

뉴스 기반의 주가 변동폭 예측 시스템에 관한 연구

류재학, 조성혁, Md Habibur Rahman, 장영민

국민대학교

jhak@kookmin.ac.kr, hang8709@kookmin.ac.kr, habib182@kookmin.ac.kr, yjang@kookmin.ac.kr

A Study on the News-based stock price fluctuation prediction system

Ryu Jae Hak, Cho Seong Hyeok, Md Habibur Rahman, Jang Yeong Min

Kookmin Univ.

요약

본 논문은 회사에 대한 긍정적/부정적 뉴스가 주가에 미치는 영향을 학습시킨 뒤 새로 입력되는 뉴스 데이터가 회사에 긍정적인지 부정적인지를 판단하고, 과거의 뉴스 감정과 주가 등락 데이터를 기반으로 새로운 주가 등락을 예측한다. LSTM 기반 주가 예측이 기존의 주가 데이터만으로 미래의 주가를 올바르게 예측하기 어렵다는 점을 개선하고자 하였다.

I. 서론

현재 넓게 분포되어있지만 눈에 띄지 않는 알고리즘 우위 영역은 주식 시장이다. 새로운 거래 알고리즘이 등장해 월스트리트의 거래 방식을 바꾸기도 하며, 알고리즘 거래는 미국 전체 거래의 약 70%, 인도와 같은 신흥 경제국에서는 약 40%의 거래량을 차지하는 것으로 추정된다.[1] 더 나아가 단순 거래뿐만이 아닌 주가를 예측하는 알고리즘 또한 활발하게 연구되었고, 그 중 하나가 LSTM을 이용한 주가 예측이다. 그러나 주가는 실수 체계의 값이고 무작위성이 강하기 때문에 LSTM 모델에서의 학습은 불확실한 중간값이나 의미 없는 값을 반환하는 경향이 강해진다.[2]

그러나 주가는 완전히 무작위적으로 결정되는 값은 아니다. 주가에 영향을 미치는 요소가 존재하고, 그 중 하나는 뉴스에 포함된 기업의 정보이다. 일일 웹 직, 간접 뉴스의 상승/하락 추세는 일일 평균 수익률에 영향을 미친다. 주가에 영향을 미칠 수 있는 기업에 대한 특정 사건이 발생하였을 때 웹 뉴스를 통해 투자자들의 심리가 좌우되고, 이는 곧 주가의 변화를 야기할 수 있음을 뜻한다.[3] 뉴스의 데이터가 주가의 등락을 온전히 결정하지는 않지만, 등락의 방향성을 잡기 위한 유의미한 데이터로서 가치를 갖게 된다.

본 논문에서는 크롤링한 뉴스 데이터를 기반으로 뉴스의 감정(긍정적/부정적)의 단계와 과거 뉴스에서 보여주었던 주가 변동폭의 경향성을 학습시켜 주가 자체의 완벽한 예측이 아닌 주가의 변화 추세를 예측하는 것을 목적으로 하고 있다.

II. 본론

2.1 주가 예측 모델 구현 과정

Kaggle 데이터(<https://www.kaggle.com/aaron7sun/stocknews>)를 기반으로[5] 주가 정보를 모은다. 이후 뉴스 제목을 크롤링하여 정규화를 거치고, VADER(Valence Aware Dictionary and sEntiment Reasoner) 감정 분석을 이용해 텍스트 분석 과정을 진행했다. 결과는 다음과 같다.

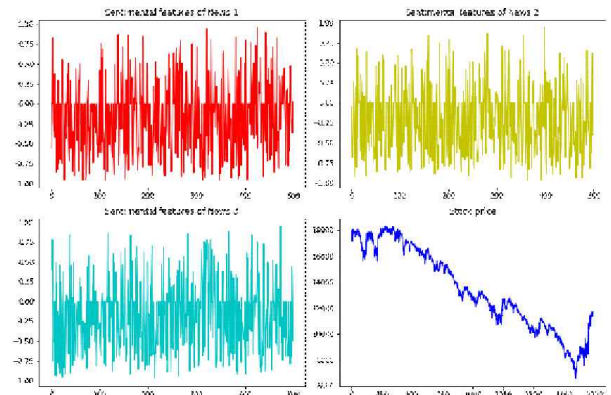


그림 1. 텍스트 특징 추출 데이터 및 주가 데이터의 그림

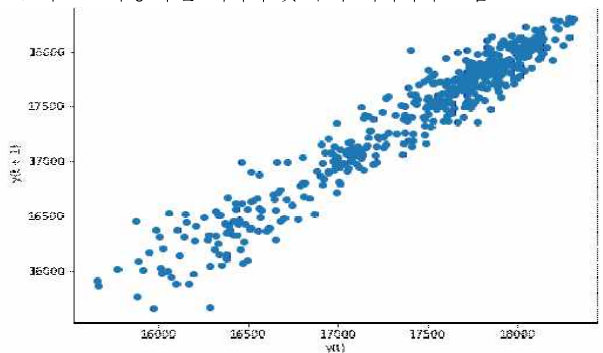


그림 2. 현재 단계(t)와 다음 단계($t+1$)데이터의 상관관계 또한 현재 단계의 $y(t)$ 와 다음 단계의 상관관계 $y(t+1)$ 를 확인하여, 양의 상관관계를 가짐을 결론하였다.

2.2 VADER 감정 분석

VADER(Valence Aware Dictionary and sEntiment Reasoner)는 규칙 기반 감정 분석 라이브러리 및 어휘를 뜻한다. Amazon의 Mechanical Turk를 통해 얻은 데이터를 기반으로 감정 샘플이 없더라도 감정 분류를

빠르게 실행할 수 있다. 또한 VADER 알고리즘은 다른 머신러닝 및 딥러닝 접근 방식과 비교했을 때 계산 면에서 효율적이며, 내장된 학습 라이브러리로 구성되어 있어 모델을 별도로 학습하지 않는다.

VADER 알고리즘은 주관성과 극성 개념을 사용한다. 극성은 문장에 표현된 긍정적 혹은 부정적 감정을 -1과 +1 사이로 나타낸 부동 소수점 값이며, 주관성은 문장의 주관적인 의견이나 감정 또는 판단을 나타내는 0에서 +1 사이의 부동 소수점 값이다. 감정 관련 단어의 사전에 기반하여 하는 감정 분석 유형에 속하며, 이 접근 방식에서 사전의 각 단어는 해당 단어가 얼마나 긍정적인지 부정적인지에 따라 등급이 매겨진다.

VADER 분석의 결과는 데이터의 텍스트가 복잡한 경우 많은 이점을 갖는다. 기존의 수동적 단어 평가의 단점이었던 고비용과 시간 소모 문제를 해결하고, 정확도를 높여 많은 양의 텍스트 데이터에서도 결과를 빠르고 정확하게 반환한다.

VADER는 텍스트 묶음을 분석할 때 텍스트의 단어가 사전에 있는지 확인 후, 3개의 감정 항목과 종합 결과값을 생성한다. 3개의 감정 항목은 각각 부정적, 중립적, 긍정적 텍스트의 비율을 나타내며, 종합 결과값은 -1에서 1 사이의 범위로 표준화된 감정 항목의 합계이다. [4] 이번 실험에서, 데이터셋의 문장 중 "U.N. Chief Admits He Removed Saudi Arabia From Child-Killer List Due to Extortion"에 VADER 알고리즘을 적용했을 때, 다음과 같은 결과가 출력되었다.

감정	결과값
긍정	0.155
부정	0.0
중립	0.845
종합	0.296

표 1. VADER 알고리즘 적용 결과

2.3 실험 과정

실험은 다음과 같은 과정으로 진행되었다. 먼저 인터넷에서 하루 단위로 3개의 주요 뉴스 헤드라인을 크롤링 후 VADER 감성 분석을 거친다. 이러한 시계열 데이터를 종합하여 입출력 행렬의 형태로 변환한다. 그리고 이 데이터 중 70%는 학습 데이터, 10%는 검증 데이터, 20%는 훈련 데이터로 분배하였다. 이후 LSTM 모델에 데이터를 훈련시켰다. 모델 매개변수는 다음과 같다.

매개변수	값
레이어	LSTM, Dropout
뉴런 개수	100, 50
손실 함수	Mean Square Error
옵티마이저	Adaptive Moment Estimation
에포크	50
배치 크기	25

표 2. 훈련 모델의 매개변수

III. 결론

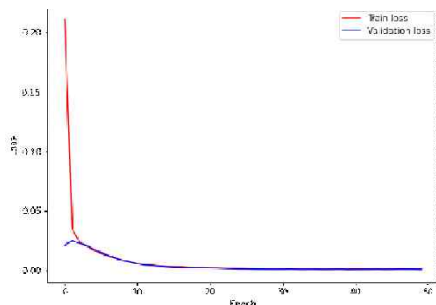


그림 3. 훈련 횟수에 따른 훈련 세트 손실과 검증 세트 손실

실험 결과 훈련을 반복함에 따라 훈련 세트 손실과 검증 세트 손실이 매우 비슷해 과적합을 걱정할 필요 없이 최적점에 도달했음을 확인했다.

성능 측정 지표인 mse, mae, mape를 측정한 결과 mse=0.000847, mae=0.023270, mape=0.108286의 결과가 도출되었다.



그림 4. 예측값과 실제값의 비교 그래프

훈련을 거듭함에 따라 미래 주가의 정확한 수치를 예측하는 능력은 다소 부족하지만, 주가 변동의 방향성과 유사한 추세를 보임을 확인하였다.

본 논문에서는 기존의 LSTM 주가 예측 모델의 단점을 보완하기 위해, 뉴스 데이터에 기반한 기업의 감정 평가를 이용하였다. 주가의 등락을 예상하는 모델을 제안하고, 이를 실제로 실험하여 주가 변동의 방향성을 성공적으로 예측함을 확인했다. 더욱 정확한 예측을 위해 뉴스의 감정 외에 추가적인 종속변수가 필요함이 예상된다.

ACKNOWLEDGMENT

참 고 문 헌

- [1] Cameron Turner, "The Future of Algorithmic Trading", <https://www.experfy.com/blog/the-future-of-algorithmic-trading>, May 2021
- [2] TBACKING, "LSTM 기반 주가 예측의 실상", <https://tbacking.com/2020/08/16/lstm-%EA%B8%B0%EB%B0%98-%EC%A3%BC%EA%B0%80-%EC%98%88%EC%B8%A1%EC%9D%98-%EC%8B%A4%EC%83%81>, Aug 2020
- [3] Sang Soo Kim, "A Study on the Relation of Web News and Stock Price", 2012, p. 201
- [4] Monali Yadav, "Sentimental Analysis on Audio and Video using Vader Algorithm", 2019, p.192
- [5] Sun, J. "Daily News for Stock Market Prediction", <https://www.kaggle.com/aaron7sun/stocknews>, 2016