뉴스 키워드 노출 빈도수에 따른 기업 주가 영향 분석

Analysis of the impact of corporate stock prices according to the frequency of exposure to news keywords

요약

최근 주식시장을 향한 사람들의 관심이 커지면서, 주식의 가치를 판단할 때 어떤 것을 참고해야 하는지에 대한 고민 또한 증가하고 있다. 주가를 움직이는 원인에는 여러 요소가 있지만, 그중 뉴스가 큰 영향을 준다고 알려져 있다. 따라서 우리는 기업 혹은 관련 업종에 관한 뉴스의 빈도수와 주가 간의 상관관계가 존재한다는 가설을 세워, 이를 뒷받침하기 위한 빅데이터 분석 연구를 진행한다.

1. 서론

최근 주식시장을 향한 사람들의 관심이 커지고 있다. 금융투자협회에 따르면 현재 주식거래로 활동하고 있는 계좌는 2021년 3월 19일 기준 40,067,529개로 집계됐다.[1] 주식거래 활동 계좌는 코로나19 사태 직후인 지난해 3월 6일 3천만 개를 넘어선 뒤, 약 1년 만에 1천만 개가 늘었다. 올해 들어서도 주식 열풍이 이어지면서 지난해 말 3,548만 개보다 500만 개 가까이 늘었다. 이러한 관심의 증가는 사람들이 주식의 가치를 판단할 때 어떤 것을 참고해야 할지 고찰하게 한다.

주가는 여러 요인에 의해 움직이지만, 그중 뉴스가 큰 영향을 주는 경우를 볼 수 있다. 뉴스는 사회의 여러 사건, 각국의 경제, 정치 상황, 기업의 행보 등 여러 정보를 전달하는 매체이다. 현재 스마트폰이 대중화되면서 사람들은 자신이 원하는 주제의 뉴스를 언제 어디서나 실시간으로 접할 수 있게 되었고, 이는 사람들이 뉴스의 흐름에 더욱 민감하게 반응하며 빠르게 변화하는 경제를 손쉽게 따라잡을 수 있게 만들어 주었다. 예를 들어, 2020년 코로나로 인해 진단 키트의 수요가 늘어나자 코로나 진단 키트를 전문적으로 만드는 기업인 씨젠의 노출이 뉴스에서 증가하기 시작했다.[2] 그에 따라 해당 기업의 주가가 같이 증가했음을 확인할 수 있다. 또 다른 예시로, 2021년 초 현대자동차와 애플과의 자동차 협업 이슈가 제기됨에 따라 현대자동차의 주가가 상승[3]하는 모습을 보였다. 하지만 기업에서 직접 이 사실을 부정하고 뒤이어 협상이 결렬되었다는 뉴스가 여럿 등장하자 주가가 곧바로 하락하는 추세[4]를 보였다.

따라서 본 연구에서는 주가와 관련된 뉴스의 노출 빈도수와 주가 간의 상관관계에 대해 분석하고자 한다.

2. 관련 연구

2.1. KoNLPy

KoNLPy(Korean NLP in Python)[5]은 한국어 정보처리를 위한 파이썬 패키지이다. KoNLPy 패키지 안에는 카이스트 semantic web 연구실에서 개발한 Hannanum, 서울대 IDS 연구실의 Kkma, Shineware에서 개발한 Komoran, Mecab-ko등이 포함되어 있다. 이를 이용하여 뉴스에 포함된 한국어 정보들을 원하는 품사에 따라 처리할 수 있다. 샘플데이터를 각 분석기를 이용해 분석해 본 후 소요 시간과 분석 정확도를 고려하여 적절한 분석기를 선택한다.

2.2. KNU 한국어 감성 사전

감성 사전은 웹사이트나 소셜미디어에서 특정 주제에 대한 여론이나 정보를 수집, 분석해 평판을 도출하는 빅데이터 처리 기술이다. KNU 한국어 감성 사전[6]은 군산대학교에서 개발한 한국어 감성 사전으로 표준국어

대사전 뜻풀이의 감성을 Bi-LSTM을 사용해 분류하였다. 해당 사전을 통해 추출한 뉴스 키워드의 감성을 분류한다.

2.3. TF-IDF모델

단어의 빈도수를 세어 수치화하는 방법에는 DTM(문서 단어 행렬: Document-Term Matrix)과 TF-IDF(단어 빈도 - 역문서 빈도: Term Frequency - Inverse Document Frequency)[7]가 있다. DTM은 서로 다른 문서를 비교하기 위해 다수의 문서에서 등장하는 각 단어의 빈도를 행렬로 표현하는 방식이며, 단순히 모든 단어에 대해 빈도수에 기반을 두어 접근하기 때문에 불용어로 인한 유사도 오류가 발생할 수 있다는 단점이 있다. 따라서 불용어와 중요한 단어에 대해 가중치를 별개로 매길 수 있는 TF-IDF를 본 프로젝트에서 사용한다.

TF(단어 빈도: term frequency)는 특정한 단어가 문서 내에 얼마나 자주 등장하는지를 나타내는 값으로, 이 값이 클수록 문서에서 중요하다고 생각할 수 있다. 하지만 단어 자체가 문서 내에서 자주 사용될 때도 TF 값이 크게 나와 단어의 중요도로 인해 나온 결과와 단순한 단어의 흔한 사용으로 나타난 결과를 구별할 수가 없다. 따라서 단어가 흔하게 사용된 정도를 DF(문서 빈도: document frequency)라고 하고, 이 값의 역수를 IDF(역문서 빈도: inverse document frequency)라고 정한 뒤 이를 이용해 TF값의 결과를 보완한다.

TF-IDF는 TF와 IDF를 곱한 값이다. TF-IDF를 이용해 정제된 데이터의 상대적인 가중치를 구하여 그래프로 나타낸다.

3. 설계

3.1. 데이터 수집, 정제

수집할 자료를 크게 업종, 종목, 종합지수별로 구분한다. 코스피 상위 50종목의 기업명과 그 기업이 속한 업종 그리고 종합주가지수인 코스피를 네이버에서 검색했을 때 나오는 뉴스 제목을 크롤링하여 데이터를 수집한다. KoNLPy를 사용하여 명사 단위로 수집한 뉴스 제목에 들어있는 단어를 추출한 뒤, 수치 혹은 의미를 알 수 없는 한 글자 단어들을 제거한다. TF-IDF를 이용하여 문서 내 단어들의 중요도 가중치를 추출한다. 자료의 형태는 엑셀로 지정한다.

3.2. 종목별 특화 감성 사전 구축

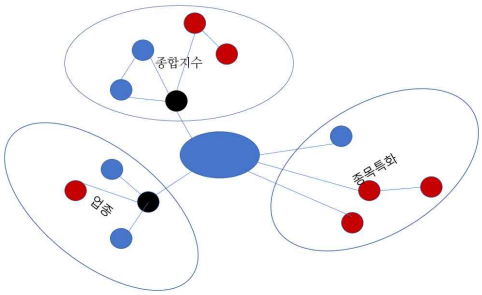

그림 1. 감성 사전 구조도

KNU 한국어 감성 사전을 이용해 수집한 키워드들의 감성을 판별한다. 그 후 판별한 결과를 주가에 적용했을 때 적절한지 비교한 후, 해당 종목에 맞게 다시 조정하여 종목에 특화된 감성 사전을 구축한다.

구축 후 각 종목을 구성하는 키워드의 가중치를 비교하여 중심이 어디인지 판별하고 키워드의 감성에 따라 주가의 경향을 판단한다.

3.3. 데이터 그래프 표현

언어는 파이썬을 사용하며, 그래프의 형태는 Basic Plotting 선 그래프로 설정한다. 가로축은 날짜로 설정하고, 세로축에는 주가 변동과 단어 빈도를 표 안에 함께 표현하여 노출 증감의 빈도와 주가의 흐름을 표시한다. 그래프의 모양 흐름의 유사도를 통해 둘 사이의 연관성을 파악한다.

4. 결론 및 기대효과

하루에 수많은 뉴스가 쏟아진다. 우리는 이러한 뉴스 중 어떤 것이 주가와 연관이 있고 영향을 주는지 찾기가 어렵다. 따라서 이런 주가와 연관된 키워드를 분석하는 프로젝트를 진행하여 직접 그래프를 도출해 비교함으로써 뉴스와 주식의 연관성을 파악할 수 있을 것이다. 어떠한 뉴스 키워드가 성행했는지, 그것이 주식시장에 무슨 영향을 끼쳤는지를 확인할 수 있다. 또한 과거의 뉴스와 주가의 연관성을 분석한 것을 통해 미래의 주가도 예측할 수 있을 것이다.

5. 참고문헌

1) 김영배, “주식계좌 수 4천만개 돌파···1년만에 1천만개 늘어”, 한겨레, 2021년 3월 23일자.

2) 이영란, “"코로나 진단키트 주목"...씨젠 · 바디텍메드, '장중 주가 껑충', 초이스경제, 2020년 10월 8일자.

3) 이영란, “美 애플과 협력?...현대차 · 기아차, '장중 주가 급등'”, 초이스경제, 2021년 2월 3일자.

4) 권유정, “현대차그룹, 애플카 쇼크에 시총 13조원 증발”, 조선비스, 2021년 2월 8일자.

5) 박은정, 조성준. “KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지.” 제26회 한글 및 한국어 정보처리 학술대회 논문집, p. 133-134, 2014.

6) 박상민, 나철원, 최민성, 이다희, 온병원. ” Bi-LSTM 기반의 한국어 감성사전 구축 방안.” 지능정보연구, 24(4), p.219, 2018.

7) Won Joon Yoo, Introduction to Deep Learning for Natural Language Processing, Wikidocs.