https://doi.org/10.7840/kics.2022.47.11.1900

감성 라벨링을 이용한 주식 트레이딩 시스템 개발

박 명 석◆, 김 재 윤°

Developing a Trading System for the Stock Market Using Sentiment Labeling

Meyong-seok Park◆, Jaeyun Kim°

요 약

주식 시장의 방향성을 예측한 기존 연구들은 주로 정형 데이터를 활용하였다. 본 연구는 기존에 활용되지 못했던 비정형 데이터 (뉴스와 커뮤니티 데이터)를 이용해 시장 참여자들의 감성을 분석하여 이를 활용한 감성 라벨링 기반 트레이딩 시스템을 제안하였다. 감성 사전 구축 시 주식 시장에 적합한 감성 사전을 구축하기 위해 코사인 유사도를 활용하여 새로운 감성 사전을 구축했다. 뉴스 데이터는 개별 종목의 감성을 대표하기 위해 사용했으며, 커뮤니티 데이터는 개인 투자자들의 감성을 대표하기 위해 사용했다. 개별 종목 감성과 개인 투자자들의 감성을 종합적으로 고려하기 위해 뉴스와 커뮤니티 데이터를 혼합하여 사용했다. 제안한 감성 라벨링 방법의 유용성을 살펴보기 위해 국내에 상장된 기업들 중 시가총액 상위 20위까지 분석 대상으로 선정하였으며, 감성분석을 위한 학습모형은 KoBERT와 LSTM 모형을 적용하였다. 트레이딩 결과 감성 라벨링 방법이 기존 주가의 등락을 예측하는 방법인 Up/Down 라벨링 방법보다 트레이딩 성과 지표들을 비교 분석한 결과, 트레이딩 성과가 개선되는 것을 확인하였다.

Key Words : Unstructured data, Sentiment Analysis, Trading System, Deep Learning, FinTech

ABSTRACT

Most of the studies predicting the direction of the stock market have been used structured data. This study analyzed the sentiment of market participants using unstructured data (news and community data) that were not previously used and developed an sentiment labeling-based trading system using them. In order to build an sentiment dictionary suitable for the stock market when constructing an sentiment dictionary, a new sentiment dictionary was constructed using cosine similarity. News data were used to represent the sensibilities of individual stocks, and community data were used to represent the sensibilities of individual investors. News and community data were combined to comprehensively consider individual stock sentiment and individual investor sentiment. As a result of the experiment, it was confirmed that the sentiment labeling method performed better than the Up/Down labeling method, which is a method of predicting fluctuations in stock prices. Through this study, it was confirmed that using sentiment analysis of the market and individual investors can make excess profits in the stock market.

※ 본 연구는 2021년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과 (2021-0-01399)와 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구입니다 (NRF-2022R1A2C1092808). 또한 순천향대학교 학술연구비 지원으로 수행하였음.

◆ First Author : Soonchunhyang University Department of Future Convergence Technology, pmsk980122@sch.ac.kr, 학생회원
° Corresponding Author : Soonchunhyang University Department of Big Data, kimym38@sch.ac.kr, 정회원
논문번호 : 202208-181-0-SE, Received August 15, 2022; Revised September 17, 2022; Accepted September 22, 2022

Ⅰ. 서 론

오랫동안 주식 시장의 방향성을 예측하는 대다수의 연구들은 주로 미리 정한 형식과 구조에 따라 저장된 정형 데이터 (structured data)를 이용하였다. 주식 시장에서의 정형 데이터는 주식의 가격 (시가, 고가, 저가, 종가)과 거래량, 기본적 지표, 기술적 지표 등을 의미한다. 하지만 다양한 요인들로 영향을 받는 주식 가격을 정형 데이터만으로 활용하여 예측하는 것은 쉽지 않은 일이다. 이에 정형 데이터 외의 기존에 활용되지 못한 비정형 데이터 (unstructured data)를 활용한 주식 시장의 방향성 연구들이 늘어나고 있다. 여기서 비정형 데이터란 정형 데이터와 반대로 정의된 구조 없이 형성된 데이터이다. 비정형 데이터의 예로 비디오/텍스트/이미지 데이터 등이 있다[1].

본 연구에서는 시장 참여자들의 감성을 분석해 주가의 방향성을 예측하고 이를 활용한 트레이딩 전략을 구축하기 위해 감성 라벨링 기법을 제안한다. 감성 분석 기법이란 텍스트의 긍정, 부정 등의 감성을 분류하는 분석으로 오피니언 마이닝의 한 부분이다. 감성 분석 기법을 활용해 단어의 감성적인 특성을 부여할 수 있는 감성사전 구축이 가능하다. 감성 사전은 범용 감성사전과 도메인 맞춤형 감성 사전으로 나눌 수 있다. 범용 감성사전은 일반적으로 보편적인 상황에 적용이 가능한 극성 값을 부여한 사전을 말하며 도메인 맞춤형 감성사전에 비해 빠르게 사전 구축이 가능하다. 하지만 문장의 의미에 따라 단어의 극성이 달라질 수 있어 오류 발생 가능성이 존재한다. 이를 극복하기 위해 도메인 맞춤형 감성사전을 구축할 수 있다. 도메인 맞춤형 감성사전은 단어에 해당하는 극성 값을 부여하는 건 물론 해당 분야에 긍정, 부정적으로 느껴지는 사건에 대해서도 극성 값을 부여할 수 있다[2].

주식 시장에서 감성 분석을 활용한 대표적 연구는 다음과 같다. 김유신[3] 등은 오피니언 마이닝을 통해 뉴스 컨텐츠를 분석하기 위해 주가지수의 등락을 예측하는 지능형 투자 의사결정 모형을 제시했다. 모형 검증을 위해 오피니언 마이닝 결과와 주가 등락 간의 관계를 통해 분석하고 결과값과 주가지수 등락이 유의한 관계를 가짐을 확인했다. 김재봉과 김형중[1]은 기존 연구가 뉴스의 형식과 한정된 어휘만을 사용했기 때문에 이를 개선하기 위해 증권 전문 사이트인 팍스넷 (Paxnet)의 게시 글을 분석 대상으로 주식 시장 맞춤형 감성사전을 구축했다. 장은하 등[4]은 뉴스 감성 분석과 다양한 거시경제 지표를 고려해 효과적인 지표 조합을 찾고 미국 다우존스지수를 예측했다. 뉴스 정보의 감성 분석을 진행하기 위해 자연어처리 기법인 BERT와 NLTK VADER를 사용했다. 주가 예측 모델로는 Long Short-Term Memory 모형을 적용하여 가장 효과적인 지표 조합의 성능을 확인하였다. 김명진 등[5]은 SNS의 댓글 데이터가 주식 가격의 변동에 영향을 주는지 실험하고 1시간 후의 가격 변동 방향과 가격 변동의 폭에 대한 예측력을 지니는지 확인했다. 실험은 20개 종목을 대상으로 진행했으며 13개 종목에 대해서 주가 이동 방향을 50% 이상의 정확도를 가짐을 확인했고 16개 종목에 대해서 주가 변동폭을 50% 이상의 정확도를 가짐을 확인했다. 그러나 앞서 소개한 연구들은 비정형 데이터를 뉴스에 국한시키거나 상대적으로 짧은 길이의 데이터를 가지고 분석을 진행하였다.

본 논문에서는 개별 종목에 해당하는 뉴스와 커뮤니티(네이버의 종목토론방)의 텍스트 데이터를 수집한 후, 개별 종목마다 적합한 감성사전을 구축하였다. 구축된 감성사전을 바탕으로 주가 방향성에 대한 라벨링을 위해 감성 라벨링 방법을 제안하였으며, 이를 활용한 트레이딩 전략 및 시스템을 개발하였다. 제안 방법의 검증을 위해 시가 총액 20위 종목에 대한 뉴스와 커뮤니티 텍스트 데이터를 학습 데이터로 활용하였다. 감성 분석을 위한 모델로는 KoBERT과 LSTM 알고리즘을 적용하였으며, 트레이딩 성과를 비교분석하였다.

본 논문은 다음과 같이 구성된다. 2장에서는 감성 분석을 활용한 주가 예측의 선행 연구와 KoBERT 모형의 이론적인 내용을 간단히 설명한다. 3장에서는 본 논문에서 제안하는 트레이딩 전략 및 시스템에 대한 프레임워크를 설명한다. 4장에서는 본 논문에서 제안한 트레이딩 시스템의 성과를 비교 분석하며, 5장에서는 연구 결과를 요약하고 추후 연구에 대해 논의한다.

Ⅱ. 관련 연구

2.1 감성 분석을 활용한 주가 예측

감성분석을 활용한 주가 예측에 관한 대표적인 연구들은 다음과 같다. 김세완 등[6]은 소셜 감성을 긍정 및 부정적 의견으로 구분하여 이들 의견이 개별 기업의 주식수익률에 미치는 영향이 비대칭적인지 분석하였다. 이를 분석하기 위해 트위터 의견이 충분한 기아차, 아모레퍼시픽, 포스코, 한국전력 등 4개 기업을 선정하였다. 분석 결과 트위터의 긍정 또는 부정적 의견이 주식수익률에 비대칭적으로 영향을 미치는 것을 확인하였으며, 트위터의 의견을 투자자 심리 대용 변

수로 활용할 수 있음을 보여주었다. 김다예와 이예인[7]은 주가 등락에 주목해 주가 등락을 예측할 수 있는 감성사전을 구축하는 새로운 방법을 제안하였다. 25,000여건의 증시 뉴스를 수집하고 문맥을 고려하기 위해 Word2Vec을 적용하였다. 뉴스 감성분석을 실시하여 KOSPI 종가 지수를 예측한 경우, 약 50%의 정확도를 보여주는 것을 확인하였다. Jin 등[8]은 투자자의 감성을 고려하기 위해 심리지수 (sentiment index)를 제안하였으며 LSTM을 이용하여 주가를 예측한 경우에 예측 정확도가 개선됨을 확인하였다. Ren 등[9]은 텍스트 데이터가 투자자의 심리를 반영한다고 보고 서포트 벡터 머신 (Support vector machine, SVM)을 활용한 감성 분석을 진행하였다. 감성 분석을 통해 상하이50 (SSE50)지수의 등락 예측 정확도가 18.6% 상승하여 89.93%까지 개선됨을 확인하였다. 이러한 결과는 투자자의 감성이 주식시장의 선행 지표 중 하나로 간주 될 수 있다는 결과를 보여준다. Bharathi and Geetha[10]은 주식 뉴스를 긍정, 부정, 중립으로 구분하고 투자자들의 행동을 예측하여 투자자들의 매매 타이밍을 예측하는 연구를 진행하였다. 실험 결과로는 ID3, C4.5 알고리즘과 이동 평균 지표와 비교했을 때, 14.43%의 정확도 개선 효과를 확인하였다. Alkubaisi 등[11]은 트위터 데이터를 활용해 긍정, 부정, 중립으로 극성을 라벨링하고 분류를 위해 하이브리드 나이브 베이즈 분류기를 사용하였다. 실험 결과 90.38% 라는 높은 정확도가 나오는 것을 확인하였다. 앞서 언급한 연구들은 감성 분석을 통해 단순히 주가의 등락만을 예측하였다. 그러나 본 연구는 주가의 등락 예측보다는 감성 분석을 활용한 트레이딩 전략을 제안하고 그 성과를 분석하고자 한다.

2.2 KoBERT

BERT (Bidirectional Encoder Representations for Transformers)은 Google에서 개발된 딥러닝 모델로 자연어처리 분야에서 가장 우수한 성능을 보이는 모델 중 하나이다. BERT는 트랜스포머에 기반한 모델로 사전학습 후 fine-tuning을 통해 문제를 해결한다. BERT의 입력은 총 3가지의 임베딩으로 표현된다. 첫 번째, 토큰 (Token) 임베딩은 토큰의 의미 표현을 의미하며, Word piece 임베딩 방식을 사용한다. 이러한 방식은 Out of Vocabulary 처리에 효과적이며 정확도 상승효과가 있다. 두 번째, 세그멘트 (Segement) 임베딩은 문장과 문장을 문장 구분자로 이어주는 표현이다. 세 번째, 포지션 (Position) 임베딩은 단어들의 절대적인 위치 정보를 담고 있다[4]. 본 연구에서는 BERT 모델 중 KoBERT모델을 사용하여 감성 분류를 진행하였으며, KoBERT 모델은 한국어 처리 성능을 높이기 위해 한국어 데이터를 사전학습을 진행한 BERT 모델이다.

Ⅲ. 연구 방법

본 연구의 프레임워크는 Figure 3.1과 같으며 총 3단계로 구성된다. Step 1에서는 뉴스와 커뮤니티 데이터를 크롤링을 통해 해당 종목과 관련된 텍스트 데이터 (뉴스 및 게시판)를 수집한다. 수집된 데이터를 바탕으로 개별 기업에 해당하는 감성사전을 구축한 후, 각 텍스트의 감성을 라벨링하여 학습 데이터를 구성한다. 감성 라벨링 방법은 벤치마크 감성사전과 코사인 유사도를 측정해 개별 기업 종목에 해당하는 감성사전을 구축하고, 구축된 감성 사전으로 각 텍스트의 긍정/부정 점수를 부여한다. 그리고 일자별로 해당하는 텍스트 데이터는 한 개가 아니기 때문에 각 텍스트의 긍정/부정 점수를 바탕으로 일별 감성을 표현할 수 있는 일별 감성 라벨링을 진행한다. Step 2에서는 구축된 학습 데이터를 바탕으로 KoBERT와 LSTM 알고리즘을 활용해 감성 분석을 위한 학습을 진행한 뒤 긍정/부정 또는 상승/하락에 해당하는 예측값을 생성한다. Step 3에서는 각 모델의 예측값을 바탕으로 거래 전략을 수립하고 트레이딩 시뮬레이션을 진행한다. 트레이딩 시뮬레이션을 진행한 결과는 트레이딩 성과 지표를 통해 평가한다.

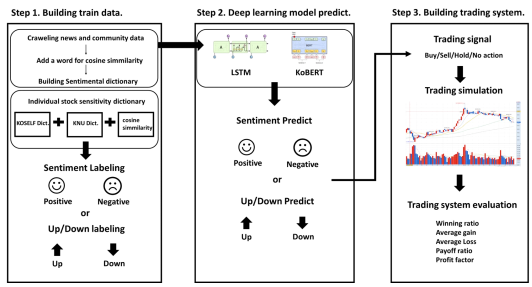

그림 1. 제안 모델 프레임워크
Fig. 1. The framework of the proposed model

3.1 데이터 수집

본 연구에서 수집한 주식 종목은 국내에 상장된 기업들 중 시가총액 20위까지 분석 대상으로 선정하였다. 해당 종목들의 데이터 수집을 위해 Python의 Beautiful Soup 패키지를 사용하였으며, 종목별 매일 경제 뉴스 데이터와 네이버 주식 토론방 데이터를 크

표 1. 주식 종목 리스트
Table 1. Stock list

Stock Name	Stock Ticker
Samsung Eelectronics	005930
SK hynix	000660
Cellrtrion	068270
LG chemical	051910
Hyundai Motor	005380
KB Financial	105560
LG H&H	051900
LG Eelectronics	066570
Naver	035420
Posco	005490
SK Innovation	096770
KIA	000270
Samsung SDI	006400
Samsung C&T	028260
Samsung Biologics	207940
Shinhan Financial	055550
NCsoft	036570
Kakao	035720
Hyundai Mobis	012330
SK	034730

롤링하였다. 뉴스 데이터의 구성은 기사 제목과 기사 본문이며, 커뮤니티 데이터는 종목 토론방 게시글의 제목으로 구성하였다. 실험에 사용된 총 데이터 기간은 2018년 1월 1일부터 2020년 12월 30일까지이며 제안 모델의 검증을 위해 2018년 1월 2일부터 2019년 12월 30일까지 학습 기간으로, 2020년 1월 1일부터 2020년 12월 30일까지 테스트 기간으로 설정하였다. 본 연구에서 사용한 주식 종목 리스트는 표 1과 같다.

3.2 감성사전 구축

3.2.1 감성사전

벤치마크 감성사전으로 두 개의 감성사전을 사용한다. 첫 번째로 조수지 등[12]이 제안한 KOSELF 감성사전을 활용하였으며, 이는 기업 재무 분석에 필요한 감성사전을 구축 및 검증하였다. 2016년부터 2018년까지 한국에서 발행된 약 2만 개의 애널리스트 보고서를 사용해 감성사전을 구축했으며, 보고서 별로 5개 감성사전을 통해 계산한 부정어, 긍정어 비율 등과 목표주가와의 관계를 검증하였다. 두 번째로 KNU 감성사전이며[13] 이는 특정 도메인에서 사용되는 긍·부정보다는 기본 감정 표현을 나타내는 긍·부정로 구성된다. 긍정 표현 예로는 '감동받다', '가치 있다'와 보편적인 부정 표현의 예로는 '그저 그렇다', '열받다' 등이 존재한다. 각 도메인의 감성사전을 빠르게 구축하기 위한 기초 자료로 활용하기 위해 개발되었다.

위와 같은 감성사전들은 주식시장에 특화된 감성사전이 아니기 때문에 주식시장이라는 도메인에 알맞은 감성사전을 구축하고 개별 주식 종목마다 해당하는 긍·부정 단어들이 차이가 있을 것이기 때문에 개별 주식 종목에 해당하는 감성사전 구축이 목표이다.

3.2.2 코사인 유사도

코사인 유사도는 두 벡터 사이의 코사인 각도를 측정함으로써 두 벡터가 얼마나 유사한가를 측정하는 방법이다. 두 벡터가 같은 방향을 가리키고 있는가를 측정하고 방향이 같을수록 유사함을 의미한다. 두 벡터가 비트 벡터일 때, 식 (1)과 같이 표현된다. 코사인 유사도는 -1에서 1사이에 값을 가질 수 있고 1에 가까울수록 유사함을 의미한다.

$$smilarity(A,B) = \frac{A \cdot B}{|A| \times |B|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (1)$$

본 연구에서는 분석하고자하는 대상 기업의 뉴스와 커뮤니티 글에서 단어을 추출한 뒤 벤치마크 감성사전에 등록된 긍부정 단어들과 코사인 유사도를 측정해 코사인 유사도 0.75이상인 긍부정 단어들을 추가했다. 코사인 유사도 0.75이상인 긍정 단어 사례는 '떡상', '상승장', '부유' 등이 있고 부정 단어 사례는 '떡락', '하락장', '가난' 등이 있다.

3.3 감성 라벨링

3.3.1 감성 스코어링

각 텍스트의 긍정/부정 점수는 해당 날짜 뉴스와 커뮤니티 안에 구축한 감성 사전 긍정 단어 숫자만큼 +1을 하고 부정 단어 숫자만큼 +1을 하는 방식으로 식 (2)와 식(3)과 같이 부여된다.

$$postive\ \ score = \frac{\sum_{i=1}^{n} postive\ word}{\sum_{i=1}^{n} postive\ word + \sum_{i=1}^{n} negative\ \ word} > 0.6 \quad (2)$$

$$negative\ \ score = \frac{\sum_{i=1}^{n} negative\ \ word}{\sum_{i=1}^{n} postive\ word + \sum_{i=1}^{n} negative\ \ word} < 0.4 \quad (3)$$

3.3.2 일별 감성 라벨링

각 텍스트마다 긍/부정 라벨링을 진행한 후 일별 감성을 측정하기 위해 일별 감성 라벨링을 진행했다. 일별 감성 라벨링은 한국은행의 뉴스 심리지수(NSI,News Sentiment Index) 공식을 이용하였다. 뉴스심리지수(NSI)는 한국은행이 뉴스 기사에 나타난 경제 심리를 일 단위로 지수화한 것이다. 뉴스심리지수(NSI) 값이 100을 초과하면 긍정, 100미만이면 부정으로 일별 감성을 라벨링 했다. 뉴스심리지수의 공식은 식(4)와 같다.

$$NSI = \frac{positive\ sentences - negative\ \ sentences}{positive\ sentences + negative\ \ sentences} \times 100 + 100 \quad (4)$$

3.4 Up/Down 라벨링

Up/Down 라벨링은 T시점의 가격과 T+1 시점의 가격 차이를 바탕으로 양의 값을 가지면 Up, 음의 값을 가지면 Down으로 라벨링하는 방법이다. Up/Down 라벨링 방법은 기존 연구에서 주가 등락을 예측하기 위해 사용하는 라벨링 방법이다. 본 연구에서는 각 종목의 수정 종가를 바탕으로 Up/Down 라벨링을 했고 텍스트 데이터가 주가 등락 여부를 예측하여 수익을 낼 수 있는지 확인하고 감성 라벨링과 비교하기 위해 진행했다. 식(5)와 식(6)은 Up/Down 라벨링 공식이다.

$$Up = price_{t+1} > price_t \quad (5)$$

$$Down = price_{t+1} < price_t \quad (6)$$

3.5 트레이딩 신호 생성 및 시뮬레이션

학습된 KoBERT와 LSTM 모형을 테스트 데이터를 통해 예측값을 생성하고 생성된 예측값을 이용하여 트레이딩 신호를 생성한다. 예측값은 라벨링 방법에 따라 달라지며 긍정 (또는 상승)/부정 (또는 하락)으로 예측한다. 본 연구에서는 2가지 방법 (감성 라벨링과 Up/Down 라벨링)으로 라벨링을 진행하여 트레이딩 성과를 비교분석한다. 감성 라벨링은 긍정, 부정으로 라벨링이 진행되며 예측값이 긍정이면 매수 포지션, 예측값이 부정이면 매도 포지션으로 정의하였다. Up/Down 라벨링은 상승, 하락으로 라벨링이 진행되며 예측값이 상승이면 매수 포지션, 예측값이 하락이면 매도 포지션이다. 트레이딩 전략은 매수 포지션이 발생하고 나서 다음 매도 포지션이 발생하기 전까지는 주식을 보유하고, 매도 포지션이 발생하고 다음 매수 포지션이 발생하기 전까지는 거래를 하지 않는다. 성과 평가를 위해 매수 신호가 발생할 때 다음날 시가로 매수하고, 매도 신호가 발생할 때 다음날 시가로 매도한다. 실거래와 유사한 결과를 도출하기 위해 거래 수수료는 0.015%를 부과하여 실험을 진행했다. 표 2는 트레이딩 신호 생성의 예시이다.

표 2. 거래 신호 생성 예시
Table 2. Example of a process of generating a trading signals

Time	Label	Signal
1	Positive (Up)	Buy
2	Negative (Down)	Sell
3	Negative (Down)	No action
4	Positive (Up)	Buy
5	Positive (Up)	Holding
6	Negative (Down)	Sell
7	Negative (Down)	No action
8	Negative (Down)	No action
9	Negative (Down)	No action
10	Positive (Up)	Buy

3.6 데이터 전처리

본 연구에서는 형태소 분석기 Python의 mecab 패키지를 사용해 데이터 전처리를 진행하였다. 텍스트 전처리는 명사 및 형용사를 추출하였으며, 불용어 및 한 글자 단어는 모두 제거하였다. 또한 커뮤니티 데이터에서 광고글은 실험 결과에 영향을 미칠 수 있으므로 삭제했다. 그리고 주말 및 공휴일에 나온 뉴스 기사는 다음 개장일에 영향을 미칠 것이라고 판단하여 다음 개장일로 수정하였고 오후 3시 이후 업로드된 기사도 마찬가지로 다음 거래일자로 변경하였다.

3.7 딥러닝 모델 하이퍼 파라미터

본 연구에서 사용한 KoBERT와 LSTM 모형의 하이퍼 파라미터는 실험에 사용하였던 컴퓨팅 자원을 고려해 조정하였으며, KoBERT모형에 사용한 파리미터는 표 3과 같다. 그리고 LSTM의 파라미터는 epoch를 100으로, batch size는 10으로 설정했다.

표 3. KoBERT 하이퍼 파라미터
Table 3. KoBERT hyper parameter

No.	Parameter name	Parameters
1	max_seq_len	100
2	batch size	24
3	warmup_steps	0.1
4	epoch	3
5	max_grad_norm	1
6	logging_steps	200
7	learning rate	5e-5

Ⅳ. 연구 결과

4.1 모형 검증 방법

본 연구에서 제안한 모델의 성능을 평가하기 위해 트레이딩 평가 지표를 계산한다. 트레이딩 평가 지표는 거래 횟수(합계), 승률, 평균 수익, 평균 손실, Payoff ratio, Profit factor로 총 6개로 평가한다. 식 (7)은 평균 수익 ($\overline{w}$)과 평균 손실 ($\overline{L}$)을 나타낸다. 평균 수익은 총 수익을 수익거래 횟수로 나눈 값이며 평균 손실은 총 손실을 손실거래 횟수로 나눈 것이다. 식 (8)은 payoff ratio (P_r)과 profit factor (P_f)을 나타낸다. payoff ratio는 평균 수익을 평균 손실으로 나눈 값이며 profit factor는 총 수익을 총 손실로 나눈 값이다. 따라서 payoff ratio와 profit factor가 1보다 커야 장기적으로 수익이 창출할 수 있다.

$$\overline{w} = \frac{\sum w}{N_w}, \quad \overline{L} = \frac{\sum L}{N_L} \tag{7}$$

$$P_r = \frac{\overline{w}}{\overline{L}}, \quad P_f = \frac{\sum w}{\sum L} \tag{8}$$

4.2 트레이딩 시뮬레이션 결과 및 성과 비교

뉴스와 커뮤니티의 감성을 분류하기 위해 딥러닝 모델인 LSTM과 KoBERT 모델을 적용하였으며, 데이터 라벨링 방법은 제안 방법인 감성 라벨링과 전통적으로 주가 방향성 예측에 활용되고 있는 Up/Down 라벨링을 사용한 트레이딩 성과를 비교분석 하였다. 수집한 뉴스 데이터와 커뮤니티 데이터의 특성에 차이가 있을 것으로 판단되어 실험을 뉴스 데이터만 사용한 경우, 커뮤니티 데이터만 사용한 경우, 뉴스와 커뮤니티 데이터를 모두 사용한 경우로 구분하여 트레이딩 성과를 분석하였다.

표 4는 감성 라벨링 방법과 Up/Down 라벨링 방법을 이용한 평균적인 트레이딩 결과이다. 첫 번째로 감성 라벨링 방법을 이용한 KoBERT의 경우 커뮤니티 데이터만 사용했을 때 평균적으로 Payoff ratio와 Profit factor의 값이 1.56, 1.26으로 가장 높았으며, 평균 트레이딩 횟수가 11.4 로 한 달에 한 번 정도 거래되는 것을 확인할 수 있었다. 뉴스 데이터만 사용한 경우는 커뮤니티 데이터만 사용한 경우보다 트레이딩 성과가 비교적 낮지만 거래 횟수가 약 3배 늘어나는 것을 확인할 수 있었다. 뉴스와 커뮤니티 데이터 모두 사용한 경우에는 Payoff ratio와 Profit factor의 값이 각각 1.19, 1.17로 측정되었으며, 뉴스만 사용한 경우와 큰 차이를 보이지 않음을 알 수 있었다.

표 4. 라벨링 방법에 따른 트레이딩 시스템 평균 결과
Table 4. Average results of the trading systems according to labeling method

	No. trades	Win ratio	Average gain	Average loss	Payoff ratio	Profit factor
Sentiment KoBERT						
Only news	37.95	0.51	7945.17	8457.04	1.06	1.16
Only community	11.4	0.51	9133.01	6354.61	1.56	1.26
News + Community	39.9	0.48	10070.2	9003.43	1.19	1.17
Sentiment LSTM						
Only news	14.90	0.49	7993.81	6731.81	1.24	1.75
Only community	8.45	0.50	6538.14	6901.73	1.20	0.92
News + Community	18.4	0.45	8741.86	8157.11	1.28	1.44
Up/Down KoBERT						
Only news	35.05	0.48	8571.90	9062.04	1.10	1.01
Only community	13.25	0.29	4979.68	5417.39	1.12	0.57
News + Community	29.25	0.50	10655.4	8723.31	1.30	1.06
Up/Down LSTM						
Only news	42.7	0.47	7776.30	7936.33	1.00	0.79
Only community	17.65	0.38	6090.05	6655.01	0.81	0.44
News + Community	33.25	0.47	8955.63	9577.87	0.95	0.75

반면에 LSTM을 통해 감성분석을 진행한 경우에는 전반적으로 KoBERT 보다 상대적으로 적은 평균 거래 횟수로 보였으며, 보다 우수한 성과가 도출되었음을 알 수 있다. 뉴스 데이터만 사용했을 경우 Payoff ratio와 Profit factor가 1.24, 1.75로 측정되었으며, 뉴스와 커뮤니티 데이터를 모두 사용한 경우에는 Payoff ratio와 Profit factor가 1.28, 1.44의 값을 얻었다. 하지만 이러한 결과는 KoBERT와 LSTM을 활용한 트레이딩 성과가 통계적으로 차이가 있다고 볼 수는 없다. 단, KoBERT와 LSTM의 알고리즘과 상관없이 뉴스와 커뮤니티 데이터를 모두 사용한 결과가 뉴스, 커뮤니티만 사용한 트레이딩 결과보다 덜 민감한 편임을 알 수 있다. 이러한 결과는 감성 라벨링 방법을 활

용시 뉴스와 커뮤니티 데이터를 모두 활용하는 것이 보다 안정적인 트레이딩 성과를 도출함을 알 수 있다.

두 번째로, 대다수의 기존 연구에서 사용하고 있는 Up/Down 라벨링 방법을 적용한 트레이딩 성과이다. KoBERT의 경우 뉴스와 커뮤니티 데이터를 모두 학습 데이터로 이용한 결과 Payoff ratio와 Profit factor 값이 1.30과 1.06으로 가장 우수한 결과를 보였다. LSTM의 경우에는 뉴스 데이터만 사용했을 때 Payoff ratio와 Profit factor의 값이 1.00, 0.79로 가장 우수한 성과를 도출하였다. 하지만 Profit factor의 값이 1보다 작으므로 장기적으로는 손실이 발생될 가능성이 높기 때문에 트레이딩 시스템으로 적합하지 않음을 알 수 있다. 그리고 전반적으로 Up/Down 라벨링 방법은 감성 라벨링에 비해 트레이딩 성과가 낮음을 알 수 있었다.

V. 결론 및 향후 연구

본 연구는 비정형 데이터를 활용해 시장 참여자들의 감성을 분석하여 이를 활용한 감성 라벨링 기반 주식 트레이딩 시스템을 개발하였다. 감성 사전 구축 시 각 기업별로 주식 시장에 알맞은 감성 사전을 구축하기 위해 코사인 유사도를 활용해 새로운 맞춤형 감성 사전을 구축했다. 또한 텍스트 데이터가 주가의 등락을 예측할 때 수익이 나는지도 실험했다. 뉴스의 경우 개별 종목 시장 감성을 대표하기 위해 사용했으며, 커뮤니티의 경우 개인 투자자들의 감성을 대표하기 위해 사용했다. 개별 종목 시장 감성과 개인 투자자들의 감성을 종합적으로 고려하기 위해 뉴스와 커뮤니티 데이터를 합하여 사용했다. 실험 결과는 감성 라벨링 방법을 진행했을 때 Up/Down 라벨링 방법의 비해 모든 성능이 좋음을 확인했다. 이는 본 논문에서 제안한 개별 종목에 해당하는 주식 감성사전이 주식 시장이라는 도메인에 텍스트의 감성을 잘 분석했다고 사료된다. 또한 시장과 개인 투자자들의 감성분석을 이용하는 것이 주식 시장에서 초과 수익을 낼 수 있음을 확인하였다.

본 연구의 한계점은 다음과 같다. 첫 번째는 감성 스코어링 시에 임계값을 0.6, 0.4로 임의로 결정했다는 것이다. 임계값을 변경한다면 입력 데이터의 크기에 변화가 생길 수 있다. 두 번째는 KoBERT 모형과 LSTM 모형의 하이퍼파라미터 값을 컴퓨팅 자원의 한계로 임의로 조정하였기 때문에 성능의 변화가 생길 수 있다. 한계점을 보완한 연구는 다음과 같다. 감성 스코어링 시 단순히 긍정, 부정으로 라벨링하지 않고 중립을 추가하여 다중 분류를 하는 연구가 필요하다. 그리고 딥러닝 모형의 하이퍼 파라미터를 임의로 조정하지 않고 Grid Search 방법과 같은 파라미터를 최적화하는 알고리즘을 이용하면 성능 개선이 가능하다. 또한 국내 주식 시장뿐만 아니라 해외 주식 시장에도 적용하여 본 연구에서 제안하는 방법의 유용성을 확인할 필요가 있다.

References

[1] J. Kim and H. Kim, "A domain-specific sentiment lexicon construction method for stock index directionality," *J. Digital Contents Soc.*, vol. 18, no. 2, pp. 585-592, Jun. 2007. (https://doi.org/10.9728/dcs.2017.18.3.585)

[2] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Lang. Technol.*, vol. 5, no. 1, pp. 1-167, 2012. (https://doi.org/10.2200/S00416ED1V01Y201204HLT016)

[3] Y. Kim, N. Kim, and S. Jeong, "Stock-index invest model using news big data opinion mining," *JIIS*, vol. 18, no. 2, pp. 143-156, Jun. 2012. (https://doi.org/10.13088/jiis.2012.18.2.143)

[4] E. Jang, H. Choi, and H. Lee, "Stock prediction using combination of BERT sentiment analysis and macro economy index," *J. Korea Soc. Comput. and Inf.*, vol. 25, no. 5, pp. 47-56, 2020. (https://doi.org/10.9708/jksci.2020.25.05.047)

[5] M. Kim, J. Ryu, D. Cha, and M. Sim, "Stock price prediction using sentiment analysis: from "Stock discussion room" in Naver," *J. Soc. for e-Busin. Stud.*, vol. 25, no. 4, pp. 61-75, 2020. (https://doi.org/10.7838/jsebs.2020.25.4.061)

[6] S. Kim, J. Park, Y. Kim, and H. Ham, "Asymmetric effect of social sentimental on an individual stock price return," *Inf. Syst. Rev.*, vol. 22, no. 4, pp. 59-74, 2020. (https://doi.org/10.14329/isr.2020.22.4.059)

[7] D. Kim and Y. Lee, "News based stock market sentiment lexicon acquisition using Word2Vec," *J. Bigdata*, vol. 3, no. 1, pp. 13-20, 2018.

(https://doi.org/10.36498/kbigdt.2018.3.1.13)

[8] Z. Jin, Y. Yang, and Y. Liu, “Stock closing price prediction based on sentiment analysis and LSTM,” *Neural Comput. and Appl.*, vol. 32, no. 13, pp. 9713-9729, 2020. (https://doi.org/10.1007/s00521-019-04504-2)

[9] R. Ren, D. D. Wu, and T. Liu, “Forecasting stock market movement direction using sentiment analysis and support vector machine,” *IEEE Syst. J.*, vol. 13, no. 1, pp. 760-770, 2018. (https://doi.org/10.1109/JSYST.2018.2794462)

[10] S. Bharathi and A. Geetha, “Sentiment analysis for effective stock market prediction,” *Int. J. Intell. Eng. and Syst.*, vol. 10, no. 3, pp. 146-154, 2017. (https://doi.org/10.22266/ijies2017.0630.16)

[11] G. A. A. J. Alkubaisi, S. S. Kamaruddin, and H. Husni, “Stock market classification model using sentiment analysis on twitter based on hybrid naive bayes classifiers,” *Comput. Inf. Sci.*, vol. 11, no. 1, pp. 52-64, 2018. (https://doi.org/10.5539/cis.v11n1p52)

[12] S. Cho, H. Kim, and C. Yang, “Building the Korean Sentiment Lexicon for Finance (KOSELF),” *KJFS*, vol. 50, no. 2 pp. 135-170, 2021. (https://doi.org/10.26845/KJFS.2021.04.50.2.135)

[13] S. Park, C. Na, M. Choi, D. Lee, and B. On, “KNU Korean Sentiment Lexicon: Bi-LSTM-based method for building a Korean Sentiment Lexicon,” *JIIS*, vol. 24, no. 4, pp. 219-240, 2018. (https://doi.org/10.13088/jiis.2018.24.4.219)

박 명 석 (Meyong-seok Park)

2021년 2월 : 순천향대학교 빅데이터공학과 졸업

2022년 3월~현재 : 순천향대학교 미래융합기술학과 석사과정

<관심분야> 빅데이터, AI, 텍스트 마이닝, 강화학습

김 재 윤 (Jaeyun Kim)

2009년 8월 : 국민대학교 비즈니스IT 학과 졸업

2015년 8월 : 연세대학교 정보산업공학과 박사

2018년 3월~현재 : 순천향대학교 빅데이터공학과 조교수

<관심분야> 금융빅데이터분석, 머신러닝, 메타휴리스틱

[ORCID:0000-0001-7855-8969]