Stock Prediction and Prediction Accuracy Improvement using Sentiment Analysis and Machine Learning based on Online News

Yunsoo Lee, Hosung Ryu, and Hohyun Lee
Data Science Lab, Paul Math School
Goesan County, Republic of Korea
Okyunsoo1234@gmail.com, rhs4878@naver.com, hohyun.lee@pmath.org

Abstract

Investment strategy and predicting technique appeared to analyze pattern of stock market to get economic gain. However, predicting the flow of stock index is quite difficult because stock market contains uncertain factors. To overcome this problem, varieties of methodology is going along. And with 'big data', varieties of atypical data come out with social media. Therefore, in this research paper we predicted fluctuation of stock price with using 'News data'. We used morpheme analysis and sentimental analysis to make digitalize it. Next with this data we applied machine learning and made predicting model. Finally, we got prediction rate and F1 score.

Keywords
Machine Learning, Sentiment Analysis, Online news, Stock Index, text data, Stock-Prediction

1. Introduction

Stock is related to money therefore it is becoming an object of attention to not only professional but also ordinary people. Furthermore, as age of low interest rate is continuing, an importance of stock prediction is gradually standing out.

At present, varieties of research papers are trying with new scientific technology to predict stock index. Fundamental analysis and technical analysis were typical stock-predicting methodologies. Fundamental analysis predicts by analyzing the financial structure, market prospect, inherent value. Contrarily technical analysis predicts by analyzing statistical information such as volume, moving average, trend. Nowadays, as technology develops and big data appears, the range of technical analysis has extended. Technological method became a part of technical analysis. Machine learning, sentimental analysis and text mining are typical example of technological methods. Technological method is using for purpose of rapid and accurate calculation of the vast amount of data. A notable feature of technological method is that it can use text data to independent variable. Recently, online news, which comprehensively contains social issue, are using on stock-prediction.

Therefore, this research paper predicted stock index in using online new data. First, select target companies to predict and collect online news data which is related in target company. Second, digitalize new data and apply it on machine learning and make predicting model. Finally, we analyzed relationship between data and predicting model and compared accuracy rate.

2. Related Research

Sentimental analysis is the part of text mining technology which analyzes semi-structured and atypical text and extracts practical information that is defined as the natural language processing technology which analyzes people's subjective opinions about a certain kind of brand, event and product. Text mining is able to be applied to Social Network Service which contains a lot of subjective information and news which consist of objective information.

Morpheme is defined as 'the minimum unit of morphologic level of a language' and the meaning of a language would be lost if further unit is reduced. Morpheme analysis is the process of separating given sentence by morphemic units and then giving them a part of speech or meaning which fits each morpheme such as predicate, noun, adjective, adverb and postposition. In this paper, for the sentimental analysis of atypical text, a process of analyzing morphemes and giving them polarities appropriate for each morpheme and grasping polarities of the whole texts is carried out.

Sentimental analysis is for analyzing a polarity which judges whether given text data means the positive or the negative in certain writing. In this process, if the research is being done by using the sentimental dictionary which contains frequency and positive index of each word, the accuracy of the sentimental analysis will rise. Jong- Sung Song(2011) found out that carrying out an research with an exclusive sentimental dictionary for each field not an universal sentimental dictionary can increase the accuracy of the sentimental analysis and Dong-Sung Kim(2015) confirmed that if sentimental dictionary is expanded as time passes an increase in the accuracy of sentimental analysis will happen, but if sentimental dictionary is neglected without any further expansion that dictionary will lose its efficiency as time goes by. In this research, based on the method of building sentimental dictionary of Dong-Yeong Kim(2015), an exclusive sentimental dictionary for stock prediction for each company is made to conduct sentiment analysis.

So far constant researches for stock prediction have been conducted in various ways. Byeong-Su Jo(2009) found out that despite the differences among each news, the top 10 news presented every year affect the stock index of securities market and Jang-Yeun Um (2015) and Il-Ji Choe conducted the research on prediction for stock fluctuation through text mining based on news, and Yu-sin Kim(2012) used the sentimental analysis method to analyze news and produce investment information, and based on this, Yu-sin Kim suggested intelligent investment decision-making model for stock prediction. Additionally, R.Tushar and Saket(2012) analyzed the correlation between social site based on the sentimental analysis about certain company and result of short-term market through extensive social data, and found out that public opinions which show positive and negative perspectives clearly hugely affect stock index of each company. Therefore, this paper is to predict stock by creating prediction model by mechanically educate data set after sentimental-analyzing articles published in economy section in Naver news and refining the data.

3. Stock Prediction Model

3.1 Stock Prediction by Online News

In this paper, stock prediction was tried through news data which is atypical data, which have close relationship with stock, and news data of 'NAVER', the biggest search engine portal site in South Korea, were used. Data were collected within economy section in NAVER news and 'articles published in newspaper', and it was planned to create stock prediction model after sentimental-analyzing collected data and use them for mechanical education. The process of conducting this research is shown in [Figure 1].

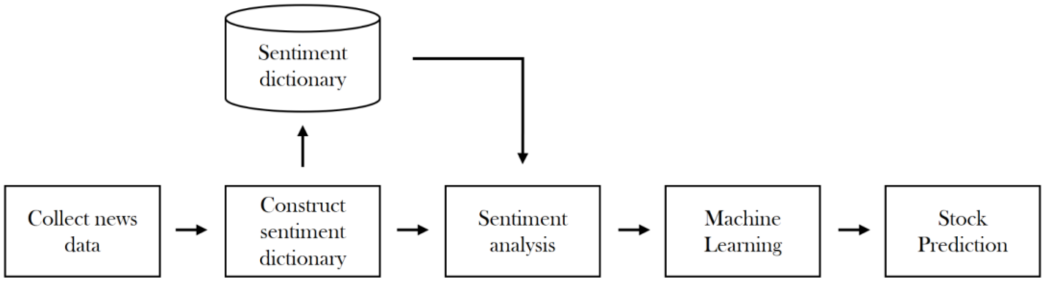


Figure 1. Stock Prediction Model

The process of conducting this research is quoted from stock prediction model of Dong-Yeong Kim(2015), and prediction for stock fluctuation was made through 4 steps in order of data collection, building of sentimental dictionary, sentimental analysis and mechanical education. In data collection, one keyword for each chosen sample company is selected and news data which has those keywords are collected. Collected data will go through a morpheme analysis and will be sorted to the words that correspond to nouns and category of assumed noun, and the sorted words of highest frequency of 1% will be registered in sentimental dictionary. Subsequently, frequency, positive figure and positive index about each word will be calculated to complete sentimental dictionary for each company. In the next step, based on sentimental dictionary built for each company, sentimental analysis was applied to refined data and they were used for mechanical education. In mechanical education, multiple prediction models were made through the process of machine learning of each classification method by using refined data and finally, after verification for functions accuracy of models will be compared to each other.

3.2 Data Collection

Web news data are generated in real time, as well as by a variety of routes and it might be hard to collect only data necessary for a study due to extremely wide range of their kinds. In order to use this kind of web news data the work of automated system is required to extract only necessary information. In this paper, through package Beautifulsoup(4.4.1) of computer programming language Python(Ver. 3.4), data-collecting program and database is set, and using these, only necessary data is extracted. In collected text data, irrelevant information like publisher, advertisement and copyright is included so memo program 'Notepadd++' was used to refine data. Collection process of web news data is shown in [Figure 2], and a part of programming code is shown in [Figure 3].

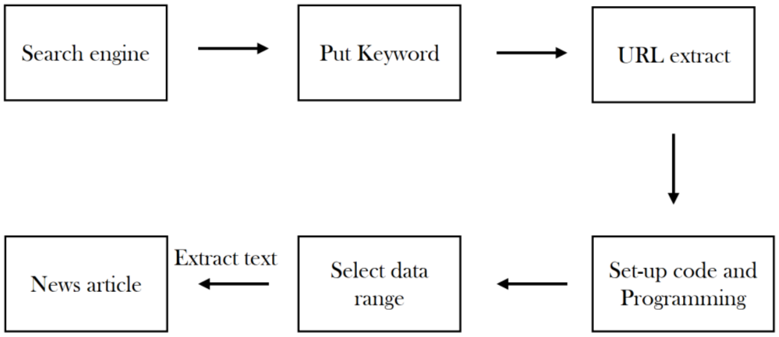
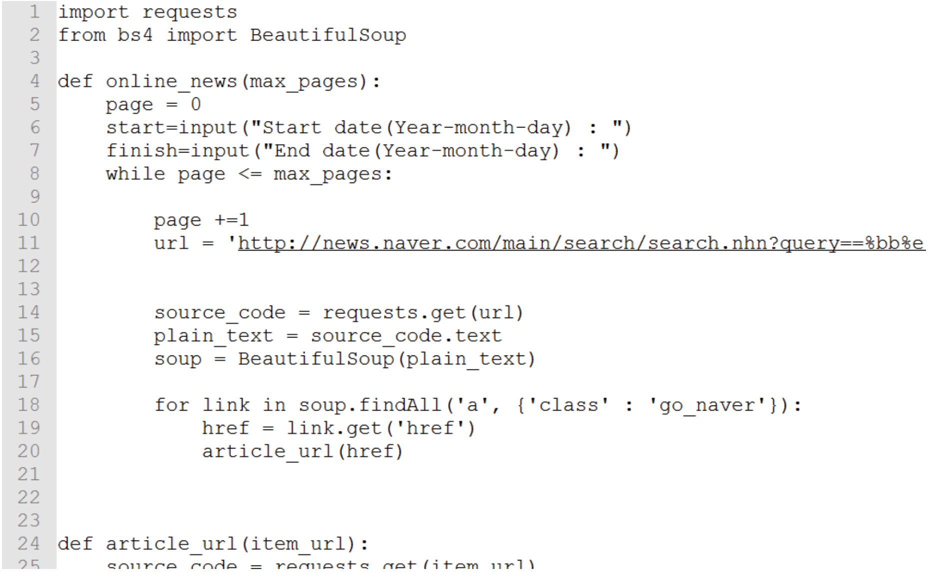


Figure2. Data Collection Procedure



```
import requests
from bs4 import BeautifulSoup

def online_news(max_pages):
    page = 0
    start=input("Start date(Year-month-day) : ")
    finish=input("End date(Year-month-day) : ")
    while page <= max_pages:

        page +=1
        url = 'http://news.naver.com/main/search/search.nhn?query==%bb%e


        source_code = requests.get(url)
        plain_text = source_code.text
        soup = BeautifulSoup(plain_text)

        for link in soup.findAll('a', {'class' : 'go_naver'}):
            href = link.get('href')
            article_url(href)



def article_url(item_url):
    source_code = requests.get(item_url)
```

Figure3. Part of Python Code

3.3 Morpheme Analysis

In this paper, as a system for analyzing morpheme for sentimental analysis, Kkokkoma® morpheme analyzer which is developed by research team of IDS in Seoul National University and open-source morpheme analyzer were selected. Kkokkoma morpheme analyzer divides parts of speech of Korean letter into noun, predicate, determiner, adverb, interjection, postposition, ending, affix, root, mark and 'nor Korean' and classify each detailed parts of speech. 'Noun' which is not sorted in detail will be sorted into normal noun, proper noun, general dependent noun and unit-dependent noun. Moreover, in case of 'analysis impossible', it could be sorted into range of assumed noun, assumed predicate and range of analysis impossible, but the range of assumed predicate and analysis impossible are eliminated from the tag chart. In this research, morphemes that correspond to nouns were extracted based on the assumption that it will have biggest effect compared to other morphemes when the noun which means practical concept among various parts of speech performs sentimental analysis, and morphemes that correspond to the range of assumed noun which includes a lot of loanword nouns were also extracted.

3.4 Sentiment Dictionary

Text data is separated into units of morphemes through analysis morphemes, and related part of speech for each morpheme is given. Processed morphemes like above will be given a polarity figure for each word and through this, sentimental dictionary will be built. Sentimental dictionary is a dictionary of collection of polarity figures for each word and depending on how it is being built the accuracy of sentimental analysis fluctuates. As the current status of sentimental dictionary, there is 'dictionary for English/Chinese sentimental analysis' which is developed by the English sentimental dictionary SentiWordNet(http://sentiwordnet.isti.cnr.it/) and the knowledge-based database site of China hownet(http://keenage.com/html/c_index.html), and researches related to developing sentimental dictionary are being carried out actively around the world including Korea. However, sentimental dictionary developed by individual researchers is only used in certain fields and not universally used, and even if it is open-source it is inappropriate to generalize it and use it for researches. For preventing this and improving the accuracy of the dictionary researchers need to develop a sentimental dictionary autonomously that can be used in research field. The order of building a sentimental dictionary is stated as follows. First, morphemically analyze the collected data. Second, extract the candidate words that will be registered in the dictionary. Third, calculate frequency of the words. Fourth, extract the words of highest frequency of top 1% and calculate positive index(Dong-Yeong Kim, 2015). Frequency represents the number of register of a word on all of news articles and positive figure is calculated by adding numbers of cases of increase in that day's stock when news that have corresponding keyword were published(Dong-yeong Kim, 2015). The process of building sentiment dictionary is shown in [Figure 4].

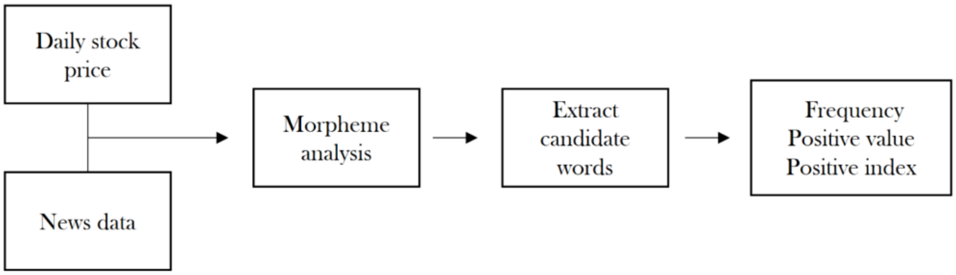


Figure 4. Process of building an sentiment dictionary

At this time, the equation of positive index is stated as follows.

$$include(i, j) = \begin{cases} 1(if\ the\ word\ j\ contained\ in\ news\ article\ j) \\ 0\,(In\ other\ cases) \end{cases}$$

$$NSP(j) = \begin{cases} 1(if\ the\ daily\ stock\ price\ increased, after\ news\ article\ j\ published) \\ 0\,(In\ other\ cases) \end{cases}$$

$$positive(i) = \sum_{j=1}^{n} \{include(i,j) \times NSP(j)\}, n = number\ of\ \ whole\ news\ article$$

Because it was believed that higher accuracy could be made if the words that have relatively higher frequency among extracted candidate words through morpheme analysis were chosen to be registered in the dictionary and use them for machine learning so the words that have highest frequency of top 1% were selected as the subjects for analysis. Finally, calculate positive index through frequency and positive figure of each word and build the sentimental dictionary. Positive index is the figure of positive figure divided by frequency and the equation is stated as follows.

$$P(i) = \frac{\sum_{j=1}^{n} \{include(i,j) \times NSP(j)\}}{\sum_{j=1}^{n} include(i,j)}$$

3.5 Sentiment Analysis

By using the building method of sentimental dictionary built in the paragraph 4 sentimental analysis of web news data was conducted. First, after performing the process of morpheme analysis to extract the morphemes of nouns and the range of assumed nouns collected data will be compared with words on sentimental dictionary to calculate positive index of corresponding data. Positive index of text will be presented as arithmetic average figure which is divided by the number added from positive index of morphemes extracted from corresponding data(Dong-yeong Kim,2015). The equation for positive index of data is stated as follows.

$$match(i,j) = \begin{cases} 1(if\ the\ word\ \ j\ which\ is\ contained\ in\ text\ i, exist\ in\ sentment\ dictionary) \\ 0\,(In\ other\ cases) \end{cases}$$

$$PT(i) = \frac{\sum_{j=1}^{n} \{match(i,j) \times P(j)}{\sum_{j=1}^{n} match(i,j)}, n = number\ of\ \ text\ in\ \ i$$

Going further from just calculating the positive index of data, calculate positive index per each day. Positive index per each day is presented as arithmetic average figure divided by the number added from positive index of texts registered on corresponding day, and equation for positive index per day is stated as follows.

$$DP(k) = \frac{\sum_{j=1}^{n} PT(i)}{n},\ n = number\ text\ appeared\ in\ date\ k$$

Frequencies, positive and positive indexes of each word were calculated to understand the positive index of text of news article and by using this, refine the data to use for mechanical education. Finally, by using frequency and positive index of each word per each day shown in database, positive index per day can be drawn and it is able to complete data set by adding items about fluctuation of the next day’s stock. Because this research predicts only fluctuation of stock it shows only whether stock increases or decreases in the item of the next day’s stock fluctuation. In database, 1 indicates increase and 0 indicates decrease. [Figure 5] is the process of overall sentimental analysis.

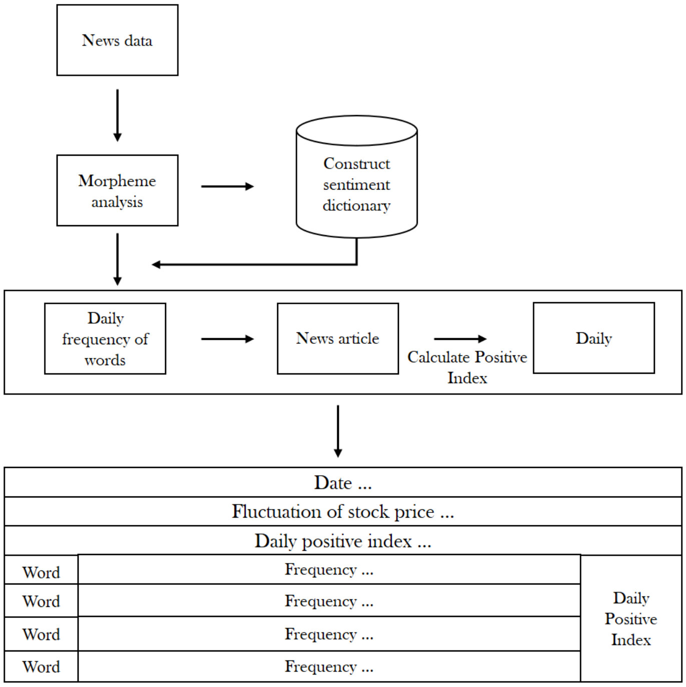


Figure 6. Process of overall sentiment analysis

3.6 Machine Learning

Human behavior is derived from past experience not unconsciousness. By getting new information and studying over and over, they can bring themselves the best outcome from next actions. What brought this human characteristics to technology is called 'Machine learning' which is a serial process of extracting characteristics of collected data and test them to optimize or self-develop, and it is frequently used in prediction areas(Lee, 2016).
Data for conducting machine learning consist of class label which confirms whether attribute set and respective data are time series, whole number or real number and through this kind of property, predicting dependent variable with independent variable data is the objective of machine learning(Dong-yeong Kim, 2015). This research predicted stock after creating prediction model using the method of decision-making tree, logistic regression analysis, SVM and artificial nerve network which are based on the theory of machine learning as it is stated in [Figure 7].

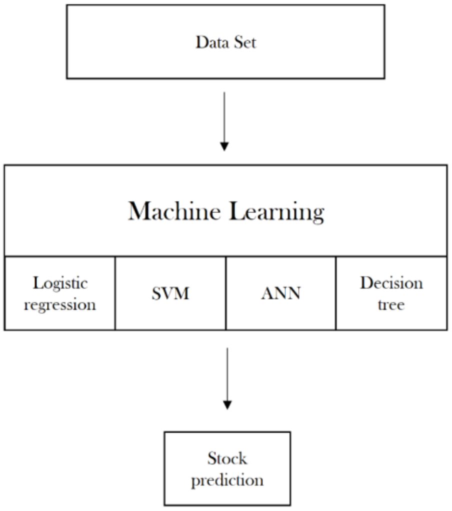


Figure 7. Process of Machine learning

4. Experiment

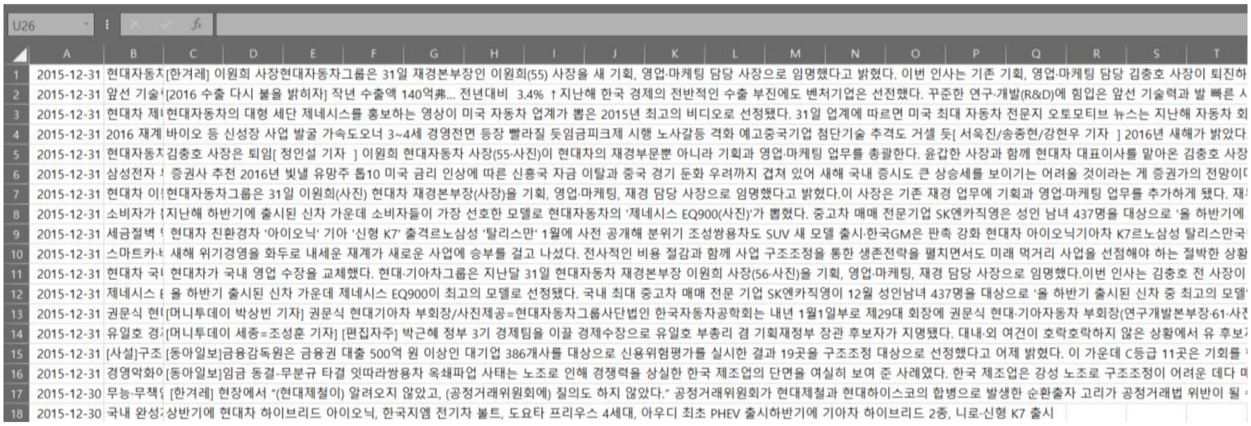
4.1 Experiment Protocol

In this paper, the research was conducted by choosing 4 companies that are listed on the KOSPI as the subjects, and news articles published from January 1th, 2015 to December 31th, 2015 were used for the research. The research was carried out based on the contents of chapters 2 and 3 and the order of experiment is data collection, morpheme analysis, building sentimental dictionary, sentimental analysis and machine learning, according to the stock prediction model in chapter 3. Through data refined on each step, prediction models were made and analyses on results of predictions were conducted.

4.2 Data Collection

Web news data extracted only articles published in newspaper which were in economy section of NAVER news using 'parser' which is the news developed autonomously by using 'python', and collected articles of 62,765. It was believed that news related to economy would reflect the flow of stock market the most so the news that were published in economy section of NAVER news were selected as data subjects.

[Figure 8] is the arrangement of texts of collected news articles in MS Excel, there are the dates when the news was published in the first row and dates are written in the order from December to January. There are the contents of articles of news published on corresponding dates in the second row. [Figure 8] is a part of collected news articles and [Table 1] is the number of news articles collected by each company.



	A	B	C
1	2015-12-31	현대자동차	[한겨레] 이원희 사장현대자동차그룹은 31일 재경본부장인 이원희(55) 사장을 새 기획, 영업·마케팅 담당 사장으로 임명했다고 밝혔다. 이번 인사는 기존 기획, 영업·마케팅 담당 김충호 사장이 퇴진하
2	2015-12-31	앞선 기술	[2016 수출 다시 불을 밝히자] 작년 수출액 140억弗... 전년대비 3.4% ↑지난해 한국 경제의 전반적인 수출 부진에도 벤처기업은 선전했다. 꾸준한 연구·개발(R&D)에 힘입은 앞선 기술력과 발 빠른
3	2015-12-31	현대차 제	현대자동차의 대형 세단 제네시스를 홍보하는 영상이 미국 자동차 업계가 뽑은 2015년 최고의 비디오로 선정됐다. 31일 업계에 따르면 미국 최대 자동차 전문지 오토모티브 뉴스는 지난해 자동차 회
4	2015-12-31	2016 재계	바이오 등 신성장 사업 발굴 가속도오너 3~4세 경영전면 등장 빨라질 듯임금피크제 시행 노사갈등 격화 예고중국기업 첨단기술 추격도 거셀 듯[서욱진/송종현/강현우 기자] 2016년 새해가 밝았다.
5	2015-12-31	현대자동차	김충호 사장은 퇴임[정인설 기자] 이원희 현대자동차 사장(55·사진)이 현대차의 재경부문뿐 아니라 기획과 영업·마케팅 업무를 총괄한다. 윤갑한 사장과 함께 현대차 대표이사를 맡아온 김충호 사장
6	2015-12-31	삼성전자	증권사 추천 2016년 빛낼 유망주 톱10 미국 금리 인상에 따른 신흥국 자금 이탈과 중국 경기 둔화 우려까지 겹쳐 있어 새해 국내 증시도 큰 상승세를 보이기는 어려울 것이라는 게 증권가의 전망이
7	2015-12-31	현대차 이	현대자동차그룹은 31일 이원희(사진) 현대차 재경본부장(사장)을 기획, 영업·마케팅, 재경 담당 사장으로 임명했다고 밝혔다.이 사장은 기존 재경 업무에 기획과 영업·마케팅 업무를 추가하게 됐다. 재
8	2015-12-31	소비자가	지난해 하반기에 출시된 신차 가운데 소비자들이 가장 선호한 모델로 현대자동차의 '제네시스 EQ900(사진)'가 뽑혔다. 중고차 매매 전문기업 SK엔카직영은 성인 남녀 437명을 대상으로 '올 하반기에
9	2015-12-31	세금절벽	현대차 친환경차 '아이오닉' 기아 '신형 K7' 출격르노삼성 '탈리스만' 1월에 사전 공개해 분위기 조성쌍용차도 SUV 새 모델 출시·한국GM은 판촉 강화 현대차 아이오닉기아차 K7르노삼성 탈리스만국
10	2015-12-31	스마트카·	새해 위기경영을 화두로 내세운 재계가 새로운 사업에 승부를 걸고 나섰다. 전사적인 비용 절감과 함께 사업 구조조정을 통한 생존전략을 펼치면서도 미래 먹거리 사업을 선점해야 하는 절박한 상황
11	2015-12-31	현대차 국	현대차가 국내 영업 수장을 교체했다. 현대·기아차그룹은 지난달 31일 현대자동차 재경본부장 이원희 사장(56·사진)을 기획, 영업·마케팅, 재경 담당 사장으로 임명했다.이번 인사는 김충호 전 사장이
12	2015-12-31	제네시스	올 하반기 출시된 신차 가운데 제네시스 EQ900이 최고의 모델로 선정됐다. 국내 최대 중고차 매매 전문 기업 SK엔카직영이 12월 성인남녀 437명을 대상으로 '올 하반기 출시된 신차 중 최고의 모델'
13	2015-12-31	권문식 현	[머니투데이 박상빈 기자] 권문식 현대기아차 부회장/사진제공=현대자동차그룹사단법인 한국자동차공학회는 내년 1월1일부로 제29대 회장에 권문식 현대·기아자동차 부회장(연구개발본부장·61·사진
14	2015-12-31	유일호 경	[머니투데이 세종=조성훈 기자] [편집자주] 박근혜 정부 3기 경제팀을 이끌 경제수장으로 유일호 부총리 겸 기획재정부 장관 후보자가 지명됐다. 대내·외 여건이 호락호락하지 않은 상황에서 유 후보자
15	2015-12-31	[사설]구조	[동아일보]금융감독원은 금융권 대출 500억 원 이상인 대기업 386개사를 대상으로 신용위험평가를 실시한 결과 19곳을 구조조정 대상으로 선정했다고 어제 밝혔다. 이 가운데 C등급 11곳은 기회를
16	2015-12-31	경영악화	[동아일보]임금 동결-무분규 타결 잇따라쌍용차 옥쇄파업 사태는 노조로 인해 경쟁력을 상실한 한국 제조업의 단면을 여실히 보여 준 사례였다. 한국 제조업은 강성 노조로 구조조정이 어려운 데다 미
17	2015-12-30	무능·무책	[한겨레] 현장에서 "(현대제철이) 알려오지 않았고, (공정거래위원회에) 질의도 하지 않았다." 공정거래위원회가 현대제철과 현대하이스코의 합병으로 발생한 순환출자 고리가 공정거래법 위반이 될
18	2015-12-30	국내 완성	상반기에 현대차 하이브리드 아이오닉, 한국지엠 전기차 볼트, 도요타 프리우스 4세대, 아우디 최초 PHEV 출시하반기에 기아차 하이브리드 2종, 니로·신형 K7 출시

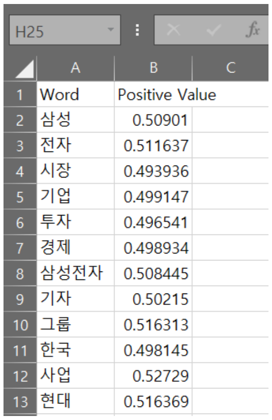
Figure 8. Part of News Articles

Table 1. Amount of collected news data for each company

Company	Amount of collected news data
A	27347
B	24795
C	3858
D	11265

4.3 Building an Sentiment Dictionary

Using news article data and stock fluctuation data, calculate frequency, positive figure, positive index and positive index per day in order and build sentimental dictionary for each company. News article is the python the programming language, and stock fluctuation data were extracted from 'Yeong Ung Mun' which is HTS(Home Trading System) of Ki Um Securities Inc. the Internet stock firm. Through preprocessing of data candidate words will be extracted and those have highest frequency of top 1% will be registered in sentimental dictionary and positive figure of each word will be calculated. Among sentimental words, those of only one letter such as '억' and '판' are mostly used to refer to units or something else so they were judged to be inappropriate for having sentiments which makes them be eliminated from sentimental dictionary. Furthermore, morphemes of numeral like '1200', '200' and '31km' are thought to indicate only quantity, distance and degree thus they were also eliminated from the dictionary. Through the process above candidate words will be narrowed and by calculating positive figure(PF), positive index and frequency of each word, sentimental dictionary for sentimental analysis will be finalized. [Figure 9] is a part of the sentiment dictionary and [Table 2] shows the number of candidate words for each company and the number of words that are registered in real sentimental dictionary.



	A	B	C
1	Word	Positive Value	
2	삼성	0.50901	
3	전자	0.511637	
4	시장	0.493936	
5	기업	0.499147	
6	투자	0.496541	
7	경제	0.498934	
8	삼성전자	0.508445	
9	기자	0.50215	
10	그룹	0.516313	
11	한국	0.498145	
12	사업	0.52729	
13	현대	0.516369	

Figure 9. Part of Sentiment dictionary

Table2. Number of Words for each company

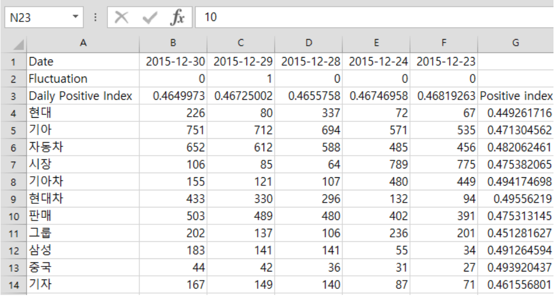
Company	Candidate words	Sentiment words
A	237408	1983
B	214012	1968
C	58677	556
D	142133	1122

4.4 Sentiment Analysis

Based on the method of sentimental analysis of Dong-yeong Kim suggested in chapter 3 sentimental analysis applies to sentimental dictionary of each company. The analysis was conducted in automated fashion by changing suggested equation of sentimental analysis to function of MS Excel and with positive index per day, frequency of sentimental words per day, positive index(PI) of sentimental word, item of stock fluctuation and opening day of stock market, data set for machine learning for each company was set. [Figure 10] is a part of the data set and input PI per day and PI of word are presented as the figures that are rounded off to four decimal places from their original figures. From second row of second column whether stock of that day has increased or decreased is indicated, 0 means down and 1 means up. From first row of fourth column there are words that are registered in the dictionary and from second row of fourth column frequency of sentimental words per date are written.

The data subject of this paper, NAVER news, is published every day regardless of weekends and holidays but data about the next day's stock fluctuation do not get renewed on weekends and holidays. Even though, if news' are used on the day when there is no data on stock fluctuation it might cause a waste of data and inappropriate outcome of sentimental analysis so in case of above, from the day when data on stock fluctuation to the day when that data will be generated, news data were added into one datum. In this research, the research was carried out with the belief that the news article about a certain date would affect the stock price on that day so for an example for the situation above, it is written as follows. News published on December 25th, 2015 was expected to affect the stock price on that day but because December 25th is Christmas which is a holiday data for the next day's stock cannot be generated. As an alternative, sentimental analysis could be practiced by adding data of stock price from 25th to 28th. Thanks to this, the amount of data of the day when the holiday ends and Monday could be twice to 4 times as big as other days.

In this paper, four sample companies' stock market was open 248 times in January 2nd, 2015 to December 31th, 2015 and New Year holiday excluded from the analysis date. [Table 3] is fluctuating distribution of each company per stock, the [Figure 11] is the average of the daily positive index for each company and entire data. Daily positive index average is 0.473 of the total data, and company A 0.502, B company 0.426, C company 0.493, D company were derived the value of 0.471. Signed value shows the actual values to the nearest point on the fourth digit.



N23 10

	A	B	C	D	E	F	G
1	Date	2015-12-30	2015-12-29	2015-12-28	2015-12-24	2015-12-23	
2	Fluctuation	0	1	0	0	0	
3	Daily Positive Index	0.4649973	0.46725002	0.4655758	0.46746958	0.46819263	Positive index
4	현대	226	80	337	72	67	0.449261716
5	기아	751	712	694	571	535	0.471304562
6	자동차	652	612	588	485	456	0.482062461
7	시장	106	85	64	789	775	0.475382065
8	기아차	155	121	107	480	449	0.494174698
9	현대차	433	330	296	132	94	0.49556219
10	판매	503	489	480	402	391	0.475313145
11	그룹	202	137	106	236	201	0.451281627
12	삼성	183	141	141	55	34	0.491264594
13	중국	44	42	36	31	27	0.493920437
14	기자	167	149	140	87	71	0.461556801

Figure 10. Part of data set

Table3. Number of Fluctuation for each company

Fluctuation Company	Rise	Fall
A	122	126
B	106	142
C	124	124
D	113	132

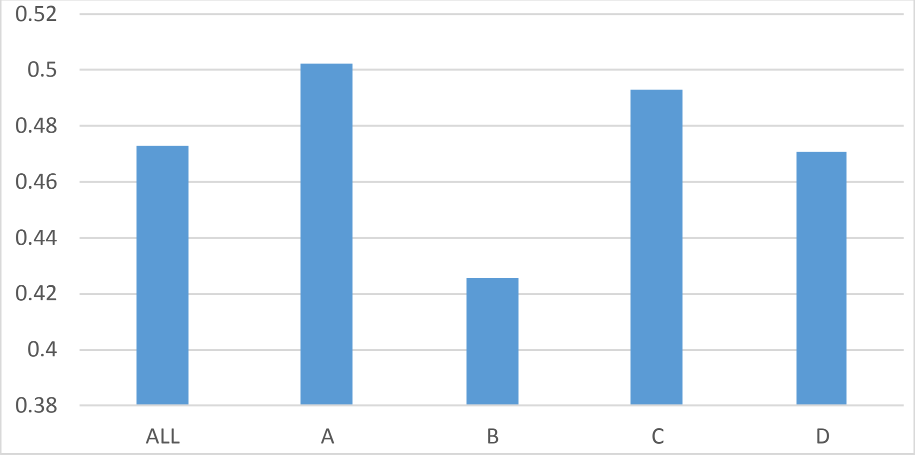


Figure 11. Average of the Daily Positive Index

4.5 Machine Learning

Depending on the stock price prediction model presented in Chapter 3, the data set purified in each step and used to utilize the machine learning model generate a prediction model which is corresponding to each machine learning classification techniques. In this paper, we use big data analysis package, RapidMiner®, which measures the performance of a data set through machine learning. RapidMiner providing Logistic Regression, SVM, Decision Tree, Artificial Neural Network of machine learning classifiers and not only basic functions, addition text mining, web mining technology which are specialized in sentimental analysis. In this paper, January to September data, 2015 is for performance evaluation data (test set), October to December data, 2015 is for performance comparison data (validation set) and performance verification method was 10-fold cross-validation.

4.6 Result

Proceeding 10-fold cross-validation after split performance verification data as learning data and evaluation data, it shows more than 70% in a number of model accuracy and F1 score closer to 0.7. In the result sector, D company's value was comparatively low that average accuracy was 60% and F1 score was near from 0.5. Even, classification technique, decision tree recorded significantly low F1 score 0.3707. The comparison of precision and F1 score based on A company, the accuracy and F1 score of artificial neural network was 77.78% and 0.7766. And the accuracy and F1 score of decision tree was 75.73% and 0.7550. Through this result, artificial neural network shows most good performance and decision tree turn up worst performance. The artificial neural network is appropriate for process of 10-fold cross-validation. The list of the accuracy of each classification techniques with descending order, AHN(77.78%), SVM(77.7%), Logistic regression(76.77%), decision tree(75.73%). [Table 4], [Figure 12], [Figure 13] exhibited each company's accuracy and F1 score.

After machine learning performance verification data, the accuracy is close to 70% of all four companies showed up and A, B, C company got F1 score of average near from 0.7. B company's F1 score doesn't reach to 0.5 in a large number of classification techniques. The list of arithmetic mean value of the accuracy and F1 score of the prediction model for each companies is as follow as. Order of each classification techniques' accuracy is Decision tree(71.425%), Logistic regression(69.638%), AHN(69.043%), SVM(66.663%) and Logistic regression(0.645) and SVM(0.645) got equal value in F1 score and Decision tree(0.634), AHN(0.598) stand in a line. Test result, came out highest accuracy from Decision tree 71.425% and came out highest F1 score 0.645 from Logistic regression and SVM. Accuracy and F1 score of verification data were compiled in [Table 5], [Figure 14] and [Figure 15]

Table 4. Accuracy and F1 Score for Each Company (Test Period)

Classification techniques	Measure	A	B	C	D
Logistic Regression	Accuracy(%)	76.77	75.73	73.34	59.22
	Fl score	0.7656	0.7379	0.7283	0.5355
ANN	Accuracy(%)	77.78	72.82	71.84	61.17
	Fl score	0.7766	0.7348	0.7503	0.6079
Decision Tree	Accuracy(%)	75.73	70.87	69.42	54.37
	Fl score	0.7550	0.6539	0.6725	0.3707
SVM	Accuracy(%)	77.7	75.73	75.24	61.17
	Fl score	0.7766	0.7174	0.7174	0.6086

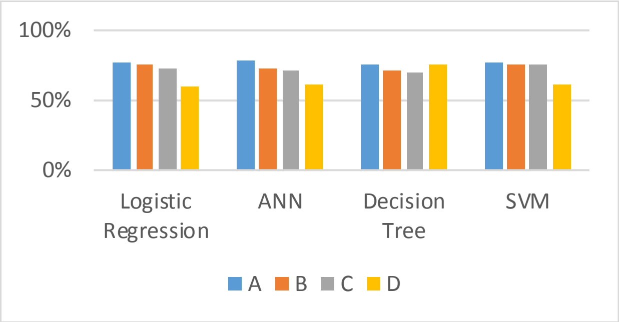


Figure 12. 10-Validation accuracy

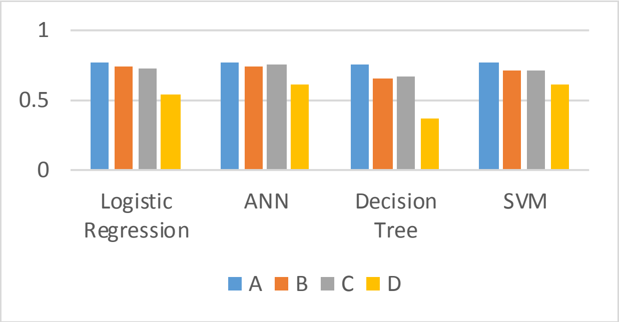


Figure 13. 10-Validation F1 Score

Table 5. Accuracy and F1 Score for Each Company (Validation Period)

Classification	Measure	A	B	C	D
Logistic Regression	Accuracy(%)	71.42	57.14	76.19	73.8
	Fl score	0.7035	0.4094	0.7407	0.7255
ANN	Accuracy(%)	78.57	61.9	64.28	71.42
	Fl score	0.7597	0.4603	0.4457	0.7255
Decision Tree	Accuracy(%)	78.57	61.9	80.95	64.28
	Fl score	0.7846	0.3550	0.8055	0.5906
SVM	Accuracy(%)	76.19	59.52	57.14	73.8
	Fl score	0.7846	0.5073	0.5728	0.7136

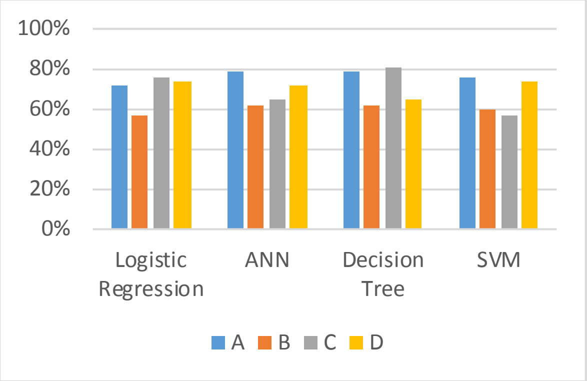


Figure 14. Validation Set Accuracy

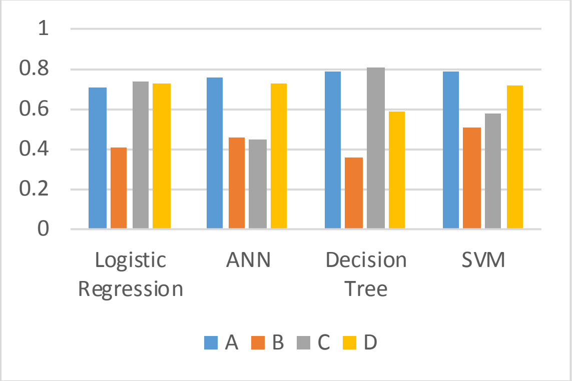


Figure 15. Validation Set F1 Score

5. Conclusion

In this paper, to predict rise and downfall in the next day's stock fluctuation news data were sentimental-analyzed and the method of machine-learning data set was used. Data-collecting program was developed autonomously using computer programming language, and collected data are used to build a sentimental dictionary by getting refined in MS Excel of database. To use news articles which are atypical data for machine learning, the work of digitizing data by applying the method of sentimental analysis to the data refined based on the sentimental dictionary. Next, digitized data set was machine-learned to create prediction model. As the result of the experiment, looking at the average accuracy and average F1 score of each method of classification, high accuracies were checked in the order of Logistic regression(average accuracy -70.451%, average F1 score-0.668), Artificial neural network(average accuracy -69.973%, average F1 score-0.658), SVM(average accuracy -69.561%, average F1 score-0.675), Decision tree(average accuracy -69.511%, average F1 score-0.623). Following the result of the experiment, the method of classification which showed the highest predictability among prediction models built in this paper was Logistic regression analysis, therefore it is considered to be suitable to use that analysis in the process of machine learning. However, in the process of decuple cross verification, when doing the decuple cross verification with arithmetic average figure of each average accuracy and average F1 score of SVM recording 72.46%, 0.705, it looks clear that use of SVM is more suitable.

Through this research, one method for stock price prediction is suggested which is suggested through news data that are the most closely related to stock among atypical data which are considered as the subject for big data analysis recently. By using the morpheme analyzer it was possible to improve the accuracy of sentimental analysis with the building of optimized sentimental dictionary for each company and was possible to apply the method of sentimental analysis in the process of refining data to use for machine learning. Lastly, after making a prediction model by machine-learning the data set, the operation of comparing the performances of models was put into action.

References

Kim, D., A study on Stock-price Prediction Model using Sentiment Analysis and machine Learning Based on News Articles, *Soongsil University*, Seoul, 2015.

Kim, Y., News Big Data Opinion Mining Model for Prediction KOSPI Movement, *Kookmin University*, Seoul, 2013

Seo, J., Polarity Classification using the Features of the Product of Opinion Mining, *Seoul National University of Science and Technology*, Seoul, 2014

Lee, K., and Lee, H., A Study on the Combined Decision Tree(C4.5) and Neural Network Algorithm for Classification of Mobile Telecommunication Customer, *Chungnam University*, Chungchungnam-do, 2014

Gu, Y., Comparative Analysis of Prediction Taekwondo Trainee's Defection using Decision Tree and Logistic Regression, *Hanyang University*, Seoul, 2007

Kim, D., Study on the Lexicon Development for Public Opinion Trend Analysis on Social Media: a Case Study of Twitter Opinion on Nuclear Power, *Hanyang University*, Seoul, 2015

Park, J., Sentiment Analysis of Twitter Data by Using Support Vector Machine Learning, *Yonsei University*, Seoul, 2015

Sa, G., Hotel service quality evaluation using sentiment analysis of online reviews, *Gyeongbuk University*, Gyeongsangbuk-do, 2016

Song, E., The Sensitivity Analysis for Customer Feedback on Social Media, *Namseoul University*, Seoul, 2015
Um, J., Stock fluctuation prediction using the ARIMA model and text mining, *Soongsil University*, Seoul, 2015