사이드 프로젝트 논문

양방향 LSTM 기반 주가 예측 알고리즘 설계 및 성능 분석

Bidirectional LSTM-based predictive design and analysis performance

김민수

2023. 11. 12

목차

Ⅰ. 서론

A. 금융 AI 시장 전망

금융 시장이 발전함에 따라 우리가 흔히 알고 있는 국내 종목 외에도 다양한 파생 금융 상품들이 거래되고 있고 최근에는 가상화폐가 등장하면서 금융 시장이 열풍을 이끌었다. 이토록 금융 시장이 급격하게 발전하고 거대해짐에 따라 금융 자산 투자에 대한 많은 신규 고객들이 유입되고 있고 투자에 대한 관심도가 높아지고 있다. 다수의 신규 투자자들의 급증은 변동성 증가로 이어질 수 있으며 특히 실전 투자 경험이 부족하거나 기본 펀더멘털을 고려하지 않고 단순히 투기적으로 혹은 군집 행동적인 요소로 투자하는 신규 투자자들의 경우 수습하지 못할 정도의 금융자산의 손실까지 초래할 수 있다. 하지만 우리는 투자를 하지 않을 수 없는 상황에 놓였다. 소비자 물가지수는 지난 20년간 꾸준하게 상승하고 있으며, 최근 2020년도부터는 훨씬 급격하게 상승하는 추세를 보이는 가운데 한국 금융시장에서 정기예금이율이 2020년도 0.80%로 최저치를 찍었고, 최근 10년간 3.00%를 넘는 상품은 거의 사라져 찾아볼 수 없게 되었다. 계속되는 물가 상승과 금리 인하로 저금리 시대에 목돈 마련과 노후 대비를 위해서 금융 투자가 필수가 된 상황이다. 실제로 주식 시장은 경제, 기업의 실적 등 외부 요인들에 영향을 받는 것은 물론이고 주가는 랜덤워크 속성이 있기 때문에 수익을 내는 전문 투자자들과 금융 애널리스트들도 매년 변화하는 주식 시장에서 항상 성공하는 것은 아니다. 그럼에도 주식 열풍과

함께 주식 투자는 고무적인 일임을 알 수 있다.

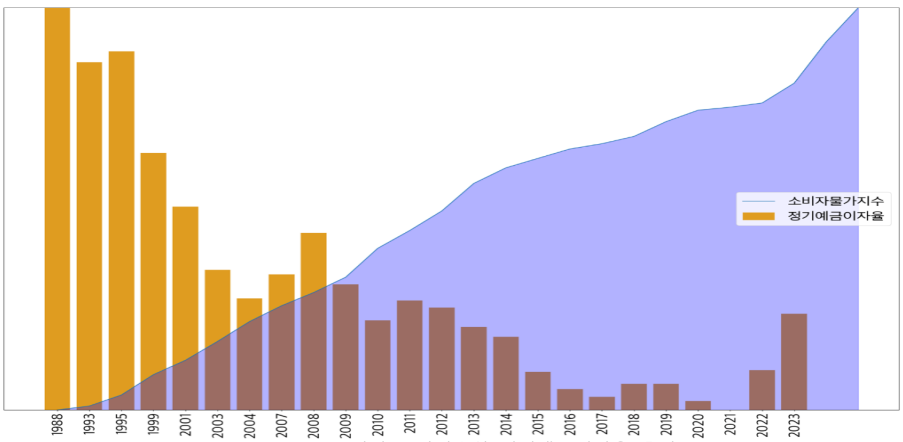


Figure 1.1 소비자 물가지수와 정기예금이자율 추이

그리고 금융 분야에서까지 다양한 인공지능들이 도입이 가속화되고 있으며 국내 금융분야 인공지능 시장규모는 2019년 3천억 원에서 2021년에 6천억 원으로 45.8% 증가하였으며, 이후 2026년까지 연평균 38.2% 성장하여, 3조 2천억 원의 시장을 형성할 것으로 전망된다. 그리고 금융 분야의 발전과 더불어 주식 거래에서 까지 인공지능 도입 사례들이 등장하면서 관심이 고조되고 있다. 2016년 3월 구글 딥마인드사의 바둑 인공지능 프로그램인 알파고가 이세돌 9단과의 바둑 매치 이후 최근에는 인간과 인공지능의 실전 주식 투자에 대한 주제로 많은 대회들이 관심을 끌고 있으며, 인공지능과 자문 전문가의 합성인 로보어드바이저(robo-adviser)에 대한 관심도 높아지고 있다. 하나은행에 따르면 2025년 국내 로보어드바

이저 시장 규모가 30조원으로 성장할 것으로 전망이 나왔다.

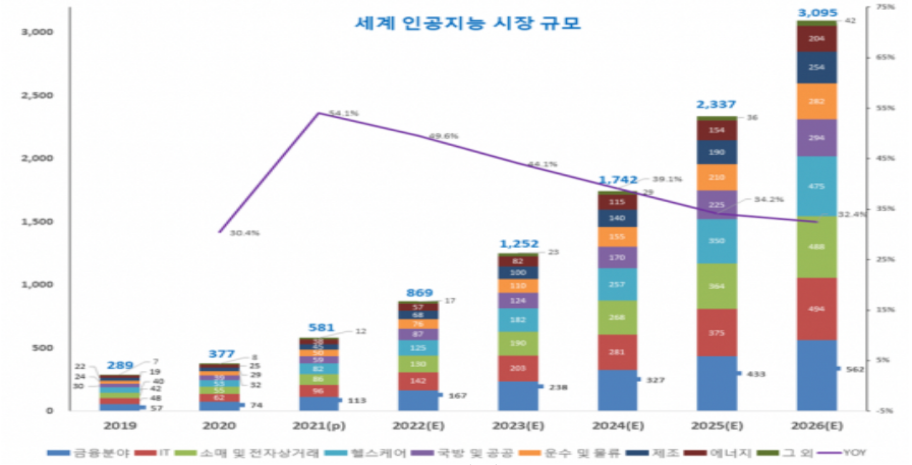


Figure 1.2 자료: "Artificial Intelligence(AI) Market - Global Forecast To 2026", MarketandMarkets(2021)

최근 2016년도 3월 국내에서도 '금융자문업 활성화 방안'을 수립하고 로보 어드바이저에 대해서도 자문 운용 허용 방침을 밝힘에 따라 투자 자문업 진입 장벽이 완화되고 금융서비스의 디지털 혁신 가속화를 전망하는 가운데 투자자들의 대거 유입과 딥러닝 기술이 연계된 연구들이 많이 이뤄지고 있지만 아직까지 불확실한 주식시장을 예측하는 것은 쉽지 않은 일이다. 본 연구는 딥러닝 모델을 활용한 알고리즘을 구현하고 실제 주가 예측에 적용하여 실제 주가와 딥러닝 모델이 예측한 주가의 차이를 비교 분석 하고 추후 개선된 딥러닝 모델을 고안하기 위함이다. 기준 날짜 국내 코스피 상위 종목 10개에 대해서 분석하고자 한다. 분석 기간은 2022

년 11월 10일부터 2023년 11월 10일 1년을 기간으로 한다.

Ⅱ. 본 론

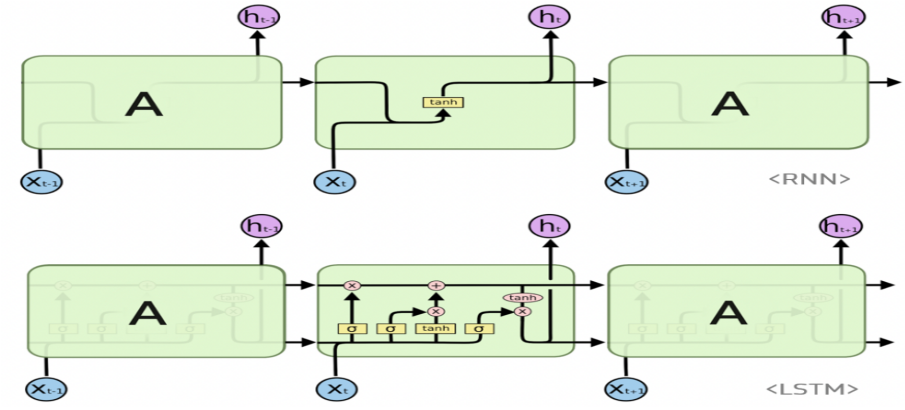
A. 제안 알고리즘의 개요

1. LSTM

본 논문에서는 주가 예측에서 최종적으로 사용될 Bidirectional LSTM 모델의 개념을 이해하고 구현을 위해서 먼저 베이스라인이 되는 LSTM 모델에 대한 기본 개념을 이해하고자 한다. Figure 2.1과 같이 RNN(Recurrent Neural Network)은 관련 정보와 그 정보를 사용하는 지점 사이 거리가 멀 경우 역전파 과정에서 그래디언트가 점차 줄어들어 학습 능력이 크게 저하되는 그래디언트 소실(Vanishing Gradient Problem) 문제가 발생한다. 이 문제를 극복하기 위해서 RNN의 은닉층(Hidden Layer)에 일종의 컨베이어 벨트 역할을 하는 셀 상태(Cell State)를 추가하여 고안된 것이 바로 LSTM(Long Short Time Memory)이다.

Figure 2.1 <RNN>과 <LSTM> 구조

Cell State는 은닉 상태(hidden State)와 마찬가지로 이전 시점의 Cell State를 다음 시점으로 넘겨준다. Cell State의 역할은 다른 게이트(Gate)들과 함께 작용하고 정보를 선택적으로 활용할 수 있도록 한다. LSTM 블록은 망각 게이트(Forget Gate), 입력 게이트(Input Gate), 출력 게이트(Output Gate) 총 3개의 게이트가 더한 Cell State로 구성되어 있다. LSTM 셀의 수식은 다음과 같습니다.



$$ft = \sigma(Wxh\_fxt + Whh\_fht-1 + bh\_f)$$

$$it = \sigma(Wxh\_ixt + Whh\_iht-1 + bh\_i)$$

$$ot = \sigma(Wxh\_oxt + Whh\_oht-1 + bh\_o)$$

$$gt = \tanh(Wxh\_gxt + Whh\_ght-1 + bh\_g)$$

$$ct = ft \odot ct-1 + it \odot gt \quad ht = ot \odot \tanh(ct)$$

ft는 과거의 정보를 잊기 위한 게이트로 망각 게이트(Forget Gate)라고 부르고, $ht-1$과 xt를 받아 시그모이드를 취해준 값이 바로 Forget Gate가 출력하여 내보내는 값이 된다. 시그모이드 함수의 출력 범위는 0과 1 사이이므로 값이 0이면 이전 상태의 정보는 잊고 값이 1이면 이전 상태의 정보를 기억하게 된다. 반면에 $it \odot gt$는 현재 정보를 기억하기 위한 게이트로 입력 게이트(Input Gate)라고 부르고, xt와 $ht-1$를 입력으로 받아 시그모이드를 취하고, 같은 입력으로 tanh를 취한 다음 ⊙(hadmard product) 연산을 한 값이 바로 Input Gate가 출력하여 내보내는 값이 된다. 앞서 과거 정보를 잊기 위해서 ft, 현재 정보를 기억하기 위해서 $it \odot gt$를 구하고 $ft \odot ct-1$ 계산을 하여 이전 시점의 Cell State 정보를 얼마나

유지할지 구했다. 최종적으로 이 과정들을 모두 더하여 현재 시점의 Cell State를 업데이트(Update)하고 tanh 함수를 취해준 뒤 출력으로 사용한다. 그러나 이 값을 그대로 출력으로 사용하지 않고 출력이 얼마나 중요한지 조절하기 위해 은닉 상태(Hidden State)와 현재 입력에 대해 시그모이드 함수를 취하여 0과 1사이의 값으로 만든 후 출력하고자 하는 신호와 곱해서 크기를 조절 후 출력을 한다. 이 부분이 출력 게이트(Ouput Gate)에 해당하는 것이다.

2. Bidirectional LSTM

일반적인 LSTM은 순방향 즉 왼쪽에서 오른쪽으로만 정보를 추출하기 때문에 시퀀스 데이터가 시간 순으로 입력되며 결과가 직전 패턴에 수렴하는 한계가 존재한다. 그래서 Figure 2.2와 같이 역방향으로도 정보를 추출하면 더 많은 정보를 추출할 수 있고 수렴 문제도 해결할 수 있다는 점에서 제안된 모델이 우리가 흔히 부르는 양방향 LSTM(Bidrectional LSTM) 이다. 양방향 LSTM은 기존의 순방향 LSTM 모델에 역방향으로 전달하는 은닉층(Hidden Layer)을 추가하여 Forward, Backward 2개 역할을 할 수 있는 레이어(Layer)를 만들고 Forward Layer에는 입력이 순방향 순서로, Backward Layer에는 역방향 순서로 들어가며, 최종 은닉 상태에서는 두개의 LSTM 계층의 은닉 상태를 Concat하여 벡터를 출력한다. 벡터에는 포함하고 있는 정보가 점점 많아지고 대응하는 정보의 주변 정보들을 균형 있게 담게 된다.

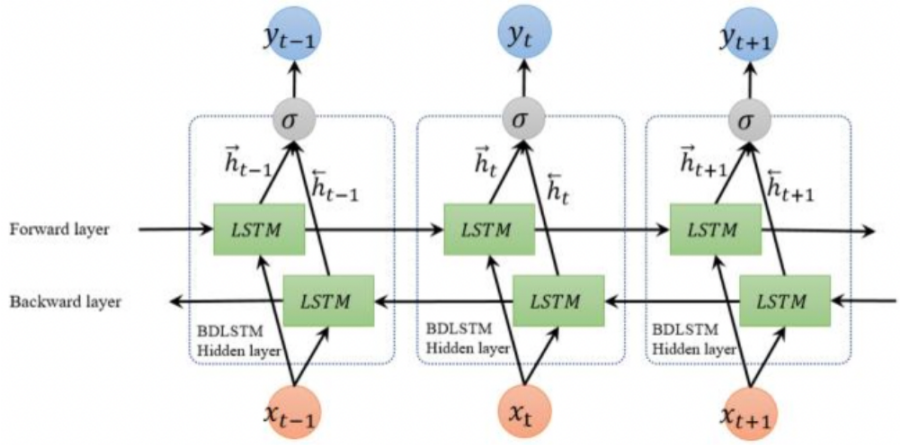


Figure 2.2 <Bidirectional LSTM> 구조

보통 Bidrectional LSTM은 양방향으로 정보를 처리할 수 있어서 앞뒤 문맥 정보를 필요로 하는 자연어 처리 분야나 음성 인식, 언어 번역에서도 흔히 사용되고 있다. 본 연구에서는 Bidrectional LSTM이 각 시점에서 은닉 상태(Hidden State)가 이전 시점과 미래 시점의 정보를 모두 갖는다는 이점을 시계열 분야에서도 적용할 수 있다고 보았고, 이를 시계열의 한 분야인 주가 예측에서 주가의 과거 시점과 미래 시점의 정보가 향후 주가를 예측에 있어 더 많은 영향을 줄 것으로 기대한다. 실제로 기존의 RNN, LSTM모델 보다 더욱 우수한 예측 성능을 보였다.

A. 데이터 분석 및 구축

1. 크롤링 알고리즘

Figure 3.1 크롤링 진행 화면

주가 예측을 하기 위해서는 가장 먼저 종목 관련 정보들을 불러와야 한다. 데이터 수집 알고리즘 구성은 다음과 같다. 크롤링(Crawling) 기술을 사용하여 네이버 증권의 종목 데이터와, 종목 관련 네이버 뉴스 데이터들을 수집하는 알고리즘을 각각 구성한다. WebDriver를 이용해서 웹브라우저를 띄우고 각 크롤링 작업을 자동화한다. WebDriver는 웹페이지를 자동으로 탐색하고 원하는 정보를 수집할 수 있는 툴이다. 증권 종목 데이터의 경우 Figure 3.1 화면과 같이 체크 박스의 초기화 값을 해지하고 '거래량', 'ROE', 'PER', 'PBR', '외국인비율'에 해당되는

체크 박스를 선택하게 한다. 그리고 체크 값에 일치하는 코스피 상위 종목 10개에 대해서 진행한다. KRX(한국거래소) 사이트 URL에 Request 모듈로 요청을 보낸 후 종목의 OHLCV (Open, High, Low, Close, Volumn)와 시가총액등 통계 정보를 수집하는 과정을 자동화한다. 종목 뉴스 데이터의 경우 네이버 플랫폼 검색창에 종목 코드를 입력하였을때 나열되는 종목 관련 뉴스들을 기준 날짜별로 최대 5개의 뉴스만 수집하도록 한다. 마찬가지로 웹 브라우저를 띄우고 진행하며 종목 코드와 데이터 수집기간이 입력된 네이버 플랫폼 검색창 URL을 찾고 Request 모듈로 요청을 보낸 후 정렬된 뉴스들을 불러오고 뉴스 제목을 가져오는 과정을 자동화한다. 본 연구 데이터 수집기간은 2022년 11월 10일부터 2023년 11월 10일 1년을 기간으로 한다.

2. 분석 기술 적용

가. 기술적 분석

실제로 대부분의 전문 투자가들은 종목에 대해서 투자 의사 결정을 내릴 때 일반적으로 기술적 분석과 펀더멘털 분석 두가지 분석 유형을 참고한다. 기술적 분석이란 오직 주식의 가격 변동과 거래량 등 시장 활동에서 생성된 통계를 분석하여 시장 동향을 예측하는 방법이다. 기술적 분석의 목표는 과거 시장 데이터에서 향후 가격 변동을 예측하기 위해 가격의 움직임 패턴과 추세를 파악하고 매매 신호를 감지하는 것이다. 과거 시장 데이터에서 패턴과 신호를 식별하기 위해 차트, 보조지표, 오실레이터

(Oscillators) 등 다양한 도구와 기법을 사용한다. 또한 기술적 분석은 주관적인 특성을 갖고 있으며 각 분석 지표 마다 각각 다른 해석을 가져올 수 있고 최대한 많은 분석 지표를 활용한다면 더 많은 해석들을 저장할 수 있을 것이고 최고의 결과를 반영할 수 있을 것이다. 하지만 사람이 수많은 분석 지표들을 머리에 담고 해석까지 하는 것은 한계가 있다. 따라서 본 연구에서는 최대한 많은 분석 지표들을 피쳐로 생성한 후 학습 데이터에 포함시킬 것이다. Ta-Lib 라이브러리를 통해서 분석 지표들을 생성할 수 있다. 분석 지표들은 크게 주기 지표(Cycle Indicators), 모멘텀 지표(Momentum Indicators) 등 8가지 종류가 있고, 학습 데이터를 shape 했을때 총 154개가 생성된 것을 확인하였다. 또한 Rolling과 Lagging 등 추가적으로 19개의 피처를 생성하였다. 그리고 "수정주가"에 대해서도 고려해 봐야할 것이다. 기업의 배당, 무상증자, 무상감자, 액면분할 등이 있는 경우 주가가 순간의 연속성을 잃고 단층으로 인해 변질된다. 예시로 기업이 배당을 배분하게 되면 배당락이 발생하여도 수정주가 적용이 되지 않은 종가는 배당락으로 인해 떨어지는 주가를 그대로 두는 반면 수정주가 적용이 된 수정종가의 경우 배당락이 발생되지 않은 것처럼 수정된다. 따라서 수정주가(Adjusted price)를 적용하여 왜곡된 주가를 보정하고 분석 해야한다. 정확한 수정주가적용을 하려면 배당금(Dividend), 분할비율(Split Ratio)등의 정보가 필요하나 본 연구에서는 수정주가적용이 되지 않은 데이터들만이 존재하고 필요한 정보가 제공되지 않았으므로 비슷한 결과를 가져오기 위해 알고리즘을 구성하였습니다.

본 연구에서는 역산한 기준가와 전일의 종가 사이에 값의 차이가 발생하는 경우 알 수 없는 외부 영향이 미칠 것으로 생각할 수 있고 역산한 기준가와 전일 종가 사이의 비율을 통해 수정주가 비율을 계산하여 종가(Close)외 나머지 시가(Open), 저가(Low), 고가(High), 거래량(Volume)에도 적용한다.

나. 감성 분석

주가는 가격 및 거래량과 같은 과거 주가의 움직임 외에도 어떤 외부 요인에 따라 크게 달라질 수 있다. 특히 주식의 가치는 기업활동과 직접적, 간접적으로 연관이 있다. 기업이나 경제 상황을 분석할때 펀더멘탈(Fundamental)이라는 방식과 이에 대비되는 용어로 센티멘탈(Sentimental)이라는 방식이 있다. 펀더멘탈은 기업가치평가와 재무제표 같은 객관적인 사실과 이성적인 것이라면 센티멘탈은 외부 요인으로 인한 투자자들의 직관적, 감정적 분위기에 주목한다. 특히 기업에 대한 긍정적인 전망에 관한 기사가 실리면 보통 그 기업의 주가가 상승하곤 한다. 그렇듯 주가의 방향은 과거 주가의 행보 외에도 주식을 거래하는 사람들의 인식에 영향을 받기 때문에 소셜미디어나 뉴스 언론과 같은 매체에 따라서 결정될 수 있다. 하지만 기업에 관한 뉴스 기사들은 비정형 데이터에 해당하기 때문에 텍스트 마이닝(Text-mining) 분석을 통해 시장의 기대치를 측정하거나 긍정/부정과 같은 감정을 수치로 정량하는 작업을 해야한다. 이 과정을 감성분석(Sentiment Analysis)이라고 하며, 본 연구는 실

제로 특정 종목 관련 뉴스 기사와 해당 주가의 변동성에 대해서 분석한 결과, 뉴스 기사에 긍정적인 신호 혹은 부정적인 신호가 포함되었을 경우 다음날의 주가가 상승세 혹은 하락세를 보이는 사실을 발견했고, 감성분석을 진행하여 감성점수를 피쳐 형태로 추출 후 기술적 지표와 결합할 것이다.

$$\text{단어의 감성점수} = \frac{\sum_{i}^{n} \text{뉴스에서 단어가 등장한 날의 종목 등락률}}{\text{모든 뉴스에서 단어가 등장한 총 횟수}}$$
$$(n : \text{전체 뉴스의 개수})$$

$$\text{뉴스의 감성점수} = \frac{\sum_{i}^{n} \text{단어의 감성점수}}{n}$$
$$(n : \text{뉴스에 등장한 단어의 개수})$$

$$\text{이날 뉴스의 감성점수} = \frac{\sum_{i}^{n} \text{뉴스의 감성점수}}{n}$$
$$(n : \text{이날 뉴스의 개수})$$

Figure 3.2 감성점수 계산 과정

감성분석을 하기 전 앞서 크롤링을 통해 얻어진 뉴스 제목들에는 특수기호, 공백 등 우리가 분석하려는 뉴스에서 중요하지 않은 문자들이 존재한다. 그리고 해당 종목에 관한 뉴스에서는 당연하게 그 종목의 이름이 자주 언급될 것이다. 이는 우리가 분석하려는 뉴스의 잘못된 감성점수를 구

하게 될 수 있다. 따라서 앞서 말한 과정을 처리하는 커스텀 토크나이저를 구현하고 감성분석을 하기 위한 좋은 조건으로 구성시킨다.

분석하기 좋은 조건이 완성되면 Figure 3.2 계산 과정을 통해 뉴스의 감성점수를 도출한다. 본 연구에서 사용하는 감성점수 공식의 컨셉은 다음과 같다. 먼저 단어별 감성사전을 만든다. 어떤 단어가 등장할때 종목의 등락률의 평균을 계산하고 각 단어의 감성점수를 구한다. 이는 각 뉴스에서 등장하는 단어가 평균적으로 종목이 상승세일때 자주 등장한다면 단어가 가지는 감성점수는 양수를 가질 것이고 하락장일때 자주 등장하는 단어는 음수의 감성점수를 가지게 될 것이다. 코스피 등락률은 { (오늘의 종가 - 어제의 종가) / 어제의 종가 * 100 } 계산식을 통해 구할 수 있다. 그리고 당연하게 각 뉴스에는 여러가지 단어들이 등장할 것이다. 앞서 구한 단어들의 감성점수를 모두 더하고 단어들의 등장 횟수를 나눠주어 각 뉴스의 감성점수를 구한다. 그리고 각 날짜마다 뉴스들의 감성점수를 모두 더하고 뉴스들의 수만큼 나눠주면 각 종목의 이날 감성점수를 구할 수 있다. 이렇게 얻어진 감성점수는 강한 상승세에서 자주 등장하는 단어가 뉴스에 많이 포함될 수록 뉴스의 감성점수는 더 큰 점수를 받을 것이다. 반면 단어가 약한 상승세, 강한 하락세일때 자주 등장한다면 감성점수는 더 낮은 마이너스 점수를 받을 것이다.

다. 기능 중요도

앞서 데이터를 포함하면 할수록 좋다고 가정하였다. 하지만 단지 이론적인 가정일 뿐이다. 실제 주가를 예측하는데 어떤 변수가 어떻게 영향을 끼치는지 얼마나 영향을 끼치는지 아는 것이 중요하다. 우리는 학습된 모형에 대하여 예측 관점에서 각 변수들의 영향력을 구할 수 있다. 기계학습 알고리즘이 풀 수 있도록 선형회귀 문제로 접근한다. 본 연구에서는 일반적으로 성능이 가장 좋으며 트리 기반 앙상블 학습에서 가장 각광받고 있는 XGBoost 알고리즘을 사용한다. XGBoost 알고리즘에는 변수 중요도(Feature Importance)를 자체적으로 사용할 수 있도록 구성되어있다. 변수 중요도는 예측 모형의 각 변수가 예측력에 기여한 정도를 측청하는 방법이다. XGBoost는 가장 대표적인 MDI(Mean decrease impurity) 변수 중요도를 사용하고 MDI는 각 변수가 split될 때 불순도(Impurity) 감소분의 평균을 중요도로 정의한다. 하지만 변수 중요도 측정 기준별로 중요 변수가 달라지기 때문에 일관성(Consistency)이 없는 한계가 있다. 그래서 문제들을 보완하기 위해 고안된 방법이 Shap Value 이다. 본 연구에서는 Shap Value를 사용할 것이다. Shape Value는 기계 학습 모델의 출력을 설명하기 위한 게임 이론적 접근 방식이고 Shape Value는 여러번 수행하며 반복적인 결과물의 평균을 도출하기 때문에 일관성(Consistency)이 잘 유지된다.

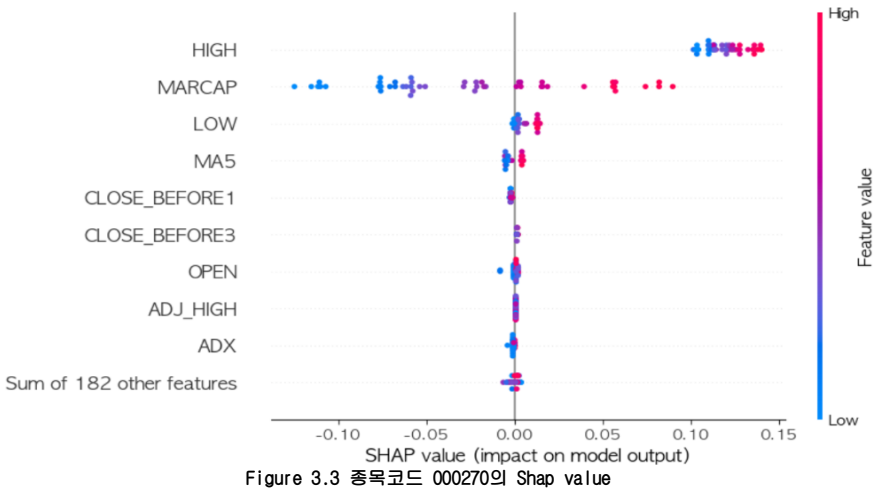


Figure 3.3 종목코드 000270의 Shap value

Figure 3.3 그림과 같이 모든 종목에 대해서 각 변수들의 중요도를 확인하기 위해서 중요도를 플롯(Plot)화 하였다. 종목들마다 각 변수들이 주가에 영향을 미치는 정도가 달랐으며, 전반적으로 시가총액(MARCAP) 변수는 거의 모든 종목에서 다른 변수들과 달리 비교적 큰 영향력을 미치는 것을 확인했다. 하지만 전체적으로 변수들의 영향력은 미미한 것으로 나타났다. 또한 영향력이 0인 변수들이 존재하였고, 이는 모델 학습에서 도움되지 않는 변수로 제외 한다. 주가의 특성상 데이터에 빈번한 이상치들로 인한 큰 변동성과 불균형적인 주가로 인해 패턴을 효과적으로 포착하지 못하는 것이 요인이다. 최근 코로나19 장기화, 길어지는 국가들 간에

전쟁 발발 및 긴장 고조 등 당면한 복합 위기가 불확실한 시장 혼란까지 가중되는 요인으로 판단된다.

라. 기능 공학

앞서 기능 중요도 기법을 통해 알고리즘에 기반하여 학습에 도움되는 기능들만 선택했다. 하지만 아직도 종목 마다 데이터가 가지고 있는 기능들이 너무 많다. 딥러닝 모델이 기능들을 모두 학습하기에는 자원 비용이 크게 들고 "차원의 저주(Curse of dimensionality)"가 발생할 수 있다. 차원의 저주란 공간의 차원이 증가함에 따라 데이터의 밀도가 급격히 감소하고 이로 인해 모델의 성능이 저하되는 것을 말한다. 이처럼 고차원 공간은 과적합의 위험을 증가시키고 고차원 공간의 훈련 모델은 계산 비용이 많이 들게 된다. 따라서 딥러닝 모델이 학습하기 위한 최적의 조건으로 구성해주기 위해서 스택형 오토인코더를 구축하고 주성분 분석도 진행한다. 먼저 VAE(Variational AutoEncoder) 알고리즘을 사용하여 상위 수준 기능을 추출한다. VAE 알고리즘 구조는 Figure 3.3과 같습니다. 요약하자면 알고리즘에는 인코더(Encoder)와 디코더(Decoder) 구조로 나뉘고 주어진 주식 시장 특성들을 잠재 공간으로 압축(인코딩)하여 잠재 공간에서 가져온 샘플로부터 임의의 노이즈를 더하고 새로운 특성들을 생성(디코딩)하는 작업을 통해 입력 데이터를 재구성하고 Reparameterize는 확률론적 요소를 도입함으로써 부드럽고 연속적인 잠재 공간을 학습한다. 손실함수로는 잠재 공간을 표준 정규 분포에 가깝게 정규화 하는 KL(Kullback-Leibler)발산과 디코더에서 재구성의 정도를 측청하는 로그

가능성 항 두가지 항을 조합하고 최종 레이어에서는 0과 1사이의 값을 출력하도록 시그모이드(Sigmoid) 활성화 함수를 사용한다.

학습 과정을 거쳐 데이터의 패턴과 관계를 캡쳐하고 재구성 오류를 비교하여 노이즈 처리, 이상 처리를 할 수 있고 잠재 공간에서 샘플을 생성하고 특징을 학습하는 과정에서 시나리오 생성과 점차적으로 주식 시장의 요소들이 의미있는 방향으로 구성되도록 한다. 또한 PCA(Principal Component Analysis) 기법을 앞서 Stacked VAE 알고리즘과 하이브리드(Hybrid)하여 각 방법의 고유한 장점과 한계를 보완한다. PCA는 주성분 분석이라고 불리고 직교 선형 변환으로 정의된다. 동일하게 높은 기능의 벡터를 찾는 것이 주 목표이고 고차원 데이터를 저차원 데이터로 바꿔주는 역할을 한다. PCA기법을 통해 많은 입력 특성으로 인한 모델 성능 저하 원인을 해결 하도록 한다. 두 기법은 모두 차원 축소 기술이지만 작동 방식이 다르다. VAE의 경우 복잡하고 계층적인 표현을 학습하는 심층 생성 모델인 반면, PCA는 데이터의 주요 구성 요소를 캡쳐하는 선형 기술이다. 이를 결합하여 잠재적으로 데이터의 선형 및 비선형 관계를 모두 캡쳐할 수 있다.

3. 모델 학습 및 성능 결과

모든 전처리 과정을 마친 테이블을 시계열 딥러닝 모델이 학습할 수 있도록 슬라이딩 윈도우 알고리즘의 적용하여 데이터를 가공한다. 슬라이딩 윈도우 알고리즘은 연속된 데이터에서 일정 크기의 윈도우를 설정하고,

윈도우를 하나씩 이동하면서 데이터를 처리하는 알고리즘을 말합니다. 주가는 과거 가격의 영향을 받는 시간적 종속성을 가지고 불안전한 장일수록 주가는 단기 추세 혹은 지역적 패턴을 보입니다. 이는 슬라이딩 윈도우 알고리즘을 적용하기에 타당하고, 본 연구에서는 고정 윈도우 사이즈를 120으로 설정하였고 미래 5일을 예측할 수 있도록 출력 윈도우를 5으로 설정하였습니다. BiLSTM(Bidirectional LSTM) 모델을 이용해서 종목마다 개별 모델링을 하고 미래 5일을 예측하도록 한다. 그리고 학습하기 전 주의해야할 사항이 있다. 종목의 종가, 거래량, 시가총액 등 특성들은 각자 다른 단위를 가지고 있다. 시가총액 처럼 규모가 큰 특성들은 더 중요한 변수 혹은 영향력이 높은 변수로 인식될 수 있다. 이들의 규모 차이를 해소하기 위해 데이터 정규화(Normalization) 작업이 필요하다. Figure 3.4 공식을 이용하여 종목별로 Min-Max 정규화를 진행한다.

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Figure 3.4 Min-Max Normalization

이제 모든 준비가 끝났고 모델 학습에 앞서 모델에 대한 하이퍼 파라미터 탐색은 각 파라미터마다 특정 단위로 범위를 설정하고 수동적인 실험을 통해 최적의 파라미터 조합을 선정하였다. 파라미터는 다음과 같이 고정

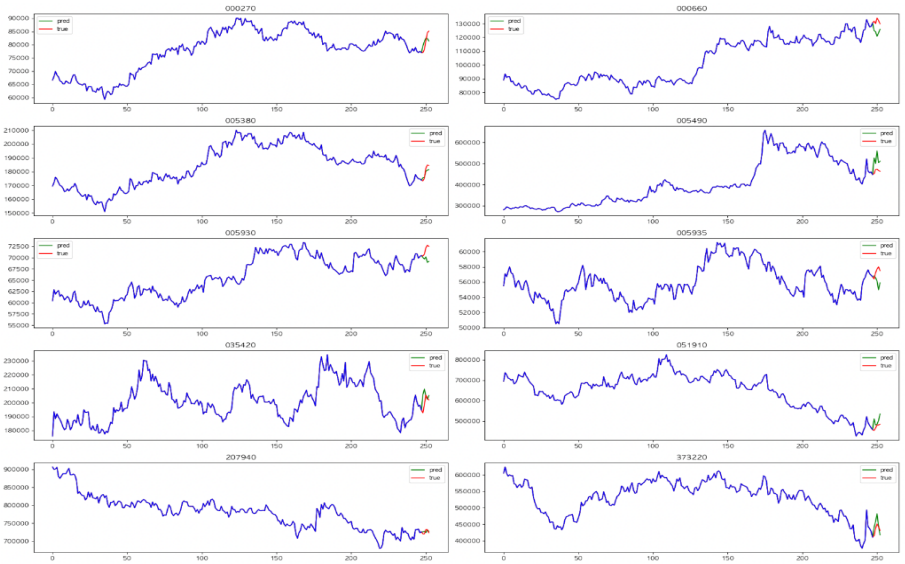
하였다. { num_epochs = 500, learning_rate = 0.01, batch_size = 64, hidden_size = 50, n_layers = 1, dropout = 0 }

그리고 학습기간은 동일하게 2022년 11월 10일부터 2023년 11월 10일 1년치 데이터를 각 종목별로 학습을 진행하였고 성능 비교를 위해 미래 주가의 기간은 주식 시장이 개장하는 날로 하여금 2023년 11월 13일부터 2023년 11월 17일 5일간으로 설정한다. 종목별로 미래기간의 실제 주가의 방향성과 딥러닝 알고리즘이 예측한 미래기간의 주가의 방향성을 한눈에 확인하기 위해서 Figure 4.1 그림과 같이 시각화(Visualization)를 하였다. 종목 코드 000270, 005380, 035420, 207940, 373220 종목들의 경우 비교적 미래 주가의 흐름을 잘 예측하였지만 종목코드 000660, 005930, 005935 종목들의 경우에는 최근 과거 주가 패턴에 너무 치중되어 있어서 주가의 방향성을 전혀 예측하지 못한 것으로 확인하였다.

Figure 4.1 주가 예측 결과

Ⅲ. 결론

주가를 예측함에 있어 합당한 근거를 토대로 분석 기술과 딥러닝 모델을 적용하여 이론상으로는 미래의 주가를 정확하게 예측할 수 있을 것으로 보이지만 역시나 주식 시장은 불확실하고 휘발성이 큰 성질을 가지기 때문에 완벽에 가깝게 예측하는 것은 당연히 어려웠고 본 연구에서 사용된 감성분석에 대해서도 문제점들이 많았다. 우선 수집된 뉴스들의 수가 너무 적었기 때문에 감성점수가 긍정이야하는 부분들이 부정쪽으로 편향되는 부분들이 있었고 그 결과 왜곡된 예측 결과를 가져올 수 있었다. 또한



뉴스들의 제목만을 가지고 즉 한정적인 데이터만을 가지고 감성점수를 매긴다는 점 그리고 주가의 상승 또는 하락 요인에 도움이 되지 않은 뉴스들이 있었고 문제들을 다시금 보안을 해야할 필요가 있다. 그리고 본 연구에서 수집된 정보들만으로는 주가 예측에 한계가 있었다. 주가는 다양한 요인에 영향을 받으며 경제 성장률(GDP) 시장 지수 등 경제, 시장 지표들과, 기업에 대한 영업이익, 재무재표등 기술적 분석도 고려하고 향후 비선형적인 역학 관계와 복잡한 패턴을 식별할 수 있도록 다양한 딥러닝 모델들과 복잡한 모델 구조를 구성하고 하이퍼 파라미터 튜닝을 통해 성능 고고능 성도화도 수행해야된다고 생각된다. 제시한 과정들을 모두 적용한다면 완벽에 가까운 예측은 보장할 수 없지만 주가의 흐름정도는 예측이 가능할 것으로 생각된다.

Ⅴ. 참고문헌

◇

https://koapy.readthedocs.io/en/latest/notebooks/getting-historical-stock-price-data.html

◇ https://towardsdatascience.com/aifortrading-2edd6fac689d#7682

◇ https://github.com/seyoongit/news