논문 2022-17-07

뉴스 감성 앙상블 학습을 통한 주가 예측기의 성능 향상

(An Accurate Stock Price Forecasting with Ensemble Learning Based on Sentiment of News)

김 하 은, 박 영 욱, 유 시 은, 정 성 우, 유 준 혁*
(Ha-Eun Kim, Young-Wook Park, Si-eun Yoo, Seong-Woo Jeong and Joonhyuk Yoo)

Abstract : Various studies have been conducted from the past to the present because stock price forecasts provide stability in the national economy and huge profits to investors. Recently, there have been many studies that suggest stock price prediction models using various input data such as macroeconomic indicators and emotional analysis. However, since each study was conducted individually, it is difficult to objectively compare each method, and studies on their impact on stock price prediction are still insufficient. In this paper, the effect of input data currently mainly used on the stock price is evaluated through the predicted value of the deep learning model and the error rate of the actual stock price. In addition, unlike most papers in emotional analysis, emotional analysis using the news body was conducted, and a method of supplementing the results of each emotional analysis is proposed through three emotional analysis models. Through experiments predicting Microsoft's revised closing price, the results of emotional analysis were found to be the most important factor in stock price prediction. Especially, when all of input data is used, error rate of ensembled sentiment analysis model is reduced by 58% compared to the baseline.

Keywords : Stock Price Prediction, Sentiment analysis, Macro economic indicator

Ⅰ. 서 론

정확한 주가 예측은 매우 어려운 문제이다. 효율적 시장 가설 (Efficient-market hypothesis) [1]에 따르면 현재 얻을 수 있는 모든 정보가 주가에 빠르게 반영된다고 가정한다. 이에 따라 지속적으로 생성되는 모든 새로운 정보를 얻을 수 없기 때문에 주가 예측은 매우 어려운 문제라고 결론 내어진다. 또 무작위 행보 이론 (RandomWlak Theory) [2]에 따르면 불확실성을 가지는 과거 추세를 바탕으로 미래 가치를 예측하는 것은 거의 불가능하다고 결론 내려졌다. 이렇게 주식시장은 불확실성 (uncertainty), 비선형 (nonlinear), 동적 (dynamic), 비모수적 (nonparametric), 노이즈 환경 (noisy environment) 특성 때문에 매우 도전적인 과제로 간주되어 [3], 과거부터 현재까지 주식시장 예측을 위해 다양한 연구가 진행되고 있다.

주가 예측은 이전에 관측된 정보를 바탕으로 미래 가치를 측정하는 시계열 분석에 초점을 둔다. 전통적인 방식은 이전 데이터와 이후 데이터를 각각 독립 변수와 종속 변수로 가정하여 이들 사이의 양적 관계만을 얻는 것을 목표로 하는 선형 모델을 사용 하였다 [4]. 하지만 이러한 방법은 상황 정보와 투자자의 심리와 같은 주식시장을 움직이는 다양한 영향 요인은 무시하기 때문에 복잡하고 불확실한 주식시장을 예측하기엔 어려움이 있다. 또한 주식시장 분석을 위한 데이터는 크기가 매우 크고 비선형적이다. 이를 처리하기 위해 대규모 데이터 세트 안에 있는 숨겨진 패턴과 복잡한 관계를 식별할 수 있는 모델이 필요로 하였지만, 기존의 선형 분석 방법으로는 거의 불가능 하였다.

최근 자연어 처리 분야는 딥러닝 기술을 사용함으로써 급속도로 발전하였고 이미 사람의 자연어 처리 능력을 뛰어넘었다. 이는 인공 신경망의 한 종류인 순환 신경망의 발전 때문이다. 순환 신경망은 스스로 반복하면서 이전 단계에서 얻은 정보를 내부의 메모리를 이용해 시변적 동적 시퀀스 형태의 특징들을 처리할 수 있다 [5, 6]. 이를 통해 전통적인 주가분석 방법의 문제점인 대규모 데이터의 비선형적 패턴 추출을 자동으로 하게 되어 일부분 해결되었다. 따라서 주가 예측을 위해 딥러닝 모델을 사용하여 다수의 연구가 진행되고 있고, 최근에는 다양한 입력데이터를 사용하여 주가를 예측하고 있다.

주식 가격은 여러 가지 요인들이 내포돼 상호작용하며 책정된다. 그 중 거시적 경제지표는 실업률, 이자율, 물가, 국민소득, 환율, 통화량, 국제수지 등 국가차원의 경제 상황 전반을 뜻하며 이는 한 나라 경제 전체의 움직임을 보여주고, 이들의 변동이 주가 변동의 주요 요인이라고 밝혀졌다 [7]. 이러한 이유로 최근 거시적 경제를 이용한 딥러닝 주가 예측 연구가 다양하게 진행되고 있다 [8-10]. 또 투자자들의

*Corresponding Author (joonhyuk@daegu.ac.kr)
Received: Oct. 11, 2021, Revised: Nov. 11, 2021, Accepted: Dec. 6, 2021.
H.E. Kim : Daegu University (B.S. Student)
Y.W. Park : Daegu University (B.S. Student)
S.E. Yoo : Daegu University (B.S. Student)
S.W. Jeong: Daegu University (M.S. Student)
J. Yoo : Daegu University (Full Prof.)
※ 이 연구는 2021년도 정부 (과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2020R1A2C1014768 중견연구자지원사업).

심리가 주가 시장 투자에 중요한 역할을 할 수 있다. 왜냐하면 인간은 기본적으로 비이성적으로 판단하고 행동한다. 이는 주식시장을 비합리적으로 만들고 주식시장을 설명하는 과학적 원칙을 반드시 따르지 않도록 만든다. 최근의 연구들은 투자자들의 심리가 주식 시장 투자에 중요한 역할을 할 수 있다는 것을 보여주었다. 예를 들어, 이메일의 내용과 다우존스 지수와의 연관성에 대한 실험을 진행하였고, 이메일의 내용이 주식 시장 변화를 예측하는데 도움이 된다는 것을 발견했다. 또 온라인 리뷰와 주식 거래량 사이의 상관관계 또한 확인했다 [11]. 한편으로 뉴스와 분석자료 등 다양한 참고문헌이 투자자의 심리변화에 중요한 역할을 하여 의사결정에 영향을 미치는 것을 확인했다 [12]. 이렇듯 심리와 주가는 강한 상관관계를 보여준다. 특히 뉴스에서 보도되는 기업의 호재와 악재는 투자자의 심리를 크게 좌지우지하게 되며 이는 주가에 상당부분 반영되는 것으로 보인다. 따라서 뉴스 감성분석을 통해 해당 글의 긍정 또는 부정적 감성의 수치화는 주가 예측에 매우 중요한 특징으로 간주된다. 이에 따라 감성분석을 통한 주가 예측 연구도 다양하게 진행되고 있다 [13-15].

하지만 지금까지의 연구들은 개별적으로 진행이 되었기에 각각의 방법들의 객관적인 비교가 어려우며 이들이 주가예측에 미치는 영향에 대한 연구도 아직 미흡한 것으로 보인다. 따라서 본 논문에서는 거시 경제지표와 뉴스 감성분석 결과값의 각 특징들이 주가에 미치는 영향도를 평가하고 다양하게 조합함으로써 그 중 가장 효과적인 조합의 딥러닝 주가 예측 모델을 찾고 그것을 최적의 주가 예측 모델로 제안하고자 한다.

본 논문에서 주가 예측에 사용한 딥러닝 모델은 방대한 시계열 데이터에서 데이터 안의 숨겨진 패턴을 찾아낼 수 있고 과거의 데이터가 희석되는 문제점을 해결했다고 알려진 LSTM을 사용하였다. 주가예측을 위해 사용되는 특징들은 다음과 같다. 먼저 금과 미국 경제와 관련이 있는 일본 (JPN/USD), 중국 (CNY/USD), 유럽 (EUR/USD)의 화폐 교환비를 사용한다. 이는 주가에 영향을 미치는 거시 경제요인 중 금값과 환율이 가장 대표적인 값으로 하기 때문이다 [4]. 이들 거시 경제지표들은 각각의 값들이 축척 (sacle)의 차이가 심하다. 따라서 학습시 규모가 큰 특징에 높은 중요도를 두는 문제점을 해결하기 위해 최소-최대 정규화 방법을 통해 각각의 값들의 규모를 일정하게 맞춰주었다. 뉴스 감성분석으로는 뉴스기사의 본문을 사전기반의 감성분석과 트랜스포머 금융 데이터 기반의 단어 임베딩 모델을 통해 감성분석을 진행한 후에 나온 결과 값을 사용하고 Word2Vec을 통해 뉴스기사에 있는 단어들의 유사도를 파악하여 전체적인 감성분석에 도움을 주었다.

본 논문에서는 미국의 대표기업인 마이크로소프트 (MSFT)사의 수정종가를 예측하는 실험을 통해 실제 주가와 모델 예측값의 오류율을 바탕으로 각각의 특징값들의 영향력을 분석하였고, 최종적으로 모든 특징값들을 조합하여 사용했을 때 에러율이 가장 낮았으며 이를 최적의 주가 예측 모델인 SPP-MVBE로 제안하였다.

Ⅱ. 관련 연구

1. 전통적인 주가 예측 모델

주가예측을 위해 주로 사용되는 시계열 분석 방법은 변수의 과거 값의 선형 조합을 이용하여 관심 있는 변수를 예측하는 자기 회귀모델 (AR), 데이터의 평균값 자체가 시간에 따라 변화하는 경향을 시계형 모형으로 구성한 이동 평균모델 (MA), 그리고 자기회귀와 이동평균으로 확률적 시계열을 표현하는 자기 회귀 및 이동 평균모델 (ARMA)을 사용한다 [4]. 이러한 방법들은 주식시장의 비선형 특성을 분석할 때 매우 비효율적이기 때문에 주가 예측에 부적절하다.

2. 딥러닝을 사용한 주가 예측 연구

Cho는 양방향 LSTM 및 강화학습 모델을 제안함으로써 주식 시장의 방향성을 예측한 모델을 제안하였다. 5개의 주가 데이터와 11개의 거시 데이터를 활용하여 양방향 LSTM 및 강화학습으로 학습된 딥러닝 모델을 통해 가상 투자환경에서 삼성전자 주가 예측에 대한 실험을 진행하였다. 실험결과 주가 데이터와 기술지표만을 입력 데이터로 사용했을 때는 약 1.2% 정도의 수익을 내었지만, 거시경제 변수를 추가했을 때는 약 4.4%의 수익을 보여주어 주가 데이터와 기술지표만 입력 데이터로 사용하는 것보다, 거시 경제지표를 함께 입력데이터로 하여 예측하는 것이 더 나은 것을 실험을 통해 입증하였다 [8].

Kim은 SNS에서 수집된 댓글 데이터가 주식의 미래 가격의 변동에 미치는 영향에 대한 연구를 진행하였다. 네이버 주식토론방에서 20개 종목에 대한 6개월 간의 댓글 데이터를 사용하였으며. 이 데이터가 1시간 후의 주가 변동의 방향과 폭에 대한 예측력을 가지는지 실험한다. 예측 모델은 딥러닝 모델인 LSTM과 CNN을 활용하였다. 20개 종목 중 13개 종목에서 미래의 주가 이동 방향을 50% 이상의 정확도로 예측할 수 있다는 결과를 얻었고, 16개 종목에서 미래의 주가 변동폭을 50% 이상의 정확도로 예측할 수 있다는 결과를 얻었다 [13].

이와 같이 주가 예측을 위한 딥러닝 연구가 지속적으로 진행되고 있다. 하지만 해당 연구들은 각기 다른 입력데이터를 사용하여 진행되고 있기에 객관적인 비교가 어렵고, 각 데이터가 주가 예측에 미치는 영향이 잘 나타나고 있지 않은 문제점이 있다.

Ⅲ. SPP-MVBE

본 논문에서 제안하는 SPP-MVBE은 그림 1과 같이 주가데이터와 거시경제지표를 정규화 하는 전처리 모듈과 뉴스 기사의 앙상블된 감성분석 예측 모듈을 가지고 있으며 각 모듈의 결과값을 LSTM 기반 주가 예측기의 입력으로 하여 주가를 예측한다.

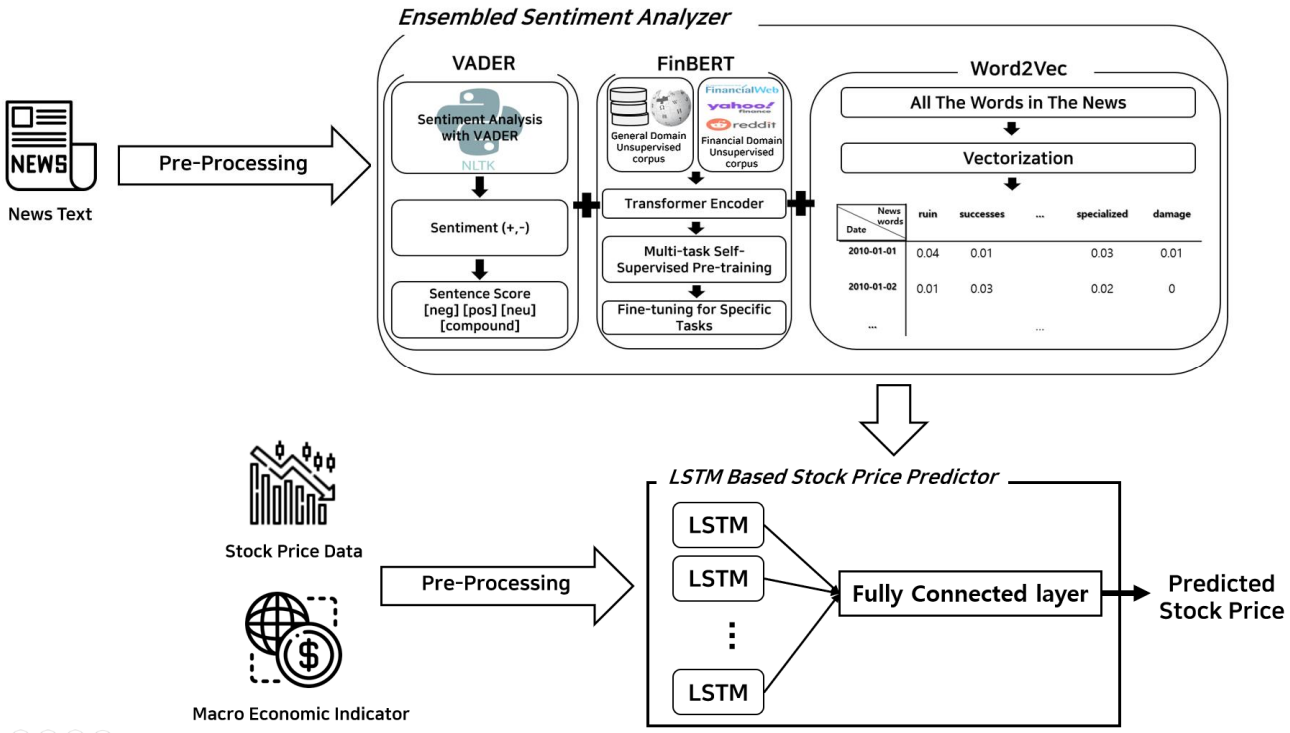

그림 1 SPP-MVBE 구조

Fig. 1. SPP-MVBE Architecture

1. 주가 데이터 및 거시경제지표 전처리

웹크롤링을 통해 얻은 주가 데이터와 거시 경제지표 데이터에서 예측에 필요한 값들만 추출하고 정제할 필요성이 있다. 먼저 주식시장이 열리지 않는 공휴일과 주말의 데이터를 제외하였다. 처리 후엔 각 데이터에서 평균값들과 달리 차이가 매우 큰 이상치 데이터를 발견할 수 있었으며, 각 데이터 사이의 축척이 심하게 차이나는 것을 발견할 수 있었다. 이상치 데이터는 딥러닝 모델 학습에 방해를 줄 수 있고, 각 데이터 중 규모가 큰 데이터의 영향력이 매우 커지는 문제점이 생긴다. 따라서 모든 특징들을 동일한 정도의 축적으로 반영되도록 최소-최대 정규화 (MIN-MAX Normalization)기법을 적용하였다. 최소-최대 정규화는 표준편차의 값이 작고 이상치의 영향을 작게 해준다는 결과가 있기 때문에 각 특징들이 동일한 영향력을 가진다고 할 수 있다 [16]. 이렇게 정규화된 값을 LSTM 모델의 입력 값으로 전달한다.

2. 앙상블 학습 기반 뉴스 감성 분석

감성분석은 각 문장에 나타난 패턴을 이용해 글의 감성을 분류하거나 수치화하여 객관적인 정보로 바꾸는 기술을 뜻한다. 좁은 의미로 감성분석은 텍스트 상의 긍정적, 부정적 감성을 분석하는 것이다. 본 논문에서는 우리는 감성분석시 대다수의 이런 연구에서 뉴스 제목을 사용하는 것과는 다르게 본문 전체를 사용한다. 이는 뉴스 기사의 제목은 흥미를 불러일으키기 위해 자극적인 문구를 사용하기 때문에 제목을 통해 감성분석을 진행할시 극단적인 경우 본문의 내용과 반대가 되는 감성분석의 결과를 얻을 수 있는 문제점이 있기 때문이다. 또한 뉴스 기사만 읽었을 때 유의미한 감성분석 키워드가 없을 수 있는 반면 본문에서는 감성분석 키워드가 다수 존재하기 때문에 감성분석에 있어 뉴스 본문을 사용하는 것이 매우 강건하고 효과적인 감성분석을 할 수 있다.

본 논문에서 사용하는 감성분석 모델은 사전 기반의 감성분석기와 금융 데이터 기반의 트랜스포머 모델을 앙상블하여 사용한다. 추가로 감성분석을 보완하기 위해 단어 임베딩 기술을 활용하여 뉴스 본문의 각 단어의 유사도를 계산한다.

2.1 사전 기반의 감성분석기 : VADER

VADER (Valence Aware Dictionary and sEntiment Reasoner)는 집단지성 기반으로 제작한 사전 기반 감성 분석 (Lexicon-based Sentiment Analysis) 도구이다 [17]. 사전 기반 방식은 사전 형식으로 작성된 긍정 (Positive) 단어와 부정 (Negative) 단어에 대한 점수 리스트에서 문장 혹은 문서에 속한 단어를 토대로 해당 문장이나 문서에 대한 점수를 합산하여 감성 점수를 산출하는 방법이다. VADER의 감성사전에는 긍정어 3,345개, 부정어 4,172개의 총 7,517개의 단어가 있으며 이중에는 이모티콘, 축약어, 속어 등도 포함되어 있다. 각 단어는 선별된 10명의 평가자들이 4가지 규칙에 따라 단어마다 0을 중립 (Neutral)으로 긍정이나 부정과 같은 속성에 따라 실수 값인 점수를 부여하였다. 이렇

표 1. VADER 감성점수 결과 비교값
Table 1. VADER sentiment score result

	Compound	Positive	Negative
RMSE (%)	0.01074	0.01652	0.01553

게 생성된 감성사전은 다양한 방식으로 사용되어 해당 문서의 전체 감성 점수를 계산한다. VADER의 장점은 이모티콘, 비속어, 축약어, 감정 기호 등에 대한 감정 분석을 할 수 있고, 감정의 강도를 수치화하여 상대적 비교가 가능한 것이다.

VADER 감성사전을 기반으로 뉴스 본문의 감성 점수 계산방법은 다양하다. 본 연구에선 주가 예측에 적합한 방법을 선정하기 위해 3가지의 방법을 비교하여 가장 오류율이 낮은 방법을 사용하였다. 각기 다른 계산방법을 비교하기 전에 먼저, 수집된 뉴스기사의 본문 내용을 마침표를 기준으로 문장별로 분리하고 텍스트들의 어간을 추출한 후 감성분석에 큰 의미가 없는 단어인 불용어를 제거하는 텍스트 전처리 과정을 거쳤다. 그다음 전체 단어에서 긍정어의 비율로 감성 점수를 평가하는 Positive 방법, 전체 단어에서 부정어의 비율로 감성 점수를 평가하는 Negative 방법, 그리고 부정적 동사, 대문자, 수식어의 위치, 느낌표 등에 따라 다른 가중치를 부여하여 감성 점수를 평가하는 Compound 방법 각각을 비교하였다. 표 1은 각각의 방법으로 계산된 감성 점수로 주가 예측을 진행하였을 때 생기는 오류율이다. 표에 따라 compound 방법이 가장 낮은 오류율을 보이기 때문에 본 논문에선 compound방법을 채택하였다.

2.2 금융 트랜스포머 모델 : FinBert

BERT (Bidirectional Encoder Representations for Transformers) 모델은 최근 다양한 자연어처리 분야에서 가장 우수한 성능을 보였다 [18]. BERT는 주의 메커니즘을 활용한 트랜스포머 (Transformer)에 기반을 둔 모델로 대규모 위키피디아 (Wikipedia) 기사 말뭉치를 사용해 사전 훈련 되었고 미세조정 (Fine-tuning)을 통해 여러 가지 자연어처리 문제에 많이 활용되고 있다.

BERT의 핵심적인 아이디어는 마스크된 언어 모델과 다음 문장 예측이다. 마스크된 언어 모델은 입력으로 사용되는 문장의 토큰에서 무작위하게 몇 개의 토큰을 Mask 토큰으로 변환시키고 이를 Transformer 구조에 넣어 주변 단어의 문맥만을 보고 Mask된 단어를 예측하는 모델이다. 그리고 다음 문장 예측은 사전 훈련시 두 문장을 동시에 입력으로 하여 문장 B가 이전 문장 A뒤에 자연스럽게 연결이 되는 문장인지 예측하도록 학습된다. 이렇게 BERT 구조는 양방향으로 동작하기 때문에 기존의 다음 단어 예측에 비해 양방향 예측이 가능하고 각 어휘들 간의 좀 더 긴 (long-term) 의존관계를 잘 포착할 수 있도록 해주는 장점을 가지고 있다.

FinBERT는 금융분야를 위한 BERT이다 [19]. 2008년과 2010년 사이에 발행된 Reuters TRC2 데이터 셋의 1.8백만 개의 뉴스 기사로 구성된 금융 텍스트 말뭉치에 대해 사전

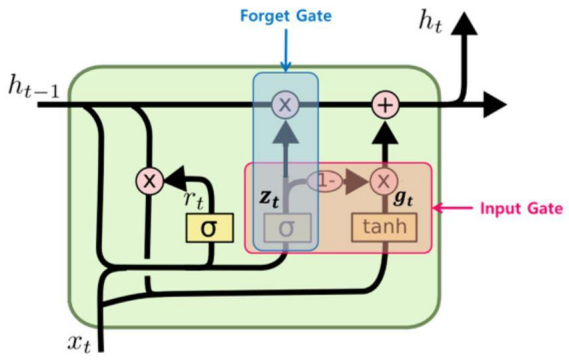

그림 2. LSTM 구조 [5]
Fig. 2. LSTM Architecture

교육되어 있다. FinBERT 모델은 다른 BERT와 비교하여 금융 분야 텍스트에서 텍스트 분류 작업의 정확도가 15% 향상되어 금융 분야에서 적합한 BERT모델인 것이 입증되었다. 따라서 본 논문에서는 주가 예측을 위해 금융관련 뉴스에서 감성분석을 진행하기 때문에 FinBERT를 사용하기 적합하다.

본 논문에서는 공개되어 있는 사전 학습된 FinBERT 모델을 사용하여 감성분석을 진행한다. 먼저 뉴스 기사 본문을 VADER와 동일한 전처리과정을 진행한 후에 한 문장씩 FinBERT의 입력으로 한다. 각 문장은 긍정, 부정. 중립으로 분류되며 문장의 분류 결과가 부정일 시 -1 중립이면 0 긍정이면 +1의 가중치를 주어 전체 점수를 더하고 본문의 문장 수만큼 나누어 평균 감성 점수를 계산하여 뉴스 기사의 감성 점수를 계산한다.

2.3 단어 임베딩 : Word2Vec

Word2vec은 각 단어들이 텍스트 내에서 가지는 의미를 다차원의 벡터 값을 통해 수치적으로 표현하는 방법이다 [20]. Word2vec은 특정 임베딩 공간상에서 같은 맥락을 갖는 단어들이 가까운 거리를 가진다는 분포 가설 (Distributional Hypothesis) 에서 출발한다 [21, 22]. Word2vec은 주어진 문장을 구성하는 단어들의 전후 관계를 학습하여 단어의 의미를 내포하고 있는 벡터 값을 통해 자질들을 수치화한다. 이는 기존의 통계적인 방식과는 다르게 별도의 유사도 계산이나 차원 축소 과정 없이 변별적인 특징을 내포하고 있는 벡터 값으로 단어를 수치화 할 수 있다. 따라서 주어진 문장에 대한 문법적 해석이 가능하며, 단어의 거리를 통해 의미론적 추론도 가능해진다.

특히 감성분석에서는 특정 단어들의 조합이 감성을 결정하는데 영향을 주기 때문에 단어 간의 관계를 파악하는 것은 중요하다. 어떤 단어가 함께 등장하고, 문장의 맥락이 어떠한 지에 따라 문장의 감성이 달라지기 때문에 동일한 문장이라도 표현하는 감성의 값이 달라질 수 있다. 예를 들면 작품 감상에 있어 '졸리다'는 부정적인 의미이지만 침대 사용 후기에 있어 '졸리다'는 긍정적인 의미이다.

본 논문에서 감성분석을 위해 사용하는 VADER는 글의

표 2. 주식 데이터 수집 목록
Table 2. Stock Data Index

	date	open	adj_close	close	high	low
0	2018-03-27	94.940	89.47	89.47	95.139	88.51
1	2018-03-26	90.610	93.78	93.78	94.000	90.40
2	2018-03-23	89.500	87.18	87.18	90.460	87.08
3	2018-03-22	91.265	89.79	89.79	91.750	89.66
4	2018-03-21	92.930	92.48	92.48	94.050	92.21

맥락을 고려하지 않고, FinBERT는 맥락을 고려하지만 손실되는 정보가 생긴다. 따라서 뉴스 본문에서 사용되는 모든 단어들을 Word2Vec을 통해 임베딩 벡터로 변환함으로써 맥락에 대한 정보를 보완한다.

3. LSTM 기반 주가 예측

거시경제 특성을 반영한 앙상블된 감성분석을 통해 생성된 시계열 데이터에서 의미 있는 패턴을 추출하여 주가 예측을 하기 위해 LSTM을 사용한다. LSTM (Long Short Term Memory)은 순환신경망의 일종으로, 은닉층의 과거의 정보가 마지막까지 전달되지 못하는 장기의존성 문제를 해결하기 위해 설계된 신경망이다 [5]. LSTM은 그림 2와 같이 새로운 구조의 메모리 셀을 가지고 있다. 메모리 셀에서는 최근의 정보를 뜻하는 단기 상태를 과거의 정보를 뜻하고 장기 상태라고 불리는 곳에 세 가지 주요 요소를 거쳐 정보를 더하거나 제거하며 일련의 정보들을 다음 메모리 셀로 전달한다. 먼저, 포겟 게이트라고 불리는 시그모이드 모듈을 통해 셀 스테이트에서 어떤 정보를 버릴지 결정한다. 그 후 새로운 정보가 메모리 셀에 저장될지를 결정하기 위해 시그모이드 모듈과 tanh 모듈로 구성되어 있는 인풋 게이트를 통해 어떤 값을 업데이트할지 결정하고, 마지막으로 아웃풋 게이트라고 불리는 tanh모듈로 전달되어 출력값을 결정하게 된다.

LSTM은 시계열 예측에 있어 우수한 성능을 입증하고 있다. LSTM이 무작위 포레스트, 심층신경망 및 로지스틱 회귀 분류기와 같은 상태기억이 없는 분류 방법보다 시계열 분석에 더 적합하다는 것을 증명하였다 [23]. 또한 강수량 시계열 예측 문제에 대해 LSTM을 사용하여 기존의 다른 강수량 예측 모델보다 우수한 성능을 보여주었다 [24]. 물론 관련연구에서 소개한 것과 같이 주가 예측에서도 우수한 모델인 것이 여러 실험에서 입증되었기에 본 논문에서도 LSTM을 활용한다 [8-10, 13-15].

본 논문에서는 LSTM의 셀은 200개로 구성하였고, 메모리셀의 입력 시퀀스의 수를 뜻하는 timestep를 2로 하였다. 그 후 dropout과 종가를 예측하기 위해 완전연결계층을 하나 추가하여 최종 출력을 하나의 값이 되도록 전체 딥러닝 모델을 구성하였다. 최종적으로 예측된 주가와 실제 주가의 mean squeared error를 최소화하는 방향으로 최적화를 진행하며 학습하게 된다.

Ⅳ. 실험 결과

본 연구에서는 주가 데이터, 뉴스 데이터, 거시적 경제 지표 데이터를 사용한다. 먼저 주가 데이터는 캐나다 데이터 공유 회사 Quandle의 API를 이용해 표 2과 같이 마이크로소프트사의 10년 (2010.01.01. - 2019.12.31.)중 공휴일과 주말을 제외한 총 2010일 간의 날짜 (date), 시가 (open), 수정종가 (adj_close), 종가 (close), 고가 (high), 저가 (low)데이터를 수집하였다. 수정 종가는 주식 분할, 배당 및 권리 제공 등 기업 행동을 고려한 뒤 해당 주식의 가치를 반영하기 위해 종가를 수정한 것으로 과거 수익률을 조사하거나 과거 실적을 상세하게 분석할 때 자주 사용된다.

뉴스 데이터는 캐나다 비즈니스 신문사 Financial Post의 마이크로 소프트가 키워드인 기사 중 조회수가 가장 높은 기사 하나를 BeautifulSoup 라이브러리를 이용해 크롤링을 진행하여 뉴스 날짜와 기사 본문을 수집하였다. 주가 데이터와 동일한 기간의 뉴스기사만을 수집하기 위해 주가 데이터의 날짜 라벨을 기준으로 데이터 전처리 과정을 진행하여 총 2010개의 기사를 수집하였다.

거시적 경제 지표는 Yahoo Finance의 Yahoofinancials Module을 이용해 수집하였다. 국제 금값과 환율은 미국 경제와 관련이 있는 일본 (JPN/USD), 중국 (CNY/USD), 유럽 (EUR/USD)의 화폐 교환비를 각각 수집하였고 마찬가지로 주가 데이터 수집 날짜를 기준으로 데이터를 수정하였다.

종합적으로 주가 데이터는 마이크로소프트사의 수정 종가를 사용하였고, 금값과 일본 (JPN/USD), 중국 (CNY/USD), 유럽 (EUR/USD) 환율비 3개를 하나의 거시 경제지표 데이터로 칭하였다. 그리고 뉴스 기사 본문을 앞서 소개한 3가지 감성분석 모듈을 통해 앙상블된 감성 분석 결과를 사용하였다. 이와 같은 데이터를 전체 2010일 중 70%인 1407일을 학습데이터로 사용하였고, 나머지 30%인 603일의 수정종가를 평가 데이터로 하였다.

실험 환경은 본 연구를 위해 구동한 컴퓨터 서버의 하드웨어 환경으로 CPU intel i5, RAM 12GB, SSD 256GB이고 사용된 딥러닝 라이브러리 및 소프트웨어는 Tensorflow의 고수준 버전 Keras와 CUDA 10.0 등이 사용되었다. 총 300 epoch동안 학습이 진행되었고, 0.001의 학습율과 배치 크기는 512를 사용하였다.

주가 예측에 미치는 영향을 거시경제지표와 각각의 감성분석 모델들이 평가하기 위해 LSTM기반 주가 예측기의 입력 변수 여러 데이터 조합들 중 최적의 조합을 찾기 위해 성능 평가 방법으로 평균 제곱근 오차 (RMSE)를 측정하였다. 이는 오류값을 실제 값과 유사한 단위로 다시 변환하기에 해석이 다소 용이해진다.

표 3. 거시경제지표와 감성분석에 따른 마이크로소프트사의 주가 예측결과의 오류율
Table 3. Microsoft's stock price error rate for macro economic indicator and sentiment analysis

	거시 경제지표 (M)	VADER (V)	FinBERT (B)	Word2Vec (E)	RMSE (%)
Baseline					0.0226
SPP-M	√				0.1474
SPP-V		√			0.0107
SPP-B			√		0.0103
SPP-E				√	0.0102
SPP-MV	√	√			0.0109
SPP-MB	√		√		0.0108
SPP-ME	√			√	0.0101
SPP-VB		√	√		0.0103
SPP-VE		√		√	0.0107
SPP-BE			√	√	0.0106
SPP-MVB	√	√	√		0.0097
SPP-MVE	√	√		√	0.0106
SPP-VBE		√	√	√	0.0099
SPP-MVBE	√	√	√	√	**0.0095**

표 3은 거시 경제지표와 각각의 감성분석에 따른 주가예측 오류율을 나타낸다. 거시 경제지표는 금값과 3가지의 환율비를 뜻하며 VADER, FinBERT, Word2Vec 3가지의 감성분석 결과값을 다양하게 조합하여 실험을 진행하였다.

Baseline은 입력데이터로 수정종가만을 사용하였을 때의 오류율이다. 그리고 거시 경제지표만을 단독으로 사용하였을 때보다 감성분석 결과 값을 조합하여 사용하였을 때 훨씬 더 작은 오류율을 관측할 수 있었다. 이를 통해 감성분석이 주가 예측에 있어 거시경제보다 더 중요한 지표라는 것을 확인하였다. 거시 경제지표와 VADAR와 FinBert 2가지의 감성분석 모두를 사용하였을 때 오류율이 0.0097로 처음으로 0.01 아래의 오류율을 달성하였고, Word2Vec까지 추가하여 보완하면 Baseline의 오류율에서 약 58% 감소한 것을 확인하였다.

본 논문에서는 실험을 통해 거시경제지표와 3가지 감성분석결과를 앙상블화한 모델을 사용하였을 때 최적의 주가예측이 가능한 것을 확인하였다. 그림 3은 최적의 모델로 제안된 SPP-MVBE모델의 주가 예측 그래프이다. 초록색 선은 학습데이터 1407일을 뜻하며 파란색 선은 평가데이터 603일을 뜻한다. 빨간색 선은 추론결과를 나타내는데 2018년 이후 2번의 큰 하락폭을 가졌을 때를 제외하고는 매우 유사하게 주가를 예측하는 것을 확인할 수 있다.

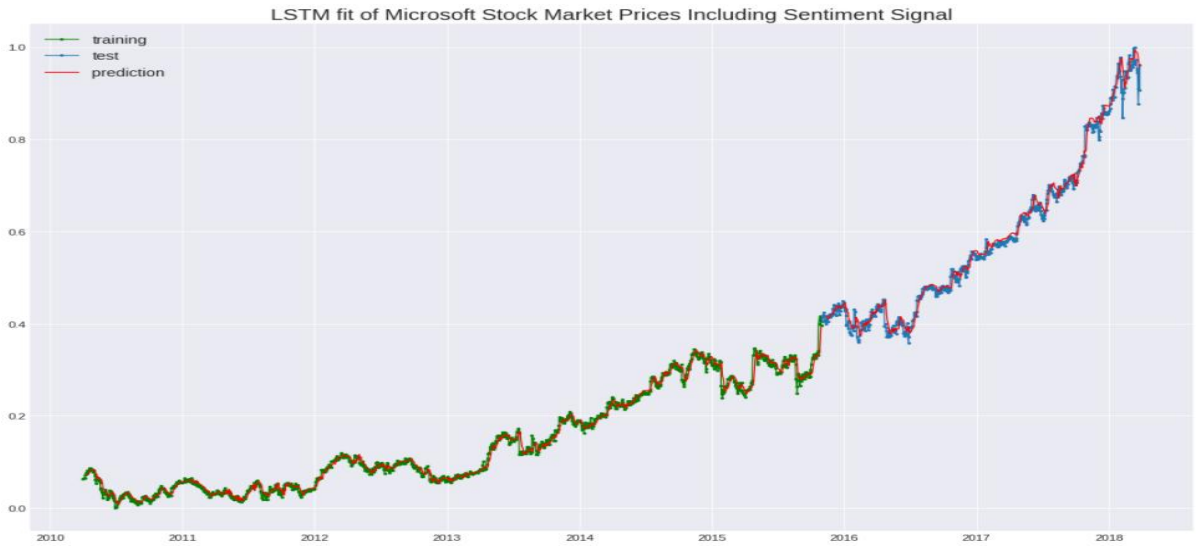

그림 3. SPP-MVBE 예측 그래프
Fig. 3. Prediction Graph Using SPP-MVBE

V. 결 론

본 논문에서는 기존에 매우 도전적인 문제로 제기되었던 주가예측을 위해 거시경제지표와 뉴스기사의 본문을 이용한 최적의 주가예측 모델인 SPP-MVBE을 제안하였다. SPP-MVBE 모델은 거시경제 지표 (금, 일본 (JPN/USD), 중국 (CNY/USD), 유럽 (EUR/USD)의 화폐 교환비)와 투자자의 심리에 영향을 미치는 뉴스기사 본문의 감성분석 (VADER, FinBert, Word2Vec)의 결과값을 앙상블하여 LSTM 모델에 입력으로하여 주가를 예측한다. 각각의 특징들을 다양하게 조합함으로써 실험을 진행하여 각 요소의 영향력을 오류율로 평가하였으며, 최종적으로 모든 특징들을 LSTM 모델의 입력으로 하는 SPP-MVBE 모델에서 수정종가만 사용하였을 때의 에러율 보다 약 58% 낮은 에러율을 관측할 수 있어 해당 모델의 우수성을 입증하였다. 이번 연구에선 거시경제 지표와 뉴스기사의 감성분석 결과만을 사용하였지만 추후 연구에선 주가의 다양한 기술적지표인 거래량, 모멘텀 지표, 추세 지표들의 주가에 미치는 영향을 추가적으로 연구할 예정이다.

References

[1] E. F. FAMA, "Efficient Capital Markets a Review of Theory and Empirical Work," The Fama Portfolio, pp.76-121, 2021.

[2] M. D. Godfrey, C. W. J. Granger, O. Morgenstern, "The Random Walk Hypothesis of Stock Market Behavior," Kyklos, Vol. 17, No. 1, pp. 1-30, 1964.

[3] T. Gao, Y. Chai, Y. Liu, "Applying Long Short Term Momory Neural Networks for Predicting Stock Closing Price," 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS). pp. 575-578, 2017.

[4] J. G. De Gooijer, R. J. Hyndman, "25 years of IIF time Series Forecasting: a Selective Review," International Journal of Forecasting Vol. 22, No. 3, pp. 443 - 473, 2006.

[5] S. Hochreiter, J. Schmidhuber, "Long Short-Term Memory," in Neural Computation, Vol. 9, No. 8, pp. 1735-1780, 1997.

[6] K. Cho, B. Van Merriënboer, D. Bahdanau, "On the Properties of Neural Machine Translation: Encoder-decoder Approaches," arXiv preprint arXiv:1409.1259., 2014.

[7] Y. S. Kim, J. S. Cha, "Analysis of the Relationship Between Stock Prices and Macroeconomic Variables," Bank Of Korea, pp. 99-12 ,1999 (in Korean).

[8] P. S. Cho, "The Stock Forecasting Model using Bidirectional LSTM and Reinforcement Learning," PhD Thesis. Hanyang University. 2021 (in Korean).

[9] D. H. Shin, K. H. Choi, C. B. Kim, "Deep Learning Model for Prediction Rate Improvement of Stock Price Using RNN and LSTM," The Journal of Korean Institute of Information Technology, Vol. 15, No. 10, pp. 9-16, 2017 (in Korean).

[10] S. H. Hong, "A Research on Stock Price Prediction Based on Deep Learning and Economic Indicators," Journal of Digital Convergence Vol. 18. No. 11, pp. 267-272, 2020 (in Korean).

[11] W. Antweiler, M. Z. Frank, "Is all that talk just noise?," The Information Content of Internet Stock Message Board," J Finance Vol. 59, No. 3, pp. 1259 - 1294, 2004.

[12] M. Baker, J. Wurgler, "Investor Sentiment and the Cross Section of Stock Returns," The Journal of Finance, Vol. 61, No. 4, pp. 1645-1680, 2006.

[13] M. J. Kim, J. H. Ryu, D. H. Cha, M. K. Sim, "Stock Price Prediction Using Sentiment Analysis: from "Stock Discussion Room" in Naver," The Jounal of Society for e-Business Studies ,Vol. 25, No. 4, pp. 61-75, 2020 (in Korean).

[14] J. G. Koh, G. Y. Lee, I. J. Son, Y. R. Gwon, "Stock Price Index Prediction Program Using Deep Learning Techniques," Proceedings of the Korean Society of Computer Information Conference, pp. 525-526, 2021 (in Korean).

[15] S. H. Hong, "A Study on Stock Price Prediction System Based on Text Mining Method Using LSTM and Stock Market news," Journal of Digital Convergence, Vol. 18, No. 7, pp. 223-228, 2020 (in Korean).

[16] E. Alickovic, A.. Subasi, "Normalized Neural Networks for Breast Cancer Classification," In: International Conference on Medical and Biological Engineering. Springer, Cham, pp. 519-524, 2019.

[17] C. Hutto, E. Gilbert, "Vader: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," Proceedings of the International AAAI Conference on Web and Social Media. Vol. 8. No. 1, pp. 216 - 225, Jan. 2015.

[18] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2018.

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

[20] T. Mikolov, K. Chen, G. Corrado, J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv preprint arXiv:1301.3781., 2013.

[21] I. S. Kang, "A Comparative Study on Using SentiWordNet for English Twitter Sentiment Analysis,"

Journal of Korean Institute of Intelligent Systems, Vol. 23, No. 4, pp. 317-324, 2013 (in Korean).
[22] T. Mikolov, K. Chen, G. Corrado, J. Dean, “Efficient Estimation of Word Representations in Vector Space,” arXiv preprint arXiv:1301.3781, 2013 (in Korean).
[23] T, Fischer, C. Krauss, "Deep Learning with Long Short-term Memory Networks for Financial Market Predictions," European Journal of Operational Research, Vol. 270, No. 2, pp. 654-669, 2018.
[24] X. Shi, Z. Chen, H. Wang, D. Y. Yeung, “Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting,“ In: Advances in Neural Information Processing Systems, pp. 802 - 810, 2015.

Ha-Eun Kim (김 하 은)

2018~Department of Artificial Intelligence from Daegu University (B.S.)

Career: 2018~B.S. Student, Daegu University
Field of Interests: Artificial intelligence
Email: kheun1411@naver.com

Young-Wook Park (박 영 욱)

2016~Department of Artificial Intelligence from Daegu University (B.S.)

Career: 2016~B.S. Student, Daegu University
Field of Interests: Artificial intelligence
Email: duddnr1612@gmail.com

Si-Eun Yoo (유 시 은)

2018~Department of Artificial Intelligence from Daegu University (B.S.)

Career: 2018 ~ B.S. Student, Daegu University
Field of Interests: Artificial intelligence
Email: ykw8077@naver.com

Seong-Woo Jeong (정 성 우)

2020 Artificial Intelligence from Daegu University (B.S.)
2021~IT Convergence Engineering from Daegu University (M.S.)

Career: 2021 Master Student, Daegu University
Field of Interests: Computer vision, Deep learning, Federated Learning, Knowledge Distillation
Email: learningsteady0j0@gmail.com

Joonhyuk Yoo (유 준 혁)

1993 Electrical and Electronic Engineering from POSTECH (B.S.)
1995 Electrical and Electronic Engineering from POSTECH (M.S.)
2007 Computer Engineering from University of Maryland at College Park, MD USA (Ph.D.)
2009~Department of Artificial Intelligence at Daegu University (Professor)

Career:
1995~2000 Embedded System Engineer, Samsung Co. Ltd.
2005~2007 Embedded Software Engineer, NASA CREAM
2008~2009 Research Professor, Korea University

Field of Interests: Machine Learning, Computer Vision, On-Device AI, Cyber-Physical Systems
Email: joonhyuk@daegu.ac.kr