JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 1, pp. 51-58, January 2026
https://doi.org/10.9708/jksci.2026.31.01.051

Big Data Text Mining-Based Stock Prediction System

Tai-Sung Hur*, Ariunjargal Amintsog**

*Professor, Dept. of Computer Science, Inha Technical College, Incheon, Korea
**Student, Dept. of Computer Science, Inha Technical College, Incheon, Korea

[Abstract]

This study presents an end-to-end system that transforms multilingual financial texts (Korean news and English forums) into quantifiable trading signals using domain-specific language models FinBERT and KR-FinBERT. The key innovation is an adaptive volatility threshold mechanism where trading signals are triggered when the 5-day EWMA of sentiment indices crosses dynamic bands defined by ±1 standard deviation of a 60-day rolling window. This approach automatically adjusts sensitivity to market volatility, enhancing strategy robustness across diverse conditions. Through rigorous backtesting on KOSPI stocks and Apple (January 2020 - June 2025) with realistic transaction costs (0.1% one-way) and execution delays, the active strategy consistently outperforms passive buy-and-hold benchmarks. Notably, Apple and Kia achieved Sharpe ratios of 1.42 and 1.13, respectively, with significantly lower maximum drawdowns. Event studies and parameter sensitivity analysis confirm the statistical and economic significance of this NLP-based algorithmic trading approach.

▸ **Key words**: Financial NLP, Sentiment Analysis, Algorithmic Trading, Backtesting, KR-FinBERT

[요 약]

본 연구는 다국어 금융 텍스트(한국어 뉴스 및 영어 포럼)를 정량화된 거래 신호로 변환하는 엔드-투-엔드 시스템을 제안한다. 금융 도메인 특화 언어 모델인 FinBERT와 KR-FinBERT를 활용하여 문장 단위 감성 분석을 수행하고, 이를 일간 감성 지수로 집계한다. 핵심 혁신은 감성 지수의 5일 EWMA가 60일 롤링 표준편차(±1σ)의 동적 밴드를 교차할 때 거래 신호를 발생시키는 적응형 변동성 임계값 메커니즘이다. 이는 시장 변동성에 따라 민감도를 자동 조절하여 전략의 강건성을 확보한다. 2020년 1월부터 2025년 6월까지 KOSPI 주식과 Apple을 대상으로 현실적인 거래 비용(편도 0.1%)과 실행 지연을 반영한 엄격한 백테스트 결과, 액티브 전략은 수동적 벤치마크 대비 우수한 성과를 보였다. 특히 Apple과 기아는 각각 샤프 비율 1.42, 1.13을 기록했으며, 이벤트 연구 및 민감도 분석을 통해 본 전략의 통계적·경제적 유의성을 확인하였다.

▸ **주제어**: 금융 자연어처리, 감성 분석, 알고리즘 트레이딩, 적응형 임계값, 백테스트, KR-FinBERT

• First Author: Tai-Sung Hur, Corresponding Author: Ariunjargal Amintsog
*Tai-Sung Hur (tshur@inhatc.ac.kr), Dept. of Computer Science, Inha Technical College
**Ariunjargal Amintsog (202547007@itc.ac.kr), Dept. of Computer Science, Inha Technical College
• Received: 2025. 11. 17, Revised: 2025. 12. 17, Accepted: 2025. 12. 29.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

디지털 혁명 이후 금융 시장은 정보의 생성 및 전파 속도가 기하급수적으로 증가하는 패러다임 전환을 맞이하고 있다. 특히 뉴스, 기업 공시, 애널리스트 리포트, 소셜 미디어 등에서 매일같이 쏟아지는 방대한 비정형 텍스트 데이터는 시장 참여자들의 심리와 행동에 지대한 영향을 미치는 핵심 변수로 부상했다 [1][7]. 2024년 말 기준 국내 개인 주주 수가 8,430만 명으로 10년 연속 증가세를 보이는 등 주식 시장 참여가 대중화되었지만 [2], 대다수의 개인 투자자들은 전문 지식의 부재와 정보 처리 능력의 한계로 인해 정보의 홍수 속에서 유의미한 정보를 선별하고 노이즈를 필터링하는 데 극심한 어려움을 겪고 있다. 특히 초보 투자자들은 복잡한 투자 전략과 시장 트렌드를 스스로 이해하기 어렵고, 뉴스나 커뮤니티 등 다양한 정보 속에서 노이즈를 걸러내기 어려우며, 감정에 흔들려 충동적 매매로 자산 손실 위험이 커지는 현실에 직면해 있다.

이러한 배경 하에, 본 연구는 빅데이터와 최신 인공지능(AI) 기술을 접목하여 비정형 텍스트에 내재된 시장의 집단적 감성(sentiment)을 정량적으로 포착하고, 이를 기반으로 객관적이고 재현 가능한 주식 거래 신호를 생성하는 자동화 시스템을 개발 및 제안한다. 최근 연구들은 X(Twitter), Facebook 등 SNS 데이터 마이닝이 주가 변동 예측에 효과적임을 입증하였으며, Reddit, 네이버 등 다양한 소셜 플랫폼에서 투자 신호를 추출하는 연구가 증가하고 있다 [1][7]. 본 시스템은 전통적인 기본적·기술적 분석을 보완하는 강력한 대안 데이터(alternative data)로서 텍스트의 가치에 주목하며, 특히 금융이라는 특수 도메인의 언어적 뉘앙스를 깊이 있게 이해할 수 있는 금융 특화 자연어처리(NLP) 모델을 적용하여 분석의 정밀도를 극대화하고자 한다 [5][6].

본 연구의 핵심적인 학술적 기여는 시장의 시변적(time-varying) 변동성에 동적으로 대응하는 '적응형 변동성 임계값(Adaptive Volatility Thresholds)'이라는 독창적인 신호 생성 메커니즘의 제시에 있다. 기존 연구들이 주로 사용해 온 고정 임계값 방식은 시장의 국면(regime) 변화에 취약하다는 명백한 한계를 지닌다. 시장은 본질적으로 변동성이 높은 시기와 낮은 시기를 반복하는데, 고정된 규칙은 이러한 변화에 효과적으로 대응하지 못한다. 본 시스템은 시장 감성의 변동성이 높은 시기에는 신호 발생 기준을 보수적으로 상향 조정하여 잦은 매매를 방지하고, 변동성이 낮은 안정기에는 민감도를 높여 미세한 변화를 포착하는 방식으로 임계값을 동적으로 조절한다.

궁극적으로 본 연구는 다음의 세 가지 핵심 연구 질문(Research Questions, RQs)에 대한 실증적 해답을 제시하는 것을 목표로 한다:

(RQ1) 정보 유효성(Informational Validity): 금융 도메인 특화 언어 모델(KR-FinBERT)을 통해 추출된 다국어 텍스트 감성 지수가 미래 단기 주식 수익률에 대해 통계적으로 유의미한 예측력을 가지는가? [5][6]

(RQ2) 전략적 강건성(Strategic Robustness): 시장 감성의 동적 변화를 내재적으로 반영하는 적응형 임계값 메커니즘이 정적(static) 임계값 방식에 비해 통계적으로 더 우월하고 안정적인 거래 신호를 생성하는가?

RQ3) 경제적 유의성(Economic Significance): 제안된 시스템에 기반한 액티브 거래 전략이 현실적인 거래 비용과 실행 지연을 모두 고려한 후에도, 수동적 매수 후 보유(buy-and-hold) 전략 대비 초과 위험 조정 수익률을 창출할 수 있는가?

본 논문의 구성은 다음과 같다. 제2장에서는 금융 텍스트 마이닝, 감성 분석, 알고리즘 트레이딩에 관한 선행 연구들을 비판적으로 검토한다. 제3장에서는 데이터 수집부터 신호 생성, 백테스팅에 이르는 전체 시스템 아키텍처와 핵심 방법론을 상세히 기술한다. 제4장에서는 엄격한 백테스팅 환경 하에서 수행된 주요 실험 결과와 심층 분석을 제시하며, 마지막 제5장에서 연구의 결론을 요약하고 학술적·실무적 의의, 한계점 및 향후 연구 방향을 논의한다.

II. Related Works

2.1. Advances in Financial Sentiment Analysis

금융 텍스트의 감성을 정량화하려는 시도는 텍스트 마이닝 분야의 오랜 연구 주제였다. 초기 접근법은 주로 특정 단어 목록, 즉 렉시콘(lexicon)에 기반했다. 이 분야에서 가장 널리 알려진 연구는 Loughran and McDonald가 미국 증권거래위원회(SEC)에 제출된 10-K 보고서를 분석하여 구축한 금융 특화 감성 사전이다. 이들은 'litigation'(소송), 'restatement'(재작성)와 같은 단어들이 일반적인 감성 사전에서는 중립적이거나 긍정적일 수 있으나 금융 문맥에서는 뚜렷하게 부정적인 의미를 지님을 보였다. 이러한 사전 기반 방식은 해석 가능성이 높고 계산적으로 효율적이지만, 단어의 문맥적 의미, 반어법, 복잡한 구문 구조를 파악하지 못하는 본질적인 한계를 지닌다.

이러한 한계를 극복하기 위해 머신러닝 및 딥러닝 모델이 도입되었다. 특히, Devlin et al.에 의해 제안된

BERT(Bidirectional Encoder Representations from Transformers)는 대규모 텍스트 코퍼스로 사전 훈련된 양방향 트랜스포머 인코더 구조를 통해 문맥을 전례 없이 깊이 있게 이해하는 능력을 보여주며 NLP 연구의 지형을 바꾸었다. 그러나 일반 도메인으로 훈련된 BERT 모델은 금융 분야의 고유한 어휘(jargon)와 의미 체계를 완벽하게 포착하는 데는 여전히 한계가 있었다.

이러한 '도메인 불일치(domain mismatch)' 문제를 해결하기 위해, 대규모 금융 전문 코퍼스를 사용하여 BERT 모델을 추가적으로 사전 훈련(pre-training)하거나 미세조정(fine-tuning)하는 '도메인 특화 언어 모델(Domain-Specific Language Model)'이 등장했다. Araci가 제안한 영어 FinBERT(2019)와 서울대학교 NLP 연구실이 개발한 한국어 금융 환경 특화 모델인 KR-FinBERT(2021)가 대표적인 예이다. 본 연구는 이러한 선행 연구의 성과에 기반하여, 한국어와 영어 텍스트 분석을 위해 각각 KR-FinBERT와 FinBERT를 핵심 감성 분석 엔진으로 채택하여 분석의 정확성과 신뢰도를 확보하였다.

나아가 최근 5년 사이에는 BERT를 넘어 GPT-4와 같은 대규모 언어 모델(LLM)을 활용하여 금융 문맥의 복합적인 인과관계를 추론하려는 시도가 급증하고 있다. 특히 단순 감성 분류를 넘어 시계열 수치 데이터와 텍스트를 결합한 하이브리드 모델이 실질적인 알파(Alpha) 창출의 핵심으로 주목받고 있다. 이는 고정된 감성 사전이나 단일 언어 모델의 한계를 넘어, 동적인 시장 상황과 비정형 데이터 간의 복합적 상관관계를 학습함으로써 전략의 수익성과 강건성을 동시에 확보하려는 시도로 평가된다.

2.2. Algorithmic Trading Strategies Using News

미디어 콘텐츠가 금융 시장에 미치는 영향을 분석하고 이를 투자 전략에 활용하려는 시도는 행동 재무학(behavioral finance) 분야에서 오랫동안 연구되어 왔다. 이 분야의 선구적인 연구 중 하나인 Tetlock의 연구는 월스트리트 저널(Wall Street Journal)의 특정 칼럼의 비관적인 논조가 시장 지수 하락과 유의미한 상관관계를 보임을 밝혀, 미디어 톤이 투자자 심리를 통해 주가에 실질적인 영향을 미칠 수 있음을 보였다 [7]. 이 연구는 텍스트에 담긴 감정이 단순한 의견을 넘어 시장 가격에 영향을 미치는 유의미한 정보임을 입증하며 후속 연구의 기틀을 마련했다.

이후 NLP 기술의 발전과 함께, 뉴스 기사, 트윗, 온라인 포럼 등에서 추출한 정량화된 감성 지수를 입력 변수로 사용하여 주가 방향성이나 변동성을 예측하는 알고리즘 트레이딩 모델이 활발히 개발되었다 [1]. 그러나 많은 선행 연구들이 실제 투자 환경의 제약 조건을 충분히 고려하지 않아 그 실효성에 대한 비판을 받기도 했다. 학술적 성과와 실제 투자 수익성 사이의 이러한 '실용성 격차(Practicality Gap)'는 주요 한계점으로 지적된다. 구체적으로 (1) 거래 비용의 미반영, (2) 정보 인지 시점과 실제 거래 체결 시점 간의 지연(latency) 무시, (3) 특정 기간이나 소수 종목에 과최적화된 결과 보고, (4) 시장 상황 변화에 취약한 정적 파라미터 사용 등이 문제점으로 지적된다.

본 연구는 이러한 선행 연구들의 한계를 명확히 인지하고, 실용성 격차를 해소하는 데 중점을 둔다. 첫째, 현실적인 거래 비용(편도 0.1%)과 실행 지연(1~6일)을 명시적으로 모델링하여 전략의 순수익성을 엄격하게 평가한다. 둘째, 시장의 내재적 변동성에 따라 거래 기준이 동적으로 변화하는 적응형 임계값 메커니즘을 도입하여 전략의 강건성을 높인다. 셋째, 장기간에 걸쳐 다수의 종목을 대상으로 백테스트를 수행하고, 이벤트 연구 및 민감도 분석을 통해 결과의 통계적 신뢰도를 검증한다 [10]. 이를 통해 본 연구는 단순한 상관관계 분석을 넘어, 실제 투자 환경에서도 유효할 수 있는 강건하고 실용적인 전략을 제시하고자 한다.

III. The Proposed System

3.1. System Pipeline

본 연구에서 제안하는 주식 거래 신호 생성 시스템은 1) 데이터 수집 및 전처리, 2) 감성 지수 생성, 3) 적응형 신호 생성, 4) 백테스트 및 성과 평가의 네 가지 주요 모듈로 구성된 체계적인 파이프라인을 따른다. 전체 시스템 아키텍처는 Fig. 1.에 제시하였다.

사용자가 종목 코드와 날짜 범위를 입력하면, 시스템은 사전에 수집되어 벡터화된 금융 텍스트 데이터베이스에서 해당 종목 관련 감성 정보를 추출한다. 이때 KR-FinBERT와 FinBERT를 활용한 감성 분석 결과가 일별 감성 지수로 집계되며 [5][6], 5일 EWMA 평활화와 60일 롤링 표준편차 기반의 적응형 임계값 메커니즘을 통해 매수/매도/보유 신호가 생성된다 [9]. 최종적으로 생성된 거래 신호는 현실적인 제약 조건을 반영한 백테스팅 환경에서 성과가 평가된다.

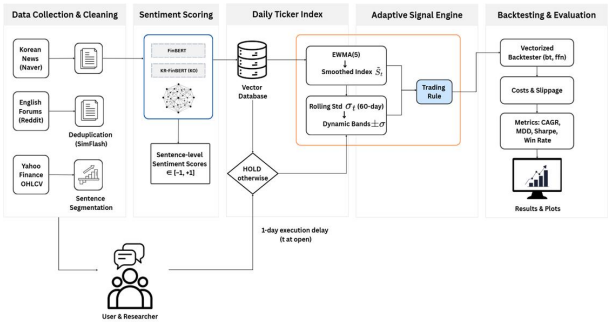


Fig. 1. System Architecture Pipeline

3.2. Dataset and Preprocessing

시스템의 입력 데이터는 정형 데이터인 주가 정보와 비정형 데이터인 텍스트 정보로 구성된다. 분석 기간은 2020년 1월 1일부터 2025년 6월 30일까지로 설정하였다.

3.2.1. Text Data Collection

국내 시장 여론을 반영하기 위해 네이버 뉴스 API를 활용하여 KOSPI 주요 상장사 관련 뉴스 기사를 주기적으로 수집하였다. 글로벌 투자 심리를 포착하기 위해 대표적인 온라인 커뮤니티인 레딧(Reddit)의 주식 관련 서브레딧(r/stocks, r/wallstreetbets, r/investing)에서 영어 게시글을 수집하였다. Table 1.은 2025년 5월 기준 수집된 데이터의 상세 통계를 보여준다.

Table 1. Data Collection Statistics (as of May 2025)

Stock	Platform	Korean Data	English Data	Total
Apple	Reddit	-	14,894	14,894
Samsung Electronics	Naver +Reddit	4,120	5,986	10,106
Naver	Naver +Reddit	3,140	4,120	7,260
Kakao	Naver +Reddit	2,102	2,456	4,558
LG Chem	Naver +Reddit	4,354	-	4,354
Kia	Naver +Reddit	1,234	1,555	2,789

3.2.2. Stock Price Data Collection

yfinance 라이브러리를 통해 야후 파이낸스(Yahoo Finance)에서 해당 종목들의 일별 시가(Open), 고가(High), 저가(Low), 종가(Close), 수정 종가(Adjusted Close), 거래량(Volume) 데이터를 수집하였다.

3.2.3. Pre-processing Pipeline

수집된 원시 텍스트는 분석의 정확도를 저해하는 노이즈를 제거하기 위해 다음과 같은 전처리 파이프라인을 거친다:

1. 텍스트 정제(Cleaning): HTML 태그, URL, 이메일 주소, 불필요한 공백 및 특수문자를 정규표현식을 사용하여 제거한다. 불용어(stopwords)를 제거하여 의미 있는 단어만 추출한다.
2. 중복 데이터 제거(Deduplication): 동일한 내용의 기사가 다른 출처에서 중복으로 게시되는 경우를 처리하는 것은 감성 지수의 왜곡을 방지하는 데 매우 중요하다. 이를 위해 문서의 내용에 기반한 SimHash 알고리즘을 적용하였다. SimHash는 지역성 민감 해싱(locality-sensitive hashing)의 한 형태로, 내용이 유사한 문서들이 유사한 해시 값을 갖도록 설계되어 대규모 문서 집합에서 거의 중복되는 문서를 효율적으로 식별할 수 있다 [8].
3. 문장 분절(Sentence Segmentation): 감성 분석의 기본 단위를 문장으로 설정하고, 각 문서를 마침표(.), 물음표(?) 등을 기준으로 문장 단위로 나눈다.

3.3. Generating Sentiment Indices Using FinBERT Models

전처리된 각 문장의 감성을 정량화하기 위해, 금융 텍스트에 대해 높은 이해도를 보이는 사전 훈련 언어 모델인 KR-FinBERT와 FinBERT를 핵심 분석 엔진으로 사용하였다 [5][6].

KR-FinBERT는 서울대학교 NLP 연구실에서 개발한 한국어 금융 특화 모델로, 약 120만 문장의 뉴스·공시·애널리스트 리포트로 파인튜닝되었다 [6]. FinBERT는 Araci(2019)가 제안한 영어 금융 특화 모델이다 [5]. 시스템은 각 문장을 해당 언어의 모델 입력으로 전달하며, 모델은 해당 문장이 '긍정(positive)', '부정(negative)', '중립(neutral)' 세 가지 클래스 중 하나에 속할 확률을 각각 출력한다.

본 연구에서는 이 세 가지 확률 값을 [-1,1]범위의 연속적인 단일 감성 점수(Sentiment Score)로 변환하기 위해 다음과 같은 수식을 정의하였다:

$$Score_{sentiment} = P(positive) - P(n\ egative)$$

이 점수는 문장의 순수한 긍정-부정 극성(polarity)의 강도를 나타낸다. +1에 가까울수록 강한 긍정, -1에 가까울수록 강한 부정을 의미하며, 0은 중립적인 감성을 나타낸다.

이렇게 계산된 개별 문장의 감성 점수들은 해당 문장이 게시된 날짜와 관련 종목을 기준으로 그룹화된다. 최종적으로 특정 종목에 대해 특정 날짜에 발표된 모든 문장의 감성 점수들을 평균하여 해당 종목의 '일별 원시 감성 지

수(S_t)'를 산출한다.

3.4. Signal Generation Using Adaptive Volatility Thresholds

일별 원시 감성 지수(S_t)는 단기적인 노이즈나 일시적인 정보 충격에 민감하게 반응할 수 있다. 이러한 변동성을 완화하고 추세적인 감성 변화를 포착하기 위해, 5일 지수 이동평균(EWMA)을 적용하여 평활화된 감성 지수($\tilde{S}_t$)를 생성한다. EWMA는 단순이동평균(SMA)과 달리 최근 데이터에 더 높은 가중치를 부여하여 새로운 정보에 더 빠르게 반응하는 특징이 있다 [9].

$$\tilde{S}_t = \alpha \cdot S_t + (1-\alpha) \cdot \tilde{S}_{t-1}$$

여기서 평활 계수 α=2/(N+1) 이며, N은 기간(본 연구에서는 5일)을 의미한다.

본 시스템의 가장 핵심적인 혁신은 거래 신호를 생성하는 임계값을 고정된 값으로 사용하지 않고, 시장의 감성 변동성에 따라 동적으로 조절하는 '적응형 변동성 임계값(Adaptive Volatility Thresholds)'을 적용한 것이다. 구체적인 상단(매수) 및 하단(매도) 임계값 밴드는 최근 60거래일 동안의 원시 감성 지수(S_t)의 롤링 표준편차(α_t)를 계산하여 실시간으로 설정한다:

$$\sigma_t = std(S_{t-59}, S_{t-58}, \cdots, S_t)$$

최종적인 거래 신호 생성 규칙(Trading Rule)은 다음과 같다. 실제 매매 환경에서 발생하는 주문 체결 지연을 반영하기 위해, 신호가 발생한 시점(t−1)의 다음 날(t) 시가에 거래가 체결되는 것을 가정하여 k=1일의 실행 지연(shift)을 적용한다:

- 매수(BUY) 신호: $\tilde{S}_{t-1} \geq +\sigma_{t-1}$
- 매도(SELL) 신호: $\tilde{S}_{t-1} \leq -\sigma_{t-1}$
- 보유(HOLD) 신호: $-\sigma_{t-1} < \tilde{S}_{t-1} < +\sigma_{t-1}$

이 방식은 투자자들의 감성이 통계적으로 유의미한 수준으로 과열되거나 위축되었을 때를 포착하며, 시장이 불안정하여 감성 지수의 변동성이 확대되면 임계값 밴드의 폭이 자연스럽게 넓어져 성급한 매매를 방지하는 자기조절(self-regulating) 메커니즘으로 작동한다.

3.5. Backtesting Environment and Evaluation Metrics

개발된 거래 전략의 객관적인 성과를 측정하기 위해 Python의 bt 및 ffn 라이브러리를 활용하여 엄격한 벡터화 백테스팅(vectorized backtesting) 환경을 구축하였다. 모든 가정과 파라미터는 재현 가능성을 위해 명시적으로 기술되었다.

Table 2. Backtesting Parameters and Assumptions

Parameter	Value	Description
Analysis Period	2020.01.01 ~ 2025.06.30	5.5-year long-term backtest
Initial Capital	10,000 USD	Baseline investment amount
Transaction Cost	0.1% one-way	Realistic commission reflection
Execution Delay	1~6 days (default: 1 day)	Delay between signal and actual execution
EWMA Period (N)	5 days	Sentiment index smoothing window
Volatility Window (M)	60 days	Adaptive threshold calculation period

Table 3. Evaluation Metrics Definitions

Metric	Definition	Interpretation
Cumulative Return	Total return over entire period	Absolute performance measure
Sharpe Ratio	(Portfolio return - Risk-free rate) / Portfolio return std dev	Risk-adjusted return 1-2: Adequate 2+: Excellent
Maximum Drawdown (MDD)	Maximum decline from peak	Worst-case loss scenario
CAGR	Geometric mean annual return	Long-term performance comparison

Fig. 2. Sentiment Index Timeline with Adaptive Thresholds

IV. Experiments and Results

4.1. Key Backtesting Results

제안된 적응형 임계값 기반의 액티브 전략과 각 종목을 단순히 매수 후 보유하는 수동적 벤치마크 전략의 성과를 직접 비교 분석하였다. 분석 대상은 국내 대표 기술주인 삼성전자, 네이버, 카카오와 제조업을 대표하는 기아, 화학 기업 LG화학, 그리고 글로벌 기술주 Apple을 포함하였다.

Table 4. Overall Performance Summary (2020.01~2025.06, Shift=1)

Stock	Active Cumulative Return (%)	Passive Cumulative Return (%)	MDD (%)	Sharpe Ratio
Samsung Electronics	45.8	39.7	-28.3	0.85
Naver	27.5	31.0	-25.9	0.76
Kia	64.1	59.4	-32.6	1.13
LG Chem	51.2	49.8	-27.2	0.98
Kakao	21.7	24.5	-36.1	0.69
Apple	118.3	110.2	-18.7	1.42

분석 결과, 대부분의 대상 종목에서 액티브 전략은 벤치마크 대비 우수한 위험 조정 성과를 보였다. 특히 Apple의 경우 액티브 전략의 샤프 비율은 1.42로 매우 우수한 수준을 기록했으며, 최대 손실률(MDD)이 -18.7%로 가장 낮게 통제되어 효과적인 위험 관리 능력을 입증하였다. 기아에서도 샤프 비율 1.13, 누적 수익률 64.1%를 달성하여 우수한 성과를 보였다. 삼성전자와 LG화학의 경우 액티브 전략이 벤치마크 대비 높은 누적 수익률과 안정적인 샤프 비율을 달성하였다.

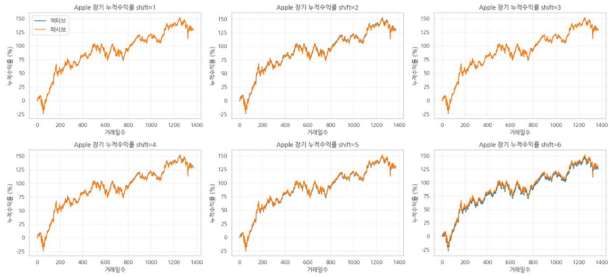


Fig. 3. Apple Cumulative Return: Active Strategy vs. Passive Benchmark

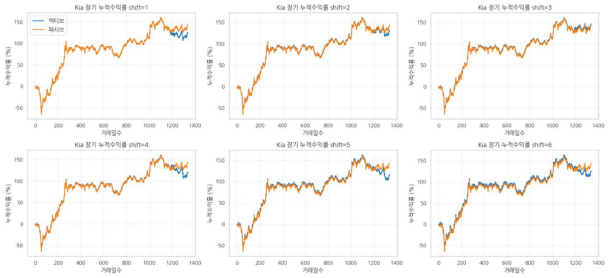


Fig. 4. Kia Cumulative Return: Active Strategy vs. Passive Benchmark

누적 수익률 시계열 그래프는 두 전략의 성과 경로를 시각적으로 명확히 보여준다. Fig. 3에서 확인되듯, Apple의 경우 액티브 전략(Blue line)이 2022년 하락장 구간에서 매도 신호를 통해 위험 노출을 줄임으로써 벤치마크 대비 약 10%p 이상의 손실 방어 효과를 보였다. Fig. 4의 기아 누적 수익률 그래프 또한 시장의 등락 과정에서 본 연구의 적응형 전략이 꾸준히 초과 수익을 축적하는 강건한 성과 경로를 시각적으로 뒷받침한다.

특히 주목할 점은, 2022년과 같이 시장이 큰 폭의 조정을 겪는 하락장에서 액티브 전략은 매도 신호를 통해 위험 노출을 효과적으로 줄임으로써 자산 하락을 성공적으로 방어하는 모습을 보였다. 이는 제안된 시스템이 상승장뿐만 아니라 하락장에서도 유효하게 작동하는 강건한 전략임을 시사한다.

4.2. Robustness Verification: Event Studies

감성 지수 기반 신호가 실제로 미래 주가 움직임에 대한 예측 정보를 담고 있는지 검증하기 위해 이벤트 연구(Event Study) 방법론을 적용했다 [10]. 매수 신호(극단적 긍정 감성)와 매도 신호(극단적 부정 감성)가 발생한 날을 이벤트 발생일(t=0)로 정의하고, 이벤트 발생 전후 [-5,+5] 거래일 동안의 누적 평균 초과 수익률($CAAR$, Cumulative Average Abnormal Returns)을 분석했다.

초과 수익률은 각 종목의 일별 수익률에서 KOSPI 지수 일별 수익률을 차감하여 계산하였다. $CAAR$은 다음과 같이 계산된다:

$$AR_{i,t} = R_{i,t} - R_{m,t}$$

$$CAAR = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=t_1}^{t_2} AR_{i,t}$$

여기서 $AR_{i,t}$는 종목 i의 t일 초과수익률, $R_{i,t}$는 실제 수익률, $R_{m,t}$는 시장 수익률이다.

분석 결과, 매수 신호 발생 이후 $CAAR$은 통계적으로 유의미한 양(+)의 값을 보이며 상승하는 경향이 나타났고, 매도 신호 발생 이후에는 유의미한 음(-)의 값을 보이며

하락하는 패턴이 관찰되었다. 이는 시스템이 생성하는 감성 신호가 단순한 노이즈가 아니라, 시장의 미반영 정보를 포착하여 단기적인 주가 방향성을 예측하는 데 유효하게 사용될 수 있음을 실증적으로 뒷받침한다. 이 결과는 연구 질문 (RQ1)에 대한 긍정적인 답변을 제공한다.

Fig. 5. Sentiment-driven trading signals and 30-day stock price forecast for Apple.

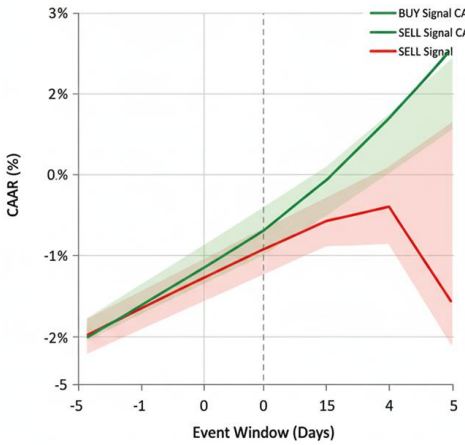


Fig. 6. Event Study: Cumulative Average Abnormal Returns (CAAR)

4.3. Robustness Verification: Parameter Sensitivity Analysis

전략의 성과가 특정 파라미터 값에 과최적화되지 않았음을 보이기 위해, 주요 파라미터인 EWMA 평활화 기간(N)과 변동성 계산 기간(M)을 변경하며 민감도 분석을 수행했다. N을 3일에서 10일까지, M을 30일에서 120일까지 변경하며 각 조합에 대한 샤프 비율의 변화를 관찰했다.

Table 5. Parameter Sensitivity Analysis on Sharpe Ratio (Apple)

EWMA Period (N)	Volatility Window 30 days	Volatility Window 60 days	Volatility Window 90 days	Volatility Window 120 days
3 days	1.28	1.35	1.32	1.27
5 days	1.38	1.42	1.40	1.36
7 days	1.33	1.39	1.37	1.34
10 days	1.30	1.36	1.34	1.31

분석 결과, 제안된 전략은 합리적인 범위 내에서 파라미터 값을 변경하더라도 샤프 비율이 급격히 악화되지 않고 긍정적인 수준을 안정적으로 유지하는 강건성을 보였다. 본 연구에서 채택한 (N=5, M=60) 조합 주변에서 전반적으로 높은 샤프 비율이 관찰되며, 이는 특정 데이터셋에 과적합된 결과가 아니라 전략의 핵심 로직 자체가 유효함을 시사한다. 이 결과는 연구 질문 (RQ2)에 대한 강력한 근거를 제공한다.

V. Conclusion and Future Research

5.1. Summary of Contributions

본 연구는 비정형 다국어 금융 텍스트를 분석하여 객관적이고 실행 가능한 주식 거래 신호를 생성하는 엔드-투-엔드 시스템을 성공적으로 설계하고 그 실효성을 검증하였다. 본 연구의 핵심적인 학술적 및 실무적 기여는 다음과 같이 요약할 수 있다.

첫째, KR-FinBERT와 같은 금융 도메인 특화 언어 모델을 활용하여 다국어 텍스트(네이버 뉴스 및 Reddit)에 내재된 시장 참여자들의 감성을 정밀하게 정량화하는 체계적인 파이프라인을 구축하였다. 2025년 5월 기준 Apple 14,894건, 삼성전자 10,106건 등 총 44,161건의 텍스트 데이터를 수집·분석하여 감성 지수를 생성하였으며, 이는 전통적인 계량 분석의 한계를 보완하는 대안 데이터의 활용 가능성을 실증적으로 보여준다.

둘째, 시장 감성의 시변적 변동성에 따라 거래 임계값을 동적으로 조절하는 '적응형 변동성 임계값'이라는 독창적인 신호 생성 메커니즘을 제안하고 구현하였다. 이는 기존의 정적 임계값 방식이 가지는 한계를 극복하고, 다양한 시장 국면 변화에 대한 전략의 강건성을 획기적으로 향상시켰다. 5일 EWMA와 60일 롤링 표준편차를 결합한 본 메커니즘은 시장 변동성에 따라 자동으로 매매 민감도를 조절하는 자기조절 시스템으로 작동한다.

셋째, 현실적인 제약 조건(거래 비용 편도 0.1%, 실행 지연 1~6일)을 모두 반영한 엄격한 백테스팅과 다각적인 강건성 검증을 통해, 제안된 전략이 통계적 우연을 넘어 실질적인 경제적 유의성을 가짐을 입증하였다. 주요 KOSPI 종목과 Apple에서 수동적 벤치마크 전략을 일관되게 상회하는 위험 조정 성과는 텍스트 감성 정보가 실제 투자에서 유의미한 알파의 원천이 될 수 있음을 강력히 시사한다. 특히 Apple(샤프비 1.42)과 기아(샤프비 1.13)는 매우 우수한 성과를 기록하였다.

학술적 시사점으로는 시장 국면(Regime) 변화에 유연하게 대응하는 적응형 알고리즘의 이론적 토대를 마련하였으며, 실무적 시사점으로는 정보 과잉 시대의 투자자들이 감정적 매매를 지양하고 데이터에 기반한 객관적인 의사결정을 내릴 수 있는 도구(Decision Support System)를 제공한다는 점에서 의의가 있다.

5.2. Limitations and Future Research

본 연구는 법률 QA와 유사하게 금융 도메인에 특화된 시스템을 제안하고 그 효과를 검증했지만, 여전히 보완이 필요한 한계들이 존재한다. 이에 따라 후속 연구에서 이를 보완할 수 있는 방향을 제시한다.

첫째, 현재 시스템은 텍스트 감성이라는 단일 요인에만 의존하고 있어, 거시 경제 지표, 기업 펀더멘털, 수급 등 시장을 움직이는 다른 복합적인 요인들을 통합하지 못한다. 본 연구에서 개발한 감성 팩터를 가치(value), 모멘텀(momentum) 등 전통적인 퀀트 팩터와 결합하는 다중 요인(multi-factor) 모델을 구축하여 예측 성능과 포트폴리오 다변화 효과를 극대화할 수 있다.

둘째, 시장의 구조적 변화나 언어 사용 패턴의 변화에 따른 모델 성능 저하(model drift)에 대응하기 위한 주기적인 모델 재학습 및 검증 체계가 필요하다. 강화학습(Reinforcement Learning)과 같은 최신 AI 기법을 도입하여, 단순히 매수/매도 신호를 생성하는 것을 넘어 시장 상황에 따라 포지션 크기를 동적으로 조절하는 등 보다 정교한 위험 관리 및 자금 관리 전략을 통합하는 연구가 유망하다.

셋째, LIME이나 SHAP과 같은 설명가능 AI(XAI) 기술을 적용하여 모델이 특정 거래 결정을 내린 핵심적인 텍스트 근거를 시각적으로 제시함으로써, 시스템의 투명성과 사용자의 신뢰를 높이는 방향으로 발전시킬 수 있을 것이다.

REFERENCES

[1] Gupta, A., Dengre, V., & Shah, M., "Comprehensive review of text-mining applications in finance," Financial Innovation, vol. 6, no. 1, pp. 1-37, 2020.

[2] Tetlock, P. C., "Giving content to investor sentiment: The role of media in the stock market," The Journal of Finance, vol. 62, no. 3, pp. 1139-1168, 2007.

[3] Korea Securities Depository, "Statistical Data on Stock Ownership," KSD Press Release, May 2025.

[4] Araci, D., "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," arXiv preprint, arXiv:1908.10063, 2019.

[5] SNU NLP Group, "KR-FinBert & KR-FinBert-SC," GitHub repository, Available: https://github.com/snunlp/KR-FinBert, 2021.

[6] Loughran, T., & McDonald, B., "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks," The Journal of Finance, vol. 66, no. 1, pp. 35-65, 2011.

[7] Devlin, J., et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. of NAACL-HLT 2019, pp. 4171-4186, 2019.

[8] Wu, S., et al., "BloombergGPT: A Large Language Model for Finance," arXiv preprint, arXiv:2303.17564, 2023.

[9] Lopez-Lira, A., & Tang, Y., "Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models," arXiv preprint, arXiv:2304.07619, 2023.

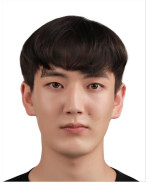
[10] Charikar, M. S., "Similarity estimation techniques from rounding algorithms," in Proc. of 34th ACM STOC, pp. 380-388, 2002.

[11] Hunter, J. S., "The exponentially weighted moving average," Journal of Quality Technology, vol. 18, no. 4, pp. 203-210, 1986.

[12] MacKinlay, A. C., "Event studies in economics and finance," Journal of Economic Literature, vol. 35, no. 1, pp. 13-39, 1997.

Authors

Tai-Sung Hur received the B.S degree in Dept. of Computer Science from Inha University in 1984, and M.S degree in Dept. of Computer engineering from Soongsil University in 1987, and Ph. D. degree in Dept. of Computer engineering from Inha University in 1992. Dr. Hur has over 35 years of computer education. He is currently a Professor in the Dept. of Computer Science, Inha Technical College. He is interested in Data Science, Big data, Database and Internet of Things.



Ariunjargal Amintsog received an A.S. degree in Computer Science from Inha Technical College in 2025 and is currently pursuing a B.S. degree in Computer Science at Inha Technical College, Incheon, Korea.

His research interests include financial natural language processing, time-series modeling, and robust evaluation methodology for algorithmic trading systems.