NGHIỆM THU ĐỀ TÀI NCKHSV CẤP TRƯỜNG

Xây dựng ứng dụng hỗ trợ đầu tư chứng khoán bằng trí tuệ nhân tạo và phân tích cơ bản

Nhóm sinh viên thực hiện:

Nguyen Quang Huy - 523H0140

Nguyen Tran Nhat An - 523H0115

GVHD: TS. Trịnh Hùng Cường

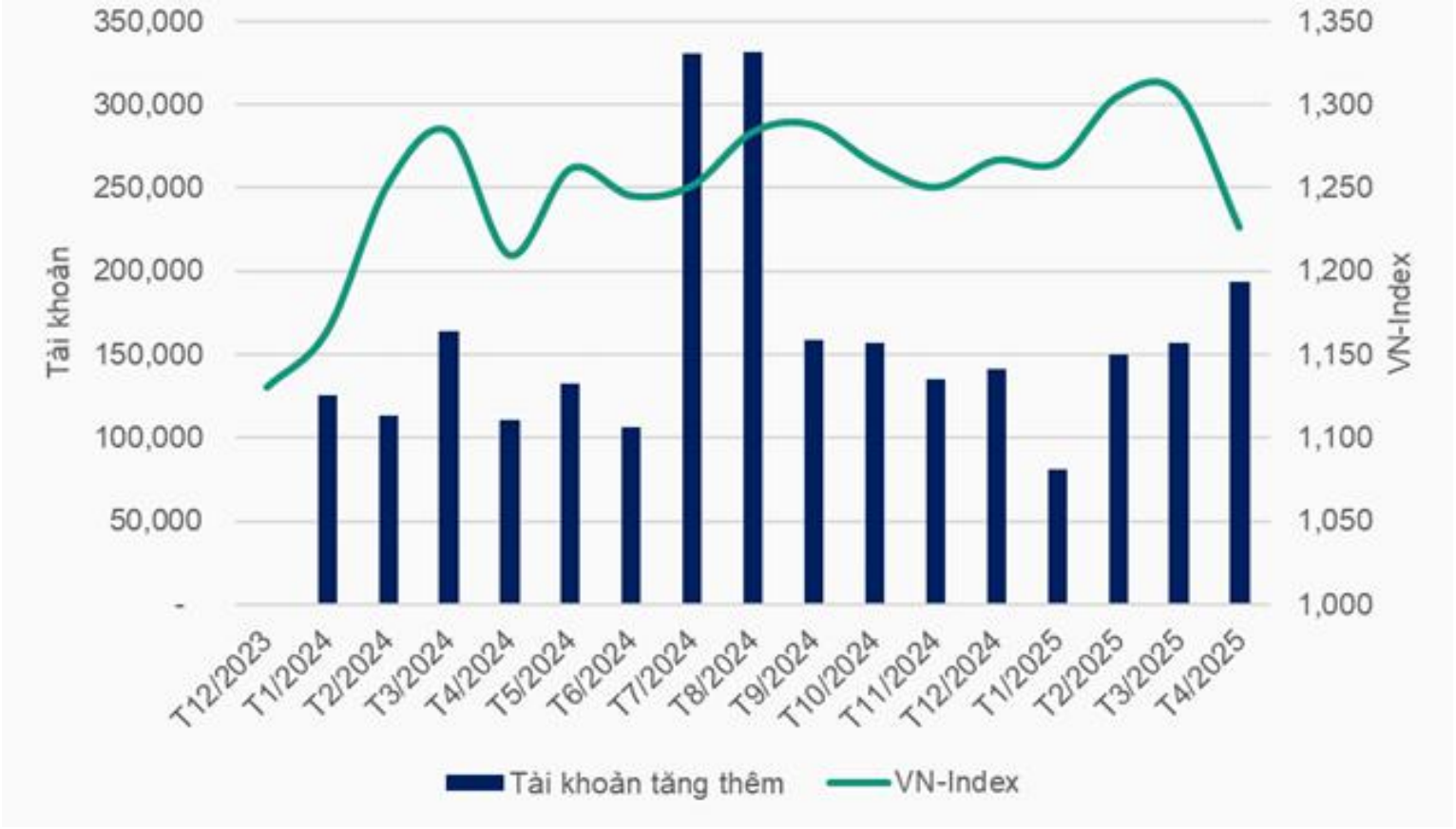
Tổng quan (Agenda)

1. Đặt vấn đề và Lí do chọn đề tài.
2. Mục tiêu và Phạm vi nghiên cứu.
3. Phương pháp luận - Gán nhãn dữ liệu
4. Kiến trúc hệ thống - Quy trình thực hiện.
5. Kết quả thực nghiệm.
6. Kết luận & Hướng phát triển.

1. Đặt vấn đề và Lý do chọn đề tài (1)

- Thị trường chứng khoán Việt Nam đang phát triển mạnh mẽ, thu hút lượng lớn nhà đầu tư mới.

Hình 1.1.1 (Báo cáo)



1. Đặt vấn đề và Lý do chọn đề tài (2)

- **Vấn đề:** Đa số nhà đầu tư mới thiếu kinh nghiệm, dẫn đến quyết định cảm tính và rủi ro cao.
- **Khoảng trống:** Các công cụ hiện tại ở Việt Nam chủ yếu tập trung vào phân tích kỹ thuật, bỏ qua dữ liệu định tính (tin tức, báo cáo) vốn rất quan trọng.
- **Giải pháp:** Nghiên cứu này xây dựng một hệ thống toàn diện, kết hợp cả AI dự báo và phân tích cơ bản để giải quyết khoảng trống trên.

2. Mục tiêu và Phạm vi nghiên cứu (1)

- **Mục tiêu chính:** Xây dựng hệ thống hỗ trợ đầu tư, kết hợp mô hình AI (LSTM, PhoBERT) để phân tích cả dữ liệu định lượng và định tính.
- **Phạm vi nghiên cứu:**
  - Không gian: Thị trường chứng khoán Việt Nam.
  - Đối tượng: Các cổ phiếu trong rổ VN30.
  - Thời gian: 2010 đến nay.

3. Phương pháp luận - Gán nhãn dữ liệu (1)

- **Thách thức lớn:** Xây dựng mô hình phân tích cảm xúc tiếng Việt cần bộ dữ liệu gán nhãn chất lượng, nhưng gán nhãn thủ công rất tốn kém và mang tính chủ quan.
- **Giải pháp đột phá:** Áp dụng phương pháp "Giám sát yếu" (Weak Supervision) với framework Snorkel.
- **Logic:** Kết hợp nhiều quy tắc gán nhãn (Label Function) để tự động tạo nhãn cho bộ dữ liệu.

3. Phương pháp luận - Gán nhãn dữ liệu (2)

- **Các quy tắc gán nhãn:**
  - **Ngưỡng ảnh hưởng** (T): Hệ số quyết định ảnh hưởng.
    - pct > T: Tích cực
    - pct < -T: Tiêu cực
    - -T ≤ pct ≤ T: Trung tính

    với **pct** là các hệ số sinh lời được tính bằng phương pháp biến động giá tuyệt đối hoặc lợi nhuận bất thường.

3. Phương pháp luận - Gán nhãn dữ liệu (3)

- **Các quy tắc gán nhãn:**
  - **LF1** - Dựa trên biến động giá trị tuyệt đối: Tỉ suất lợi nhuận của một cổ phiếu, xác định sau khi tin tức được công bố.

$$pct = \frac{price\ (t+1)\ -\ price\ (t)}{price\ (t)}$$

  - $price\ (t+1)$ là giá cổ phiếu sau khi tin tức được công bố.
  - $price\ (t)$ là giá cổ phiếu trước ngày công bố.

3. Phương pháp luận - Gán nhãn dữ liệu (4)

- **Các quy tắc gán nhãn:**
  - **LF2** - Dựa trên lợi nhuận bất thường (Alpha): Phần chênh lệch giữa lợi nhuận thực tế của một cổ phiếu so với lợi nhuận kỳ vọng.

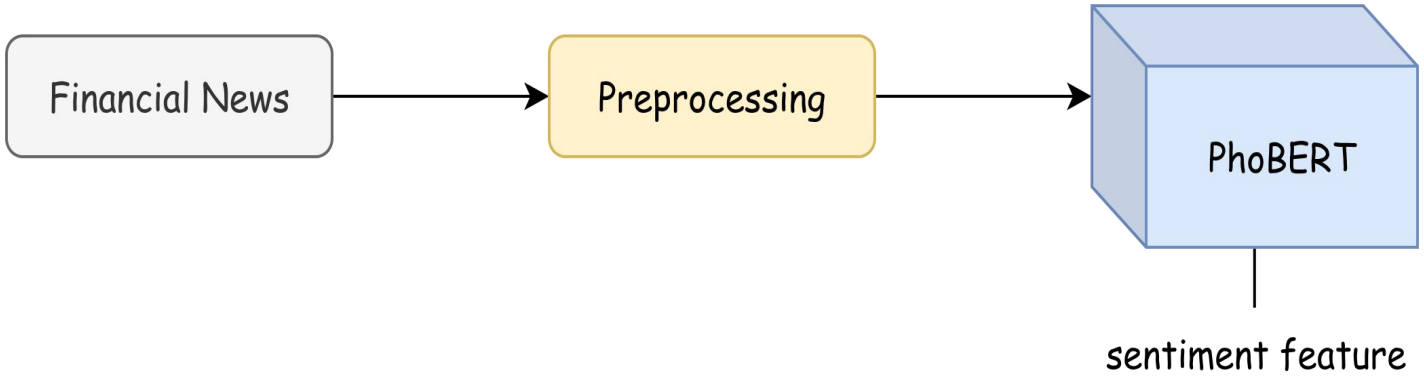
$$\alpha = R_p - R_f - \beta \times (R_m - R_f)$$

    - R_p: tỉ suất sinh lời thực tế.
    - R_m: tỉ suất sinh lời thị trường.
    - R_f: tỉ suất sinh lời phi rủi ro.
    - β: hồi quy tuyến tính lợi nhuận cổ phiếu & lợi nhuận thị trường.

3. Phương pháp luận - Gán nhãn dữ liệu (5)

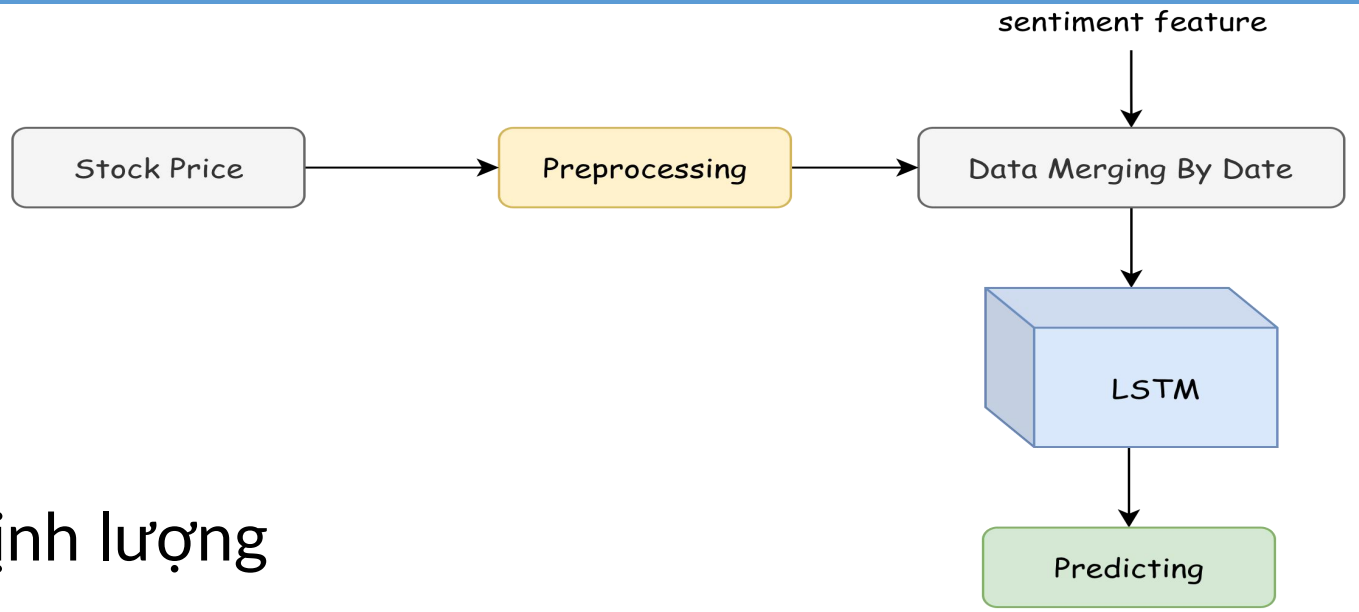
- **Các quy tắc gán nhãn:**
  - **LF3** - Dựa trên các từ khóa tài chính
    - “lợi nhuận”
    - “tăng trưởng”
    - “thua lỗ”
    - “khả quan”
    - . . .

4. Kiến trúc tổng thể của hệ thống - QTXLDL (1)



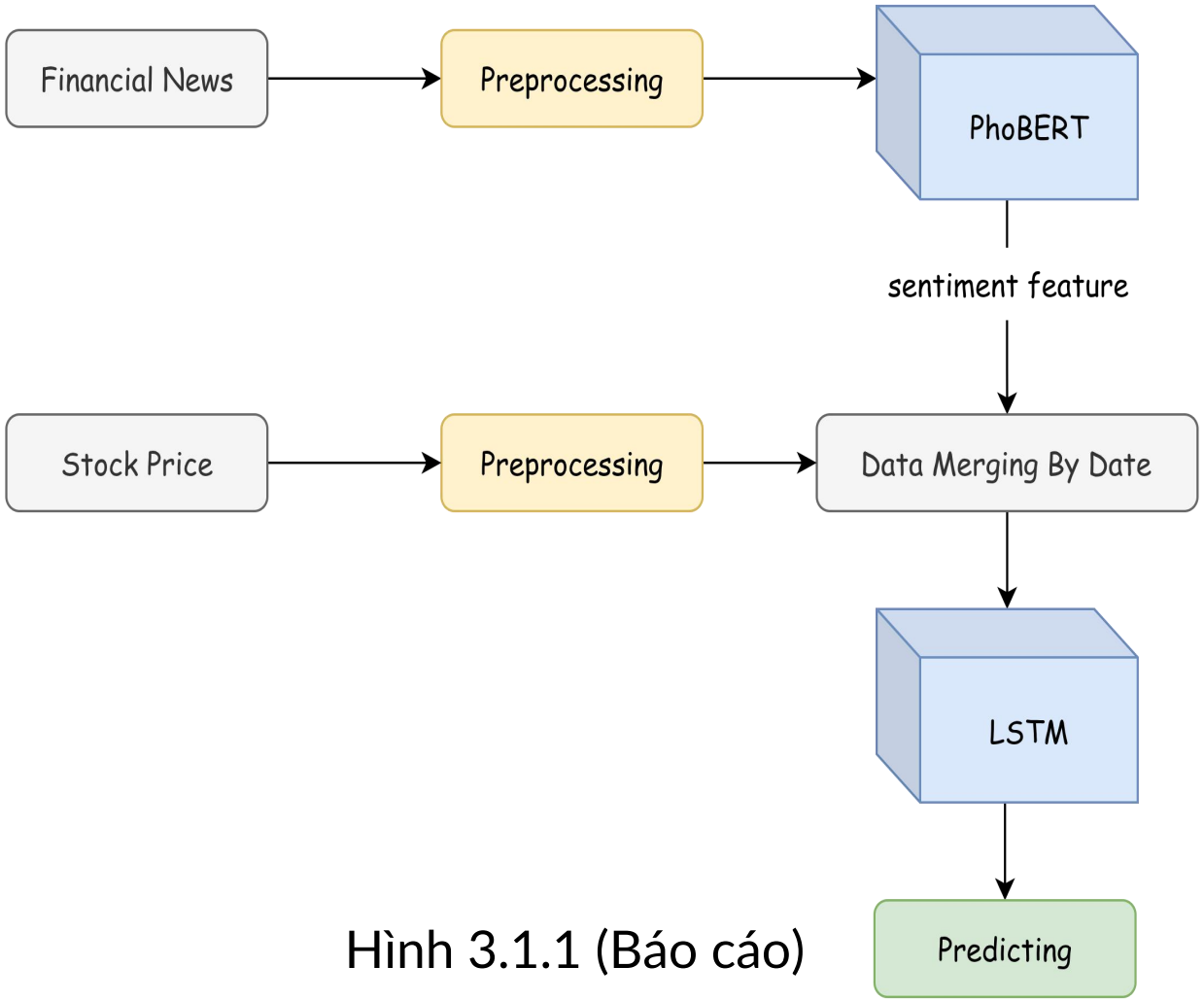
- Dữ liệu định tính
  - Thu thập các báo cáo tài chính, nhận định chuyên gia, v.v.
  - Nguồn: CafeF, VnExpress, Báo Đầu Tư.
  - Qua các bước xử lý và gán nhãn đã sẵn sàng để huấn luyện.

4. Kiến trúc tổng thể của hệ thống - QTXLDL (2)



- Dữ liệu định lượng
  - Thu thập giá cổ phiếu thông qua VnStock.
  - Qua các bước xử lý, kết hợp với đặc trưng cảm xúc và gán nhãn đã sẵn sàng để huấn luyện.

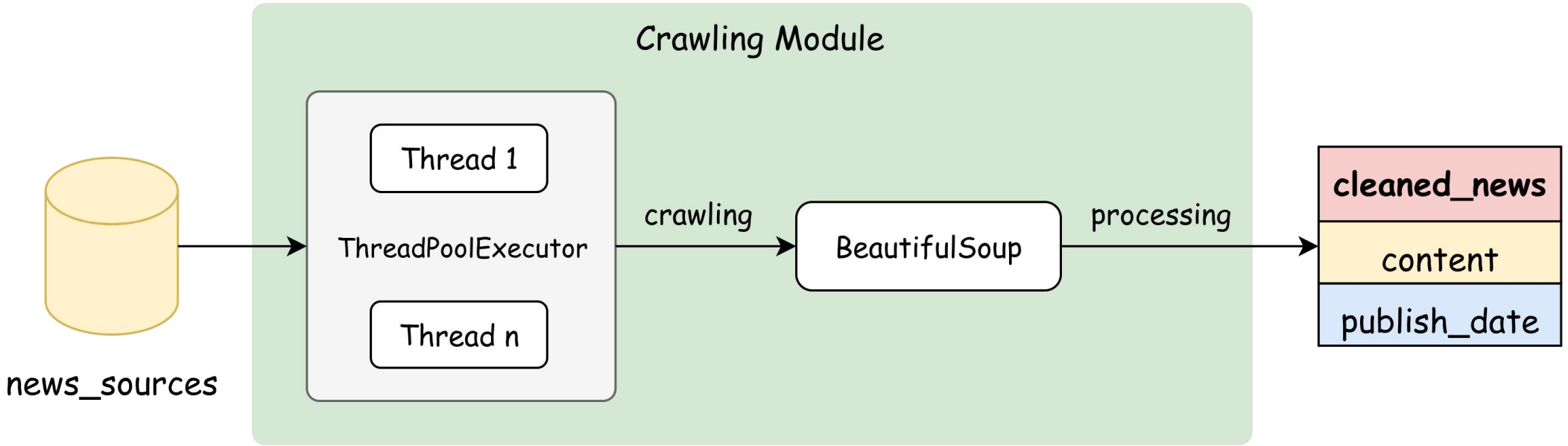
4. Kiến trúc tổng thể của hệ thống - QTXLDL (3)



Hình 3.1.1 (Báo cáo)

4. KTTTCHT- Quy trình xử lý dữ liệu (4)

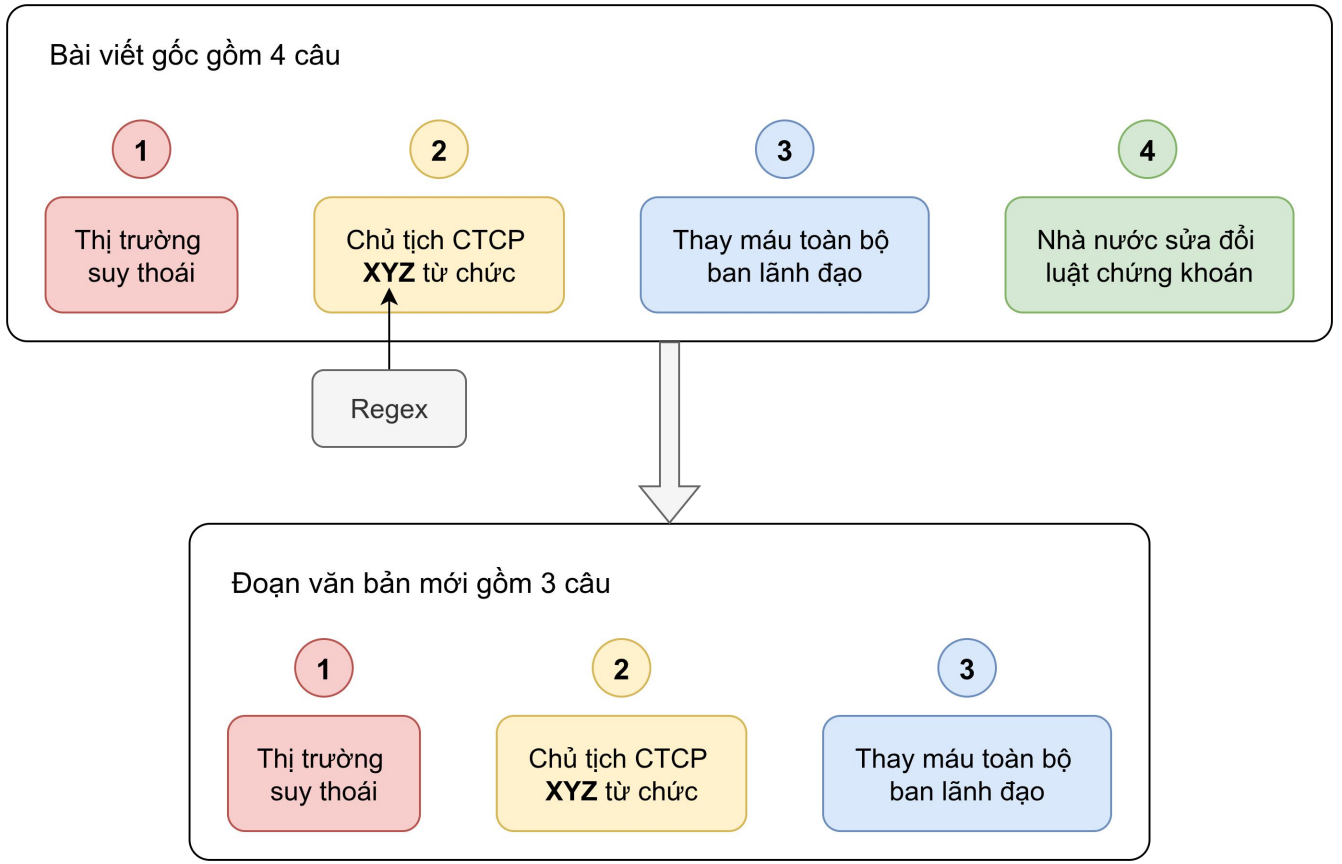
- Giai đoạn thu thập và tiền xử lý dữ liệu

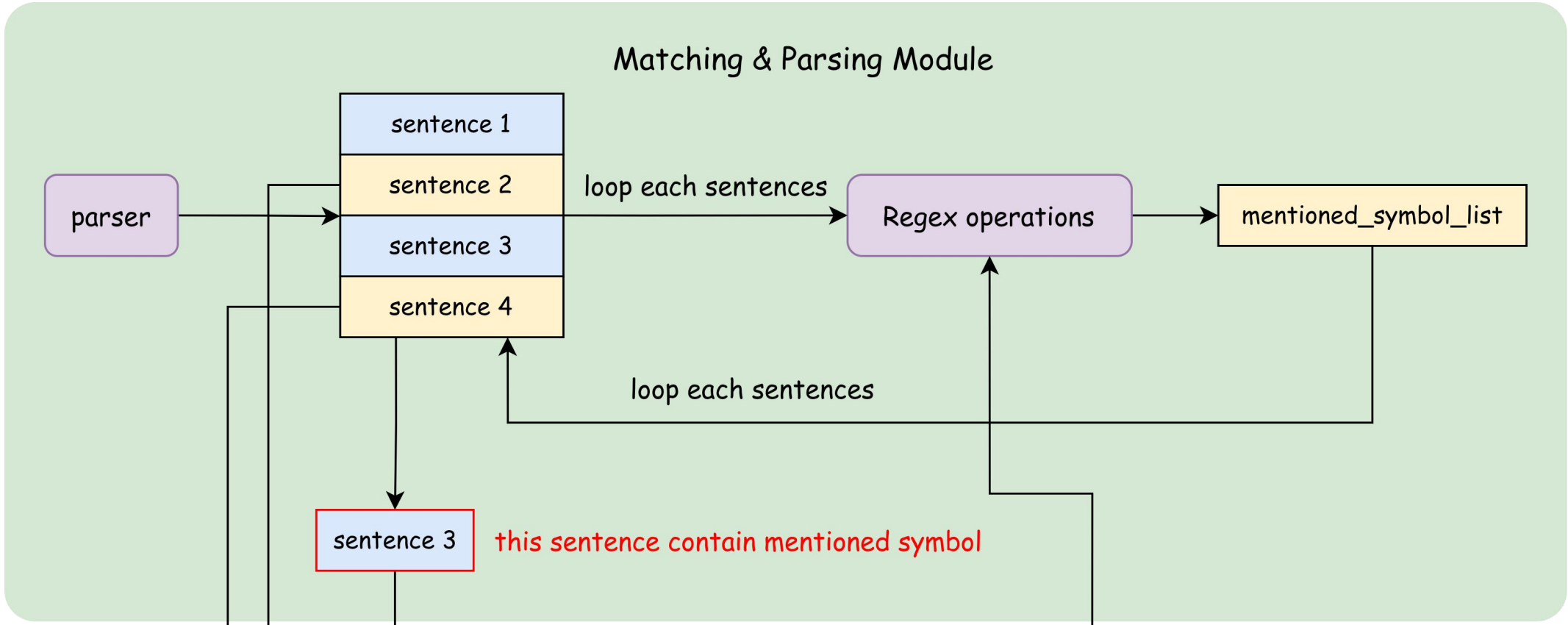


Hình 3.2.1 (Báo cáo)

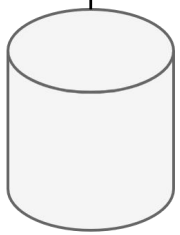
4. KTTTCHT- Quy trình xử lý dữ liệu (5)

- Giai đoạn khớp mã và trích xuất ngữ cảnh





parsed_paragraph		
sentence 2	publish_date	mentioned_symbol
sentence 3	publish_date	mentioned_symbol
sentence 4	publish_date	mentioned_symbol



vn30_symbols_list

5. Kết quả thực nghiệm

- content

6. Kết luận và đề xuất hướng nghiên cứu tiếp theo

- content