

# FIND-A 12기 파이널 컨퍼런스

**목차**

**1. 배경**

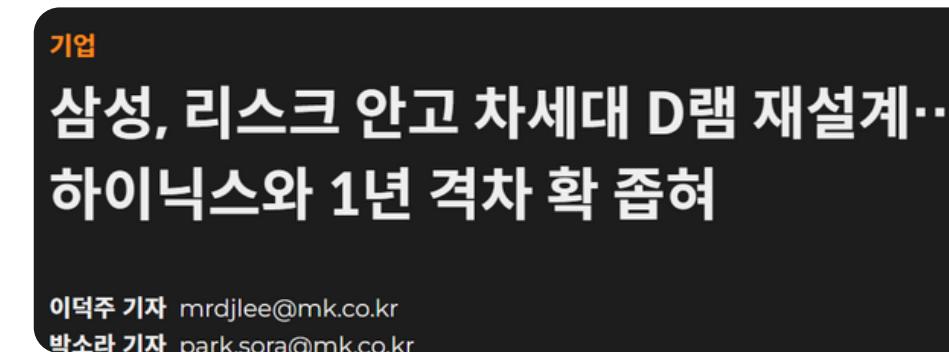
**2. 문제 정의**

**3. 방법론**

**4. 결과**

**5. 한계 및 발전**

## [뉴스 데이터 - 주가 관계]



삼성전자  
005930 • KOSPI  
**110,500 원 ▲4,200 (3.95%)**  
NXT After Market 110,700 ▲4,400 (4.14%) • 장마감 ⓘ

- 금융 시장에서는 뉴스가 주가에 영향을 준다는 직관은 명확함
- 그러나 실제 분석에서는 다음 문제가 존재
  - 하나의 뉴스가 여러 기업·여러 이벤트를 동시에 포함
  - “이 뉴스가 정확히 어떤 종목의 어떤 요인에 영향을 줬는지” 명시하기 어려움

따라서, 뉴스-주가 관계는 분명하지만, 뉴스가 너무 복합적이라 ‘원인-결과’를 구조적으로 설명하기 어렵다.

## [Ontology(온톨로지)]

: 개체와 관계를 정의하여, 서로 다른 분류 체계를 가진 데이터를 '같은 의미 체계'로 묶어 지식 그래프로 나타내는 기술

### 여러 형태를 가진 데이터

- 뉴스: 자연어 텍스트
- 주가: 숫자 시계열
- 이벤트: 카테고리 / 라벨
- 거시지표: 숫자 시계열

→ 뉴스를 단순 '텍스트'가 아닌 (기업, 이벤트, 시점)으로 분해하여 이를 거시지표 데이터, 주가 데이터와 하나로 통합하는 '지식그래프'를 생성하여 뉴스-주가 관계를 분석한다.

## 배경

### [벤치마킹 - 토스증권 'AI 시그널']

: 뉴스·공시 데이터를 AI로 분석해 시장 변동의 원인을 요약 및 설명하는 서비스

이는

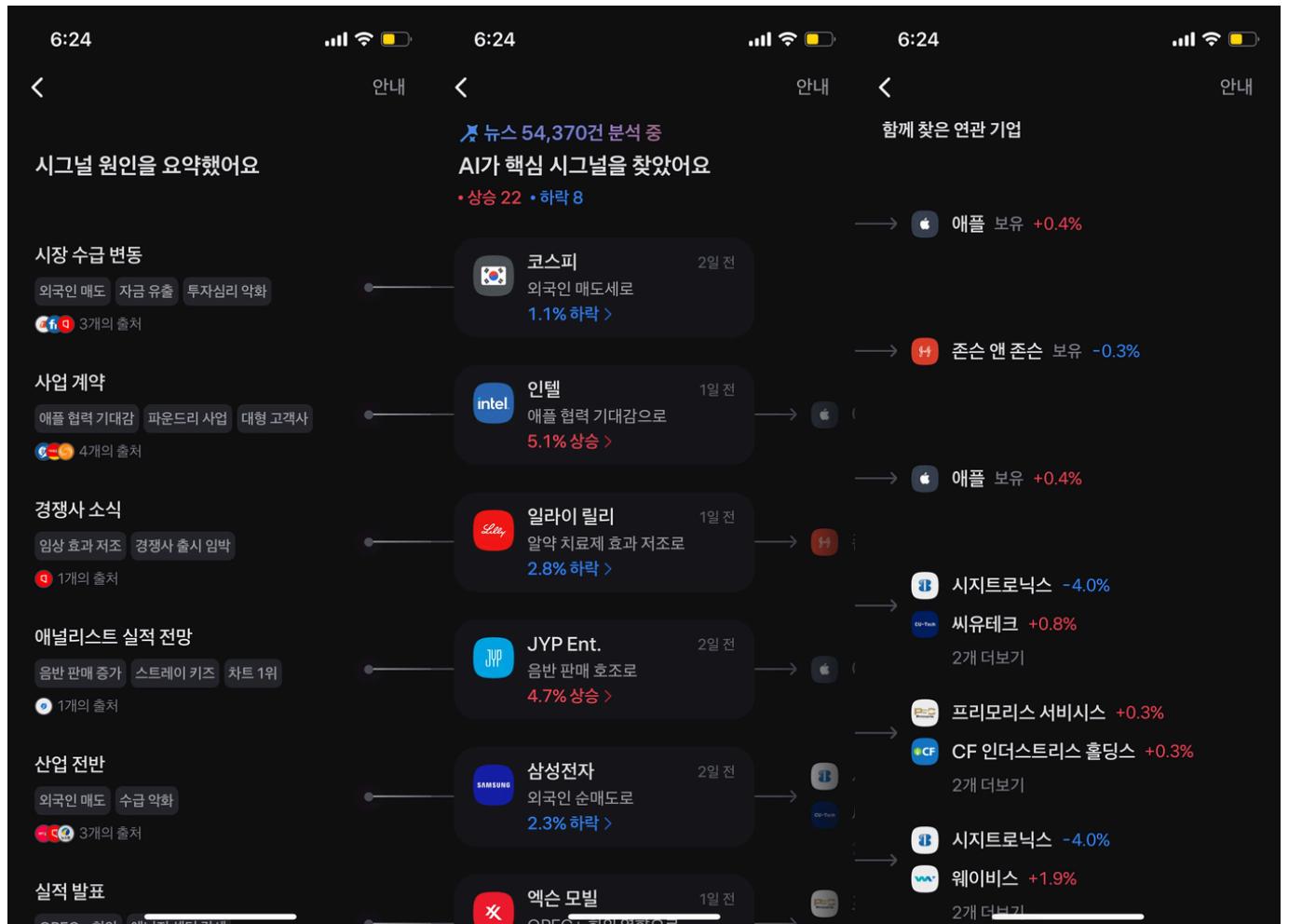
- 실무에서 뉴스와 주가를 단순히 보여주는 것을 넘어서 '**왜 움직였는지**'를 **이유를 설명하려는 시도**
- 개인 투자자에게 **정보 통합 + 원인 중심 해석**이 중요해졌음을 보여주는 사례

→ 투자자에게 중요한 것은, **정보 간의 관계와 원인에 대한 이해**이다

이를 온톨로지를 기반으로 더 구조적으로 다뤄보자 함

### AI 기반 시장 분석 서비스 'AI 시그널'

AI가 시장 변동의 이유를 읽어드려요



## 문제 정의

# [분석 플로우]

뉴스 텍스트



키워드 기반 이벤트 추출



주가 시계열과 회귀 분석

자연어의 뉴스데이터

뉴스별 7개의 event 여부 확인

빈도, event 기반 feature,  
거시지표를 통해 주가 변동 회귀분석



온톨로지 적재, 그래프 생성



자연어로 변환

Neo4j로 적재,  
NetworkX로 그래프 생성

뉴스 - 주가 변동 관계를  
자연어로 설명

## 문제 정의

# 뉴스 데이터 - 주가 관계

## 가정

- t 시점에 발행된 뉴스만 t-1와 t간의 주식의 종가의 변화에 영향을 미친다.
- 주식 종목마다 종가의 변동에 영향을 미치는 X변수가 다르다.

회귀식

$$r_t = \alpha + \beta_1 N_t + \beta_2 E1_t + \cdots + \beta_8 E7_t + \gamma_1 M1_t + \cdots + \gamma_K MK_t + \varepsilon_t$$

## Output

- 종목별 최적 회귀선을 찾고, 각 feature가 y(주식 종가 변화율)에 얼마나 영향을 미치는지 확인
- 종목별로 어떤 feature이 제일 유의한지 확인
- 사용자가 이해하기 쉽도록 자연어로 변환

## 방법론

# 데이터 수집 / 변환

3가지 이종 데이터셋의 표준화 및 병합

- 종목별 뉴스 데이터
- 거시지표 데이터
- 종목별 주가 데이터

\*분석 대상 종목 선정 기준은 **KOSPI 시가총액 상위 10위 종목**

최종 데이터셋:

final 주식 종목별 데이터셋 (회귀식 학습데이터셋)				
날짜	변수1	변수 2~8 (이벤트 E1 ~ E7 발생 여부)	거시지표들...	주가 (y)
YYYY-MM-DD	각 날짜 뉴스 개수	0 or 1	#	전날 대비 변화율

event별 키워드

E1_EARN	실적/가이던스	실적, 영업이익, 컨센서스, 매출영업이익률, 어닝, 원가
E2_ORDER	수주/계약	수주, 계약, 납품, 발주, 공급, 공급계약, 수주잔고, 납기, 발주
E3_POLICY	정책/규제	정책, 규제, 정부, 법안, 제도, 행정, 감독, 공공, 당국
E4_PRODUCT	제품/기술	출시, 개발, 기술, 공개, 상용화, 연구, 혁신, 특허
E5CAPEX	투자/증설	투자, 증설, 공장, 설비투자, 캐파, 신설, 이전, 생산라인
E6_MA	M&A/지분	인수, 합병, M&A, 지분, 피인수, 흡수, 분할
E7_RISK	사고/리스크	사고, 중단, 리콜, 결함, 안전사고, 생산중단, 회수, 품질이슈, 화재, 폭발, 벌금, 부도, 논란, 소송, 분쟁, 피해



## 방법론

# Feature Selection & Analysis

### 1) OLS Regression

- 유의미하지 않은 피쳐들
- 다중공선성 → 매크로 데이터 삭제

### 2) 회귀 → 분류 문제

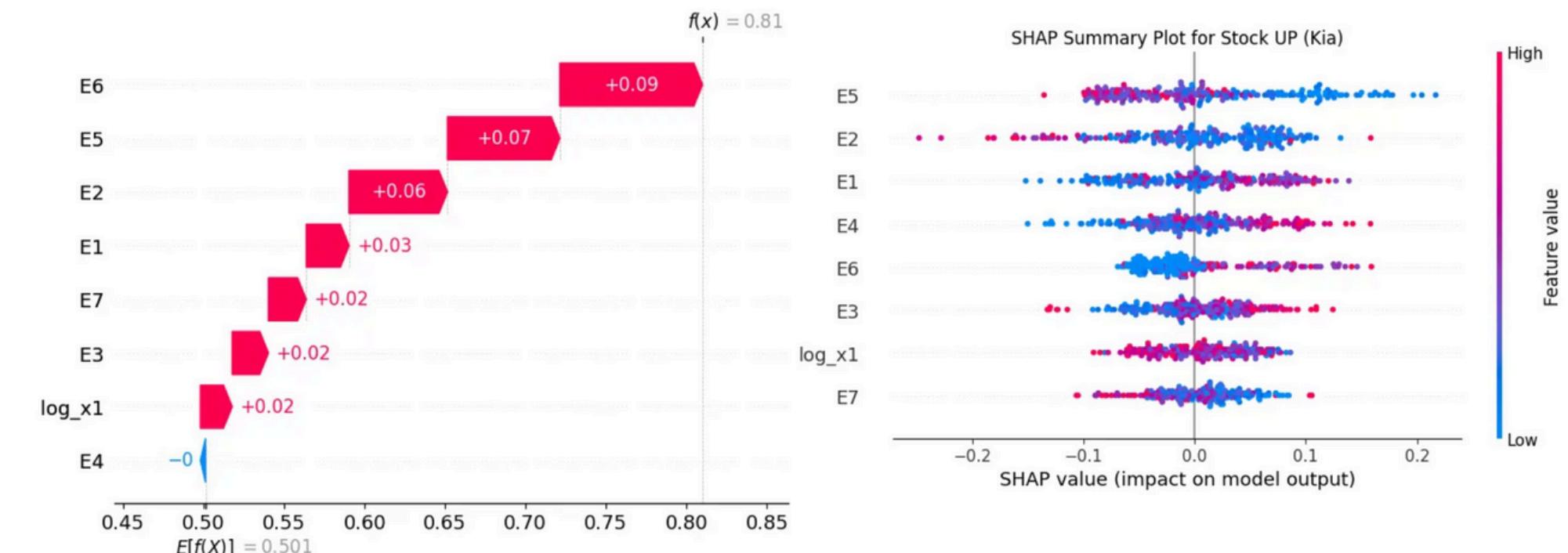
- y를 상승과 하락의 이진 분류로 변환
- 로지스틱 회귀와 결정 트리를 사용

### 3) SHAP

- 분류 문제가 성능이 회귀보다는 더 좋았으나 전반적으로 저조한 성능
- 설명력에 더 초점을 맞추기 위해 결정트리와 SHAP 지수를 최종적으로 사용

	Date	x1	E1	usd_krw	y
0	2024-12-10	42	0.047619	1432.60	-0.645161
1	2024-12-11	25	0.200000	1428.50	-1.818182
2	2024-12-12	43	0.069767	1429.47	7.010582
3	2024-12-13	29	0.000000	1434.82	-1.112485
4	2024-12-16	28	0.071429	1437.86	-0.500000

--- [Random Forest 모델 성능] ---					
Classification Report:					
	precision	recall	f1-score	support	
0	0.45	0.45	0.45	20	
1	0.59	0.59	0.59	27	
accuracy				0.53	47
macro avg	0.52	0.52	0.52	47	
weighted avg	0.53	0.53	0.53	47	

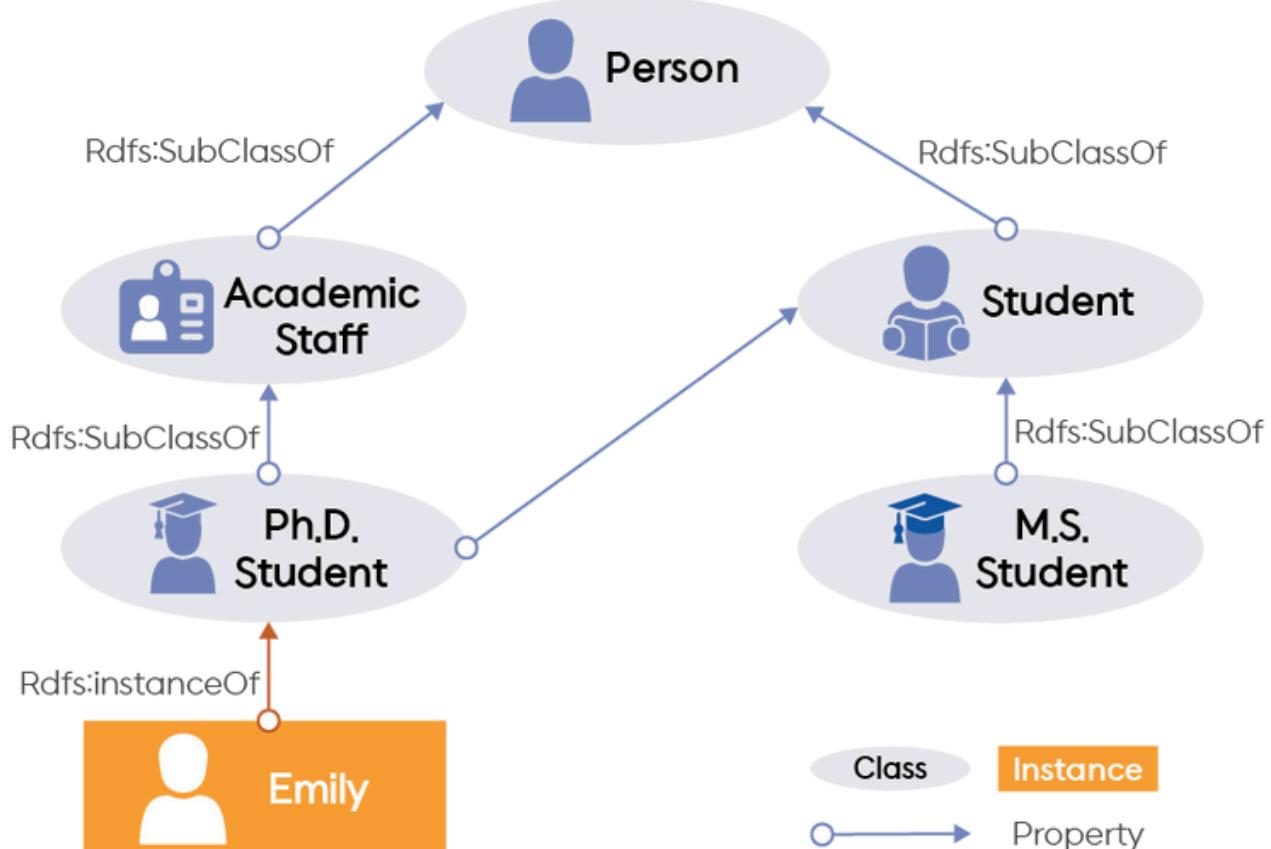


## 온톨로지 기반 데이터 구조화

단순히 데이터를 저장하는 것이 아닌,  
의미를 담아 저장하는 과정

→ 단순한 스키마가 아니라, 서로 다른 개념들이  
어떤 관계로 연결되는지를 명시적으로 표현하는 과정

각 데이터가 갖는 의미에 따라 개체를 나누고,  
개체 간 관계를 정의하여 구조화 진행



## 온톨로지 기반 데이터 구조화

“Neo4j” 프로그램을 사용

**Neo4j**는 Property Graph Model을 구현한 네이티브 그래프 데이터베이스

- 노드 / 관계 / 속성을 데이터 구조로 저장
- 관계가 단순한 JOIN 결과가 아닌 물리적으로 저장되는 객체
- 노드 - 관계 - 노드가 디스크게 직접 연결 구조로 저장되어 큰 용량의 데이터 크기에도 빠른 탐색 속도를 지님
- **Cypher Query**를 통해 원하는 형태의 조회 결과 커스터마이징 가능



Cypher Query 예시

```
MATCH (c:Company)-[:HAS_SHAP]->(s:Shap)-[:ON_DATE]->(d:Day)
WHERE d.date = '2025-01-03'
RETURN s.E1, s.E2;
```

- Neo4j의 AuraDB 클라우드 환경을 활용하여 적재

## 온톨로지 기반 데이터 구조화

### Company

SK Hynix

### Event

- E1(실적/가이던스)
  - SK하이닉스, HBM 매출 급증…연간 실적 가이던스 상향
- E2(수주/계약)
  - 엔비디아, HBM 공급 확대…SK하이닉스 수혜 기대
- E3(정책/규제)
  - 미·중 기술 규제 변수에 SK하이닉스 불확실성 지속
- E5(투자/증설)
  - SK하이닉스, 신규 생산라인 증설 검토

### Day

데이터 양, 주기성 등을 고려하여  
'2024.12.10 ~ '2025.12.08 사이 주식  
개장일 총 240일 데이터 사용

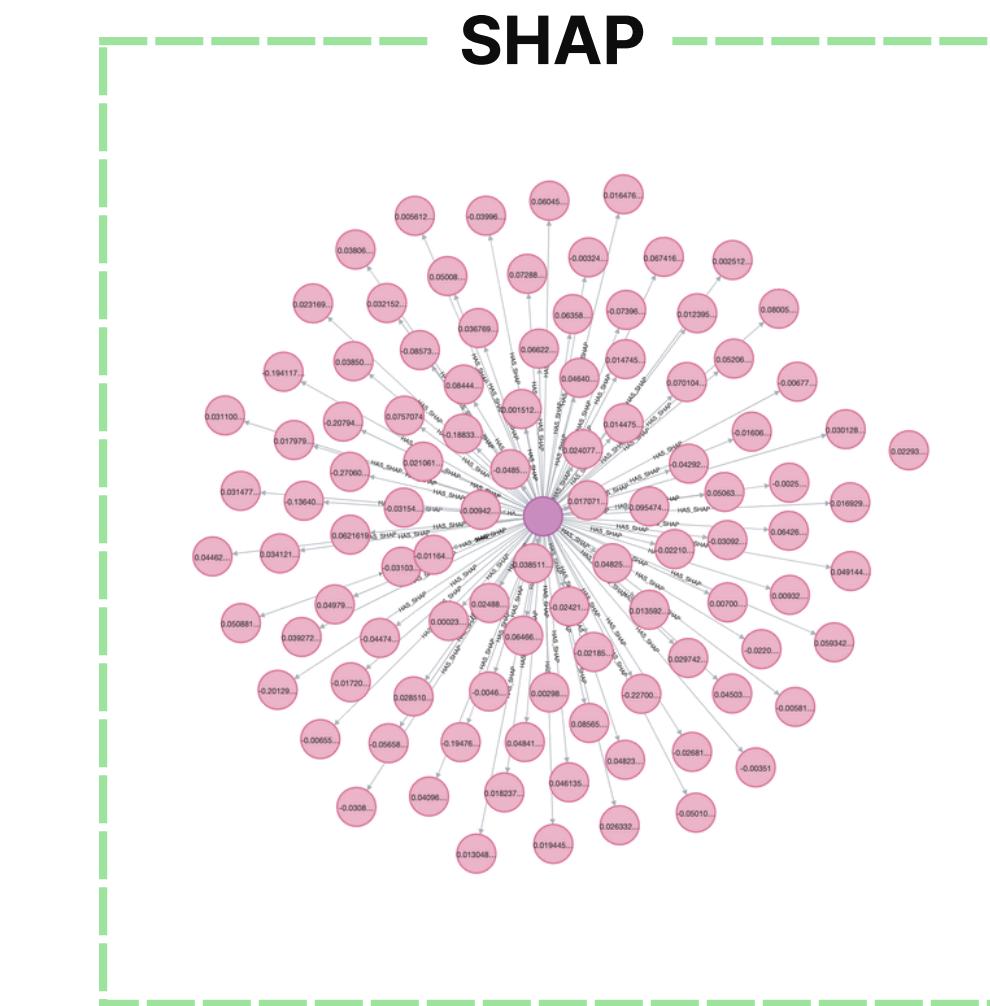
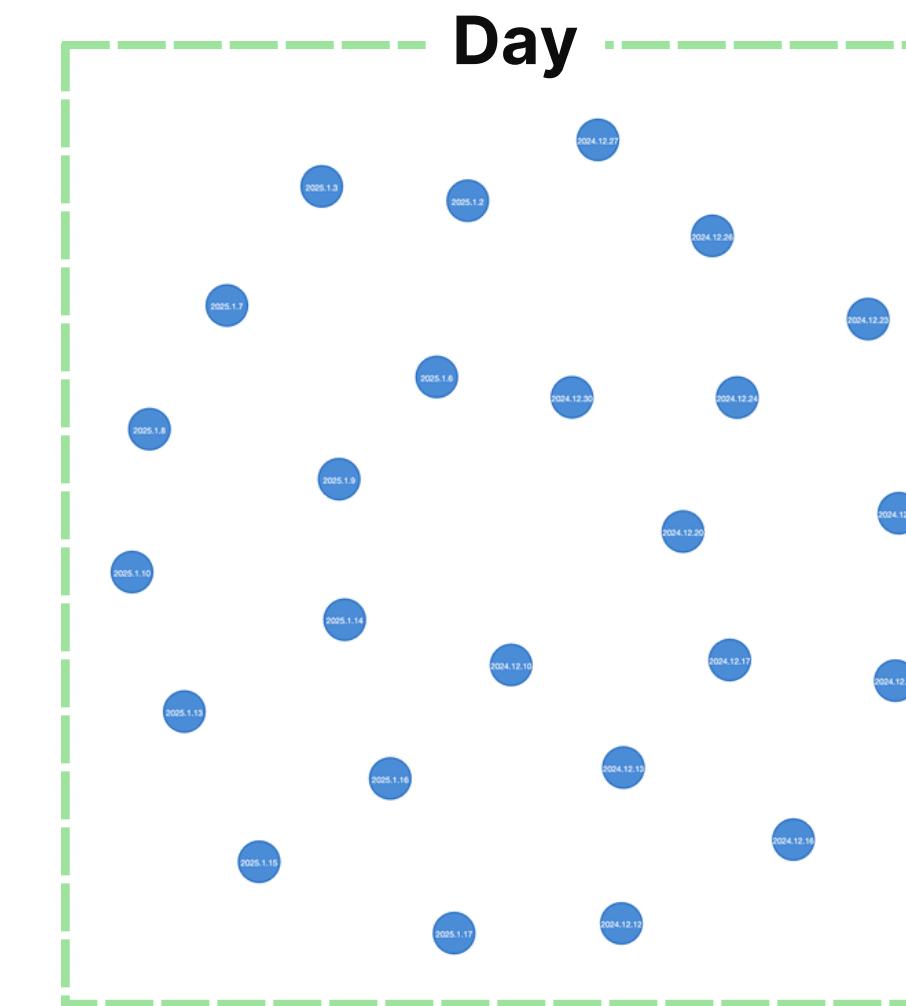
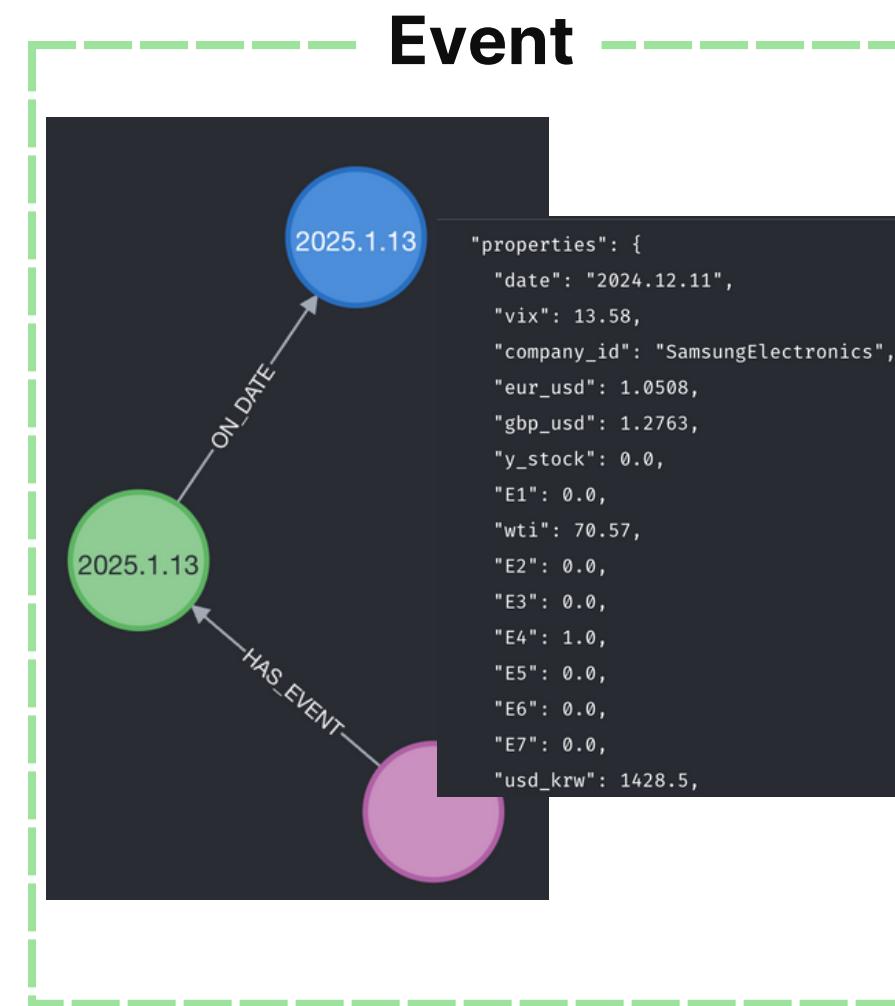
### SHAP

특정 날짜/특정 종목에서,  
뉴스로 집계된 이벤트 토픽(E1~E7)이  
목표 변수  $y$ 에 얼마나 기여했는지를 나타냄

삼성전자의 '2024.12.25  
E1 SHAP Value가 -0.074였다면,  
해당 Event 변수는 평균 대비 약 0.074만큼  
'하락 방향'으로 기여한다고 해석

## 온톨로지 기반 데이터 구조화

Company  
SK Hynix



## 방법론

# 온톨로지 기반 데이터 구조화

“2025년 3월 4일 ‘두산에너빌리티’ 주가 변화량(y)에 이벤트 변수가 기여한 정도를 알고 싶다면?”

```
MATCH (c:Company {company_id: 'DoosanEnerbility'})  
MATCH (c)-[:HAS_EVENT]->(e:Event)-[:ON_DATE]->(d:Day  
{date_norm: '2025-03-04'})  
MATCH (c)-[:HAS_SHAP]->(s:Shap)-[:ON_DATE]->(d)  
RETURN  
    c.company_id AS company_id,  
    d.date AS date,  
    e.y_stock AS y_stock,  
    s.E1 AS shap_E1,  
    s.E2 AS shap_E2,  
    s.E3 AS shap_E3,  
    s.E4 AS shap_E4,  
    s.E5 AS shap_E5,  
    s.E6 AS shap_E6,  
    s.E7 AS shap_E7;
```

company_id	date	y_stock	shap_E1	shap_E2
"DoosanEnerbility"	"2025.3.4"	-3.617026466	-0.040171954	0.020961481
shap_E3	shap_E4	shap_E5	shap_E6	shap_E7
0.083331189	-0.01613472	0.002692868	0.001133302	0.03120668

## 자연어 변환

앞서 Cypher 쿼리를 통해 조회한 특정 날짜, 특정 종목의 변수 기여도를 정형화된 그래프 or 테이블로 확인했다면  
→ LLM을 활용하면 자연어로 좀 더 명확하고 이해하기 쉬운 답변을 받을 수 있다.



Neo4j AuraDB URI,  
Username, Password 연동

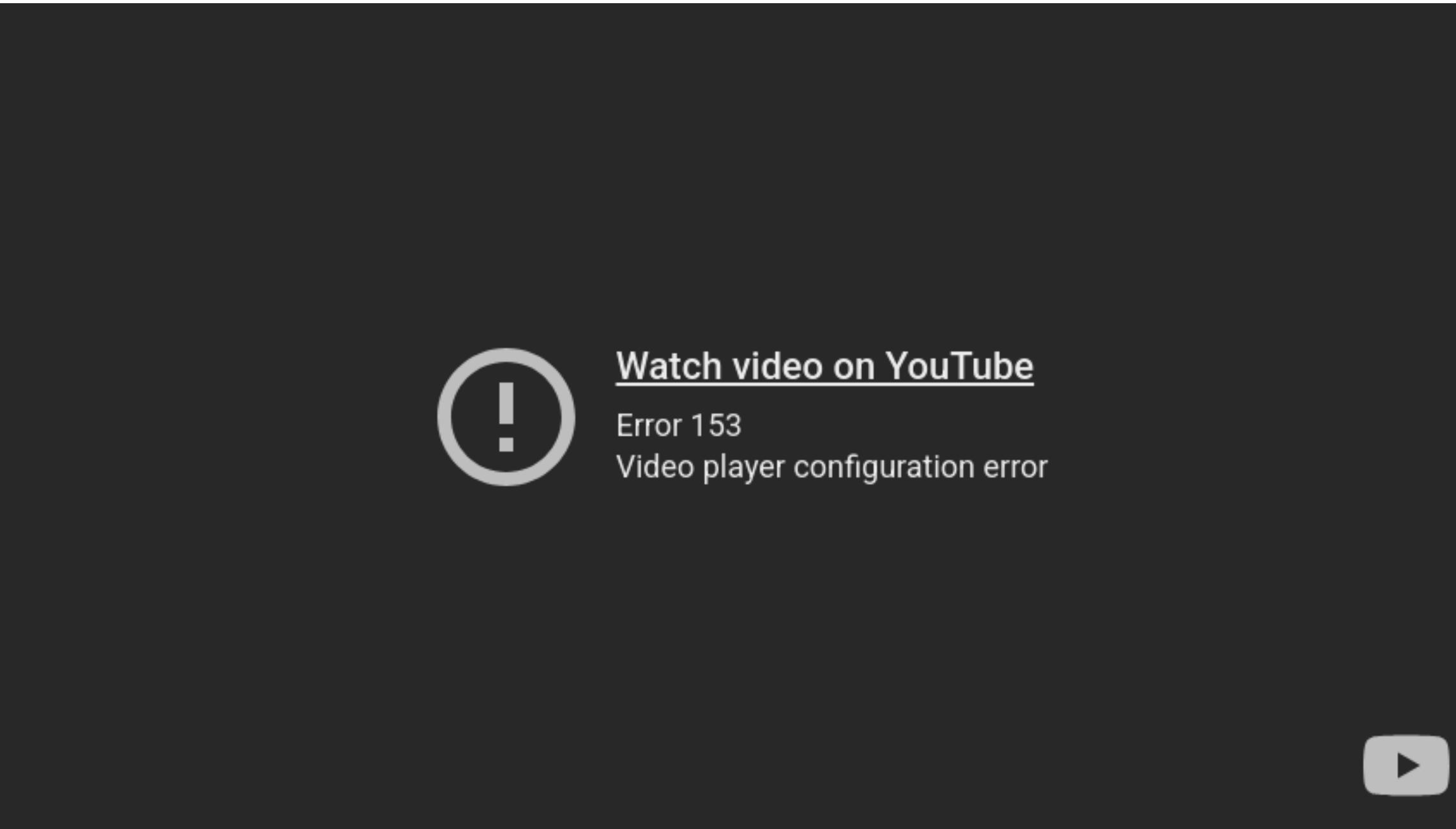


Cypher 쿼리를 활용해 특정 날짜(date),  
특정 종목(company\_id)가 포함된 Shap 노드 조회



조회한 Shap 노드를 Json 형태로 출력 후  
LLM에 자연어 분석 요청

결과



<https://youtu.be/14qN1algOT4?si=uLj7R4TEidpCly66>

## 한계 및 발전

### 한계

1. **뉴스와 주가 변화율 간의 시차(lag) 반영 불가.** t시점의 뉴스가 t-1시점과 t시점 사이의 종가 변화율에만 영향을 미친다고 가정하여, 뉴스 효과의 지연 반응이나 누적 효과를 충분히 반영하지 못함.
2. **주가 이동의 방향은 일부 설명 가능하나 변동폭을 설명하기에는 부족.** 뉴스 정보의 특성 상, 크기 정보보다는 부호 정보에 강하기 때문에 크기 설명보다는 방향성 설명이 더 적합하다고 판단했음.
3. **SHAP값 해석 불안정성.** SHAP 도구를 통해 어떤 이벤트(ex. E5-투자/증설, E6-M&A/지분)가 상대적으로 중요했는지는 확인 가능하나, 전체 모델의 성능이 낮아 해석 신뢰도에 한계가 있음.
4. **온톨로지의 구조적 장점이 모델 성능으로 직결되지는 않음.** 온톨로지는 이종 데이터 간의 관계 정의와 의미 통합에서 장점을 가짐. 하지만, 그 자체로 예측 성능을 보장하지는 않음.

## 한계 및 발전

### 발전

1. **시차(lag)를 반영한 시계열 확장** - 본 프로젝트에서는  $t$ 시점의 뉴스가 즉시 반영된다고 가정, 하지만 실제 시장에서는 뉴스 효과가 지연되거나 누적되어 나타날 수 있음. → **뉴스 및 이벤트 변수를 다중 시차(lag)로 확장**
2. **맥락 기반 뉴스 데이터 분석 도입** - 현재는 이벤트(ex. 실적, 정책, 리스크 등) 발생 여부를 기반으로 뉴스를 분석하고 있으나, 이는 뉴스의 중요도·강도·맥락 차이를 충분히 반영하지 못함. → **문맥 정보를 반영하는 언어 모델을 활용**
3. **뉴스 데이터 이외의 feature 확장** - 현재는 뉴스 데이터만을 바탕으로 설명하려고 하나 이는 명확한 한계가 있음. → 뉴스 데이터와 매크로 변수를 적절히 조합하여 **더 높은 설명력을 추구**

# EOD