

뉴스 감성 분석을 통한 글로벌 자동차 배터리 기업

Top 5 주가 변동 현황 분석



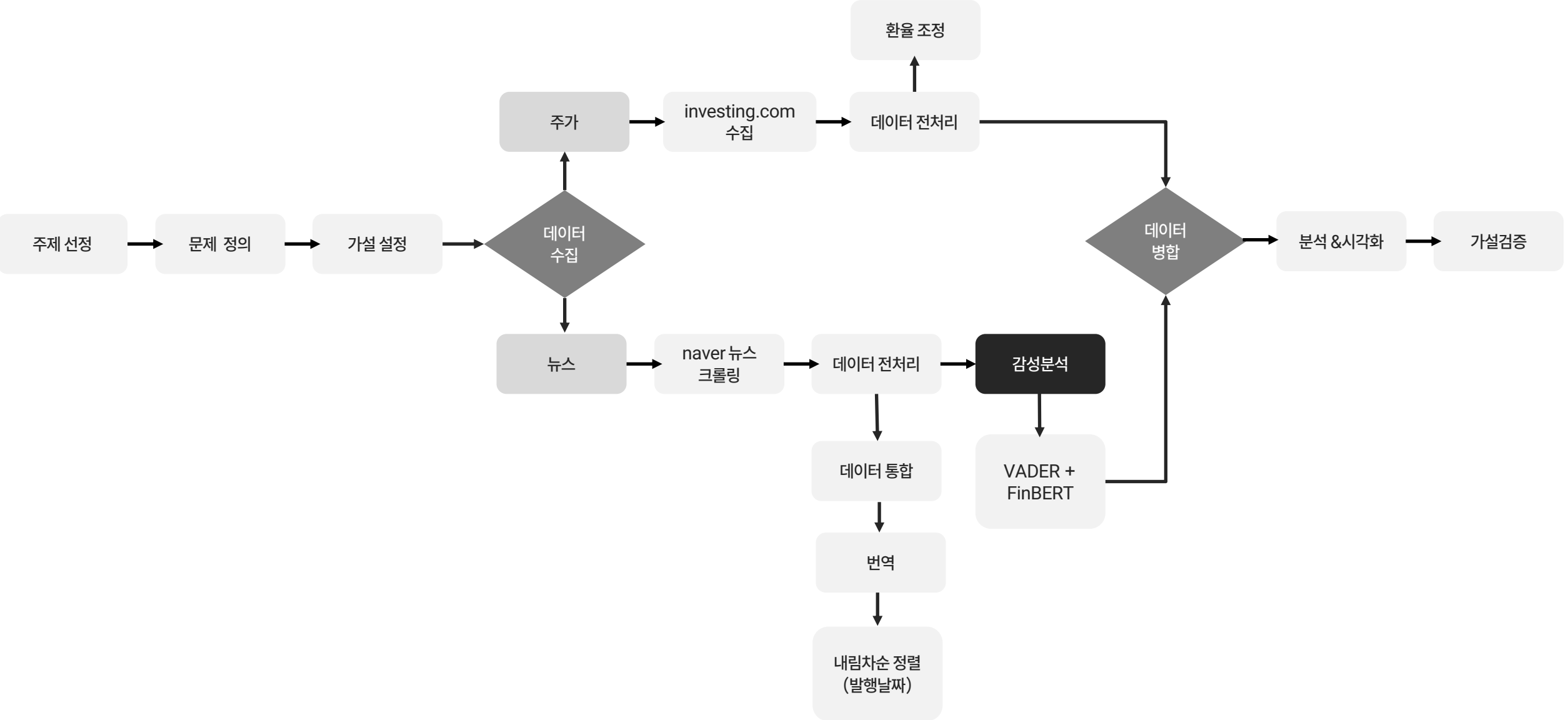
| Project Overview

- 프로젝트 목표 : 네이버 뉴스 감성 분석과 주가변동 현황 추이 분석
- 문제정의 : 언론이 주가에 미치는 영향 문제 : 최신 6개월 네이버 뉴스 감성 분석을 활용해, 글로벌 자동차 배터리 기업 Top 5 의 주가 변동 현황을 파악하고, 언론이 주가에 미치는 영향을 분석한다.
- 분석 대상 기업 : BYD , CATL, 테슬라, 파나소닉, LG에너지솔루션
- 사용 데이터
 - 뉴스 데이터 : 네이버 뉴스
 - 주가 데이터 : investing.com
- 사용 기술: Python, Sentiment Analysis, 자연어 처리(NLP) 모델

기업	Number of articles	Number of stock data	Total
BYD	9388	121	9509
CATL	9771	121	9892
테슬라	19101	124	19225
파나소닉	9881	121	10002
LG에너지솔루션	19245	120	19365

keyword	Number of articles
전기 자동차	12445
일론 머스크	10038
ev	8785
BYD	5257
2차전지	4392
리튬	4131
테슬라	3986
CATL	2731
Electric vehicle	2254
LG 엔솔	1630
Tesla	1593
Elon Musk	787
Lg energy solution	721
Lithium	597
LG 에너지 솔루션	500
파나소닉 배터리	471
Panasonic battery	28

| Project Flow-chart



뉴스 데이터



1. 데이터 수집

- ① 기사 출처 : 네이버 뉴스
- ② 수집 기간 : 2024-07-01 ~ 2024-12-24
 - * 12/26 이후 날짜 수집 불가 (ex. 1일전, 3주전)
- ③ 수집 방법 : 파이썬 라이브러리를 사용하여 크롤링
 - * 키워드 선정(영어 + 한글) → 키워드별 기사 수집
 - 기업명 : BYD, CATL, LG 에너지 솔루션 Tesla, Panasonic Battery 등
 - 관련 키워드 ; Elon Musk, EV, electric vehicle, 전기차, 2차전지, 리튬 등.
- ④ 수집 컬럼 : keyword, title, link, summary, date

2. 데이터 전처리

- ① summary 컬럼 영어로 번역 → 감성분석 정확도를 위한 작업
 - * 파이썬 : 구글 번역 패키지 설치 후 진행 (pip install googletrans==4.0.0-rc1)
 - * google spreadsheets : 보조 Tool로 사용 (=googletranslate(\$H2,"ko","en"))
- ② 감성분석 : summary 컬럼에 대해 진행
 - * 전처리 : 특수문자 제거
 - * 감성분석 : VADER + FinBERT
 - VADER : 사전 기반 방법으로 긍정적, 부정적, 중립적 감정을 분석
 - 빠르고 효율적, 단순한 텍스트 분석에 좋으나, 특정 도메인 문맥 파악에 한계
 - FinBERT : 금융 분야 특화 모델로 주식/금융 관련 텍스트에서 강점

주가 데이터



1. 데이터 수집

- ① 출처 : **Investing.com**
- ② 수집 기간 : 2024-07-01 ~ 2024-12-25 (약 6개월)
 - * 기사 수집 일자에 맞춰 수집 기간 조정
- ③ 수집 방법 : investing.com 기업명 검색 → 과거 데이터 다운로드
 - * 기업명 : BYD, CATL, LG 에너지 솔루션 Tesla, 파나소닉
- ④ 수집 컬럼 : date, 종가, 시가, 고가, 저가, 거래량, 변동 %

2. 데이터 전처리

- ① **기업명** 컬럼 추가
- ② **화폐 단위** 통일
 - * 환율 데이터 ; investing.com → 과거 환율 데이터 다운로드
 - * 기준 화폐 : USD
 - * 일자별 환율 '**종가**'로 기준 통일
 - * BYD(위안화), CATL(위안화), LG 에너지 솔루션 (원화), 파나소닉(엔화)

| Sentiment Analysis

‘텍스트 감성분석’이란? 텍스트에 담긴 의견이나 **감성, 평가, 태도** 등의 주관적인 정보를 **컴퓨터를 통해 분석**하는 과정
자극적인 요소가 다분한 뉴스 제목이 아닌 **요약** 칼럼을 사용하여 텍스트 감성분석 진행

크롤링한 뉴스 데이터.csv

키워드	제목	발행날짜	요약
LG 엔솔	"중국 뒷발 뚫었다"...LG엔솔, 전기차용 LFP 배터리 첫 대규모 수주	2024.7.2	LG에너지솔루션과 ... (중략) ... LFP 배터리 공급계약을 체결 했다. LG에너지솔루션이 처음으로 전기차용 리튬인산철(LFP) 배터리 대규모 수주에 성공 했다. ...
LG 엔솔	해외로 ‘성큼성큼’ 중국 배터리의 약진 ...‘K-배터리’ 빨간불 켜지나	2024.7.4	LG에너지솔루션은 5.9% 성장한 33.3GWh(점유율 25.6%)로 중국 CATL에 이어 글로벌 2위를 차지 했다.
LG 엔솔	K-배터리, 2분기도 ‘한파’ ...투자속도 조절·ESS 확대로 실적 방어	2024.7.8	8일 금융정보업체 에프앤가이드에 따르면 LG에너지솔루션, 삼성SDI 등 국내 배터리 셀 제조사들의 2분기 실적이 이달 발표된다.
LG 엔솔	현대차·삼성전자·금호타이어 주가 상승세...SK이노·에코프로는 하락...	2024.7.8	LG에너지솔루션(373220)은 35만 5500원을 기록했다. 이는 전 거래일 증가 35만 7500원 대비 0.55%가 하락한 수치 다.
⋮	⋮	⋮	⋮

input

VADER + FinBERT

output

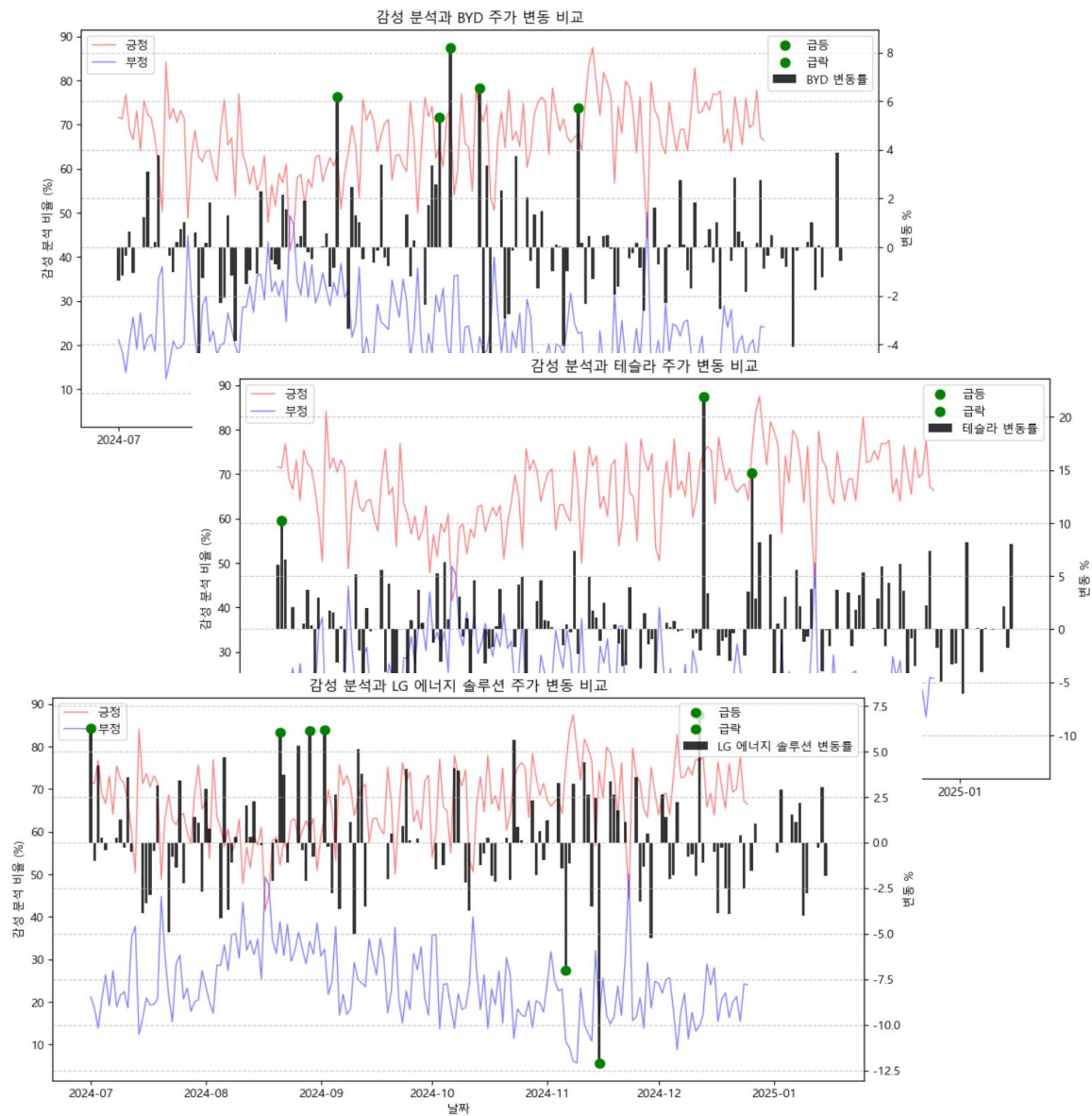
감성점수	감성결과
0.54	Positive
0.86	Positive
0	Neutral
-0.70	Negative
⋮	⋮

감성점수 사전 정의

- VADER’s Sentiment score > 0.05 : 긍정
- VADER’s Sentiment score < -0.05 : 부정
- VADER’s |Sentiment score| <= 0.05 → FinBERT’s Sentiment score

Hypothesis Test # 변동률

가설: **긍정적** 기사보다 **‘부정’**적 기사가 **‘주가 변동률’**에 미치는 영향이 더 클 것이다.



[분석방법 - 회귀분석]

* cf. 회귀분석: 어떤 요소(X)가 결과(Y)에 얼마나 영향을 주는지 알아보는 방법

- ✔ 부정적인 기사 비율이 높을수록 주가 변동률 변화가 커진다.
- 부정적 기사 비율(Negative)의 **회귀 계수: 0.1794**
- **p-value = 0.0027** (유의수준 0.05 이하 → 유의미함)
- 즉, 부정적인 기사가 많을수록 주가 변동률이 커지는 경향이 있음.

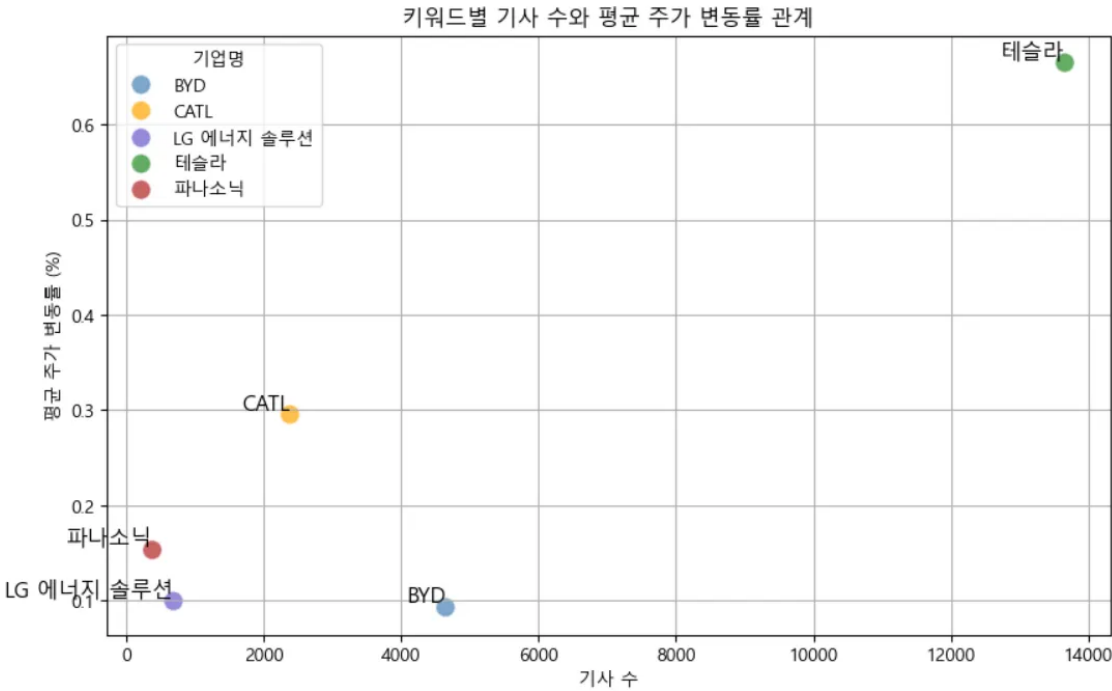
- ✔ 긍정적인 기사(Positive) 비율도 유의미한 영향이 있지만, 부정 기사보다 영향력이 작다.
- 긍정적 기사 비율(Positive)의 **회귀 계수: 0.1662**
- **p-value = 0.0036** (유의미한 영향이 있음)
- 긍정적인 기사도 주가 변동률과 관련 있지만, 부정 기사보다 영향이 약함.

- ✔ 결정 계수(R^2 , 모델 설명력) 확인
- **$R^2 = 0.076$** → 감성 분석 데이터만으로는 주가 변동률을 충분히 설명하기 어려움.
- 즉, 주가 변동률에는 감성 분석 외에도 금리, 실적 발표, 거시 경제 등의 추가적인 요인이 영향을 미친다

- ❗ 결론: **부분적인 YES**
- ✔ 부정적인 기사가 주가 변동률에 미치는 영향이 긍정적인 기사보다 크다.
- ✔ 다만, 감성 분석만으로는 주가 변동을 완벽하게 설명할 수 없고, 추가적인 요인도 고려해야 한다.

Hypothesis Test # 변동률

가설: 동일 키워드에 대한 기사 수가 많아질수록 주가에 미치는 영향은 작아질 것이다.



✓ 분석 방법

cf. Pearson 상관분석: 두 값이 같이 오르내리는지(연관이 있는지) 알아보는 방법

- 기사 수와 주가 변동률 간의 관계 분석 (Pearson 상관분석)
- 기업별 기사 수와 평균 주가 변동률 비교 (산점도 시각화)

!! 결론: 가설 기각 (NO)

✓ 결론이 가설과 반대됨

✓ 기사 수가 많아질수록 주가 변동률이 작아지는 것이 아니라, 오히려 변동성이 커지는

경향이 있음

✓ 상관계수(0.8875)는 강한 양의 상관관계를 의미하며, 기사 수가 증가할수록 주가 변동률도 증가함

✓ p-value(0.0445)는 0.05 미만으로 통계적으로 유의미한 관계가 있음을 나타냄

✓ 즉, 기사 수와 주가 변동률은 관련이 있으며, 기사 수 증가가 변동률 감소로

이어진다는 가설은 틀림

Hypothesis Test # 국가간

가설: 미국 시장의 주가 변동 폭이 동아시아 시장의 주가 변동 폭보다 클 것이다.

가설 설정 배경: 미국 기업 Tesla의 주가 변동은 동아시아 기업들보다 주가 변동이 잦을 것이라 생각했기 때문이다.

시장	주가 변동%	표준편차
동아시아	0.19	2.58
미국	0.64	4.56



Levene's Test

- 두 개 이상의 그룹 간 분산이 동일한지(Homoscedasticity, 등분산성)를 검정하는 방법
- 미국과 동아시아 시장의 주가 변동성이 통계적으로 차이가 있는지 검정하는 데 적합

Levene	21808.93
P-value	0

p-value < 0.05이면 두 시장의 변동성이 통계적으로 다르다
가설이 통계적으로 유의미하다 but 대립가설 검증은 불가

t - Test

- 두 그룹 간의 평균 차이가 통계적으로 유의미한지 확인하기 위해 사용되는 가설 검정 방법
- 주로 두 독립적인 그룹 간의 평균을 비교(양측검정)하거나, 하나의 그룹의 평균이 특정 값과 다른지(단측검정)를 확인

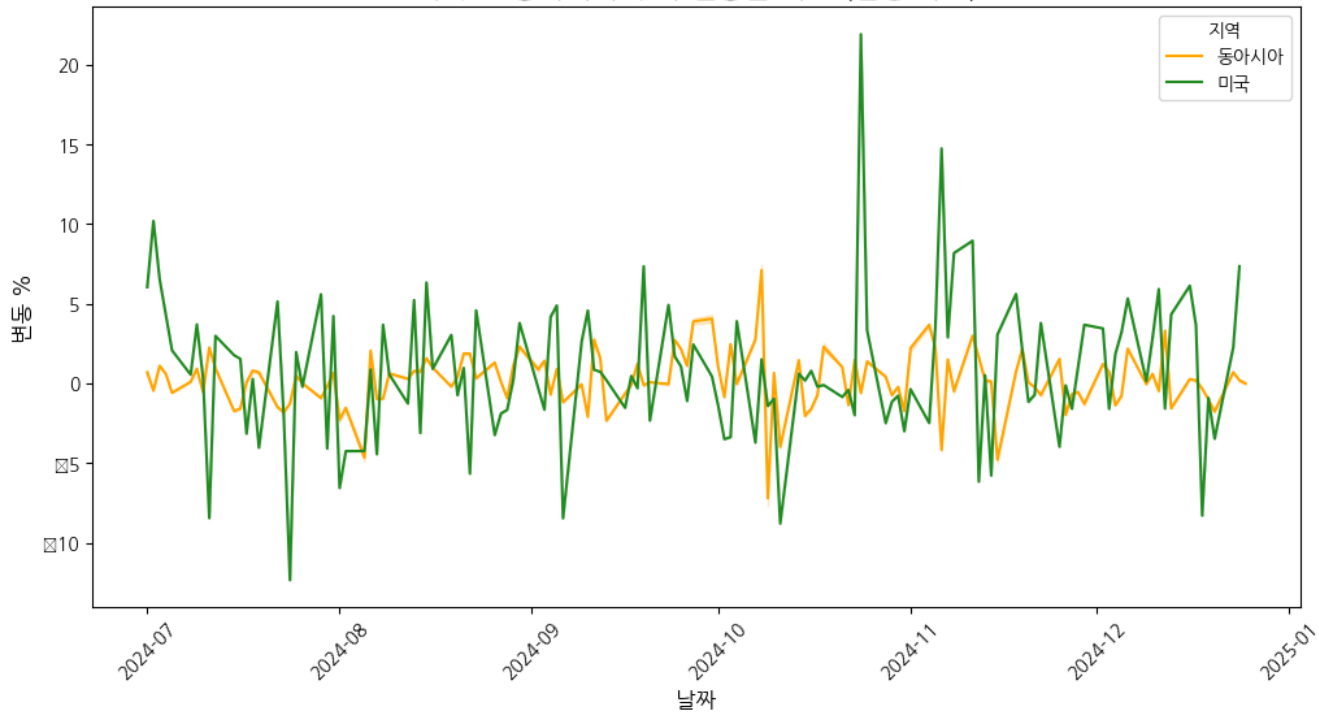
t (양측검정)	23.08
P-value	0

p-value < 0.05이면 두 시장의 평균 변동 차이가 있다
가설이 통계적으로 유의미하다

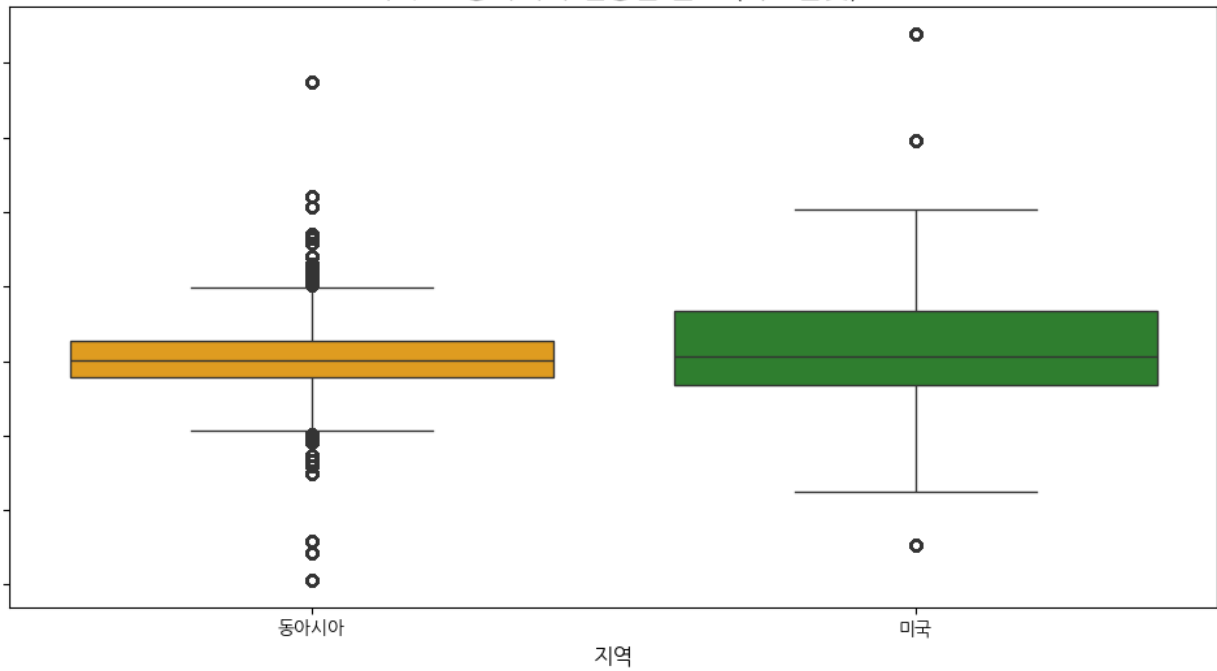
Hypothesis Test # 국가간

가설: 미국 시장의 주가 변동 폭이 동아시아 시장의 주가 변동 폭보다 클 것이다. → YES

미국 vs 동아시아 주가 변동률 비교 (선형 차트)



미국 vs 동아시아 변동률 분포 (박스플롯)



- 미국 시장은 변동이 심하고, 급격한 변동이 여러 번 발생
- 동아시아 시장은 비교적 안정적인 흐름을 유지하며 큰 변동없이 일정한 패턴

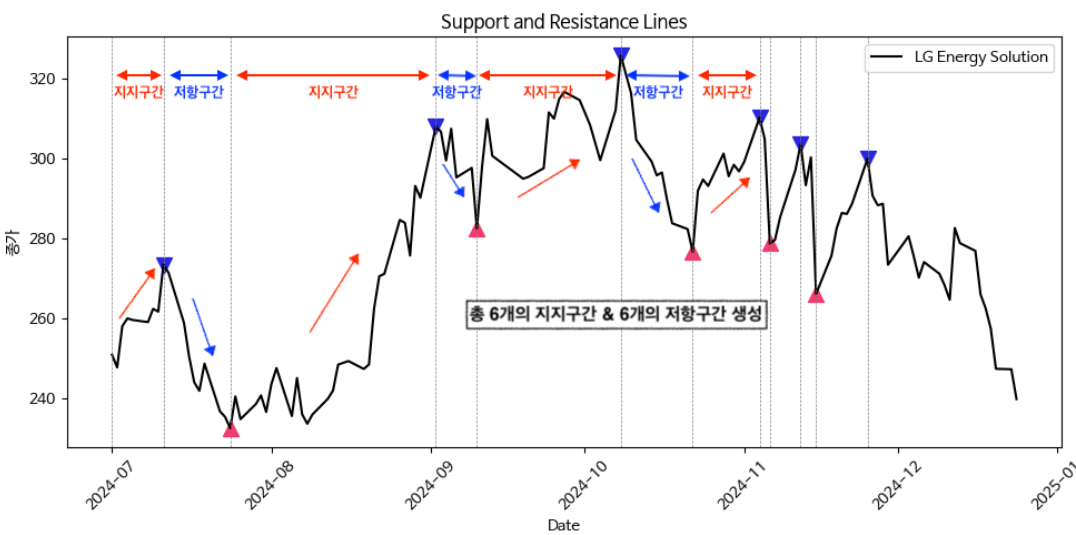
미국시장
변동률 분포가 더 넓으며 이상치(극단적인 변동)가 적음
중앙값이 상대적으로 높고, 상하위 사분위 범위(IQR)도 동아시아보다 크다.

동아시아 시장
변동성이 미국보다 낮고 안정적인 분포
이상치가 미국보다 많고, 변동률 범위도 좁다.

Hypothesis Test #종가

가설 : 종가가 상승/하락하는 구간에는 뉴스의 각 긍정/부정의 비율이 높을 것이다.

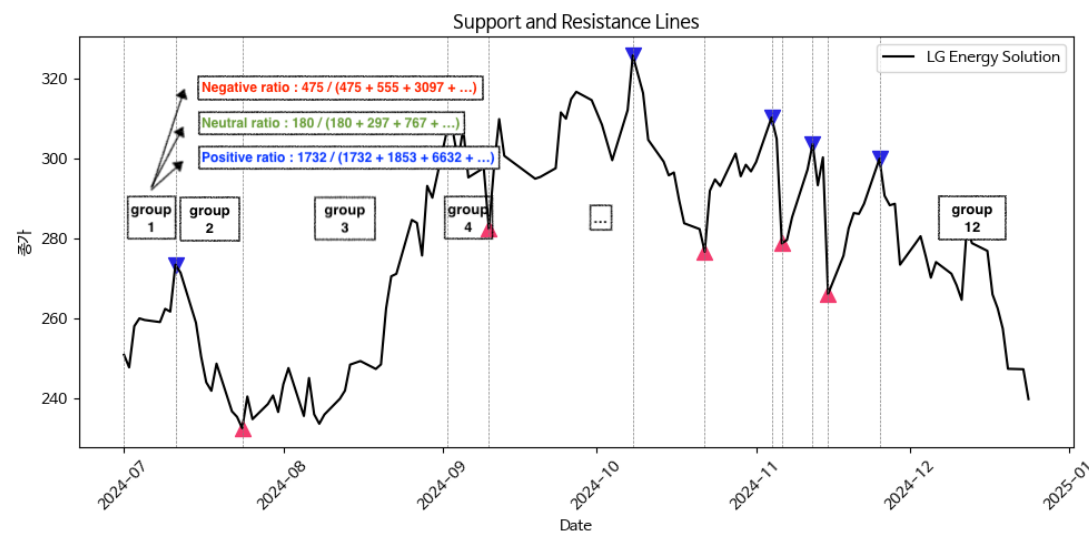
1. 저항구간과 지지구간 설정



저항구간 : 하락하기 시작하는 가격 수준

지지구간 : 상승하기 시작하는 가격 수준

2. 구간별 최대 감성점수 추출



	감성결과	ratio
group		
1	Negative	0.103951
2	Negative	0.064438
3	Neutral	0.260388
4	Positive	0.046517
5	Positive	0.126447
6	Neutral	0.063712

→ 지지구간 인 group1 에서는 Negative의 비율이 가장 높음

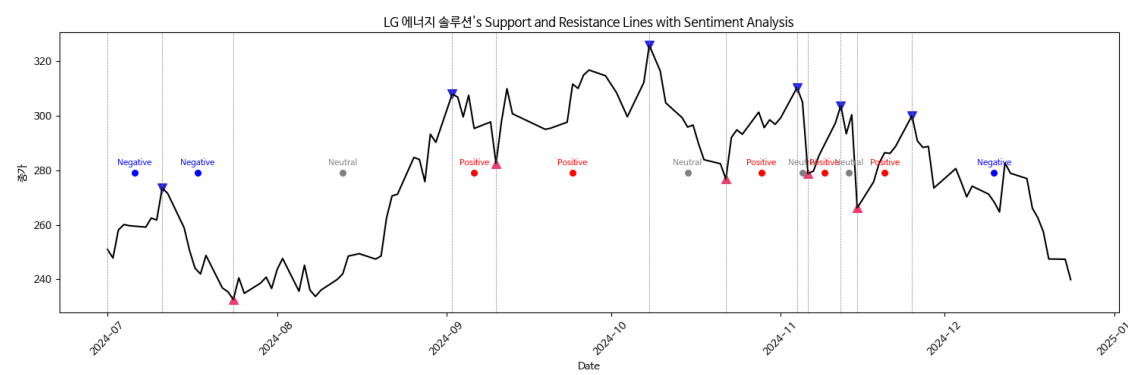
→ 저항구간 인 group2 에서는 Negative의 비율이 가장 높음

⋮

Hypothesis Test # 종가

가설 : 종가가 상승/하락하는 구간에는 뉴스의 각 긍정/부정의 비율이 높을 것이다.



LG Energy Solution



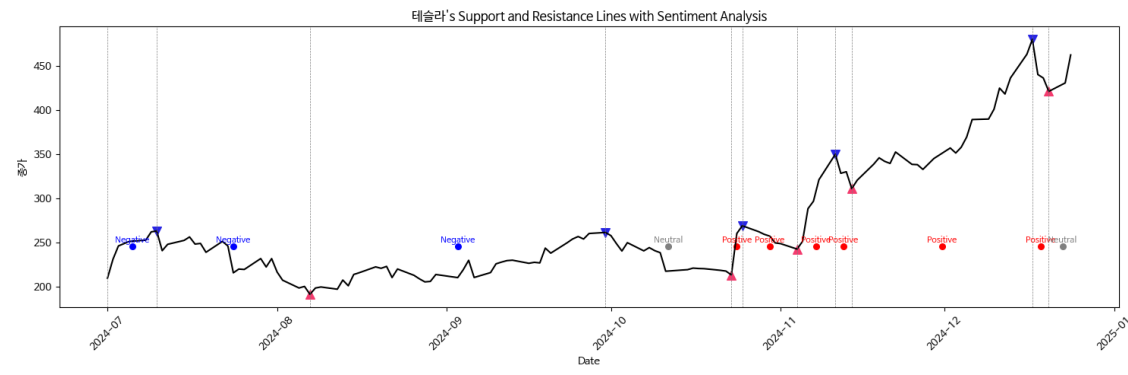
		count
Type	감성결과	
저항구간	Negative	2
	Neutral	3
	Positive	1
지지구간	Negative	1
	Neutral	1
	Positive	4

검증 결과

테슬라를 제외한 4개의 기업에서

- 종가의 상승구간 에는 긍정의 비율이 높고
- 종가의 하락구간 에는 부정의 비율이 높음

Tesla

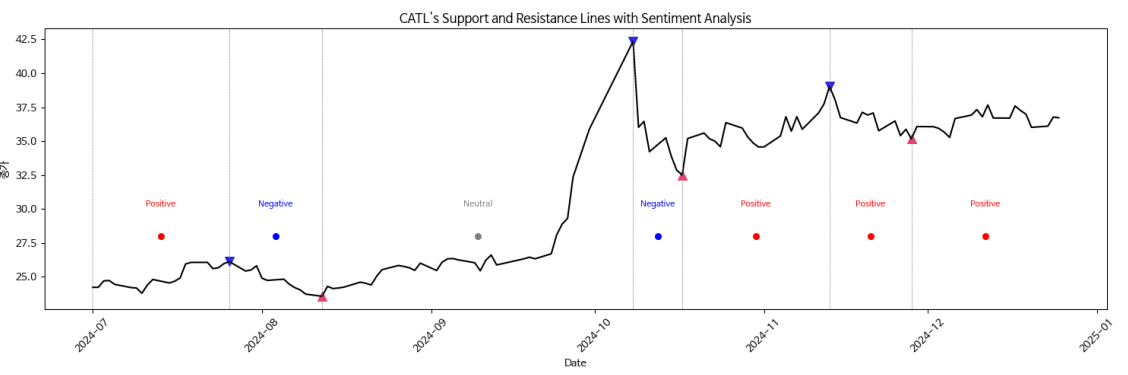


		count
Type	감성결과	
저항구간	Negative	1
	Neutral	1
	Positive	3
지지구간	Negative	2
	Neutral	1
	Positive	3

(테슬라의 경우 종가가 낮을 때에는 Negative가,
종가가 높을 때에는 Positive가 분포)

→ YES

CATL



		count
Type	감성결과	
저항구간	Negative	2
	Positive	1
지지구간	Neutral	1
	Positive	3

Hypothesis Test # 증가

가설 : 네이버 뉴스 감성점수는 테슬라 증가 변동에 20%의 영향력을 끼칠 것이다.

가설 설정 배경 : 테슬라의 CEO 성향을 고려했을 때, 언론이 주가변동에 적어도 20% 정도의 영향을 끼칠 것이라 판단



회귀분석의 필요성

$$Y = aX + b$$

Y : 테슬라 증가
a : 회귀계수
X : 감성점수
b : 절편

- 1st 구현 과정
1. 전처리

- 날짜별 평균 감성점수 + 증가
2. 가공

- 기업별 데이터 분리
 - 날짜 형식 변환 (datetime)
3. 회귀분석 & 시각화

- 각 기업별 회귀분석 진행
 - 그래프 생성

전처리 → 기업 관련 키워드 필터링 추가



 채택
1st 결과

회귀계수	276.32
절편	188.23
R^2	0.27

 기각
2nd 결과

회귀계수	173.35
절편	223.83
R^2	0.14



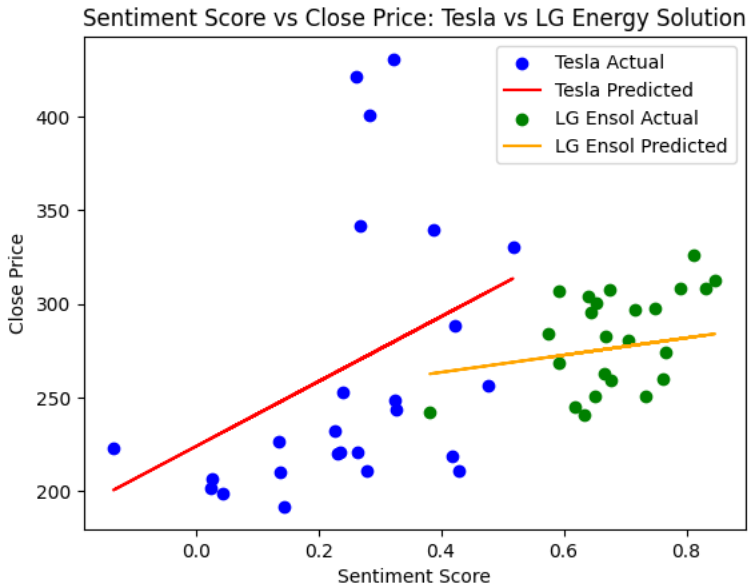
LG 에너지 솔루션

회귀계수	45.93
절편	245.10
R^2	0.10

Hypothesis Test # 증가

가설 : 네이버 뉴스 감성점수는 테슬라 증가 변동에 20%의 영향력을 끼칠 것이다.

테슬라 vs LG 에너지 솔루션



- 기울기 : 테슬라가 감성 점수와 증가 사이에 더 강한 상관관계가 존재
- 상관관계: 테슬라의 점들이 회귀선 근처에 밀집되어 있어 상관관계가 강하다고 볼 수 있음. 반면에, LG 에너지 솔루션의 점들은 회귀선 주변에 균등하게 흩어져 있어 감성점수와 증가 간의 상관관계가 약하거나 없음을 의미함.
- 결과 : 테슬라 투자자들이 뉴스 감성에 더 민감하게 반응할 가능성이 있음을 시사하며, LG 에너지 솔루션의 경우 뉴스감성 보다는 다른 요인(기업 실적, 산업 동향 등)에 의해 더 크게 영향을 받을 가능성이 있음을 시사.

X축 : 감성점수(Sentiment Score)

Y축 : 증가(Close Price)

점(Scatter) : 파란색(테슬라의 감성점수와 주가 데이터),

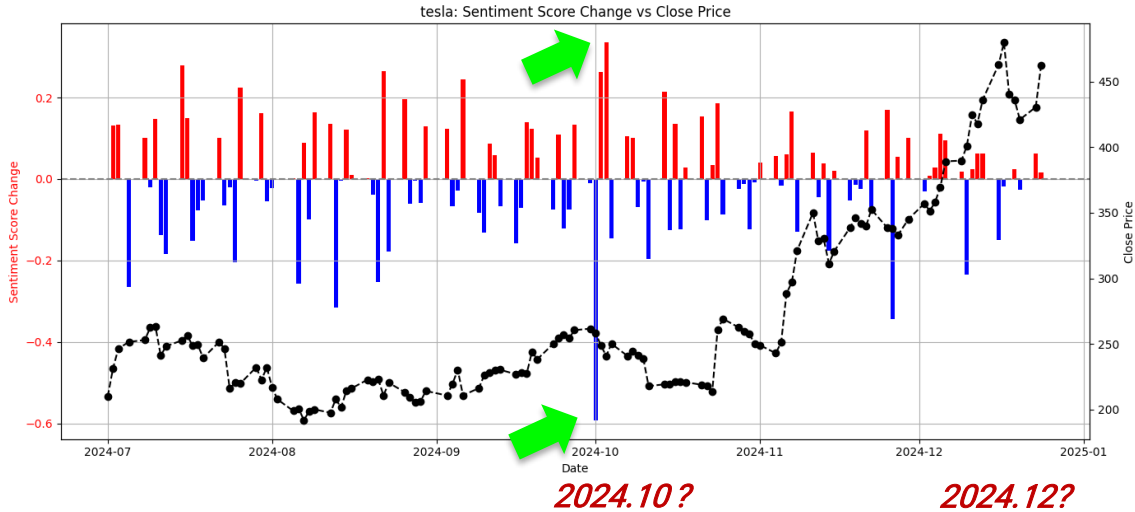
초록색(LG 에너지 솔루션의 감성점수와 주가 데이터)



즉, **네이버 뉴스**가 국내 기업인 LG 에너지 솔루션 보다
테슬라 증가변동에 더 큰 영향을 끼침을 알 수 있음

감성점수 vs 종가변동(line chart) _ 테슬라

24.07.01 ~ 24.12.24



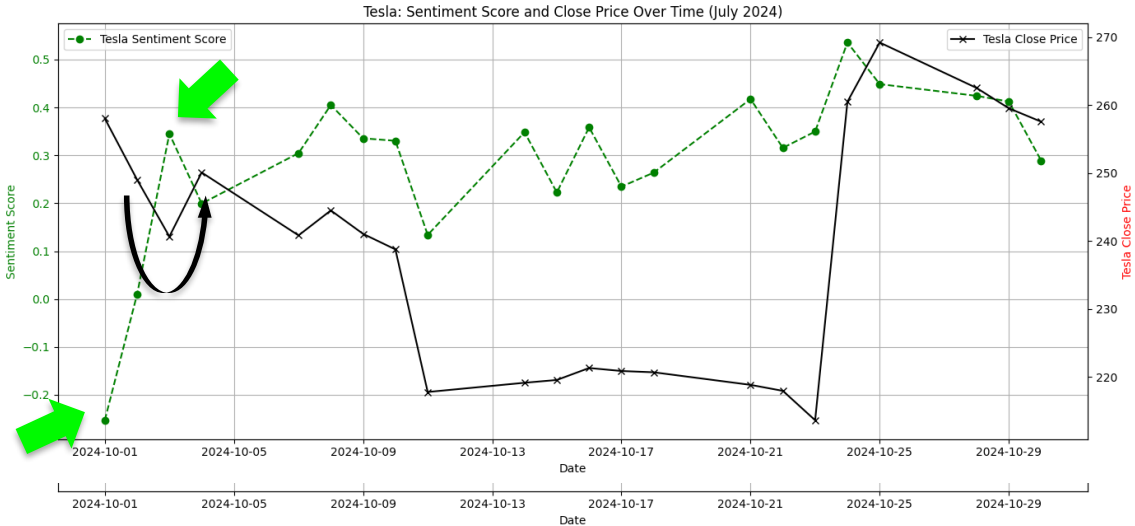
📌 감성점수 vs 주가변동 시각화로 어떤 걸 알 수 있나?

- 감성 점수가 변할 때 **테슬라**의 주가가 **어떻게 반응**하는지 확인 가능
- 감성 점수가 상승할 때 주가도 상승하는 **패턴**이 있는지 분석 가능

✅ 얻을 수 있는 인사이트

- **감성 점수 급락** 후 주가 하락이 동반되면, 부정적 뉴스가 투자자 심리에 영향을 준 것일 수 있음
- **감성점수 급등** 후 주가 상승이 동반되면, 긍정적 뉴스가 투자자 심리에 영향을 준 것일 수 있음

24.10.01 ~ 24.10.30



10월

- 감성점수가 급락하여 종가가 떨어지는 가 싶다가 감성점수가 곧바로 급등하자 종가가 하방을 지지하고 상승
 - ✓ **극단치의 감성점수**는 종가변동에 영향을 주는 것으로 보임
 - ✓ **비극단치의 감성점수 vs 종가변동** 간 필연적인 패턴이 보이지는 않음

12월

- **감성점수와는 별개로** 종가 급등함 → 감성점수 외 요인(정치상황)들도 큰 영향을 끼친다는 걸 알 수 있음

| Modeling

- 시계열 예측 : 시계열 예측은 시간에 따라 수집된 이전 데이터 포인트를 기반으로 미래 값을 예측하는 도구
- 예측 모델로 LSTM 사용
- LSTM은 장기간 동안 중요한 패턴을 효과적으로 학습하고 기억

Train data : 47186개

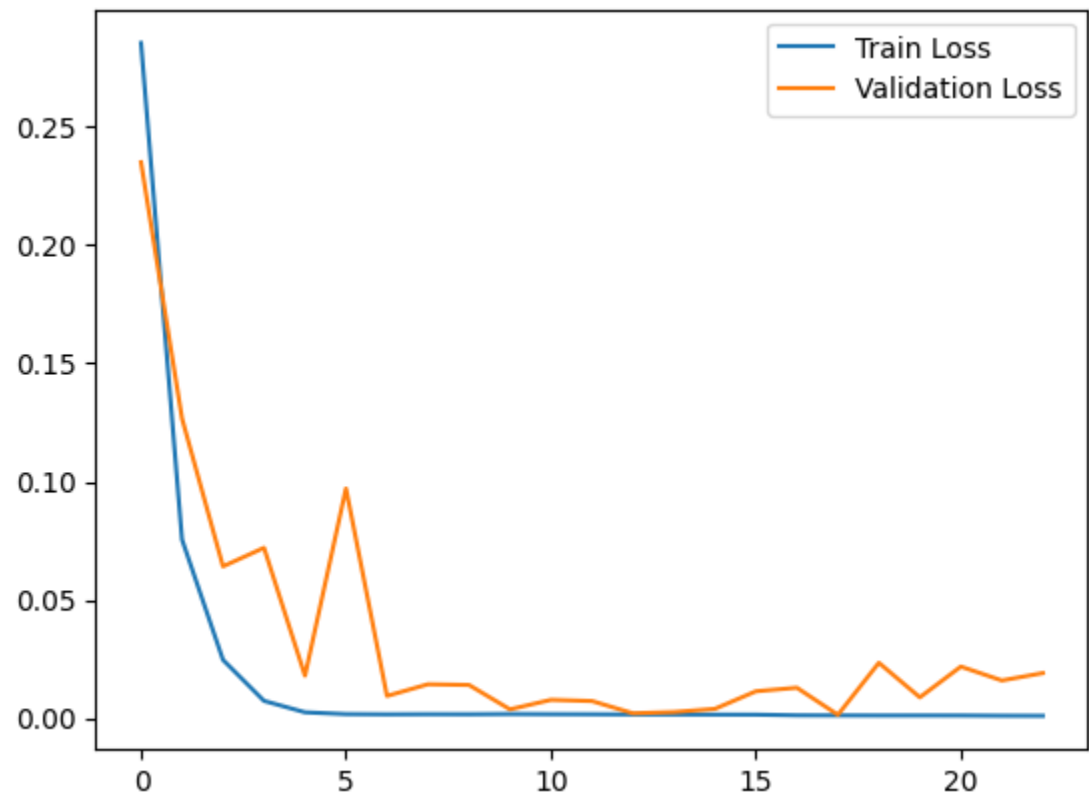
Test data : 11797개

- 성능 평가 지표

RMSE : 오차의 제곱 평균을 루트 씌운 값으로, 예측값과 실제값 간의 평균적인 차이를 측정

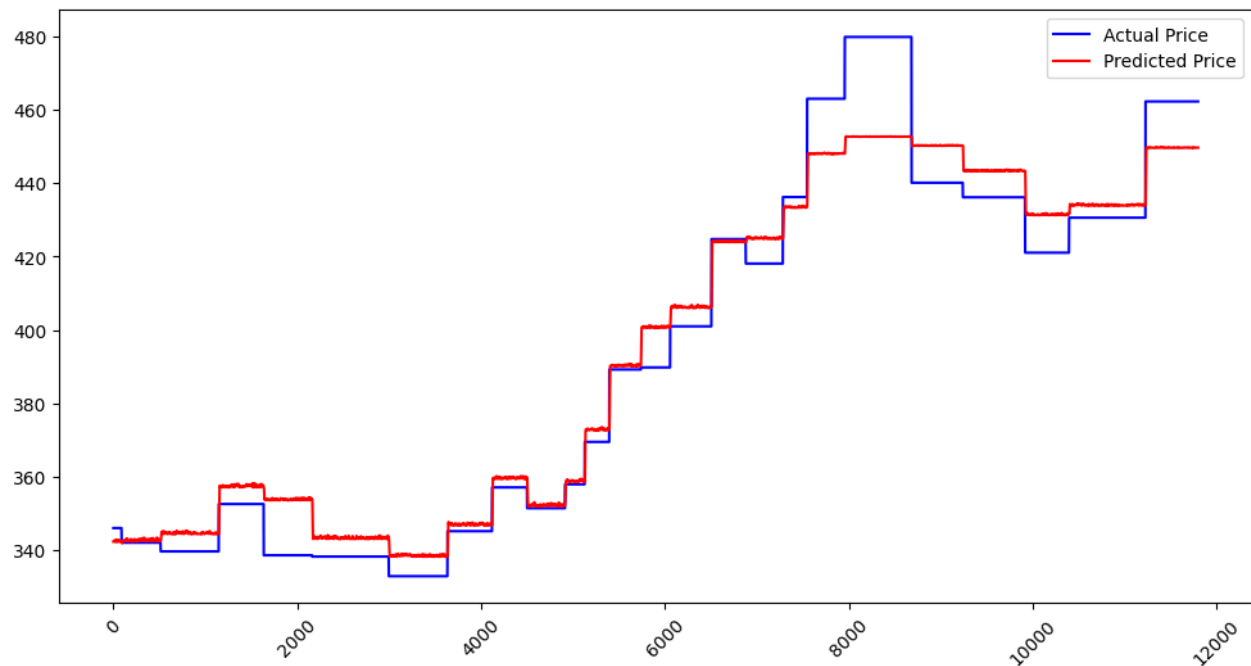
MAE : 예측값과 실제값의 차이의 절대값의 평균

모델 학습 과정



- 초반에는 Train Loss와 Validation Loss가 급격히 감소
- 약 5~6 Epoch 이후에는 거의 수렴하면서 안정적으로 학습됨
- 일부 구간에서 Validation Loss가 불규칙하게 상승하는 모습이 보임
- 이는 데이터 샘플링 편차 또는 모델의 과적합(overfitting) 가능성을 의미할 수 있음
- 과적합 가능성 낮음
- Train Loss와 Validation Loss가 거의 비슷한 수준으로 수렴하여, 심각한 과적합(overfitting) 문제는 보이지 않음.

실제 주가(Actual Price) vs 예측 주가(Predicted Price)



- 예측값(빨간색)이 실제 주가(파란색)의 흐름을 대체로 따라가고 있음
- 가격이 상승하는 구간과 하락하는 구간이 유사하게 반영됨
- 주가는 계단식으로 변동(급등, 급락)하는 반면, 예측값은 다소 완만하게 움직임
- 이는 LSTM 모델이 단기 변동성을 포착하는 데 한계가 있음을 나타냄
- 고점과 저점 예측 정확도모델이 전반적인 상승·하락 패턴은 맞추지만, 극단적인 고점과 저점에서 오차가 발생

| Modeling

- 성능 평가

RMSE: 9.9328 / MAE: 7.4190

RMSE :

- 현재 RMSE값이 9.9328이므로, 모델이 예측한 주가가 실제 주가와 평균적으로 약 9.93달러 차이가 남
- RMSE는 큰 오차에 더 민감하기 때문에 일부 예측값이 벗어난 경우 이 값이 커질 수 있음

MAE :

- 현재 MAE 값이 7.4190이므로, 모델이 예측한 주가는 실제 주가와 평균적으로 약 7.42달러 차이가 난다고 볼 수 있음

결론 :

- 평균적으로 7~10달러 정도의 오차가 발생하는 것으로 볼 수 있음
- 테슬라 주가가 일반적으로 수백 달러 수준에서 변동한다면 이는 비교적 괜찮은 성능
- 추가적인 데이터 특징 반영, 모델 구조 개선, 최적의 시계열 길이 설정 등의 방법을 고려하면 예측 성능을 더욱 향상시킬 수 있겠음

언론이 주가에 미치는 영향에 대한 분석을 통해 ...

- 뉴스 감성 점수와 주가 변동 간의 상관관계를 분석하면, 특정 이벤트가 주가 변동성에 미치는 영향을 더 명확히 파악할 수 있을 것이다.
- 언론이 투자자 심리에 미치는 영향을 분석해 포트폴리오 조정과 리스크 관리에 활용할 수 있는 가능성을 제시
- 뉴스 감성 점수가 주가를 어느 정도 예측할 수 있지만, 기업 실적, 정책 변화, 산업 동향 등 다른 요인도 큰 영향을 미침
- 주가 상승 구간에서는 긍정 뉴스가, 하락 구간에서는 부정 뉴스가 많아 투가 의사결정에 뉴스 감성이 중요한 역할을 할 수 있음을 보여줌
- 다중 회귀 분석 등 더 정교한 모델링과 외부 요인 데이터를 추가하면 예측력을 더욱 높일 수 있을 것임

| Limitations & Lesson Learned

Limitations

- 기업별 키워드 필터링에 따라 각 가설에 대한 결론이 상이함
- 뉴스 키워드 필터링에 따라 주가 민감도가 달라 질 수 있음
- 코딩 실력이 미흡하여, AI 에 대한 의존도가 높음
- 통계지식이 부족하여, 가설 검증 시 놓치는 부분이 많음
- 감성분석에 대한 정확도 파악하기 힘들
- 수많은 데이터 양에 비해 실제로 모든 데이터를 활용하지 못함

Lesson Learned

- 데이터 전처리를 어떻게 하는 지에 따라 분석 결과가 달라짐
- 명확한 결론을 얻기 위해 첫 가설 설정을 잘 하는 것이 중요
- 데이터 분석 프로젝트의 핵심은 '결과의 성공 여부' 가 아닌, 문제정의의 과정과 해결하기 위한 접근 방식 및 논리를 구축하는 데 있다고 느낌
- 데이터를 시각화 할 때, 전달하고자 하는 메시지를 명확히 표현할 수 있도록 필요한 부분만 선별하여 직관적인 차트 유형을 선택해야 함
- 끊임없는 팀 소통이 필요하나, 각자의 의견을 잘 정리하여 정확히 전달 하는 것이 중요함

The background of the image is a dark blue financial chart. It features a candlestick chart with blue bars representing price movements. Overlaid on the candlesticks are several line graphs in green, purple, and pink, each with circular markers at data points. Numerical values are scattered across the chart in a light green font, including 44.291, 63.772, 26.4, 44.870, 20.556, 12.002, and 30.381. The text 'Thank You' is centered in a large, white, italicized font with a subtle blue glow. The overall aesthetic is high-tech and professional, typical of a financial presentation or a trading-related graphic.

Thank You