

빅데이터 처리 기말 프로젝트

뉴스 감정 예측 및

주가와의 상관관계 분석

2024.12.05

202044034 | 왕건



주가 예측이란?

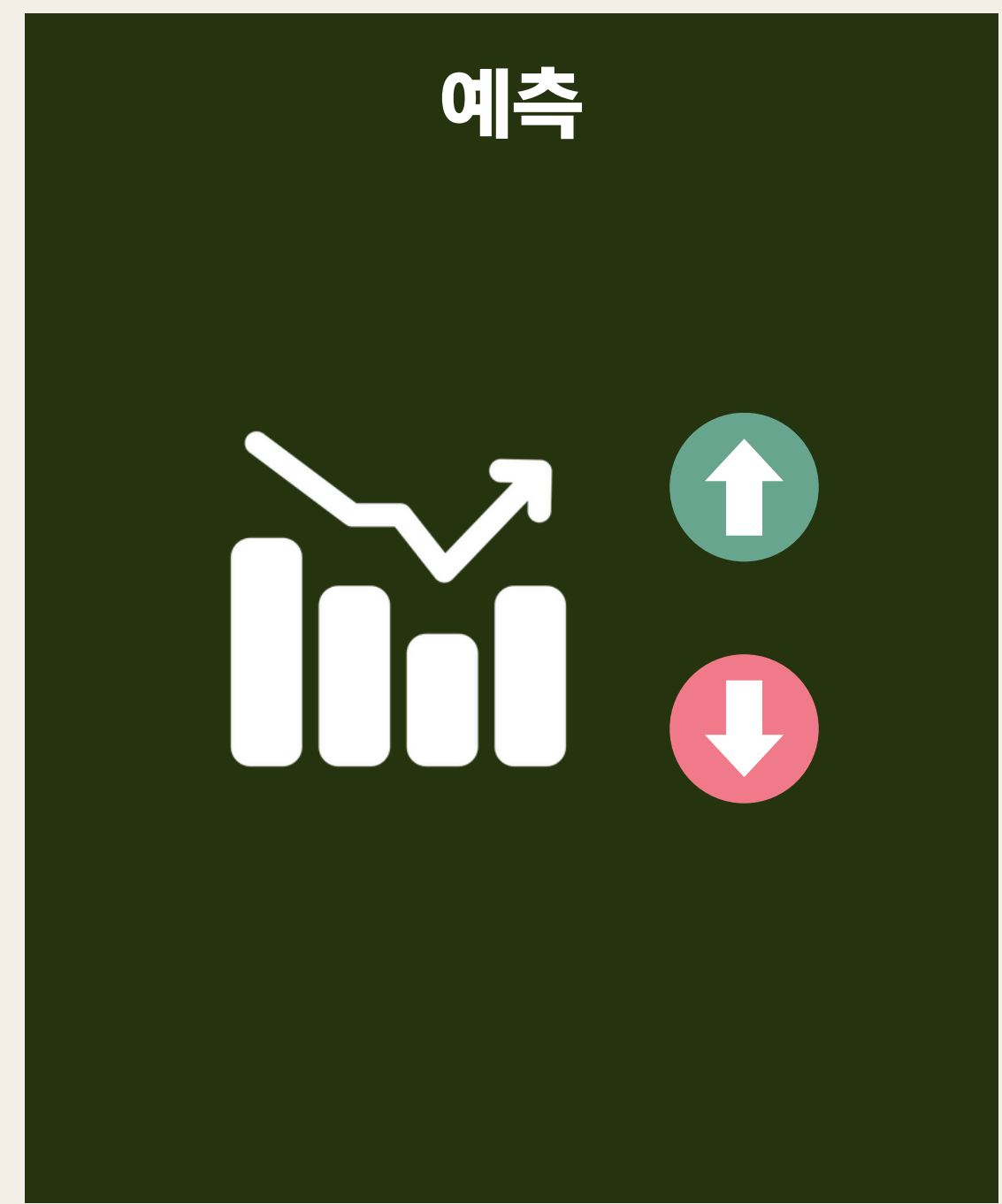
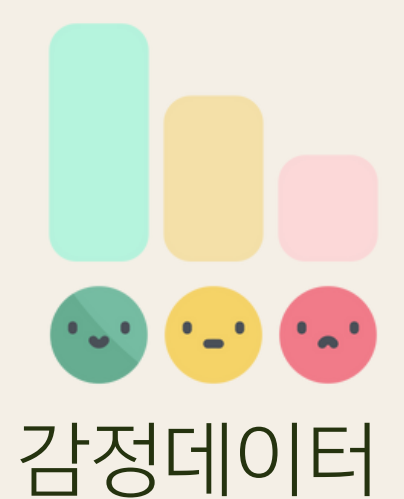
기술적 분석



멀티 모달 학습



+



프로젝트 흐름

1 뉴스 데이터 수집

- 제약사항
- 해결방법

2 뉴스 데이터의 감정 예측 모델 구축

- 모델 학습
- 데이터 증강
- Fine-Tuning
- 감정 예측

3 뉴스 감정 점수와 주가 데이터의 상관관계 분석

- 상관관계 시각화
- WordCloud 시각화

뉴스 데이터 수집

제약사항

- 1.네이버 뉴스 크롤링 명시적 금지
- 2.제한 된 API요청

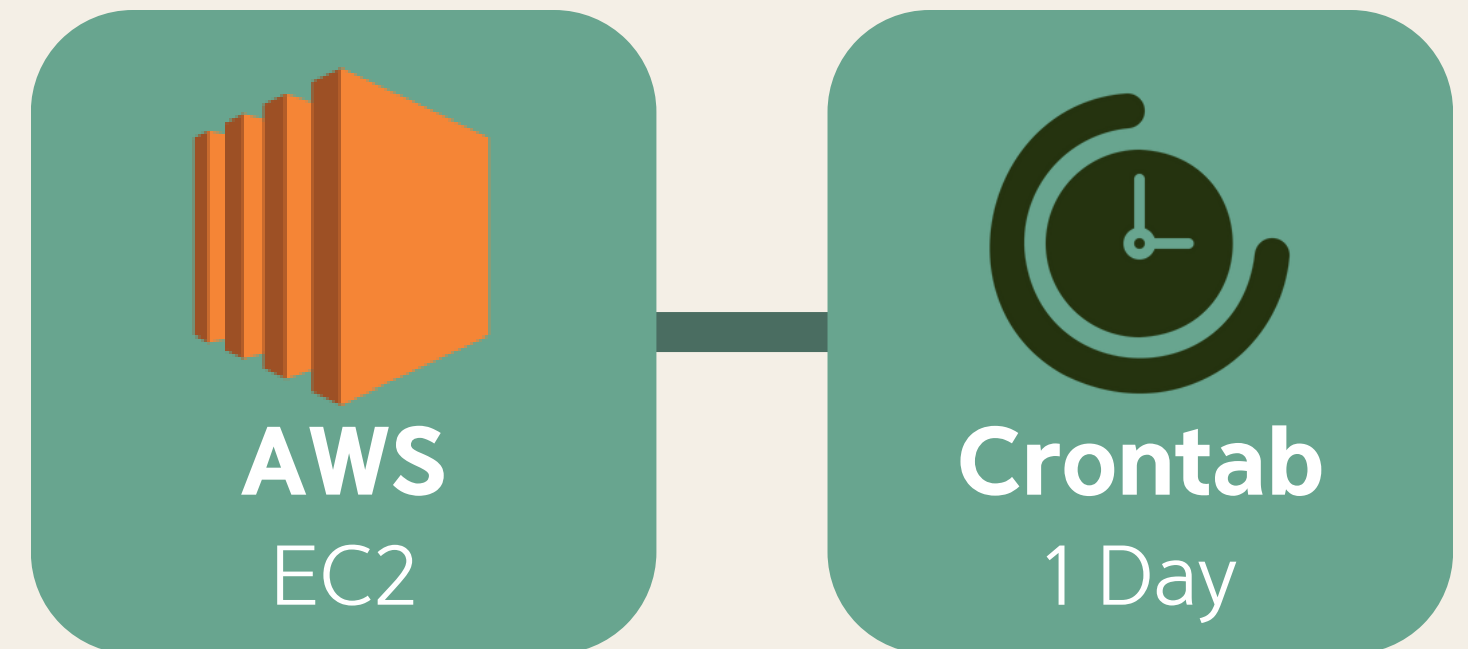
User-agent: *
Disallow: /

NAVER OpenAPI

*Max: 1000건

해결방법

AWS EC2 인스턴스에서 크론탭(Crontab)을 사용하여
API요청 스케줄러를 통한 데이터셋 확보



뉴스 데이터 감정 예측 모델 구축

KLUE/BERT-base

- 구글의 BERT 모델의 아키텍처
- 한국어 자연어 처리(NLP)에 특화 된 모델
- Transformer 구조의 양방향 인코더 - 문맥을 양방향으로도 이해 가능

GPT

어제 카페 갔었어 거기 사람 많더라



BEAT




어제 카페 갔었어 사람 많더라



뉴스 데이터 감정 예측 모델 구축

금융 뉴스 문장 감성 분석 데이터셋




Github: [ukairia777/finance_sentiment_corpus](https://github.com/ukairia777/finance_sentiment_corpus)

labels	sentence	kor_sentence
 neutral	neutral,"According to Gran, the company has no plans to move all production to Russia, although that is where the company is growing."	"Gran에 따르면, 그 회사는 회사가 성장하고 있는 곳이지만, 모든 생산을 러시아로 옮길 계획이 없다고 한다."
 negative	"The international electronic industry company Elcoteq has laid off tens of employees from its Tallinn facility... "	,"국제 전자산업 회사인 엘코텍은 탈린 공장에서 수십 명의 직원을 해고했으며... "
 positive	"With the new production plant the company would increase its capacity to meet the expected increase in demand and... "	"새로운 생산공장으로 인해 회사는 예상되는 수요 증가를 충족시킬 수 있는 능력을 증가시키고... "

뉴스 데이터 감정 예측 모델 구축

데이터 전처리

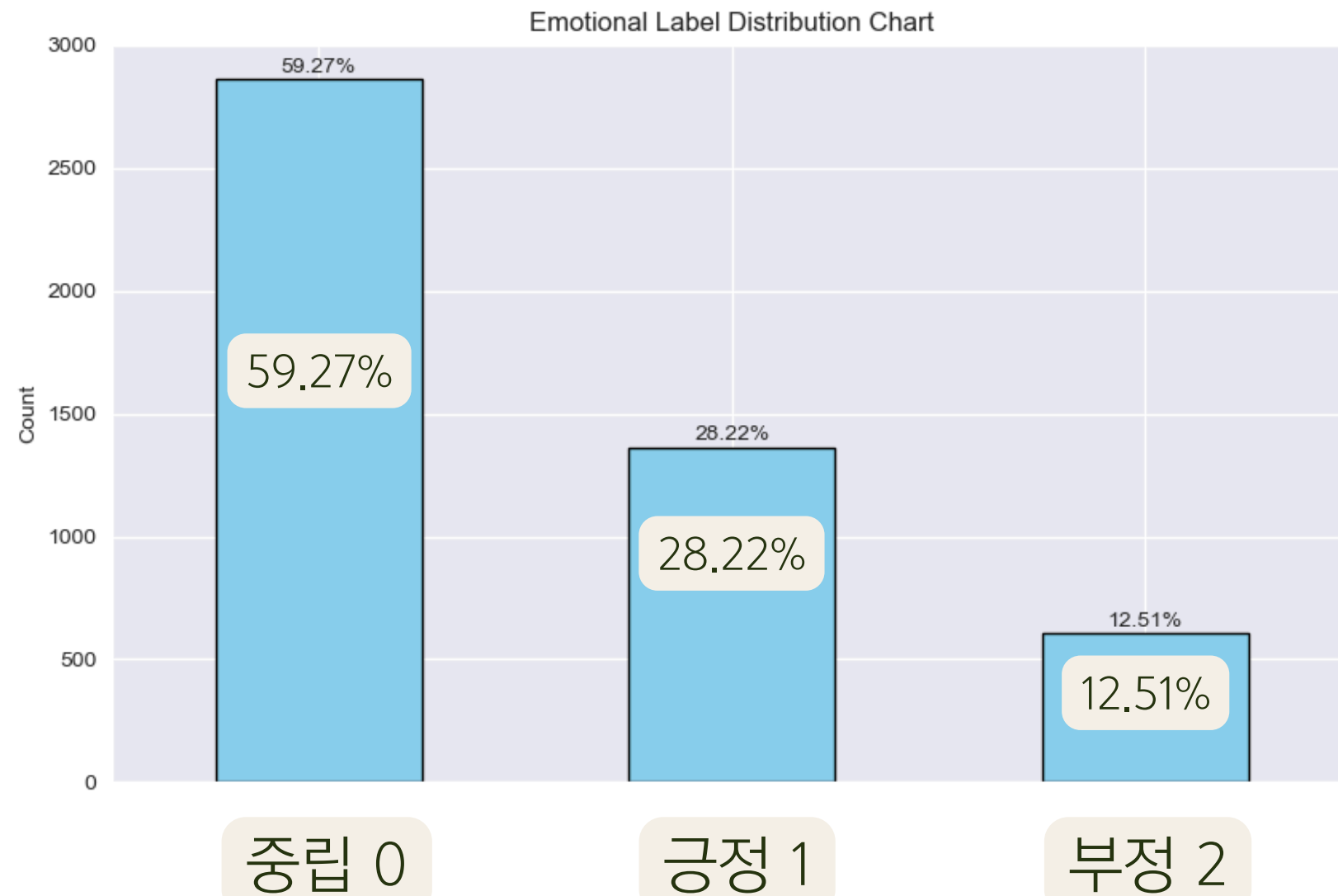
레이블 변환, 영문 데이터 제거

labels	sentence	kor_sentence
 0	neutral, "According to Gran, the company has no plans to move all production to Russia, although that is where the company is growing."	"Gran에 따르면, 그 회사는 회사가 성장하고 있는 곳이지만, 모든 생산을 러시아로 옮길 계획이 없다고 한다."
 1	"The international electronic industry company Elcoteq has laid off tens of employees from its Tallinn facility..."	,"국제 전자산업 회사인 엘코텍은 탈린 공장에서 수십 명의 직원을 해고했으며..."
 2	"With the new production plant the company would increase its capacity to meet the expected increase in demand and..."	"새로운 생산공장으로 인해 회사는 예상되는 수요 증가를 충족시킬 수 있는 능력을 증가시키고..."

뉴스 데이터 감정 예측 모델 구축

EDA(탐색적 데이터 분석)

데이터 감정 비율 불균형 - 클래스 가중치 설정



```
# 클래스 가중치 계산
class_weights = compute_class_weight(
    class_weight='balanced',
    classes=np.unique(train_data['labels']),
    y=train_data['labels']
)
```

0: 0.562254259501966,
1: 1.1938775510204083,
2: 2.6052631578947367

뉴스 데이터 감정 예측 모델 구축

토큰나이저 | BertTokenizer

BERT 모델 입력에 맞게 토큰화 및 정수 인코딩
최대 길이 128 설정, 패딩 및 트리밍 적용

문장 1: 올리페카 칼라스부오가 이사회 부회장으로 선출되었다.

토큰: ['올리', '##페', '##카', '칼라', '##스', '##부', '##오', '##가', '이사회', '부회장', '##으로', '선출', '##되', '##었', '##다', '.']

정수 인코딩: [2 4705 2743 2127 17278 2255 2144 2168 2116 9314 5875 6233
6940 2496 2359 2062 18 3]

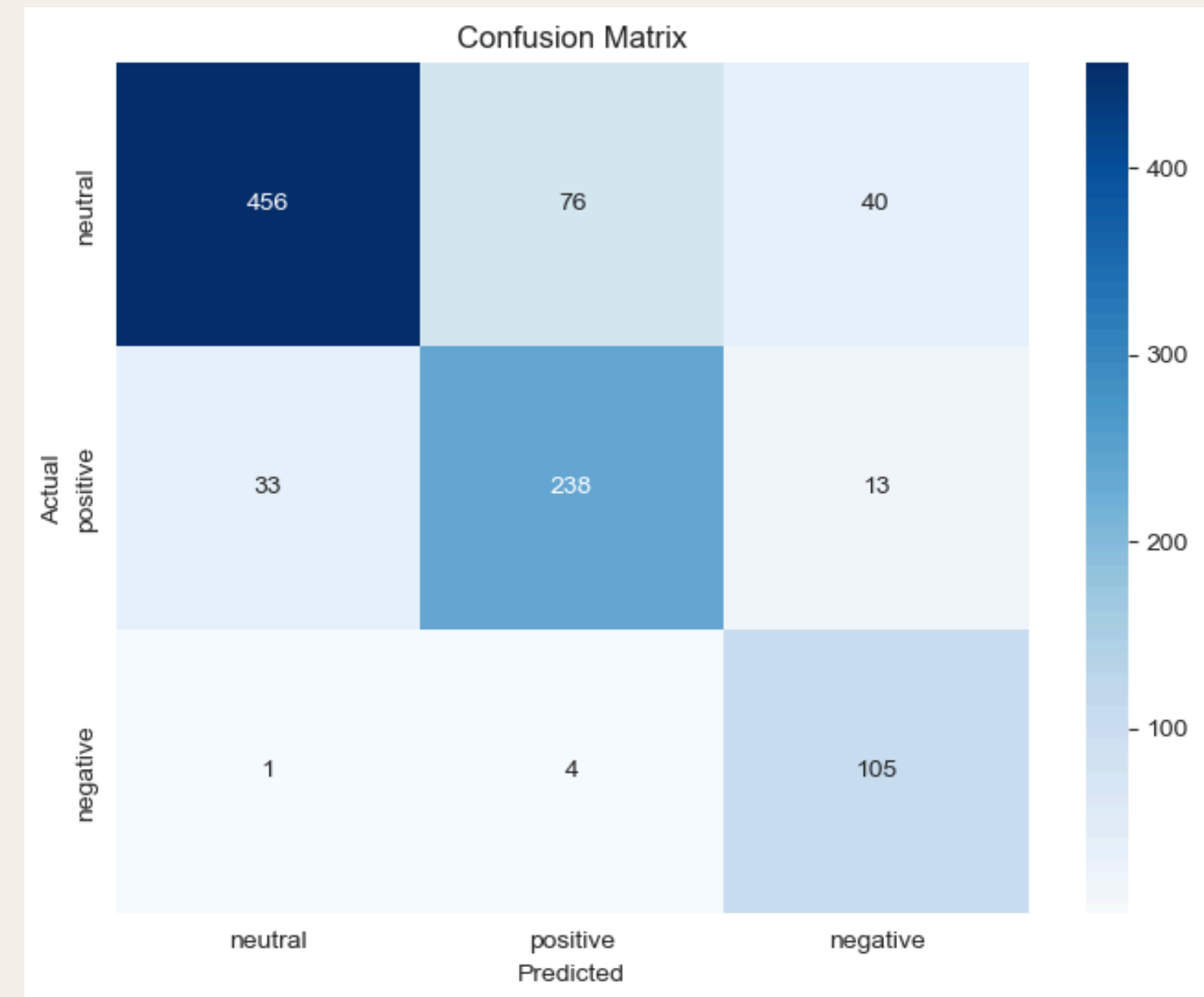
뉴스 데이터 감정 예측 모델 구축

학습 결과

검증 데이터에 대한 모델 평가

- 손실: **0.4612**
- 정확도: **0.8271**

혼돈 행렬 시각화



뉴스 데이터 감정 예측 모델 구축

다수의 주체가 담긴 문장 이해

뉴스 데이터 특성 상 문장에 다른 기업과의 비교와 같은 다수의 주체가 담긴 문장이 많으며, 해당 문장에 대한 **감정 예측 성능이 저하**되는 문제가 발견

문장	실제 감정	예측 감정
삼성전자가 실적 발표에서 긍정적인 결과를 보였으나, SK하이닉스는 부진했다...	부정	중립
SK하이닉스는 실적 상승을 기록했지만 삼성전자는 다소 실망스러운 실적을 발표했다...	긍정	부정
SK하이닉스의 실적 발표 발표에서 클라우드 부문의 성장 둔화가 우려를 불러일으켰다, 아마존은 반도체 부문 시장에서 강세를 보였다...	부정	긍정

뉴스 데이터 감정 예측 모델 구축

다수의 주체가 담긴 문장 이해 / Fine-Tuning

기업명 토큰 추가

문장 1: 삼성전자는 기대 이상의 실적을 발표했지만
<SK하이닉스>는 오히려 부진한 결과를 보였다

토큰: ['삼성전자', '##는', '기대', '이상', '##의', '실적', '##을', '발표', '##했', '##지만',
'<SK하이닉스>', '는', '오히려', '부진', '##한', '결과', '##를', '보였', '##다']

K-TACC을 사용한 데이터 증강 (94 => 536)

문장 1-1: 삼성전자는 **당초 기대치** 이상의 실적을 발표했지만
<SK하이닉스>는 **반대로 어느정도로** 부진한 결과를 보였다

문장 1-2: 삼성전자는 **기대치** 이상의 깜짝 실적을 발표했지만
<SK하이닉스>는 오히려 부진한 **실적 결과**를 보였다

뉴스 데이터 감정 예측 모델 구축

다수의 주체가 담긴 문장 이해

Fine-Tuning을 추가로 진행한 결과,
다수의 주체가 포함된 문장에서 **감정 예측 성능이 유의미하게 개선**

Fine-tuning 이후 성능

문장	실제 감 정	예측 감 정
삼성전자가 실적 발표에서 긍정적인 결과를 보였으나, SK하이닉스는 부진했다...	부정	부정
SK하이닉스는 실적 상승을 기록했지만 삼성전자는 다소 실망스러운 실적을 발표했다...	긍정	긍정
SK하이닉스의 실적 발표 발표에서 클라우드 부문의 성장 둔화가 우려를 불러일으켰다, 아마존은 반도체 부문 시장에서 강세를 보였다...	부정	부정

뉴스 데이터 감정 예측 모델 구축

최신 네이버 뉴스 데이터 감정 점수 예측

2024.11.04 ~ 2024.11.27 (8681건)

	pub_date	description	sentiment
0	2024-11-04 09:00:00	코스피 시가총액 상위 10개 종목 중에서는 <SK하이닉스>000660 121 삼성전...	1
1	2024-11-04 09:00:00	시가총액 상위 종목 중 삼성전자137 <SK하이닉스>170 등 반도체주와 현대차16...	0
2	2024-11-04 09:00:00	KB증권은 지난 달 23일부터 AI 실적속보를 통해 삼성전자 <SK하이닉스> 현대차...	0

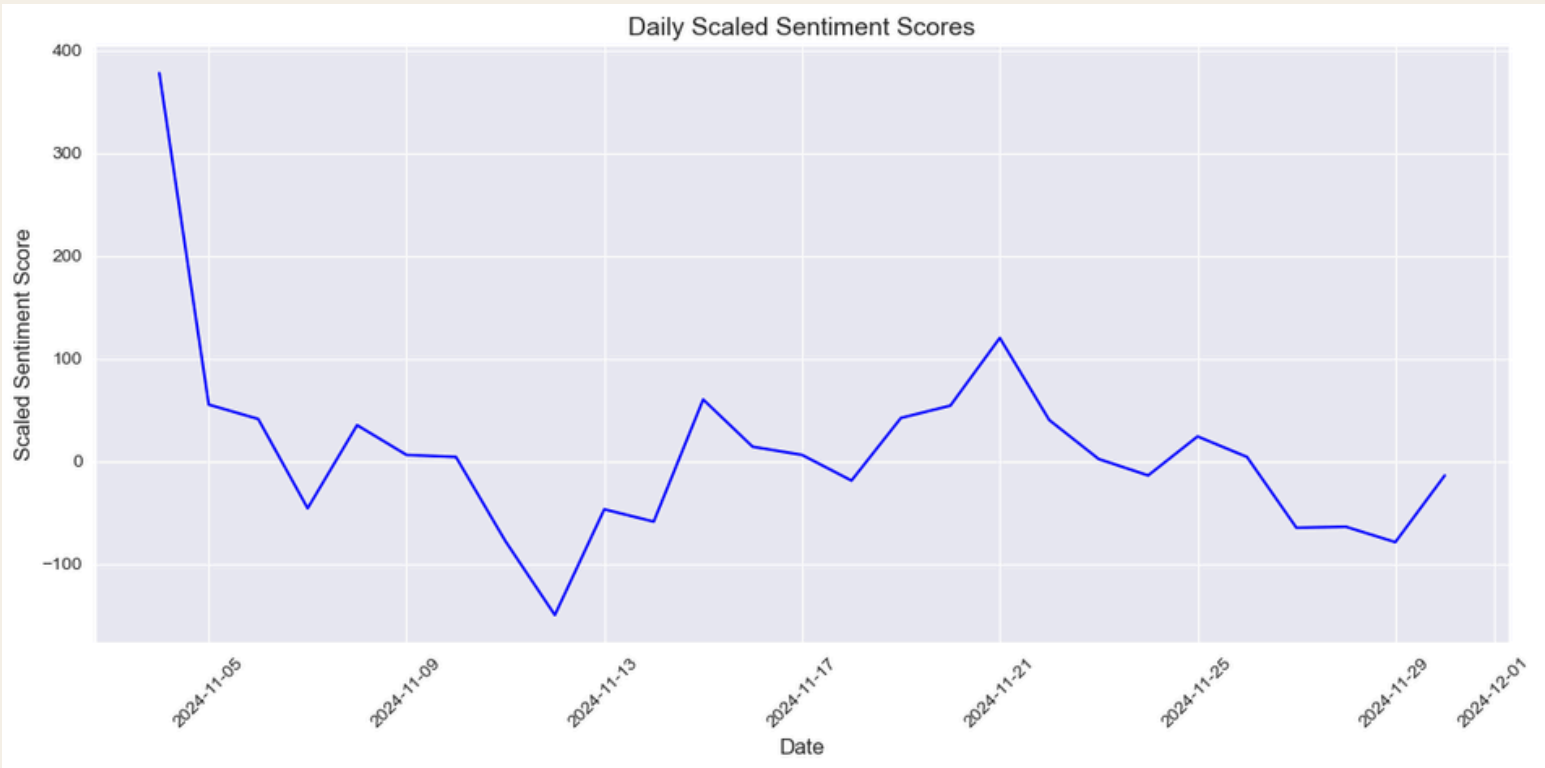
감정 레이블을 점수로 매핑

긍정(1) → +1

부정(2) → -1

중립(0) → 0

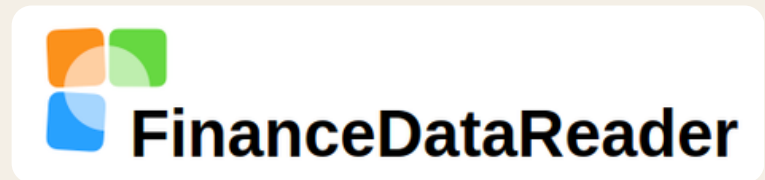
	day	sentiment_score
0	2024-11-04	378
1	2024-11-05	55
2	2024-11-06	41
3	2024-11-07	-46
4	2024-11-08	35



감정 점수와 주가의 상관관계 분석

데이터 전처리 / 상관관계 분석 준비

FinanceDataReader 라이브러리로 SK하이닉스의 종가 데이터 추출



```
# SK Hynix(000660) 주가 데이터 불러오기
stock_prices = fdr.DataReader('000660', '2024-11-04', '2024-11-29')['Close']
```

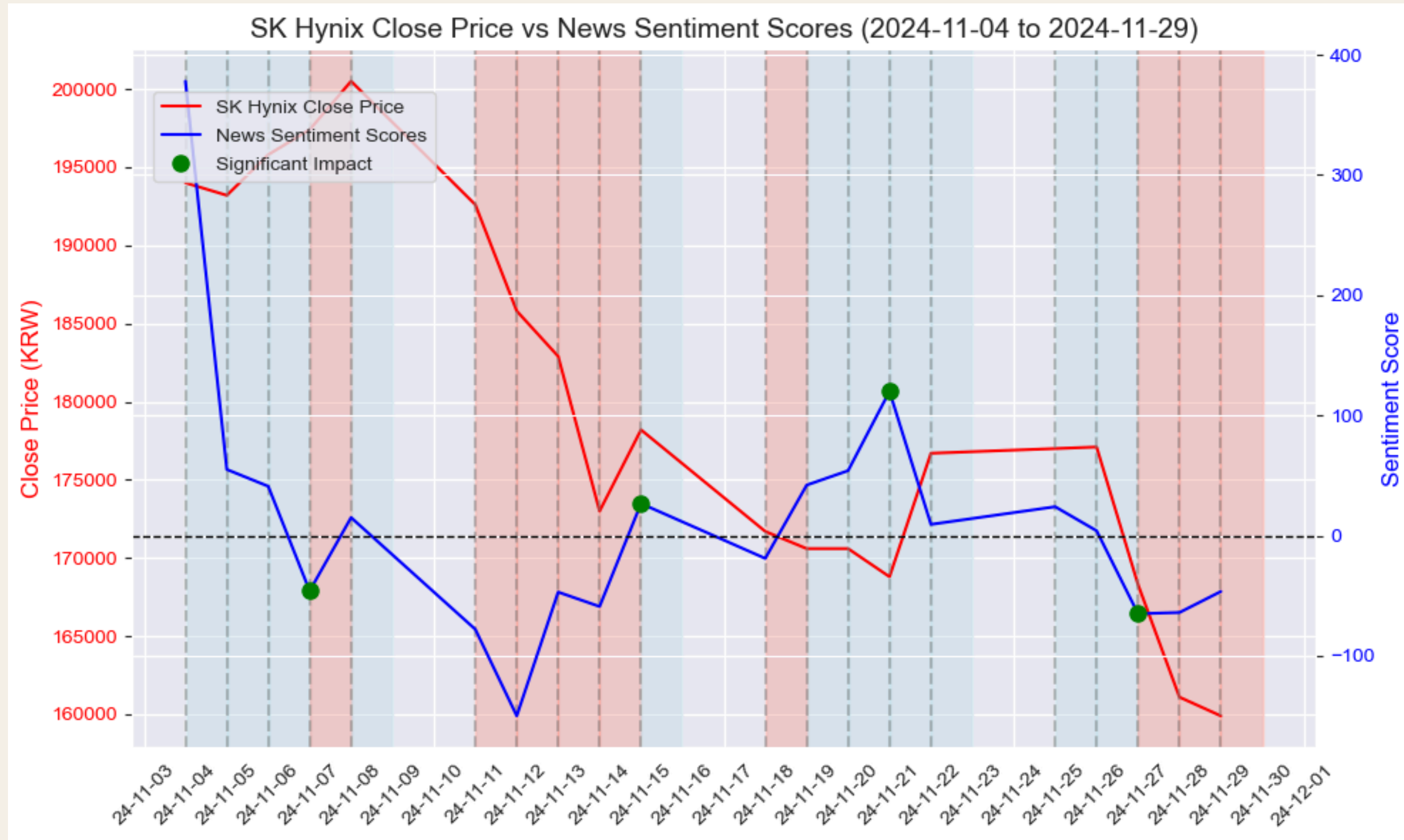
종가, 감정 점수 데이터 병합

```
주가 데이터 :
Date
2024-11-04    194000
2024-11-05    193200
2024-11-06    195800
2024-11-07    197400
2024-11-08    200500
Name: Close, dtype: int64
```

	Close	sentiment_score
2024-11-04	194000	378.0
2024-11-05	193200	55.0
2024-11-06	195800	41.0
2024-11-07	197400	-46.0
2024-11-08	200500	15.0

감정 점수와 주가의 상관관계 분석

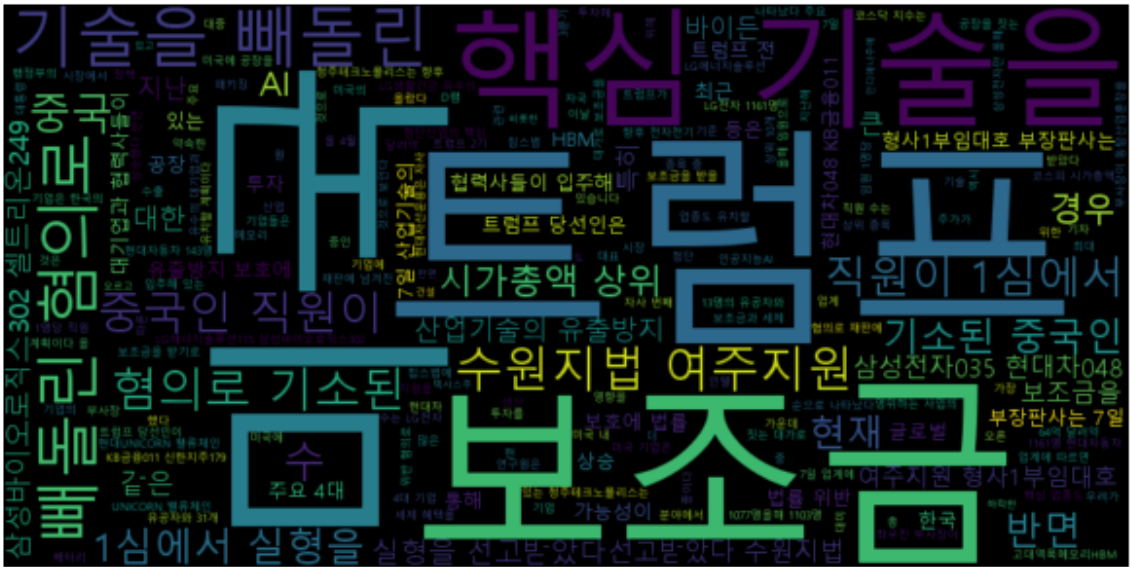
상관관계 분석 시각화



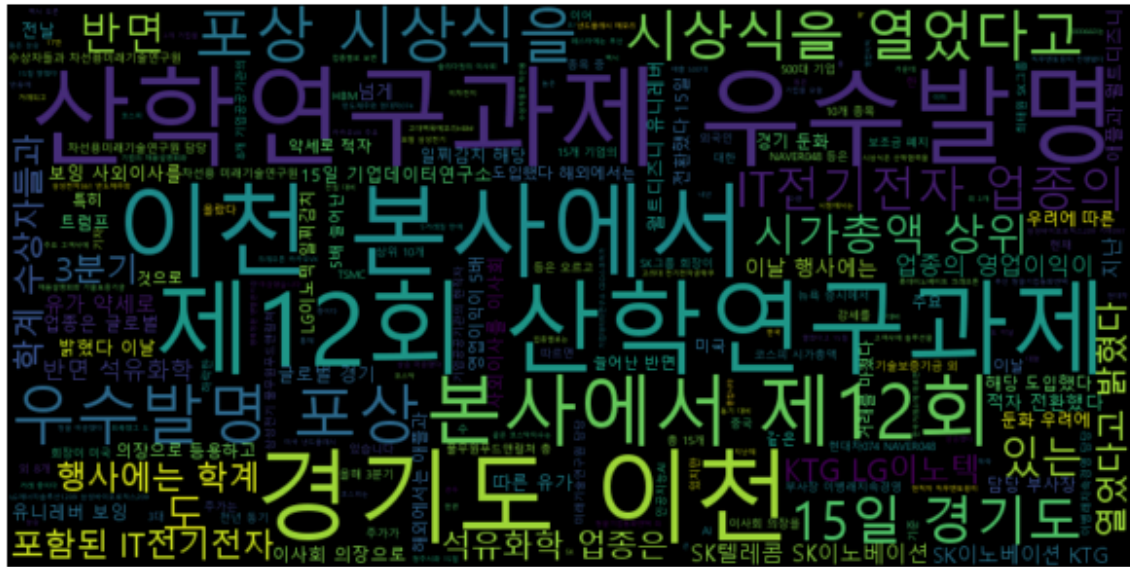
감정 점수와 주가의 상관관계 분석

상관관계 분석 시각화 – WordCloud

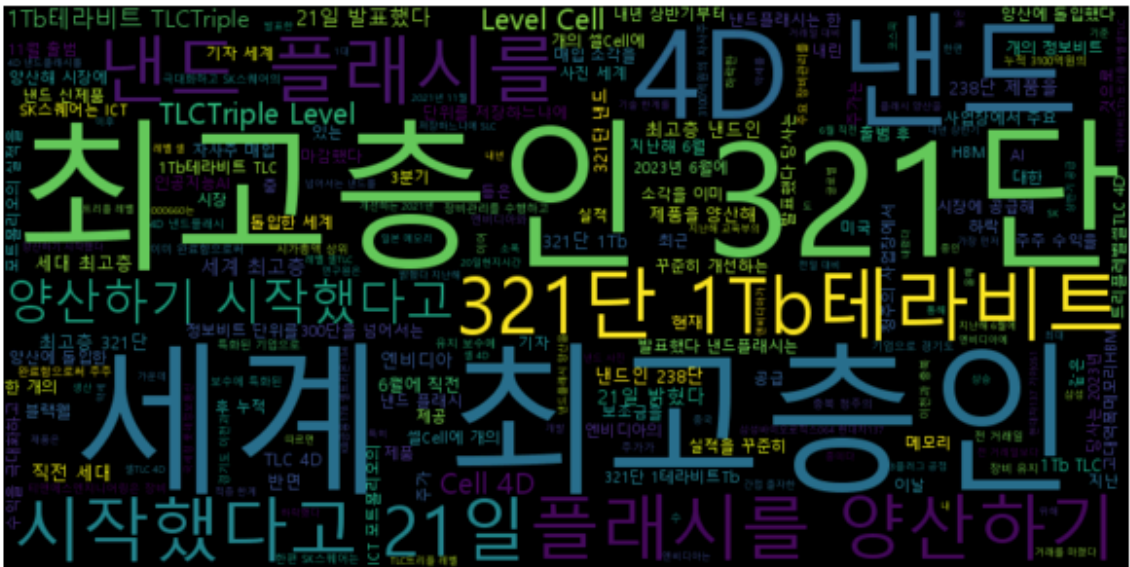
Word Cloud for 2024-11-07 (Sentiment Score: -46.00)



Word Cloud for 2024-11-15 (Sentiment Score: 26.67)



Word Cloud for 2024-11-21 (Sentiment Score: 120.00)



Word Cloud for 2024-11-27 (Sentiment Score: -65.00)



결과 해석

1 뉴스 감정 예측의 정확도

뉴스 감정 점수가 주가 데이터와 어느정도 같은 흐름을 보임,
11.07의 부정적인 뉴스 다음날 주가 하락, 11.21의 긍정적인 뉴스 다음날 주가 상승

2 뉴스 감정과 주가 데이터의 상관관계

뉴스 감정 점수와 주가 간의 상관계수 계산 값은 0.2로 다소 약한 상관관계,
두 변수 간의 관계는 비선형적, 뉴스 특성 상 서로 반대적인 상관관계가 나타날 가능성 이 큼

3 분석의 한계

수집된 뉴스 데이터의 양이 적고 분석 기간이 짧아 정확한 상관관계를 도출하는 데 어려움,
데이터의 양이 충분했다면 예측까지 진행할 수 있었을 거 같아 아쉬움

감사합니다

End of Document

2024.12.05

20204434 왕건

