**<뉴스 키워드 노출 빈도수에 따른 기업 주가 영향 분석>**

2017104003 양지수

2017104039 한서흔

**요약**

최근 주식시장을 향한 사람들의 관심이 커지면서 주식의 가치를 판단할 때 어떤 것을 참고해야 할지에 대한 고민 또한 증가하고 있다. 우리는 기업 혹은 관련 업종에 관한 뉴스의 빈도수와 주가 간의 상관관계가 존재한다는 가설을 세웠으며, 이를 뒷받침하기 위한 빅데이터 분석 연구를 진행한다. 사용하는 기술에는 자연어 처리, TF-IDF 모델이 있다.

**1. 서론**

**1.1 연구배경(3/15)**

최근 주식시장을 향한 사람들의 관심이 커지고 있다. 금융투자협회에 따르면 현재 주식거래로 활동하고 있는 계좌는 2021년 3월 9일 기준 39,243,554개로 집계됐다. 이는 작년 말 35,495,401개에서 두 달여 만에 약 375만 개 늘어난 수치이다. 신규 주식 투자자가 급증한 작년 한 해 증가량(612만8781개)의 절반 이상을 두 달여 만에 채운 셈이다. 이러한 관심의 증가는 사람들이 주식의 가치를 판단할 때 어떤 것을 참고해야 할지 고찰하게 한다.

주가는 여러 요인에 의해 움직이지만 그 중 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 대중화 되면서 사람들은 자신이 원하는 주제의 뉴스를 언제 어디서나 실시간으로 접할 수 있게 되었고, 이는 사람들이 뉴스의 흐름에 보다 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. 그에 따라 해당 기업의 주가가 같이 증가했음을 확인할 수 있다. 또 다른 예시로, 2020년 초 기아자동차와 애플과의 자동차 협업 이슈가 제기됨에 따라 기아자동차의 주가가 상승하는 모습을 보였다. 하지만 기업에서 직접 이 사실을 부정하고 뒤이어 협상이 결렬되었다는 뉴스가 여럿 등장하자 주가가 곧바로 하락하는 추세를 보였다.

따라서 우리는 주가와 관련된 뉴스의 노출 빈도수와 주가간의 상관관계가 있다고 판단하였고, 이 판단을 확실시하고 근거를 뒷받침하기 위한 빅데이터 분석 연구를 진행하고자 한다.

**1.2 연구목표**

주가와 관련된 뉴스의 데이터를 직접 수집하고 내용을 분석해, 뉴스의 핵심 키워드와 주가 동향의 연관성을 찾는 연구를 진행하는 것을 목표로 한다.

첫째로 코로나 사태로 인해 주식 시장에 많은 변수가 생겼기 때문에, 보다 안정적인 상황이었던 2019년부터 시작하여 2021년 3월까지 작성된 뉴스를 수집한다. 수집한 뉴스의 제목에서 키워드를 추출해 주가와 직접적으로 연관이 있는 핵심 키워드를 선별한다.

둘째로 코스피200 기업들의 업종을 분류해 기업의 주가 변동에 대한 데이터 베이스를 구축하고, 각 업종에 속한 기업들의 주가의 중간값을 활용하여 해당 업종의 평균적인 주가 변동그래프를 만든다. 그리고 그래프에 나타난 주가의 변화가 뉴스와 어떠한 연관이 있는지를 파악한다.

마지막으로 뉴스는 사회적 이슈와 기업의 행보 등 여러 정보를 가지고 있다. 이런 정보는 투자자들에게 그 기업에 투자 여부에 대한 근거가 된다. 따라서 우리는 뉴스 키워드의 노출 정도가 개인과 기관, 그리고 외국인투자자의 매매와 연관성이 있는지 탐구한다.

**2. 관련연구**

**2.1 사용 기술**

**2.1.1 자연어 처리(NLP)**

자연어란 우리가 일상 생활에서 사용하는 언어이다. 자연어 처리란 이러한 자연어의 의미를 분석하여 컴퓨터가 처리할 수 있도록 하는 일을 말한다. NLP 기술은 기계번역, 대화체 질의응답 시스템 대화시스템, 정보검색, 말뭉치 구축, 시맨틱웹, 텍, 딥러닝, 그리고 빅데이터 분석 분야뿐만 아니라 인간의 언어정보처리 원리와 이해를 위한 언어학과 뇌인지 언어정보처리 분야까지 핵심적인 요소로 작용하고 있다. 크롤링한 뉴스 제목을 형태소 분석을 통해 형태소 단위로 분리하여 텍스트 데이터를 정제한다.

**2.1.2 TF-IDF모델**

TF-IDF는 단어의 빈출도(Term Frequency)를 이용하여 해당 단어가 문서 내에 얼마나 자주 등장하는지를 추출한다. 정보 검색과 텍스트 마이닝에서 이용하는 가중치로, 여러 문서로 이루어진 문서군이 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치이다.

TF(단어 빈도)는 특정한 단어가 문서 내에 얼마나 자주 등장하는지를 나타내는 값으로, 이 값이 높을수록 문서에서 중요하다고 생각할 수 있다. 하지만 단어 자체가 문서 내에서 자주 사용 되는 경우, 이것은 그 단어가 흔하게 등장한다는 것을 의미한다. 이것을 DF(문서 빈도)라고 하며, 이 값의 역수를 IDF(역문서 빈도)라고 한다. TF-IDF는 TF와 IDF를 곱한 값이다. TF-IDF를 이용해 정제된 데이터의 상대적인 가중치를 구하여 그래프로 나타낸다.

**2.2 기존 연구의 문제점과 해결방안**

기존 연구에서는 빅데이터 분석 기법 중 오피니언 마이닝을 이용하여 비정형 뉴스 컨텐츠의 감성을 분석하고 수치화하여 뉴스가 주식 시장에 미칠 영향을 분석하고, 이를 기반으로 투자정보를 생성하는 지능형 투자의사결정 모형을 제시하였다.

다만 이 연구에는 한계점이 존재했는데, 먼저 뉴스와 주가를 비교하기에 충분한 데이터 양을 사용하지 않았다. 또한 개별종목의 주가가 아닌 종합주가지수를 비교의 대상으로 하고 있어, 기업에 대한 투자의사결정 정보를 제공하지 못했다.

위의 문제에 대한 해결방안을 다음과 같이 마련했다. 먼저 해당 연구에서 진행한 데이터보다 많은 양의 데이터를 분석한다. 기존연구에서는 3개월간의 뉴스를 조사하였지만 본 연구에서는 2019년 1월부터 2021년 3월까지 총 2년 3개월에 해당하는 뉴스를 조사할 예정이다. 분석할 데이터의 폭이 넓어짐에 따라 결과에 대한 정확도는 자연스럽게 높아질 것으로 예상된다. 또한 형태소로 정제한 데이터들을 TF-IDF모델을 이용해 키워드 가중치를 구하고 이를 종합지수가 아닌 개별 주가와 연관지어, 뉴스키워드와 기업간의 연결점을 직접적으로 찾고 이를 투자의사결정에 참고가 되는 정보가 되는지 탐구한다.

**3. 프로젝트내용**

2019년부터 시작하여 2021년 3월까지 네이버 포털 사이트에 게시된 뉴스의 제목 데이터를 각 뉴스 카테고리 별로 크롤링을 통해 수집한다. KoNLPY패키지를 사용해 수집한 데이터를 자연어 처리하여 형태소를 분석한다. 접속사, 조사와 같은 불필요한 단어는 제거하고 사용 가능한 단어들을 추출한다. 또한 같은 포털 사이트 내에서 기업을 검색한 뒤 나오는 뉴스의 제목을 크롤링하여 기업별로 핵심 키워드를 정리한다.

추출한 뉴스의 키워드와 기업의 키워드를 비교하여 둘 간의 연관성을 확인하고, 기업별로 각각 관련된 뉴스 키워드를 저장하는 데이터베이스를 생성한다.

데이터베이스에 저장된 단어들을 TF-IDF모델을 이용하여 각 단어 별로 가중치를 매긴 뒤, 상위 목록 내에서 통상적으로 사용하는 긍정·부정적 의미를 가진 키워드를 분류한다.

최종적으로 선별된 키워드를 날짜별로 그래프화하여 노출 빈도의 증감을 나타내, 주가 그래프의 증감과 관련성이 있는지 확인한다.

**4. 향후 일정, 역할분담**

|  |  |
| --- | --- |
| 날짜 | 내용 |
| ~3월 31일 | 기초 조사서 제출(3/29), 뉴스 크롤링 |
| 4월 1일~14일 | 자연어 처리 |
| 4월 15일~20일 | 자연어 처리, 중간보고서 작성 |
| 4월 21일~30일 | 중간 보고서 제출, TF-IDF 처리 |
| 5월 1일~22일 | TF-IDF 처리 |
| 5월 23일~31일 | 결과물 점검 및 최종보고서 작성 |
| 6월 1일~ | 최종 발표 |

**5. 결론, 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. 특히 코로나로 인해 큰 폭으로 요동 친 주가 상황과 맞물려 어떠한 뉴스 키워드가 성행 했는지, 그것이 주식 시장에 무슨 영향을 끼쳤는지를 확인할 수 있다. 또한 과거의 뉴스와 주가의 연관성을 분석한 것을 통해 미래의 주가도 예측할 수 있을 것이다.

**6. 참고**

1. 김유신 외 2인, 뉴스와 주가 빅데이터 감성분석을 통한 지능형 투자의사결정모형

2. 김형수, 우리말 자연어 처리 기술 - 과거와 현재

3. 이성직, 김한주, TF-IDF의 번형을 이용한 전자뉴스에서의 키워드 추출 기법