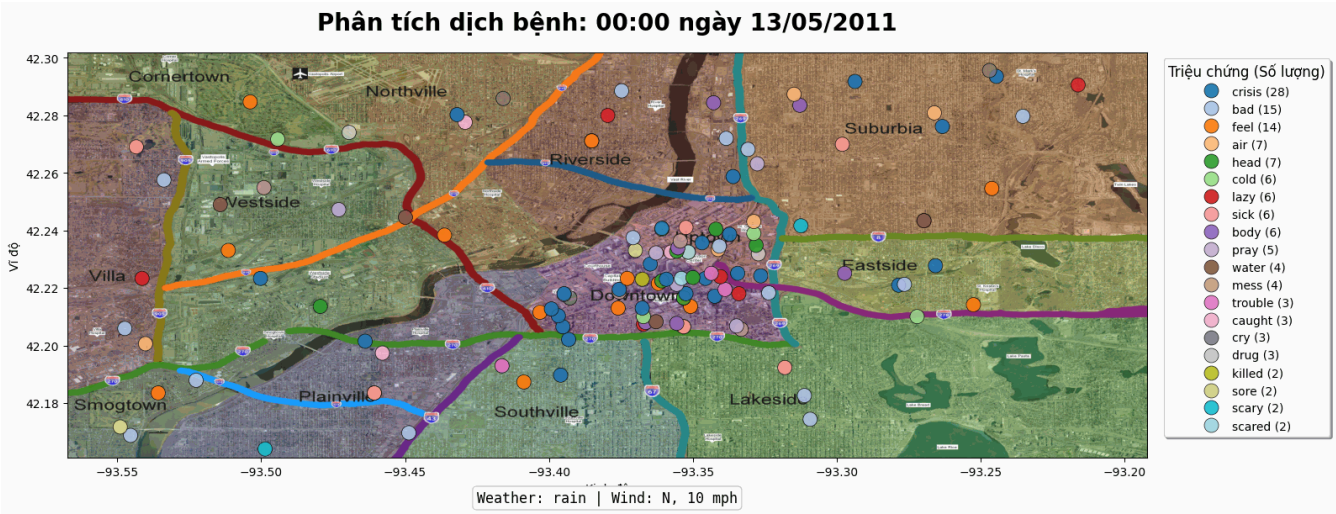
Sự kiện: Không có ghi nhận về tai nạn hay dịch bệnh diện rộng.

Chủ đề: Thảo luận xoay quanh đời sống thường nhật, giao thông và giải trí.

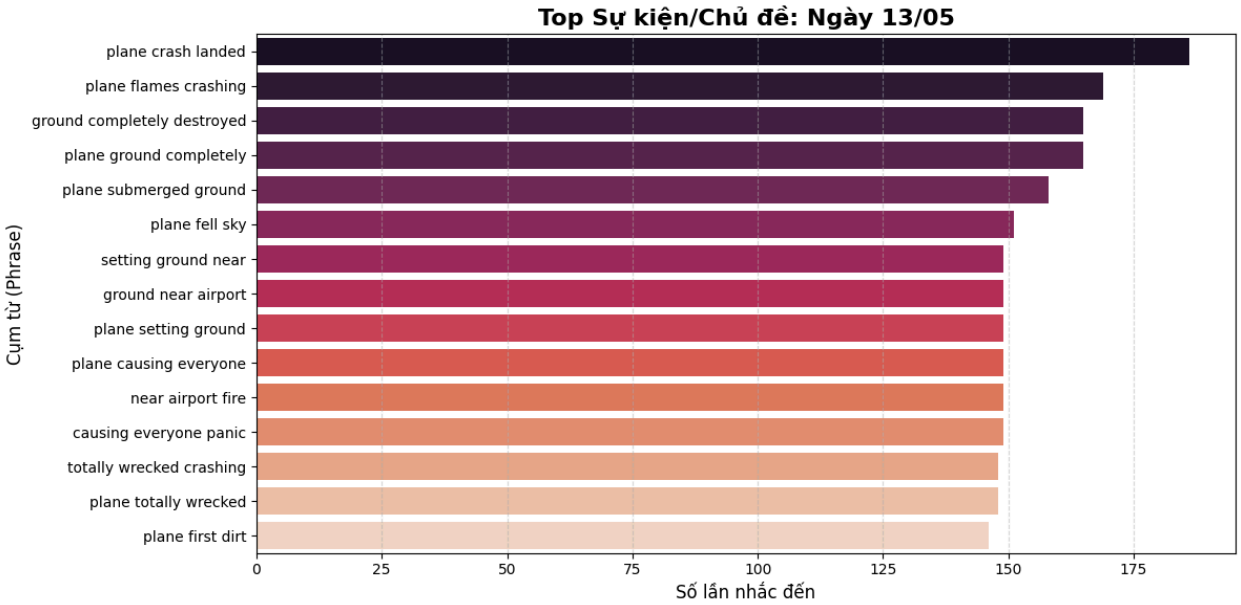


2.Ngày 13/05

Sự kiện: Máy bay rơi và bốc cháy tại Sân bay (Góc Tây Bắc). Xuất hiện một cụm "dịch" dày tại khu vực Sân bay. Gió thổi từ Bắc xuống Nam làm xuất hiện điểm nóng tập trung dày đặc ở **Downtown** (nằm ngay phía Nam Sân bay) và lan dần xuống **Southville**. Cơn gió này đã biến một tai nạn cục bộ thành thảm họa diện rộng

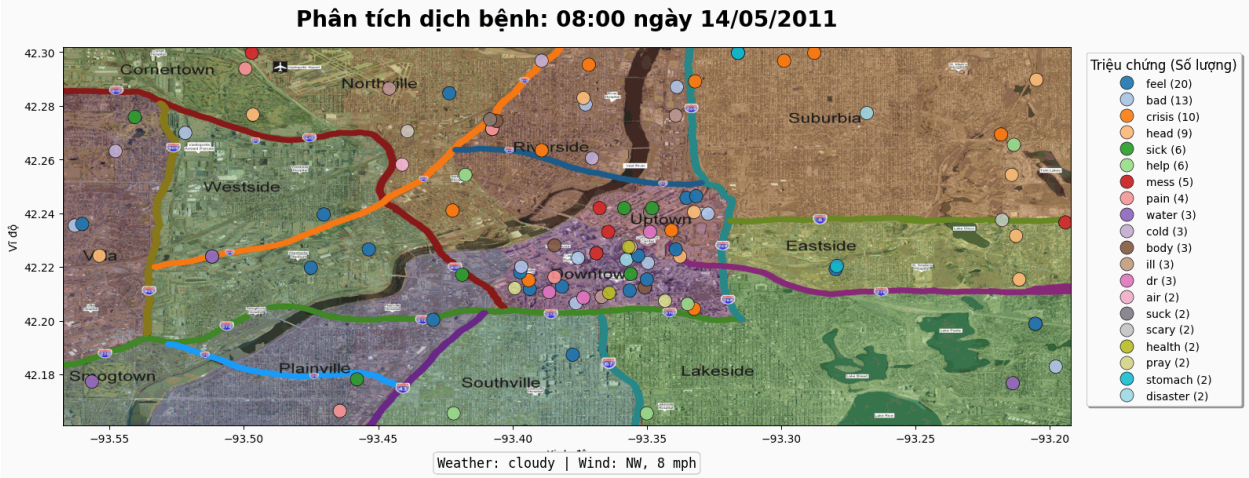


- Hàng loạt tweet xác nhận từ nhân chứng hiện trường:
  - *'i just saw a crashed plane looking at a wasteland'* (id 79809)
  - *'looking at a plane that is totally wrecked from crashing'* (85389)
  - *'plane that is in flames'* (80557)

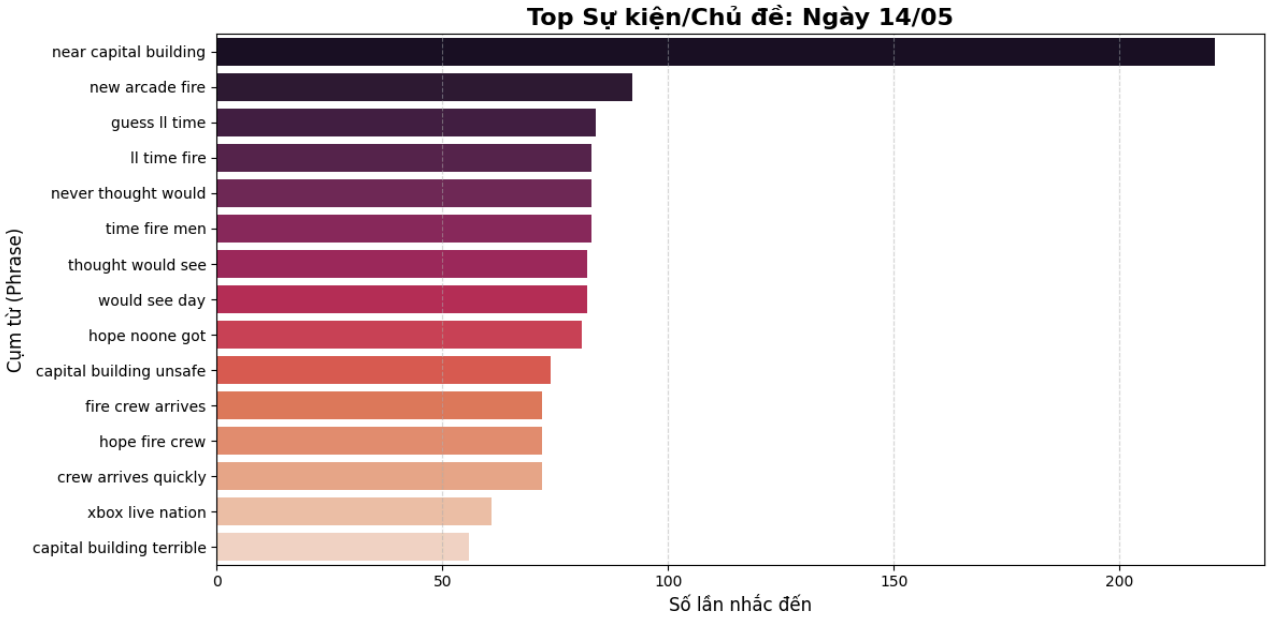


3.Ngày 14/05

Sự hồi phục sau thảm họa máy bay: Cơn mưa đêm 13/05 đã làm không khí. Các triệu chứng ngộ độc khí (Ho, Rát mắt) giảm mạnh trên toàn thành phố trong đêm 13/5 và rạng sáng 14/5.



Sự kiện chính: Hỏa hoạn tại Capital Building.



Mốc Thời gian	Diễn biến (Dựa trên Log)	Bằng chứng Tweet (Verbatim)
09:01	**BẮT ĐẦU:** Phát hiện cháy	*"near the capital building is a destructive fire emergency services go!"*
09:02	**NGUYÊN NHÂN:** Nghi vấn chập điện	*"an **electrical fire** on the second story of a building forced the temporary closure"*
09:41	**CAO TRÀO:** Hoảng loạn & Sơ tán	*"capital building unsafe people are **running for their lives**"*
12:30	**CẢNH SÁT CAN THIỆP:** Sơ tán	*"there was smoke and **policemen telling us to evacuate** the building"*
15:07	**TIN ĐỒN:** Nghi vấn trộm cướp	*"is there a fire in the building? an **escaped robber**?"* (Tin đồn vô căn cứ do thấy cảnh sát)
15:51	**KIỂM SOÁT:** Dập lửa	*"huge fire in five-storey building: at least 50 firefighters are tackling a huge blaze"*

Trích xuất Log điển hình:

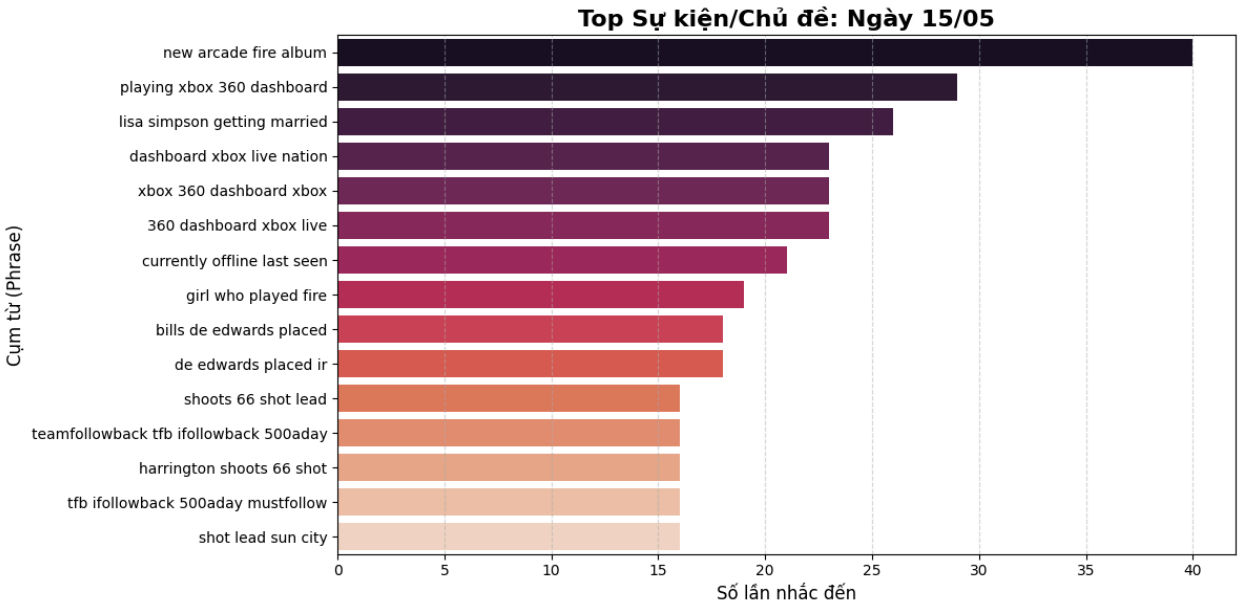
- *"an electrical fire on the second story of a building forced the temporary closure"* (83969)
- *"this building is on fire guys... smoke and policemen telling us to evacuate"* (89341)

4. Ngày 15/05

Dữ liệu ngày 15/05 cho thấy không có làn sóng lây nhiễm nào:

Chủ đề cộng đồng:

- Các cuộc thảo luận trên mạng xã hội tập trung vào các tin tức giải trí.
- Các từ khóa liên quan đến các sự kiện hay tai nạn không còn xuất hiện nhiều, cho thấy sự kiện đã trôi qua và không để lại di chứng môi trường kéo dài.



II. LỌC KEYWORDS

Mục tiêu: Quét và phân tích ngữ nghĩa các từ khóa liên quan đến tội phạm, vũ khí và hành vi khả nghi trên toàn bộ dữ liệu 4 ngày.

- Phát hiện sớm (early detection) các sự kiện liên quan đến:
  - Tội phạm nghiêm trọng
  - Khủng bố, bạo lực có tổ chức
  - Âm mưu, chuẩn bị gây dịch bệnh
- Giảm nhiễu thông tin từ các nội dung:
  - Giải trí, công nghệ, đời sống cá nhân
  - Ẩn dụ ngôn ngữ, nói đùa, meme
- Phân biệt rõ tin tức chính thống và hành vi/ý định mờ ám ở giai đoạn tiền sự kiện

Các chủ đề được phân loại:

1. Tội phạm
2. Khủng bố & vụ nổ
3. Vũ khí sinh học/hóa học
4. Hoạt động khả nghi

Các phát hiện:

- Các sự kiện tội phạm đường phố không liên quan đến một dịch bệnh nào
- Các vụ khủng bố, nổ chủ yếu là tin tức thế giới
- Vụ nổ tại địa phương do rò rỉ khí, gây ra thương vong tuy nhiên không tạo ra dịch bệnh, xảy ra tại vùng trung tâm(Downtown)

Thời gian	Giai đoạn	ID	Nội dung sự kiện
20:31	**XÁC NHẬN RÒ RỈ**	847	**"**gas leak at camosun lansdowne campus fire crews on scene**"**. Đây là tọa độ chính xác của nguồn phát: **Camosun Lansdowne**.
20:51	**Vụ nổ/Tiếng động lớn**	109363	Người dân nghe thấy tiếng nổ lớn (*"hearing grenades"*), thực chất là tiếng nổ do áp suất khí hoặc cháy nổ tại hiện trường rò rỉ.
23:15	**Xác nhận Cộng đồng**	17752	Cư dân thảo luận về việc kiện tụng công ty BP (*"suing bp for that gas leak"*), xác nhận sự cố rò rỉ là nguyên nhân chính.
23:46	**Phản ứng An ninh**	72437	Trực thăng cảnh sát (*"cop chopper"*) xuất hiện giám sát hiện trường.
00:54 (14/5)	**Cứu hộ Tổng lực**	126216	Huy động quy mô lớn: *"a thousand cop cars, 8 emts, 40 ambulances"*. Chứng tỏ số lượng nạn nhân thương vong rất lớn.

- Về vũ khí sinh học: củng cố các bằng chứng về vụ *Rò rỉ khí gas tại Camosun Lansdowne, ngoài ra còn các vụ lẻ tẻ khác*

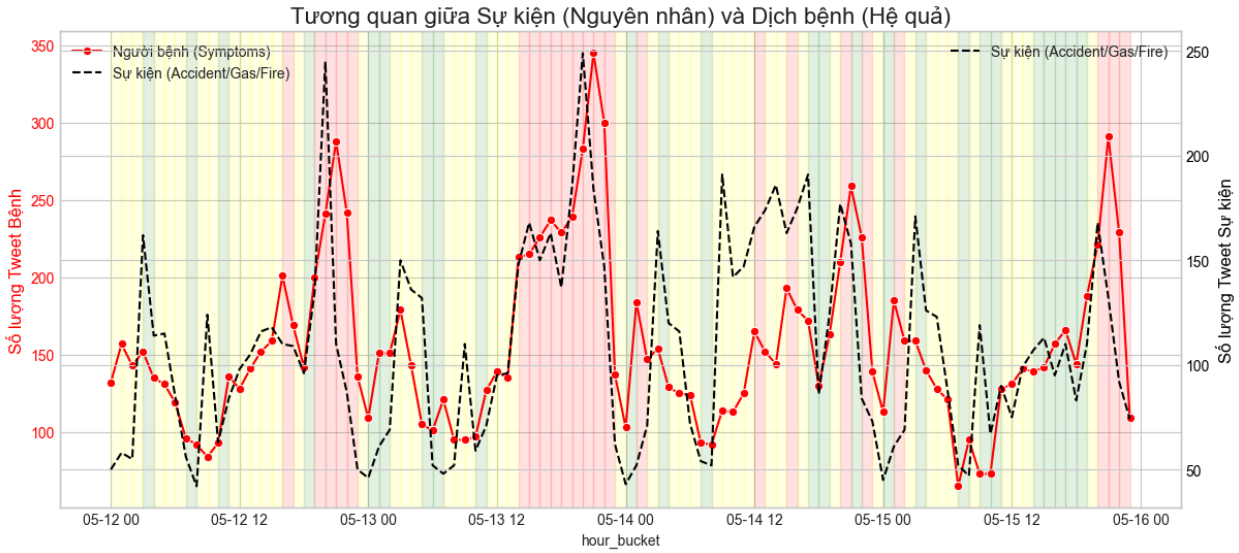
ID	Thời gian	Tọa độ	Nội dung Microblog
129289	12/05 01:22	42.23776 93.34204	a house a few blocks from us exploded tonight. **gas leak**?
10549	12/05 03:41	42.16729 93.20145	at least i like the smell of gasoline...
989	13/05 20:10	42.21004 93.3684	observatory; finding a way to stop ricin in its deadly path

847	13/05 20:31	42.19117 93.39756	**gas leak at camosun lansdowne campus fire crews on scene** (QUAN TRỌNG NHẤT)
17752	13/05 23:15	42.17053 93.3692	okay who all is suing bp for that **gas leak**??? (Xác nhận)

III. KẾT LUẬN

Tổng hợp dữ liệu từ ngày 12 đến ngày 15/05 cho thấy:

- Không có các sự kiện liên quan đến khủng bố hay một dịch bệnh lây lan, chỉ xuất hiện các triệu chứng thông thường
- Biểu đồ số lượng các blog xuất hiện nhiều vào các khung giờ tối không phải là sự bùng nổ dịch bệnh mà là hành vi người dùng mạng xã hội
    - Ban ngày mọi người bận đi làm, đi học nên ít blog.
    - Buổi tối (Sau 18:00): thời gian rảnh nên mọi người lướt mạng xã hội, đăng bài nhiều hơn.Do tổng lượng bài đăng (Total Volume) tăng vọt vào buổi tối, nên xác suất xuất hiện các từ khóa bất kỳ (kể cả từ "sick", "tired",”feel”) cũng sẽ tăng theo một cách tự nhiên.
    - Và ở cuối ngày, con người thường cảm thấy mệt mỏi về thể chất sau một ngày làm việc. Họ sẽ tweet: *"Tired from work"* , *"My head hurts from studying"* , *"Sick of this traffic"*.nên khi sử dụng các keyword để lọc thì sẽ cho các blog này vào mục bệnh tật, dù thực tế đó chỉ là sự than vãn đời thường chứ không phải bệnh lý.
    - Với các bệnh thông thường (cảm cúm nhẹ), triệu chứng thường có xu hướng nặng hơn hoặc được cảm nhận rõ hơn vào buổi tối/đêm: Sốt thường tăng về chiều tối. Khi cơ thể nghỉ ngơi, não bộ tập trung hơn vào các cơn đau nhức mà ban ngày bị sự bận rộn che lấp.



6. Diễn giải & kết luận

6.1. Phát hiện chính

- Sự kiện máy bay rơi và bốc cháy tại Sân bay (Góc Tây Bắc) đầu buổi chiều 13/5. Triệu chứng phổ biến nhất là Ho, Rát mắt (do ngộ độc khí), tập trung dày đặc ở khu vực Downtown (ngay phía Nam Sân bay) và lan dần xuống Southville.Có mối tương quan mạnh giữa hướng gió Bắc xuống Nam và sự lan truyền của các từ khóa triệu chứng, biến tai nạn cục bộ thành thảm họa diện rộng.
- Vụ nổ tối 13/5 gây ra nhiều thương vong nhưng không tạo làn sóng dịch bệnh
- Hỏa hoạn tại Capital Building, tuy nhiên sự kiện này không tạo ra làn sóng dịch bệnh lây lan.

6.2. Giới hạn

- Dữ liệu microblog chứa nhiều nhiễu, cảm xúc cá nhân không phản ánh chính xác bệnh lý.
- Dữ liệu vị trí có thể không chính xác tuyệt đối.

6.3. Đề xuất tiếp theo

- Kết hợp thêm dữ liệu từ bệnh viện (nếu có).
- Sử dụng mô hình NLP tiên tiến (BERT/LSTM) để phân loại triệu chứng chính xác hơn thay vì chỉ dựa vào từ khóa.

7.Tham khảo

- IEEE VAST Challenge 2011 Documentation.
- Gemini
- Google Search