석사학위 논문

뉴스 감성 분석과 시계열 예측 기반의 주가 등락 예측

Stock Fluctuation Prediction based on News Sentiment Analysis and Time Series Prediction

2017년 6월

숭실대학교 대학원

융합소프트웨어학과

김 기 준

석사학위 논문

뉴스 감성 분석과 시계열 예측 기반의 주가 등락 예측

Stock Fluctuation Prediction based on News Sentiment Analysis and Time Series Prediction

2017년 6월

숭실대학교 대학원

융합소프트웨어학과

김 기 준

석사학위 논문

뉴스 감성 분석과 시계열 예측 기반의 주가 등락 예측

지도교수 이 수 원

이 논문을 석사학위 논문으로 제출함

2017년 6월

숭실대학교 대학원

융합소프트웨어학과

김 기 준

김 기 준의 석사학위 논문을 인준함

심 사 위 원 장 김 명 호 인

심 사 위 원 김 계 영 인

심 사 위 원 이 수 원 인

2017년 6월

숭실대학교 대학원

목 차

표 목 차

그 림 목 차

국문초록

뉴스 감성 분석과 시계열 예측 기반의 주가 등락 예측

김기준

융합소프트웨어학과

숭실대학교 대학원

주가 예측은 경제학, 경영학, 통계학, 컴퓨터공학 등의 다양한 분야에서 많이 연구되어 왔다. 그러나 주가의 추이는 일반적으로 비선형적이며 매우 복잡한 양상을 보인다. 이러한 주가를 예측하기 위해 기존에는 과거 주가로부터 시계열 패턴으로 예측하는 방법과 주식 시장 관련 텍스트 데이터로부터 텍스트 마이닝 기법을 통하여 예측하는 방법 등을 이용하였다. 시계열 패턴 기반 예측 방법은 시계열 분석 혹은 머신러닝을 이용해서 주가를 예측하여, 텍스트 마이닝 기반 예측 방법은 뉴스 데이터 혹은 소셜 데이터에 내포되어 있는 감성을 분석하여 주가를 예측한다.

본 논문에서는 뉴스 데이터를 이용한 감성 분석 모듈과 주가 데이터를 이용한 시계열 예측 모듈을 결합한 주가 등락 예측 모델을 제안한다. 감성 분석 방법은 유사한 뉴스를 기반으로 주식뉴스에 대한 감성 수치를 추출하며, 시계열 예측 방법은 LSTM 모델을 이용하여 익일의 종가를 예측한다. 주가 등락 예측 모델은 감성 수치와 익일의 종가를 결합하여 당일 종가 대비 익일의 종가 등락을 예측한다. 본 연구에서 제안하는 방

법과 기존 방법의 성능을 비교한 결과, 제안 방법의 예측 정확도가 기존 방법에 비해 약 4.8% 상향된 결과를 보였다.

의미 있을 정도의 상향 수준??

ABSTRACT

Stock Fluctuation Prediction based on Deep Learning and News Sentiment Analysis

김기준

Department of Software Convergence

Graduate School of Soongsil University

Stock price prediction has been widely studied in various fields such as economics, business administration, statistics, and computer science. However, stock price trends are generally nonlinear and very complex. In order to predict the stock price, traditional studies used prediction methods from the past stock price in a time series pattern or through the text mining technique from the stock market related text data. The time-series pattern based on predicting method is uses time-series analysis or machine learning. The predicting method based on text mining conducts stock price prediction by analyzing sentiment on news data or social data.

In this paper, we propose a stock fluctuation prediction model that combines sentiment analysis module with news data and time series predicting module using stock price data. The sentiment analysis module extracts sentiment values for stock news based on similar

news and predicts the closing price of next day by using LSTM model. The stock fluctuation prediction model combines the sentiment value and the closing price of the next day to forecast the closing price of the next day relative to the closing price of the day. As an experimental result, the prediction accuracy of the proposed method is improved by about 4.8% compared with the conventional method.

제 1 장 서 론

1.1 연구 배경 및 개요

주식 시장의 분석과 주가 예측은 여러 분야에서 지속적으로 연구되어 왔다. 주가 예측은 다양한 접근 방법을 통하여 연구되어 왔는데, 대표적인 연구 방법으로는 시계열 예측을 이용한 방법과 텍스트 마이닝을 이용한 방법 등이 있다.

시계열 예측 방법은 주식 데이터를 사용함으로써 예측 모델을 생성하여 주가를 예측하며, 텍스트 마이닝 기반 예측 방법은 뉴스 데이터 내의 감성 등의 특징을 추출한 후 이를 이용하여 주가를 예측한다[1-13]. 시계열 예측 방법으로는 통계 모델을 이용한 연구와 인공지능을 이용한 연구 방법 등이 있다. 통계 모델을 이용한 연구 방법은 과거 주가에 ARIMA(Autoregressive Integrated Moving Average)와 같은 시계열 모델을 적용하여 주가를 예측한다[1]. 인공지능을 이용한 연구 방법은 인공신경망, SVM, 유전자 알고리즘 등의 방법을 적용하여 주가 지수를 예측한다[2, 3, 4]. 최근 인공지능 분야에서 딥러닝(Deep Learning)이 주목받고 있다. 딥러닝은 다층구조 형태의 인공신경망을 기반으로 하는 기계학습의 한 분야로, 기존의 인공신경망 알고리즘에서 발생하던 문제들을 부분적으로 해소하였다.

텍스트 마이닝 기반 예측 방법은 뉴스, SNS와 같은 텍스트 데이터에 내포되어 있는 감성(Sentiment)을 분석하여 주가 예측을 수행한다. 감성 분석은 감성 사전을 구축하여 이를 이용하는 방법과 딥러닝을 활용한 방법 등이 있다[5].

이러한 배경에서 본 연구에서는 단일 모델을 이용하여 주가 예측을 수행하는 것 보다는 다중 모델을 이용하여 주가 예측을 하는 것이 예측 정

확도를 더 높일 수 있다는 가정 하에, 딥러닝 기반의 시계열 예측 방법과 딥러닝 기반의 감성 분석을 결합하여 주가 등락을 예측하는 방법을 제안한다.

본 연구에서 제안하는 딥러닝 기반 시계열 예측 방법은 딥러닝 알고리즘 중 시계열 데이터 처리에 많이 쓰이는 LSTM(Long Short Term Memory)을 이용한다. 또한, 딥러닝 기반 감성 분석은 주가 등락 예측을 위해 딥러닝 기반의 문서 벡터화를 통해 문서간의 유사도를 구함으로써 주식 관련 뉴스의 긍/부정을 추출하여 예측 모델을 생성한다.

1.2 논문의 구성

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 본 연구와 관련된 연구들을 소개한다. 3장에서는 본 논문에서 제안하는 주가 등락 예측 모델을 설명한다. 4장에서는 기존 연구의 방법과 본 논문의 제안 방법을 실험을 통하여 비교한다. 마지막 5장에서는 결론과 향후 연구를 기술한다.

제 2 장 관련 연구

2.1 시계열 예측 기반 주가 예측

시계열 예측 기반 주가 예측 연구는 다음과 같다. Adewumi[1]는 ARIMA(Autoregressive Integrated Moving Average)를 적용하여 종가를 예측하는 방법을 제안하였다. 박강희[6]는 경제지표들의 상호 연관성 및 다차원적인 인과관계를 고려하고, 그래프 구조에 기반한 Semi-Supervised Learning을 적용하여 종목별 ① 주가를 예측하는 방법을 제안하였다. Zuo[7]는 Bayesian Network를 이용하여 ② 주가수익비율(P/E ratio)를 예측하는 방법을 제안하였다.

2.2 텍스트 마이닝 기반 주가 예측

뉴스 데이터 혹은 소셜 데이터를 이용한 텍스트 마이닝 기반 주가 예측 연구는 다음과 같다. Bollen[8]은 트위터 데이터에 대해 두 가지 방법으로 Public Mood를 측정하였다. 각각의 Public Mood는 Opinion Finder와 Google-Profile of Mood States(GPOMS)를 이용하여 긍/부정과 6개의 감정(Calm, Alert, Sure, Vital, Kind, Happy)으로 측정되었다. 이후 긍/부정과 6개의 감정 변수에 Granger Causality Analysis와 Self-Organizing Fuzzy Neural Network(SOFNN)을 적용하여 다우존스 산업평균지수(Dow Jones Industrial Average)의 종가 등락을 예측하는 방법을 제안하였다. 정지선[9]은 종목의 뉴스 데이터에 대해 감성 분석을 적용하여 개별 기업의 주가 변화를 예측하는 방법을 제시하였다. Schumaker[10]는 S&P 500 관련 종목에 해당하는 주식 뉴스에서 반복하는 고유명사를 추출하고, 뉴스가 배포된 시점의 주가 대비 20분 뒤의 주

가에 대한 등락을 예측하는 방법을 제안하였다.

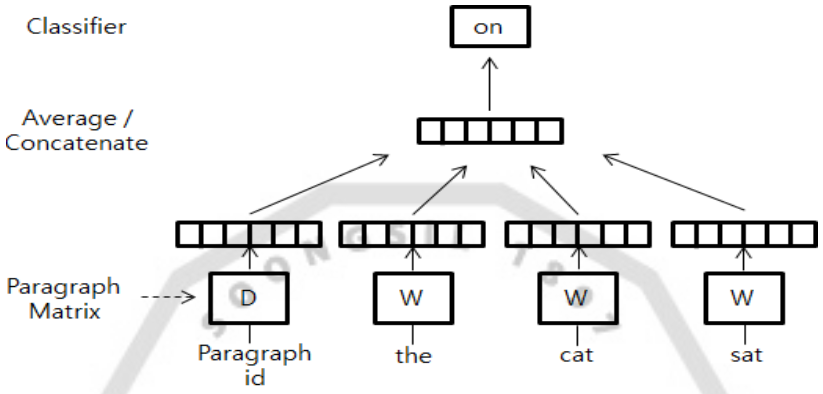
2.3 시계열 예측 및 텍스트 마이닝 기반 주가 예측

시계열 예측 및 텍스트 마이닝 기반 주가 예측을 결합하여 주가를 예측한 연구는 다음과 같다. Agarwal[11]는 Non-linear Model(Recurrent Neural Network)와 Linear Model(Autoregressive moving average)를 결합하여 배당수익을 예측하는 방법을 제안하였다. 안성원[12]은 뉴스 데이터에서 추출한 특성을 주가의 변동폭과 대비하여 전일 종가 대비 당일 종가의 등락율이 ±2% 이상인 뉴스를 Naïve Bayesian 분류기를 이용하여 분류한 다음, RSI(Relative Strength Index)를 계산하여 과매수/매도 구간일 때 가중치를 부여하여 신규 뉴스가 주가 등락에 미치는 영향을 예측하는 방법을 제안하였다. 엄장윤[13]은 뉴스 데이터를 이용한 감성 사전 기반 감성 분석 방법과 ARIMA 모형을 결합하여 주가 등락을 예측하는 방법을 제안하였다.

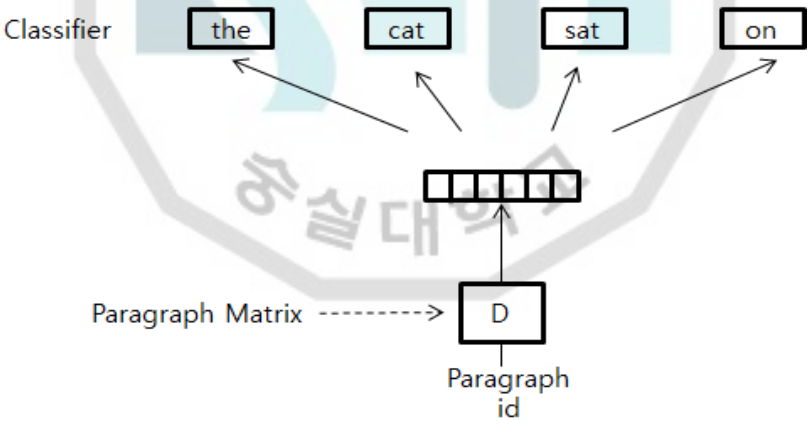
2.4 Doc2Vec

Word2Vec와 같은 Word Embedding 방법의 등장은 자연어처리(NLP) 분야의 여러 문제에 응용되고 있다. Word Embedding은 대용량의 말뭉치를 비지도 방식으로 학습하여, 차원 축소 및 추상화를 통해 문서에 등장하는 단어를 벡터로 표현하는 것이다[14]. Word2Vec은 Word Embedding 학습 방법 중 하나로 학습 시간을 비약적으로 단축시켰으며[15], 단어 뿐만 아니라 문장, 문단에 대한 임베딩 연구도 진행되고 있다. 대표적으로 Milkolov[16]가 제안한 Doc2Vec은 순서와 의미를 내포한 하나의 고유한 벡터로 문장, 문단, 문서를 표현하는 알고리즘이다. Doc2Vec은 Word2Vec의 방법과 거의 유사하며 Doc2Vec은

DM(Distributed Memory) 모델과 DBOW(Distributed bag of words) 모델로 구성되어 있다. [그림 2-1]과 [그림 2-2]는 Doc2Vec 두 가지 모델의 구조를 도식화한 것이다.



[그림 2-1] DM(Distributed Memory) 모델



[그림 2-2] DBOW 모델

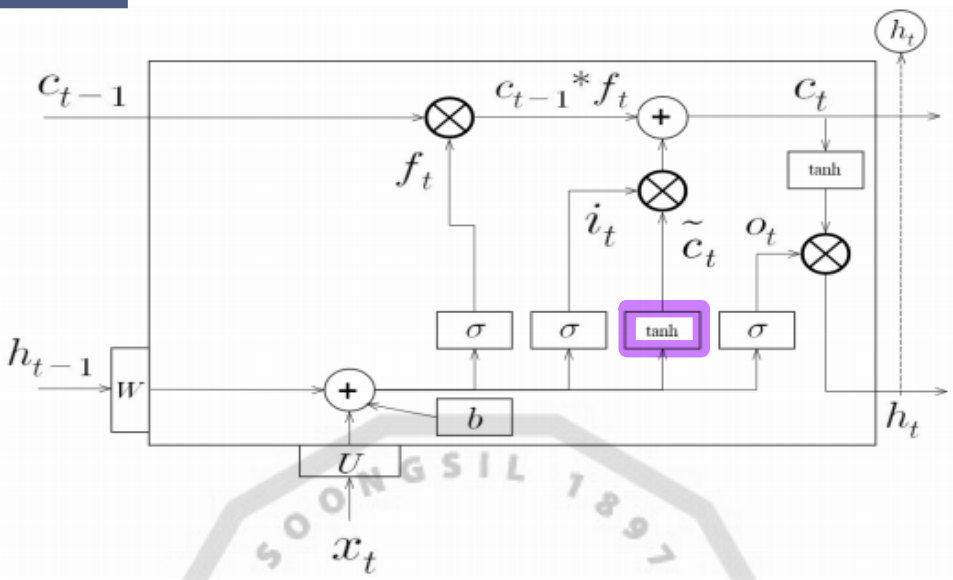
[그림 2-1]과 [그림 2-2]에서 D는 문서의 벡터이다. Doc2Vec은 문단 단위로 단어 시퀀스를 Window Size만큼 학습시키고, 학습은

SGD(Stochastic Gradient Descent) 방식을 사용한다. DM 모델은 [그림 2-1]의 “the cat sat”과 같은 Paragraph Vector를 입력으로 받아 Projection Layer에서 Paragraph Vector와 문서 벡터를 학습한다. 이후 연결된 벡터는 “on”을 예측하는데 사용된다. 반면에 DBOW 모델은 하나의 문서가 주어지면 해당 문서에 포함된 “the cat sat on”과 같은 단어들을 예측하는 모델이다.

2.5 LSTM(Long Short Term Memory)

LSTM은 RNN(Recurrent Neural Network)의 한 종류로 Hochreiter[17]에 의해 제안된 방법이며, 기존의 RNN의 문제점으로 지적된 Gradient Vanishing 문제를 극복한 모델이다. LSTM은 딥러닝 알고리즘 중 시계열 처리에 용이한 알고리즘이다. LSTM은 무언가를 학습하는 것보다 오랫동안 정보를 기억하는데 더 초점을 맞춘 알고리즘이다. [그림 2-3]은 LSTM의 구조 다이어그램이다. LSTM은 기존의 RNN과 마찬가지로 순환적인 구조를 가지며, LSTM 블록 내부는 기억 소자(Memory Cell)와 입력 게이트(Input Gate), 잊기 게이트(Forget Gate), 출력 게이트(Ouput Gate) 총 3개의 게이트로 구성되어 있다.

추후 구조 도식화나 수식 보여줄 때 참고



[그림 2-3] LSTM 구조 다이어그램

잊기 게이트는 기존의 셀 상태(Cell State)를 얼마나 잊어버릴지 결정하는 유닛이다. [식 2-1]은 잊기 게이트인데, h_{t-1}, x_t와 편향의 가중합에 시그모이드 함수를 씌운 형태이다. x_t는 입력값을 나타내며, h_{t-1}는 이전 스텝의 히든 스테이트이다. 히든 스테이트는 네트워크 메모리와 같은 것으로, 과거의 사근 스텝들에서 일어난 일들에 대한 정보를 담고 있다. 잊음 게이트는 [0, 1] 값을 나타내며, 0으로 갈수록 이전 셀 상태를 모두 잊겠다는 의미고, 1로 갈수록 모두 기억한다는 의미이다.

$$f_t = \sigma(W_f \cdot \ [h_{t-1}, x_t] + b_f)$$

[식 2-1] 잊기 게이트

입력 게이트는 얼마만큼 새로운 입력 정보를 셀 상태로 가져갈지를 정하는 유닛이다. 새로운 입력 정보는 잊음 게이트와 동일하게 h_{t-1}, x_t와 편향의 가중합으로 계산된다. [식 2-2]는 입력 게이트를 계산하는 식이다. [식 2-3]은 tanh 층을 통해 셀 상태에 더해질 수 있는 후보 값을 만든다.

$$i_t = \sigma(W_i \cdot \ [h_{t-1}, x_t] + b_i)$$

[식 2-2] 입력 게이트

$$\widetilde{C_t} = \tanh(W_c \cdot \ [h_{t-1}, x_t] + b_c)$$

[식 2-3] 새로운 후보 값

[식 2-4]는 잊음 게이트, 입력 게이트, 후보 값을 합침으로써 이전 셀 상태 C_{t-1}을 새로운 셀 상태 C_t로 갱신하는 수식이다.

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t}$$

[식 2-4] 현재 셀 상태

출력 게이트는 현재 셀 상태 C_t에서 어떤 부분 h_t를 출력할지 결정한다. [식 2-5]와 [식 2-6]은 이에 대한 내용을 수식으로 표현한 것이다.

$$o_t = \sigma(W_o \cdot \ [h_{t-1}, x_t] + b_o)$$

[식 2-5] 출력 게이트

$$h_t = o_t * \tanh(C_t)$$

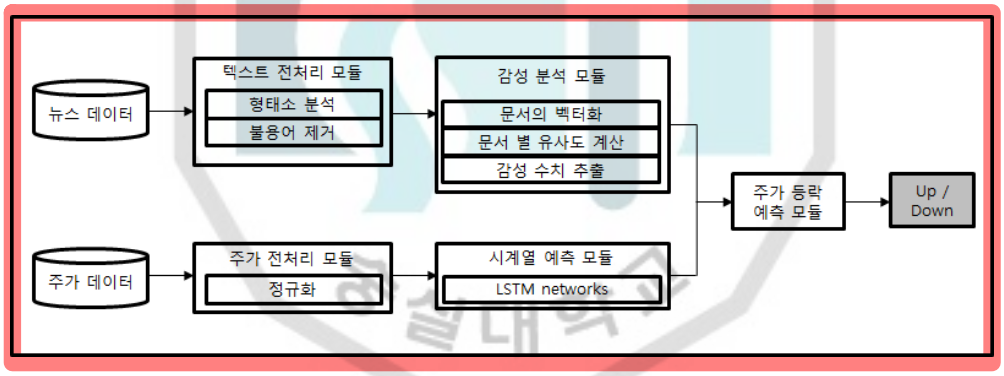
[식 2-6] 히든 스테이트

제 3 장 제안 방법

본 연구에서는 주식 관련 뉴스 데이터 및 주가 데이터를 통해 당일 종가 대비 익일 종가 등락을 당일 폐장 전에 예측하는 방법을 제안한다. 제안 방법은 단일 모델을 이용한 예측 방법이 아닌 다중 모델을 결합하여 주가를 예측하는 방법이다.

3.1 제안 방법의 개요

본 연구에서 제안하는 주가 등락 예측 모델에 대한 전체 시스템 구조도는 [그림 3-1]과 같다.



[그림 3-1] 주가 등락 예측 모델의 구조도

제안 모델은 텍스트 전처리 모듈, 감성 분석 모듈, 주가 전처리 모듈, 시계열 예측 모듈, 주가 등락 예측 모듈로 구성되어 있다. 텍스트 전처리 모듈은 텍스트 데이터를 감성 분석 모듈에 적용하기 위하여 불용어 제거 및 형태소 분석 과정을 수행한다. 감성 분석 모듈은 주가 예측을 위하여 전처리된 문서들로부터 감성 수치를 추출한다. 주가 전처리 모듈은 주가를 시계열 예측 모듈에 적용하기 위하여 데이터를 정규화하는 작업을 수

정처기 요약본 참고

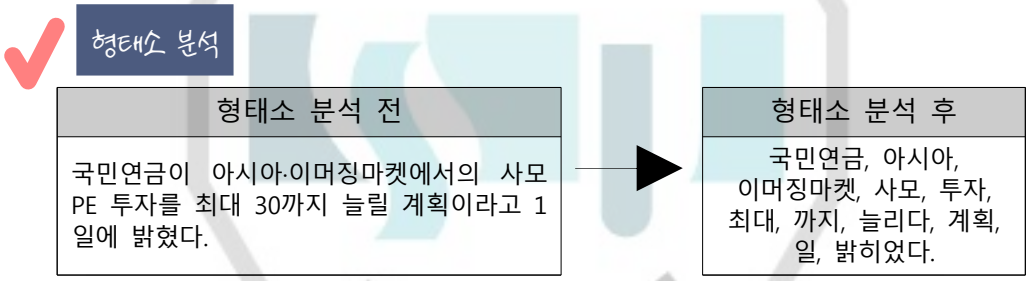
이 얘기 아니었음! 그냥 Min-Max Scaler

행한다. 시계열 예측 모듈은 전처리된 주가 데이터에 대해 딥러닝을 이용하여 익일의 종가를 예측한다. 마지막으로 주가 등락 예측 모듈은 감성 분석 모듈과 시계열 예측 모듈을 결합하여 당일 종가 대비 익일 종가 등락을 예측한다.

3.2 텍스트 전처리 모듈

텍스트 전처리 모듈은 수집된 뉴스 데이터를 분석하기 위한 기본적인 전처리 작업을 수행하며, 형태소 분석과 불용어 제거 두 단계로 구성된다.

형태소 분석은 텍스트 데이터 내 문장에서 명사 및 서술어를 추출한다. [그림 3-2]는 형태소 분석 예시를 나타낸 것이다.



[그림 3-2] 형태소 분석 예시

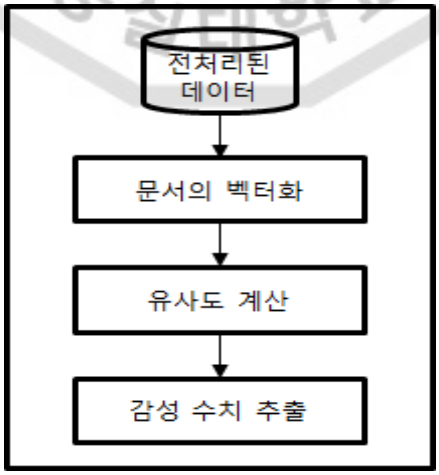
불용어 제거는 수집된 데이터가 올바르게 감성 분석 모듈에 적용되도록 정제를 수행한다. 텍스트 데이터에는 특수문자와 광고 문구 등이 많이 분포되어 있다. 특수문자와 광고 문구는 감성분석에 악영향을 끼칠 수 있기 때문에 정제해주는 작업이 필요하다. [표 3-1]은 불용어 처리 예시이다.

[표 3-1] 불용어 처리 예시

불용어 항목	불용어 예시
하나의 음절을 가진 단어 제거	그, 저, 내, 외 등
숫자 삭제	1, 200, 2016 등
한자 제거	春節, 男心, 檢察 등
기간을 나타내는 단어 제거	지난달, 최근, 올해 등
특수문자와 광고문구 삭제	▶나만의 변호사 찾기

3.3 감성 분석 모듈

감성 분석 모듈은 주가 예측을 위해 뉴스 문서들에 대하여 극성을 추출하고, 추출된 극성을 토대로 당일의 감성 수치를 계산하는 작업을 수행한다. [그림 3-3]은 감성 분석 절차를 도식화한 것이다. 첫 번째로 텍스트 전처리 모듈을 통해 전처리된 문서를 벡터화한다. 그 다음 벡터화된 학습 문서와 입력 문서의 유사도를 계산한 다음 입력 문서에 대한 극성을 추출한다. 입력 문서의 극성은 가장 유사한 학습 문서의 극성으로 정의된다.



[그림 3-3] 감성 분석 절차

3.3.1 문서 벡터화

대부분의 기계학습 알고리즘은 학습을 위해 컴퓨터가 이해할 수 있게 텍스트 데이터를 벡터 공간상에 포현하는 것이 필요하다. 학습에 있어서 DBOW 모델에 비해 DM 모델의 성능이 더 우수하다고 보고되므로, 본 연구에서는 문서의 벡터 표현 생성 시 DM 모델을 사용한다.

3.3.2 문서 별 유사도 계산

기존의 감성 사전 기반의 감성 분석은 주식 뉴스 데이터를 이용하여 감성 사전을 구축한 후, 이를 기반으로 감성 수치를 추출하였다[13]. 하지만 논문에서는 감성 사전을 이용하지 않고 유사도를 기반으로 문서의 극성을 추출한 후, 극성을 토대로 감성 수치를 계산한다. 유사도 기반의 감성 분석을 하는 이유는 감성 사전 기반의 감성 분석은 어휘자원에 의존도가 높아서 구축에 많은 노력이 필요하기 때문이다. 반면에 유사도 기반의 감성 분석은 유사한 문서를 토대로 문서의 극성을 추출하기 때문에 감성 사전 기반의 방식 보다는 좋은 성능을 얻을 수 있다. 그 이유는 유사한 문서간의 극성은 이상치(Outlier)를 제외하고는 큰 차이가 없기 때문이다.

본 연구에서 유사도 계산은 코사인 유사도(Cosine Similarity)를 이용한다. 코사인 유사도는 두 벡터 사이의 유사성을 측정하는 방법 중 하나이며, 이에 대한 정의는 [식 3-1]과 같다. [식 3-1]에서 $d^{training}$은 학습 문서이며, d^{input}은 입력 문서이다. 벡터화된 학습 문서와 입력 문서는 [식 3-1]를 통해 유사도가 계산된다. 여기서 학습 문서란 기계학습을 이용할 때 이용하는 학습 문서가 아닌 입력 문서의 극성을 정의할 때 기준이 되는 문서를 말한다. [표 3-2]는 [식 3-1]을 통하여 입력 문서와 학습 문서들 간의 유사도를 구한 예시이다.

$$Similarity(d^{training}, d^{input}) = \frac{d^{training} \cdot d^{input}}{\|d^{training}\|\|d^{input}\|}$$

[식 3-1] 코사인 유사도

[표 3-2] 입력 문서와 학습 문서들 간의 유사도 예시

입력 문서	학습 문서	유사도
CEO 직속 최고혁신책임자 신설 2총괄 9사업본부 40팀으로 개편 한화생명은 1일 본사 조직개편 및 인사를 실시했다. 이번 조직개편은 혁신과 미래 먹거리에 방점이 찍혔다.	아시아경제 한화투자증권이 영업 경쟁력 강화와 효율성·전문성 제고를 위해 조직개편 및 인사이동을 실시한다고 21일 밝혔다.한화투자증권은 서비스 선택제 도입 효과를 극대화하기 위해 리테일본부 조직을 정비했다.	0.62
	한화그룹이 한화테크윈을 인수할 때만해도 특수부문과 항공기엔진부문은 나눠져 있었다. 그러나 한화테크윈은 조직을 개편해 방산부문에 항공부문을 합쳤다. 압축기와 발전기 등 에너지장비사업도 항공·방산부문에 포함됐다.	0.54
	한화투자증권(003530)이 영업현장 중심으로 조직개편을 실시한다. 내년 흑자전환을 위해 리테일 부문을 강화해 영업력을 키우겠다는 복안으로 풀이된다.	0.48

3.3.3 감성 수치 추출

마지막으로 감성 수치 추출은 문서별 유사도 계산 작업을 토대로 계산된 유사도 중 유사한 k개의 학습 문서들의 극성을 입력 문서의 극성으로 정의된다. 학습 문서는 장중에 배포된 뉴스만을 사용한다. [식 3-2]는 학습 문서 $d^{training}$의 극성을 추정하는 식이며, [식 3-3]은 입력 문서 d^{input}의 극성 $Polarity(d^{input})$을 추정하는 식이다. 입력 문서의 극성을 [식 3-2]와 같이 추정하지 않는 이유는 입력 문서는 유사도 기반으로 극성을 추정하기 때문이다. 또한 학습 문서는 입력 문서의 극성을 정의하기 위한 비교 대상 문서인데, [식 3-2]처럼 추정할 경우 당일 입력 문서의 극성들은 모두 같은 값이 정의되기 때문에 예측 정확도에 부정적인 영향을

차이점?
본 연구에서는 장중에 배포된 뉴스를 분석에 사용

특정 뉴스 매체를 정해서 영향성을 분석해볼 필요성도 고려해봐야 할 듯 (ex. 매경, 한경)

끼칠 수 있으므로 문서의 극성을 [식 3-3]을 이용해서 추정한다. [표 3-3]은 입력 문서의 극성을 추정하는 예시이다.

$$Polarity(d^{training}) = \begin{cases} 1, & \text{학습 문서 } d^{training}\text{이 배포된 날짜의 종가가} \\ & \text{전일 종가보다 상승한 경우} \\ 0, & otherwise \end{cases}$$

[식 3-2] 학습 문서의 극성 추정 식

음... 이 부분 아직 이해 안 감ㅠ

$$Polarity(d^{input}) = \begin{cases} Polarity(d_k^{training}), & \arg max_k \\ & (Simiarity(d^{input}, d_k^{training}) \\ 0, & otherwise \end{cases}$$

[식 3-3] 입력 문서의 극성 추정 식

[표 3-3] 입력 문서의 극성 추정 예시

입력 문서	학습 문서	유사도	극성
CEO 직속 최고혁신책임자 신설 2총괄 9사업본부 40팀으로 개편 한화생명은 1일 본사 조직개편 및 인사를 실시했다. 이번 조직개편은 혁신과 미래 먹거리에 방점이 찍혔다.	아시아경제 한화투자증권이 영업 경쟁력 강화와 효율성·전문성 제고를 위해 조직개편 및 인사이동을 실시한다고 21일 밝혔다.한화투자증권은 서비스 선택제 도입 효과를 극대화하기 위해 리테일본부 조직을 정비했다.	0.62	1
	한화그룹이 한화테크윈을 인수할 때만해도 특수부문과 항공기엔진부문은 나눠져 있었다. 그러나 한화테크윈은 조직을 개편해 방산부문에 항공부문을 합쳤다. 압축기와 발전기 등 에너지장비사업도 항공·방산부문에 포함됐다.	0.54	0
	한화투자증권(003530)이 영업현장 중심으로 조직개편을 실시한다. 내년 흑자전환을 위해 리테일 부문을 강화해 영업력을 키우겠다는 복안으로 풀이된다.	0.48	0

[표 3-3]에서 입력 문서와 가장 유사한 학습 문서는 첫 번째 문서이다. [식 3-3]에 의하여 [표 3-3]의 입력 문서의 극성은 입력 문서와 가장 유사한 첫 번째 문서의 극성으로 추정한다.

날짜별 감성 수치도 계산

날짜 t에 배포된 입력 문서 집합 DOC_t에 대한 감성 수치 $Score(DOC_t)$는 DOC_t에 속한 입력 문서들의 극성의 평균값이며, [표 3-4]는 [식 3-4]를 이용하여 날짜 t의 감성 수치 $Score(DOC_t)$를 계산하는 예시이다.

$$Score(DOC_t) = \frac{\sum_{d^{input} \in DOC_t} Polarity(d^{input})}{|DOC_t|}$$

[식 3-4] 날짜 t에 대한 감성 수치

[표 3-4] $Score(DOC_t)$ 계산 과정

입력 문서 보도 시간	본문	극성
2016-07-01 11:33	상반기 브렉시트 영국의 유럽연합 탈퇴를 끝으로 온갖 악재를 견뎌온 국내 증시가 하반기 변동성 장세 속에서 한줄기 상승 모멘텀을 기대하고 있다.	0
2016-07-01 12:56	삼성전자가 7일 실적발표를 앞두고 오름세다.	1
2016-07-01 13:15	1일 중국 증시가 소폭 상승해 마감했다.	1
$Score(DOC_{2016-07-01}) = \frac{0+1+1}{3} = 0.67$		

3.4 주가 전처리 모듈

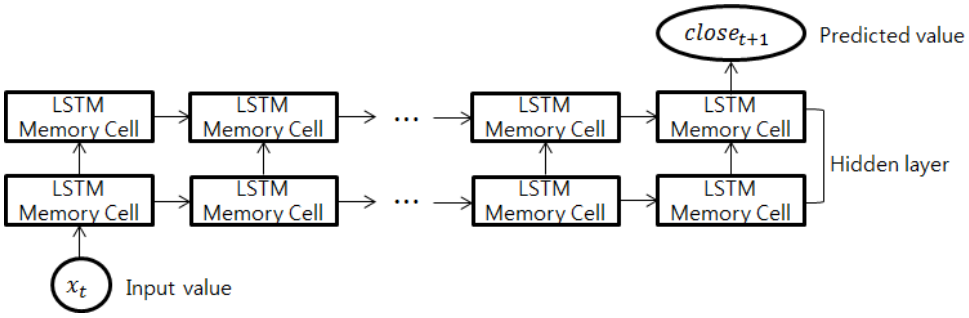
주가 전처리 모듈은 주가 데이터를 시계열 예측 모듈에 적용하기 위해, 입력 데이터를 [0, 1]로 정규화해주는 작업을 수행한다. 본 연구에서는 정규화를 위하여 Min-Max Normalization을 사용하며 그 수식은 [식 3-5]와 같다.

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

[식 3-5] Min-Max Normalization

3.5 시계열 예측 모듈

본 모듈은 주가 전처리 모듈을 통하여 전처리된 데이터가 입력으로 주어졌을 때 LSTM을 통하여 익일의 종가를 예측한다. 시계열 예측 모듈은 LSTM을 기반으로 주어진 입력 데이터를 통해 예측하고자 하는 익일의 종가를 목표로 하여 손실값을 계산하고 최소화하는 과정을 통해 학습된다. 입력 데이터는 전처리된 데이터와 타임 스텝(Time Step)의 형태로 구성한다. 시계열 예측 모듈은 2개의 층으로 쌓여있으며, 각각은 LSTM 블록으로 구성되어 있다. [그림 3-4]는 익일의 종가를 예측하는 시계열 예측 모듈을 도식화한 그림이다.



[그림 3-4] LSTM을 이용한 시계열 예측 모듈

3.6 주가 등락 예측 모듈

본 모듈은 감성 분석 모듈을 통해 계산된 특정일에 대한 감성 수치와 시계열 예측 모듈을 통해 예측된 익일의 종가를 이용하여 로지스틱 회귀(Logistic Regression) 모형을 생성하고 당일 대비 익일 종가의 등락(상승, 하락)을 예측한다. [식 3-5]는 주가 등락 예측 모델로, x_1은 감성 분석 모듈을 통해 추출된 당일의 감성 수치이며, x_2는 시계열 예측 모듈을 통해 예측된 익일의 종가이다.

상승 or 하락 정도가 아니라 상승할지 하락할지만 예측?

$$x_1 = Score(DOC_t)$$
$$x_2 = Close_{t+1}$$
$$\ln\left(\frac{p}{1-p}\right) = a + b_1 x_1 + b_2 x_2$$

[식 3-6] 주가 등락 예측 모델

제 4 장 실험 및 결과

4.1 실험 데이터

본 연구에서 실험으로 사용한 데이터는 '주가 데이터'와 '뉴스 데이터'이다. 뉴스 데이터는 네이버 증권에 실시간 뉴스 탭에 있는 경제 뉴스를 수집한 것이며, 주가 데이터는 Yahoo Finance에서 수집하였다. 수집된 뉴스 및 주가 데이터 예시는 [표 4-1]과 [그림 4-1]과 같으며, [표 4-2]는 주가 데이터 특징에 대한 설명이다.

[표 4-1] 수집된 뉴스 데이터 예시

제목	보도 날짜	신문사	본문
(꾜증시 오후) 상하이 종합 강세...일본 휴장	2013-12-31 17:19	이데일리	IPO 랠리에 상승 마감
(뉴욕전망대) 한산한 시장	2013-12-31 17:19	이데일리	엄지현 기자 올해 마지막 장이 열리는 31일 현지
정유 흑자구간 진입...SK이노베이션 빛볼까	2013-12-31 16:30	매일경제	경기 회복 조짐으로 석유화학 제품 수요가 많아
증시 폐장일 오후 4시 '올빼미 공시' 속출	2013-12-31 15:50	파이낸셜뉴스	2시간 30분간 102개 공시

Date	Open	High	Low	Close	Volume
2016-01-04	1954.469971	1954.520020	1918.760010	1918.760010	359000
2016-01-05	1911.930054	1937.569946	1911.930054	1930.530029	446500
2016-01-06	1934.250000	1934.250000	1911.609985	1925.430054	594600
2016-01-07	1915.709961	1926.410034	1901.239990	1904.329956	393000
2016-01-08	1889.420044	1918.250000	1883.819946	1917.619995	430200
2016-01-11	1897.180054	1907.430054	1892.689941	1894.839966	328800
...					

DACON 데이터와 형태 동일

[그림 4-1] 수집된 주가 데이터 예시

[표 4-2] 주가 데이터 특징

항목	설명
$Open_t$(시초가)	t일에 주식 시장이 개장 했을 때의 주가
$High_t$(고가)	t일의 고가
Low_t(저가)	t일의 저가
$Mean_t$(평균)	t일의 고가와 저가의 평균
$Close_t$(종가)	t일에 주식 시장이 폐장 했을 때의 주가
$Individual_t$(개인)	t일의 개인 순매수
$Foreign_t$(외국인)	t일의 외국인 순매수
$Organization_t$(기관)	t일의 기관 순매수
$Volume_t$(거래량)	t일의 총 거래량

제안 방법의 평가를 위해 뉴스 데이터와 주가 데이터를 Training Set과 Test Set로 나누었다. Training Set은 제안 모델을 학습하는데 사용되며, Test Set은 평가에 사용하는 데이터이다. [표 4-3]은 학습 데이터 및 평가 데이터를 요약한 것이다.

[표 4-3] 수집 데이터 요약

항목	기간	KOSPI일 수	배포된 뉴스 건 수
Training Set	2010. 01. 01. ~ 2013. 12. 31.	933일	555,766건
Test Set	2014. 01. 01. ~ 2014. 12. 30.	243일	20,725건

최근 5년치 데이터를 분석에 사용

4.2 평가 방법

예측 모델이 아니라 분류 모델?

본 연구에서는 제안하는 방법의 성능 평가를 위하여 정확도(Accuracy)를 사용하였다. [표 4-4]는 예측한 결과와 실제 결과 간의 비교를 위한 혼동 행렬이다.

[표 4-4] 혼동 행렬 표

	P'(Predicted)	N'(Predicted)
P(Actual)	TP(True Positive)	FN(False Negative)
N(Actual)	FP(False Positive)	TN(True Negative)

TP(True Positive)와 FN(False Negative)는 예측 모델이 실제로 알맞게 예측한 것이다. TN(False Negative)와 FP(False Positive)는 예측 모델이 잘못 예측한 것이다. 정확도는 [식 4-1]과 같이 계산된다.

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

[식 4-1] 정확도 계산식

4.3 실험 환경 설정

감성 분석 시 문서를 벡터화하기 위해 본 연구에서는 gensim 라이브러리를 사용하였다[18]. gensim은 문서의 벡터 표현 생성 시 문서에 포함된 단어들의 벡터 표현도 함께 생성한다. [표 4-5]는 Doc2Vec 파라미터 설정값을 나타낸 것이다.

[표 4-5] Doc2Vec 파라미터 설정

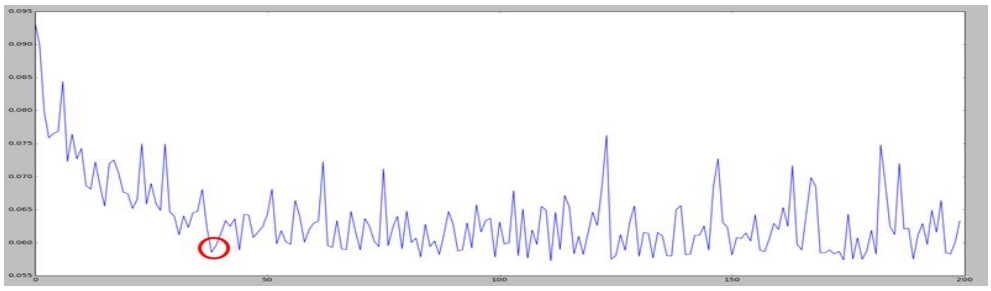
매개변수	설명	설정값
alpha	학습률 초기값	0.025
min_count	발생 빈도수가 설정값보다 작은 단어는 무시	1
size	출력하고자 하는 벡터의 차원 수	300
workers	Doc2Vec 모델을 훈련시킬 때 사용할 worker thread 개수	7
seed	랜덤 숫자 생성기	1234

시계열 예측 모듈을 수행하기 위해서 본 연구는 Keras 라이브러리를 사용하였다. Keras는 Theano, Tensorflow 라이브러리를 래핑(Wrapping)한 라이브러리로, 다양한 머신러닝 알고리즘을 사용할 수 있다.

4.4 실험 결과

4.4.1 시계열 예측 모듈 파라미터 최적화

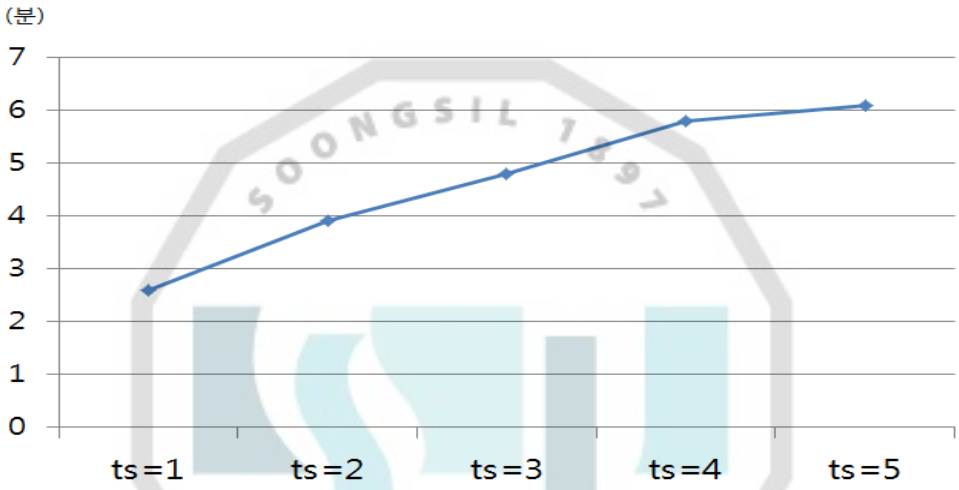
시계열 예측 모듈의 최적의 예측 정확도를 얻기 위해 은닉층(Hidden Layer)의 뉴런 수를 변경해가면서 진행하였고, [그림 4-2]는 앞서 설명한 뉴런 수를 변경하면서 RMSE(Root Mean Square Error)를 출력한 예시이다.



[그림 4-2] 뉴런 수에 대한 시계열 예측 모듈의 RMSE

4.4.2 시계열 예측 모듈 예측 정확도

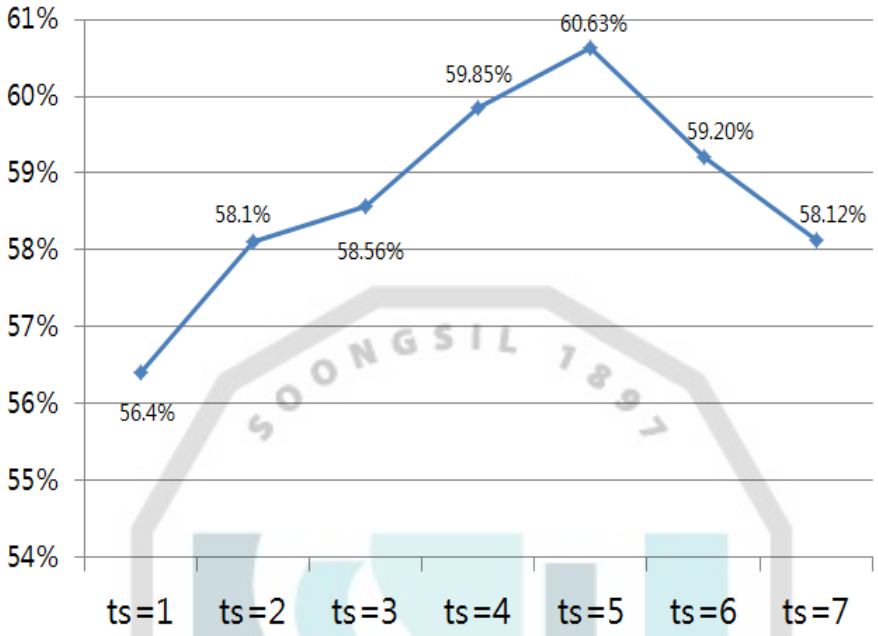
[그림 4-3]은 Time Step 변경에 따른 시계열 예측 모듈의 소요 시간을 나타낸 것이다. [그림 4-3]을 통해 알 수 있듯이 Time Step이 증가할수록 시계열 예측 모듈의 실행 시간이 증가한다. [그림 4-3]에서 x축은 Time Step이며, y축은 시계열 예측 모듈 실행시간이다.



[그림 4-3] Time Step 변경에 따른 시계열 예측 모듈 실행시간

[그림 4-4]는 시계열 예측 모듈의 Time Step 변경에 따른 익일의 종가등락 예측 정확도 결과이다. [그림 4-4]를 통해 Time Step이 5일 때 예측 정확도가 가장 높은 것을 확인할 수 있다. Time Step이 5일 때 예측 정확도가 가장 높은 이유는 주식 데이터의 카오스적인 성질을 잘 포착해서 Time Step이 5이하일 때보다 예측 정확도가 높은 것으로 추측할 수 있다. 그러나 [그림 4-4]의 결과를 통해 Time Step가 6부터는 예측 정확도가 낮아짐을 확인할 수 있는데, 무조건적인 Time Step 수 증가가 더 좋은 결과를 보장하지 않는다는 것을 확인할 수 있다. [표 4-6]은 Time Step이 5일 때의 주가 등락 예측 결과를 혼동 행렬로 나타낸 것이

며 시계열 예측 모듈의 정확도는 60.63%로 측정되었다.



[그림 4-4] Time Step 변화에 따른 시계열 예측 모듈 예측 정확도

[표 4-6] Time Step=5일 때 주가 등락 예측 결과

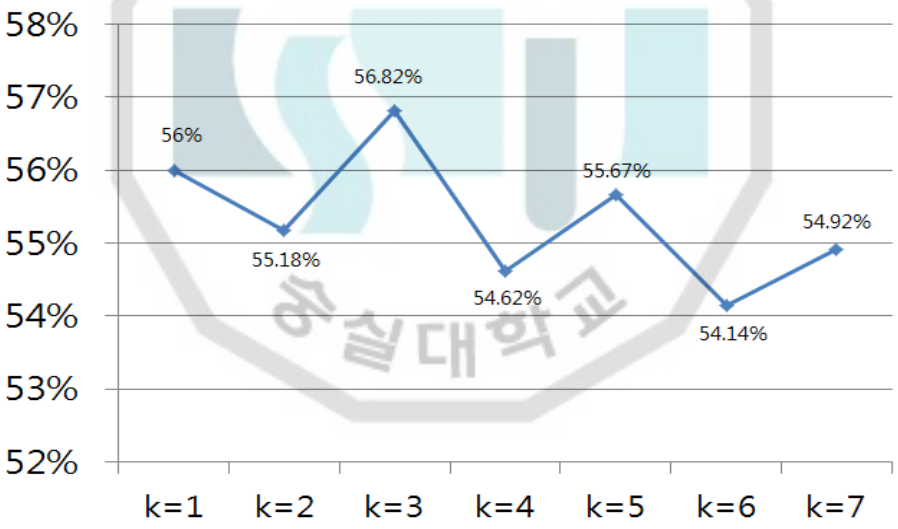
예측 / 실제	상승	하락
상승	77	45
하락	51	71

4.4.3 감성 분석 모듈 예측 정확도

본 연구에서 제안하는 감성 분석 모듈로 익일의 종가 등락을 예측한 결과는 [그림 4-5]와 같다. [그림 4-5]는 유사한 문서 개수 k개를 변경해

가면서 주가 등락을 예측하였을 때의 예측 정확도를 그래프로 표현한 것이다. [그림 4-5]에 따르면 k가 홀수일 경우가 짝수인 경우보다 예측 정확도가 높은 것을 확인할 수 있다. k가 짝수일 때 예측 정확도가 낮은 이유는 본 연구에서 중립(Neutral)을 고려하지 않고 단순히 극성을 이진(Binary) 분류하였기 때문에, k가 홀수인 경우에 비해 예측 정확도가 낮은 것으로 추정된다. k가 3일 때의 예측 정확도가 56.8%로 가장 높은 것으로 측정되었다. k가 4이상일 때는 예측 정확도가 낮아지는데, 이 결과로부터 유사도가 낮은 문서까지 포함하면 오히려 예측에 악역향을 미치는 것을 알 수 있다. [표 4-7]은 k=3인 경우의 예측 결과에 대한 혼동 행렬이다.

한계점 명시



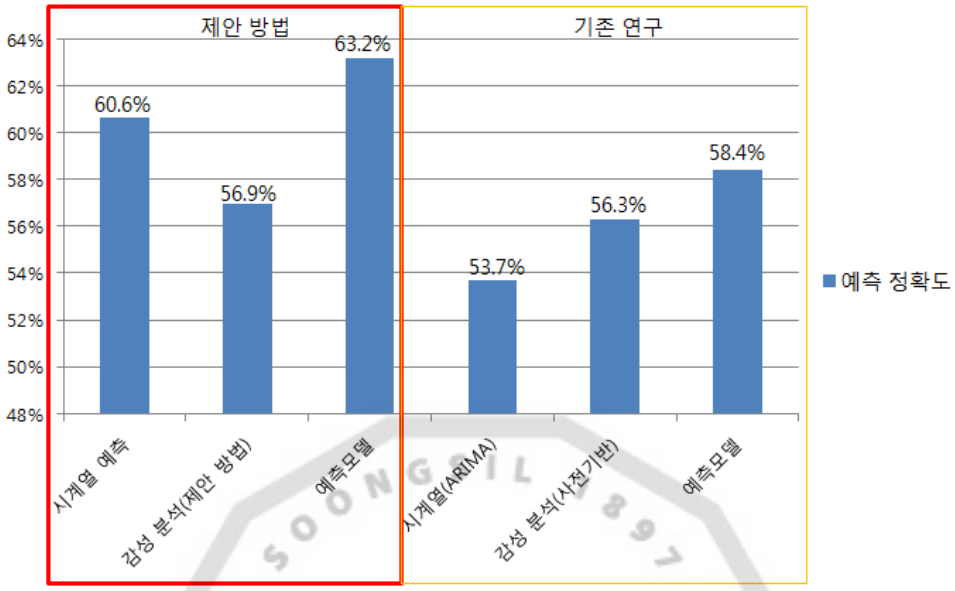
[그림 4-5] k에 따른 감성 분석 모듈 예측 정확도

[표 4-7] k=3일 때 주가 등락 예측 결과

실제 \ 예측	상승	하락
상승	72	50
하락	55	67

4.4.4 예측 모듈 별 예측 정확도

[그림 4-6]은 본 연구의 예측 정확도와 비교 연구[13]의 예측 정확도를 비교한 것이다. 비교 연구는 당일 개장부터 당일 폐장 1시간 전에 보도된 뉴스로부터 감성 사전을 구축하여 감성 수치를 추출하고, KOSPI 데이터로부터 ARIMA 모형을 토대로 예측한 결과를 결합하여 당일 종가 대비 익일의 종가 등락을 당일 폐장 전에 예측하였다. 본 연구는 주식 관련 뉴스 데이터와 주가 데이터를 통해, 당일 종가 대비 익일 종가 등락을 당일 폐장 전에 예측하였다. [그림 4-6]에 따르면 제안 모델의 예측 정확도가 비교 연구의 예측 정확도 보다 약 4.8% 높은 것을 확인할 수 있다. 그 이유는 제안 모델과 비교 연구는 두 개의 모듈을 결합하여 주가 등락을 예측하는데, 각각의 모듈의 예측 정확도가 비교 연구의 예측 정확도보다 높았기 때문이다. 본 연구에서 시계열 예측은 비교 연구 대비 6.9% 향상되었고, 감성 분석은 비교 연구 대비 0.3% 향상되었다.



[그림 4-6] 제안 방법과 기존 연구와의 예측 정확도 비교

제 5 장 결론 및 향후 계획

본 연구에서는 당일 종가 대비 익일의 종가를 당일 장 마감 전에 예측하기 위하여 감성 분석 모듈과 시계열 예측 모듈을 결합한 주가 등락 예측 모델을 제안하였다. 제안 모델은 유사한 뉴스를 토대로 주식 뉴스의 긍/부정을 토대로 감성 수치를 추출하고, LSTM 기반의 시계열 예측 모듈을 통해 익일의 종가를 예측한다. 이 후 주가 등락 예측 모듈을 통해 당일 종가 대비 익일의 종가 등락을 예측한다.

비교 연구와 동일한 데이터로 평가 결과, 제안 방법의 성능은 비교 연구보다 약 4.8%의 높은 예측 정확도를 보였다. 감성 분석 모듈은 유사한 학습 문서 개수 k가 3일 때 가장 우수한 성능을 보였고, 시계열 예측 모듈은 Time Step이 5일 때 가장 우수한 성능을 보였다.

향후 연구로 주식 전문가들이 사용하는 이동 평균(Moving Average) 및 이동 평균 수렴 분산(Moving Average Convergence Divergence)과 같은 기술 지표의 고려가 필요하다. 또한 주가 등락을 강한상승/상승/하락/강한하락 등으로 세분화하여 예측하는 것이 필요하다. 본 연구에서는 실험데이터로 KOSPI 종합 지수를 사용하여 KOSPI 등락 여부를 예측하였는데, 제안 방법은 개별 종목의 주가 등락 예측에도 확장 가능하다. 또한 향후연구를 통해 문서의 극성을 긍/부정으로만 정하는 것이 아니라 중립을 고려하는 방안과 등락을 이진으로 예측하는 것이 아닌 실수값으로 등락률을 예측하는 방법을 연구할 계획이다.

참고문헌

[1] Ariyo, A. A., Adewumi, A. O., & Ayo, C. K. (2014). Stock price prediction using the ARIMA model. In Computer Modelling and Simulation (UKSim), 2014 UKSim-AMSS 16th International Conference on (pp. 106-112). IEEE.

[2] 김광용, 이경락. (2008). 인공지능시스템을 이용한 주가예측에 대한 연구. 대한경영학회지, 21(6), 2421-2449.

[3] Pai, P. F., & Lin, C. S. (2005). A hybrid ARIMA and support vector machines model in stock price forecasting. Omega, 33(6), 497-505.

[4] 김상호, 김동현, 한창희, 김원일. (2008). 주가지수 관계와 유전자 알고리즘을 이용한 주식예측. 한국지능시스템학회 논문지, 18(6), 781-786.

[5] 서상현, 김준태. (2016). 딥러닝 기반 감성분석 연구동향. 한국멀티미디어학회지, 20(3), 8-22.

[6] 박강희, 신현정. "Semi-Supervised Learning을 이용한 주가예측." 한국경영과학회 학술대회논문집, (2010.10): 110-116.

[7] Zuo, Y., & Kita, E. (2012). Stock price forecast using Bayesian network. Expert Systems with Applications, 39(8), 6729-6737.

[8] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of computational science, 2(1), 1-8.

[9] 정지선, 김동성, 김종우. (2015). 온라인 언급이 기업 성과에 미치는 영향 분석. 지능정보연구, 21(4), 37-51.

[10] Robert P. Schumaker(2009), “A quantitative stock prediction

system based on financial news", Information Processing and Management 45, pp. 571-583.

[11] Rather, A. M., Agarwal, A., & Sastry, V. N. (2015). Recurrent neural network and a hybrid model for prediction of stock returns. Expert Systems with Applications, 42(6), 3234-3241.

[12] 안성원, 조성배. (2010). 뉴스 텍스트 마이닝과 시계열 분석을 이용한 주가예측. 한국정보과학회 학술발표논문집, 37(1C), 364-369.

[13] 엄장윤, 이수원. (2015). ARIMA 모형과 텍스트 마이닝을 이용한 주가 등락 예측. 숭실대학교 일반대학원 석사학위논문

[14] Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. Journal of machine learning research, 3(Feb), 1137-1155.

[15] Mikolov, T., Yih, W. T., & Zweig, G. (2013, June). Linguistic Regularities in Continuous Space Word Representations. In Hlt-naacl (Vol. 13, pp. 746-751).

[16] Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on Machine Learning (ICML-14) (pp. 1188-1196).

[17] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[18] Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.