뉴스를 열심히 보면 주식을 잘할 수 있을까?

주가와 증권뉴스 감성 상관관계

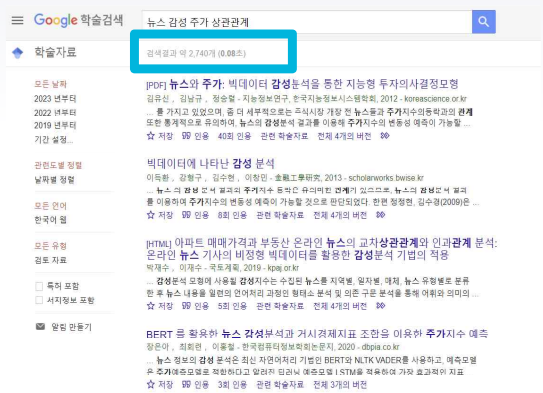
금준호, 김란, 노두호, 안지윤, 최규진

목차

주가와 증권뉴스 감성 상관관계

1 프로젝트 개요

주가와 증권뉴스 감성 상관관계



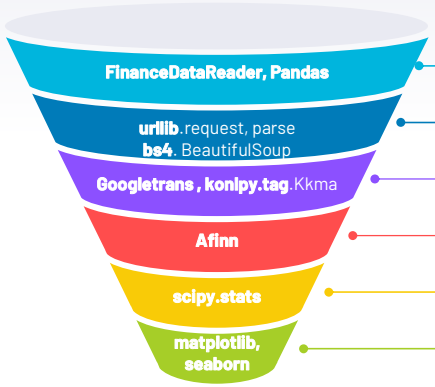
*'그 결과 **뉴스 컨텐츠의 감성분석 결과값과 주가지수 등락과는 유의한 관계를 가지고 있었으며**, 좀 더 세부적으로는 주식시장 개장 전 뉴스들과 주가지수의 등락과의 관계 또한 통계적으로 유의하여, 뉴스의 감성분석 결과를 이용해 주가지수의 변동성 예측이 가능할 것으로 판단되었다'*

뉴스와 주가 : 빅데이터 감성분석을 통한 지능형 투자의사결정모형

*The count of the positive and negative sentiment of news articles for each day and variance of adjacent days close price along with historical data is used for prediction purpose and **an accuracy ranging from 65.30 to 91.2 % achieved with various machine learning techniques.***

Efficacy of News Sentiment for Stock Market Prediction (2019)

데이터 분석 계획



1. 삼성전자 2년치 주가 **데이터프레임 생성**
2. 네이버 증권 뉴스 제목, 첫 문단 **크롤링**
3. **한영번역**과 텍스트 **전처리**
4. **감성분석**
5. **상관관계** 분석
6. 데이터 **시각화**

2 데이터 분석

데이터 분석 모델

a. 종가 **높고 낮은 30일 직전 3일 뉴스 분석**

2년 내 주식 최고가, 최저가 찍기 직전 3일 뉴스 감성분석

b. 수익률 **큰 30일 직전 3일 뉴스 분석**

주가 자체보다 많이 오르고 내린, 변동량이 큰 기간이 더 상관관계가 클 것이라 가설

c. 당일 뉴스 **감성분석**

365일 뉴스 감성 분석값과 수익률과의 상관관계 분석

1. 뉴스가 주가에 반영되는 데에 **3**일 걸린다고 가정
2. 유의미한 표본양을 **30**주로 계산

1. 주가 데이터프레임 생성

```python
def Finace_data(code, name):
    # 데이터 얻기
    data = fdr.DataReader(code,
    start='2021-01-01',
    end='2022-12-31')

    # 결측치 제거
    data.dropna(inplace=True)


#함수 실행
Finace_data('005930', 'Samsung')
```

```
Date,Open,High,Low,Close,Volume,Change
2022-01-03,79400,79800,78200,78600,13502112,0.003831417624521105
2022-01-04,78800,79200,78300,78700,12427416,0.0012722646310432406
2022-01-05,78800,79000,76400,77400,25470640,-0.016518424396442133
2022-01-06,76700,77600,76600,76900,12931954,-0.006459948320413411
2022-01-07,78100,78400,77400,78300,15163757,0.018205461638491 46
2022-01-10,78100,78100,77100,78000,9947422,-0.003831417624521105
2022-01-11,78400,79000,78000,78900,13221123,0.011538461538461497
2022-01-12,79500,79600,78600,78900,11000502,0.0
2022-01-13,79300,79300,77900,77900,13889401,-0.012674271229404344
2022-01-14,77700,78100,77100,77300,10096725,-0.007702182284980763
2022-01-17,77600,77800,76900,77500,8785122,0.002587322121604174
2022-01-18,77600,77800,76600,77000,9592788,-0.006451612903225823
2022-01-19,76500,76900,76100,76300,10598290,-0.009090909090909038
2022-01-20,76200,76700,75900,76500,9708168,0.002621231979030192
2022 01 21,75800,75800,74700,75600,15774888, 0.011764705882352 9
2022-01-24,75400,75800,74700,75100,13691134,-0.006613756613756627
2022-01-25,74800,75000,73200,74000,17766704,-0.014647137150466 06
2022-01-26,73900,74400,73100,73300,12976730,-0.009459459459459474
2022-01-27,73800,74000,71300,71300,22274777,-0.027285129604365577
2022-01-28,71300,73700,71200,73300,21367447,0.028050490883590573
2022-02-03,74900,74900,73300,73300,17744721,0.0
2022-02-04,74300,74600,73400,74000,12730034,0.009549795361528002
2022-02-07,73500,73600,72400,73000,14240838,-0.013513513513513487
2022-02-08,73800,74200,73000,73500,11736666,0.006849315068493178
```

a. 종가 기준 모델

```
highest30 = ssdf.nlargest(n=30,columns='close',keep='all')
highest30.sort_values(by=['date'])
highest14_dates = [["2021-01-06", "2021-01-08"], ["2021-01-09", "2021-01-11"], ["2021-01-12", "2021-01-14"], …]
lowest16_dates = [["2022-06-29", "2022-07-01"], ["2022-07-04", "2022-07-06"], ["2022-09-05", "2022-09-07"], …]
dates = highest14_dates + lowest16_dates
```

b. 수익률 기준 모델

```
def top_N_changes(ssdf, N):
    ssdf['Abs_Change'] = ssdf['change'].abs()
    top_N = ssdf.sort_values(by='Abs_Change',ascending=False).head(N)
    top_N = top_N.drop('Abs_Change', axis=1)
    return top_N

top30 = top_N_changes(samsung, 30)
# 결과값 :['2021-01-08','2021-01-18',...]
```

c. 당일 뉴스 기준 모델

```
Holidays = [] # 공휴일 지정
current_date = datetime.datetime.strptime(start_date, "%Y%m%d") #string to datetime
end_date = datetime.datetime.strptime(end_date, "%Y%m%d")

while current_date <= end_date:
    if current_date.weekday() >= 5 or current_date.strftime("%Y-%m-%d") in holidays:
```


2. 네이버 증권 크롤링

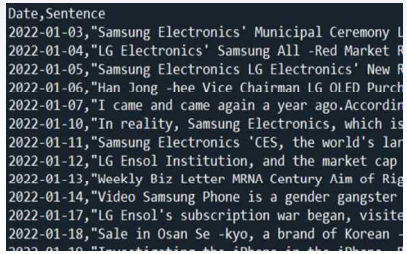
```
def crawling(date, lastpage):
    titles = [];     urls = [];     summaries = []

    for page in range(1, lastpage+1):
            url = URL
            response = requests.get(url)
            soup = BeautifulSoup(response.text, "html.parser")

            for article in soup.find_all("dd", class_="articleSubject"):
                title = article.a.text.strip()
                title = sub("[^ㄱ-ㅎㅏ-ㅣ가-힣 | A-Z |a-z ]","", title)
                url = article.a["href"]
                if title not in t_list:
                    titles.append(title)
                if url not in urls:
                    urls.append(url)

            for summary in soup.find_all("dd", class_="articleSummary"):
                summary_str = sub("[^ㄱ-ㅎㅏ-ㅣ가-힣 | A-Z |a-z ]","", summary_str)
                summary_str = summary_str.strip()
                summaries.append(summary_str)

        return titles, summaries
```

```
URL =
f"https://finance.naver.com/news/news_search.naver?rcdate=&q=%BB%EF%BC%BA%C0%FC%C0%DA&x=0&y=0&sm=all.basic&pd=4&stDateStart={date[0]}&stDateEnd={date[1]}&page={page}"
```

3. 한영번역과 전처리, 감성분석

```
def TranslatorFunc (news):
    eng_news = translator.translate(news, dest='en', src='ko')
    return eng_news

def clean_text(texts) :
    texts_re = [sub('[,.?!:;]', '', st) for st in texts]
    texts_re2 = [sub('[@#$%^&*()]', '', st) for st in texts_re]
    texts_re3 = [sub('[^a-z|A-Z]', ' ', st) for st in texts_re2]
    texts_re4 = [sub('\s+', ' ', st) for st in texts_re3]
    texts_re5 = [sub('[0-9]', '', st) for st in texts_re4]
    return texts_re5
```

```
afinn = Afinn(language='en')
sentiment = afinn.score(eng_news)

news_df = ['les gens pensent aux chiens','i hate flowers','he is
kind and smart','we are kind to good people']

scores = [afn.score(article) for article in news_df]
sentiment = ['positive' if score > 0
             else 'negative' if score < 0
             else 'neutral' for score in scores]
```

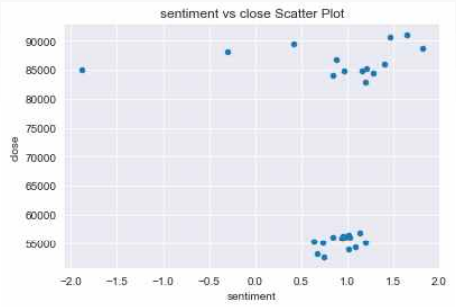


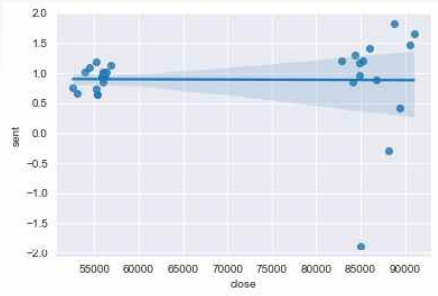
```
Date,Sentence
2022-01-03,"Samsung Electronics' Municipal Ceremony L
2022-01-04,"LG Electronics' Samsung All -Red Market R
2022-01-05,"Samsung Electronics LG Electronics' New R
2022-01-06,"Han Jong -hee Vice Chairman LG OLED Purch
2022-01-07,"I came and came again a year ago.Accordin
2022-01-10,"In reality, Samsung Electronics, which is
2022-01-11,"Samsung Electronics 'CES, the world's lar
2022-01-12,"LG Ensol Institution, and the market cap
2022-01-13,"Weekly Biz Letter MRNA Century Aim of Rig
2022-01-14,"Video Samsung Phone is a gender gangster
2022-01-17,"LG Ensol's subscription war began, visite
2022-01-18,"Sale in Osan Se -kyo, a brand of Korean -
```

	topic	scores	sentiments
0	les gens pensent aux chiens	0.0	neutral
1	i hate flowers	-3.0	negative
2	he is kind and smart	3.0	positive
3	we are kind to good people	5.0	positive

3 상관분석 및 시각화

a. 종가 기준 모델

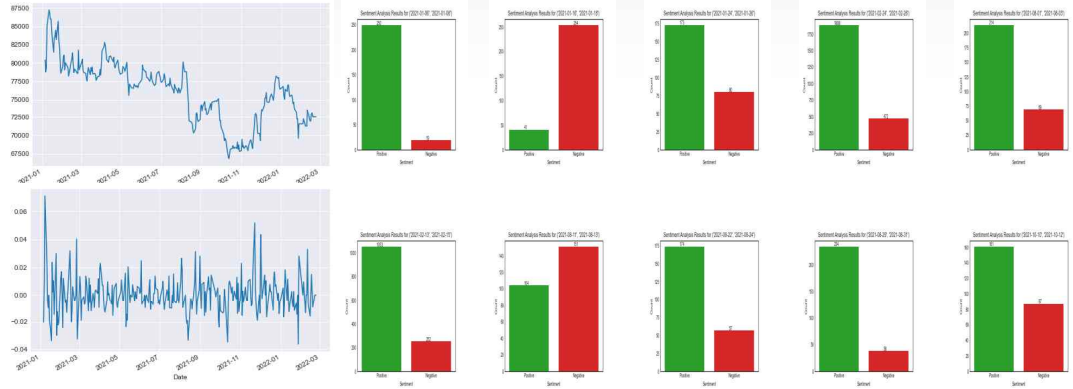




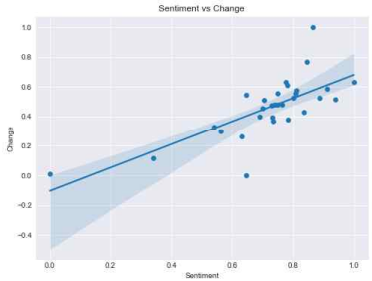
상관계수 : -0.0117, P-value : 0.9512

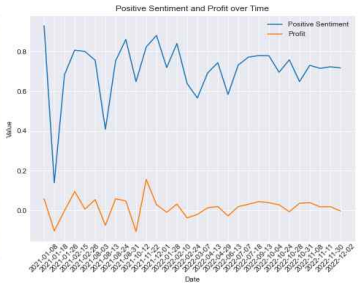
b. 수익률 기준 모델

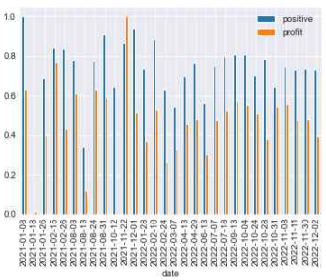
2021 - 01 ~ 2022 - 12 까지 수익률 변동폭이 큰 기간 **감성분석** 그래



b. 수익률 기준 모델







산점도 그래프 라인 그래프 막대 그래프

상관계수 : 0.73001, P-value : 0.000005

c. 당일 뉴스 감성분석 모델

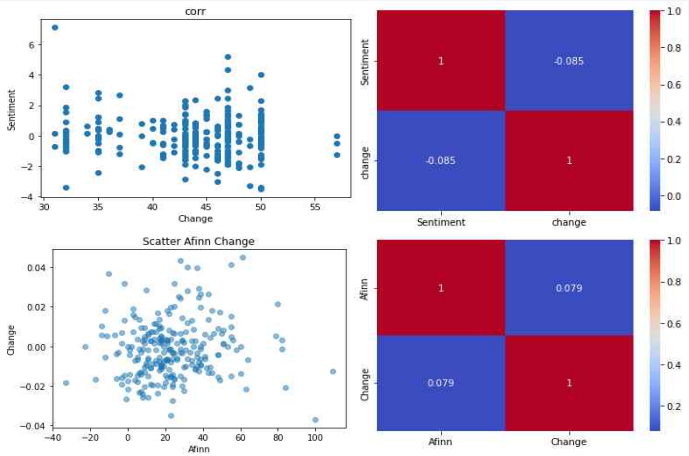
Affin 사용으로 결과 점수화

```
mean_sentiment = merged_df['Sentiment'].mean()
# 43.891
median_sentiment = merged_df['Sentiment'].median()
# 45.0

# 감성점수가 40점 이상이면 긍정, 미만이면 부정으로 나타냄
df_translated['Pos_Neg'] = df_translated['Afinn'].apply(lambda x:
'Positive' if x >= 40 else 'Negative')

# 실제로 예측 성공한 날 True, 실패한 날 False로 출력
merged_df['score'] = np.where((merged_df['Sentiment'] >= 40) &
(merged_df['change'] >= 0), 'TRUE', 'FALSE')
```

Date,Afinn,Pos_Neg	Sentiment	score
2022-01-03,17.0,Negative	47.0	TRUE
2022-01-04,51.0,Positive	47.0	TRUE
2022-01-05,30.0,Negative	39.0	FALSE
2022-01-06,24.0,Negative	47.0	TRUE
2022-01-07,21.0,Negative	31.0	FALSE
2022-01-10,-6.0,Negative	...	...
2022-01-11,15.0,Negative	47.0	TRUE
2022-01-12,60.0,Positive	47.0	FALSE
2022-01-13,109.0,Positive	47.0	TRUE
2022-01-14,17.0,Negative	47.0	FALSE
2022-01-17,34.0,Negative	49.0	FALSE
2022-01-18,41.0,Positive		

2021년 예측 성공한 날 : 95/248 → 정확도: 약 38%
2022년 예측 성공한 날 : 137/246 → 정확도: 약 56%

c. 당일 뉴스 감성분석 모델



2021년:
상관계수: **-0.085,**
p-value : 0.403

2022년:
상관계수: **0.079,**
p-value : 0.215

결과 정리

a. 종가 높은 30일 직전 3일 뉴스 분석

상관계수 : -0.0117,
P-value : 0.9512

→ 통계적으로 유의하지 않음

b. 수익률 큰 30일 직전 3일 뉴스 분석

상관계수 : 0.73001,
P-value : 0.000005

→ **통계적으로 유의**

c. 당일 뉴스 감성분석

2021년: - 0.053, 0.403
2022년: 0.079, 0.215

→ 통계적으로 유의하지 않음

공통 가설

H0 : 뉴스가 주가 변동율에 영향을 끼치지 않는다.
H1 : 뉴스가 주가 변동율에 영향을 끼친다.

4 결론 및 의의

한계점

주가데이터

수익률이 낮은 기간 상관관계가 애매

뉴스데이터

여러 뉴스사에서 비슷한 기사 업로드

감성분석

- 한영번역 한계
- 같은 단어라도 긍정/부정 가능
- 주식 특화 단어사전 부재

상관관계 분석

뉴스 감성 외에도 증시에 영향을 주는 여러 변수들 존재

발전방향

주가데이터

수익률과 변동률이 큰 주식을 선택

뉴스데이터

비슷한 뉴스 기사의 중복을 제거

감성분석

영어권 주가정보 분석

상관관계 분석

다른 변수들 같이 고려한 상관관계 분석 가능

감사합니다

금준호, 김란, 노두호, 안지윤, 최규진