[데이터 수집 및 시각화]

주식 종목 뉴스 **감정분석**과 LSTM을 활용한 **주가예측**

2018314010 IT경영학과 남도형

Data visualization

Contents

01 연구 설명 – 연구 흐름

KOSPI 상위 25개
종목 뉴스 크롤링

KOSPI 상위 25개
종목 뉴스 전처리

**학습 완료된 모델인
KR-FinBert-SC을
활용한
감정분석**

**LSTM을 활용한
주가 예측**

01 연구 설명 – 연구 목적

주식 종목 뉴스 **감성분석**을
활용해 투자자들에게
뉴스 기사 감성 정보를 제공

특정 종목의 뉴스 **감정 분석**
결과와 주가 정보를 가지고
LSTM 모델을 활용해
주가 예측

주식 시장에 영향을 미치는 감성 정보를 활용하여 투자 전략을 구성하는 데 도움

01 연구 설명 – 선행 연구

DACON 제1회 KRX 금융 빅데이터 활용 아이디어 경진대회

금융 뉴스 및 SNS 감성분석을 통한 주가예측

- 금융 뉴스, SNS등의 자연어 데이터를 크롤링 하고 다양한 모델에 학습시킨 연구(다음주식 NEWS, 다음토론방, 네이버 주식 NEWS, 네이버 토론방, 유튜브 제목)
- SP based LSTM 모델

https://dacon.io/competitions/official/235914/codeshare/5361?page=1&dtype=random

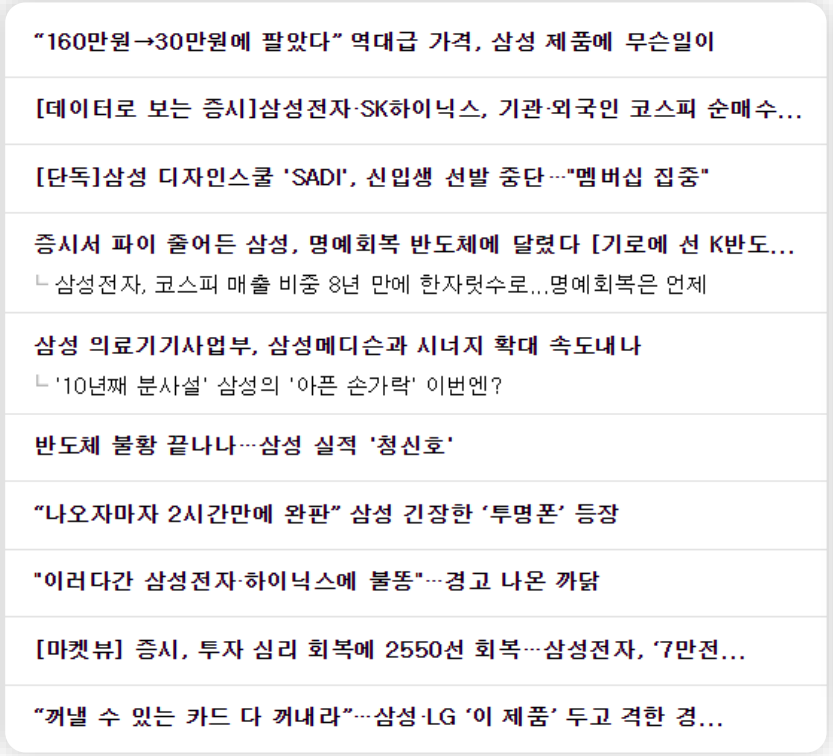
KR-FinBert-SC

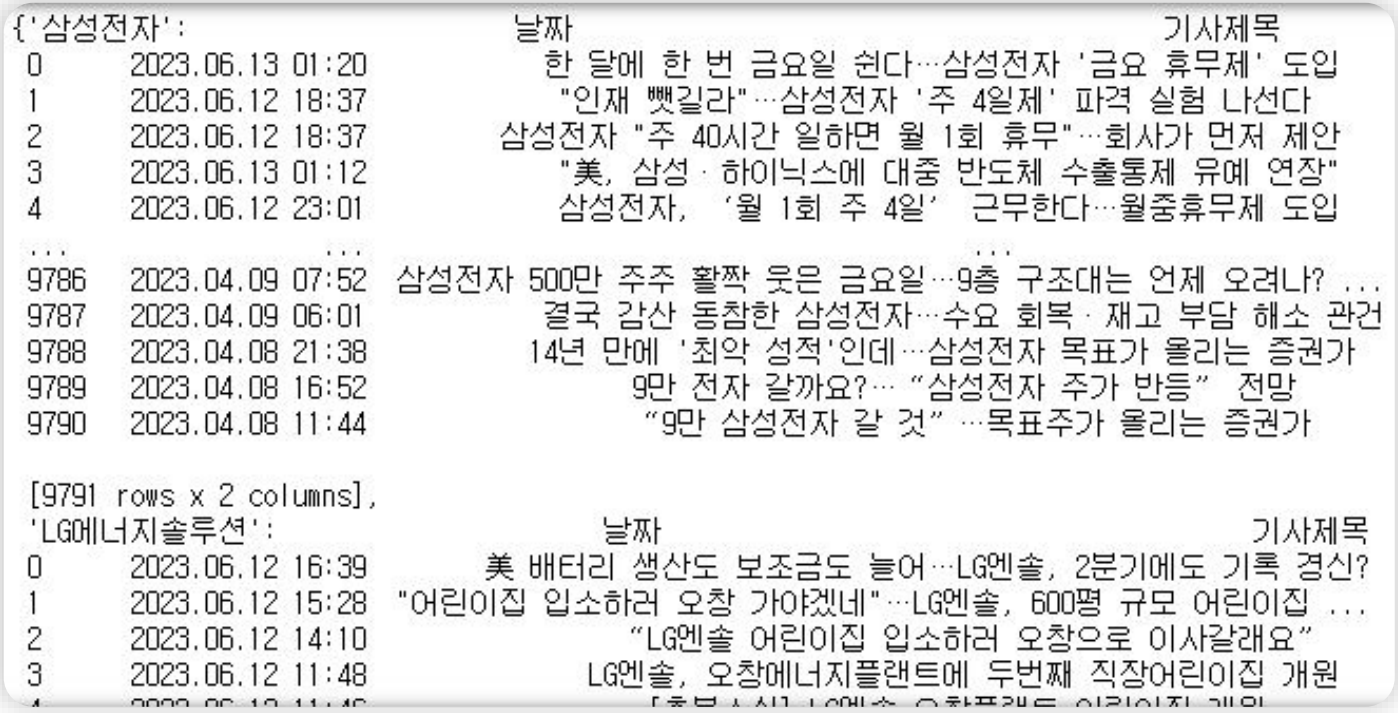
KR-FinBert 모델에 네이버 종목 뉴스 전이학습

- 구글에서 개발한 언어 모델
- 한국어 금융 데이터를 분석하기 위해 개발된 BERT 모델
- 전이학습으로 파이낸셜타임스, 한국경제 등 72개 매체의 기업 관련 경제 기사와 키움 증권, 삼성 증권 등 16개의 증권사의 애널리스트 리포트 추가

https://huggingface.co/snunlp/KR-FinBert-SC

02 연구 내용 - 크롤링





{'삼성전자':

	날짜	기사제목
0	2023.06.13 01:20	한 달에 한 번 금요일 쉰다…삼성전자 '금요 휴무제' 도입
1	2023.06.12 18:37	"인재 뺏길라"…삼성전자 '주 4일제' 파격 실험 나선다
2	2023.06.12 18:37	삼성전자 "주 40시간 일하면 월 1회 휴무"…회사가 먼저 제안
3	2023.06.13 01:12	"美, 삼성·하이닉스에 대중 반도체 수출통제 유예 연장"
4	2023.06.12 23:01	삼성전자, '월 1회 주 4일' 근무한다…월중휴무제 도입
...	...	...
9786	2023.04.09 07:52	삼성전자 500만 주주 활짝 웃은 금요일…9층 구조대는 언제 오려나? ...
9787	2023.04.09 06:01	결국 감산 동참한 삼성전자…수요 회복·재고 부담 해소 관건
9788	2023.04.08 21:38	14년 만에 '최악 성적'인데…삼성전자 목표가 올리는 증권가
9789	2023.04.08 16:52	9만 전자 갈까요?… "삼성전자 주가 반등" 전망
9790	2023.04.08 11:44	"9만 삼성전자 갈 것" …목표주가 올리는 증권가

[9791 rows x 2 columns],
'LG에너지솔루션':

	날짜	기사제목
0	2023.06.12 16:39	美 배터리 생산도 보조금도 늘어…LG엔솔, 2분기에도 기록 경신?
1	2023.06.12 15:28	"어린이집 입소하러 오창 가야겠네"…LG엔솔, 600평 규모 어린이집 ...
2	2023.06.12 14:10	"LG엔솔 어린이집 입소하러 오창으로 이사갈래요"
3	2023.06.12 11:48	LG엔솔, 오창에너지플랜트에 두번째 직장어린이집 개원

네이버 증권 - 종목 뉴스

전체 크롤링된 뉴스 수: 145557개

02 연구 내용 – 전처리

특수문자 제거 전

날짜	기사제목
2023.06.12 16:39	美 배터리 생산도 보조금도 늘어...LG엔솔, 2분기에도 기록 경신?
2023.06.12 15:28	"어린이집 입소하러 오창 가야겠네"...LG엔솔, 600평 규모 어린이집 ...
2023.06.12 14:10	"LG엔솔 어린이집 입소하러 오창으로 이사갈래요"
2023.06.12 11:48	LG엔솔, 오창에너지플랜트에 두번째 직장어린이집 개원
2023.06.12 11:46	[충북소식] LG엔솔 오창플랜트 어린이집 개원
2023.06.12 13:38	LG에너지솔루션 오창 에너지플랜트, '키즈&SOL어린이집' 개원
2023.06.12 13:18	LG엔솔 오창 에너지플랜트, '키즈&SOL어린이집' 개원
2023.06.12 11:19	LG에너지솔루션 오창 에너지플랜트, '키즈&SOL어린이집' 개원
2023.06.12 09:46	LG에너지솔루션, 분할 후 첫 신용등급 평가...AA급 우량채 지위 확보
2023.06.11 14:31	[이번주 추천주] "상승세 LG엔솔·현대미포조선에 올라탈 만"
2023.06.11 14:20	[데이터로 보는 증시]삼성전자·LG엔솔, 기관·외국인 주간 코스피 순매.
2023.06.11 11:37	[주간추천주]2Q 실적 보는 증권가...삼성전자·LG엔솔 '주목'
2023.06.11 08:54	코스피 주간 외국인 순매수 1위 LG에너지솔루션
2023.06.11 07:02	"드디어 60만원 고지"...LG엔솔 개미들 '환호'
2023.06.10 07:50	[유안타證 주간추천주]삼성전자·LG에너지솔루션·두산
2023.06.09 18:27	[시그널] LG엔솔 첫 신용등급 'AA'...1조 회사채 발행 청신호
2023.06.09 17:25	LG엔솔, 수입차 딜러와 제휴 배터리 관리사업 강화한다
2023.06.09 10:31	LG엔솔, 수입차 딜러 7곳과 '전기차 배터리 관리 사업' 확대
2023.06.09 08:45	LG엔솔, 수입차 딜러사와 전기차 배터리 관리 사업 확대

특수문자 제거 후

2023.06.12 16:39	美 배터리 생산도 보조금도 늘어LG엔솔 2분기에도 기록 경신
2023.06.12 15:28	어린이집 입소하러 오창 가야겠네LG엔솔 600평 규모 어린이집
2023.06.12 14:10	LG엔솔 어린이집 입소하러 오창으로 이사갈래요
2023.06.12 11:48	LG엔솔 오창에너지플랜트에 두번째 직장어린이집 개원
2023.06.12 11:46	충북소식 LG엔솔 오창플랜트 어린이집 개원
2023.06.12 13:38	LG에너지솔루션 오창 에너지플랜트 키즈SOL어린이집 개원
2023.06.12 13:18	LG엔솔 오창 에너지플랜트 키즈SOL어린이집 개원
2023.06.12 11:19	LG에너지솔루션 오창 에너지플랜트 키즈SOL어린이집 개원
2023.06.12 09:46	LG에너지솔루션 분할 후 첫 신용등급 평가AA급 우량채 지위 확보
2023.06.11 14:31	이번주 추천주 상승세 LG엔솔현대미포조선에 올라탈 만
2023.06.11 14:20	데이터로 보는 증시삼성전자LG엔솔 기관외국인 주간 코스피 순매
2023.06.11 11:37	주간추천주2Q 실적 보는 증권가삼성전자LG엔솔 주목
2023.06.11 08:54	코스피 주간 외국인 순매수 1위 LG에너지솔루션
2023.06.11 07:02	드디어 60만원 고지LG엔솔 개미들 환호
2023.06.10 07:50	유안타證 주간추천주삼성전자LG에너지솔루션두산
2023.06.09 18:27	시그널 LG엔솔 첫 신용등급 AA1조 회사채 발행 청신호
2023.06.09 17:25	LG엔솔 수입차 딜러와 제휴 배터리 관리사업 강화한다
2023.06.09 10:31	LG엔솔 수입차 딜러 7곳과 전기차 배터리 관리 사업 확대
2023.06.09 08:45	LG엔솔 수입차 딜러사와 전기차 배터리 관리 사업 확대

02 연구 내용 – 전처리

KOSPI 상위 25개 종목

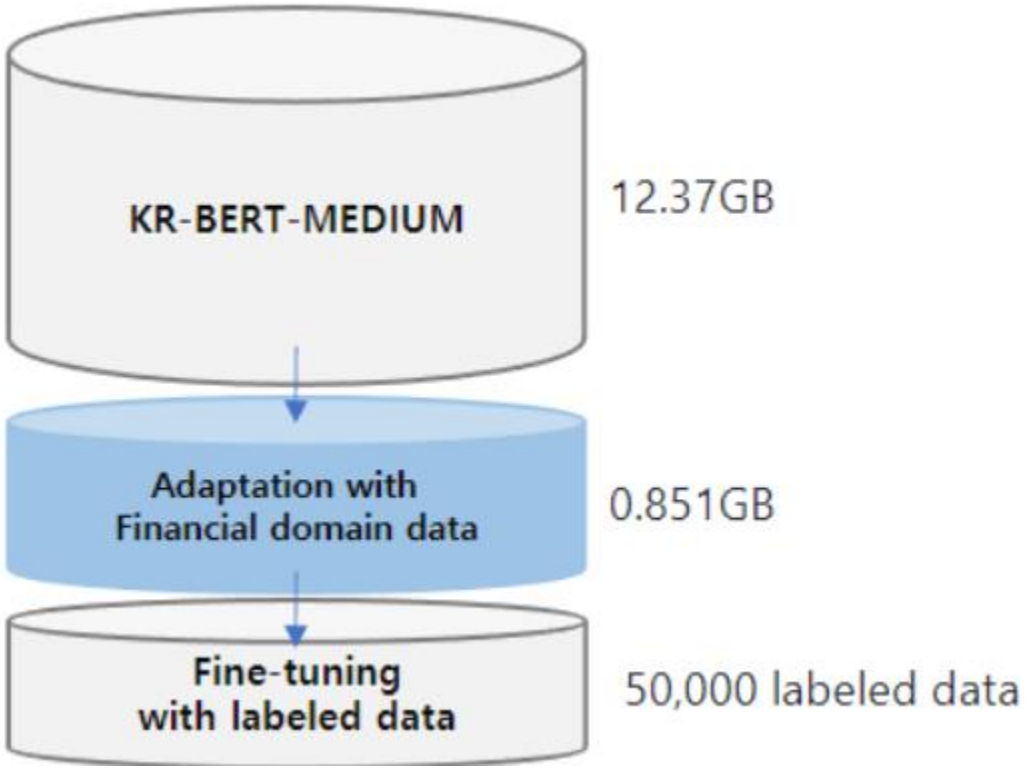
	Code	Name
~~0~~	~~005930~~	~~삼성전자~~
1	373220	LG에너지솔루션
2	000660	SK하이닉스
3	207940	삼성바이오로직스
4	051910	LG화학
5	006400	삼성SDI
~~6~~	~~005935~~	~~삼성전자우~~
~~7~~	~~005380~~	~~현대차~~
~~8~~	~~005490~~	~~POSCO홀딩스~~
~~9~~	~~035420~~	~~NAVER~~
10	000270	기아
11	003670	포스코퓨처엠
~~12~~	~~035720~~	~~카카오~~
13	068270	셀트리온
14	012330	현대모비스
~~15~~	~~066570~~	~~LG전자~~
16	028260	삼성물산
17	105560	KB금융
~~18~~	~~096770~~	~~SK이노베이션~~
~~19~~	~~055550~~	~~신한지주~~
20	003550	LG
21	032830	삼성생명
22	323410	카카오뱅크
23	034730	SK
~~24~~	~~086790~~	~~하나금융지주~~

- 2022년 6월 데이터가 있는 종목들만 선택
- 통계적 유의성을 확립하기 위해 최소 1년의 데이터가 확보된 종목만 선정

	Code	Name
0	373220	LG에너지솔루션
1	000660	SK하이닉스
2	207940	삼성바이오로직스
3	051910	LG화학
4	006400	삼성SDI
5	000270	기아
6	003670	포스코퓨처엠
7	068270	셀트리온
8	012330	현대모비스
9	028260	삼성물산
10	105560	KB금융
11	096770	SK이노베이션
12	003550	LG
13	032830	삼성생명
14	323410	카카오뱅크

02 연구 내용 – 감정분석

KR-FinBert-SC



KR-FinBert는 최대 512개, 훈련 배치 크기 32개, 학습률 5e-5로 550만 단계에 대해 훈련되며 NVIDIA TITAN XP를 사용하여 모델을 훈련하는 데 67.48시간이 걸림.

데이터 세트 : 콘텐츠가 포함된 440,067개의 뉴스 제목과 11,237개의 분석가 보고서가 포함
총 데이터 크기 : 약 13.22GB

Model	Accuracy
KR-FinBert	0.963
KR-BERT-MEDIUM	0.958
KcBert-large	0.955
KcBert-base	0.953
KoBert	0.817

뉴스 데이터 토큰화

```
from transformers import AutoTokenizer, AutoModelForSequenceClassification

tokenizer = AutoTokenizer.from_pretrained("snunlp/KR-FinBert-SC")

model = AutoModelForSequenceClassification.from_pretrained("snunlp/KR-FinBert-SC")
```

```
# 토큰화된 텍스트 리스트를 출력
for tokens in tokenized_texts:
    print(tokens)
```

```
['SK', '##이', '##노', '##베이', '##션', '행복', '##나', '##눔', '사랑', '##잇', '##기', '대', '##면', '봉사', '##활동', '3년', '##만에', '재개']
['SK', '##이', '##노', '##베이', '##션', '3년', '##만', '대', '##면', '봉사', '재개']
['SK', '##이', '##노', '##베이', '##션', '홀', '##몸', '어르신', '돌', '##봄', '자원', '##봉', '##사', '재개']
['SK', '##온', '드디어', '돈', '번', '##다', '##연', '##기', '##금', 'SK', '##이', '##노', '##베이', '##션', '미리', '찜']
['백', '##조로', '탈', '##바', '##꿈', 'SK', '##온', '품', '##고', '##S', '##K', '##이', '##노', '##베이', '##션', '훨', '##훨']
['특징', '##주', 'SK', '##이', '##노', '##베이', '##션', '자', '##회사', '수익', '##개', '##선', '기대', '등에', '강', '##세', '##종합']
['특징', '##주', 'SK', '##이', '##노', '##베이', '##션', '자', '##회사', '수익', '##개', '##선', '기대', '등에', '강', '##세']
['울산', '찾', '##은', '권', '##기', '##섭', '고용', '##차', '##관', 'SK', '##이', '##노', '##베이', '##션', '상', '##생', '사례', '타', '기업', '##에', '확산']
['SK', '##이', '##노', '##베이', '##션', '사회적', '##가', '##치', '3', '##조', '창출', '##전', '##년', '##보다', '12', '##5']
['SK', '##이', '##노', '##베이', '##션', '사회적', '##가', '##치', '3', '##조', '창출', '##전', '##년', '대비', '12', '##5']
['SK', '##이', '##노', '##베이', '##션', '지난해', '사회적', '가치', '3', '##조', '##3', '##8', '##3', '##억원', '##전', '##년', '##비', '12', '##5']
['SK', '##이', '##노', '##베이', '##션', '사회적', '가치', '성과', '3', '##조원', '##역', '##대', '최대']
['데이터', '##로', '보는', '증', '##시', '##S', '##K', '##이', '##노', '##베이', '##션', '##삼', '##성', '##전자', '기관', '##외', '##국인', '주', '##간', '코스피']
['코스피', '주', '##간', '기관', '순', '##매', '##수', '1위', 'SK', '##이', '##노', '##베이', '##션']
```

0은 중립, 1은 긍정, 2는 부정

```
from transformers import pipeline

classifier = pipeline("sentiment-analysis", tokenizer=tokenizer, model=model)
```

sentiment_mapping = {'neutral': 0, 'positive': 1, 'negative': 2}

2023-05-18 9:45	LG에너지솔루션	"도전적 R&D 응원" LG엔솔 발명왕 포상	0
2023-05-18 8:55	LG에너지솔루션	LG에너지솔루션 '2023 발명왕' 시상식	0
2023-05-18 8:38	LG에너지솔루션	LG엔솔, '발명의 날' 맞아 특허활동 우수 임직원 포상	0
2023-05-18 8:34	LG에너지솔루션	LG에너지솔루션, 사내 '발명왕' 선발..."R&D 적극 장려"	0
2023-05-19 9:46	LG에너지솔루션	LG엔솔 첫 회사채 발행 착수...2차전지 성장성에 흥행 기대	1
2023-05-19 9:07	LG에너지솔루션	LG엔솔, 호주 기업 보유 북미 광산서 5년간 리튬정광 25% 받는다	1
2023-05-19 8:53	LG에너지솔루션	LG엔솔, 북미 리튬광 업체에 지분 투자...원재료 공급망 강화	1
2023-05-17 18:13	LG에너지솔루션	LG엔솔 전기車 합작공장 건설 중단... 加 주-연방정부 마찰에 장기화 우...	2
2023-05-17 16:25	LG에너지솔루션	LG엔솔-스텔란티스 합작공장 건설 중단 장기화 우려..加 정부간 핑퐁...	2

02 연구 내용 – 주가예측

FinanceDataReader

Date	Open	High	Low	Close	Volume	Change
2022-06-13	414000	420500	413500	415000	267003	-0.023529
2022-06-14	411000	432500	410500	426500	417106	0.027711
2022-06-15	426000	427500	417000	420000	275832	-0.015240
2022-06-16	424500	439500	424500	427000	365836	0.016667
2022-06-17	419500	430000	417000	425500	1211937	-0.003513
...	...	...	...	...	...	...
2023-06-08	601000	602000	588000	591000	286608	-0.021523
2023-06-09	597000	610000	597000	609000	449434	0.030457
2023-06-12	614000	614000	603000	612000	231274	0.004926
2023-06-13	612000	613000	604000	607000	194004	-0.008170

2022년 6월 13일 ~ 2023년 6월 13일
시가, 고가, 저가, 종가, 거래량, 주식 가격 변동률 사용

날짜를 기준으로 병합

Emotion.csv

Date	Emotion
2023-06-12 16:39	1
2023-06-12 15:28	0
2023-06-12 14:10	0
2023-06-12 11:48	0
2023-06-12 11:46	0
2023-06-12 13:38	0
2023-06-12 13:18	0
2023-06-12 11:19	0
2023-06-12 9:46	1
2023-06-11 14:31	1
2023-06-11 14:20	0
2023-06-11 11:37	0
2023-06-11 8:54	1
2023-06-11 7:02	1
2023-06-10 7:50	0

날짜별로 그룹화하고 가장 많은
Emotion값 선택

병합 완료 데이터

Date	index	Open	High	Low	Close	Volume	Change	Emotion
2022-06-13	0	414000	420500	413500	415000	267003	-0.023529	1
2022-06-14	1	411000	432500	410500	426500	417106	0.027711	1
2022-06-15	2	426000	427500	417000	420000	275832	-0.015240	1
2022-06-16	3	424500	439500	424500	427000	365836	0.016667	0
2022-06-17	4	419500	430000	417000	425500	1211937	-0.003513	0
...	...	...	...	...	...	...	...	...
2023-06-08	245	601000	602000	588000	591000	286608	-0.021523	1
2023-06-09	246	597000	610000	597000	609000	449434	0.030457	1
2023-06-12	247	614000	614000	603000	612000	231274	0.004926	0
2023-06-13	248	612000	613000	604000	607000	194004	-0.008170	0

02 연구 내용 – 주가예측

Train, valid, test 데이터셋 나누기

전체 데이터셋 크기: 250

- 학습데이터 : 전체 데이터셋의 80% (200)
- 검증데이터 : 전체 데이터셋의 20% 중 40%(20)
- 테스트데이터 : 전체 데이터셋의 남은 20% 중 60%(30)

독립변수, 종속변수

독립변수

'Open', 'High', 'Low', 'Volume', 'Change', 'Emotion'

종속변수

'Close'

02 연구 내용 – 주가예측

- MinMaxScaler를 사용하여 데이터를 정규화

- 각 주식에 대한 학습, 검증, 테스트 데이터 셋의 크기 확인

```
*****Train shapes for LG에너지솔루션
(200, 6) (200, 1)
*****Validation shapes for LG에너지솔루션
(20, 6) (20, 1)
*****Test shapes for LG에너지솔루션
(30, 6) (30, 1)
*****Train shapes for SK하이닉스
(200, 6) (200, 1)
*****Validation shapes for SK하이닉스
(20, 6) (20, 1)
*****Test shapes for SK하이닉스
(30, 6) (30, 1)
*****Train shapes for 삼성바이오로직스
(200, 6) (200, 1)
*****Validation shapes for 삼성바이오로직스
(20, 6) (20, 1)
*****Test shapes for 삼성바이오로직스
(30, 6) (30, 1)
```

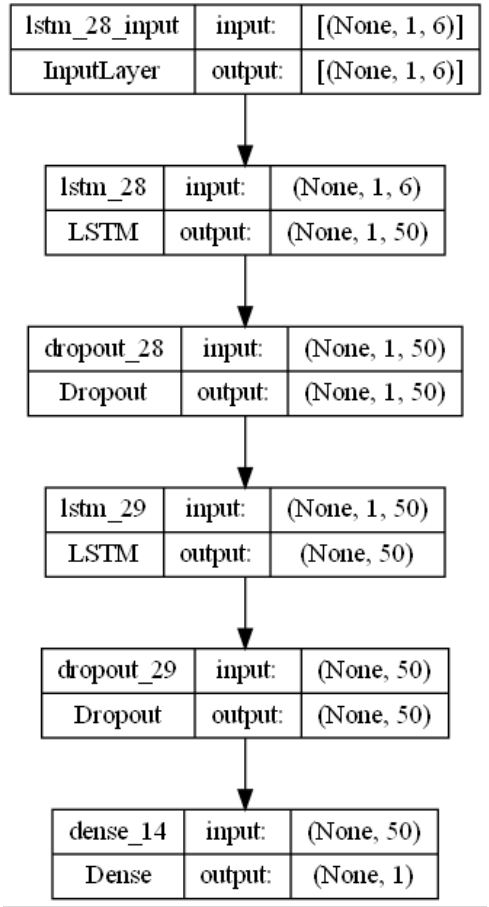
- 데이터셋에 대해 Train,Validation, Test 데이터를 생성

- 생성된 generator 객체를 각 데이터 셋에 대해 저장

```
Train samples for LG에너지솔루션: 200
Validation samples for LG에너지솔루션: 20
Test samples for LG에너지솔루션: 30
Train samples for SK하이닉스: 200
Validation samples for SK하이닉스: 20
Test samples for SK하이닉스: 30
Train samples for 삼성바이오로직스: 200
Validation samples for 삼성바이오로직스: 20
Test samples for 삼성바이오로직스: 30
Train samples for LG화학: 200
Validation samples for LG화학: 20
Test samples for LG화학: 30
```

03 분석 결과 – 주가예측

LSTM 내부 layer 아키텍처



03 분석 결과 – 주가예측

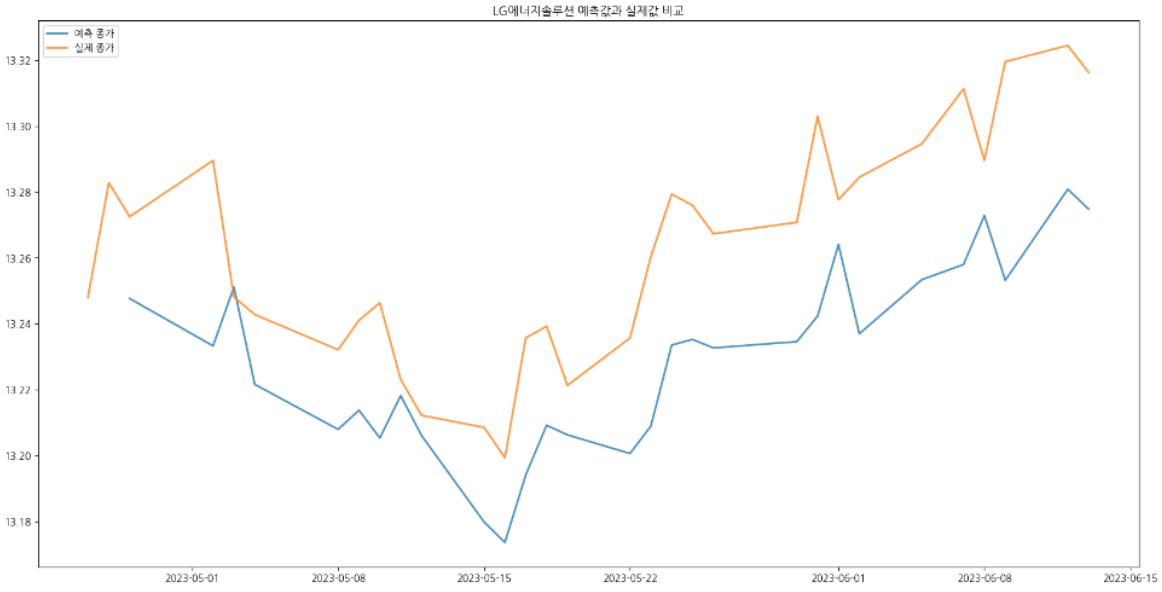
model.compile(loss='mean_squared_error', optimizer='adam')

손실함수로 평균 제곱 오차(MSE)를 사용
Adam은 경사 하강법의 한 종류로, 학습 속도를 조절하면서 가중치를 업데이트 하는 방법

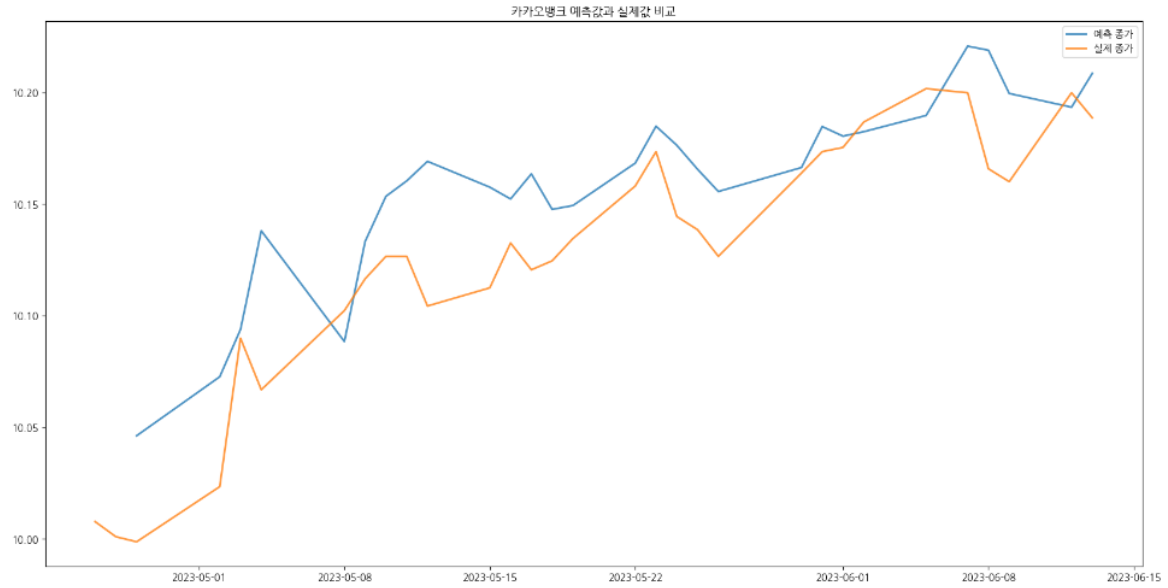
```
198/198 [==============================] - 6s 4ms/step - loss: 0.0679

C:₩Users₩dong_₩AppData₩Local₩Temp₩ipykernel_42820₩1173966349.py:30: UserWarning: `Model.predict_generator` is deprecated and will be re
moved in a future version. Please use `Model.predict`, which supports generators.
  test_predict = model.predict_generator(test_generator)

198/198 [==============================] - 7s 5ms/step - loss: 0.0195

C:₩Users₩dong_₩AppData₩Local₩Temp₩ipykernel_42820₩1173966349.py:30: UserWarning: `Model.predict_generator` is deprecated and will be re
moved in a future version. Please use `Model.predict`, which supports generators.
  test_predict = model.predict_generator(test_generator)

198/198 [==============================] - 7s 4ms/step - loss: 0.0654

C:₩Users₩dong_₩AppData₩Local₩Temp₩ipykernel_42820₩1173966349.py:30: UserWarning: `Model.predict_generator` is deprecated and will be re
moved in a future version. Please use `Model.predict`, which supports generators.
  test_predict = model.predict_generator(test_generator)

198/198 [==============================] - 7s 4ms/step - loss: 0.0420

C:₩Users₩dong_₩AppData₩Local₩Temp₩ipykernel_42820₩1173966349.py:30: UserWarning: `Model.predict_generator` is deprecated and will be re
```

03 분석결과 – 주가예측

LG에너지솔루션 예측값과 실제값 비교

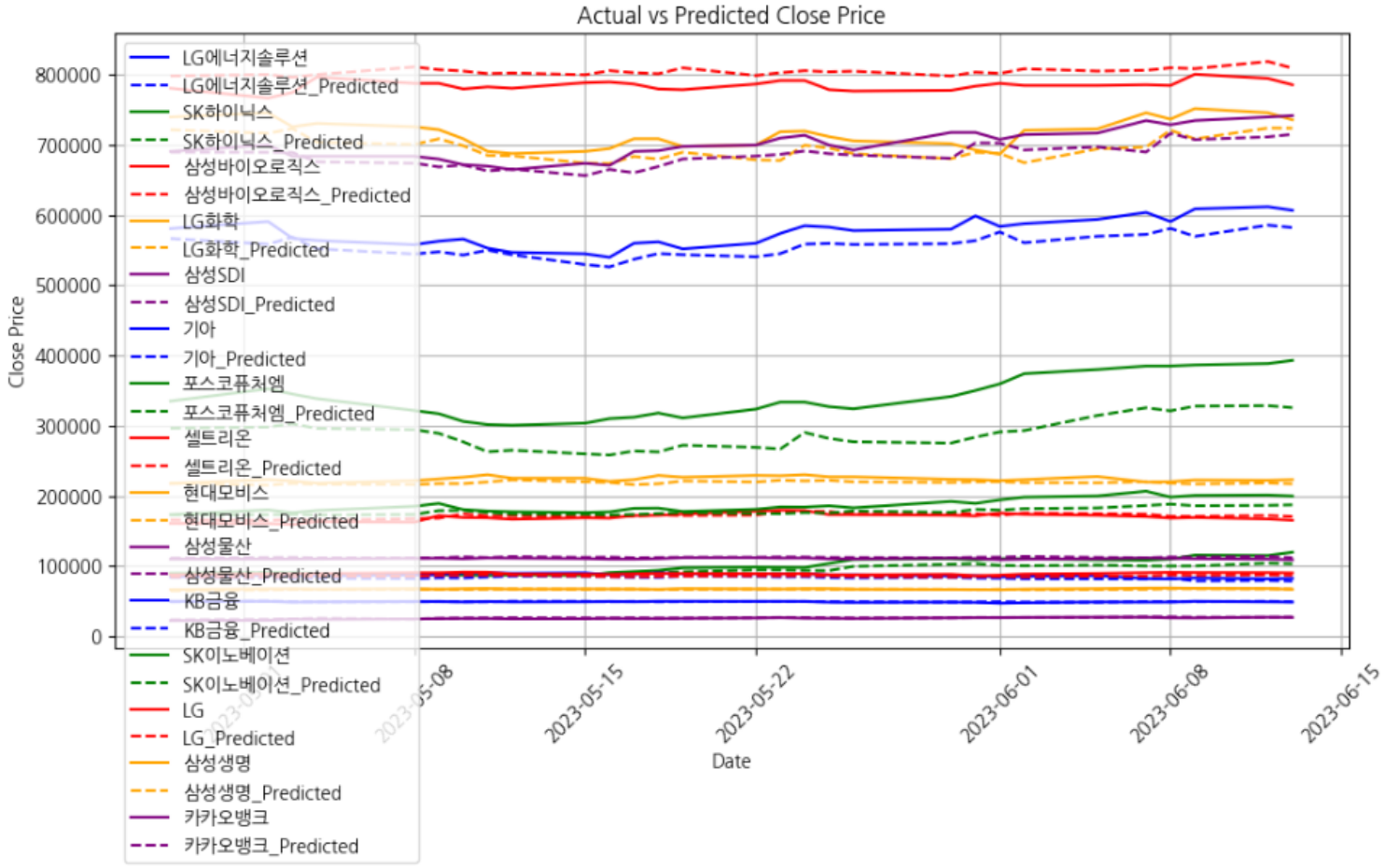


카카오뱅크 예측값과 실제값 비교



03 분석 결과 - 주가예측

전체 종목 실제값과 예측값 종가 비교



04 시사점 및 한계점

시사점

- 높은 Accuracy(0.963)를 가진 KR-FinBert 모델을 사용함으로써 주식 관련 뉴스의 감성분석에 효과적으로 적용이 가능했다.
- 주가 예측 모델 : LSTM을 통해 주가 예측을 수행하여 투자자들이 유용하게 활용 가능하다.
- 감성분석과 주가 예측의 결합 : 주식 시장에 영향을 미치는 감성 정보를 활용하여 투자 전략을 구성하는 데 도움을 줄 수 있다.

한계점

- 네이버 종목 뉴스뿐만 아니라 다양한 사이트의 텍스트를 분석했으면 더 유의미한 분석이 가능했을 것 같다.
- 더 많은 데이터를 확보하고 컴퓨터 H/W 자원이 좋았다면 모델 성능 향상에도 도움이 됐을 것 같다.
- 독립변수와 종속변수를 더 추가했다면 다양한 시각화를 할 수 있었지 않을까하는 아쉬움이 있다.

감사합니다.