

Tài liệu Thiết kế Kỹ thuật: Module Phân tích Tin tức (Quick Check)

Dự án: ITAPIA

Module: News Analysis (Quick Check)

Ngày: 24/07/2025

1. Mục tiêu và Triết lý Thiết kế

Mục tiêu của module này là xử lý các tin tức tài chính (tiêu đề và tóm tắt) một cách nhanh chóng để **trích xuất các tín hiệu có cấu trúc, có thể hành động được (actionable signals)**. Các tín hiệu này sẽ phục vụ làm đầu vào cho Decision Maker Agent.

Triết lý thiết kế tuân thủ các nguyên tắc của quy trình "Quick Check":

- Tốc độ:** Toàn bộ quá trình phân tích phải diễn ra nhanh chóng trên hạ tầng CPU.
- "Đủ tốt":** Cung cấp các tín hiệu có giá trị cao mà không cần đến sự hiểu biết sâu của các mô hình LLM lớn.
- Giải thích được:** Output phải có cấu trúc rõ ràng, cung cấp "bằng chứng" cho các kết luận của nó, duy trì triết lý cốt lõi của ITAPIA.

2. Kiến trúc và Công nghệ Dự kiến

Module sẽ bao gồm 4 chức năng chính, mỗi chức năng sử dụng các công nghệ đã được lựa chọn cẩn thận để cân bằng giữa hiệu suất và độ chính xác.

Chức năng	Công cụ/Công nghệ	Lý do Lựa chọn
1. Phân tích Tình cảm	AnkitAI/distilbert-base-uncased-financial-news-sentiment-analysis (Mô hình Transformer)	Đã được fine-tune sẵn trên dữ liệu tài chính, cho độ chính xác cao về ngữ nghĩa. Tận dụng mô hình state-of-the-art mà không cần tự huấn luyện. Mức tiêu thụ tài nguyên (~600-800MB RAM) là chấp nhận được.
2. Nhận dạng Thực thể (NER)	spaCy (với model en_core_web_sm/md)	Cực kỳ tối ưu về tốc độ trên CPU. Cung cấp khả năng nhận dạng các thực thể phổ biến (ORG, PERSON, MONEY...) một cách đáng tin cậy.
3. Đánh giá Tác động	Quy tắc dựa trên Từ khóa (Keyword-based Rules)	Cách tiếp cận nhanh nhất và hiệu quả nhất. Dễ dàng định nghĩa và mở rộng các quy tắc để xác định mức độ quan trọng của tin tức dựa trên sự hiện diện của các thuật ngữ tài chính cốt lõi (ví dụ: "lợi nhuận", "mua lại", "FDA").
4. Trích xuất "Bằng chứng"	Nêu bật Từ khóa Tình cảm (Sentiment Keyword Highlighting)	Thay thế cho SHAP (vốn quá chậm cho Quick Check). Sử dụng một từ điển tình cảm tài chính (ví dụ: Loughran-McDonald) để tìm và trích xuất các từ/cụm từ có sắc thái mạnh nhất, cung cấp một lời giải thích "đủ tốt" và tức thời.

3. Luồng Xử lý và Cấu trúc Output

Đầu vào: Một bản tin (tiêu đề + tóm tắt).

Luồng xử lý:

- Bản tin được đưa vào FinBERT để tính toán điểm số và nhãn tình cảm (Positive, Negative, Neutral).
- Đồng thời, bản tin được đưa vào spaCy để trích xuất một danh sách các thực thể có tên.
- Bản tin được quét qua một bộ quy tắc từ khóa để xác định mức độ tác động (High, Medium, Low).
- Bản tin được quét qua từ điển tình cảm để tìm ra các từ khóa "bằng chứng" mạnh nhất.

Đầu ra (Output): Kết quả của 4 chức năng trên sẽ được tổng hợp thành một đối tượng JSON (được định nghĩa bởi Pydantic schema) cho mỗi bản tin. Đây là "hợp đồng dữ liệu" mà module này cung cấp cho phần còn lại của hệ thống.

Ví dụ về cấu trúc Output cho một bản tin:

```
{
  "news_uuid": "abc-123-def-456",
  "sentiment": {
    "score": 0.85,
    "label": "Positive"
  },
  "key_entities": [
    { "text": "Apple", "type": "ORG" },
    { "text": "iPhone 16", "type": "PRODUCT" }
  ],
  "impact_assessment": {
    "score": 0.9,
    "label": "High"
  },
  "evidence": {
    "method": "Keyword Highlighting",
    "positive_keywords": [
      { "word": "vượt kỳ vọng", "weight": 4},
      { "word": "doanh số kỷ lục", "weight": 4}
    ],
    "negative_keywords": []
  }
}
```

4. Tích hợp và Tầm nhìn

- **Aggregation:** Một sub-module tổng hợp sẽ nhận vào một danh sách các output trên (cho nhiều tin tức) và tạo ra một báo cáo tóm tắt (NewsSentimentSummary) cho một ticker hoặc sector.
 - **Deep Dive:** Các kỹ thuật phức tạp và tốn kém hơn như **SHAP Explainer** cho **FinBERT** sẽ được dành riêng cho quy trình "Deep Dive", nơi người dùng có thể yêu cầu một phân tích sâu, chi tiết ở cấp độ từ ngữ cho một bản tin cụ thể.
-