Stock Fluctuation Prediction based on Internet Stock News Analysis

성균관대학교 일반대학원

데이터사이언스 융합학과

민재홍, 손지훈, 전재준

목차

표목차

그림목차

논문요약

Stock Fluctutation Prediction based on Internet Stock News Anlaysis

주가 예측은 다양한 학문에서 다양한 방법으로 시도되어 왔다. 과거의 주가 그래프를 이용한 시계열 분석 방법, 장부가치 대비 주가가 낮은 종목을 투자하는 가치투자 등 다양한 방법으로 주가 상승을 예측하여 수익을 거두려 하였다. 하지만 주식시장에는 급격한 시장 변화, 자연재해, 투자자의 군중심리 등 다양한 변수들이 작용되어 주가의 등락을 예측하기 쉽지않다.

본 연구에서는 이러한 다양한 변수들을 통제하고 보다 정확하게 주가를 예측하기 위해서 인터넷에 실시간으로 업로드 되는 증권뉴스를 통해서 주가 등락을 예측하고자 하였다. 해당 종목의 당일 인터넷 뉴스기사를 수집하고, 뉴스기사 속 단어들이 주가에 미치는 영향을 분석하여 익일의 주가 등락을 예측하여 보았다. 이와 더불어 뉴스기사 외에 변화를 줄 수있는 다양한 지표들을 대입하여 더 정확한 주가 등락 예측을 하였다. 연구결과 본 연구에서는 71% 의 정확성을 보여주었다.

주제어 : 주가 예측, 딥러닝 , 석유/화학, 증권뉴스

제1장. 서 론

1.1 연구 배경 및 개요

주가 예측은 다양한 학문에서 다양한 방법으로 시도되어 왔다. 과거의 주가 그래프를 이용한 시계열 분석 방법, 장부가치 대비 주가가 낮은 종목을 투자하는 가치투자 등 다양한 방법으로 주가 상승을 예측하여 수익을 거두려 하였다. 하지만 주식시장에는 급격한 시장 변화, 자연재해, 투자자의 군중심리등 다양한 변수들이 작용되어 주가의 등락을 예측하기 쉽지 않다.

본 연구에서는 이러한 다양한 변수들을 통제하고 보다 정확하게 주가를 예측하기 위해서 인터넷에 실시간으로 업로드 되는 증권뉴스를 통해서 주가 등락을 예측하고자 하였다. 해당 종목의 당일 인터넷 뉴스기사를 수집하고, 뉴스기사 속 단어들이 주가에 미치는 영향을 분석하여 익일의 주가 등락을 예측하여 보았다. 이와 더불어 뉴스기사 외에 변화를 줄 수 있는 다양한 지표들을 대입하여 더 정확한 주가 등락 예측을 하였다.

필자들은 해당연구를 위해 매일경제 증권코너에서 석유/화학 시가총액 상위 5개 업체들의 1년치 주가 데이터를 수집했으며, 해당업체의 관련뉴스는 네이버 증권 페이지에서 수집했다. 수집한 뉴스기사는 BOW, TF-IDF 2가지 방법을 이용해 임베딩했으며, 신문기사 외에도 석유/화학 업종에 영향을 미칠 수 있는 kospi지수, 환율, Nasdaq지수, Dow 지수, 유가지수(WTI)를 결합하여 데이터를 구성했다. 모델 구축은 딥러닝, RandomForest, SVM 등의 방법을 채택하여 예측했는데, 본연구에서 71%의 정확도로 등락 여부를 예측할 수 있었다.

1.2 논문의 구성

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 데이터 수집 방법 및 일련의 전처리 과정과 구축 모형에 대해서 설명한다. 3장에서는 본 연구를 통해 제안하는 최적의 주가 등락 예측 모델을 설명하며, 실제 데이터에서 정확도가 어떻게 달라지는지 설명한다.

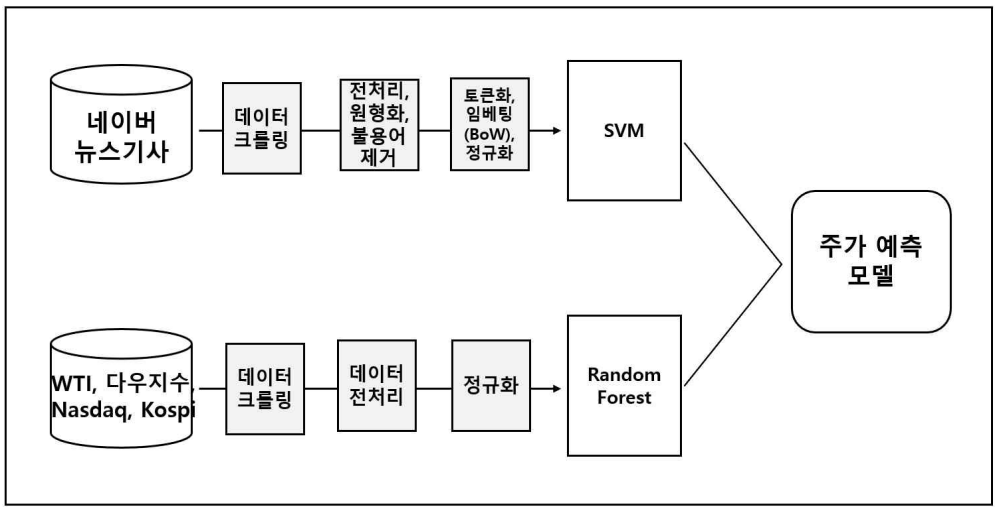
제2장. 본 론

본 연구에서는 특정 종목과 관련된 뉴스 기사 및 유의미한 지표를 이용해 익일의 주가 등락을 예측하는 모델을 제시한다. 뉴스기사가 실시간으로 올라오는 점을 착안하여, 주식의 시가를 기준으로 상승 여부를 예측한다. 필자들은 신문기사와 각종 지표를 이용한 다중모델을 제안한다.

2.1 연구 배경 및 개요

1) 연구 방법의 개요

뉴스 기사는 네이버 증권 코너에서 크롤링 하여 수집하였다. 네이버 증권 코너에는 1년치 뉴스기사만 제공되기 때문에, 데이터 수집을 위해 3개 상장사의 증권 뉴스를 수집한다. 단, 기사에 따른 영향도가 비슷한 종목을 채택하여 모델의 객관성을 증대 시킨다. 본 연구에서는 유사한 흐름을 보이는 화학 관련 주(LG화학, 롯데케미칼, 한화케미칼) 3개사 데이터를 수집했다.



[그림1.1] 주가 등락 예측모델 구조도

2.2 데이터 수집

1) 수집 대상 종목 선정

뉴스 기사는 네이버 증권 코너에서 수집했다. 네이버 증권 코너에는 1년치 뉴스만 제공되기 때문에, 충분한 데이터 수집을 위해 3개 상장사의 증권 뉴스를 수집했다. 단, 기사에 따른 영향도가 비슷한 종목을 채택하여 모델의 객관성을 증대 시킨다. 본 연구에서는 유사한 흐름을 보이는 화학 관련 주(LG화학, 롯데케미칼, 한화케미칼) 3개사 데이터를 수집했다.

2) 수집 방법 선정

신문기사의 경우 Python Beautiful-soup 패키지를 이용해 네이버 증권 코너 (출처: https://finance.naver.com/)를 크롤링 했으며, 3개 상장사에 대한 1년치 기사 (18년5월~19년5월) 17716개를 크롤링 했다.

그 외 주가에 영향을 줄 수 있는 지표(WTI, 환율, 나스닥, Kospi지수 등)는 포털 사이트를 통해 수집했다. (출처: https://kr.investing.com/indices/major-indices)

[표2.1] 신문기사 데이터 수집 예시

날짜	기사	상승여부
2018.06.01	롯데그룹은 사업장별로 정기적인 화재 지진 테러 등에 대비한 방재훈련을 하고 있다(이하 생략)	1

[표2.2] 경제 지표 데이터 수집 예시

날짜	Nasdaq	Dow	환율	WTI	Kospi	상승여부
2018.06.01	74455.58	24620.79	1070	22820	24550	1

2.3 데이터 전처리

1) 원형화 및 불용어 제거

Konlpy에서 제공되는 자연어 처리기(OKT) 를 이용해 기사에 포함된 단어들을 원형화 처리했으며, 모델과 관련없는 단어(기자이름, 숫자단위 등) 800여개를 불용어로 선정하여 제거했다.

[표2.3] 불용어 처리 예시

불요어 종류	예시
숫자 및 특수문자 삭제	▲, 123
대명사 삭제	기자 이름, 특정 인물에 대한 언급
광고성 기사	광고 링크
하나의 음절로 구성된 단어	채, 명, 등

[표2.4] 불용어 처리 후 단어 빈도순위 변화

처리 전	처리 후
화학, 등, 사업, 것, 부회장, 기술, 수 …	경제, 코스피, 롯데, 케미칼, 화학 …

2) 원형화 및 중요 품사 추출

의미있는 단어만 선별하기 위해, 문서 내 단어들을 원형화 하고, 가장 유의미한 품사인 명사와 동사만 추출한다.

처리전
외국인 투자자는 4일 거래소에서 하이닉스 현대차 등을 중점적으로 매도한 것으로 나타났다 외국인 투자자의 순매도 상위 2개 종목은 하이닉스 현대차 등이다.

처리후
외국인 투자자 거래소 하이닉스 현대차 중점 매도

[그림 2.3] 원형화/품사추출 예시

3) 임베딩

본 연구에서는 신문 기사를 임베딩하는 방법으로 BoW와 TF-IDF를 채택했다.

- BoW

BoW(Bag of Word) 방식은 위 전처리된 각 기사의 단어들을 토큰화한 후, 하나의 단어 저장소를 구축하여 각 날짜별 통합된 기사별로 저장소에 있는 단어들을 얼마나 포함하고 있는지 빈도를 계산하여 활용하였다. 이 방식으로 활용한, 중복이 제거된 모든 단어의 수는 4504이며, 날짜별 통합된 기사묶음은 560개이다. 통합 묶음별 4504개 단어들을 포함하고 있는 빈도 테이블을 통해 예측모델링을 실시하였다.

	경제	코스피	롯데	...	투자자
회사1 Day_1	3	2	0	...	1
회사1 Day_2	0	1	0	...	0
....	...	...	...	...	0
회사3 Day_n	0	2	1	...	2

- TF-IDF

TF-IDF(Term Frequency - Inverse Document Frequency) 방식은 여러 문서로 이루어진 문서군이 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치이다. 문서의 핵심어를 추출하거나, 검색 엔진에서 검색 결과의 순위를 결정하거나, 문서들 사이의 비슷한 정도를 구하는 등의 용도로 사용할 수 있다. TF(단어 빈도, term frequency)는 특정한 단어가 문서 내에 얼마나 자주 등장하는지를 나타내는 값으로, 이 값이 높을수록 문서에서 중요하다고 생각할 수 있다. 하지만 단어 자체가 문서군 내에서 자주 사용되는 경우, 이것은 그 단어가 흔하게 등장한다는 것을 의미한다. 이것을 DF(문서 빈도, document frequency)라고 하며, 이 값의 역수를 IDF(역문서 빈도, inverse document frequency)라고 한다. TF-IDF는 TF와 IDF를 곱한 값이다.

$$tf(t,d) = 0.5 + \frac{0.5 \times f(t,d)}{\max f(w,d) : w \in d}$$

$$idf(t,D) = \log \frac{|D|}{|d \in D : t \in d|}$$

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D)$$

$|D|$: 전체 문서의 수

$|d \in D : t \in d|$: 단어 t가 포함된 문서의 수

특정 문서 내에서 단어 빈도가 높을수록, 그리고 전체 문서들 중 그 단어를 포함한 문서가 적을수록 TF-IDF값이 높아진다. 따라서 이 값을 이용하면 모든 문서에 흔하게 나타나는 단어를 걸러내는 효과를 얻을 수 있다. 따라서 본 연구에서는 위 TF-IDF를 활용하여 560개 통합 묶음별 4412개 단어들을 포함하는 TF-IDF 테이블을 통해 예측 모델링을 실시하였다.

	경제	코스피	롯데	...	투자자
회사1 Day_1	0.04	0.78	0.01	...	0.56
회사1 Day_2	0.53	0.23	0.13	...	0.21
....	...	...	...	...	...
회사3 Day_n	0.22	0.55	0.32	...	0.54

2.4 데이터 모델 구성

본연구에서는 모델은 svm, 랜덤 포레스트의 혼합 모델을 제안한다.

- 신문기사를 이용한 주가 상승 예측 모델 : Bow & SVM with linear kernel
- 경제지표를 이용한 주가 상승 예측 모델 : RandomForest with 1500 Trees

제3장. 실험 및 결과

3.1 실험 데이터

총 560일의 신문기사를 수집, 509일(90% 수준)의 기사 데이터를 traindata로, 51일치 기사분을 testdata로 분류했다.(Train set의 10%를 Validation으로 활용)

	Data size
Train set	BoW : (509, 4504) / TF-IDF (509, 4412)
Test set	BoW : (51, 4504) / TF-IDF (51, 4412)

[표3.1] trainset & dataset

	Data size
Train set	(509, 5)
Test set	(51, 5)

[표3.2] 경제지표 trainset & dataset

3.2 평가 방법

본 연구에서는 예측값에 대한 평가 방법으로 Accuracy를 사용하였다. 주가의 상승과 하락 예측여부의 일치정도를 산술하여 평가하는 방법이다.

[그림3.1] 오차 행렬

	Predict P	Predict N
Actual P	TP	FN
Actual N	FP	TN

[그림3.2] Accuracy 산술식

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

3.3 모델별 결과비교

[표3.3] 뉴스기사만 사용한 예측모델(BoW)

	accuracy	precision	recall	F1
DNN	0.49	0.85	0.17	0.29
RF	0.63	0.63	0.46	0.53
SVM	**0.67**	**0.71**	**0.46**	**0.56**

[표3.4] 뉴스기사만 사용한 예측모델(TF-IDF)

	accuracy	precision	recall	F1
DNN	0.61	0.62	0.21	0.31
RF	0.63	0.47	0.55	0.51
SVM	0.47	0.36	0.65	0.46

[표3.5] 경제 지표를 사용한 예측모델 : RandomForest 채택

	accuracy	precision	recall	F1
DNN	0.67	0.63	0.60	0.61
RF	**0.66**	**0.79**	**0.58**	**0.67**
SVM	0.56	0.75	0.29	0.41

[표3.6] 뉴스기사(SVM) + 경제지표(RandomForest)를 포함한 예측 모델

	accuracy	precision	recall	F1
SVM + RF	0.71	0.68	0.69	0.74

표 3.6 모델의 경우 표3.3, 표3.5 모델을 혼합하여 사용하였다. 특정일에 대해 등락 확률을 각각 산출했으며 더 높은 확률을 최종 예측값으로 선정했다.

Ex. 뉴스기사(SVM) -> 상승확률(55%), 하락확률(45%)

경제지표(RandomForest) -> 상승확률(20%), 하락확률(80%) 일 때,

경제지표가 예측한 80%를 채택하여 익일 주가가 하락할 것으로 판단.

제4장. 결과 해석 및 향후 계획

본 연구에서는 익일의 주식 시가를 예측하기 위한 뉴스 기사와 지표를 이용한 이중 모델을 제안한다. 뉴스기사만 이용해 모델을 구축했을 때, Embedding방법, 모델에 따라 예측력이 다르게 나타났는데, Bow방식으로 SVM 모델을 구현했을 때 가장 높은 예측력을 보였다. 반면, 경제지표만을 이용해 모델을 구현했을 때 Randomforest 방법이 가장 잘 예측 하는 것으로 나타났다. 본 연구에서는 두 모델을 혼합한 방법으로 모델을 구현해 보았는데, 2% 정도 예측력이 개선될 수 있었다.

기존 전문가들의 경우 경제 지표를 이용한 이동 평균(Moving Average)을 주로 활용하고 있는데, 추가 연구에 접목해 볼 것이 필요할 것으로 생각된다. 또한 본 연구에서는 데이터 부족으로 인해 충분한 신뢰성을 확보하지 못했으나, 뉴스를 통한 주가등락 예측이 가능함을 확인했으며, 향후 충분한 데이터를 이용해 분석할 경우 보다 신뢰성 높은 모델을 구현할 수 있을 것으로 전망한다.

참고 문헌

[1] Word2vec, [Online]. Available: https://en.wikipedia.org/ wiki/Word2vec

[2] 유은순.최건희.김승훈(2015) TF-IDF와 소설 텍스트의 구조를 이용한 주제어 추출 연구.KCI

[2] 김유신.김남규.정승렬 (2012) 뉴스와 주가 빅데이터 감성분석을 통한 지능형 투자의 결정모형.지능정보연구

[3] 성노윤.남기환(2017) 온라인뉴스 및 거시경제 변수를 활용한 주가예측.한국지능정보시스템학회

[4] 송유정.이재원.이종우. 텐서플로를 이용한 주가예측에서 가격-기반 입력 피셔의 예측성능평가.정보과학회 컴퓨의 실제 논문지

[5] 안성원.조성배(2010) 뉴스 텍스트 마이닝과 시계열 분석을 이용한 주가예측.한국컴퓨터종합학술대회