

긍정적 뉴스에 따른 주기변동 분석

신태현^o, 백재순^{*}, 김성진(교신저자)^{*}

^o명지전문대학교 ICT융합공학과,

^{*}명지전문대학교 ICT융합공학과

e-mail: destro21@naver.com^o, {hisoon99, ict214548}@mjic.ac.kr^{*}

Analysis of stock price fluctuations due to positive news

Shin-Tae Hyun^o, Jai-Soon Baek^{*}, Sung-Jin Kim(Corresponding Author)^{*}

^oDept. of ICT Convergence Engineering, Myongji College,

^{*}Dept. of ICT Convergence Engineering, Myongji College

● 요약 ●

언젠가 한 번쯤 “뉴스에서 발생하는 소식들은 주기변동에 어떤 영향을 줄까?” 하고 궁금해 본 적이 있을 것이다. 본 논문에서는 이러한 궁금증을 해소하기 위해 뉴스기사에 따른 주식시장의 변동을 연구분석한 내용을 소개한다. 이를 위해 뉴스 기사를 크롤링하고 자연어 처리 기술을 통해 긍정적 기사와 부정적 기사를 분류하고 해당 날짜의 주기변동 그래프를 분석하여 뉴스에서의 긍정적 요소와 부정적 요소가 주식시장에 얼마나 영향을 주는지 머신러닝 알고리즘을 통해 알아보았다.

키워드: 크롤링(Crawling), 뉴스(News), 주식(Stock), 긍정(Positive), 부정(Negative), 모델링(Modelling), 단순회귀분석(Simple Regression Analysis), 다중회귀분석(Multiple Regression Analysis)

I. Introduction

본 연구는 긍정적인 뉴스와 주가 변동 간의 상관 관계를 분석하는데 초점을 맞추고 있다. 연구 결과에 따르면 긍정적인 뉴스가 주가에 어떠한 영향을 미치는지를 규명함으로써 금융 시장 참여자들에게 유용한 정보를 제공할 수 있을 것으로 기대된다. 연구에서는 자연어 처리 기술과 머신 러닝 알고리즘을 활용하여 대량의 긍정적 뉴스 데이터를 분석하고, 해당 데이터와 주가 지표 간의 상관 관계를 도출하였다. 이를 통해 향후 긍정적인 뉴스 발표에 따른 주가의 추이를 예측하는데 도움이 될 것으로 기대된다.

II. Preliminaries

1. 대상기업 선정

주기변동을 분석하기 위해 하나의 기업을 선정하여 집중적으로 분석할 필요가 있었다. 방대한 양의 주식시장의 전체를 분석하는 것은 어려움이 있기 때문이다. 따라서 주기변동폭이 큰 편이며 최신 뉴스기사의 양이 충분한 카카오톡 대상기업으로 선정하고 카카오톡에 대한 뉴스 기사를 네이버 뉴스 사이트에서 크롤링하여 텍스트 데이터를 수집하였다.

2. 텍스트 전처리의 필요성

수집한 데이터를 원하는 데이터로 가공하기 위해서는 데이터에 존재하는 수많은 텍스트를 전처리하는 과정이 필요하다. 크롤링을 통해 수집한 뉴스기사의 텍스트를 자연어 처리 기술을 활용하여 정제한다. 불필요한 정보를 제거하고, 형태소 분석을 통해 단어를 추출하여 각각의 단어를 분류할 수 있게 준비하였다.

3. 텍스트 감성분석의 필요성

해당 뉴스기사가 긍정적 뉴스인지 부정적 뉴스인지 판단해서 분류하려면 자연어 처리 기술을 통해 추출한 단어들을 긍정적인 단어와 부정적인 단어로 분류하여 뉴스기사에 있는 단어들의 감성을 분석하는 과정이 필요하다. 군산대학교의 KNU 한국어 감성사전을 이용하여 감성분석을 실시하였다.

4. 분석 모델링

뉴스기사와 주기변동의 상관관계를 분석하기 위해 다중회귀분석, 단순회귀분석, 그리고 GRU 모델을 구축하여 뉴스의 감성과 주가 변동 간의 관계를 모델링하는 과정이 필요하다.

5. 예상결과 및 기대효과

본 연구에서는 카카오에 대한 긍정적 뉴스와 주가 변동 간의 관계를 탐색하고, 감정 분석을 통해 모델을 학습시켜 예측력을 검증할 것으로 기대된다. 이를 통해 긍정적 뉴스가 주가에 미치는 영향을 정량적으로 파악하여 향후 투자 전략에 도움이 되는 결과를 얻을 것으로 기대된다.

III. The Proposed Scheme

1. 데이터 수집

카카오 주가의 변동폭이 본격적으로 활발해지는 2020년부터 22년까지의 데이터를 대상 기간으로 정하고 Yahoo Finance에서 해당 기간 카카오 기업의 주식시장 데이터를 수집하였고 전자공시시스템에서 카카오 기업의 분기별 재무제표를 수집하였다.

2. 크롤링

Python의 BeautifulSoup 패키지를 이용하여 20-22년도의 네이버 뉴스 기사를 크롤링한다. 카카오라는 단어는 기업명으로는 쓰이는 단어가 아니기 때문에 필요 없는 데이터까지 수집되는 상황을 방지하기 위해 카카오의 주식종목번호인 035720을 포함하여 크롤링을 진행하였다.

3. 데이터 전처리

수집한 뉴스기사의 헤드라인을 긍정, 부정, 중립으로 구분하여 분류하기 위해 자연어 처리 기술을 통한 전처리 작업을 진행하였다. 먼저 헤드라인에 포함된 특수문자를 제거한다. 특수문자가 제거된 뉴스기사 헤드라인을 Morphs 함수를 사용하여 형태소 단위로 나누어 추출을 진행하였다.

반복되거나 중복되는 단어들을 제거하고 한 글자로 이루어진 단어들은 앞 뒤 문맥을 알아야 판단이 가능하기 때문에 제거한다.

4. 감성사전

군산대학교의 KNU 한국어 감성사전을 이용하여 추출된 각 단어들 이 긍정, 부정, 중립 중에 어떤 의미를 가지고 있는지 판단하고 분류한다. '상승', '올라갔다' 등의 긍정적인 의미를 가진 단어를 1점, '약세', '날아갔다' 등의 부정적인 의미를 가진 단어를 -1점, '나란하', '시기충액' 등 여러 의미로 사용될 수 있는 단어를 중립으로 분류하여 0점을 부여하는 방식으로 각 날짜별로 점수를 도출하였다.

5. 데이터 시각화 및 모델링

분석에 필요한 데이터를 한 눈에 알아보기 쉽도록 데이터 시각화를 진행하였고 앞서 감성사전을 통해 긍정, 부정, 중립으로 분류한 텍스트 데이터 또한 시각화를 진행하여 긍정 텍스트가 가장 많은 날과 부정 텍스트가 가장 많은 날 등을 분석하였다. 분석의 결과를 알아보기

위해서 단순회귀분석, 다중회귀분석, GRU모델링 등의 다양한 방법으로 데이터분석을 진행하였다. 여러 모델링 중에 GRU모델링을 선택한 이유는 다른 모델들에 비해 정확도가 높은 편이라 선택하고 진행하였다.

IV. Conclusions

1차적으로 독립변수를 긍정텍스트 빈도, 종속변수를 일일마감주가 차이로 설정하고 단순회귀분석을 진행하였고 다른 독립변수를 추가해 보기 위해 다중회귀분석을 진행하였다. 독립변수에 시가와 거래량을 추가하여 다중회귀분석을 진행했으나 실제값과 예측값의 차이가 크고 예측 정확도가 낮게 나왔다. 다중회귀분석을 진행하였고 실제값과 예측값의 차이는 줄어들었지만 여전히 큰 편이고 예측정확도는 높아졌다. 하지만 긍정 텍스트가 많아짐에 따라 일일마감주에 유의미한 차이가 있을 것이라 예상한 것과는 다르게 다른 독립변수를 추가하는 경우에만 예측값이 실제값과 가까워지는 사실을 확인하였다.

REFERENCES

- [1] Kang Doo-won, Yoo So-yeop, Lee Ha-young, & Jeong Ok-ran. (2022). A study on deep learning-based stock price prediction using news sentiment analysis. Journal of the Korea Computer and Information Society, 27(8), 31-39.