빅데이터 처리
프로젝트 결과 보고서

- 뉴스 감정 예측 및 주가와의 상관관계 분석 -

제출일	2024. 12. 01
과목명	빅데이터 처리
담당교수	민정혜
학번	202044034
이름	왕 건

목차

서론

1. 프로젝트 배경 및 목적

1) 프로젝트 배경

현대 금융 시장에서는 기업의 주가가 경제적 요인뿐만 아니라 사회적, 정서적 요인에도 큰 영향을 받습니다. 특히, 온라인 뉴스는 투자자들의 심리를 형성하고, 이에 따라 주가 변동에 중요한 역할을 합니다. 대형 기술 기업인 SK 하이닉스의 경우, 실적 발표나 정부 규제 외에도 대중 매체에서 다루는 뉴스가 주가에 즉각적인 영향을 미칠 수 있습니다.

감정 분석은 이러한 뉴스 데이터를 정량적으로 평가하고, 긍정적 또는 부정적인 감정이 주가에 미치는 영향을 측정하는 유용한 도구로 자리잡고 있습니다. 본 프로젝트는 SK 하이닉스를 대상으로 뉴스 감정 점수와 주가 간의 상관관계를 분석하여, 투자 의사결정 과정에 도움이 될 수 있는 인사이트를 도출하고자 합니다.

2) 프로젝트 목적

뉴스 감정 예측

본 프로젝트는 뉴스 데이터를 기반으로 감정을 예측하는 모델을 개발하여, 뉴스가 포함하는 감정을 정확히 분석하고 예측하는 것을 목표로 합니다. 이를 위해, KLUE BERT 모델을 뉴스 문장에 맞게 Fine-Tuning하여, 네이버 API로 수집한 뉴스 데이터를 정확하게 분류할 수 있는 감정 예측 모델을 구축했습니다.

뉴스 감정과 주가의 상관관계 규명

뉴스 감정 점수가 주가 변동성에 미치는 영향을 분석하고, 감정 점수가 높은 날과 낮은 날의 주가 변화 패턴을 탐색합니다. 이를 위해, 수집된 뉴스의 감정 점수를 기반으로 주가 변동성을 분석하고, 뉴스 감정이 주가에 미치는 영향을 파악했습니다. 특히, 긍정적 또는 부정적인 뉴스 감정이 주가 상승 또는 하락과 어떠한 관계를 가지는지 파악하는 데 중점을 두어, 뉴스 감정이 주가 예측에 미치는 영향을 명확히 규명하고자 했습니다.

예측 가능성 탐색

뉴스 감정 데이터를 활용해 주가 변동성 예측 가능성을 탐구하고, 이를 기반으로 향후 투자 전략을 설계할 수 있는 가능성을 검토합니다.

뉴스 데이터 예측 모델 구축

1. 뉴스 데이터 수집

1) 수집 방법 개요

본 프로젝트에서는 SK하이닉스와 관련된 뉴스 데이터를 수집하기 위해 네이버 뉴스 API를 활용했습니다. 네이버의 robots.txt 파일에 따라 동적인 웹 크롤링이 명시적으로 금지된 상태였으므로, API를 통해 합법적이고 윤리적인 방식으로 데이터를 수집했습니다.

2) 제약 사항 및 해결 방법

데이터 수집 제한

네이버 뉴스 API는 한 번의 호출로 최대 1000개의 최신 뉴스만 반환하며,
과거 데이터를 수집하는 기능은 제공하지 않습니다.

해결 방안

이를 해결하기 위해 AWS EC2 인스턴스에서 크론탭(Crontab)을 사용하여 2024년 11월 3일부터 매일 데이터를 자동으로 수집하도록 설정했습니다. 이를 통해 지속적으로 최신 뉴스 데이터를 축적할 수 있었습니다.

3) 크롤링 프로그램 상세 설명

데이터 수집

네이버 뉴스 API를 통해 'SK 하이닉스' 관련 뉴스 기사를 검색하고,
뉴스 제목, 요약(description), 게시 날짜를 수집했습니다.

데이터 정제

수집된 데이터에서 HTML 태그와 특수문자를 제거하여 텍스트를 정제했습니다.

데이터 저장

수집된 뉴스 데이터는 CSV 파일로 저장되며, 중복된 데이터는 제외하고 추가하도록 구현하여 데이터의 정확성을 유지했습니다.

스케줄링

크론탭을 사용하여 매일 한 번씩 뉴스 데이터를 자동으로 축적했습니다.

5) Crontab 설정

AWS EC2 인스턴스에서 크론탭을 아래와 같이 설정하여
자동화된 뉴스 수집 작업을 진행했습니다.

```
$ crontab -e
```

```
0 0 * * * /usr/bin/python3
/home/ec2-user/stock_sentiment_predictor/news_crawler.py >>
/home/ec2-user/stock_sentiment_predictor/logs/news_crawl.log 2>&1
```

설정 내용

매일 자정에 news_crawler.py 스크립트를 실행하며,
실행 로그는 news_crawl.log 파일에 저장됩니다.

6) 수집 결과

수집 기간: 2024년 11월 4일부터 지속
저장 위치: AWS EC2 인스턴스
데이터 형식: CSV 파일
컬럼: 날짜(pubDate), 제목(title), 요약(description)

	pub_date	title	description
0	2024_11_04_09:00	코스피 장 초반 2540선 강보합코스닥도 상승세개?	코스피 시가총액 상위 10개 종목 중에서는 SK하이닉스000660 121 삼성전자005930 051
1	2024_11_04_09:00	코스피 美 대선 앞두고 2550대 강보합코스닥도 돌	시가총액 상위 종목 중 삼성전자137 SK하이닉스170 등 반도체주와 현대차164 기아233 등
2	2024_11_04_09:00	KB證 생성형 AI 기반 보고서 발간	KB증권은 지난 달 23일부터 AI 실적속보를 통해 삼성전자 SK하이닉스 현대차 HD현대건설기
	pub_date	title	description
9693	2024_11_30_13:00	예상 뛰어넘더니 개미들 열광3조 뭉칫돈 우르르 돌	이 종목은 SK하이닉스103 일본 반도체 장비주 어드반테스트96 대만 TSMC86 등 아시아 종
9692	2024_11_30_12:00	희망브리지 폭설 이재민 긴급 구호물품 5000점 지	트레이닝복세면도구속옷 등으로 구성된 구호키트와 대피소 칸막이는 SK하이닉스의 기부로 제
9691	2024_11_30_12:00	美 추가관세와 中 기술추격에 내몰린 한국 저성장	미국에 공장을 지으며 보조금을 받기로 한 삼성전자SK하이닉스 등 반도체 기업의 사업계획

이와 같은 방식으로 체계적으로 뉴스 데이터를 수집하여, 본 프로젝트에서 필요로 하는 감정 분석 및 주가와의 상관관계 분석에 활용 가능한 데이터셋을 확보했습니다.

뉴스 데이터 예측 모델 구축

2. 뉴스 감정 예측 모델 학습

1) 데이터 수집 및 전처리

데이터 수집

총 데이터: 4,846건

영문 데이터: 16명의 전문 연구자가 수동으로 라벨링하여 생성된 데이터 (Finance Phrase Bank (Moal et al., 2014))

한국어 번역 데이터: 영문 데이터를 한국어로 변역한 후 검수를 거쳐 데이터 추가 (Finance Sentiment Corpus)

레이블 분류: 긍정(Positive), 중립(Neutral), 부정(Negative)

	labels	sentence	kor_sentence
0	neutral	According to Gran, the company has no plans to...	Gran에 따르면, 그 회사는 회사가 성장하고 있는 곳이지만, 모든 생산을 러
1	neutral	Technopolis plans to develop in stages an area...	테크노폴리스는 컴퓨터 기술과 통신 분야에서 일하는 회사들을 유치하기 위해
2	negative	The international electronic industry company ...	국제 전자산업 회사인 엘코텍은 탈린 공장에서 수십 명의 직원을 해고했으며,

주요 전처리 작업

영어 뉴스 제거: sentence 열(영문)을 제거

감정 레이블 변환: 텍스트 레이블(neutral, positive, negative) → 숫자 라벨(0, 1, 2)

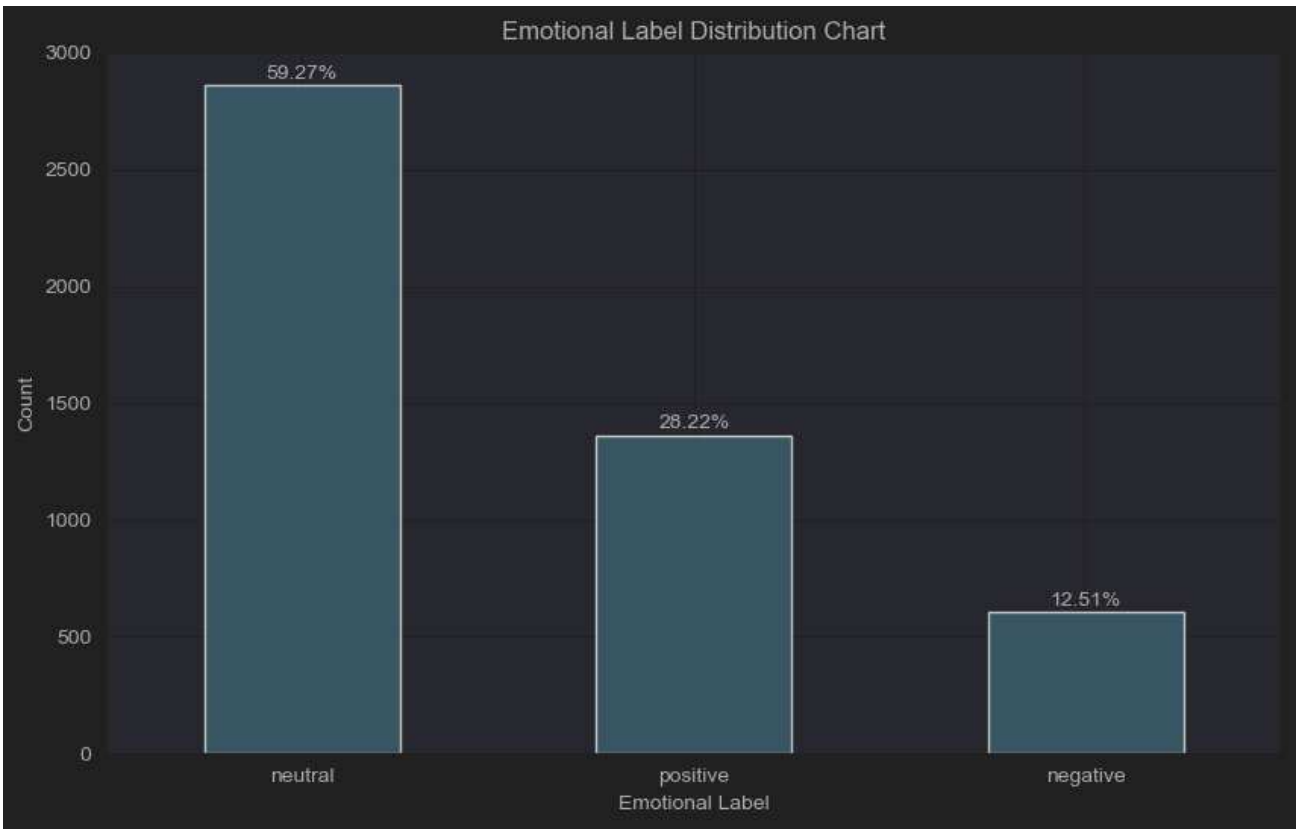
결측값 제거 및 중복 제거: kor_sentence 열 기준으로 중복된 데이터 제거

	labels	kor_sentence
0	0	Gran에 따르면, 그 회사는 회사가 성장하고 있는 곳이지만, 모든 생
1	0	테크노폴리스는 컴퓨터 기술과 통신 분야에서 일하는 회사들을 유치하
2	2	국제 전자산업 회사인 엘코텍은 탈린 공장에서 수십 명의 직원을 해고

2) EDA(탐색적 데이터 분석)

데이터셋 내 감정 레이블 분포 시각화

시각화 결과에 따라 중립(Neutral)이 가장 많고, 부정(Negative)이 가장 적어, 감정 레이블이 불균형적으로 분포되어 있습니다. 이를 해결하기 위해 sklearn의 compute_class_weight를 사용하여 클래스별 가중치를 계산하여 적용하였습니다.




```
{0: 0.562254259501966, 1: 1.1938775510204083, 2: 2.6052631578947367}
```

3) 모델 및 토크나이저 설정

사용 모델: KLUE/BERT-Base

한국어의 문법적 특성을 반영한 BERT 모델.

토크나이저: 뉴스 텍스트를 BERT 입력에 맞게 토큰화 및 정수 인코딩.

최대 길이 128로 설정, 패딩 및 트리밍 적용

```
데이터 전처리 과정:
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]
```

4) 훈련 데이터 구성

데이터 분리: 훈련 데이터 80%, 검증 데이터 20%로 분리.

클래스 가중치 적용

5) 모델 학습

Fine-Tuning: KLUE/BERT-Base 모델의 사전 학습 가중치를 사용하여 감정 데이터셋에 맞게 미세 조정.

학습 파라미터: Batch Size: 16

Learning Rate: 3e-5

Epoch: 최대 10

Dropout 설정: hidden_dropout_prob=0.2, attention_probs_dropout_prob=0.3

EarlyStopping: 2회 이상 검증 손실(val_loss)이 개선되지 않으면 학습 중단.

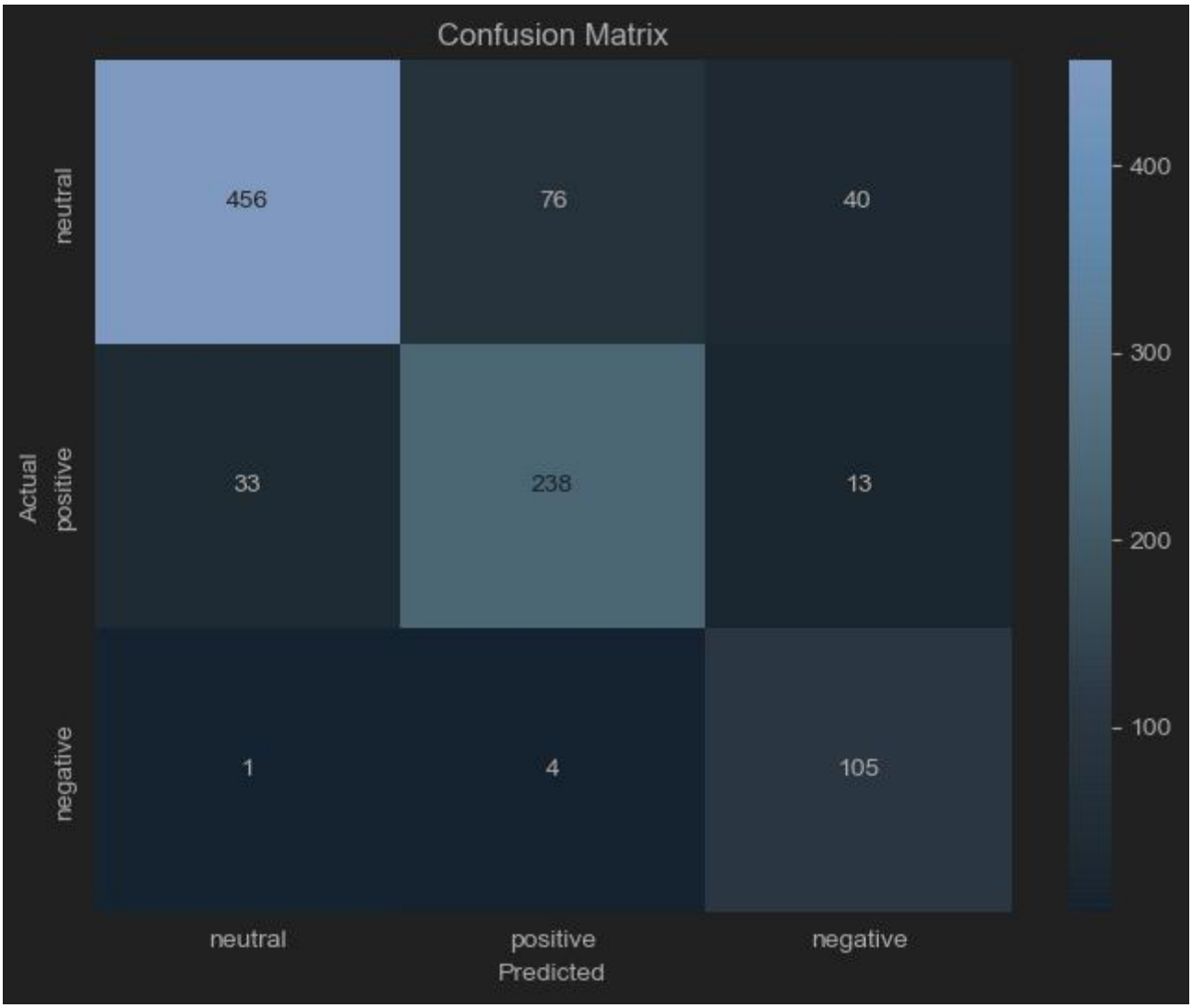
최종 결과:

```
Epoch 3/10
242/242 [==============================] - 90s 370ms/step - loss: 0.2560 - accuracy: 0.8840
  - val_loss: 0.4612 - val_accuracy: 0.8271
```

6) 결과 시각화

검증 데이터 예측 결과를 바탕으로 혼동 행렬(confusion matrix) 생성.

중립(Neutral) 클래스에서 가장 높은 정확도를 보였으며, 긍정/부정 클래스 간 혼동 발생.



7) 평가 및 저장

검증 데이터에 대한 손실: 0.4612

검증 데이터에 대한 정확도: 82.71%

학습 완료 후 모델 및 토크나이저 저장:

모델: ../sentiment_analysis_model

토크나이저: ../sentiment_analysis_model

뉴스 데이터 예측 모델 구축

3. 다수의 주체가 포함된 데이터 증강 및 Fine-Tuning

1) 학습시킨 뉴스 감정 예측 모델의 문제점 확인 및 해결 방안

문제점

뉴스 데이터 특성상 하나의 문장에 다수의 주체가 포함된 경우,감정 예측 성능이 저하되는 문제가 발견되었습니다.

해결 방안

이를 해결하기 위해 학습 데이터에서 여러 주체가 포함된 문장을 선별하여 Fine-Tuning 데이터셋을 구축했습니다. 또한, 기업명을 <SK하이닉스>의 고유 토큰으로 변환하고 모델에 새로운 토큰으로 추가했습니다. 이를 통해 모델이 특정 기업명에 따른 문맥을 더 잘 이해할 수 있도록 유도했습니다.

Fine-tuning 이전 성능

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, 아마존는 반도체 부문 시장에서 강세를 보였다...	부정	긍정

2) 데이터 증강

random_masking_insertion과 adverb_gloss_replacement로 텍스트를 변형.
원본 데이터에 대해 다섯 번 증강하며, 중복 제거와 모든 문자의"SK하이닉스" 기업명을 <SK하이닉스>로 변경.

데이터 증강 결과: 94 => 536

3) 데이터 전처리 및 모델 준비

위에서 금융 뉴스에 맞게학습시킨 KLUE BERT 모델을 사용.

<SK하이닉스>을 새로운 토큰을 추가하고 모델 임베딩 레이어를 업데이트.

```
데이터 전처리 과정:
문장 1: 삼성전자는 기대 이상의 실적을 발표했지만 <SK하이닉스>는 오히려 적기는 하지만 어느 정도로 부진한 결과를
 보였다
토큰: ['삼성전자', '##는', '기대', '이상', '##의', '실적', '##을', '발표', '##했', '##지만', '<SK하이
닉스>', '는', '오히려', '적기', '##는', '하지만', '어느', '정도', '##로', '부진', '##한', '결과', '##
를', '보였', '##다']
정수 인코딩: [    2  4798  2259  3869  3658  2079  4759  2069  3913  2371  3683 32000
   793  4312 25697  2259  3696  3875  3681  2200  6043  2470  3731  2138
  4278  2062     3]
```

4) 모델 학습

옵티마이저로 Adam을 사용하고 학습률은 1e-5로 설정했으며, 과적합을 방지하기 위해 EarlyStopping을 적용. 학습된 모델과 토크나이저는 별도로 저장하여 이후 분석 및 예측에 활용할 수 있도록 준비.

최종 결과:

```
Epoch 5/5
29/29 [==============================] - 12s 404ms/step - loss: 0.1114 - accuracy: 0.9735
- val_loss: 0.0760 - val_accuracy: 0.9737
```

5) 학습 결과

Fine-Tuning을 추가로 진행한 결과, 다수의 주체가 포함된 문장에서 감정 예측 성능이 유의미하게 개선되었습니다. 모델이 기업명에 따른 문맥을 보다 정확히 인식하고 감정을 예측할 수 있도록 학습되었습니다.

Fine-tuning 이후 성능

문장	실제 감정	예측 감정
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, 아마존는 반도체 부문 시장에서 강세를 보였다...	부정	부정

뉴스 데이터 예측 모델 구축

4. 뉴스 데이터 감정 점수 예측 및 레이블링

1) 작업 개요

뉴스 데이터 로드 및 전처리

네이버 API를 사용하여 수집한 SK하이닉스 관련 뉴스 데이터의 결측값 제거와 데이터 포맷을 정리했습니다. 이때 뉴스 감정 특성상 뉴스 내용이 중복되거나 비슷한 뉴스가 존재하나, 이 또한 해당 뉴스 내용에 대한 이슈도라고 생각하여 중복 제거를 하지 않고 데이터를 BERT 모델에 적합한 형태로 변환하였습니다.

```
결측값 여부 : True

데이터 요약
<class 'pandas.core.frame.DataFrame'>
Index: 9692 entries, 0 to 9693
Data columns (total 2 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   pub_date     9692 non-null   datetime64[ns]
 1   description  9692 non-null   object
dtypes: datetime64[ns](1), object(1)
memory usage: 227.2+ KB

총 데이터 수 : 9692
```

BERT 모델 적용

KLUE BERT 모델을 사용해 감정 분석 Fine-Tuning 완료된 모델을 불러와 description 텍스트를 토크나이저를 사용해 인코딩.

감정 레이블 예측

모델을 활용해 각 뉴스 텍스트에 대한 감정 레이블 예측 (중립0, 긍정1, 부정2).

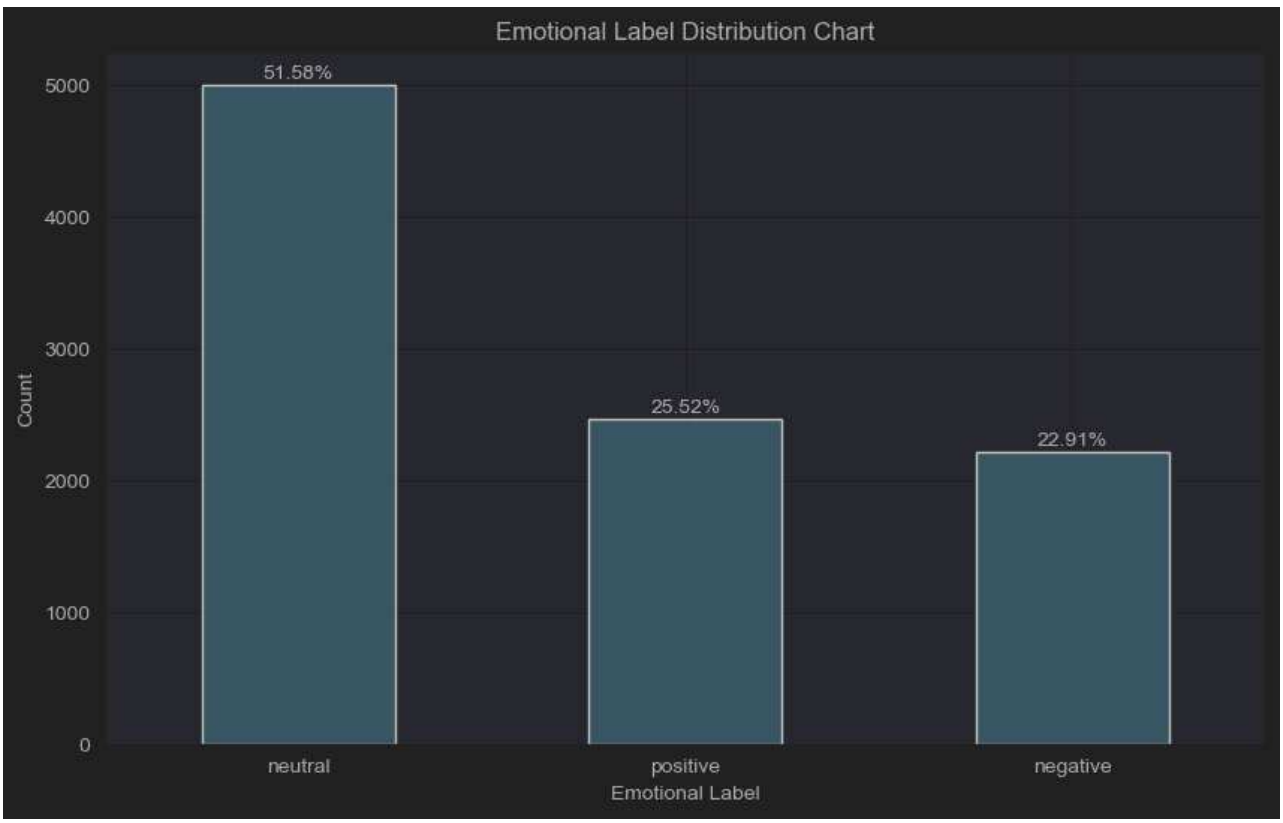
결과를 원본 데이터프레임에 sentiment 컬럼으로 추가.

```
'중립'감정 데이터
                pub_date                                        description  \
9689 2024-11-30 10:00:00  김경희 이천시장과 임해규 두원공대 총장이 말하는 반도체 산업의 속살 첨단과 전통 도...
9692 2024-11-30 12:00:00  트레이닝복세면도구속옷 등으로 구성된 구호키트와 대피소 칸막이는 <SK하이닉스>의 기...
9693 2024-11-30 13:00:00  이 종목은 <SK하이닉스>103 일본 반도체 장비주 어드반테스트96 대만 TSMC8...
'긍정'감정 데이터
                pub_date                                        description  \
9661 2024-11-30 00:00:00  이 산업의 대표적인 수혜자가 최고 성능의 5세대 HBM인 HBM3E 양산에 성공 엔...
9672 2024-11-30 07:00:00  반면 인공지능AI 시장이 개화하며 HBM 수요가 크게 늘었고 <SK하이닉스> 등 메...
9687 2024-11-30 10:00:00  국내 대표적 인공지능AI 붐 수혜기업으로 꼽히는 <SK하이닉스>와 관련 최태원 SK...
'부정'감정 데이터
                pub_date                                        description  \
9685 2024-11-30 06:00:00  미국 정부효율부 수장 임명자가 반도체과학법칩스법에 따른 반도체 보조금 지급 전반을 ...
9690 2024-11-30 10:00:00  추가 규제엔 AI인공지능칩에 필수요소인 HBM고대역폭메모리에 따른 조항도 일부 포함...
9691 2024-11-30 12:00:00  미국에 공장을 지으며 보조금을 받기로 한 삼성전자<SK하이닉스> 등 반도체 기업의 ...
```

데이터 저장 및 시각화

최종 데이터셋을 CSV 파일로 저장.

감정 레이블 분포를 시각화해 데이터의 전반적인 경향을 파악.



감정 점수와 주가의 상관관계 분석

1. 감정 점수 계산과 시각화

1) 데이터 로드 및 초기 처리

뉴스 감정데이터를 로드하고 pub_date을 datetime형식로 변환.

뉴스 데이터가 적은 2024-11-03의 데이터를 제거하여 분석에서 제외.

	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 현대치	0
2	2024-11-04 09:00:00	KB증권은 지난 달 23일부터 AI 실적속보를 통해 삼성전자 <SK하이닉스> 현(	0

2) 데이터 전처리

pub_date 컬럼에서 날짜 정보를 추출하여 day 컬럼으로 변환.

	pub_date	day
0	2024-11-04 09:00:00	2024-11-04
1	2024-11-04 09:00:00	2024-11-04
2	2024-11-04 09:00:00	2024-11-04

감정 레이블을 점수로 매핑:

긍정(1) => +1, 부정(2) => -1, 중립(0) => 0

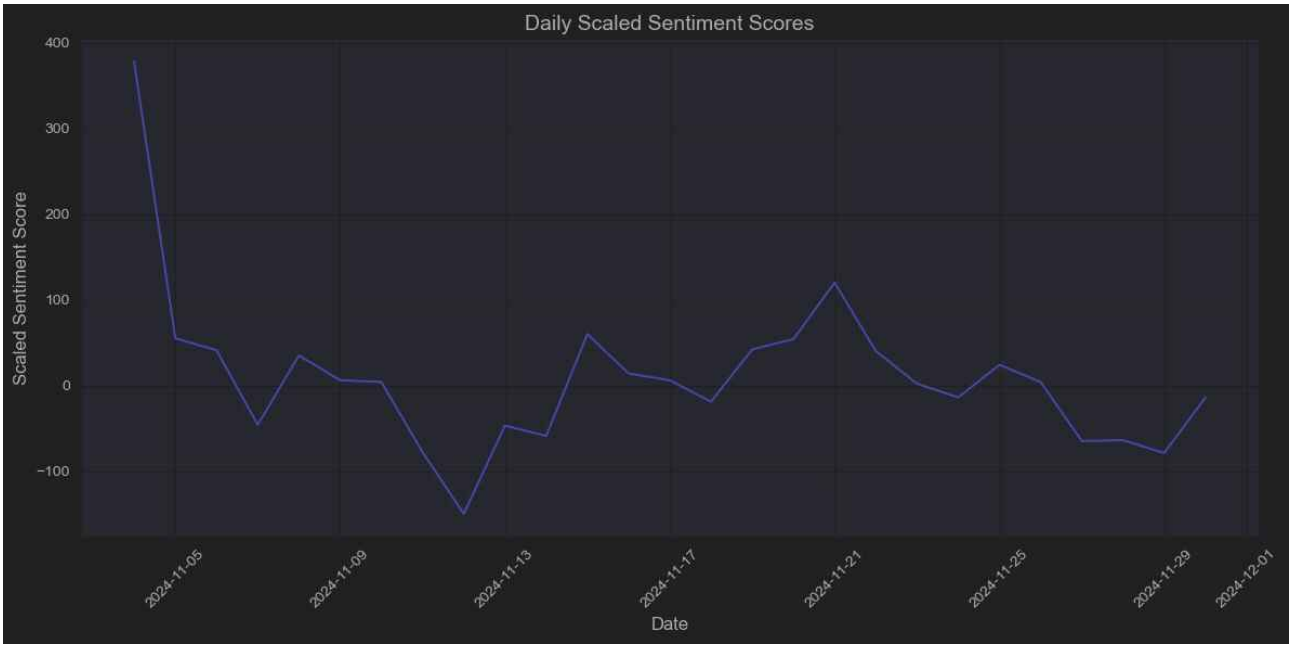
날짜단위로 그룹화:

day 컬럼을 기준으로 데이터를 그룹화하여 감정 점수 합산.

	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 점수 시각화 및 CSV저장

날짜별 감정 점수를 라인 그래프로 시각화하여 변동을 확인.



감정 점수와 주가의 상관관계 분석

2. 데이터 준비 및 전처리

1) 주가 데이터와 감정 점수 데이터 불러오기

SK하이닉스의 주가 데이터를 'FinanceDataReader' 라이브러리를 사용하여 '2024-11-04' ~ '2024-11-29'의 종가 추출.

일일 뉴스 감정 점수 데이터를 로드하여 'day' 컬럼을 인덱스로 설정.

```
주가 데이터:                    감정 데이터:
Date                                        sentiment_score
2024-11-04    194000        day
2024-11-05    193200        2024-11-04                  378
2024-11-06    195800        2024-11-05                   55
2024-11-07    197400        2024-11-06                   41
2024-11-08    200500        2024-11-07                  -46
Name: Close, dtype: int64   2024-11-08                   35
```

2) 주말 감정 점수 보정

주말에는 주가데이터 존재하지 않기 때문에 주말(토, 일)의 감정 점수가 월요일에 영향을 준다고 생각되어 금요일과 주말(토, 일)의 감정 점수를 평균 내어 금요일 점수로 업데이트하였습니다.

day	sentiment_score	sentiment_type
2024-11-04	378.000000	Positive
2024-11-05	55.000000	Positive
2024-11-06	41.000000	Positive
2024-11-07	-46.000000	Negative
2024-11-08	8.333333	Positive
2024-11-09	6.000000	Positive
2024-11-10	4.000000	Positive
2024-11-11	-78.000000	Negative
2024-11-12	-150.000000	Negative
2024-11-13	-47.000000	Negative

3) 감정 점수 구간화

시각적으로 감정 점수가 긍정인지 부정인지 직관적으로 표시하기 위해
감정 점수가 0 이상이면 긍정(Positive), 0 미만이면 부정(Negative)으로 분류.

day	sentiment_score	sentiment_type
2024-11-04	378.0	Positive
2024-11-05	55.0	Positive
2024-11-06	41.0	Positive
2024-11-07	-46.0	Negative
2024-11-08	15.0	Positive

4) 주가와 감정 점수 데이터 병합

날짜를 기준으로 주가 데이터와 감정 점수를 병합.

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

감정 점수와 주가의 상관관계 분석

3. 감정 점수와 주가의 상관관계 분석 및 시각화

1) 감정 점수 시각화

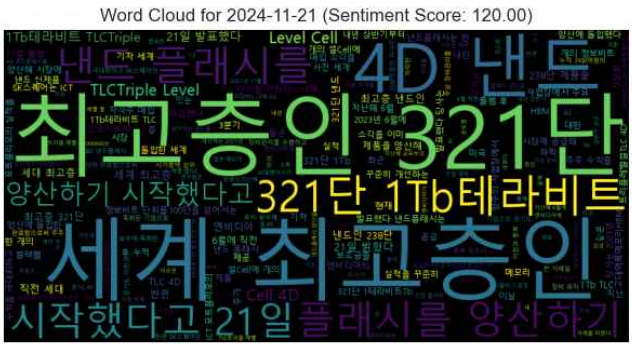
SK 하이닉스 종가와 뉴스 감정 점수를 동일한 날짜 축에서 시각화.
긍정/부정 감정 점수에 따라 배경색을 변경하여 직관적으로 표현.

2) 감정 점수와 주가 데이터 비교

감정 점수와 주가의 흐림이 비슷한 패턴을 보이며
특히, 특정 날짜(2024-11-07, 2024-11-15, 2024-11-21, 2024-11-27)에서 감정 점수와 주가 변동 간의 상관관계가 관찰됨.

3) 특정 날짜의 감정 점수와 뉴스 데이터 워드클라우드

4개의 주요 날짜를 선정하여 뉴스 데이터를 워드클라우드로 시각화.
기업명과 같은 불용어 리스트로 불필요한 단어를 제외.
감정 점수와 관련된 주요 키워드 확인.



특정 날짜의 감정 점수와 주가 변동성

2024-11-07	-46	SK하이닉스 전 직원이 핵심기술을 중국으로 빼돌려 1심에서 실형, 미국 트럼프 대통령의 보조금 삭감
2024-11-15	+26.67	산학연구과제 우수발명 포상
2024-11-21	+120	세계 최초 321단 4D 낸드플래시 양산 시작
2024-11-27	-65	주가가 계속 하락함에 따라 주주환원정책을 확대하여 주당 고정배당금을 25% 상향 발표, 하지만 최근 주가의 하락에 관해 부정적인 뉴스가 감정 점수에 큰 영향을 미침

결론 및 논의

1. 주요 발견

1) 뉴스 감정 예측의 정확도

KLUE BERT 모델을 사용하여 네이버 API로 추출한 뉴스 데이터를 감정 분석에 적용한 결과, 모델은 뉴스 문장의 맥락을 잘 반영하여 긍정적 또는 부정적 감정을 정확히 예측할 수 있었습니다. 특히 다수의 주체가 포함된 뉴스에서도 잘 작동하였으며, 예를 들어 2024년 11월 7일의 부정적인 뉴스는 다음 날 주가 하락과, 2024년 11월 21일의 긍정적인 뉴스는 주가 상승을 유도하는 경향을 보였습니다.

2) 뉴스 감정과 주가 데이터의 상관관계

뉴스 감정 점수와 주가 간의 상관계수는 0.2로 다소 약한 상관관계를 보였으나, 두 변수 간의 관계는 비선형적일 가능성이 높습니다. 주가의 반응은 뉴스 감정 점수의 크기에 따라 다르게 나타날 수 있으며, 예를 들어 매우 긍정적인 뉴스는 큰 주가 상승을 이끌 수 있지만, 경미한 긍정적인 뉴스는 영향을 미치지 않을 수 있습니다. 또한, 뉴스 특성상 오늘날의 주가에 대한 뉴스도 다수 존재하기 때문에 감정 점수와 주가의 상관관계가 아니라 주가와 감정 점수의 관계가 나타날 수도 있습니다.

2. 분석의 한계

수집된 뉴스 데이터의 양이 적고 분석 기간이 짧아 정확한 상관관계를 도출하는 데 어려움이 있었습니다. Prophet 라이브러리를 사용하여 주가 예측 결과와 뉴스 감정 데이터를 결합한 예측 결과를 비교한 결과, 감정 데이터가 적어서 의미 있는 예측을 도출하기 어려웠습니다. 데이터 양이 많고 상관관계가 명확해지면 투자자들에게 더 유용한 예측 모델을 제공할 수 있을 것으로 보입니다.