서술어 중심 감성 사전을 통한 주가 등락 예측

엄장윤, 이수원
숭실대학교 컴퓨터학과
e-mail : eom3835@gmail.com

Stock Market Prediction using Sentiment Dictionary based on Predicates

Jang-Yun Um, Soowon Lee
Dept. of Computer Department, SoongSil University

요 약

본 연구에서는 경제 뉴스로부터 서술어 중심의 감성 사전을 구축하고, 하루 동안에 배포된 뉴스를 이용해 전일 종가 대비 당일 종가의 등락을 예측하는 모델을 제안한다. 기존의 주식 도메인 관련 감성 사전을 구축하는 방식은 주가 등락에 관련된 명사를 중심으로 사전을 구축하는 방식이나 대부분의 명사는 극성 값이 중립인 경우가 많아 극성 값을 추정하기 힘들다는 문제점이 있다. 본 연구에서는 극성 값이 잘 표현되는 서술어 중심의 감성사전을 구축하고 극성 값을 자동 추출하여 주가의 등락을 예측한다. 실험 결과 기존 감성 사전을 통한 주가 예측 방법에 비하여 본 연구에서 제안하는 서술어 중심의 감성 사전을 통한 주가 예측 정확도가 높게 나타났다.

1. 서론

기업의 가치를 평가하고 나아가 국가의 가치를 평가할 수 있는 수단 중에 하나는 증권시장이다. 2010 년 한국은행 발표 자료에 의하면 1996 년 ~ 2010 년간 주식 시장의 규모가 꾸준한 상승세를 보이고 있는 추세라고 하고 있다. 따라서 증권시장을 방향성을 예측하는 일은 중요하다[1]. 현재 경제, 통계, 전산, 인공지능 분야 등에서 다양한 연구방법으로 증권시장을 예측하는 연구가 진행 중이다.

Robert P.Schumarker 에 의하면 기술적 주가 예측은 인공지능, 통계적, 수학적 기술로 가능하며 뉴스에서 패턴을 추출하고 이를 통해 주가를 예측하는 방법이 최근의 연구 추세라고 소개하고 있다[2]. 국내에서도 이와 유사하게 뉴스의 패턴을 추출해 주가를 예측하는 방법이 활발하게 연구 중이다. 대표적인 예로 매일 배포되는 경제 뉴스로 오피니언 마이닝(Opinion Mining)을 통해 감성사전을 구축하고 주가 예측 모델을 학습하여 코스피의 종가의 등락을 예측한 연구[3]와 뉴스 데이터를 이용하여 극성을 뽑고 시계열 분석을 통해 개별 종목의 종가 등락을 예측한 연구[4]를 들 수 있다.

본 연구에서는 이러한 오피니언 마이닝 기법을 이용하지만, 기존 연구 방법과의 차별성을 가지는 주가 등락 예측 모델을 제시한다. 첫째, 감정을 나타내기 힘든 명사 중심의 기존 감성사전이 아닌 동사와 형용사로 구분되는 서술어 중심의 감성사전을 구축한다. 둘째, 기존 연구에서는 뉴스의 발생 시점을 고려하지 않고 주가 등락 예측 모델을 생성하였지만, 장전에 발생한 뉴스와 장중에 발생한 뉴스가 주가에 미치는 영향이 다를 수 밖에 없기 때문에 시간을 중심으로 장중에 발생한 뉴스와 장전에 발생한 뉴스로 뉴스를 구분하여 감성사전을 구축한다.

본 논문의 구성은 다음과 같다. 2 장에서는 뉴스를 활용한 주가 등락 예측에 대한 기존 연구에 대해 소개하고 3 장과 4 장은 뉴스를 이용해 감성 사전을 자동으로 구축하는 방법과 전일 대비 당일 종가의 등락을 예측하는 모델을 제시한다. 이에 대한 실험 환경 및 실험 결과는 5 장과 6 장에서 설명한다. 끝으로 7 장에서는 본 연구의 결론 및 한계, 향후 연구에 대해 기술한다.

2. 관련 연구

뉴스를 이용한 주가 예측과 관련된 국내 연구들은 다음과 같다. 유은지, 김유신[3]은 2011 년 7 월부터 9 월까지 3 개월간의 뉴스에서 출현한 단어들에 대하여 전일 대비 종가가 상승한 경우에 대해 빈도수를 기반으로 명사 중심의 감성사전을 구축하고 전일 대비 종가의 등락을 예측하는 지능형 투자의사결정 모델을 제안하였으며, 이 모델을 통해 약 52%의 예측 정확도를 얻었다. 또한, 안성원[4]은 2005 년부터 2008 년까지

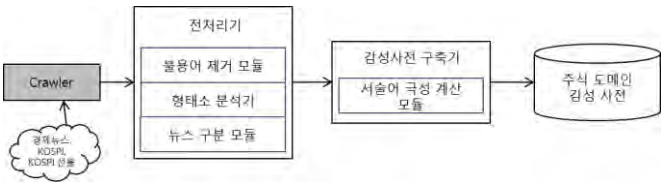
4 년간의 뉴스 데이터를 이용하여 전일 대비 당일의 종가에 대한 주가 등락폭이 +2% 상승이면 호재, -2% 하락이면 악재로 나누고 Bag of word, Naive Bayesian 분류 기법을 이용하여 전일 대비 종가 또는 다음 거래일의 시초가 등락을 예측하였다.

해외 관련 연구들을 살펴보면, Schumaker[5]은 2005 년 10 월부터 2005 년 11 월까지 5 주간의 S&P 500 관련 종목에 해당되는 경제 뉴스에서 반복되는 고유명사만을 추출하여 뉴스가 배포된 시점의 주가 대비 20 분 뒤의 주가에 대한 등락을 SVM 을 이용하여 예측하는 AZFinText 라는 모델을 제안하였으며, 이 모델을 통해 58.2%의 예측 정확도를 얻었다.

또한, Xiangyu Tang[6]는 2008 년 3 월부터 7 월 까지의 뉴스 데이터로 중국 증권 시장의 특정 업종(에너지, 정보통신, 부동산)에 대해 PR(Probability Ratio) Rank 를 계산하고 SVR(Support Vector Regression)을 통해 주가 등락을 예측하였으며, 63%의 예측 정확도를 얻었다.

3. 감성 사전 자동 구축 시스템

본 연구에서 제안하는 감성 사전 자동 구축 시스템은 (그림 1)과 같으며, 웹에서 수집된 경제 뉴스 전처리기와 감성사전 구축기 모듈로 구성된다. 전처리기는 수집된 뉴스를 배포 시간을 기준으로 장전, 장중으로 구분하고 불필요한 문자들과 광고 문구를 제거하는 과정을 수행한다. 감성사전 구축기는 수집된 뉴스에서 서술어를 추출하고 극성 계산식에 의해 감성 단어들의 극성을 부여하는 역할을 수행한다.



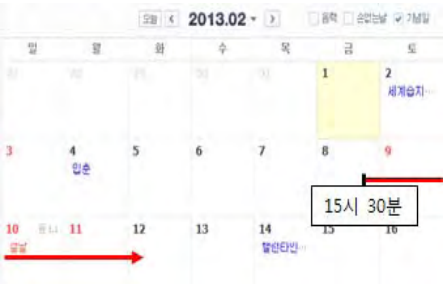
(그림 1) 감성 사전 자동 구축 시스템 구조도

3.1 전처리기

불용어 제거 모듈은 수집된 데이터가 올바르게 분석되도록 정제하는 모듈이다. 경제 뉴스는 불필요한 광고문구와 숫자, 종목코드와 같은 단어가 많다. 이로 인해 형태소 분석기가 올바르게 동작하지 못하기 때문에 데이터를 정제해주는 작업이 필요하다. 이를 위해 전처리기에서 아래와 같은 작업을 수행한다.

- 하나의 음절을 가진 단어("저","그","외" 등) 제거
- 숫자 혹은 연도와 같이 의미 없는 단어들 삭제
- 특수문자와 광고문구 삭제
- 동사, 형용사만을 추출
- 명사+동사로 구성된 경우 동사만을 추출
  예> 상승마감하다 ☞ 상승(명사)+마감하다(동사)

뉴스 구분 모듈은 개장 일에 배포된 뉴스가 아닌 경우(주말, 공휴일)를 처리하고 장중 뉴스와 장전 뉴스를 구분하는 모듈이다. 장중 뉴스는 개장 시간(09:00~15:00) 사이에 배포된 뉴스이며, 장전 뉴스는 개장 일을 기준으로 전일(D-1) 15 시부터 당일 09 시 사이에 배포된 뉴스로 구분된다. 주말과 공휴일에 대해서는 (그림 2)와 같이 다음 개장 일의 장전 뉴스로 처리한다. 예를 들어 2013 년 2 월 8 일 15:30 분에 발생한 뉴스는 주말(9 일, 10 일)과 공휴일(11 일)을 포함하기 때문에 뉴스 구분 모듈은 8 일 뉴스를 12 일 장전 뉴스로 처리하게 된다.



(그림 2) 주말, 공휴일 처리

3.2 감성 사전 구축기

전처리 작업이 완료된 뉴스 데이터로부터 장전, 장중으로 구분하여 감성사전을 구축한다. <식-1>은 감성단어에 대한 극성을 구하는 식으로, [3]에서의 주가지수 상승 예측을 위한 주제지향 감성사전 구축 방안을 적용한 것이다. Term(i,j)는 특정 뉴스 j 에 포함된 감성단어 i 에 대하여 뉴스 j 가 배포된 날짜의 종가가 전일 대비 상승(1)인지 상승이 아닌지(0)를 나타내는 변수이다. Term(i).score 는 0~1 의 범위를 갖는 감성단어 i 의 극성 값이며, 1 에 가까울수록 강한 상승을 의미한다.

$$\text{Term}(i,j) = \begin{cases} 1 & \text{특정 뉴스 } j\text{가 배포된 날짜를 "d일"이라고 가정하면 (d일 - d-1일)이 양수(+)인 경우} \\ 0 & \text{그 외의 경우} \end{cases}$$

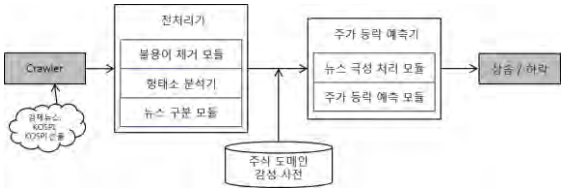
$$\text{Term}(i).\text{NumNews} = \text{Term}(i)\text{를 포함하고 있는 뉴스의 개수}$$

$$\text{Term}(i).\text{score} = \frac{\sum_{j=1}^{n} Term(i,j)}{Term(i).NumNews}$$

<식-1> 감성단어 극성 추정 식

4. 주가 등락 예측 시스템

본 연구에서 제안하는 주가 등락 예측 시스템은 (그림 3)와 같다. 3 장에서 기술한 감성 사전 자동 구축의 결과로 추출된 감성단어 리스트로부터 하루 동안 발생한 경제 뉴스의 극성 값을 계산한다.



(그림 3) 주가 등락 예측 시스템 구조도

<식-2>는 특정 뉴스에 등장하는 감성단어들의 극성 평균으로 특정 뉴스에 대한 감성 수치 News.Score 를 계산하는 식이며, <식 3>은 하루 동안 발생한 뉴스들로부터 특정일에 대한 감성 수치 DailyNews.score 를 계산하는 과정이다.

m = 특정 뉴스 하나에서 추출한 감성단어의 개수

$$\text{News.score} = \frac{\sum_{i=1}^{m} Term(i).Score}{m} * 100\%$$

<식-2> 특정 뉴스에 대한 감성 수치

k = 하루에 발생한 뉴스의 개수

$$\text{DailyNews.score} = \frac{\sum_{j=1}^{k} News(j).score}{k} * 100\%$$

<식-3> 특정일에 대한 감성 수치

<식-2>와 <식-3>을 통해 계산된 특정일에 대한 감성 수치를 이용하여 주가 등락 예측 모델을 학습하고 전일 대비 당일 종가의 상승, 하락을 예측한다. 주가 등락 예측 모델에 대한 평가 방법으로는 정확도(Accuarcy)를 이용하였으며, <식-4>와 같이 계산된다. 정확도는 주가 등락 예측 모델에서 상승과 하락으로 예측한 결과 중에서 올바르게 예측한 비율로 정의된다.

<식-4>에서 TP(True Positive)는 실제 상승인 것을 예측 모델(Logistic Regression)이 상승으로 분류한 것을, FP(False Positive)는 실제 하락인 것을 예측 모델이 상승으로 분류한 것을 의미한다. 또한, FN(False Negative)는 실제 하락인 것을 예측 모델이 상승으로 분류한 것을, TN(True Negative)는 실제 하락인 것을 예측 모델이 하락으로 분류한 것을 의미한다.

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

<식-4> 정확도 계산 방법

5. 실험 환경

본 제안방법에 대한 평가는 3 장과 4 장에서 제시한 모형에 따라 진행되었다. 데이터 수집을 위해 2010 년~2013 년까지의 경제 뉴스와 일별 KOSPI 지수를 Crawler 를 이용하여 수집하였다.

첫째, 경제 뉴스 수집을 위해 '네이버>증권>뉴스>주요뉴스' 탭에 있는 경제 뉴스 데이터를 제목, 작성날짜, 본문 순으로 수집하였다. 전체 수집된 데이터 총 60,957 건 중 감성 사전 구축을 위해 2010 년~2012 년까지의 뉴스 데이터 50,142 건을 사용하였고 테스트를 위해 2013 년의 뉴스 데이터 10,815 건을 사용하였다.

둘째, 일별 KOSPI 지수 수집을 위해 '한국증권거래소(KRX)>국내지수>일자별 지수' 탭에 있는 일별 KOSPI 를 일자, 종가지수, 등락 순으로 수집하였다. 전체 데이터 총 979 건 중 예측 모델 학습을 위해 2010 년~2012 년까지의 KOSPI 지수 747 일을 사용하였고 테스트를 위해 2013 년의 KOSPI 지수 246 일을 사용하였다. <표-1>은 수집된 데이터에 대한 요약이다.

<표-1> 수집 데이터 요약

항목	기간	뉴스 건 수	KOSPI 일 수
Tranining Set	2010.01 ~ 2012.12	50,142 건	747 일
Test Set	2013.01 ~ 2013.12	10,815 건	246 일

6. 실험 결과

본 연구에서는 서술어 중심의 감성사전을 자동으로 구축하고 Logistic Regression 을 이용하여 주가 등락을 예측하였다. 주가 등락 예측 결과는 <표-2>와 같다.

<표-2> 주가 등락 예측 결과

실제 \ 예측	상승	하락
상승	TP(77)	FN(50)
하락	FP(53)	TN(67)

기존 논문에서 제시하는 명사 중심의 감성사전과 서술어 중심의 감성사전의 구축된 결과를 살펴보면 <표-3>과 같다. 명사 중심의 감성사전은 서술어 중심 사전에 비해 상위/하위 5 개 단어에서 극성을 띄는 단어 보다는 "현대시멘트", "라덴"과 같이 새로운 용어와 인물 혹은 기업에 대한 언급이 많이 나타나고 있다. 이에 비해 서술어 중심의 감성 사전은 "오르다", "번지다"와 같이 극성을 띄기 쉬운 단어들로 구성되어 있다. 이와 같은 서술어 리스트는 등락을 예측하기에 더 적합하다고 할 수 있으며, 예측 정확도를 살펴보면 58.29%로 명사 중심의 감성사전 보다 약 6% 상승의 효과가 있음을 확인할 수 있었다.

<표-3> 명사 중심과 서술어 중심 구축 사전 비교

구분	상위 단어 (5 개)	하위 단어 (5 개)	감성단어 개수	Accuracy
명사 중심	현대시멘트 재할인율 신민당 김정 로켓	라덴 후순위채 사이드채 재선 두바	3626 개	52.69%
서술어 중심 (제안방법)	오르다 상승마감하다 연기되다 천안하다 놓치다	번지다 영업실적 겸손하다 위협하다 적자전환하다	1364 개	58.29%

7. 결론 및 향후 연구

본 연구에서는 경제 뉴스를 이용하여 감성 사전을 구축하고, 구축된 감성사전을 이용하여 주가 등락 예측 모델을 학습하여 전일 대비 당일 종가의 등락을 예측하였다. 기존 연구들은 극성 값이 중립인 경우가 많은 명사를 이용하여 감성 사전을 구축하였지만, 본 연구에서는 극성 값이 잘 표현되는 서술어 중심의 감성 사전을 자동으로 구축하였다. 명사를 이용한 감성 사전과 비교 실험을 수행한 결과 본 연구에서 제시하는 모델의 예측 정확도가 약 6% 상승하는 효과가 있었다.

향후 연구로는 “금리가 오르다”와 같이 “오르다”라는 서술어 자체에 대한 의미보다는 앞 단어인 “금리”와 연관되어 추출되는 의미에 대한 처리 등이 필요하다. 예를 들어 연관 규칙 등을 통해 단어 간 패턴을 찾아 감성사전을 구축한다면, 예측 모델의 정확도를 더 높일 수 있을 것이라 사료된다. 또한, 감성 사전의 극성 값뿐만 아니라 과거의 하락 패턴 및 상승 패턴을 분석하여 예측 정확도를 높이는 연구도 필요하다.

Acknowledgement

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신 · 방송 연구개발사업[13-912-03-003, 사회문제에 관한 도메인 별 이벤트 추출 및 예측 기술 개발], 중소기업청에서 지원하는 2014 년도 산학연협력기술개발사업(No. C0221419) 및 교육과학기술부의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2014030008).

참고문헌

[1] 금융시장 60년, 총 금융자산368배, 주식시장 10배 커진 개방 금융시대 (http://fninside.hyundaicapital.com/486)
[2] R. Schumark and Hsinchun Chen “A quantitative stock prediction system based on financial news” ACM Volume 45 Issue 5, September, 2009 Pages 571-583
[3] 유은지, 김유신 외 "주가지수 상승 예측을 위한 주제지향 감성사전 구축 방안" 한국지능정보시스템학회 2012년 추계학술대회, 2012.12, 42-49 (8 pages)
[4] 안성원, 조성배 “뉴스 텍스트 마이닝과 시계열 분석을 이용한 주가 예측” 한국정보과학회 2010 한국컴퓨터종합학술대회 논문집 제37권 제1호(C), 2010.6, 364-369
[5] R. Schumark and Hsinchun Chen "Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFinText System" ACM Transactions on Information Systems (TOIS), Volume 27, Issue 2, February 2009, Article No. 12
[6] Xiangyu Tang, Chunyu Yang, Jie Zhou "Stock Price Forecasting by Combining News Mining and Time Series Analysis" WI-IAT '09 Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01, Pages 279-282