

SA for Finance

IFin

김순호

목차

1. 프로젝트 개요
2. 프로젝트 팀 구성 및 역할
3. 프로젝트 수행 절차 및 방법
4. 프로젝트 수행 결과
5. 자체 평가 의견

1. 프로젝트 개요

- 주식정보, 신문기사, 투자자의 투자 동향, 신문기사에 대한 소비자의 의견(해시태그, X) 등을 종합적으로 판단하여 주식에 대한 정보와 소비자들의 감성(Sentiment)의 영향을 분석 => OSINT기반의 정보를 활용한 주식예측측
- 다양한 데이터(Multimodal Data) 형태를 통합(Data Fusion)
- 감성분석이 주가에 미치는 영향의 크기와 시점 등을 분석

02. 프로젝트 팀 구성 및 역할

- 팀명 : IFin
김순호 단독 팀으로 구성
- 주식시장분석 방법과 가치평가 방법에 대한 조사
- 감성분석 문헌연구
- Python Code

03. 프로젝트 수행 절차 및 방법

- Finance Prediction 방법론
- 감성분석 방법론
- 감성분석의 영향
- 데이터 융합 방법론
- AI Agent vs RAG 영향

3.1 기존 주가 예측 방법론

- **기본적 분석 (Fundamental Analysis)**

기업의 재무제표, 성장성, 산업 전망, 거시경제 지표 등을 분석

예: 매출 증가율, 이익률, 금리, 환율, 경기 사이클

- **기술적 분석 (Technical Analysis)**

과거 가격과 거래량 데이터를 기반으로 패턴을 찾음

예: 이동평균선, RSI, MACD, 캔들차트 패턴

- **계량적 모델 (Quantitative Models)**

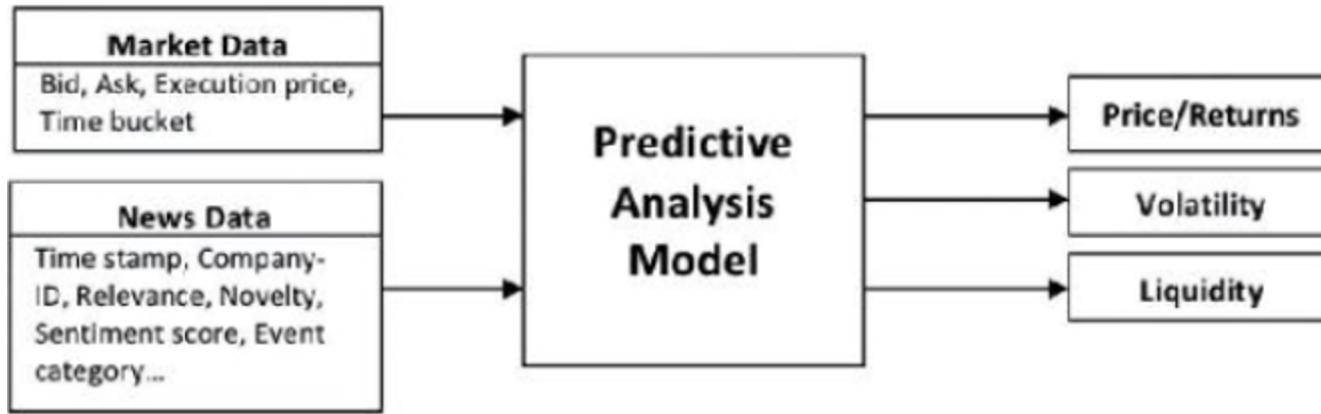
통계, 머신러닝, 인공지능을 활용해 데이터 기반 예측

예: 시계열 분석(ARIMA), 딥러닝 기반 예측

- **행동재무학적 접근 (Behavioral Finance)**

투자자 심리와 군집행동(허딩)을 고려

예: 뉴스, 소셜미디어, 투자자 감정 지수



3.2 서비스 구축 컨셉

- 증권과 관련된 최근 추세정보 수집과 분석 그리고 활용 DB구축
- LLM기반의 전문적 FinTech기반 대화형 지식 제공
- 주식 용어를 시작으로 투자 원칙 등의 설명과 적용
- 트렌드 분석을 위한 국내외 OSINT를 적용하는 시스템
- OSINT기반의 페이크뉴스 검출

3.3. 시장분석

The screenshot shows the homepage of Quiver Quantitative. At the top, there's a navigation bar with links for Home, Datasets, Strategies, Premium, Pricing, More, a search bar, and sign-in/join buttons. Below the header, a large banner features the text "Trade Like an Insider" and "Track the forces that move the markets". It includes a search bar and a section titled "Trending Now On Quiver:" with cards for Nancy Pelosi, Marjorie Taylor Greene, J. D. Vance, Congress Trading, Insider Trading, Institutional Holdings, and stock market data for NVIDIA, UNH, and GME. The main visual is a stylized 3D rendering of a person standing in front of a large screen displaying financial charts and data. At the bottom, there are sections for "Trending" news and "Just Disclosed New Holdings".

<https://www.quiverquant.com>

1. Quiver Quantitative

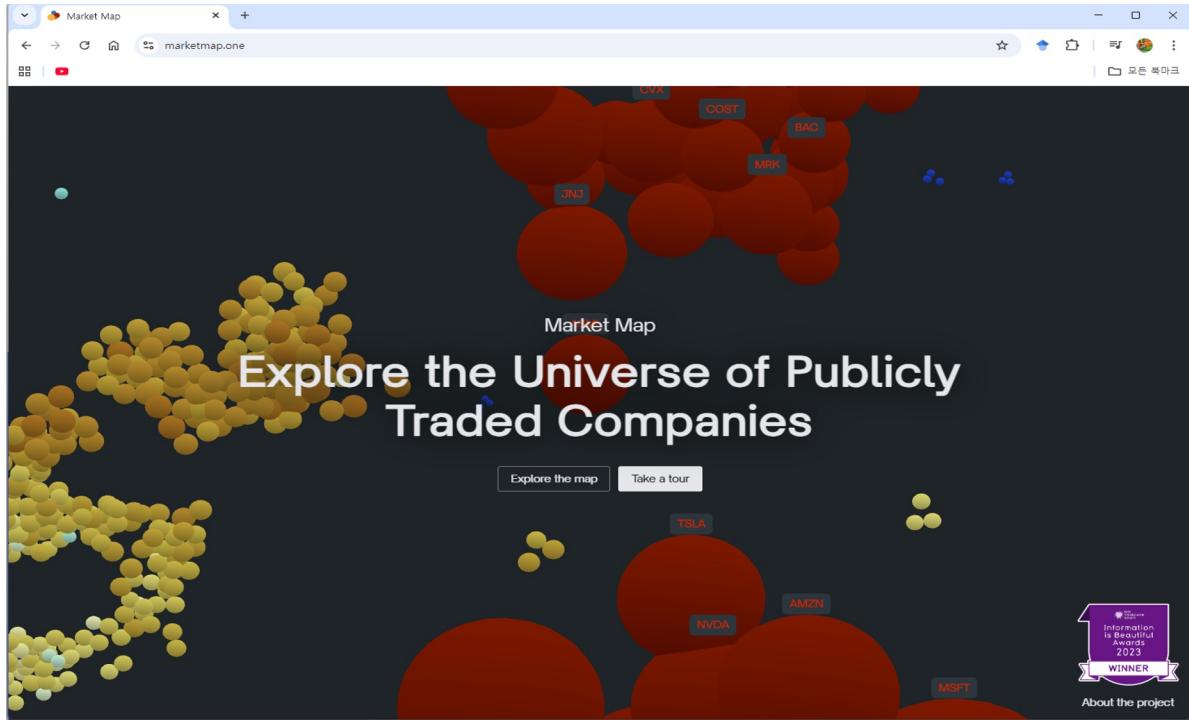
- **특징:** 미국 상장기업에 대한 비정형 공개정보를 시각화
- **데이터 출처:**
 - 정치인 거래 내역
 - 정부 계약
 - 로비 활동
 - 특허 출원
 - Reddit, Twitter 등 소셜미디어
- **장점:** 투자자들이 접근하기 어려운 정보들을 구조화하여 제공

The screenshot shows the homepage of the AlphaSense website. At the top, there's a navigation bar with links for 'AlphaSense', 'Platform', 'Solutions', 'Resources', 'About' (which is underlined), and 'Pricing'. To the right of the navigation are 'Log In', 'Customer Support', and a 'Start Free Trial' button. Below the navigation, a large banner features the text 'Accelerate your workflow with AI insights you can trust'. Underneath the banner, a blue section contains the text: 'Streamline your workflow with the most advanced enterprise-grade platform — designed for financial and business professionals who need to move quickly and make confident decisions.' It includes two buttons: 'Get started for free →' and 'Take the Tour'. A sidebar on the left says 'AlphaSense' and 'We value your privacy', followed by a cookie consent message and three buttons: 'Customize', 'Reject All', and 'Accept All'. On the right side, there's a screenshot of the Deep Research AI interface showing a conversation with an AI representative named Alphred.

<https://www.alpha-sense.com/>

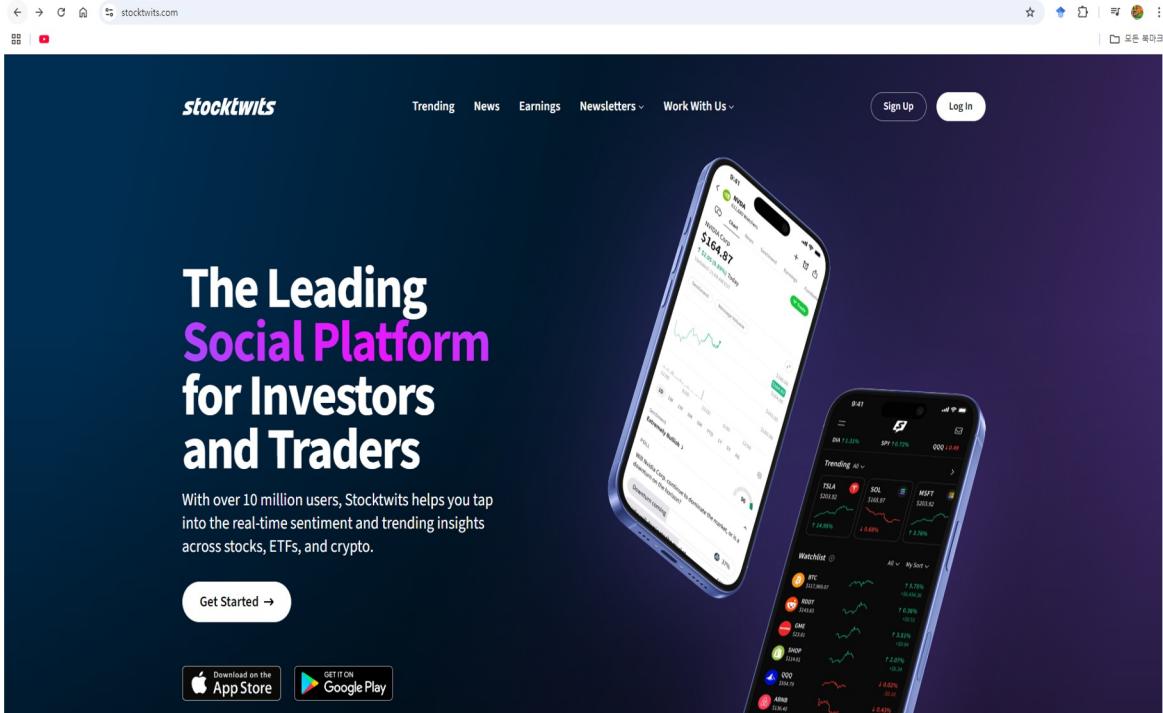
2. AlphaSense

- 특징: 기업 보고서, 뉴스, 애널리스트 리포트 등 방대한 문서 검색
- 기술: 자연어 처리 기반의 의미 검색(Semantic Search)
- 대상: 기관 투자자, 애널리스트, 기업 전략 담당자
- 장점: 실시간 정보 추적 및 경쟁사 분석에 강점



3. <https://marketmap.one/>

- **특징:** 뉴스, 소셜미디어, 공시자료 등에서 투자 신호 추출
- **기능:**
 - 감성 분석(Sentiment Analysis)
 - 이벤트 기반 알림
 - 종목별 리스크 요인 시각화
- **장점:** AI 기반으로 투자 타이밍 포착에 도움



<https://stocktwits.com/>

4. 스택위츠

- **Cashtag 시스템:** \$TSLA처럼 종목 앞에 \$를 붙여 특정 종목에 대한 대화를 쉽게 모아볼 수 있음
- **실시간 시장 심리(Sentiment):** 투자자들의 긍정/부정 의견을 집계해 시장 분위기를 한눈에 확인 가능
- **트렌드 탐색:** 어떤 종목이나 코인이 주목받고 있는지 실시간으로 확인
- **AI 기반 인사이트:** 커뮤니티 데이터를 분석해 주요 종목과 투자 아이디어를 추천
- **모바일 앱 지원:** iOS와 Android에서 무료로 사용 가능

서비스 유형 분류

플랫폼/기업	주요 기능 및 특징
Stocktwits	소셜 미디어 기반 투자자 의견 공유 플랫폼. 실시간 감정 흐름 파악 가능.
Sentimentrader	뉴스, 트위터, 포럼 등에서 투자자 심리를 분석해 시장 타이밍 도구로 활용.
RavenPack	AI 기반 뉴스 감정 분석으로 금융 시장 반응 예측. 기관 투자자 중심.
TrendSpider	기술적 분석 도구에 감정 지표를 결합. 트레이더에게 실시간 심리 흐름 제공.
MarketSmith	투자자 행동 분석과 시장 심리 지표 제공. 주식 선택과 타이밍에 도움.
Permutable.ai	증권은행 발표, 원자재 가격 변화에 대한 감정 반응 분석. 전략적 투자 판단에 활용.

- 각 정보 원천에 따라 시황에 대한 예측을 제시하는 BM
- 뉴스, SNS 등을 기반으로 하는 방식
- AI기반의 뉴스 감정분석 방식
- 기술분석과 감정지표를 융합하는 방식
- 투자자 행동분석과 결합 방식
- 거시경제 지표를 기반으로 하는 방식

국내현황

- 외국 사례는 기업형태로 존재하나, 국내의 경우 공공데이터, DART전자공시스템, KIPRIS특허정보검색서비스, 그리고 크리딧잡/사람인과 같은 형태로 존재함
- 국내에도 기업형태의 증시 정보를 OSINT형태로 제공할 수 있는 기업이 필요함

주요 한국어 Lexicon

이름	출처	특징	활용 예시
KnuSentiLex (KNU 한국어 감성사전)	군산대학교 Data Intelligence Lab	보편적 긍·부정 감성어 포함. 이모티콘까지 감성어로 등록. GitHub에서 무료 제공	감성 분석, 텍스트 마이닝, SNS 데이터 분석
부산대 Sentimental Dictionary	부산대학교	한국어·영어·중국어 감성사전 제공. 구축 방법과 모델링 방식 공개	다국어 감성 분석 연구
표준국어대사전 기반 확장 감성어 사전	국립국어원 자료 활용	표준국어대사전에서 추출된 감성어 + 축약어·신조어·이모티콘 포함	빅데이터 분석, 소셜 미디어 감성 연구

04. 프로젝트 수행 결과

4.1 정보이론

약성, 준강성, 강성(이론은 정보가 시장가격에 즉시 반영된다는 가정을 기초로 함)

=> 정보의 공개 시점과 정보의 영향시점의 차이

=> 정보는 공개되는 즉시 시장에 영향을 미친다. 다만 인지의 주체가 누구인가에 따라 영향이 다를 것인가?

=> 때문에 역할자의 수집능력, 적용능력, 개인능력 차 등에 의해 시장에 대한 영향 차이 발생생

4.2 배경

- 주식시장 동인(Market Drivers)

Market drivers are the factors that influence financial markets, such as economic indicators, corporate performance, geopolitical events, central bank policies, and **technological advancements**.

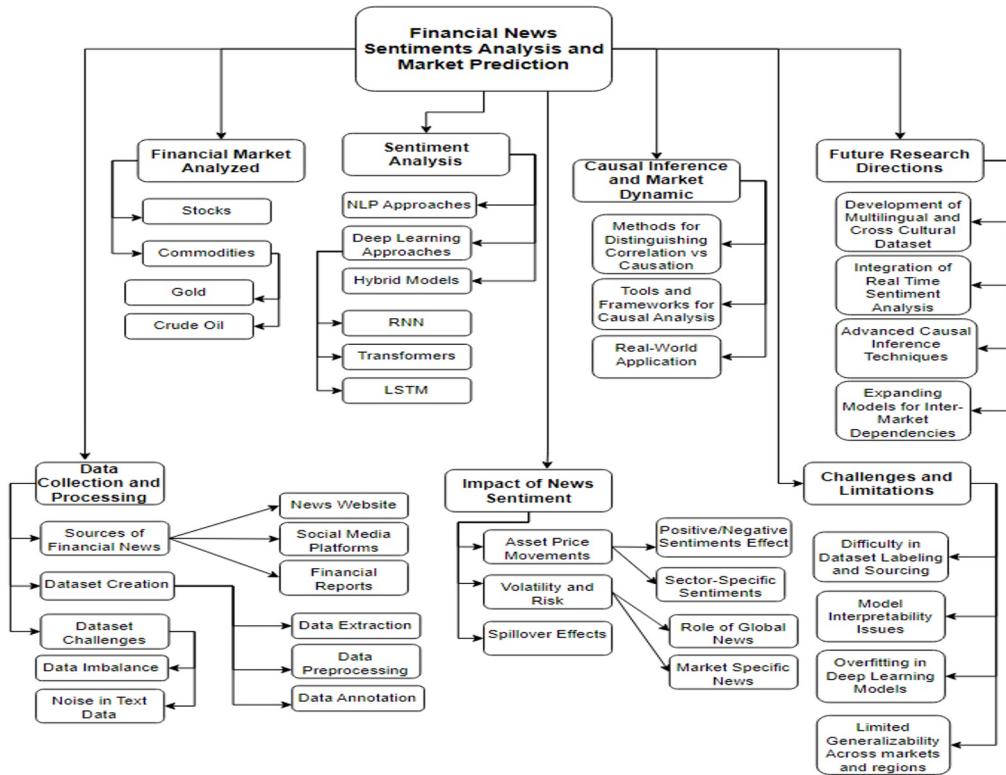
- 기존 주식의 예측 방법 : 전문가 중심의 판단으로 그들의 개인적 결정과 집단적인 평가에 의해 주도되거나 하기도 함
- Herding Forecast and Bold Forecast
- AI Agent를 기반으로 Argument Mining을 통한 Opinion의 영향에 따른 향후 주가 변동을 예상
- 투자 시점의 예측과 시장탈출 시점의 예측
- 시장의 상승과 하락 장의 예측 => 시장 변화의 시점 예측

Herding Forecast vs Bold Forecast

구분	Herding Forecast	Bold Forecast
정보 반영	기존 컨센서스 중심, 사적 정보 적음	사적 정보 적극 반영
시장 기여도	낮음 (새로운 정보 부족)	높음 (새로운 인사이트 제공)
리스크	낮음 (집단에 묻힘)	높음 (개인 책임 증가)
정확성	상대적으로 낮음	상대적으로 높음 ②
발생 조건	불확실성 낮을 때, 경력 위험 회피	경험 많고 과거 정확도 높은 분석가일 때

- **투자자 입장:** Herding Forecast는 참고용으로만 활용하고, Bold Forecast는 새로운 정보와 기회를 제공할 수 있으므로 더 주목할 가치가 있음.
- **분석가 입장:** Bold Forecast는 차별화된 전문성을 보여줄 수 있지만, 실패 시 평판 리스크가 크므로 신중한 판단이 필요.

AI Sentiment Analysis 분류



- 주식시장 정보를 수집, 분류, 분석하는 과정과 분류 및 분석 방법의 체계도

감성분석 vs. 기술분석

구분	Sentiment Feature	Technical Feature
데이터 출처	뉴스, 보고서, SNS	가격, 거래량, 차트
분석 방법	NLP, 감정 분석	수학적 계산, 통계
강점	심리 반영, 방향성	추세·패턴 반영, 매매 신호
약점	주관성, 변동성	과거 데이터 의존, 외부 요인 반영 부족

Sagala, Tommy Wijaya, Mei Silviana Saputri, Rahmad Mahendra, and Indra Budi. 2020. "Stock Price Movement Prediction Using Technical Analysis and Sentiment Analysis." *Proceedings of the 2020 2nd Asia Pacific Information Technology Conference*, January 17, 123–27. <https://doi.org/10.1145/3379310.3381045>.

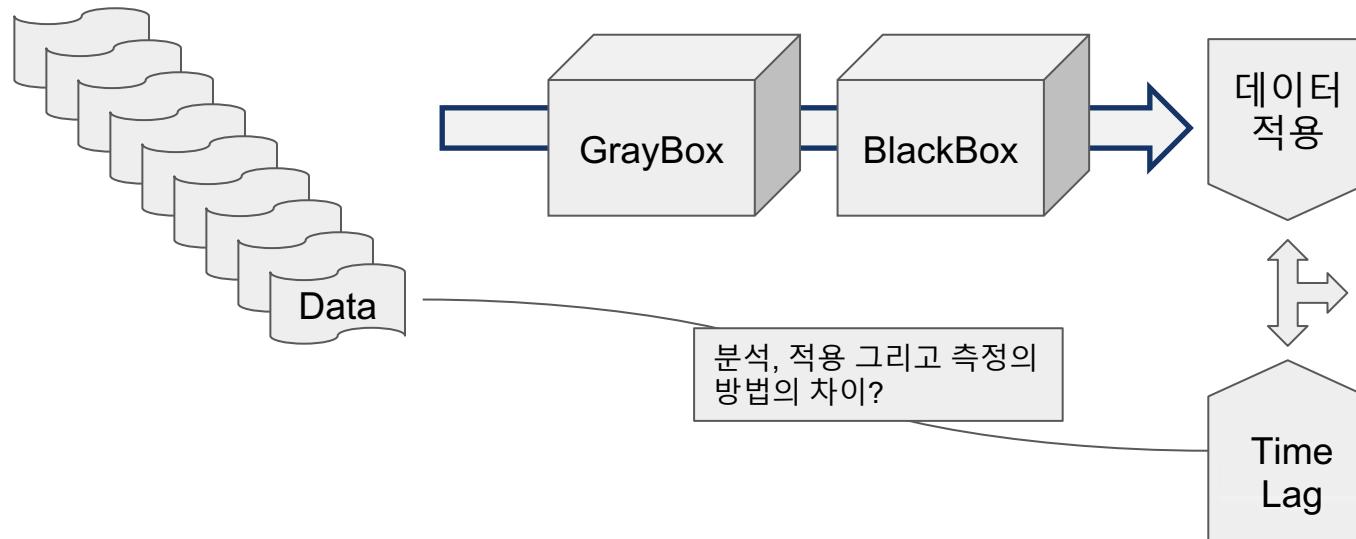
Impact Duration of Events

“Patton and Verardo (2012) noticed decay in the impact of news on asset prices and their betas on a daily timescale and further determine the complete disappearance of news effects within 2-5 days.” (MITRA, 2016, p. 9)

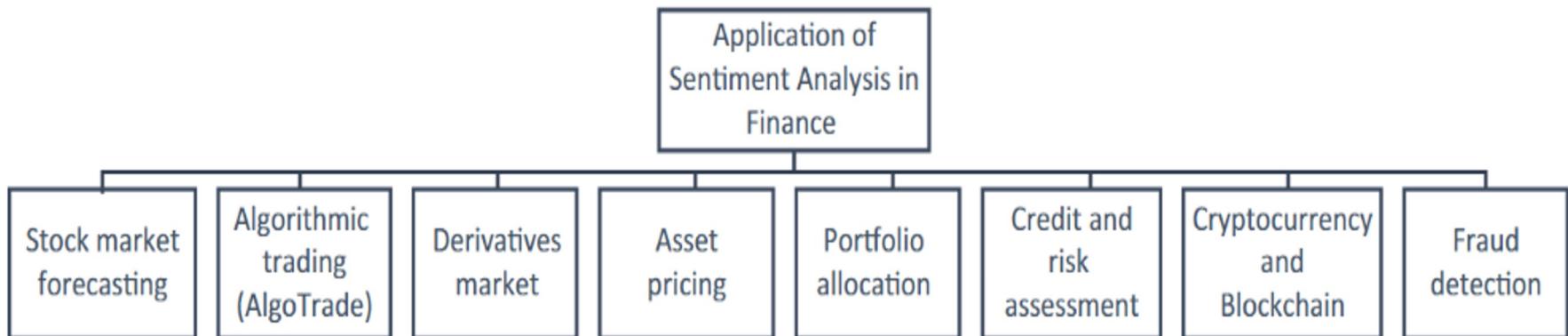
=> 감성분석의 영향은 2-5일에서 길어야 1주일이다. 이 경우 영향을 지속적으로 판단하는 방법을과 과정을 반복해야하는 문제가 발생한다. 이 문제를 어떻게 해결하는가에 따라 달리 분석의 영향을 진행할 수 있을 것이다.

- 감성과 관련된 데이터의 출현 이후 영향을 분석하기 위한 시차의 적용과 실제적인 오류의 조정문제가 발생
- 프로그램 상에서 이를 구체적 구현가능성과 영향에 대한 불확신이 존재함

데이터와 효과의 괴리(Time Lag)



감성분석 적용 분야



어의의 차이(사전적의미와 금융에서 의미 차이)

- Three-quarters of the negative words in the Harvard Dictionary are not negative words in financial narrative [7].
- Bullish words in the financial domain are sometimes labeled as neutral words in general sentiment dictionaries [1].
- Positive sentiment does not always lead to bullish market sentiment [2].

재무분야에서 사용되는 단어의 의미와 실사용되는 단어에서의 어의적인 차이가 존재함

즉 사용되는 단어 유의적인 의미 괴리가 존재함

해석하는 방법과 시장경험 부족한 경우 판단오류의 원인

금융감성사전 1

- <https://github.com/ntunlplab/Finance-NTUSD-Fin>
- NTUSD-Fin은 *National Taiwan University Sentiment Dictionary for Finance*의 약자로, 금융 소셜 미디어 데이터를 기반으로 구축된 시장 감성 사전입니다. 이 사전은 330,000개 이상의 라벨링된 금융 소셜 미디어 게시물을 크롤링하여 구축되었으며, 총 8,331 개의 단어, 112개의 해시태그, 115개의 이모지가 포함되어 있습니다.
- 데이터는 GitHub 저장소에서 공개되어 있으며, **NTUSD-Fin.zip** 파일로 다운로드할 수 있습니다.
- 저장소에는 감성 점수 계산 방식(빈도, CFIDF, 카이제곱 검정, 시장 감성 점수, 워드 벡터 등)에 대한 설명도 포함되어 있습니다.

금융감성사전 2

FinSoSent: Advancing Financial Market Sentiment Analysis through Pretrained Large Language Models

- 출처: University of Pennsylvania, Pennsylvania State University, Western New England University 연구진 발표: *Big Data and Cognitive Computing* 저널, 2024년 8월호
- 목적: 금융 시장의 방향성을 예측하기 위해 **사전학습된 대형 언어 모델(LLM)**을 활용한 감성 분석
- 특징:
 - 기존 금융 감성 사전 기반 접근을 넘어, **뉴스·소셜 미디어 데이터**를 통합적으로 분석
 - 시장 예측 정확도를 높이기 위해 LLM 기반 감성 점수 산출
 - 금융 데이터의 **비정형성(unstructured data)**을 처리할 수 있도록 설계

SA Dataset

데이터셋	설명	GitHub 주소
NTUSD-Fin	대만대학에서 구축한 금융 소셜 미디어 기반 감성 사전. 단어·해시태그·이모지 포함.	Finance-NTUSD-Fin
Financial Phrase Bank	헬싱키 대학에서 만든 금융 뉴스 문장 라벨링 데이터셋 (긍정·부정·중립).	FinancialPhraseBank
FinSenticNet	SenticNet 프로젝트에서 확장한 금융 감성 사전. 개념 수준 감성 분석 지원.	FinSenticNet
FiQA Dataset	금융 Q&A와 뉴스 헤드라인 기반 감성·주제 라벨링 데이터셋.	FiQA

<https://github.com/SenticNet/finSenticNet>

데이터셋	설명	출처
AG's News	496,835개의 뉴스 기사에서 4개 대분류 선택 (제목+설명 사용)	<u>AG Corpus of News Articles</u>
Sogou News	중국어 뉴스 코퍼스 (SogouCA + SogouCS), 2,909,551개 기사. 5개 카테고리 선택 후 Pinyin 변환	Sogou 뉴스 데이터셋
DBpedia Ontology	Wikipedia에서 추출된 DBpedia 2014 버전, 14개 클래스 선택	<u>DBpedia</u>
Yelp Reviews	Yelp Dataset Challenge 2015 제공 리뷰 데이터. 별점 예측과 긍·부정 polarity 예측 두 가지 과제 구성	<u>Yelp Dataset Challenge</u>
Yahoo! Answers	Yahoo! Webscope 프로그램 제공 Q&A 데이터셋. 10개 주요 카테고리 선택	<u>Yahoo! Webscope</u>
Amazon Reviews	Stanford SNAP 프로젝트 제공. 18년간 34,686,770개의 리뷰 포함. 별점 예측과 polarity 예측 두 가지 과제 구성	<u>Stanford SNAP Amazon Review Dataset</u>

http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html

<https://www.dbpedia.org/>

<https://business.yelp.com/data/resources/open-dataset/>

<https://snap.stanford.edu/data/web-Amazon.html>

Fin-SoMe

Issues and perspectives from 10,000 annotated financial social media data

<https://github.com/ntunlplab/Finance-Fin-SoMe>

Fin-SoMe 데이터 개요

- 데이터 규모: 10,000개 금융 트윗
- 주석 방식:
 - 은행 트레이저리 부서의 **프론트 데스크(front desk)**와 **미들 데스크(middle desk)** 전문가들이 감성 및 시장 관련성을 라벨링
 - **작성자 감성(writer sentiment)**과 **시장 감성(market sentiment)**을 구분하여 주석
- 발견된 특징:
 - 작성자가 직접 표시한 시장 감성은 종종 **잘못된 라벨**일 수 있음
 - 작성자의 감성과 실제 투자자의 시장 감성이 다를 수 있음
 - 대부분의 금융 트윗은 **근거 없는 분석 결과**를 제공
 - 투자자들이 자신의 포지션에 대한 손익 결과를 거의 기록하지 않음

=> 메일을 통해 허가를 득해야 한다(미수집)

FINGPT

DR JIGNASHA SHAH DALAL, DR SANTHILATA K. V. 2025. *ULTIMATE FINGPT FOR FINANCIAL ANALYSIS*. ORANGE EDUCATION PVT LTD.

<https://huggingface.co/FinGPT>

<https://github.com/AI4Finance-Foundation/FinGPT/fork>

- 개발 주체: [AI4Finance Foundation](#) (비영리 단체) 목표: 금융 데이터와 대형 언어 모델을 결합해 금융 AI 생태계의 민주화
- 특징:
 - 오픈소스로 누구나 접근 가능
 - Hugging Face에 모델 배포 → 연구자와 개발자가 쉽게 활용 가능
 - 금융 데이터에 맞춘 **파인튜닝(fine-tuning)**과 LoRA(저자원 적응 학습) 지원
 - 투자자 감성 분석, 금융 뉴스 요약, 주식 시장 예측 등 다양한 응용 가능

금융 평가 데이터

sources of financial opinions; we group these sources by the opinion holders: insiders (Sect. 3.1), professionals (Sect. 3.2), social media users (Sect. 3.3), and journalists (Sect. 3.4).

Opinion Holder

- Opinion Source : 관리자, 전문가, SNS사용자, 저널리스트
- 전문가 : 투자 전문가, 아마츄어 전문가
- 역대 성과자 : 투자성공자와 실패자

시장감성분석과 AI 방법론

- Cortis et al. [21] annotate market sentiment scores from -1 to 1 on both social media data and news articles, and publish an annotated dataset for SemEval-2017 Task 5. Jiang et al. [34] augment word2vec embeddings [56] with n-gram, part-of-speech, word cluster, sentiment lexicon, numeral, metadata, and punctuation features. Their ensemble model performed the best in the SemEval-2017 Task 5 social media data track. Mansar et al. [54] achieved the first place in the news article track with a convolutional neural network with features extracted using VADER [32], a rule-based sentiment analysis toolkit.
- Gaillat et al. [30] concatenate (1) the output of a long short-term memory architecture (LSTM) for encoding tweets, (2) LSTM output for the word embedding with general sentiment features, (3) VADER output, and (4) the sentiment degree from the AFINN word list [57] as features. Their model outperforms that of Jiang et al. [34] on the financial social media sentiment analysis task.
- Xing et al. [79] compare the performance of dictionary-based methods and machine learning models on the Yelp dataset [83] and their StockSen dataset. They find that all models make incorrect predictions, and point out several error types, including irrealis mood, rhetoric, dependent opinion, unspecified aspects, unrecognized words, and external references.
- Yuan et al. [82] publish a Chinese news dataset for target-based sentiment analysis, and compare the performance of several baselines. On their dataset, BERT achieves an F1 score of 79.84%.

- 감성분석은 -1 에서 1 까지 값
 - n-gram, POS 등을 적용
 - 규칙기반 분석 - VADER 적용
 - LSTM, BERT를 적용
 - Yelp, StockSen Dataset 등 활용
- => 7-80%대의 F1값을 보임

Market mood

Indicator	Data Analyzed	Sentiment Indication
Volatility Index (VIX)	Measures expected market volatility based on S&P 500 option prices.	High VIX = Fear (known as the "fear gauge"). Low VIX = Complacency/Optimism .
Put/Call Ratio (PCR)	Compares the trading volume of put options (bets on decline) to call options (bets on rise).	High PCR (more puts) = Bearish/Fear . Low PCR (more calls) = Bullish/Greed .
Market Breadth	Tracks the number of stocks advancing versus declining, or reaching new highs versus new lows.	Broad participation (many advancing stocks) = Strong Optimism . Narrow participation = Caution/Fear (hidden weakness).
Momentum and Volume	Analyses the speed of price movements and the volume associated with them.	Strong price rise on high volume = Conviction/Greed . Sharp drop on high volume = Panic/Fear .

Market Mood Index

The most popular example is the **Market Mood Index (MMI)** or the **Fear and Greed Index**.

This index typically scores the market mood on a scale (e.g., 0 to 100) and categorizes it:

Score Range	Market Mood	Interpretation
0 - 30	Extreme Fear	Oversold conditions, panic selling. Contrarian signal to Buy.
31 - 50	Fear/Caution	Negative sentiment, defensive positioning.
51 - 70	Optimism/Greed	Growing confidence, prices may be climbing too fast.
71 - 100	Extreme Greed	Overbought, euphoria, risk ignored. Contrarian signal to Sell/Hedge. ⁶

각 국의 데이터 출처 분류

주식 시장/지수	데이터 출처
독일 (DAX)	Bundesbank ²
대만 (TAIEX, TWSE)	Yahoo Finance, Capital Futures Corporation (Taipei), Bloomberg, Taiwan Economic Journal ³
일본 (NIKKEI 225)	Yahoo Finance, Bloomberg ⁴
터키 (ISE National100, BIST100)	Matriks Information Delivery Services and Electronic Data Delivery System, Central Bank of the Republic of Turkey (CBRT), Bloomberg ⁵
인도 (S&P BSE SENSEX, NIFTY 50)	Yahoo Finance, Moneycontrol, NSE website, BSE website, Quandl ⁶
미국 (S&P 500, DJIA, IXIC)	Yahoo Finance, US Energy Information Administration, https://www.google.com/search?q=OANDA.com ⁷
중국 (SSE composite index 등)	Shanghai stock exchange, Bloomberg, Stockstar, Wind database ⁸
홍콩 (HS300, HSI)	Bloomberg, Yahoo finance, Tradingeconomics ⁹
호주 (ASX200 등)	Resset database, Norgate investor services ¹⁰
멕시코	Bloomberg ¹¹
태국, 인도네시아, 필리핀 (SET50)	Thailand stock exchange ¹²
대한민국 (KOSPI, KOSPI 200)	Trading economics, Korea Securities Computing Corporation (KOSCOM), Bloomberg ¹³
크로아티아 (CROBEX)	Zagreb stock exchange website ¹⁴
비트코인 환전 (BTC)	Kaggle ¹⁵
그리스 (ASE General Index)	Bloomberg, Trading economics ¹⁶
영국 (FTSE)	Yahoo finance, London stock exchange ¹⁷

- 각 국의 주식시장 지수와 데이터의 출처가 상이함
- 즉 데이터셋의 구성의 차이에 의해 실제적으로 발생할 수 있는 감성분석에 있어서 차이가 존재할 수 있음

데이터 수집과 피처링 문제

각기 분산된 데이터 예를 들어 주제, 기업에 따라 달리 검색되는 형태를 주제의 유사성과 기업에 따른 산업군 그리고 다시 이를 종합적으로 분석하여 데이터를 가공할 수 있는 형태로 만들 수 있는가?

API를 이용한다 하더라도 유용한 정보를 취득하는 것과 취합과 응용에는 어떠한 방법이 있는가?

검색 수가 증가한다는 것이 반드시 긍정적 의미를 주는 것은 아니다. . Trends to sell on the financial market at lower prices may be preceded by periods of concern. During such periods of concern, people may tend to gather more information about the state of the market. It is conceivable that such behavior may have historically been reflected by increased Google Trends search volumes for terms of higher financial relevance.에서 보는 바와 같이 불안한 요소 또는 낮은 가격에 판매를 고민하는 경우에도 검색량은 증가할 수 있다.

또한 검색자가 반드시 투자자로 정의할 수 없다. 한국의 경우 그 비율에 대한 조사가 없으니 미국의 경우 상단수가 투자자로 간주됨으로 일정 수준 반영된다고 하지만,,,즉 상당수의 검색이 바이어스일 가능성도 있다.

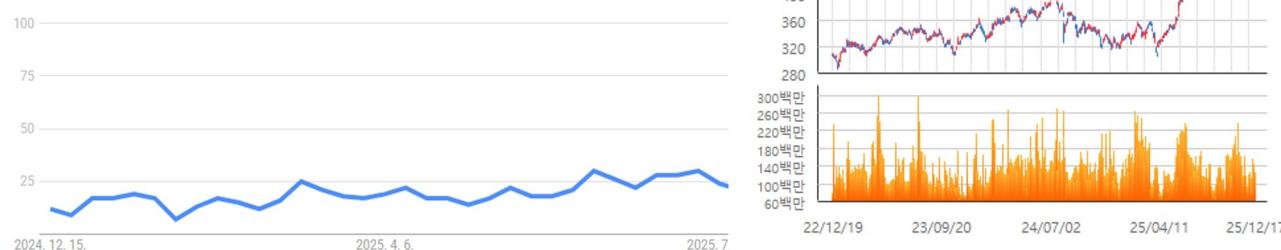
검색 수의 증가와 감소는 주가의 추세에 영향을 미친다.

A recent investigation has shown that the number of clicks on search results stemming from a given country correlates with the amount of investment in that country³¹. Further studies exploiting the temporal dimension of Google Trends data have demonstrated that changes in query volumes for selected search terms mirror changes in current numbers of influenza cases³² and current volumes of stock market transactions³³. This demonstration of a link between stock market transaction volume and search volume has also been replicated using *Yahoo!* data³⁴. Choi and Varian³⁵ have shown that data from *Google Trends* can be linked to current values of various economic indicators, including automobile sales, unemployment claims, travel destination planning and consumer confidence. A very recent study has shown that Internet users from countries with a higher per capita GDP are more likely to search for information about years in the future than years in the past³⁶.

Google Trends

1. 주제 선택
2. 시간 경과에 따른 관심도 그래프 읽기
3. 숫자 이해하기
4. 위치로 검색
5. 인기 검색어 및 급상승 검색어
- A. 거래에서 제이된 데이터

시간 흐름에 따른 관심도 변화 ⓘ



<https://trends.google.co.kr/trends/>

<https://data.krx.co.kr/contents/MDC/MDI mdiLoader/index.cmd?menuId=MDC0201010105>

Sentiment Feature

정의

- 시장 참여자의 심리와 태도를 수치화한 변수.
- 뉴스 기사, 기업 공시, 트위터·레딧 같은 소셜미디어 글에서 긍정·부정 감정 점수를 추출해 사용.

기능

- 투자자 심리 반영: 군중심리(bullish/bearish)를 모델에 포함해 가격 변동성을 설명.
- 비정형 데이터 활용: 텍스트·감정 데이터를 정량화하여 예측 모델의 입력 변수로 사용.
- 시장 방향성 예측: 긍정적 뉴스 → 매수세 강화, 부정적 뉴스 → 매도세 강화.

특징

- 비정형성: 자연어 처리(NLP), 감정 분석 기법 필요.
- 실시간성: 뉴스·SNS 데이터는 빠르게 변동, 단기 예측에 강점.
- 주관성: 데이터 출처와 분석 방법에 따라 결과가 달라질 수 있음.

Technical Feature

정의

- **과거 시장 데이터(가격·거래량)**에서 계산된 통계적·수학적 지표.
- 예: 이동평균(MA), 상대강도지수(RSI), 볼린저 밴드, MACD 등.

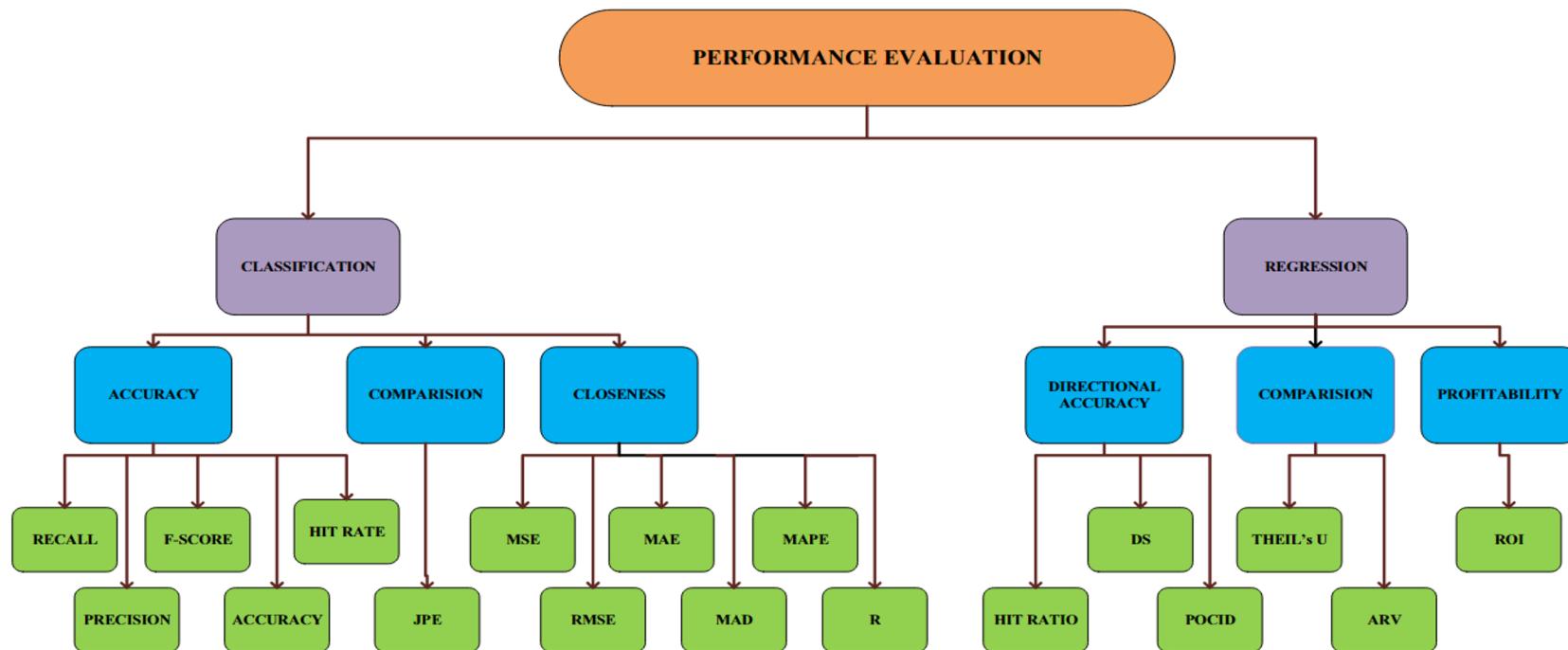
기능

- 추세 파악: 상승·하락 추세를 식별.
- 매매 신호 제공: 진입·청산 시점을 알려주는 지표.
- 패턴 기반 예측: 과거 가격 패턴이 미래에도 반복된다는 가정에 기반.

특징

- 정량적·객관적: 가격·거래량 같은 수치 데이터에 기반.
- 단기 예측에 강점: 기술적 지표는 주로 단기 매매 전략에 활용.
- 보편적 활용: 차트 분석가, 알고리즘 트레이딩 모델에서 기본적으로 사용.

Kumar, Gourav, Sanjeev Jain, and Uday Pratap Singh. 2021. "Stock Market Forecasting Using Computational Intelligence: A Survey." *Archives of Computational Methods in Engineering* 28 (3): 1069–101. <https://doi.org/10.1007/s11831-020-09413-5>.



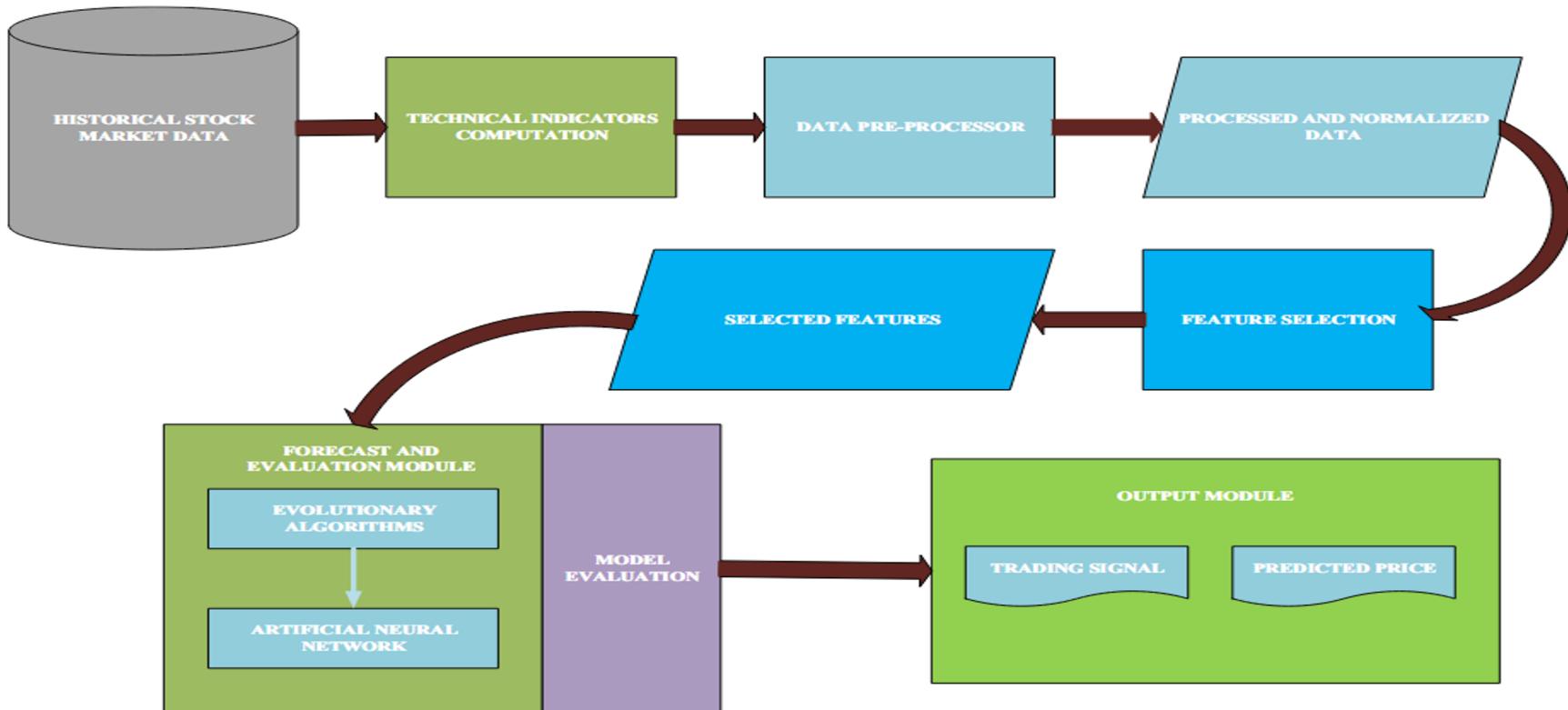


Fig. 12 Working of proposed system

감성발생 후 시간 반영 기간

The highest accuracy is accomplished when technical feature applied on time horizon (day 1). Meanwhile, the sentiment feature performs the best when implemented on 30 and 45 trading windows. 126p

=> 기술적분석은 1일 후의 변화를 예측하는 것에, 감성분석은 30-45일 이후에 적용되는 경우, 즉 단기적으로 기술적 분석이 장기적 분석은 감성분석이 유의적일 수 있다

the combination of technical analysis feature and sentiment label feature has steady performance, especially on time horizon (day 1) and trading windows of 3, 5, 10 and 15. However, the accuracy of combined feature falls on the 30 trading window.

감성분석을 위한 데이터의 종류

주식 데이터를 분석을 위한 감성 데이터를 수집하는 대상

- 전체 시장에 미치는 데이터는 거시경제, 정치, 외교(미국 등)에 관한 정보
- 개별 산업에 대한 데이터는 미,거시경제, 산업동향
- 개별 기업에 대한 데이터는 기업의 재무보고, 영업현황

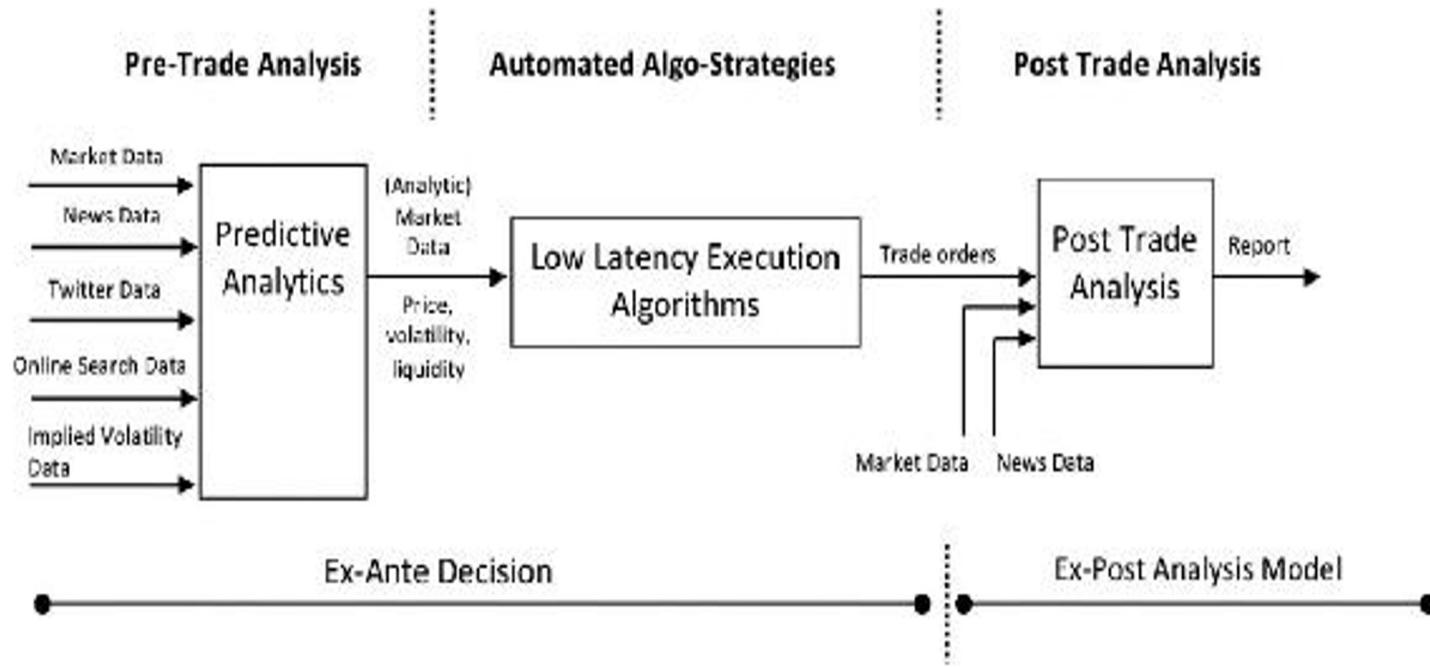


Figure 1.5.2: Information flow and computational architecture for automated trading

감성분석 기술과 모델

Techniques	Models	References
Natural Language Processing (NLP) and Sentiment Analysis	TF-IDF, BERT, LSTM (Long Short-Term Memory)	[13, 14, 18]
Deep Learning Models	LSTM, CNN, CNN-LSTM Hybrid	[64, 69, 96, 98]
Generative Adversarial Networks (GANs) and Attention Mechanisms	GANs, Attention Mechanisms	[80, 86, 98]
Investor Sentiment Indices and Econometric Models	VAR(Vector Auto-Regressive), GARCH, EGARCH	[14, 23, 32, 42, 91]
Hybrid and Real-Time Sentiment Analysis Models	SVM, ANN, Random Forests, High-Frequency Sentiment Analysis	[65, 74, 80, 90, 107]
Graph Neural Networks (GNNs) and Multi-Source Aggregated Classification (MAC)	GNNs, MAC	[86, 101, 104]

Table 10. Overview of techniques and models used for financial news sentiment analysis and market prediction

Model	References	Strengths	Weaknesses
TF-IDF	[18]	Simple, fast, good for keyword relevance	Ignores context, shallow analysis
BERT	[5]	Captures contextual and semantic meaning	Requires large annotated data, high compute
LSTM (Long Short-Term Memory)	[14, 64]	Good for sequential dependencies, time-series	Needs large datasets, prone to overfitting
CNN	[96]	Extracts local features well, efficient	Limited temporal understanding alone
Model	References	Strengths	Weaknesses
GANs	[80, 98]	Generates synthetic data, useful for sparse datasets	Training unstable, resource-intensive
VAR (Vector Auto-Regressive)	[14]	Interpretable, links sentiment with macro factors	Assumes linearity, poor for non-linear dynamics
GARCH	[91]	Models volatility effectively	Limited for structural breaks
EGARCH	[32]	Captures asymmetric volatility	Complex, parameter sensitive
SVM	[99]	Strong classifier for small datasets	Not scalable for large data
ANN	[105]	Flexible, learns complex patterns	Black-box, risk of overfitting
Random Forests	[107]	Handles non-linearities, robust	Less effective with high-dimensional text
GNNs	[104]	Models inter-stock/industry relations	Needs large relational data, heavy compute
MAC	[86]	Aggregates multiple signals, high accuracy	Complex integration, less interpretable

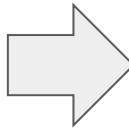
Ehsan, Aqsa, Shaista Habib, and Ann

News Sentiment Analysis Using NLP and Machine Learning for Asset Price Prediction: A Systematic Review." *VFAST Transactions on Software Engineering* 13 (3): 279–308.

<https://doi.org/10.21015/vtse.v13i3.2165>.

감성분석의 한계

Ref	Paper Title	Limitations	Future Directions
[11]	A novel approach to Predict WTI crude spot oil price: LSTM-based feature extraction with Xgboost Regressor	The model has an intricate architecture and requires significant processing time, which can be a challenge for real-time applications.	Future research could focus on developing models that use fewer processing resources while maintaining comparable levels of precision. It is also suggested to incorporate more extensive datasets and additional geopolitical/macroeconomic factors, like OPEC decisions, to improve model accuracy.
[108]	Development of a CNN-LSTM Approach with Images as Time-Series Data Representation for Predicting Gold Prices	The prediction model may not be generalizable across different time periods or markets.	Future research should incorporate a wider range of datasets and timeframes, and integrate more external economic indicators for improved performance?
[18]	Oil price volatility and new evidence from news and Twitter	Media sentiment from Twitter and news may not always reflect true sentiment, and results may vary depending on the platform used.	Future research should examine the impact of news across various platforms and focus on real-time data for volatility forecasting?
[106]	Sentiment Analysis of Financial News and its Impact on the Stock Market	The ability to predict stock movements is limited by the accuracy of the sentiment analysis and the quality of the news data.	Future directions include improving sentiment analysis techniques and incorporating additional data from diverse sources, including social media?
[85]	Predicting the Effects of News Sentiments on the Stock Market	The model is dependent on the quality and reliability of the news sentiment data, which may not accurately reflect market sentiment.	Future work should focus on developing better sentiment analysis tools and integrating more varied data sources to improve prediction accuracy?
[97]	Stock Price Prediction Using News Sentiment Analysis	The model's performance can be influenced by the quality and availability of news data.	Future research could enhance the model by integrating machine learning algorithms with real-time data and improving its generalization ability?
[105]	Efficacy of News Sentiment for Stock Market Prediction	The predictive power of news sentiment can be impacted by noisy data and biases in sentiment interpretation.	Future work could explore advanced machine learning techniques and focus on real-time prediction systems for improving accuracy?
[50]	News Headlines Categorization Scheme for Unlabeled Data	The approach requires manual selection of seed keywords, which limits its scalability and automation.	Future work could focus on automating the process of keyword selection and refining the categorization technique to handle larger datasets?
[48]	News-driven stock market index prediction based on trellis network and sentiment attention mechanism	The model requires a considerable amount of data and computation, making it difficult for real-time application.	Future research could explore reducing computational complexity, enhancing the model's real-time applicability, and investigating the impact of news sentiment in different market conditions?
[70]	News-based intelligent prediction of financial markets using text mining and machine learning	Existing models are not fully leveraging advanced language models like transformers, limiting their ability to capture deeper contextual meanings from news data.	Future directions include integrating more sophisticated NLP techniques like BERT and focusing on enhancing model adaptability across different markets?
[58]	Stock Price Prediction using Zero-Shot Sentiment Classification	Reliance on social media data introduces potential biases, and its effectiveness may vary across different industries.	Future work should aim to incorporate multiple data sources, including diverse financial news and real-time data, and further refine sentiment classification techniques?
[69]	A sentiment analysis-based machine learning approach for financial market prediction via news disclosures	The model is limited by the complexity of natural language and may not perform well with highly ambiguous text.	Future research could work on improving the model's robustness to ambiguous data and exploring the use of larger, more diverse datasets to increase accuracy?
[80]	Enhancing stock price prediction using GANs and transformer-based attention mechanisms	The models' reliance on GANs and attention mechanisms may lead to overfitting and unrealistic data generation.	Future work should explore regularization techniques and the integration of more external data sources to improve prediction accuracy and generalization?
[60]	Predictive Precision: LSTM-Based Analytics for Real-time Stock Market Visualization	The high volatility of stock data makes it challenging to achieve high prediction accuracy.	Future directions could involve improving data preprocessing methods and enhancing model adaptability to different market conditions?
[85]	Predicting the Effects of News Sentiments on the Stock Market	The sentiment analysis model is limited by the quality and granularity of the news data, which may not fully capture market sentiment.	Future research could focus on improving sentiment classification techniques and integrating social media data for more dynamic and real-time predictions?



Sentiment analysis techniques, however, have yet to demonstrate their full potential, as they are frequently hindered by problems of noisy data, subjective sentiment interpretation, and nonstationarity of financial time series

다중 감성 분석: 텍스트, 소리, 영상을 넘어 인간의 감정을 읽다

텍스트 분석 (Textual Analysis)



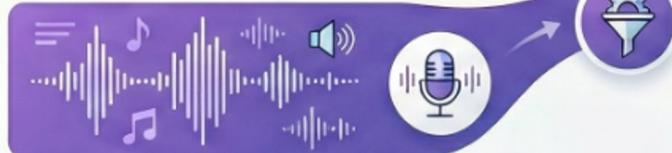
단어와 문장의 의미를 파악하여
긍정 또는 부정 극성을 식별합니다.

시각 분석 (Visual Analysis)



얼굴 표정, 몸짓 등 비언어적 신호를 통해
감정 상태를 분석합니다.

오디오 분석 (Audio Analysis)



목소리의 둔, 늦낮이, 말하는 속도 등
음성 특징으로 감점을 파악합니다.

2. 데이터 융합 (Data Fusion)

추출된 다양한 특징들을
결합하여 통합적인 감정
정보를 생성합니다.

3. 최종 감성 판단 (Final Sentiment Classification)



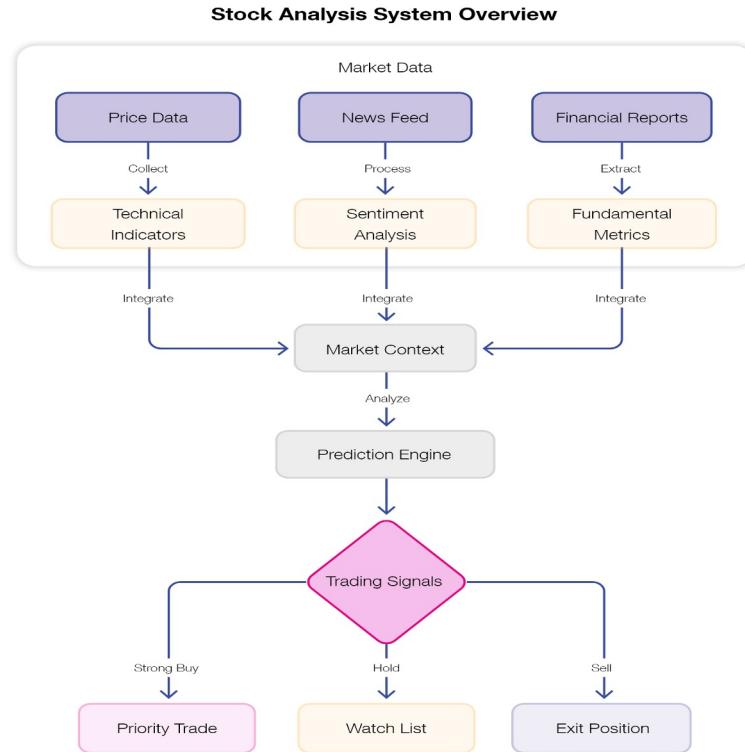
AI Agent와 통합 기대효과

When integrated into agentic AI agents, these models can autonomously analyze incoming news streams, cross-check sentiment with live market data, and adapt predictions in real time, ensuring that forecasts remain accurate and contextually relevant as market sentiment drivers evolve.

AI agent의 향후 기대효과

=> 유입된 데이터를 자동 분석하고, 시장의 실시간 데이터와 교차 확인이 가능하여 실시간 예측이 가능해짐

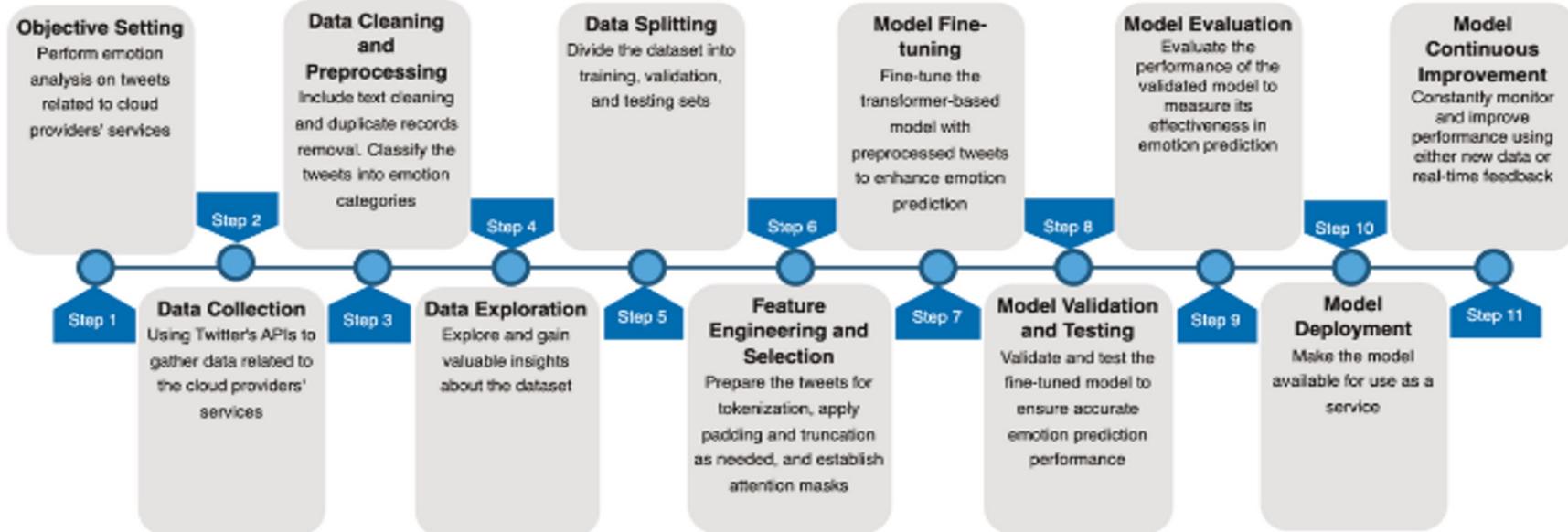
주가 분석의 흐름



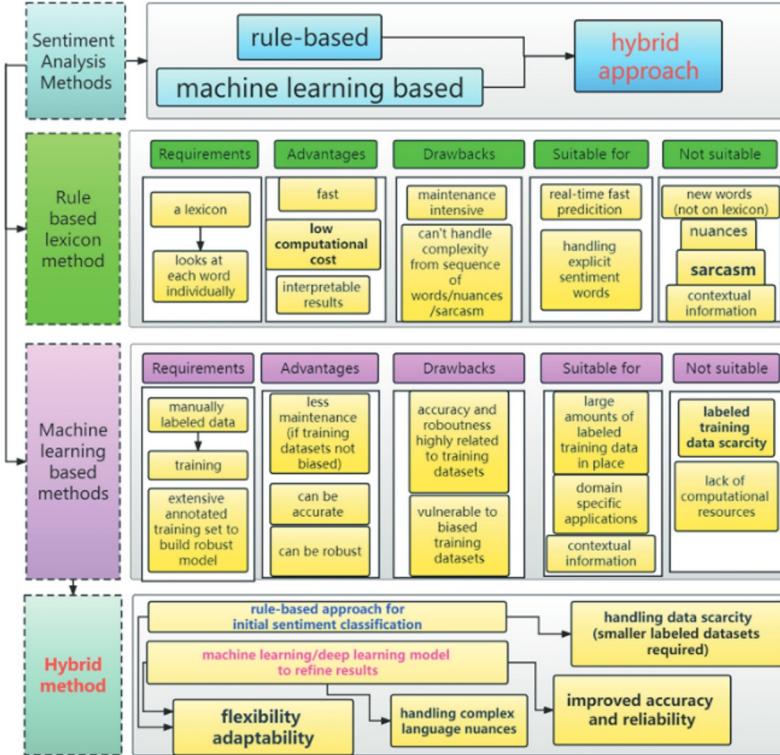
- 주식시장의 기초분석
- 주가변동 데이터
- 소비자의 태도, 의견 등 감성데이터

=> 이를 통합하여 시장에서의 맥락
(Market Context)와 통합 후 변동 예측

리서치 프로세스



감성분석의 진행방향



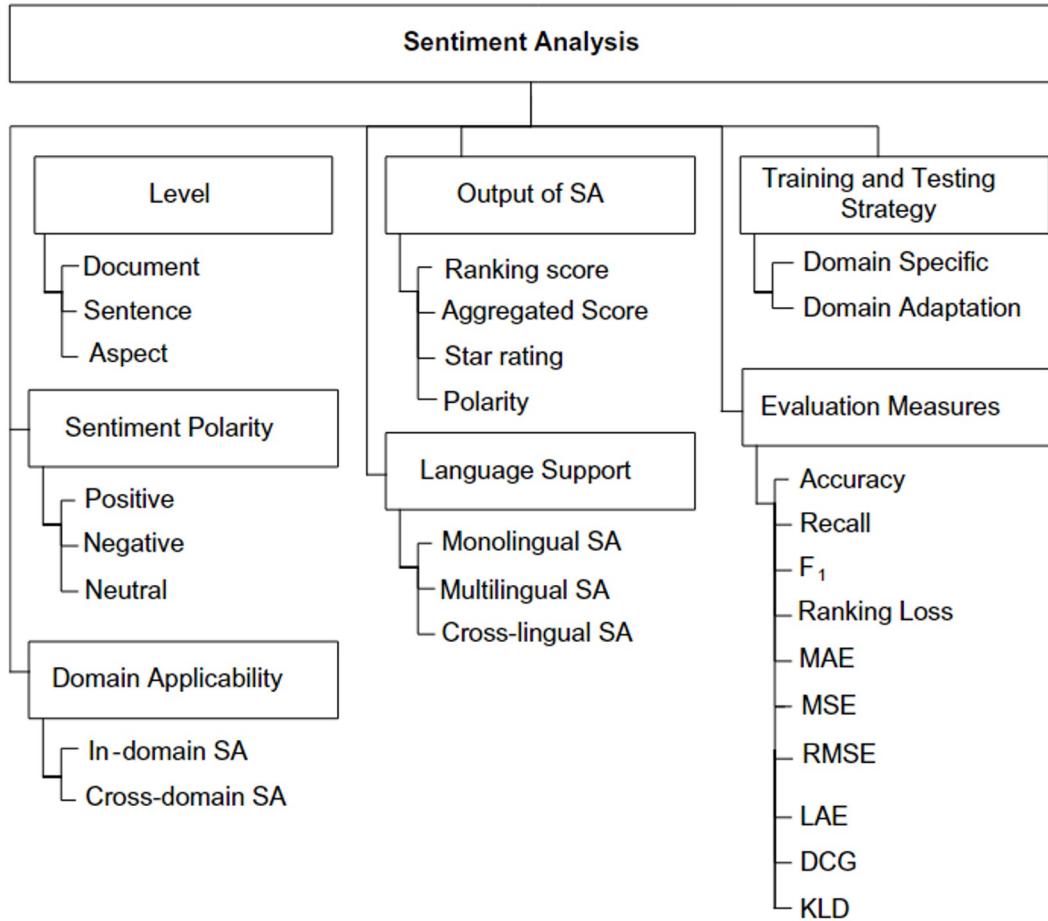
감성분석을 위한 규칙

-Rule-Based

-ML-Based

=> Hybrid Approach

- Data Fusion



- 다양한 데이터를 분석
- 감성의 분류(3단, 5단 etc.)
- 주식의 경우 다양한 거시 경제, SNS, 주식시장정보 등 다양함
- 데이터 분석 수준의 정확성과 같은 지표가 아닌 실질적인 데이터의 영향에 의한 주가변화를 측정

ML for SA

Sentiment Analysis and Graphical Taxonomy

Graphical taxonomy of sentiment analysis based on traits like levels, polarity, output, language support, applicability to domain

Text Representation

Detailed discussion on embedding models, initializing and enhancing the embedded vectors, and approximation algorithms

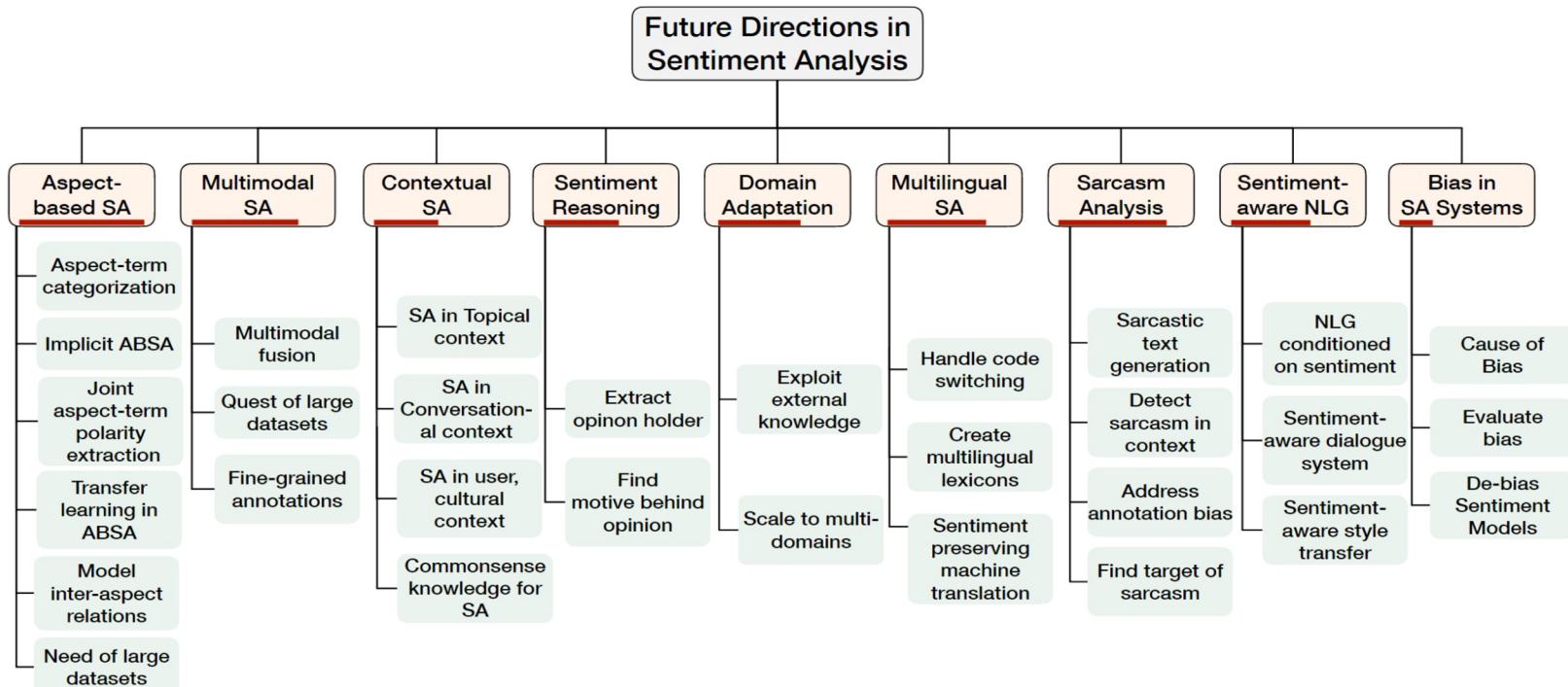
Deep Learning for Sentiment Analysis

Evaluation Metrics, Benchmarked Datasets and Tools

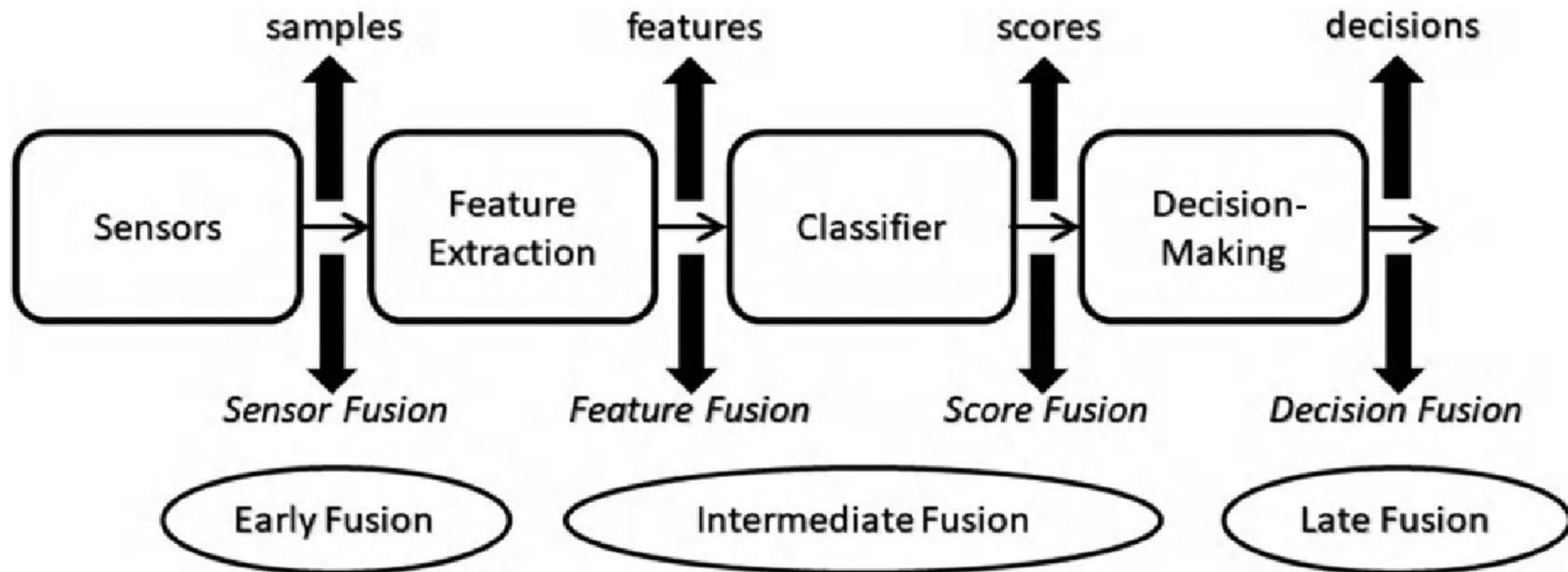
Comparative study of evaluation metrics, benchmarked datasets and online tools for sentiment analysis

State-of-the-art Deep Learning approaches for Sentiment Analysis
Comparative study of deep learning approaches being applied at various levels (granularities) for sentiment analysis

SA의 향후 방향



How to Multi-modal Data Fusion



데이터 융합 방식

Feature-level Fusion (특징 융합)

시계열 특징과 텍스트 감성 점수를 하나의 벡터로 결합하여 모델 입력으로 사용합니다.

```
import numpy as np
from sklearn.linear_model import LogisticRegression

# 예시 데이터
# 시계열 특징: [주가 변동률, 거래량]
time_series_features = np.array([[0.02, 1000], [0.05, 1500], [-0.01, 800]])
# 텍스트 감성 점수: [-1 ~ +1]
sentiment_scores = np.array([[0.8], [-0.3], [0.1]])

# Feature-level 융합
X = np.hstack((time_series_features, sentiment_scores))
y = np.array([1, 0, 1]) # 예: 주가 상승(1)/하락(0)

# 간단한 모델 학습
model = LogisticRegression()
model.fit(X, y)

print("예측:", model.predict(X))
```

Model-level Fusion (모델 융합)

시계열 모델과 텍스트 모델을 각각 학습 후 결과를 결합합니다.

```
import numpy as np
from sklearn.linear_model import LinearRegression

# 시계열 모델 (예: LSTM 대신 간단한 회귀)
time_series_X = np.array([[0.02], [0.05], [-0.01]])
time_series_y = np.array([1.1, 1.5, 0.9])
time_model = LinearRegression().fit(time_series_X, time_series_y)

# 텍스트 모델 (예: 감성 점수 기반 회귀)
sentiment_X = np.array([[0.8], [-0.3], [0.1]])
sentiment_y = np.array([1.2, 0.7, 1.0])
text_model = LinearRegression().fit(sentiment_X, sentiment_y)

# 두 모델의 결과 결합
combined_predictions = (time_model.predict(time_series_X) +
    text_model.predict(sentiment_X)) / 2

print("모델 융합 결과:", combined_predictions)
```

Decision-level Fusion (결정 융합)

각 모델이 독립적으로 예측한 결과를 **양상률(ensemble)**하여 최종 판단합니다.

```
import numpy as np
from sklearn.linear_model import LogisticRegression

# 예시 데이터
X_time = np.array([[0.02], [0.05], [-0.01]])
y = np.array([1, 0, 1])

X_text = np.array([[0.8], [-0.3], [0.1]])

# 두 개의 독립 모델
time_model = LogisticRegression().fit(X_time, y)
text_model = LogisticRegression().fit(X_text, y)

# 각 모델의 예측
pred_time = time_model.predict(X_time)
pred_text = text_model.predict(X_text)

# Decision-level Fusion (다수결 양상률)
final_pred = []
for t, s in zip(pred_time, pred_text):
    final_pred.append(1 if (t + s) >= 1 else 0)

print("결정 융합 결과:", final_pred)
```

- **Feature-level Fusion** → 입력 단계에서 특징을 합침
- **Model-level Fusion** → 모델 별로 학습 후 결과를 결합
- **Decision-level Fusion** → 독립적 예측을 양상률

감성분석을 위한 실시간 데이터의 수집 (Data Sources and Access Methods)

- **Yahoo Finance (prices, news)**
 - **Prices:** Use `yfinance` for near real-time quotes with 1-minute polling.
 - **News:** `yfinance.Ticker.news` gives recent headlines. Poll for updates.
 - Note: Yahoo's realtime WebSocket isn't public; polling is the safe route.
- **X/Twitter (live stream)**
 - **Filtered Stream API v2** provides true real-time tweets for your rules (e.g., stock cashtags).
 - Requires a developer account and Bearer Token.
- **Reddit (new posts/comments)**
 - Use `PRAW` to poll subreddits (e.g., r/stocks, r/wallstreetbets) for new submissions/comments.
 - Pushshift is not real-time anymore; stick with Reddit's API.
- **Stocktwits (sentiment-rich messages)**
 - Public endpoints for symbol streams (e.g., AAPL). Poll for new messages.
- Optional extras:
 - **YouTube comments** (poll via Data API), **News APIs** (e.g., NewsAPI), **Discord** (bot with event handlers) — include if relevant.

대표적인 트랜스포머 기반 융합 모델

① MuLT (Multimodal Transformer)

- 핵심 아이디어: 서로 다른 모달리티(예: 텍스트와 오디오)를 직접적으로 정렬(Alignment)하지 않고, **Cross-modal Attention**을 통해 한 모달리티의 정보를 다른 모달리티로 투영하여 잠재적인 상호작용을 학습합니다.
- GitHub: [yaohungt/Multimodal-Transformer](https://github.com/yaohungt/Multimodal-Transformer)
- 특징: CMU-MOSI, CMU-MOSEI 데이터셋에서 성능이 매우 뛰어나며, "비정렬(Unaligned)" 데이터 처리에 강점이 있습니다.

② MAG-BERT (Multimodal Adaptation Gate for BERT)

- 핵심 아이디어: 텍스트를 처리하는 BERT 모델 내부에 **MAG(Multimodal Adaptation Gate)**라는 장치를 삽입하여, 텍스트 정보가 흐르는 중간에 시각 및 청각 정보를 주입(Injection)하는 방식입니다.
- GitHub: [eshansurendra/MAG-BERT](https://github.com/eshansurendra/MAG-BERT)
- 특징: 언어 모델(BERT)의 강력한 성능을 유지하면서 멀티모달 정보를 효과적으로 결합합니다.

③ MISA (Modality-Invariant and Specific-Analysis)

- 핵심 아이디어: 각 모달리티를 '공통적인 특징(Invariant)'과 '고유한 특징(Specific)'으로 분리하여 학습한 뒤, 이를 트랜스포머 레이어 등으로 결합합니다.
- GitHub: [declare-lab/MISA](https://github.com/declare-lab/MISA)

통합 프레임워크

MMSA (Multimodal Sentiment Analysis Toolkit)

- GitHub: [thuiai/MMSA](https://github.com/thuiai/MMSA)
- 내용: 이 리포지토리는 **MuLT**, **MISA**, **TFN**, **LMF**, **Self-MM** 등 15개 이상의 최신 멀티모달 감성 분석 모델을 통합하여 제공합니다.
- 장점:
 - CMU-MOSI, MOSEI 데이터를 자동으로 처리해 줍니다.
 - 간단한 파이썬 코드로 여러 트랜스포머 기반 모델을 실험해 볼 수 있습니다.
 - pip install MMSA로 쉽게 설치가 가능합니다

멀티모달 데이터 통합 주요 방법론

1. Early Fusion (입력단 융합) 서로 다른 데이터를 신경망에 넣기 전에 하나로 합치는 방식입니다.

- **방법:** 텍스트에서 감성 점수(Sentiment Score)나 임베딩(Vector)을 추출한 뒤, 이를 시계열 데이터의 열(Column)로 추가합니다.
- **구현:** 주가 데이터가 [시가, 고가, 저가]라면 여기에 [감성점수]를 더해 [시가, 고가, 저가, 감성점수]라는 4차원 입력을 만듭니다.
- **장점:** 구현이 매우 간단하며 데이터 간의 초기 상관관계를 학습하기 좋습니다.

2. Intermediate Fusion (모델 내부 융합) 각 모달리티(Modality)를 전담하는 신경망을 따로 두고, 중간의 **특징 추출 단계**에서 결합하는 방식

- **구조:**
 1. **텍스트 경로:** BERT나 RoBERTa 같은 언어 모델을 통해 문맥 벡터를 추출합니다.
 2. **시계열 경로:** LSTM, GRU 또는 1D-CNN을 통해 시간적 패턴 벡터를 추출합니다.
 3. **병합(Concat):** 추출된 두 벡터를 하나로 이어 붙여(Concatenation) 최종 Dense Layer(Fully Connected)에 전달합니다.
- **장점:** 각 데이터의 고유한 특성(언어적 의미 vs 시간적 흐름)을 보존하면서 결합할 수 있습니다.

3. Cross-modal Attention (어텐션 기반 융합) 최근 트랜스포머(Transformer) 모델에서 주로 사용하는 최첨단 방식입니다.

- **원리:** 텍스트 정보가 시계열 데이터의 어느 시점에 중요한 영향을 주었는지 모델이 스스로 찾게 합니다.
- **예시:** "금리 인상"이라는 텍스트가 입력되었을 때, 모델은 시계열 데이터 중 금리 발표일 근처의 데이터에 더 높은 가중치(Attention)를 부여합니다.
- **핵심 모델:** **MuLT (Multimodal Transformer)**가 이 방식을 대표합니다.

Python Code Ex.

```
import torch
import torch.nn as nn

class FusionModel(nn.Module):
    def __init__(self):
        super(FusionModel, self).__init__()
        # 1. 시계열 처리 (LSTM)
        self.ts_model = nn.LSTM(input_size=5, hidden_size=64, batch_first=True)

        # 2. 텍스트 처리 (BERT 결과를 768차원을 가짐)
        self.text_dense = nn.Linear(768, 64)

        # 3. 결합 후 최종 예측
        self.final_layer = nn.Linear(64 + 64, 1) # 64(TS) + 64(Text)

    def forward(self, ts_data, text_embeddings):
        # 시계열 빅터 추출
        _, (ts_hidden, _) = self.ts_model(ts_data)
        ts_feature = ts_hidden[-1] # 마지막 시점의 출력

        # 텍스트 빅터 추출
        text_feature = torch.relu(self.text_dense(text_embeddings))

        # 결합 (Fusion)
        combined = torch.cat((ts_feature, text_feature), dim=1)

        return self.final_layer(combined)
```

주식시장의 감성데이터와 시계열 데이터의 융합

- **융합 대상 데이터:** 수치 데이터(정형 시계열 자료), 뉴스 기사 및 소셜 미디어 데이터(비정형 감성 자료).
- **적용 기법:**
 - **순환 신경망 (RNN/LSTM):** 주가 변화와 같은 시계열 데이터의 시간적 의존성을 포착하는 데 탁월하며, 뉴스 데이터의 시퀀스 정보와 결합하여 학습할 수 있습니다.
 - **감성 분석 (Sentiment Analysis) 및 NLP:** 뉴스나 게시글에서 긍정/부정적 감성을 추출하여 수치화한 뒤, 이를 시계열 모델의 추가 변수로 입력합니다.
 - **어텐션 메커니즘 (Attention Mechanisms):** 수많은 뉴스 정보 중 주가 변동에 가장 큰 영향을 미치는 특정 시점이나 텍스트 정보를 동적으로 가중치 있게 반영하여 예측 정확도를 높입니다.
 - **동적 베이즈 네트워크 (DBN):** 시간에 따른 변수 간의 확률적 관계를 모델링하여, 결측치가 있는 복잡한 금융 데이터 내의 인과 관계를 추론하는 데 사용됩니다.
- 시계열 데이터는 지속적인 데이터의 변화를 감성데이터는 시장의 급속한 또는 갑작스러운 변화와 소비자들의 태도를 반영

CRMKL Code_part

```
# 3. CRMKL 스탠다드 모델 정의
class CRMKL_Simplified(nn.Module):
    def __init__(self):
        super(CRMKL_Simplified, self).__init__()
        # 주가 처리를 위한 LSTM (Temporal 특징)
        self.lstm = nn.LSTM(input_size=5, hidden_size=64, batch_first=True)

        # 감성 점수 처리를 위한 MLP (Auxiliary 특징)
        self.sentiment_mlp = nn.Sequential(
            nn.Linear(1, 16),
            nn.ReLU()
        )

        # 두 정보를 합친 최종 분류기 (Fusion Layer)
        self.classifier = nn.Sequential(
            nn.Linear(64 + 16, 32),
            nn.ReLU(),
            nn.Linear(32, 2) # [하락 확률, 상승 확률]
        )

    def forward(self, stock_seq, sentiment):
        # 주가 시퀀스 처리
        _, (hn, _) = self.lstm(stock_seq)
        stock_feat = hn[-1] # 마지막 hidden state

        # 감성 데이터 처리
        sent_feat = self.sentiment_mlp(sentiment)

        # 두 특징 결합 (Fusion)
        combined = torch.cat((stock_feat, sent_feat), dim=1)
        return self.classifier(combined)
```

융합 모델 구조(Multimodal Fusion Model)

모델의 강점

- 다양한 정보 수용:** 주가의 단순한 수치 패턴뿐만 아니라, 시장의 심리(감성 데이터)와 거시적인 지표를 동시에 고려합니다.
- 비동기 데이터 처리:** 뉴스나 블로그는 매일 발생하지 않을 수 있습니다. 이를 수치화하여 MLP 입력으로 넣음으로써 시계열의 연속성을 해치지 않고 정보를 주입할 수 있습니다.
- 상승/하락 예측 최적화:** 단순 수치 예측보다 '방향성 (Classification)'을 타겟으로 잡음으로써, 실제 투자 전략 (Long/Short)에 더 직관적으로 활용 가능합니다.

삼성전자 사례

```
C:\Users\dongu\AppData\Local\Temp\ipykernel_4948\1965019332.py:12: FutureWarning: YF.download() has changed argument auto_adjust default to True
df = yf.download("005930.KS", start="2023-01-01", end="2024-12-31")
[*****100%*****] 1 of 1 completed
삼성전자 융합 모델 학습 샘플 출력: [ 0.05578613 -0.09016061]
```

05. 자체 평가 의견

- 데이터의 융합 방법은 다양함
- 파이썬을 통한 데이터의 융합은 실제적으로 가능함

향후 방향

- 데이터 수집에 제한적(각 데이터의 유료 또는 유료화)
- 수집할 수 있는 웹크롤링의 제한
- 전반적인 시장 정보 수집 등이 제한 됨
- 개별적인 기업 대상의 주가 분석은 가능한 것으로 보임
- KOSPI 등의 전반적 시장 데이터를 분석하는 방법론이 필요

참고문헌

- Albrecht, J., Ramachandran, S., & Winkler, C. 2020년. *Blueprints for Text Analytics Using Python*. O'Reilly Media, Inc.
- Chen, Chung-Chi, and Hiroya Takamura. 2025. *Agent AI for Finance: From Financial Argument Mining to Agent-Based Modeling*. SpringerBriefs in Intelligent Systems. Springer Nature Switzerland. <https://doi.org/10.1007/978-3-031-94687-5>.
- Chen, Chung-Chi, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. *From Opinion Mining to Financial Argument Mining*. SpringerBriefs in Computer Science. Springer Singapore. <https://doi.org/10.1007/978-981-16-2881-8>.
- Corazza, Marco, René Garcia, Faisal Shah Khan, Davide La Torre, and Hatem Masri. 2024. *Artificial Intelligence and Beyond for Finance*. Vol. 15. Transformations in Banking, Finance and Regulation. WORLD SCIENTIFIC (EUROPE). <https://doi.org/10.1142/q0449>.
- DR JIGNASHA SHAH DALAL, DR SANTHILATA K. V. 2025. *ULTIMATE FINGPT FOR FINANCIAL ANALYSIS*. ORANGE EDUCATION PVT LTD.
- Ehsan, Aqsa, Shaista Habib, and Arman Sohail. 2025. "Financial News Sentiment Analysis Using NLP and Machine Learning for Asset Price Prediction: A Systematic Review." *VFAST Transactions on Software Engineering* 13 (3): 279–308. <https://doi.org/10.21015/vtse.v13i3.2165>.
- Li, Jinfeng. n.d. *Sentiment Analysis Advances*.
- Liu, Bing. 2007. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Data-Centric Systems and Applications. Springer.
- Liu, Bing. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies Ser. Morgan & Claypool Publishers.
- Liu, Bing. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. First published. Cambridge University Press. <https://doi.org/10.1017/CBO9781139084789>.
- Liu, Bing. 2020. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions 2nd*. 2nd ed. Cambridge University Press. <https://doi.org/10.1017/9781108639286>.
- Mitra, Gautam. 2016년. *Handbook of Sentiment Analysis in Finance*.
- Kumar, Gourav, Sanjeev Jain, and Uday Pratap Singh. 2021. "Stock Market Forecasting Using Computational Intelligence: A Survey." *Archives of Computational Methods in Engineering* 28 (3): 1069–101. <https://doi.org/10.1007/s11831-020-09413-5>.

- Meredith, Dale. 2024. *The OSINT Handbook: A Practical Guide to Gathering and Analyzing Online Information*. 1st ed. Packt Publishing Limited.
- Mitra, Gautam. 2016. *Handbook of Sentiment Analysis in Finance*. 2nd ed. John Wiley & Sons.
- Poria, Soujanya, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2020. *Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research*. November 16. <https://doi.org/10.48550/arXiv.2005.00357>.
- Preis, Tobias, Helen Susannah Moat, and H. Eugene Stanley. 2013. "Quantifying Trading Behavior in Financial Markets Using Google Trends." *Scientific Reports* 3 (1): 1684. <https://doi.org/10.1038/srep01684>.
- Rao, Jitendra R., S. Sethu Selvi, Sudesh K. Kashyap, and Ailneni Sanketh. 2025. *Data Fusion Mathematics: Theory and Practice*. 2nd ed. CRC Press. <https://doi.org/10.1201/9781003560265>.
- Ramsay, Allan. 2023. *Machine Learning for Emotion Analysis in Python: Build AI-Powered Tools for Analyzing Emotion Using Natural Language Processing and Machine Learning*. 1st ed. With Tariq Ahmad. Packt Publishing Limited.
- Sagala, Tommy Wijaya, Mei Silviana Saputri, Rahmad Mahendra, and Indra Budi. 2020. "Stock Price Movement Prediction Using Technical Analysis and Sentiment Analysis." *Proceedings of the 2020 2nd Asia Pacific Information Technology Conference*, January 17, 123–27. <https://doi.org/10.1145/3379310.3381045>.
- Stewart, Dominic. 2010. *Semantic Prosody: A Critical Evaluation*. Routledge Advances in Corpus Linguistics 9. Taylor & Francis Group.
- Suzuki, Joe. 2023. *WAIC and WBIC with Python Stan: 100 Exercises for Building Logic*. Springer Nature Singapore. <https://doi.org/10.1007/978-981-99-3841-4>.
- Van Der Post, Hayden. 2023. *Sentiment Analysis in Trading: Python's Perspective on Market Emotions: Sentiment Analysis in Quantitative Finance*. Reactive Publishing.

article

- Bagozzi, R. P., M. Gopinath, and P. U. Nyer. 1999. "The Role of Emotions in Marketing." *Journal of the Academy of Marketing Science* 27 (2): 184–206.
<https://doi.org/10.1177/0092070399272005>.
- Kaur, Chhinder, and Anand Sharma. 2020. "Social Issues Sentiment Analysis Using Python." *2020 5th International Conference on Computing, Communication and Security (ICCCS)*, October 14, 1–6. <https://doi.org/10.1109/ICCCS49678.2020.9277251>.
- Kirtac, Kemal, and Guido Germano. 2024. "Sentiment Trading with Large Language Models." *Finance Research Letters* 62 (April): 105227.
<https://doi.org/10.1016/j.frl.2024.105227>.
- Kumar, Gourav, Sanjeev Jain, and Uday Pratap Singh. 2021. "Stock Market Forecasting Using Computational Intelligence: A Survey." *Archives of Computational Methods in Engineering* 28 (3): 1069–101. <https://doi.org/10.1007/s11831-020-09413-5>.
- Mino, Domenica, and Cillian Williamson. 2025. *Sentiment and Volatility in Financial Markets: A Review of BERT and GARCH Applications during Geopolitical Crises*. October 18. <https://doi.org/10.48550/arXiv.2510.16503>.
- Sagala, Tommy Wijaya, Mei Silviana Saputri, Rahmad Mahendra, and Indra Budi. 2020. "Stock Price Movement Prediction Using Technical Analysis and Sentiment Analysis." *Proceedings of the 2020 2nd Asia Pacific Information Technology Conference*, January 17, 123–27. <https://doi.org/10.1145/3379310.3381045>.
- Zhang, Alex L., Tim Kraska, and Omar Khattab. 2025. "Recursive Language Models." arXiv:2512.24601. Preprint, arXiv, December 31.
<https://doi.org/10.48550/arXiv.2512.24601>.