Data Engineering Project Final Presentation

뉴스데이터를 활용한 LLM 기반 포트폴리오 최적화



6조

박장환 | 윤원상 | 장미현 | 정유진 | 최인주

Contents

01 Problem Statement

02 Architecture

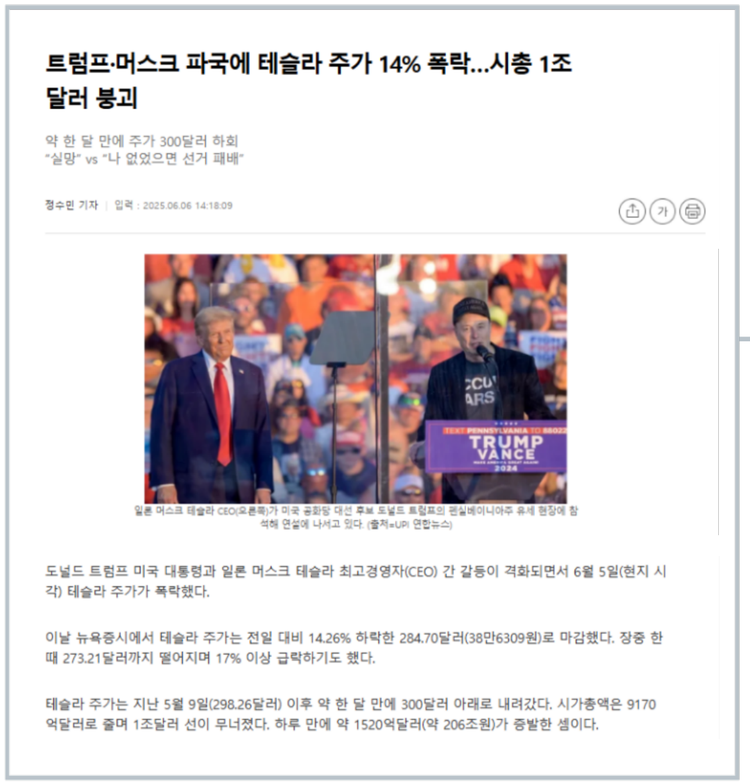
03 Data

04 Method

05 Result

06 Future Work

01. Problem Statement



트럼프·머스크 파국에 테슬라 주가 14% 폭락...시총 1조 달러 붕괴

약 한 달 만에 주가 300달러 하회
"실망" vs "나 없었으면 선거 패배"

정수민 기자 | 입력 : 2025.06.06 14:18:09

일론 머스크 테슬라 CEO(오른쪽)가 미국 공화당 대선 후보 도널드 트럼프의 펜실베이니아주 유세 현장에 참석해 연설에 나서고 있다. (출처=UPI 연합뉴스)

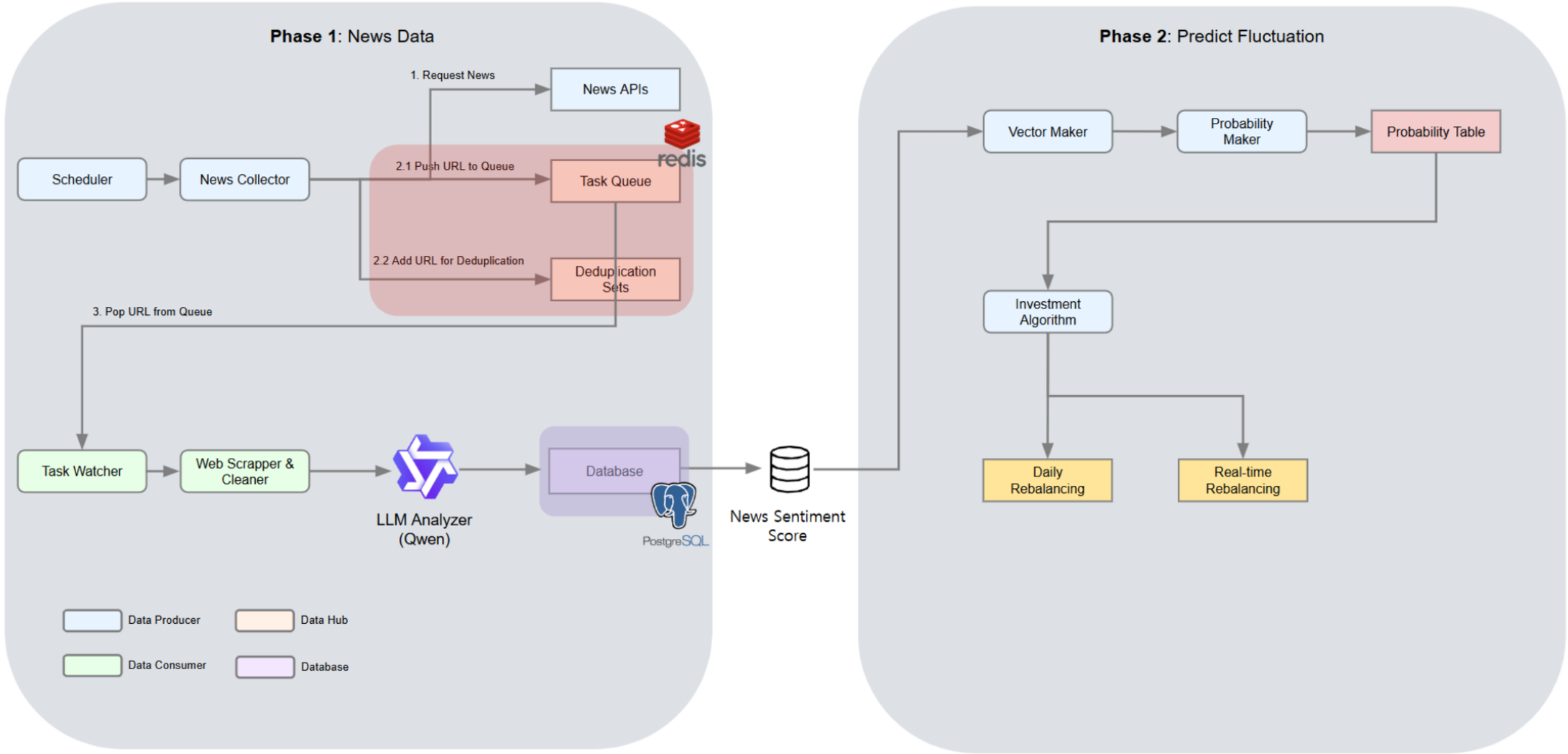
도널드 트럼프 미국 대통령과 일론 머스크 테슬라 최고경영자(CEO) 간 갈등이 격화되면서 6월 5일(현지 시각) 테슬라 주가가 폭락했다.

이날 뉴욕증시에서 테슬라 주가는 전일 대비 14.26% 하락한 284.70달러(38만6309원)로 마감했다. 장중 한때 273.21달러까지 떨어지며 17% 이상 급락하기도 했다.

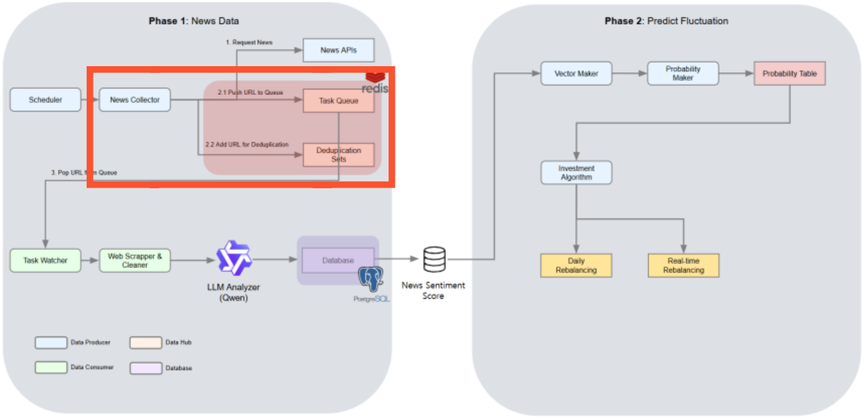
테슬라 주가는 지난 5월 9일(298.26달러) 이후 약 한 달 만에 300달러 아래로 내려갔다. 시가총액은 9170억달러로 줄며 1조달러 선이 무너졌다. 하루 만에 약 1520억달러(약 206조원)가 증발한 셈이다.

PROBLEM		SOLUTION
Information Gap 비정형 데이터의 복잡성 및 비가치화	›	**Contextual** 뉴스 맥락을 이해하는 LLM 기반 정량화
Time-Zone Gap 한국-미국 간 시차에 대한 한국 투자자의 대응 시간 한계	›	**Real-time** 24시간 시장 감시 및 자동화된 포트폴리오 최적화

02. Architecture



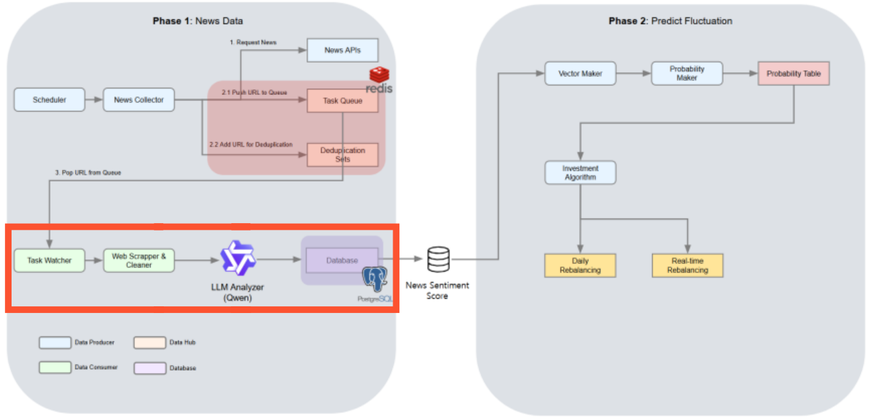
02. Architecture



Step 1 Collector & Queueing Strategy

- RSS 피드 주기적 모니터링
  - Deduplication
    - Redis Set (`yahoo : processed_urls`) 활용
    - O(1)의 빠른 조회 속도로 먼저 수집된 URL 필터링
      Redis Set의 해시 테이블 구조 사용
  - Task Scheduling
    - 새로운 URL만 Redis List (`yahoo : task_queue`)에 push

02. Architecture

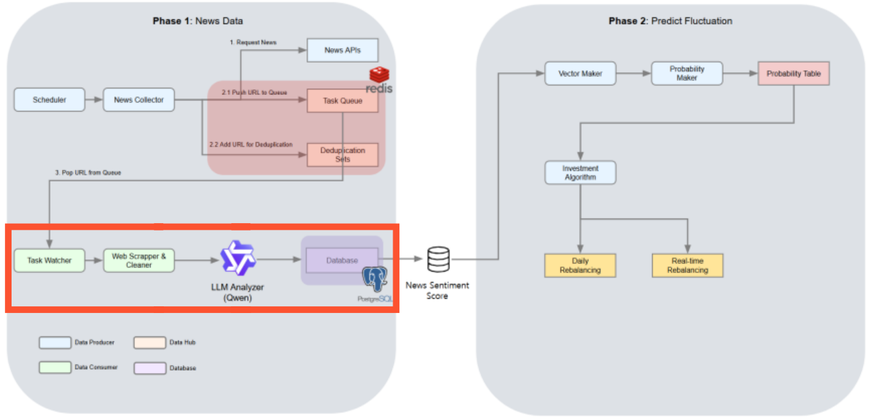


Step 2 Worker & Persistence

- Worker Mechanism:
  - BLPOP을 통한 Blocking 방식의 대기 (불필요한 Polling 방지)
  - 24시간 상시 가동 및 본문 추출

특징	Polling	BLPOP (Redis Blocking)
대기 방식	Busy Wait (주기적 확인)	Blocking (이벤트 대기)
반응 속도	Sleep 주기에 따라 지연 발생	즉시 반응 (Real-time)
네트워크	불필요한 트래픽 발생	필요한 트래픽만 발생
적합한 상황	주기가 매우 긴 배치 작업	뉴스 피드, 채팅, 실시간 처리

02. Architecture

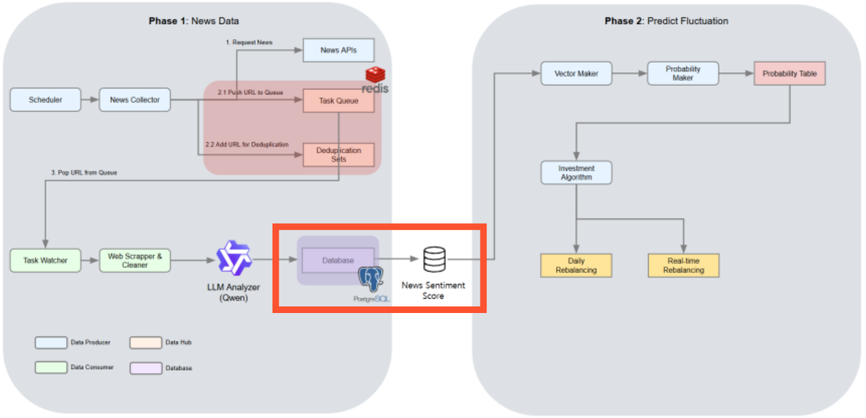


Step 2 Worker & Persistence

- Scalability (확장성):
  - Redis 큐 구조 덕분에 Worker 프로세스만 늘리면 즉시 처리량 증대 가능

- Storage Strategy (PostgreSQL):
  - Phase 1: 원문(Raw Text) + 메타데이터 저장
  - Phase 2: LLM 분석 결과 별도 업데이트 (Decoupled Analysis)

02. Architecture



Step 3 **Feed data into the model**

- 미국시간 기준 전날 16:00 ~ 당일 9:30 뉴스데이터 수집
  - LLM이 이를 분석하여 sentiment score로 변환

- 한국 시간 기준 우리가 자는 동안 모든 시사 이슈를 캐치할 수 없어, 이를 매일 아침 반영하여 매수 종목 포착

03. Data

카테고리	활용데이터	구성	범위	출처
주가 데이터	과거 주가데이터	시가, 종가	최근 2년	Massive
	실시간 주가데이터	시가, 종가	실시간	Yahoo Finance
뉴스 데이터	과거 뉴스데이터	기사 제목, 링크, 발행 시간	최근 2년 (44,179개)	Massive
	실시간 뉴스데이터	기사 제목, 전문, 발행 시간	실시간	Yahoo Finance

04. Method

Step 1 LLM을 활용한 감성분석

Massive API → 뉴스 데이터 추출 → 기사 제목, 링크, 발행시간 → 크롤링 → 기사 제목, 전문, 발행시간

프롬프트 생성

```
너는 금융 분석 전문가야.
주어진 [Nasdaq 100 기업 리스트]와 [뉴스 기사]를 분석하여,
기사가 해당 기업의 주가에 미칠 영향을 정밀하게 판단해줘.
...
  [뉴스 기사]:
  {How the US got left behind
   in the global electric car race ... }
...
```

LLM 응답 추출

```
{
  "ticker_symbol": "TSLA",
  "sentiment_classification": -1,
  "sentiment_score": -0.65,
  "relevance_score": 0.9
}
```

04. Method

Step 2

데이터 정제

- `ticker_symbol`이 NONE 즉, 특정 기업에 매칭되지 않은 뉴스 데이터 제거

published_utc	title	llm_ticker_symbol	llm_sentiment_classification	llm_sentiment_score	llm_relevance_score	full_content
2024-04-23 12:15:00	1 No-Brainer Va	NVDA	1	0.85	0.95	This Vanguard growth fund is home to som
2024-02-27 10:29:00	1 Overlooked Ai	NONE	0	0	0	DigitalOcean helps thousands of small busi
2024-07-26 07:12:00	1 Phenomenal S	ORCL	1	0.85	0.95	A long track record of cloud and informatic
2024-07-07 09:01:00	1 Phenomenal S	SMCI	1	0.85	0.95	This artificial intelligence (AI) hardware spe
2024-04-22 18:05:00	1 Phenomenal S	GOOGL	-1	-0.65	0.95	Alphabet's stock is back to setting new high
2024-04-27 15:15:00	1 Powerhouse G	NONE	0	0	0	The right investment can supercharge your
2024-11-09 10:35:00	1 Soaring Vangu	NVDA	1	0.85	0.95	Is the Vanguard Information Technology ET
2024-06-17 10:00:00	1 Stock I Would	UPST	-1	-0.85	0.95	There are obvious reasons investors should
2023-12-28 12:19:00	1 Stock That Co	MELI	1	0.85	0.95	Even if it doesn't pass Amazon, this stock co
2024-04-16 10:15:00	1 Stock That Tur	AMZN	1	0.85	0.98	By constantly innovating and expanding int
2024-01-28 11:03:00	1 Stock to Buy, '	MSFT	1	0.85	0.95	Fed meeting, January jobs report, and heavy
2024-05-27 08:25:00	1 Stock-Split Art	NVDA	1	0.95	0.98	Nvidia shares have increased 11-fold in four
2024-09-19 08:47:00	1 Stock-Split Sto	AVGO	1	0.85	0.95	This company's artificial intelligence presen
2024-01-03 10:06:00	1 Stock-Split Sto	GOOGL	1	0.85	0.95	Among Nvidia, Amazon, DexCom, Shopify, A
2023-12-04 10:21:00	1 Stock-Split Sto	GOOGL	1	0.85	0.95	Among Nvidia, Amazon, DexCom, Shopify, A
2023-12-13 12:03:00	1 Stunning Stoc	AAPL	1	0.85	0.95	The stock market is great at creating long-t
2024-10-03 09:55:00	1 Supercharged	NVDA	1	0.85	0.95	The Vanguard Information Technology ETF'
2024-10-29 09:45:00	1 Supercharged	NONE	0	0	0	The Vanguard Mega Cap Growth Index Fur
2024-06-25 09:12:00	1 Supercharged	NVDA	1	0.85	0.95	The Vanguard S&P 500 ETF includes many

04. Method

Step 3 **감성분석 결과 벡터화**

Sentiment Score
(an article of AAPL)



column of AAPL





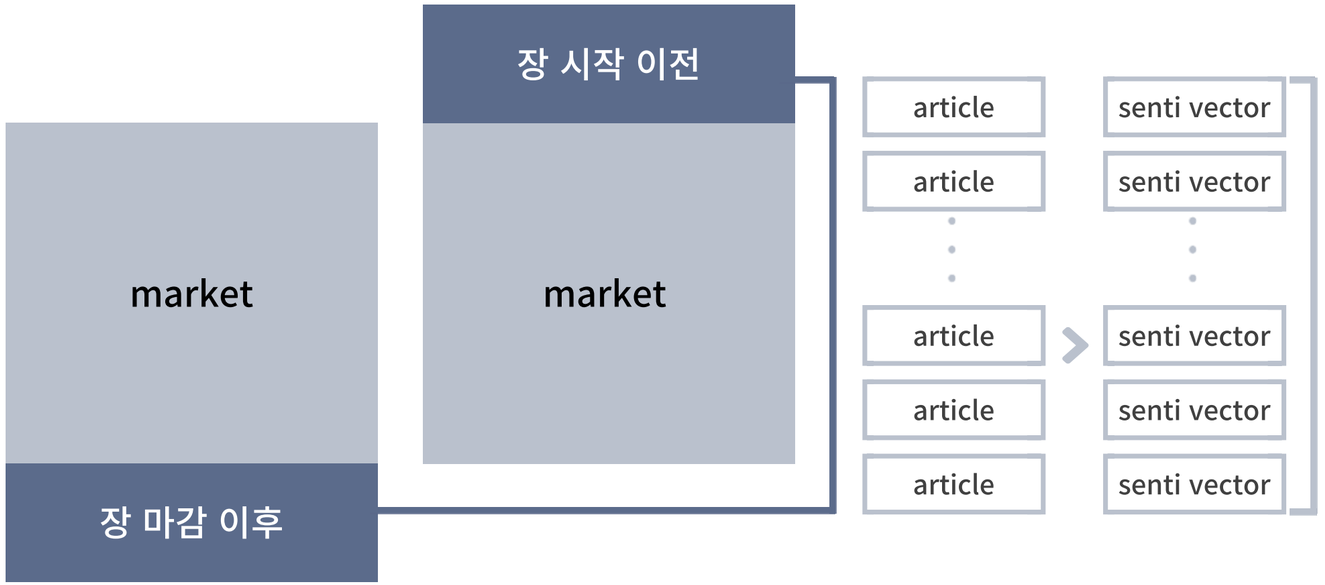
Sentiment Vector
(an article of AAPL)

1	corr(NVDA, AAPL)	corr(NVDA, MSFT)	...	corr(NVDA, GFS)
corr(AAPL, NVDA)	1	corr(AAPL, MSFT)	...	corr(AAPL, GFS)
...	...	1	...	...
corr(GFS, NVDA)	corr(GFS, AAPL)	corr(GFS, MSFT)	...	1

과거 30일 수익률 간 corr matrix

04. Method

Step 4 수익률 예측



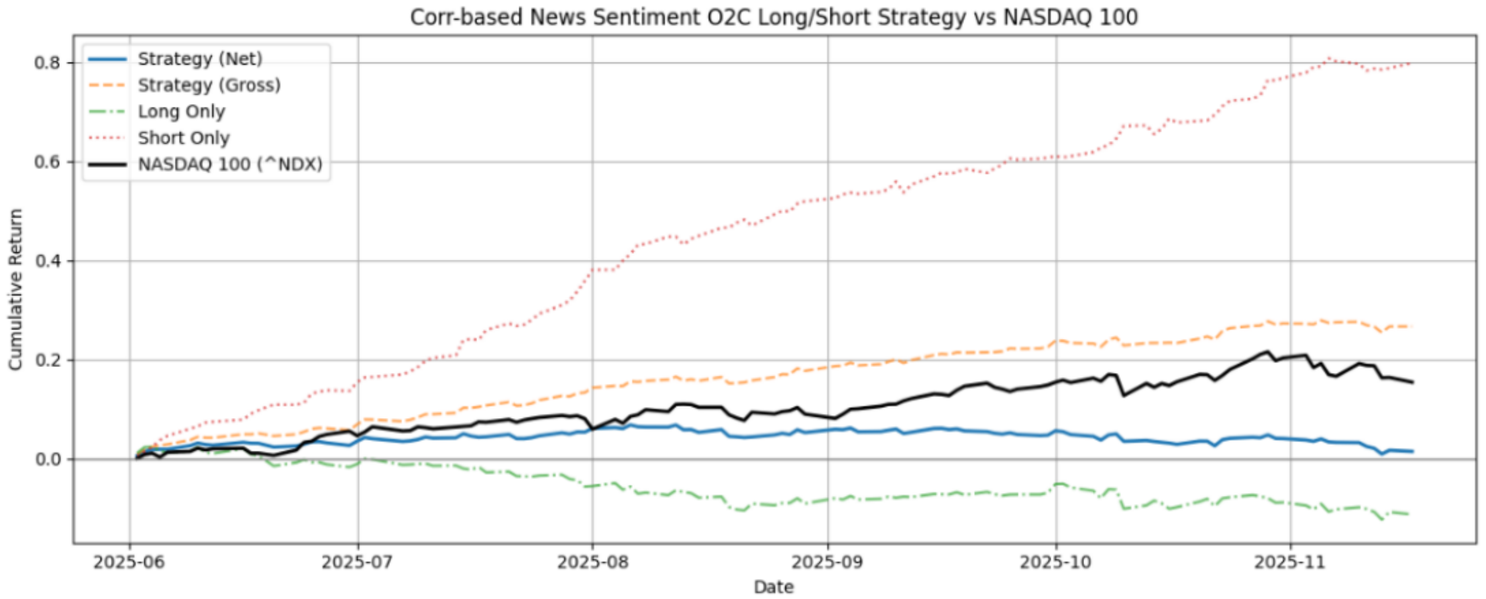
Mean of Sentiment Vector

Gradient Boosting Regressor (GBR)
: 예측 오류를 최소화하기 위해 모델을 순차적으로 구축하는 트리 기반 앙상블 방법

Open to Close 수익률 예측

05. Result

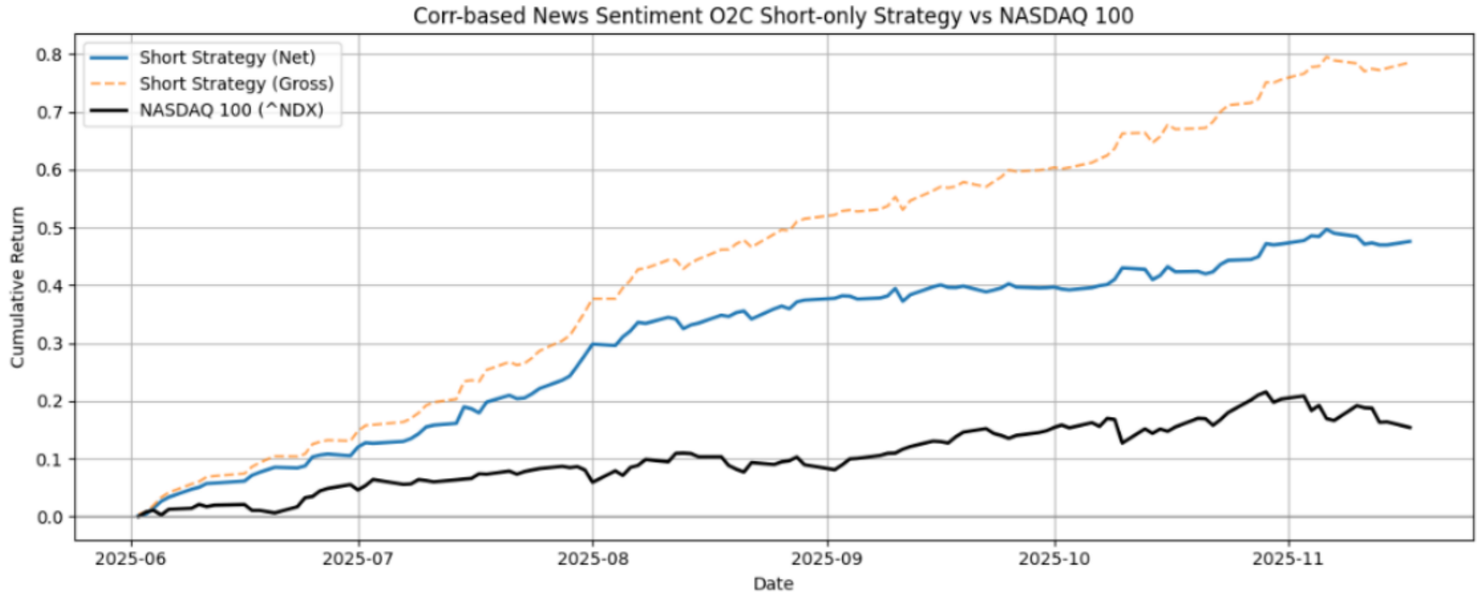
Transaction Cost = 0.25% (per every trade)
Strategy: Long Top 25%, Short Bottom 25%



- Asymmetric effect: Market(NASDAQ)은 긍정적 뉴스보다 부정적 뉴스에 더 크게 반응
- 공매도를 통한 수익 실현 가능

05. Result

Transaction Cost = 0.25% (per every trade)
Strategy: Short Bottom 25%



- Cumulative Return = 47.57%
- Turnover = 65.3%

05. Result

For example :

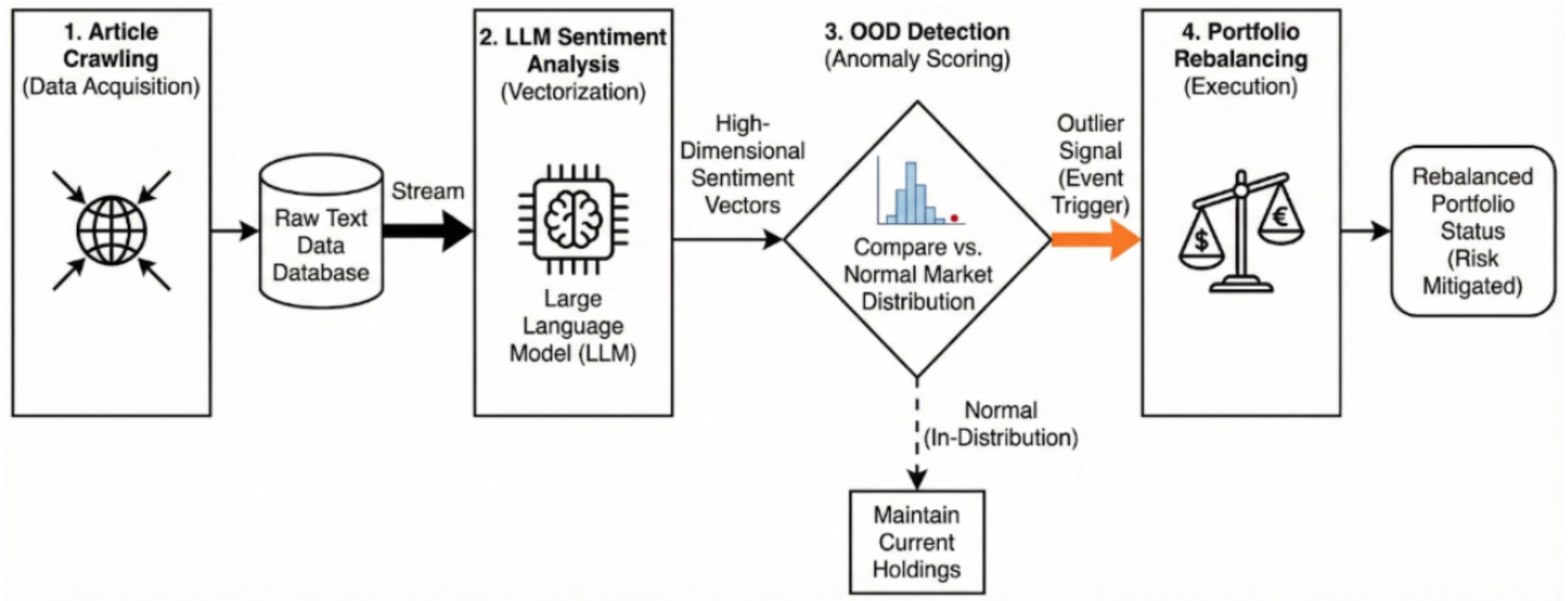
📋 마지막 거래일 숏 포지션 상세:

	trade_date	ticker	sentiment_factor	pred_ret
0	2025-11-17	MDLZ	-0.355475	-0.031477
1	2025-11-17	CCEP	-0.267741	-0.029247
2	2025-11-17	VRSK	-0.311077	-0.028933
3	2025-11-17	PEP	-0.281902	-0.028741
4	2025-11-17	ORLY	-0.307486	-0.028241
5	2025-11-17	TMUS	-0.252102	-0.026639
6	2025-11-17	BIIB	-0.154687	-0.026151
7	2025-11-17	ADP	-0.202133	-0.025908
8	2025-11-17	LIN	-0.193858	-0.024464
9	2025-11-17	MNST	-0.212565	-0.024402
10	2025-11-17	PAYX	-0.179425	-0.023687
11	2025-11-17	GILD	-0.182986	-0.023687
12	2025-11-17	EXC	-0.158306	-0.022312
13	2025-11-17	KHC	-0.144538	-0.021423
14	2025-11-17	MAR	-0.144755	-0.021423
15	2025-11-17	CTAS	-0.146744	-0.021423
16	2025-11-17	XEL	-0.237757	-0.020590
17	2025-11-17	COST	-0.080372	-0.019165
18	2025-11-17	CMCSA	-0.132018	-0.017888
19	2025-11-17	CSGP	-0.062652	-0.017806
20	2025-11-17	AZN	-0.114325	-0.016805
21	2025-11-17	FAST	-0.119936	-0.016805
22	2025-11-17	AMGN	-0.046298	-0.016336
23	2025-11-17	CHTR	-0.075867	-0.015754
24	2025-11-17	TRI	-0.040504	-0.015464

06. Future Work

실시간 리밸런싱 구현 → 단기 충격에 대한 전조 증상을 찾아내 시장 충격에 미리 대응

HOW?



Data Engineering Project Final Presentation

THANK YOU

6조

박장환 | 윤원상 | 장미현 | 정유진 | 최인주