creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

석사학위 논문

SNS와 뉴스기사의 감성분석과 기계학습을 이용한 주가예측 모형에 관한 연구

A Study on Stock Price Prediction Model Using Sentiment Analysis and Machine Learning Based on SNS and News Articles

2014년 12월

숭실대학교 소프트웨어특성화대학원

소프트웨어전공

김 동 영

석사학위 논문

SNS와 뉴스기사의 감성분석과 기계학습을 이용한 주가예측 모형에 관한 연구

A Study on Stock Price Prediction Model Using Sentiment Analysis and Machine Learning Based on SNS and News Articles

2014년 12월

숭실대학교 소프트웨어특성화대학원

소프트웨어전공

김 동 영

석사학위 논문

SNS와 뉴스기사의 감성분석과 기계학습을 이용한 주가예측 모형에 관한 연구

지도교수 최 재 현

이 논문을 석사학위 논문으로 제출함

2014년 12월

숭실대학교 소프트웨어특성화대학원

소프트웨어전공

김 동 영

김 동 영의 석사학위 논문을 인준함

심 사 위 원 장 _____ 인

심 사 위 원 _____ 인

심 사 위 원 _____ 인

2014년 12월

숭실대학교 소프트웨어특성화대학원

목 차

표 목 차

그 림 목 차

국문초록

SNS와 뉴스기사의 감성분석과 기계학습을 이용한 주가예측 모형에 관한 연구

김동영

소프트웨어전공

숭실대학교 소프트웨어특성화대학원

경제의 발전과 더불어 사람들의 주식시장에 대한 관심이 증가함에 따라 수시로 변하는 주가를 예측하기 위한 다수의 연구들이 진행되고 있다. 여러 예측방법들 가운데 최근 과학기술적 방법을 이용한 많은 연구들이 활발하게 이루어지고 있으며 사용하는 데이터 또한 다양해지고 있다. 이에 본 논문에서는 주가예측의 정확도를 높이기 위해 SNS와 뉴스기사의 감성분석과 기계학습을 이용한 주가예측 모형을 제시하였다. 제시하는 주가예측 모형은 데이터 수집, 감성사전 구축, 감성분석, 기계학습의 4단계 과정을 거쳐서 생성된다. 데이터 수집은 뉴스기사의 경우 경제관련 언론사를 대상으로, SNS의 경우 트위터를 대상으로 데이터를 수집하였다. 수집한 데이터 중 뉴스기사를 분석해 감성사전을 구축하였으며, 이를 활용해 감성분석 과정을 수행하였다. 기계학습단계에서는 감성분석을 통해 생성한 데이터를 이용하였고, 다양한 분류기법으로 예측모형을 생성하였다. 생성한 예측모형의 성능을 측정하기 위해 10중 교차검증 방식을 사용하였으며, 실험결과 다수의 방법에서 80%가 넘는 정확도

를 보였고 0.8에 가까운 F1 score를 나타냈다. 이는 오피니언 마이닝 혹은 데이터 마이닝 기법들을 활용한 기존의 연구들과 비교했을 때 상당히 진전된 결과로 볼 수 있다.

ABSTRACT

A Study on Stock Price Prediction Model Using Sentiment Analysis and Machine Learning Based on SNS and News Articles

KIM, DONGYOUNG
Major in Software
Graduate School of Software Soongsil University

Because people's interest of the stock market has been increased with the development of economy, a lot of studies have been going to predict fluctuation of stock prices. Latterly many studies have been made using scientific and technological method among the various forecasting method, and also data using for study are becoming diverse. So, in this paper we propose stock prices prediction models using sentiment analysis and machine learning based on news articles and SNS data to improve the accuracy of prediction of stock prices. Stock prices prediction models that we propose are generated through the four-step process that contain data collection, sentiment dictionary construction, sentiment analysis, and machine learning. The data have been collected to target newspapers related to economy in the case of

news article and to target twitter in the case of SNS data. Sentiment dictionary was built using news articles among the collected data, and we utilize it to process sentiment analysis. In machine learning phase, we generate prediction models using various techniques of classification and the data that was made through sentiment analysis. After generating prediction models, we performed 10-fold cross-validation to measure the performance of them. The experimental result showed that accuracy is over 80% in a number of ways and F1 score is closer to 0.8. The result can be seen as significantly enhanced result compared with conventional researches utilizing opinion mining or data mining techniques.

제 1 장 서 론

경제의 발전과 더불어 사람들의 주식시장에 대한 관심이 증가함에 따라 수시로 변하는 주가를 예측하기 위한 다수의 연구들이 진행되고 있다. 주가를 예측하기 위한 방법론은 크게 3가지로, 기본적 분석(Fundamental analysis)과 기술적 분석(Technical analysis) 그리고 과학기술적 방법(Technological methods)이 있다. 기본적 분석은 재무 정보나 수익률과 같은 기업의 과거 성과를 평가하여 주가를 예측하는 방법으로 펀드 매니저나 주식 투자자 등이 주로 사용하는 방식이다. 기술적 분석은 과거주식가격의 동향을 파악해 미래주식가격을 예측하는 방식으로 EMA (Exponential Moving Average) 등과 같은 통계적인 기법을 활용한다. 과학기술적 방법은 수많은 연산을 짧은 시간에 처리할 수 있는 컴퓨터 기술과 이와 관련된 이론들의 발전으로 인해 가능하게 된 방법이다. 대표적인 기법으로는 기계학습을 바탕으로 하는 인공신경망, 유전자 알고리즘 등이 있다. 여러 예측방법들 가운데 최근 과학기술적 방법을 이용한 많은 연구들이 활발하게 이루어지고 있다[1][3][4][6][13][14]. 과학기술적 방법을 사용한 연구들에서는 기계학습, 데이터 마이닝, 오피니언 마이닝 등의 다양한 기법들을 이용해 연구를 진행 하였고, 예측에 사용하는 데이터 또한 다양해지고 있다. 이는 빅데이터의 등장으로 인한 것으로 여러 분야에서 기존의 사용하지 않던 데이터를 연구에 이용하기 시작했고, 주가 예측 분야에서도 이러한 현상이 반영된 것이다. 때문에 기존의 주가 예측에 사용하던 기업의 재무정보나 투자정보뿐만 아니라 뉴스기사[2][3][4][10][13]나 SNS[14][15]의 데이터를 이용하는 등 다양한 데이터를 활용하여 주가를 예측하기 위한 연구를 진행하고 있다.

이에 본 논문에서는 주가 예측의 정확도를 높이기 위해 SNS와 뉴스기

사 데이터를 동시에 이용한 다수의 주가예측 모형을 생성한 후 정확성을 비교하는 연구를 진행하였다. 우선 주가예측에 사용될 표본 회사를 추출한 후 일정 기간 동안 표본 회사가 언급된 SNS 데이터와 뉴스기사를 수집한다. 수집한 데이터를 형태소 분석기를 이용해 분해한 후에 감성분석 과정을 거쳐 기사와 SNS 데이터의 긍정지수를 계산한다. 데이터들의 긍정지수와 빈도수를 예측에 사용되는 기법들을 기계학습 시키기 위한 데이터로 사용하여 예측 모형을 생성한다. 기계학습을 통해 생성된 다수의 예측 모형들과 주가와의 관계를 통계적인 기법을 이용해 분석하고 모형들 간의 예측 정확도를 비교하여 연구를 진행한다.

제 2 장 관련연구

2.1 감성분석

감성분석이란 자연어 처리와 텍스트 분석, 전산언어학 등을 이용해 텍스트 내에서 주관적인 정보를 확인하고 추출하는 기법으로 '오피니언 마이닝'이라고도 한다. 감성분석의 기본 작업은 텍스트의 극성을 긍정, 부정, 중립 등으로 분류하는 것이다. 하지만 본 논문에서는 감성분석을 주가예측에 직접적으로 이용하지 않고 데이터 마이닝의 예측기법들을 기계학습 시키기 위한 데이터를 생성하기 위해 사용하고자 한다. 때문에 감성분석의 과정 중 형태소 분석기를 사용해 텍스트를 분해한 후 감성사전과의 비교를 통해 텍스트의 긍정지수를 도출하는 과정까지만 진행하였고 전체 텍스트의 극성을 분류하는 과정은 생략하였다.

2.2 형태소 분석

형태소는 언어의 형태론적 수준에서의 최소단위를 뜻하며, 형태소 분석이란 문장의 어절을 형태소단위로 분리한 다음 각 형태소에 맞는 범주를 부여하는 과정으로 정의할 수 있다[9]. 본 논문에서 필요한 데이터를 추출하기 위해서는 수집한 뉴스기사와 SNS 데이터를 형태소 단위로 분석하는 과정을 수행해야한다. 본 연구에서는 서울대학교 IDS 연구실에서 개발한 꼬꼬마 형태소 분석기를 사용하여 수집한 데이터들의 형태소 분석을 진행하였다.

2.3 감성사전

형태소 분석을 통해 분리된 형태소의 극성을 분류하기 위해서는 개별 언어의 내용을 대상으로 단어 또는 단어의 의미 수준에서 긍정, 부정, 중립의 평가를 해놓은 감성사전을 사용한다[11]. [7]에서는 도메인의 특징을 고려하여 구축한 감성사전을 이용하는 것이 감성 평가의 정확도를 향상시킴을 나타내었고, 이를 바탕으로 [11]에서는 주식 도메인에 특화된 감성사전을 구축하고 활용하는 방안을 제시하였다. 본 연구에서는 [11]의 방법을 바탕으로 주식시장에 맞는 감성사전을 구축하였고 이를 이용해 뉴스와 SNS 데이터의 감성분석을 실시하였다.

2.4 주가예측

지금까지 주가를 예측하기 위한 다수의 연구들이 진행되어 왔다. [8]에서는 뉴스의 발생이 주가에 유의한 영향을 미침을 나타내었고, 이를 바탕으로 [4]에서는 뉴스 콘텐츠를 감성분석하여 주가 지수를 예측하는 투자의사결정 모형을 제시하였다. 제시한 모형과 주가 지수간의 관계를 통계기법을 이용하여 분석한 결과 통계적으로 유의한 관계를 가지는 것으로 확인되었다. 또한 [14]에서는 SNS의 한 종류인 트위터 사용자들의 데이터를 감성분석하여 주가 지수와의 연관성에 대한 실증적인 연구를 진행하였다. 사용자들의 데이터를 감성분석하기 위해 사용한 2가지 툴 중 하나인 Google-Profile of Mood States는 사용자의 감성을 6개의 수준(Calm, Alert, Sure, Vital, Kind, and Happy)으로 분류하였고, 이중 Calm으로 분류된 감정상태의 변화가 주가 지수의 변동과 유사한 흐름을 나타내고 있음을 발견하였다.

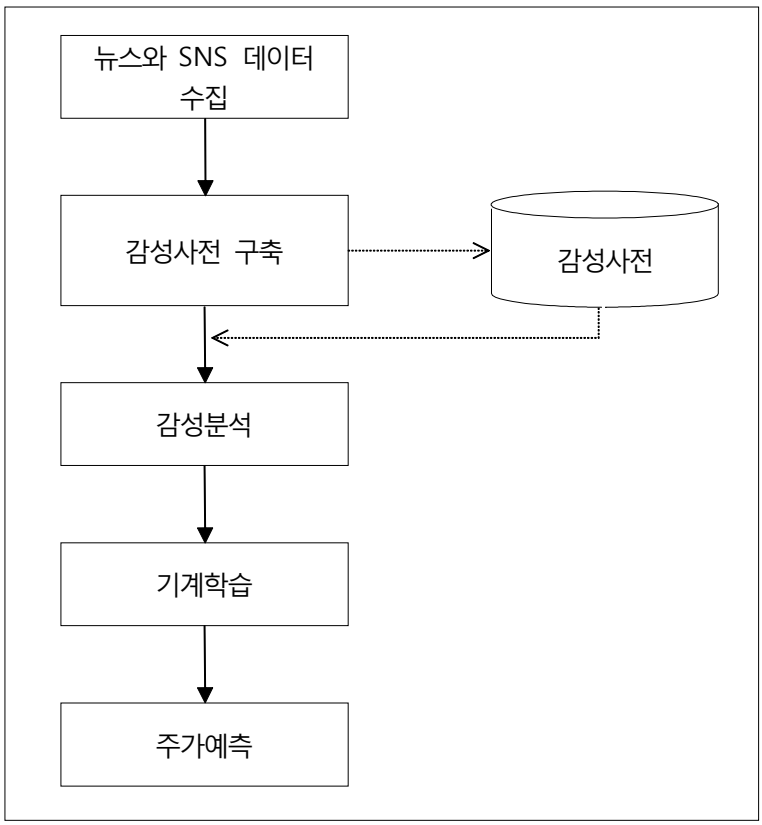
이에 본 논문에서는 뉴스기사와 SNS 사용자들의 데이터 모두를 주가 예측을 위한 데이터로 사용하여 예측의 정확도를 높이고자 한다. 또한

오피니언 마이닝을 통해 뉴스나 SNS에 담겨있는 의미를 파악하여 주가 예측을 하는 방법[3][4]이 아닌, 오피니언 마이닝의 과정 중에 발생하는 데이터를 이용해 기계학습의 과정을 거쳐 주가를 예측하고자 한다. 또한 기존의 연구들과는 다르게 주가지수의 등락을 예측하는 것이 아닌 개별 기업 주가의 등락을 다룬다. 이를 위해 본 연구에서는 KOSPI 상장 기업 중 7개의 회사를 표본으로 추출하여 실험을 진행하였다.

제 3 장 주가예측 모형

3.1 감성분석과 기계학습을 통한 주가예측 모형

본 논문에서는 주가예측의 정확도를 높이고자 기존의 연구에서 많이 사용되어오던 뉴스기사와 최근 많이 대두가 되고 있는 SNS의 데이터를 동시에 이용하였다. 또한 감성분석 기법을 적용하여 뉴스와 SNS의 데이터를 정제하였고, 이를 데이터 마이닝 분야의 여러 예측기법들을 기계학습 시키는 데에 사용하여 주가예측 모형을 생성하였다.

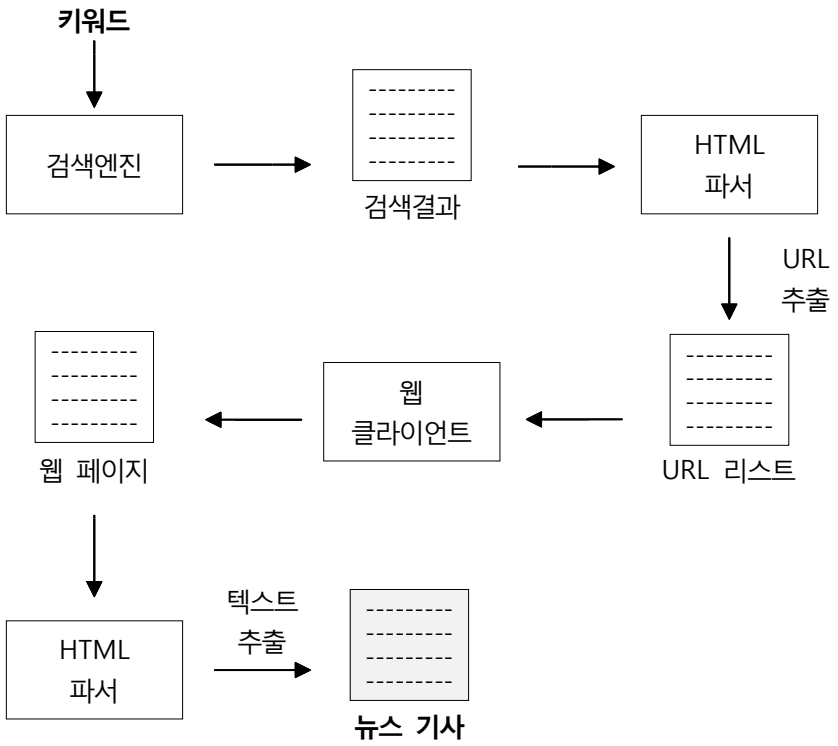


[그림 3-1] 주가예측 모형

제시하는 주가예측 모형은 [그림 3-1]과 같으며, 데이터 수집, 감성사전 구축, 감성분석, 기계학습의 4단계를 거쳐 생성된다. 데이터 수집단계에서는 일정기간동안 해당 기업과 연관되어 발생한 뉴스기사와 SNS 데이터를 수집한다. 다음단계에서는 수집한 데이터 중 뉴스기사를 분석하여 도메인에 맞는 감성사전을 구축한다. 감성분석 단계에서는 구축한 감성사전으로 수집한 데이터의 감성분석을 수행하여 기계학습을 위한 데이터를 생성한다. 기계학습 단계에서는 감성분석과정을 통해 생성된 데이터를 기반으로 기계학습 과정을 거쳐 최종적으로 예측모형을 생성한다.

3.2 데이터 수집

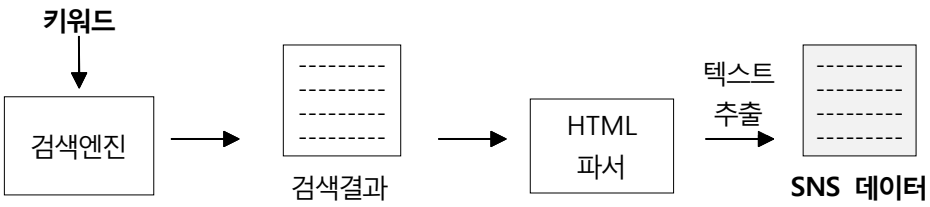
뉴스기사와 SNS 데이터는 다양한 경로를 통해 대량으로 생산된다. 때문에 이를 연구에 이용하기 위해서는 자동화된 방식에 의해 필요한 정보만을 수집하여 저장해야 한다. 이를 위해 본 연구에서는 뉴스기사와 SNS 데이터를 수집하는 프로그램과 데이터베이스를 구현하였고, 이를 이용해 자동화된 방식으로 데이터를 정제하여 수집하였다.



[그림 3-2] 뉴스기사 수집 과정

뉴스기사와 SNS 데이터의 수집과정은 [그림 3-2]와 [그림 3-3]에 각각 나타내었다. 수많은 뉴스기사와 SNS 데이터들 중 필요한 데이터를

수집하기 위해 검색엔진을 이용하여 해당 키워드가 들어간 데이터만을 추출한다. 뉴스기사의 경우 검색결과에 기사 전문이 표시되지 않기 때문에 HTML 파서를 이용해 검색결과에서 해당기사의 원문이 게시되어있는 웹페이지의 URL을 추출한다. 추출한 URL을 통해 웹페이지에 접근한 후 다시 한 번 HTML Parsing 과정을 거쳐 뉴스기사를 수집한다.



[그림 3-3] SNS 데이터 수집 과정

SNS의 경우 게시하는 글의 글자 수 제약이 있기 때문에 검색결과상에 글의 전문이 표시되어있다. 때문에 뉴스기사와는 다르게 검색결과에서 직접 텍스트를 추출해 데이터를 수집한다. 수집한 뉴스기사와 SNS 데이터는 데이터베이스에 저장하여 관리하며 형태소 분석과 감성분석 과정을 거쳐 주가를 예측하기 위한 데이터로 활용한다.

3.3 형태소 분석

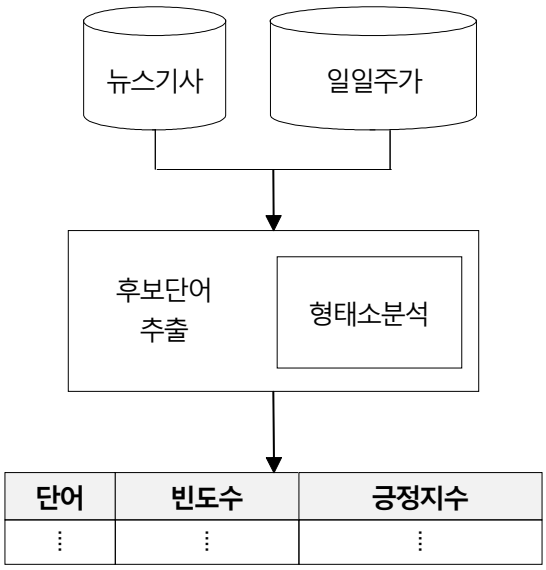
텍스트로부터 작성자의 감정이나 의견을 추출하기 위해선 텍스트를 형태소단위로 분리하여 각 형태소별 극성을 파악한 후 전체 텍스트의 극성을 분류하는 방식을 사용한다. 본 논문에서는 서울대학교 IDS 연구실에서 개발한 '꼬꼬마 형태소 분석기'를 사용하여 텍스트의 형태소를 분석하였다. 꼬꼬마 형태소 분석기는 한글 형태소의 품사를 '체언, 용언, 부사, 관형사, 조사, 감탄사, 어미, 어근, 접사, 부호, 한글 이외'의 항목으로 나누고 세부 품사를 구분 하였다. 이중 체언의 경우 명사와 수사, 대명사로 구분하였고 명사는 다시 세부적으로 보통명사, 고유명사, 일반 의존 명사, 단위 의존 명사로 구분하였다. 본 연구에서는 [표 3-1]에서 표시되어 있는 바와 같이 여러 품사 중 체언에 속하며 실질적 개념을 표시하는 명사에 해당하는 형태소들을 추출하여 연구에 활용하였다.

[표 3-1] 꼬꼬마 형태소 분석기 체언 태그표

대분류	묶음1	묶음2	태그	설명
체언	N	NN	**NNG**	**보통 명사**
			NNP	**고유 명사**
			NNB	**일반 의존 명사**
			NNM	**단위 의존 명사**
		NR	NR	수사
		NP	NP	대명사

3.4 감성사전 구축

감성분석 과정에서 극성을 분류하기 위해 사용하는 감성사전은 분석의 정확도를 높이는 데 있어 매우 큰 비중을 차지한다. 하지만 현재 한글의 감성분석을 위한 범용 감성사전이 제대로 구축되어 있지 않다. 그 뿐만 아니라 일반적인 단어들의 경우 주가의 변동을 제대로 반영하기에 적절하지 않다고 판단하여 주식시장에 특화된 감성사전을 구축하여 연구를 진행하였다. 감성사전의 구축과정은 [그림 3-4]와 같으며, 이전 과정에서 수집한 뉴스기사와 일별 주가 변동 데이터를 이용한다. 우선 수집한 뉴스기사의 형태소를 분석하여 감성사전에 실릴 후보 단어들을 추출하고 단어들의 빈도수와 긍정 값을 계산한다. 빈도수는 해당 단어가 나온 기사의 수를 합산하여 계산하고, 긍정 값은 해당 단어가 들어간 기사가 게재되었을 때 익일 주가(Next day Stock Price, NSP)가 상승한 경우의 수를 합산하여 계산한다.



[그림 3-4] 감성사전 구축을 위한 후보단어 추출

빈도수(frequency)와 긍정 값(positive)을 수식으로 표현하면 식 (2), (4)와 같다.

$$include(i,j) = \begin{cases} 1 & (\text{기사 } j \text{ 에 단어 } i \text{ 가 포함된 경우}) \\ 0 & (\text{그 외의 경우}) \end{cases} \quad (1)$$

$$frequency(i) = \sum_{j=1}^{n} include(i,j),\ n = \text{전체 기사의 수} \quad (2)$$

$$NSP(j) = \begin{cases} 1 & (\text{기사 } j \text{ 가 게재된 후 익일 주가가 상승한 경우}) \\ 0 & (\text{그 외의 경우}) \end{cases} \quad (3)$$

$$positive(i) = \sum_{j=1}^{n} \{include(i,j) \times NSP(j)\},\ n = \text{전체 기사의 수} \quad (4)$$

추출한 후보단어들 중 상대적으로 빈도수가 너무 작은 단어들은 주가 변동을 제대로 반영할 수 없다고 판단하여 평균 빈도수 이하의 단어들은 제거하였다. 마지막으로 추출한 단어들의 긍정지수를 계산하여 감성사전을 완성한다. 긍정지수는 긍정 값을 빈도수로 나누어 나타내며, 식으로 표현하면 식 (5)와 같다.

$$P(i) = \frac{\sum_{j=1}^{n} \{include(i,j) \times NSP(j)\}}{\sum_{j=1}^{n} include(i,j)},\ n = \text{전체 기사의 수} \quad (5)$$

본 논문에서 구축한 감성사전은 단어와 긍정지수의 두 가지 속성으로 이루어져 있다. 단어는 명사만을 추출하여 구성하였고, 긍정지수는 0에서 1사이의 값으로 1에 가까울수록 긍정의 의미를 나타낸다.

3.5 감성분석

구축한 감성사전을 이용하여 뉴스기사와 SNS 데이터의 감성분석을 [그림 3-5]와 같은 과정을 거쳐 진행한다. 우선 수집한 데이터의 형태소를 분석하여 명사를 추출한 후 추출한 명사와 감성사전의 단어들을 비교해 해당 텍스트의 긍정지수를 계산한다. 텍스트의 긍정지수(Positive index of Text, PT)는 해당 텍스트에서 추출한 명사들의 긍정지수를 합해 그 개수로 나눈 산술평균값으로 나타내며 수식으로 표현하면 식 (7)과 같다.

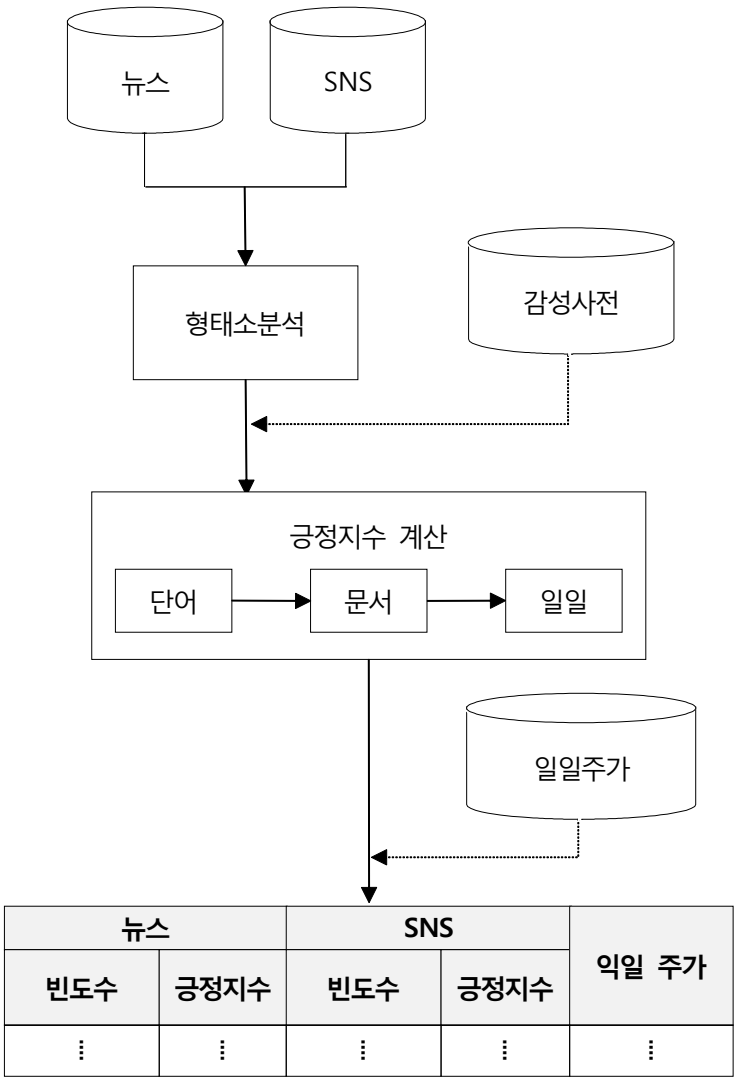
$$match(i,j)=\begin{cases} 1 & (\text{텍스트 } i \text{ 에 포함된 명사 } j \text{ 가 감성사전에 존재 할 경우}) \\ 0 & (\text{그 외의 경우}) \end{cases} \quad (6)$$

$$PT(i)=\frac{\sum_{j=1}^{n}\{match(i,j)\times P(j)\}}{\sum_{j=1}^{n}match(i,j)},\ n=\text{텍스트 } i \text{ 에 포함된 단어의 수} \quad (7)$$

텍스트의 긍정지수를 계산한 후 일별 긍정지수를 계산한다. 일별 긍정지수(Daily Positive index, DP)는 해당일자에 게재된 텍스트들의 긍정지수를 합해 그 개수로 나눈 산술평균값으로 나타낸다. 일별 긍정지수를 수식으로 표현하면 식 (8)과 같다.

$$DP(k)=\frac{\sum_{i=1}^{n}PT(i)}{n},\ n=k \text{ 일에 게재된 텍스트의 수} \quad (8)$$

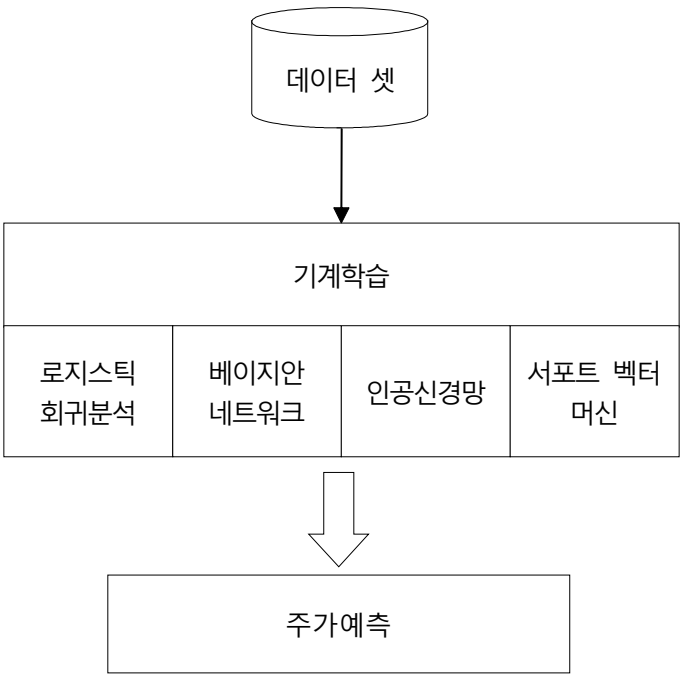
최종적으로 일별 뉴스기사의 빈도수와 긍정지수, SNS의 빈도수와 긍정지수를 도출한 후 익일 주가변동 항목을 추가해 기계학습을 위한 데이터로 사용한다. 익일 주가변동 항목은 주식가격의 오르고 내림의 정도를 표시하지 않고 오르고 내림의 여부를 표현하기 위해 up, 0, down 세 가지로 나타낸다.



[그림 3-5] 감성분석 과정

3.6 기계학습

기계학습을 위한 데이터는 속성 집합과 클래스 레이블로 이루어져 있으며, 기계학습은 속성 집합을 미리 정해진 클래스 레이블에 사상하는 목표함수를 학습하는 작업이다. 감성분석을 통해 생성된 데이터 중 뉴스기사의 빈도수와 긍정지수, SNS의 빈도수와 긍정지수는 속성 집합에 해당하며 익일주가 항목은 클래스 레이블에 해당한다.



[그림 3-6] 기계학습 흐름

본 연구에서는 [그림 3-6]에서 나타낸 바와 같이 기계학습이론을 바탕으로 하는 분류기법들을 사용하여 예측 모형을 생성하였고, 여러 분류기법들 중 로지스틱 회귀분석, 베이지안 네트워크, 인공신경망, 서포트 벡터 머신 방법을 이용하였다. 회귀분석이란 종속변수와 독립변수 사이의

상관관계를 설명하는 선형 관계식을 구하는 기법으로, 로지스틱 회귀분석은 종속변수의 형태가 범주형일 경우 사용한다. 베이지안 분류기는 베이즈 이론을 기반으로 주어진 속성집합이 특정 클래스에 속할 확률을 예측하는 분류기로, 베이지안 네트워크는 베이지안 분류기에 그래픽 이론을 결합하여 속성집합의 부분집합들이 가지는 종속관계를 표현할 수 있도록 한다. 인공신경망은 뇌기능의 특성을 컴퓨터 시뮬레이션으로 표현하여 문제 해결 능력을 가지도록 하는 수학 모델이다. 시냅스와 뉴런의 네트워크 형태로 이루어져 있으며, 뉴런의 학습을 통해 시냅스의 결합세기를 변화시켜 모델을 생성한다. 생성된 모델은 입력 신호를 받아 뉴런과 시냅스를 거쳐 하나의 출력신호를 생성하는 형태를 이루고 있다. 서포트 벡터 머신은 주어진 자료들을 분리하는 최대 마진 초평면을 찾는 방법으로, 초평면을 기준으로 속성집합을 특정 클래스로 분류한다.

4. 실험 및 결과

4.1 실험설계

실험은 KOSPI 상장사 중 7개의 기업을 대상으로 진행하였으며, 2013년 1월 2일부터 2013년 12월 30일까지의 뉴스기사와 SNS 데이터, 주가변동 데이터를 수집하여 실험에 활용하였다. 실험은 3장에서 제시한 주가예측 모형에 따라 데이터수집, 감성사전 구축, 감성분석, 기계학습 순으로 진행하였다.

4.2 데이터 수집

실험에 필요한 데이터는 자체적으로 제작한 프로그램과 데이터베이스를 사용하여 수집하였다. 뉴스기사의 경우 경제관련 기사를 중점적으로 다루는 11개의 언론사에서 경제 섹션에 게재한 기사들을 대상으로 검색하였고, 이 중 해당기업과 관련된 기사만을 추출하여 총 132,123건의 기사를 수집하였다.

SNS 데이터의 경우 다양한 SNS들 중 트위터에 게재된 트윗들을 검색 대상으로 하였으며, 이 중 해당기업과 관련된 트윗들을 추출하여 총 228,260건의 데이터를 수집하였다. 수집한 뉴스기사와 트윗의 일부를 [그림 4-1]에 나타내었고, 기업별로 수집한 데이터의 수를 [표 4-1]을 통하여 표현하였다.

date	article
2013-01-16 오후 11:55	[삼성전자, 네덜란드서 승소.."애플, 소송비 부담"] [뉴욕= 이데일리 이정훈 특파원] 네...
2013-01-16 오후 5:42	[특허괴물 표적 된 삼성·LG]작년 소송건수 2·3위… 국내 IT업체에 공세 집중삼성전자와...
2013-01-16 오후 5:30	["기온 21도 넘어가면 휴가 마케팅 시작"…'빅데이터 경영' 배운 삼성 사장단]"빅데이터...
2013-01-16 오후 5:24	[삼성전자 고효율 냉장고 인도서 CDM 사업 승인]삼성전자가 최근 유엔기후변화협약(UNFC...
2013-01-16 오후 5:13	[삼성-LG 냉장고 싸움에 소비자 "크기 상관없다" 냉담]국내 가전회사업계의 맞수인 삼성전...
2013-01-16 오후 3:51	[삼성전자, 파운드리 3위 도약 "Thank you 애플"] [머니투데이 서명훈 기자] [IC...
2013-01-16 오후 2:49	[삼성 냉장고 '유엔 청정개발체제 사업' 승인]삼성전자가 유엔기후변화협약(UNFCCC) 사...
2013-01-16 오후 2:22	[[특징주]삼성전자, 장중 150만원 하회] [아시아경제 송화정 기자]가 외국인 매도공세에 ...
2013-01-16 오전 11:16	["1억짜리 TV도 살 사람 많다"…삼성, 中 · 중동 큰손 공략]7년 연속 세계 TV 시...

date	tweet
2013-01-02	[SMNR] 삼성전자가 세계 최대 가전 전시회 CES 2013에서 세계 최초로 '에볼루션 ...
2013-01-02	삼성전자 1,535,000
2013-01-02	삼성전자 153만원~ ㅠㅜ
2013-01-02	"애플, TSMC에 차세대 프로세서 생산 이미 의뢰" : 대만 언론 "시장 예측보다 이른 ...
2013-01-02	'특허 괴물' 인터디지털 미국 국제무역위원회에 삼성전자를 3G,4G 통신 특허 침해 혐의로...
2013-01-02	이건희 삼성전자 회장은 앞만 보고 열심히 달리겠다면서 지금까지의 성공을 잊고 새로 도전해 ...
2013-01-02	애플, 모바일 기기 핵심부품인 애플리케이션 프로세서(AP)의 공급선을 삼성전자에서 대만의 ...
2013-01-02	[SMNR] 삼성전자가 1월 2일 구석 청소능력을 향상시킨 로봇청소기 '스마트 탱고 코너클...
2013-01-02	웹기획/마케팅/모바일웹/앱개발/트위터/페이스북/미디어/패밀리삼성/삼성전자/.. [트친찾기]...
2013-01-02	새해 첫날 코스피 2,030…삼성전자 157만원 fb.me/sV1WYxVN

[그림 4-1] 수집한 뉴스기사와 트윗 일부

[표 4-1] 기업별 수집 데이터 수

기업	뉴스 데이터 수	SNS 데이터 수
A	41,343	96,146
B	28,410	20,714
C	10,879	6,561
D	15,214	26,851
E	16,923	34,076
F	7,136	4,141
G	12,218	39,771

4.3 감성사전 구축

기업별로 수집한 뉴스기사와 주가변동 데이터를 이용하여 감성분석에 필요한 감성사전을 각각 구축한다. 우선 뉴스기사를 형태소분석하여 명사를 추출한 후 그 빈도수를 계산하여 후보단어들을 도출한다. 이 중 뉴스에 포함 된 해당기업명, 언론사명 등 불필요한 단어와 평균 빈도수 이하의 단어들을 제거한 후, 추출한 단어들의 빈도수와 긍정 값으로 긍정지수를 계산하여 감성사전의 구축을 마무리한다. 감성사전의 일부를 [그림 4-2]에 나타냈으며, [표 4-2]에 기업별 감성사전의 단어 수를 나타내었다.

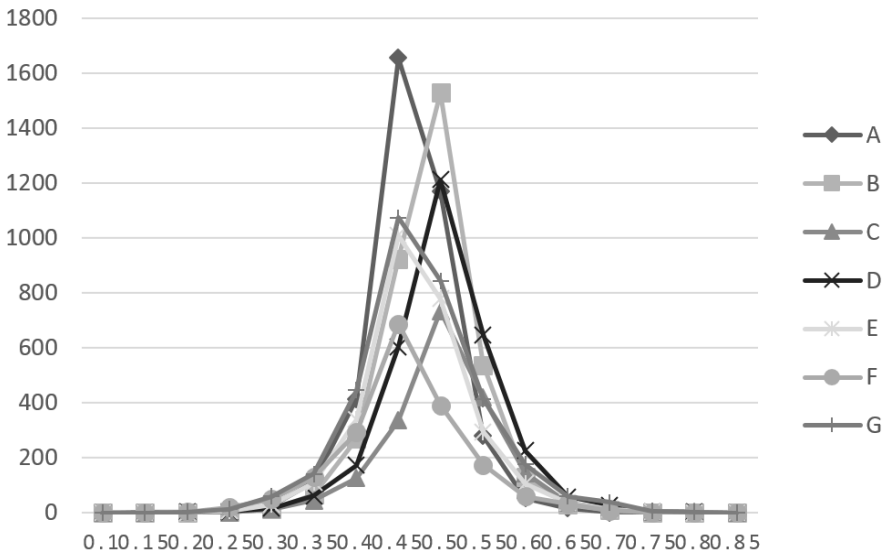
word	positive_index	word	positive_index
절벽	0.147321	패소	0.195122
민사	0.203252	증발	0.222222
시동	0.307692	촉구	0.308333
투자	0.442915	전자	0.446831
증권	0.449651	상승	0.451437
국산화	0.514793	테크	0.514811
기어	0.649275	회장단	0.626667
시리아	0.746032	베를린	0.710462

[그림 4-2] 감성사전 일부

감성사전의 긍정지수는 0에서 1사이의 값으로 표현되며 1에 가까울수록 긍정의 의미를 나타낸다. 실험을 통해 구축한 각 기업 별 감성사전의 긍정지수 분포를 [그림 4-3]에 표현하였다. 긍정지수의 분포는 기업별로 근소한 차이를 보이고 있지만 0.4에서 0.5사이 값에 집중되어 있는 거의 동일 한 형태를 이루고 있다.

[표 4-2] 기업별 감성사전 단어 수

기업	후보 단어 수	감성사전 단어 수
A	31,789	3,740
B	29,108	3,513
C	16,541	1,884
D	24,246	3,041
E	22,490	2,730
F	15,429	1,851
G	22,966	3,277



[그림 4-3] 기업별 감성사전 긍정지수 분포

4.4 감성분석

구축한 감성사전을 이용하여 각 기업별로 수집한 뉴스기사와 트윗들의 감성분석을 진행한다. 감성분석과정을 거쳐 생성된 데이터 셋은 일별 뉴스기사의 빈도수와 긍정지수, 일별 트윗의 빈도수와 긍정지수, 익일 주가변동 항목으로 구성되며, [표 4-3]에 데이터 일부를 나타내었다.

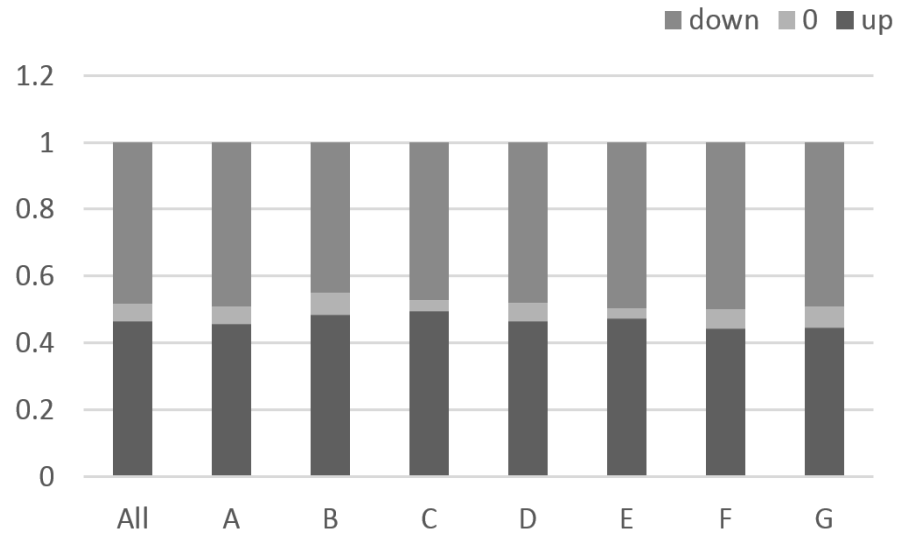
[표 4-3] 데이터 셋 일부

속성 집합				클래스 레이블
뉴스		SNS		익일주가
빈도수	긍정지수	빈도수	긍정지수	
117	0.442	369	0.438	up
183	0.440	871	0.424	down
212	0.440	542	0.430	down
265	0.440	651	0.435	0
260	0.445	492	0.435	up

수집한 뉴스기사와 트윗들은 매일 발생하지만 주식시장은 매일 개장하는 것이 아니기 때문에 익일 주가변동 데이터가 존재하지 않는 날이 발생하게 된다. 때문에 익일 주가변동 데이터가 존재하지 않을 경우 익일 주가변동 데이터가 존재할 때까지 데이터들을 합산하여 데이터 셋을 생성하였다. 예를 들어 금요일에 게재된 뉴스기사의 경우 다음날인 토요일의 주가에 영향을 미친다고 가정했지만 토요일엔 주식시장이 개장되지 않는다. 때문에 금요일에 게재된 뉴스기사는 토요일과 일요일의 기사들

과 같이 월요일 주가에 영향을 미친다고 가정하고 이들 데이터를 합산하여 하나의 데이터로 구성하였다.

2013년 한 해 동안 주식시장은 247일 개장되었고, 이 중 1월 2일을 제외한 246일에 대한 감성분석을 진행하였다. 때문에 감성분석 과정을 거쳐 생성된 대부분의 데이터 셋은 246개의 데이터로 구성되어 있지만, G 기업의 경우 경제관련 뉴스기사나 트위터 데이터가 발생하지 않은 날들이 포함되어 있어 이를 제외한 238개의 데이터로 구성하였다. 또한 일일 감성지수를 계산하는 과정에서 텍스트의 감성지수가 0인 데이터들은 제외하고 일일 감성지수를 계산하여 데이터를 생성하였다. 생성한 전체 데이터 중 익일주가가 상승한 경우는 798개, 하락한 경우는 828개, 전날과 동일한 경우는 88개로 나타났고, 전체 데이터와 기업별 데이터의 분포를 [그림 4-4]에 표현하였다. 주가가 전날과 동일한 경우의 데이터 빈도수가 비교적 적지만 주가가 상승한 경우와 하락한 경우의 데이터 수는 한쪽에 치우치지 않게 분포되어 있는 것으로 판단된다.



[그림 4-4] 클래스 레이블별 데이터 분포

4.5 기계학습

본 논문에서는 기계학습 과정을 거쳐 예측모형의 성능을 측정하기 위해 뉴질랜드 와이카토 대학에서 개발한 오픈소스 소프트웨어인 'WEKA'를 사용하였다. 2013년 1월 2일부터 9월 1일까지의 데이터는 예측 모형의 성능을 검증하기 위한 데이터(training set, test set)로 사용하였고, 9월 2일부터 12월 30일까지의 데이터는 예측 모형간의 성능을 비교하기 위한 데이터(validation set)로 사용하였다. 예측 모형의 성능을 검증하기 위한 데이터를 학습 데이터 셋과 평가 데이터 셋으로 나누어 기계학습을 수행하기에는 충분하지 않다고 판단하여 교차 검증으로 예측모형의 성능을 측정하였다[12]. 분류기법은 로지스틱 회귀분석, 베이지안 네트워크, 인공신경망, 서포트 벡터 머신 총 4가지 방법을 이용하였고, 성능 측정은 10중 교차 검증(10-fold cross-validation) 방식을 사용하였다.

분류기법 중 서포트 벡터 머신은 주어진 데이터가 선형으로 분류되는 경우와 분류되지 않을 경우 적용하는 기법이 다르다. 본 논문에서 사용한 데이터는 선형으로 분류되지 않는 경우에 속하여 비선형 서포트 벡터 머신을 사용하였다. 비선형 서포트 벡터 머신의 경우 커널함수를 이용하여 기존 데이터를 고차원 공간으로 변환한 후에, 선형 서포트 벡터 머신의 공식화를 통해 2차 계획법 문제로 표현하여 최대 마진 초평면을 찾을 수 있다. 본 연구에서는 서포트 벡터 머신을 학습하는 과정에서 SMO(Sequential Minimal Optimization) 알고리즘과 Polynomial kernel을 사용하였다. 2차 계획법 문제는 라그랑즈 승수 기법을 이용해 풀 수 있는데, 이때 사용하는 라그랑지 계수 α의 상한 값을 사용자가 지정해 주어야 한다. 'WEKA'에서는 이를 Complexity 파라미터로 지정해 두었다. Complexity 파라미터는 1부터 10까지의 수 중 최적의 성능을 내는 파라미터를 시행착오법을 통해 결정하였다. 인공신경망의 경우 다층퍼셉

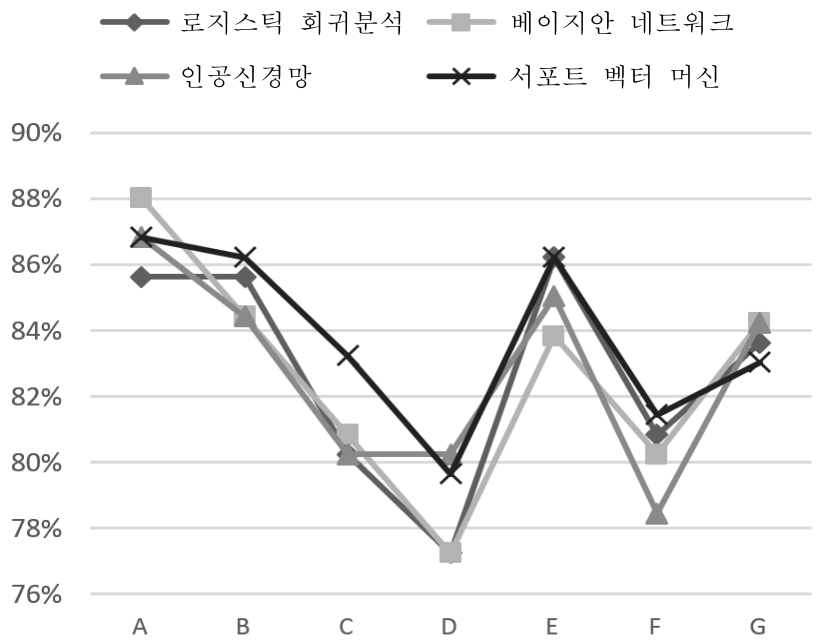
트론을 이용하였으며 이때 사용하는 은닉층의 수를 사용자가 지정해 주어야 한다. 다층퍼셉트론은 4개의 입력노드와 3개의 출력노드를 사용하였으며 은닉층의 수는 서포트 벡터 머신의 경우와 같이 시행착오법을 통해 결정하였다. 각 기업별 Complexity 파라미터와 은닉층의 수는 [표 4-4]에 나타내었다.

[표 4-4] 기업별 설정 파라미터

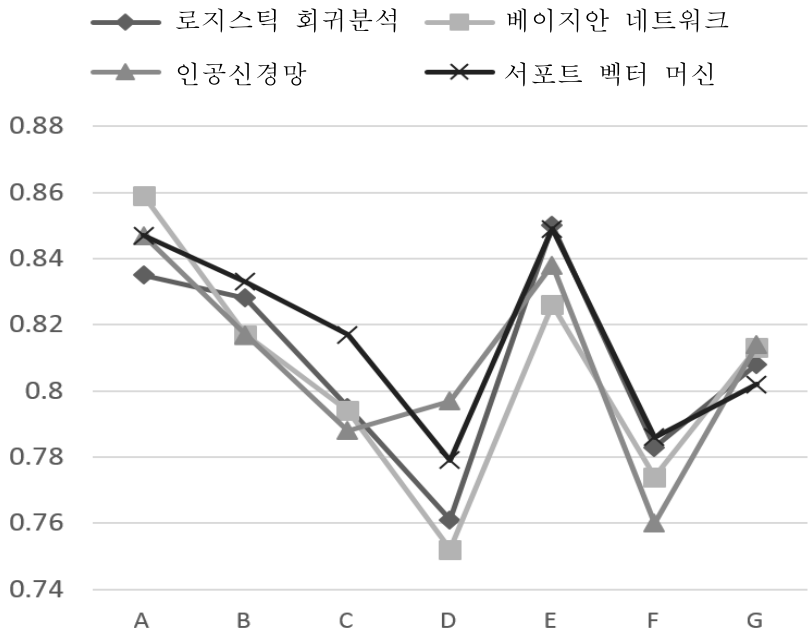
기업	Complexity 파라미터	은닉 층 수
A	3	2
B	6	1
C	2	1
D	5	5
E	9	1
F	3	9
G	1	1

4.6 결과 분석

학습 데이터 셋과 평가 데이터 셋을 이용한 10중 교차 검증 실험 결과 7개의 기업 모두 다수의 방법에서 80%가 넘는 정확도를 보였고 0.8에 가까운 F1 score를 나타내었다. 또한 예측 모형들의 정확도와 F1 score의 산술평균값을 비교한 결과 서포트 벡터 머신을 이용하였을 때 정확도 83.8%와 F1 scroe 0.8161로 가장 좋은 성능을 보였다. 기업별 정확도와 F1 score는 [그림 4-5]와 [그림 4-6]에 나타내었고 수치화된 결과를 [표 4-5]에 나타내었다.



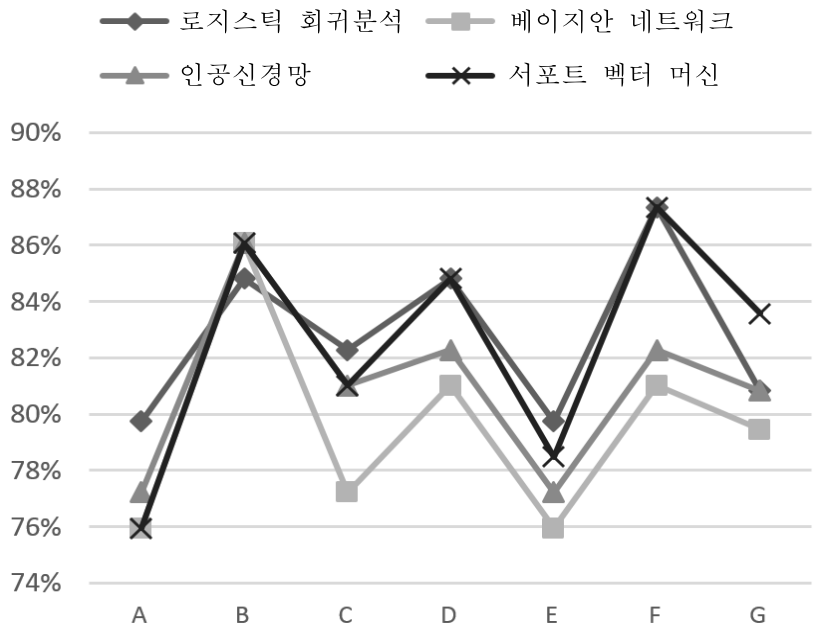
[그림 4-5] 10중 교차검증 정확도



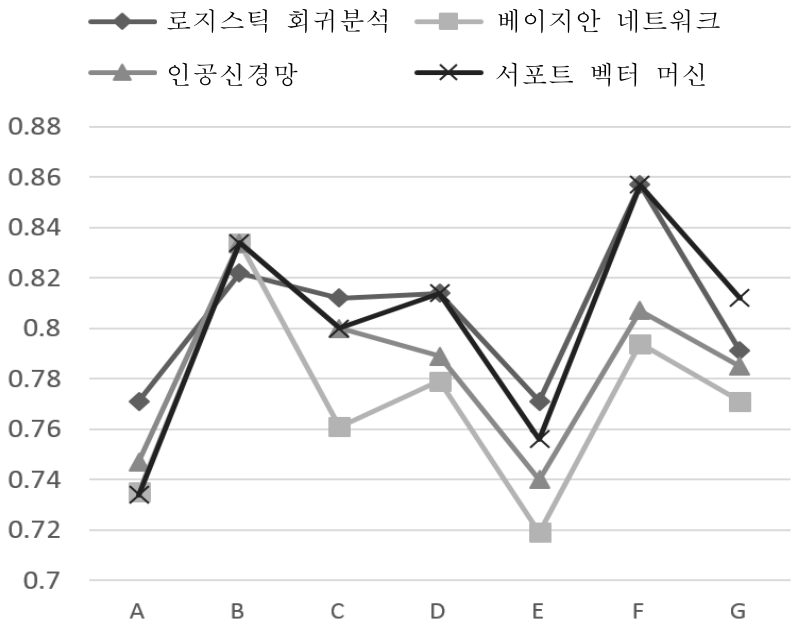
[그림 4-6] 10중 교차검증 F1 Score

검증 데이터 셋을 이용한 실험 결과 5개의 기업(B, C, D, F, G)에서는 다수의 방법에서 80%가 넘는 정확도와 0.8에 가까운 F1 score를 보였고, 2개의 기업(A, E)에서는 75%이상의 정확도와 0.75에 가까운 F1 score를 나타내었다. 또한 예측 모형들의 정확도와 F1 score의 산술평균값을 비교한 결과 로지스틱 회귀분석(정확도-82.7%, F1 score-0.805)과 서포트 벡터 머신(정확도-82.4%, F1 score-0.801)을 이용하였을 때 비슷한 수준으로 좋은 성능을 보였다. 검증 데이터 셋을 이용한 실험의 기업별 정확도와 F1 score는 [그림 4-7]와 [그림 4-8]에 나타내었고 수치화된 결과를 [표 4-6]에 나타내었다.

본 논문의 실험결과는 주가예측을 위해 뉴스 콘텐츠와 오피니언 마이닝을 이용하거나[3][13], 데이터 마이닝 기법들을 이용한 기존 연구들[5][6][10]의 결과와 비교했을 때 상당히 진전된 결과로 볼 수 있다.



[그림 4-7] 검증 데이터 셋 정확도



[그림 4-8] 검증 데이터 셋 F1 Score

[표 4-5] 10중 교차검증 결과

분류기	척도	A	B	C	D	E	F	G
로지스틱 회귀분석	정확도(%)	85.6287	85.6287	80.2395	77.2455	86.2275	80.8383	83.6364
	F1 score	0.835	0.828	0.795	0.761	0.85	0.783	0.808
베이지안 네트워크	정확도(%)	88.024	84.4311	80.8383	77.2455	83.8323	80.2395	84.2424
	F1 score	0.859	0.817	0.794	0.752	0.826	0.774	0.813
인공신경망	정확도(%)	86.8263	84.4311	80.2395	80.2395	85.0299	78.4431	84.2424
	F1 score	0.847	0.817	0.788	0.797	0.838	0.76	0.814
서포트 벡터 머신	정확도(%)	86.8263	86.2275	83.2335	79.6407	86.2275	81.4371	83.0303
	F1 score	0.847	0.833	0.817	0.779	0.849	0.786	0.802

[표 4-6] 검증 데이터 셋 결과

분류기	척도	A	B	C	D	E	F	G
로지스틱 회귀분석	정확도(%)	79.7468	84.8101	82.2785	84.8101	79.7468	87.3418	80.8219
	F1 score	0.771	0.822	0.812	0.814	0.771	0.857	0.791
베이지안 네트워크	정확도(%)	75.9494	86.0759	77.2152	81.0127	75.9494	81.0127	79.4521
	F1 score	0.735	0.834	0.761	0.779	0.719	0.794	0.771
인공신경망	정확도(%)	77.2152	86.0759	81.0127	82.2785	77.2152	82.2785	80.8219
	F1 score	0.747	0.834	0.8	0.789	0.74	0.807	0.785
서포트 벡터 머신	정확도(%)	75.9494	86.0759	81.0127	84.8101	78.481	87.3418	83.5616
	F1 score	0.734	0.834	0.8	0.814	0.756	0.857	0.812

5. 결 론

본 논문에서는 개별 기업 주가의 등락을 예측하기 위해 뉴스기사와 SNS 데이터를 바탕으로 감성분석과 기계학습기법을 사용한 예측모형을 제시하였다. 데이터의 수집은 자체적으로 개발한 프로그램으로 수행하였으며, 수집한 데이터 중 뉴스기사를 이용해 주식도메인에 맞는 감성사전을 구축하였다. 구축한 감성사전을 사용한 감성분석 과정을 거쳐 기계학습을 위한 데이터를 생성하였고, 이를 이용해 다수의 기법들을 기계학습 시켜 각각의 예측모형을 생성하였다. 생성한 예측모형들의 성능을 비교하기 위한 실험결과 로지스틱 회귀분석(평균 정확도-82.7%, 평균 F1 score-0.805)과 서포트 벡터 머신(평균 정확도-82.4%, 평균 F1 score-0.801)을 이용하였을 때, 베이지안 네트워크(평균 정확도-79.5%, 평균 F1 score-0.77)와 인공신경망(평균 정확도-80.9%, 평균 F1 score-0.786)을 이용하였을 때보다 더 좋은 성능을 보였다. 실험 결과에 미루어 봤을 때, 본 논문에서 제시한 주가 예측모형은 기계학습 단계에서 로지스틱 회귀분석과 서포트 벡터 머신 기법을 사용하는 것이 더 적합한 것으로 판단된다. 또한, 본 논문에서 제시한 주가예측 모형의 성능측정 결과는 기존의 연구들에 비해 개선된 결과로 볼 수 있다.

본 연구를 통해 최근 빅데이터의 대두로 주목받고 있는 SNS 데이터를 주가예측 분야에서 활용할 수 있는 한 가지 방안을 제시하였다. 또한 기계학습이론을 바탕으로 하는 예측기법에 필요한 데이터를 정제하는 과정에서 감성분석 기법을 이용할 수 있음을 나타내었다. 반면, 본 연구에서 제시한 예측모형은 뉴스기사와 SNS 상에서 충분히 언급되지 않는 기업들에 대해서는 그 정확도가 떨어질 것으로 예상된다. 또한 감성사전을 구축하기 위해 사용한 데이터와 기계학습에 이용하는 데이터가 같아 종속변수로 설정한 데이터

가 독립변수에 영향을 주었을 가능성을 가지고 있다. 향후 연구에서는 뉴스 기사와 SNS 데이터의 양과 예측정확도와의 상관관계에 대한 연구와 감성사전을 위한 데이터와 기계학습을 위한 데이터를 구분한 연구를 진행해 예측 모형의 한계점을 보완할 수 있는 대처 방안을 강구할 필요가 있을 것이다.

참고문헌

[1] 김광용, 이경락, 이성원, "신규상장기업의 주가예측에 대한 연구", *한국디지털정책학회*, 제11권, 제5호, pp.145-158, 2013

[2] 김상수, 남달우, 조 현, 김성희, "웹 뉴스의 양과 주가의 관계에 관한 연구", *한국IT서비스학회*, 제11권, 제3호, pp.191-203, 2012

[3] 김유신, "주가지수 예측을 위한 뉴스 빅데이터 오피니언마이닝 모형", 국민대학교 박사학위논문, 2013

[4] 김유신, 김남규, 정승렬, "뉴스와 주가: 빅데이터 감성분석을 통한 지능형 투자의사결정모형", *한국지능정보시스템학회 지능정보연구*, 제18권, 제2호, pp.143-156, 2012

[5] 박강희, 신현정, "시계열 네트워크에 기반한 주가예측", *한국경영과학회 한국경영과학회지*, 제28권, 제1호, pp.53-60, 2011

[6] 사공재현, "데이터마이닝을 이용한 주가 예측모형 비교 연구", 인제대학교 석사학위논문, 2012

[7] 송종석, 이수원, "상품평 극성 분류를 위한 특징별 서술어 긍정 부정 사전 자동 구축", *한국정보과학회 정보과학회논문지*, 제38권, 제3호, pp.157-168, 2011

[8] 송치영, "뉴스가 금융시장에 미치는 영향에 관한 연구", *한국국제경제학회 국제경제연구*, 제8권, 제3호, pp.1-34, 2002

[9] 심광섭, 양재형, "인접 조건 검사에 의한 초고속 한국어 형태소 분석", *한국정보과학회 정보과학회논문지*, 제31권, 제1호, pp.89-99, 2004

[10] 안성원, "뉴스 텍스트 마이닝과 시계열 분석을 이용한 주가예측", 연세대학교 석사학위논문, 2010

[11] 유은지, 김유신, 김남규, 정승렬, "주가지수 방향성 예측을 위한 주

제지향 감성사전 구축 방안", *한국지능정보시스템학회 지능정보연구*, 제19권, 제1호, pp.95-110, 2013

[12] 이안 위튼, 아이베 프랑크, 마크 홀, *데이터 마이닝*, 이승현 옮김, 에이콘출판주식회사, 2013

[13] 천세원, "뉴스 콘텐츠의 오피니언 마이닝을 통한 매체별 주가상승 예측정확도 비교 연구", 국민대학교 석사학위논문, 2013

[14] Johan Bollen, Huina Mao, and Xiaojun Zeng, "Twitter mood predicts the stock market", *Journal of Computational Science*, Volume 2, Issue 1, pp.1-8, 2011

[15] Taehwan Kim, Woo-Jin Jung, and Sang-Yong Tom Lee, "The Analysis on the Relationship between Firms' Exposures to SNS and Stock Prices in Korea", *Asia Pacific Journal of Information Systems*, Vol.24, No.2, pp.233-253, 2014