금융 도메인, 뉴스 데이터 감성 분석

2기 정다희

프로젝트 개요

분석 목적

금융 도메인에서 뉴스 데이터 감성 분석을 하고자 함

최종적으로, 뉴스 데이터를 기반으로 주가 예측/숨겨진 인사이트 탐색을 하고 싶음

데이터 출처

DACON, 2023 NH 투자증권 빅데이터 경진대회, “블룸버그, 나스닥과 함께 세계속으로!"

데이터셋 설명

주요 변수

	rgs_dt	tck_iem_cd	til_ifo	ctgy_cfc_ifo	mdi_ifo	news_smy_ifo	rld_ose_iem_tck_cd	url_ifo
0	2023-01-02	NGS	Hoak & Co. Urges NGS Board to Halt CEO Search ...	Investing	Fintel	Fintel reports that Hoak Public Equities, LP, ...	NGS	https://www.nasdaq.com/articles/hoak-co.-urges...
1	2023-01-02	PFX	Here's Why PhenixFIN (PFX) is Poised for a Tur...	Stocks,Investing	Zacks	PhenixFIN (PFX) has been beaten down lately wi...	PFX	https://www.nasdaq.com/articles/heres-why-phen...
2	2023-01-02	TYG	My Advice? Sell These 2 Dividend Funds in 2023	Markets,Stocks	BNK Invest	There's a disconnect setting up in the energy ...	TYG,TYG,FEN	https://www.nasdaq.com/articles/my-advice-sell...
3	2023-01-02	NVIV	France's InVivo to buy one of the world's olde...	Commodities,BioTech,US Markets	Reuters	Malteries Soufflet, a branch of French agribus...	NVIV	https://www.nasdaq.com/articles/frances-invivo...
4	2023-01-02	FEN	My Advice? Sell These 2 Dividend Funds in 2023	Markets,Stocks	BNK Invest	There's a disconnect setting up in the energy ...	TYG,TYG,FEN	https://www.nasdaq.com/articles/my-advice-sell...

til_info

뉴스 제목 내 상위 단어 분석
(예: buy, recommendation 등)

→ 주식·투자 관련 뉴스 여부 식별 가능

rgs_dt

뉴스 등록일 정보

→ 요일별 기사량 차이 분석 가능
(예: 월~목 많고, 주말 적음)

ctgy_clc

뉴스 카테고리
(예: 주식, ETF, 옵션 등)

→ 유형별 군집화 및 실적·배당 관련 뉴스 필터링 가능

데이터 전처리

값 처리

- 중복행 제거
- Url_info 컬럼 삭제
- 뉴스 데이터에는 있지만, 주가 데이터에는 없는 티커 삭제

날짜 변환

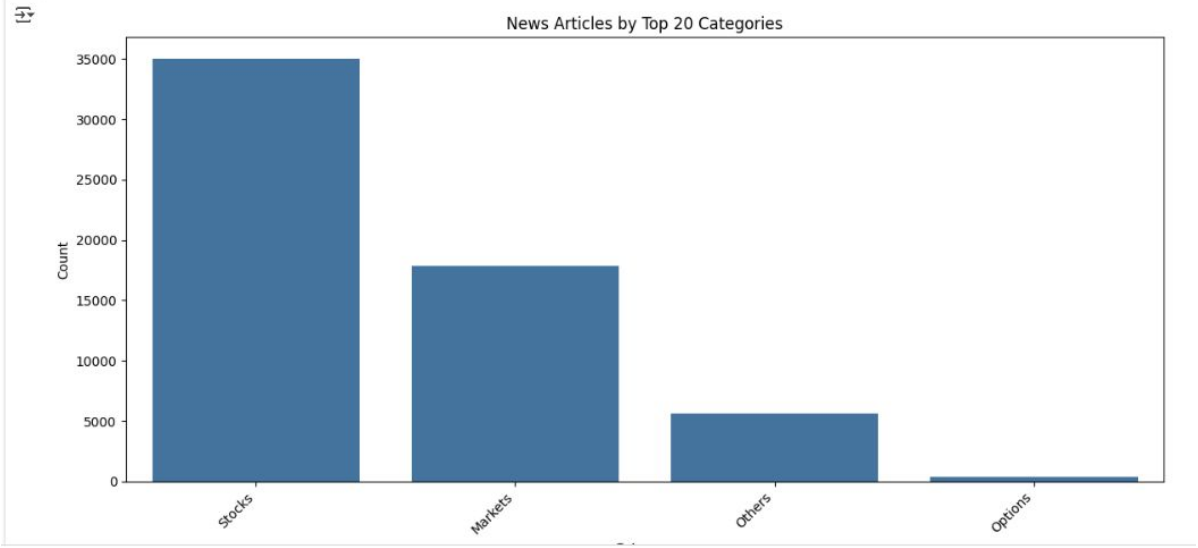
- 영업일 기준으로 날짜 변환
  → 주말 뉴스의 날짜를 다음 영업일로 조정

	rgs_dt	tck_iem_cd	til_ifo	ctgy_cfc_ifo	mdi_ifo	news_smy_ifo	rld_ose_iem_tck_cd	trd_dt
0	2023-01-02	NVIV	France's InVivo to buy one of the world's olde...	Commodities,BioTech,US Markets	Reuters	Malteries Soufflet, a branch of French agribus...	NVIV	2023-01-03
1	2023-01-02	NUVL	Are Medical Stocks Lagging DICE Therapeutics ...	Stocks,Investing	Zacks	Investors interested in Medical stocks should ...	DICE,NUVL	2023-01-03
2	2023-01-02	SCPH	scPharmaceuticals, Inc. (SCPH) Is a Great Choi...	Stocks,Investing	Zacks	"When it comes to short-term investing or trad...	SCPH	2023-01-03
3	2023-01-02	IVVD	Invivyd, Inc. (IVVD) Upgraded to Buy: Here's Why	Stocks,Investing	Zacks	Investors might want to bet on Invivyd, Inc. (...	IVVD	2023-01-03
4	2023-01-02	BELFB	Is Bel Fuse (BELFB) Stock Outpacing Its Comput...	Stocks,Investing	Zacks	For those looking to find strong Computer and ...	BELFB,PERI	2023-01-03
5	2023-01-02	NUVL	Are Medical Stocks Lagging DICE Therapeutics ...	Stocks,Investing	Zacks	Investors interested in Medical stocks should ...	NUVL	2023-01-03
6	2023-01-03	NRXP	Health Care Sector Update for 01/03/2023: XLV,...	US Markets	MTNewswires	Health care stocks were trending higher pre-be...	XLV,XLV,VHT,TXMD,SY,NRXP	2023-01-03
7	2023-01-03	SILO	Pre-market Movers: GRRR, MNPR, KALA, SONX, PEGY···	Markets	RTTNews	(RTTNews) - The following are some of the stoc...	AMAM,AMAM,EVH,GRRR,KALA,KEP,MNPR,PBR,PEGY,RMED...	2023-01-03
8	2023-01-03	SIDU	Technology Sector Update for 01/03/2023: SIDU,...	Technology	MTNewswires	Technology stocks were declining on Tuesday, w...	SIDU,SIDU,TDY,IQ	2023-01-03
9	2023-01-03	SIDU	Technology Sector Update for 01/03/2023: IDCC,...	Technology	MTNewswires	Technology stocks pared a portion of their ear...	IDCC,IDCC,SIDU,TDY,IQ	2023-01-03
10	2023-01-03	SY	So-Young Increases Aggregate Value Of Share Re...	Markets	RTTNews	(RTTNews) - So-Young International Inc. (SY) a...	SY	2023-01-03
11	2023-01-03	NVX	Australian shares commence 2023 lower; gold st...	_	Reuters	Australian shares fell on the first trading se...	NVX,RIO	2023-01-03
12	2023-01-03	RTLPO	The Necessity Retail REIT Announces Common Sto...	_	_	--The Necessity Retail REIT, Inc. announced to...	RTLPO,RTLPP,RTL	2023-01-03
13	2023-01-03	PEGY	Pre-market Movers: GRRR, MNPR, KALA, SONX, PEGY···	Markets	RTTNews	(RTTNews) - The following are some of the stoc...	AMAM,AMAM,EVH,GRRR,KALA,KEP,MNPR,PBR,PEGY,RMED...	2023-01-03
14	2023-01-03	PROF	Recent Price Trend in Profound Medical (PROF) ...	Stocks,Investing	Zacks	"Most of us have heard the dictum ""the trend ...	PROF	2023-01-03
15	2023-01-03	RDHL	RedHill Biopharma Reports Positive Data From P...	Markets	RTTNews	(RTTNews) - RedHill Biopharma Ltd. (RDHL) anno...	RDHL	2023-01-03
16	2023-01-03	ASLN	ASLAN, Thermo Fisher Join To Manufacture High ...	Markets	RTTNews	(RTTNews) - ASLAN Pharmaceuticals (ASLN) and T...	ASLN,ASLN,TMO	2023-01-03
17	2023-01-03	ATAT	Sore Thumb Indicator: 3 Chinese ADRs Pass the ...	Investing,Stocks	Zacks	With thousands of stocks open for trading each...	YUM.MOMO.KWEB.BILI.PDD.OFIN.COIN.HOOD.ATAT	2023-01-03

EDA 요약

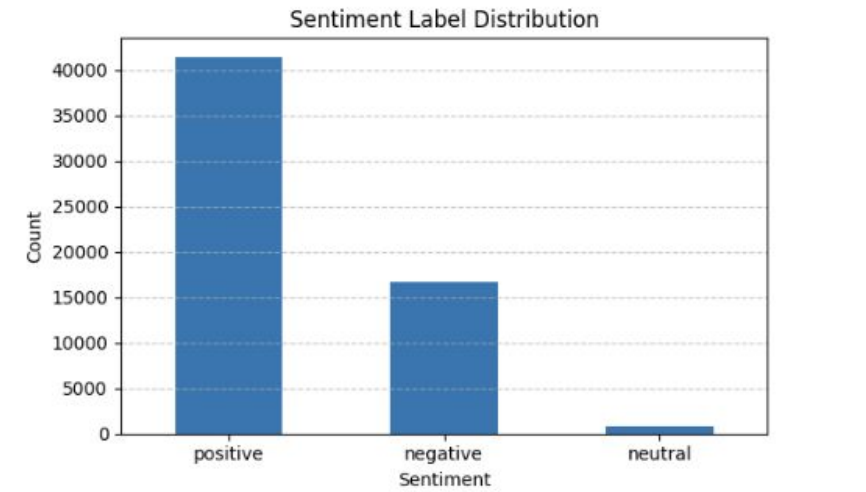
```python
# 1) 텍스트 전처리
text_data = " ".join(news_2['til_ifo'].astype(str).tolist())
text_data = re.sub(r'[^a-zA-Z ]', '', text_data.lower())

# 2) WordCloud에 사용할 stopwords 정의
my_stopwords = set(STOPWORDS)
# 내장된 영어 불용어 외에 추가로 제거할 단어
my_stopwords.update(["to", "for", "the", "and", "of", "in", "on", "q", "stocks", "stock", "earnings"])

# 3) 워드클라우드 생성 시 stopwords 인자로 전달
wordcloud = WordCloud(
    width=800,
    height=400,
    background_color='white',
    stopwords=my_stopwords  # 불용어 세트
).generate(text_data)

# 4) 시각화
plt.figure(figsize=(10, 5))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.title('Word Cloud (with Stopwords Removed)')
plt.show()
```


```python
# 카테고리별 기사 수 계산
category_counts = news_2['CategoryGroup'].value_counts()

# 예: 상위 20개 카테고리만 추출
topN = 20
top_categories = category_counts.head(topN).index

# 상위 20개 카테고리에 해당하는 데이터만 필터링
news_2_topN = news_2[news_2['CategoryGroup'].isin(top_categories)]

plt.figure(figsize=(12, 6))
sns.countplot(data=news_2_topN, x='CategoryGroup', order=top_categories)
plt.title(f'News Articles by Top {topN} Categories')
plt.xlabel('Category')
plt.ylabel('Count')
plt.xticks(rotation=45, ha='right')  # 레이블 45도 회전, 오른쪽 정렬
plt.tight_layout()
plt.show()
```



EDA 요약

```
import matplotlib.pyplot as plt

# 감성 레이블 분포 시각화
sentiment_counts = news_2['sentiment'].value_counts()

plt.figure(figsize=(6,4))
sentiment_counts.plot(kind='bar')
plt.title('Sentiment Label Distribution')
plt.xlabel('Sentiment')
plt.ylabel('Count')
plt.xticks(rotation=0)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```



Word Cloud (뉴스 제목 텍스트 기반)

상위 키워드: report, recommendation, buy, revenue, estimates 등

긍정적 투자 관련 키워드가 중심에 있음 → 종가 등 반응 변수와 연계 가능성

카테고리별 뉴스 분포

Stocks 관련 뉴스가 절반 이상 차지

Markets, Options 등 세부 주제별 차이 큼 → 카테고리 기반 피처의 영향력 예상됨

감성 분석 결과 분포

positive 비중이 가장 높음 → 전반적으로 낙관적 보도 경향

negative, neutral은 상대적으로 적음 → 뉴스의 편향 여부도 고려 필요

Feature Engineering

```
[45] # CategoryGroup
     # ctry_cfc_info 컬럼에서 특정 키워드를 찾아 대분류(Stocks, Markets, Options 등)로 분류
     def map_category_group(x):
         if not isinstance(x, str):
             return 'Others'
         x_lower = x.lower()
         if 'stocks' in x_lower:
             return 'Stocks'
         elif 'markets' in x_lower:
             return 'Markets'
         elif 'options' in x_lower:
             return 'Options'
         elif 'futures' in x_lower:
             return 'Futures'
         else:
             return 'Others'

     news_2['CategoryGroup'] = news_2['ctgy_cfc_ifo'].apply(map_category_group)

     # HasBuyWord / HasSellWord
     # 제목(til_ifo)에 'buy', 'sell', 'short' 키워드가 포함되어 있는지 여부를 0/1로 표시
     news_2['HasBuyWord'] = news_2['til_ifo'].str.contains(r'\bbuy\b', case=False, na=False).astype(int)
     news_2['HasSellWord'] = news_2['til_ifo'].str.contains(r'\bsell\b', case=False, na=False).astype(int)
     news_2['HasShortWord'] = news_2['til_ifo'].str.contains(r'\bshort\b', case=False, na=False).astype(int)

     # 결과 확인
     news_2.head()
```

Category Group
: 다중 카테고리의 군집 구조 단순화
→ 카테고리별 뉴스 트렌드 파악 및 모델 입력 간소화

Has Signal Word
: 투자 관련 핵심 키워드 반영 긍·부정 신호 포착
→ 뉴스가 시장에 미치는 긍·부정 영향력 정량적으로 분석 가능

	rgs_dt	tck_iem_cd	til_ifo	ctgy_cfc_ifo	mdi_ifo	news_smy_ifo	rld_ose_iem_tck_cd	trd_dt	date	weekday	CategoryGroup	HasBuyWord	HasSellWord	HasShortWord
0	2023-01-02	NVIV	France's InVivo to buy one of the world's olde...	Commodities,BioTech,US Markets	Reuters	Malteries Soufflet, a branch of French agribus...	NVIV	2023-01-03	2023-01-02	0	Markets	1	0	0
1	2023-01-02	NUVL	Are Medical Stocks Lagging DICE Therapeutics ...	Stocks,Investing	Zacks	Investors interested in Medical stocks should ...	DICE,NUVL	2023-01-03	2023-01-02	0	Stocks	0	0	0
2	2023-01-02	SCPH	scPharmaceuticals, Inc. (SCPH) Is a Great Choi...	Stocks,Investing	Zacks	"When it comes to short-term investing or trad...	SCPH	2023-01-03	2023-01-02	0	Stocks	0	0	0
3	2023-01-02	IVVD	Invivyd, Inc. (IVVD) Upgraded to Buy: Here's Why	Stocks,Investing	Zacks	Investors might want to bet on Invivyd, Inc. (...	IVVD	2023-01-03	2023-01-02	0	Stocks	1	0	0
4	2023-01-02	BELFB	Is Bel Fuse (BELFB) Stock Outpacing Its Comput...	Stocks,Investing	Zacks	For those looking to find strong Computer and ...	BELFB,PERI	2023-01-03	2023-01-02	0	Stocks	0	0	0

다음 단계: news_2 변수로 코드 생성 | 추천 차트 보기 | New interactive sheet

결과 요약 및 향후 계획

한계점

1. 뉴스 본문 부재
   → 감성 분석의 정밀도 한계
2. 실제 주가 데이터 미포함
   → 종속 변수로 활용 불가

인사이트

1. 제목만으로도 감정 성향과 투자 시그널을 추출 가능
2. Stocks와 같은 주제별 기사 집중도 차이
   → 카테고리 기반 모델링 가치 있음

향후 계획

1. 시계열 데이터 결합
   → 주가/변동률 변화 예측 모델 고도화
2. 뉴스 외부 데이터 (예: 종가, 거래량, 공시 등) 연동
3. 카테고리 세분화 및 감성 레이블 수작업 정제

뉴스 기반 투자 분석 → 텍스트 감성 + 시계열 반응 예측 통합 모델로 확장

QnA