Expert Systems With Applications 164 (2021) 113988

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Predicting stock movements based on financial news with segmentation

Nohyoon Seong [a], Kihwan Nam [b,*]

[a] Korea Advanced Institute of Science and Technology College of Business, 85 Hoegiro Dongdaemoon-Gu, Seoul 130-722, Republic of Korea
[b] Management Information Systems, Dongguk Business School, Dongguk University, 30, Pildong-ro, 1-gil, Jung-gu, Seoul, Republic of Korea

ARTICLE INFO

Keywords:
Stock prediction
Data mining
Machine learning
Heterogeneity
Cluster analysis

ABSTRACT

With the development of machine learning technologies, predicting stock movements by analyzing news articles has been studied actively. Most of the existing studies utilize only the datasets of target companies, and some studies use datasets of the relevant companies in the Global Industry Classification Standard (GICS) sectors. However, we show that GICS has a limitation in finding relevance regarding stock prediction because heterogeneity exists in the GICS sectors. To solve this limitation, we suggest a methodology that reflects heterogeneity and searches for homogeneous groups of companies which have high relevance. Stock price movements are predicted using the K-means clustering and multiple kernel learning technique which integrates information from the target company and its homogeneous cluster. We experiment using three-year data from the Republic of Korea and compare the results of the proposed method with those of existing methods. The results show that the proposed method shows higher predictability than existing methods in the majority of cases. The results also imply that the necessity of cluster analysis depends on the heterogeneity in the sector, and it is essential to perform cluster analysis with a larger number of clusters as the heterogeneity increases.

1. Introduction

Stock price forecasting is known to be difficult due to interactions among factors affecting stock price fluctuations (Schumaker & Chen, 2009). The effects of various factors, such as national policies, diplomatic relations, and the psychological impact of investors, must be considered when predicting stock movements. In theory, because stock prices are expected values for future performance, investors try to analyze the value before investing. Various kinds of information can be used to assess value, which in turn changes investor sentiment. In other words, investors' psychology, and new information have a high correlation, and information plays an important role when looking at the value of future performance (Campbell & Shiller, 1987; MacKinlay, 1997). That is, information is essential when estimating future stock prices. This phenomenon has become more prominent, especially as it has become possible for people to get information from the Internet where abundant information is passed on to investors. In theory, the stock market fluctuates when new information comes (Boguth, Carlson, Fisher, & Simutin, 2016). In other words, analyzing information, such as various media and time series data of formal stocks, is vital for forecasting stock prices.

With the introduction of the Big-Data era, a tremendous amount of financial news is produced every day, and it has become difficult for an individual to selectively search for information that affects stock prices in numerous news stories. Therefore, text mining techniques for automatically analyzing and predicting financial news have been developed and studied academically. For example, Bag-of-words (De Fortuny, De Smedt, Martens, & Daelemans, 2014; Nam & Seong, 2019), n-grams (Hagenau, Liebmann, & Neumann, 2013), topic modeling (Nguyen, Shirai, & Velcin, 2015), and word embedding (Kraus & Feuerriegel, 2017) have been widely used.

Generally, previous studies selected only the most relevant articles of the target companies to predict stock prices with text mining. The way to select relevant articles is usually to choose based on a ticker (Nam & Seong, 2019; Shynkevich, McGinnity, Coleman, & Belatreche, 2016), whether the target company includes its name in the headline (Nassirtoussi, Aghabozorgi, Wah, & Ngo, 2015), or whether the target company includes its name in the news body (Li et al., 2014). However, if the good news from a camera maker comes out, then the stock price for a related lens part company will increase. In other words, finding relevant companies is an important problem in prediction (Shi, Lee, & Whinston, 2016).

* Corresponding author at: Management Information Systems, Dongguk Business School, Dongguk University, 30, Pildong-ro, 1-gil, Jung-gu, Seoul, Republic of Korea.
E-mail address: namkh@dongguk.edu (K. Nam).

https://doi.org/10.1016/j.eswa.2020.113988
Received 15 November 2018; Received in revised form 7 September 2020; Accepted 7 September 2020
Available online 16 September 2020

However, only some papers investigated the prediction considering relevance between the companies (Nam & Seong, 2019; Schumaker & Chen, 2009; Shynkevich et al., 2016). Especially, Shynkevich et al. (2016) showed that it is more accurate to predict stock movements considering relevance based on the Global Industry Classification Standard (GICS) than to predict them merely based on a target company. Although research has progressed gradually on the basis of the influence within the GICS sector, it has been conducted based on the assumption that a GICS sector reflects the relevance of the stock price and all sectors have high homogeneity (Schumaker & Chen, 2009). By examining the diverse sectors, however, we find that some of GICS sectors are not a homogeneous group.

Moreover, GICS has limitations that (1) it can't dynamically reflect similarity among companies because it is updated yearly, and (2) it does not reflect aspects of the company's various products (For example, although significant amount of revenue of Amazon is from Amazon Web Service, Amazon is in the Consumer Discretionary sector). However, so far, no attempts have been made to overcome these limitations. To close the research gap, we propose a novel methodology that reflects heterogeneity and dynamically searches for homogeneous groups of companies that have high relevance.

To close the research gap, we propose a novel algorithm that reflects heterogeneity and dynamically searches for homogeneous groups of companies that have high relevance and forecasts stock movements. We use clustering analysis that makes heterogenous sectors homogeneous according to stock patterns, and build a stock prediction system using text mining in relevant companies with results of the clustering analysis. As a result, we find that the proposed method has better predictive power than conventional techniques. We also note that if there are sectors with heterogeneous characteristics, it is advantageous to divide the sectors into a large number of clusters so that they have homogeneous characteristics. If there are sectors with homogeneous characteristics, it is beneficial to divide the sectors into a small number of clusters because the companies having similar characteristics are tied together already.

In this work, we make three main contributions. First, we are able to achieve higher performance by successfully suggesting a method to find relevant companies using data mining techniques to reflect heterogeneity in the industry classification system. To the best of our knowledge, this is the first paper that has studied this component. Second, we contribute to stock prediction literature by identifying that the heterogeneity in the GICS sector (Hagenau et al., 2013). This suggests that the pre-defined classification standard, which is widely used in traditional finance and the economy (Nam & Seong, 2019; Shynkevich et al., 2016), may not be suitable for the stock price forecast. In other words, when designing a stock price prediction machine learning algorithm, it is important to understand the characteristics of stock price data closely. Finally, we contribute to market clustering literature by analyzing the level of heterogeneity to determine when to use clustering analysis to find relevant companies (Nanda, Mahanty, & Tiwari, 2010). Since heterogeneity varies among financial markets and sectors, it is important to set the criteria to fit the data. The methodology in this study proposes a method for determining the threshold.

The remainder of this paper is organized as follows. In Section 2, we show the overall flow of existing related papers that predict the impact of the stock price on qualitative information. In Section 3, we present a research model for forecasting stocks with financial news data and compare it with previous studies on the industry classification system (Schumaker & Chen, 2009; Shynkevich et al., 2016). Next, we describe the experimental results in Section 4. In Section 5, we describe implications of the research. Finally, we conclude this study and discuss limitations and guidelines for future research in Section 6.

2. Literature review

2.1. Stock price

2.1.1. Stock price theory (efficient market hypothesis)

The price of the stock goes up and down reacting to the actual transaction of the seller and the buyer in the market and finds the proper point. Therefore, the stock price is determined based on the law of supply and demand (Nassirtoussi, Aghabozorgi, Wah, & Ngo, 2014). The stock price will rise when demand for stocks increases as investors think business performance is good, and fall when supply increases as investors believe corporate performance is weak. In other words, the stock price is the expected value of the company's future performance. Diversiform information, such as historical stock prices, financial news, and postings on social networks, can influence the company's expectations regarding future performance.

Efficient Market Hypothesis (EMH) is the theory that market price accurately reflects a company's value and reacts only to new information (Fama, 1995). The EMH is divided into three categories according to the extent that information is reflected: Weak form efficient market, Semi-strong form efficient market, and Strong form efficient market.

- Weak form efficient market

In the Weak form efficient market, prices of all financial assets such as stocks, bonds and tangible assets that can be traded in the current financial market already reflect all information of the past stock price. Therefore, analysis via technical analysis cannot produce excessive profits.

- Semi-strong form efficient market

In the Semi-strong form efficient market, all public information such as past stock price changes, quarterly earnings reports, and financial news is already reflected in the price of financial assets. Therefore, forecasts using public information cannot generate excessive profits in the market.

- Strong form efficient market

In the Strong form efficient market, all public information and private information are already reflected in market prices. Therefore, prediction using any information cannot yield excessive profits in the market.

2.1.2. Test of the Weak form efficiency: Hurst Exponent

The Weak form efficient market is a state in which historical stock prices are already reflected in the current stock price. Several measurement methods have been tried to measure these market states quantitatively, and one of the concepts mainly used in econophysics is the Hurst Exponent (Zunino, Olivares, Bariviera, & Rosso, 2017).

The Hurst Exponent is a concept to measure long-term memory quantitatively (Hurst, 1951). When considering $x_i$as the point of the time series and using the DFA method, the Hurst exponent, H, is defined as follows (Lahmiri, 2015).

$$y_i = \sum_{i=1}^{N} (x_i - \bar{x}) \tag{1}$$

$$F(n) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} [y(i) - y_n(i)]^2} \tag{2}$$

$$F(n) = c{*}n^H \tag{3}$$

If the Hurst Exponent is lower than 0.5, the market trend can be interpreted as mean reverting, and it can be interpreted as a random walk if 0.5, or trending if higher than 0.5. In other words, the higher the Hurst Exponent, the more significant the impact of past prices on the current price. That is, in the market has the Hurst Exponent higher than 0.5, we can predict the future stock price with a pattern of the stock history. Therefore, if the Hurst Exponent is below 0.5, the stock market cannot be predicted using technical Analysis, which means the Weak form efficient market

As seen in Zunino et al. (Zunino et al., 2017), the Republic of Korea has the Hurst Exponent of around 0.5 which means Korean stock market follows the Weak form efficient market. Since Korean stock prices do not have long-term memory and have a random walk form, we cannot forecast stock movements using stock history. Therefore, when analyzing stock prices in the Korean market, it is meaningful to use news rather than technical indicators. Therefore, in this paper, we use data from the Korean stock market, and we forecast the stock price using financial news.

2.2. *Key related research*

There are has three main differences in the literature that constitute forecasting of financial markets based on news articles. They are text preprocessing, a machine learning algorithm, and a dataset. Then, text preprocessing includes feature extraction, feature selection and feature representation. Table 1 summarizes the papers that played an essential role in this study.

Deng et al. (Deng et al., 2011) used an algorithm for predicting stock prices using data with various characteristics such as sentiment analysis of social network service, technical analysis, and statistical characteristics of news. Generally, support vector regression (SVR) is one of the main choices for the machine learning algorithm. However, SVR has the limitation that it can only use data that has a single characteristic. Therefore, to overcome this limitation, multiple kernel learning was used, and it showed better results than SVR which only uses a single kernel.

Hagenau et al. (Hagenau et al., 2013) designed a system that automatically received corporate announcements and financial news and predicted stock prices through text mining. The authors performed Bag-of-words, 2-Gram, 2-word combination and Noun phrases for feature extraction, and chi-square and bi-normal-separation for feature selection. High accuracy rates were shown using various feature extraction methods. In this paper, we will use Chi-square feature selection and TF-IDF weighting for the Bag-of-words model as used in Hagenau et al. (Hagenau et al., 2013).

Schumaker and Chen (Schumaker & Chen, 2009) proposed the Arizona Financial Text System (AZFinText), a process for predicting stock price after constructing a dataset with news articles, trading experts, and stock quotes. Using AZFinText, the authors investigated how to construct a dataset and predict the stock price.

Also, the authors divided the financial news affecting the stock price into several groups of datasets according to the GICS system: Sector-based, Sub-Industry-based, Industry-based, Group-based, and Stock-specific news articles. The authors predicted the stock prices with each dataset using all of the related companies classified as Sector-based, Sub-Industry-based, Industry-based, Group-based, and Stock-Specific news articles. As a result, not only is Stock-Specific news effective but also Sector-based news is also effective in predicting stock prices of the target company. However, limitation is that the features at various levels are not used simultaneously. Therefore, in this paper, multiple kernel learning is used to integrate several levels of features simultaneously.

Shynkevich et al. (Shynkevich et al., 2016) developed a system for forecasting stock prices by incorporating a group of news of different levels of relevance. The authors created a group of news - Sector based, Sub - Industry-based, Industry-based, Group-based, and Stock specific news articles - based on GICS and compared them using multiple kernel learning on all of the companies. The authors found that predicting the stock price considering the relevance of other levels is better than predicting the stock price alone. However, the authors' analysis was based on the assumption that relevance would be high in the same GICS system (Shynkevich et al., 2016). However, even if they are in the same industry, relevance is not high according to our study. Therefore, in this paper, we use the market segmentation rather than merely using the GICS system to find the relevance between firms.

Fig. 1 shows the stock price forecasting process in summary. It has two research streams: technical analysis and media effect analysis (Nassirtoussi et al., 2014). Concerning technical analysis, the Korean stock market follows the weak form efficient market so that prediction

Table 1
Key Related Research.

Reference	Dataset	Feature Extraction	Feature Selection	Feature representation	Machine Learning Algorithm	Forecast Type	Other Effects
Schumaker and Chen (Schumaker & Chen, 2009)	Financial news	Bag-of-words, Noun phrases, Named entities, Proper Nouns	Minimum occurrence per document	Binary	SVR	Price value	Relevant company's financial news
Hagenau et al. (Hagenau et al., 2013)	Corporate announcement and financial news	Bag-of-words, 2-Gram, 2-word combination, Noun phrases	News frequency, Chi-square, Bi-normal-separation	TF-IDF	SVM	Up and Down	None
Shynkevich et al. (Shynkevich et al., 2016)	Financial News	Bag-of-words	Chi-square	TF-IDF	Multiple kernel learning	Up and Down	Relevant company's financial news
De Fortuny et al. (Junqué de Fortuny et al., 2014)	Financial News	Bag-of-words	–	TF-IDF	SVM	Up and Down	Technical Indicators
Groth and Muntermann (Groth & Muntermann, 2011)	Corporate disclosures	Bag-of-words	Chi-square, Information Gain	TF-IDF	Naïve Bayes, k-NN, ANN, SVM	Up and Down	None
Bollen et al. (Bollen, Mao, & Zeng, 2011)	Twitter tweets	Sentiment Analysis	Sentiment Analysis	Sentiment Analysis	SOFNN	Stock Price and Up and Down	None
Deng et al. (Deng, Mitsubuchi, Shioda, Shimada, & Sakurai, 2011)	Social network sentiment, news comment	Sentiment Analysis	Sentiment Analysis	Sentiment Analysis	Multiple kernel learning	Price return	Comment, technical analysis

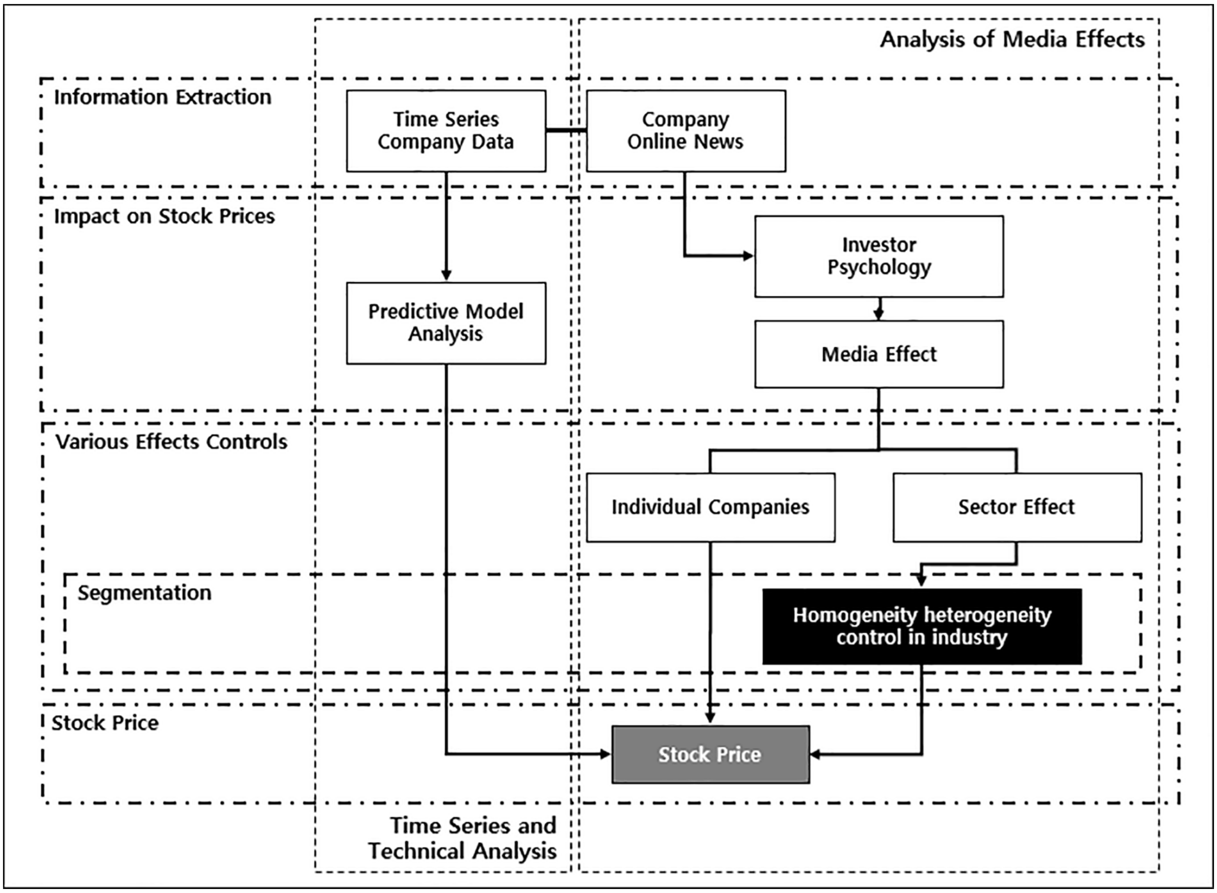

Fig. 1. The concept of stock price prediction through news.

via technical analysis cannot produce a substantial effect.

Regarding media effect analysis, financial news includes corporate-wide conditions such as financial status and economic activities of a specific company. Investors' psychology differs significantly according to the same financial news. According to Behavioral Finance and Investment Psychology theory, it has been confirmed that investor behavior can be determined by whether they feel optimistic or pessimistic about the future market value (Bollen et al., 2011). In addition, investor sentiment can have an impact on individual firms and can have an effect on the industrial sector. Therefore, the combined analysis of these industry sectors has high predictability (Shynkevich et al., 2016). In this study, the heterogeneity, which is the limitation of the existing research at the industrial level, is solved by the data mining technique and prediction is improved.

2.3. *Text preprocessing*

Once we receive the news, we need to convert the text data into meaningful data that the machine learning algorithm can use. More precisely, we must extract and represent features that affect the stock price in news data. For example, the word 'capitalization' has a profound impact on the company's stock price. In other words, words and sentences that affect the stock price have to be appropriately extracted. This process is called preprocessing. As in Hagenau et al. (Hagenau et al., 2013), text preprocessing follows three steps: Feature extraction, Feature selection, and Feature representation.

Feature extraction is the process of creating features that can predict stock price movements. This part consists of extracting words and their combinations from a document. According to Nassirtoussi et al. (Nassirtoussi et al., 2014)), the Bag-of-words approach is the most popular feature extraction method in the field of stock price prediction through text mining. In order to carry out Bag-of-words, we remove meaningless text such as e-mails, and we remove punctuation through morphological analysis and finds infinitives. At this time, sets of words are features that represent the article.

Feature selection is to select features that affect the stock price among the features found in the Bag-of-words model. Chi-square is mainly used to determine the explanatory power of features to forecast direction (Groth & Muntermann, 2011; Hagenau et al., 2013; Shynkevich et al., 2016). Chi-square is a statistical method based on the chi-square distribution and is used to verify whether the observed frequency, O_{ij}, is expected to differ significantly from the expected frequency, E_{ij}. The Chi-square value is as follows.

$$\aleph^2 = \frac{\sum(O_{ij} - E_{ij})}{E_{ij}} \tag{4}$$

If a feature appears uniformly in all documents, it does not have any meaning, and it has the same expected frequency and observed frequency so that the chi-square value is low. Likewise, if a feature appears only in positive documents, it has significant meaning, and the difference of the expected frequency and the observed frequency is high. Accordingly, a feature having a high value of Chi-square is meaningful (Hagenau et al., 2013), and we only chose meaningful data.

After extracting the influential variables through the feature selection, features should be revised to have a form suitable for machine learning. This process is called Feature Representation. If a particular word frequently comes out and the stock price rises or goes down too much, the feature should be weighted. Additionally, some words that appear only in a small number of documents rather than in all documents are more meaningful, and it should be reflected. Therefore, we used TF-IDF (Term Frequency-Inverse Document Frequency), the most commonly used form (Groth & Muntermann, 2011; Hagenau et al., 2013).

2.4. Machine learning prediction techniques

After text preprocessing is completed, the machine learning algorithm predicts the stock price based on features extracted from the text data. A variety of algorithms has been used in the existing literature: Support Vector Machine (SVM), Artificial Neural Network (ANN), k-Nearest Neighbors (kNN), and Naïve Bayes (NB). Groth and Muntermann (Groth & Muntermann, 2011) used machine learning algorithms - ANN, SVM, NB, and KNN - for risk management and investment decision-making using textual analysis and the results among algorithms were compared. The authors recommended SVM considering both the result and time efficiency. Also, Hagenau et al. (Hagenau et al., 2013) divides the market states into positives and negatives, and predicts movement using SVM, ANN, and NB with financial messages. SVM performed better than the other algorithms. In the flow of research, SVM is the most common and accurate algorithms for predicting stock movements with text data.

Recently, ensemble techniques are widely used to predict the stock movements. Ensemble techniques refer to integrating multiple algorithms at the same time to prevent overfitting and to increase predictability. Ensemble techniques also make it easier to apply the algorithm to various situations and improve accuracy.

Multiple Kernel Learning (MKL) is an ensemble technique which is the extension of the SVM (Aiolli & Donini, 2015). The most significant difference between SVM and MKL is that SVM can only use data with a single characteristic, but MKL can use various data at the same time combining multiple kernels. Deng et al. (Deng et al., 2011) predicted stock price through multiple kernel learning by integrating news text data, comments, and stock history. As a result, the combining of different features using different kernels was more effective than that of a single kernel. Shynkevieh et al. (Shynkevich et al., 2016) predicted stock prices using Multiple Kernel Learning for news data from various economic groups based on the GICS system. As a result, predictions made using the financial news of companies having relevance rather than using only corporate news showed better prediction results. However, to reduce the heterogeneity of complex systems, no existing literature increases synchronization by clustering companies, and predicts stock prices.

In this paper, we apply cluster analysis to heterogeneous groups and predict stock prices at various levels. Therefore, we used the multiple kernel learning method to use numerous datasets simultaneously.

2.5. Clustering by homogeneous group

Subdivision by the Homogeneous Group has been applied in various fields. The most active field is market segmentation.

Over fifty years ago, market segmentation was established (Smith, 1956). Since then, market segmentation has been common and widely used in many business sectors as a tool to manage various customer necessities as well as to target marketing resources (LaPlaca, 1997; McDonald & Dunbar, 2004; WEINSTEIN, 2004). The generalized theorem is that heterogeneity, as well as purchasing behavior, is manageable by putting customers with similar characteristics into sections. Out of these segments, some will be the focal point for the marketing efforts (Kalwani & Morrison, 1977; Mahajan & Jain, 1978).

Much research has been carried out from the data mining point of view to proceed with efficient segmentation. Much of the recent success of artificial intelligence can be accredited to data mining techniques and tools (Hsu, Lin, & Yeh, 2013; Martínez-López & Casillas, 2013; Singh, Gupta, Ojha, & Rai, 2013). Out of these methods, clustering has on all occasions been an investigative yet pivotal tool in the process of knowledge discovery. It has also been implemented in almost all domains where the grouping of objects with similar attributes is sensible (Herrera, Pajares, & Guijarro, 2011; Karaboga & Ozturk, 2011; Lam & Tsang, 2012). Clustering methods have garnered interest from the scientific and business community alike, and methods extending from the simplest approaches, like the k-means algorithm (Huysmans, Martens, Baesens, Vanthienen, & Van Gestel, 2006), to the most sophisticated methods, like the kernel method (Zhou, Lu, Yang, Ma, & Tuo, 2011) and the spectral approach (Shang, Jiao, Shi, Gong, & Shang, 2011). Occasionally, those that utilize such methods possess expertise and desire to implement it as a part of the clustering exercise.

In this study, we perform k-means cluster analysis in the GICS sectors based on the background knowledge of the GICS sectors to cluster companies so that they have similar attributes.

3. The proposed approach

This chapter describes in detail how to build a system for performing cluster analysis and predicting stock prices using financial news with highly relevant companies. First, we describe how to cluster relevant companies. Second, how to get textual data, preprocess it and perform feature extraction will be shown. We will also demonstrate how to use machine learning algorithms to predict the stock movement and what evaluation indicators are used. Fig. 2 shows a series of these processes.

3.1. Data

We conducted experiments with actual data to prove propositions in this paper. The dataset consists of financial news and stock price data from January 1, 2014 to December 31, 2016.

3.1.1. Stock history data and heterogeneity test

Historical stock prices are used for feature selection, company selection, group segmentation and data labeling. We received the data from KOSCOM, a company that builds an IT infrastructure on the financial industry. We could get open, close, volume, median, min, and max value of stock price for every company day by day. The return is defined as follows.

$$\mathrm{Return}_t \overset{\mathrm{def}}{=} \frac{Close_t - Open_t}{Open_t} \tag{5}$$

The companies to be used for the experiment are chosen based on the sector-level of the GICS from the KOSPI 200. To identify the clustering effect depending on the heterogeneity of the sectors, we first define the heterogeneity. We define the heterogeneity as variance from the sector index which is calculated in the same way that the stock index is calculated. As a result of examining heterogeneity, we select the Food Expenses sector with low heterogeneity and the Pharmacy sector and Material sector with high heterogeneity.

3.1.2. News data

We crawled all the financial and economic news registered in Naver, the Republic of Korea's largest portal site, which contains 10 comprehensive newspapers, 14 broadcast communication sites, 9 economic newspapers, and 33 internet news sites. A news dataset includes all online financial news available in the Republic of Korea. Therefore, it is an nice dataset to understand the impact of the financial news on stock movements. During the period, there were 1,397,800 records crawled except for duplicates. The format of news data is the title, author, post time, and contents.

According to the influence of news on the stock price, the news is labeled as positive or negative. In other words, each news item is labeled as 1 (Up) if the stock return of the day news appears is larger than 1 and 0 (Down) if the return of that is smaller than 1. For example, news for a company published at 11:00 on Monday is labeled 1 if the return is greater than or equal to 1, or 0 if the return is less than 1. However, since the Korean stock market opens at 9:00 and closes at 16:00, the news after 16:00 and the news published on holidays need to be interpreted as affecting the next day (Li et al., 2014).

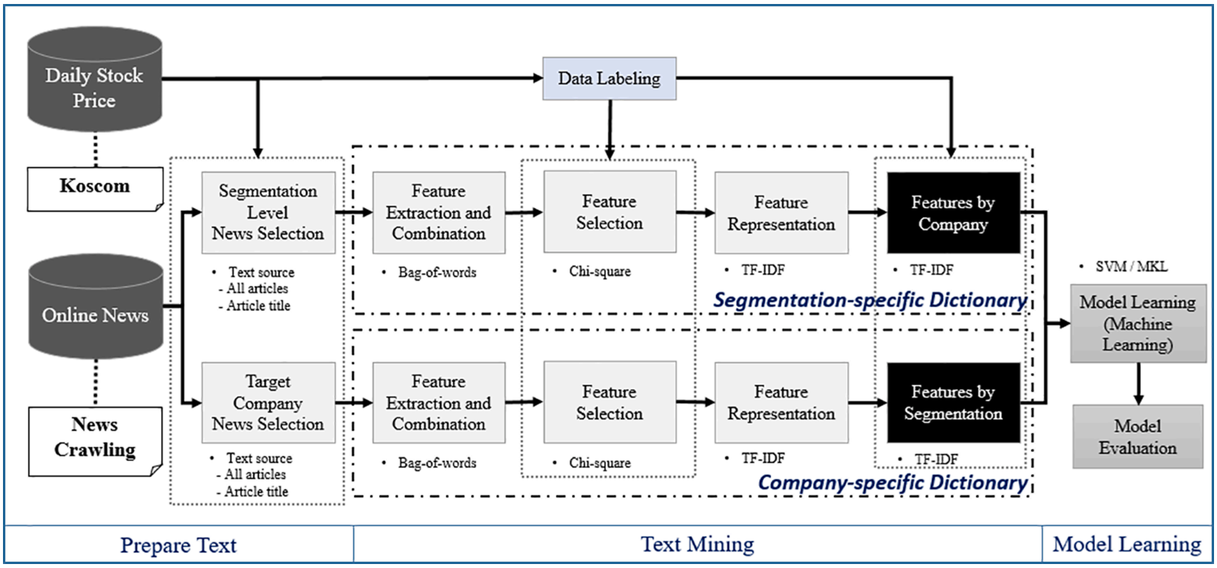

Fig. 2. Proposed approach.

To extract news that is relevant to each company, we assign news that includes the name of the company in the news body. The list of all companies used in the experiment and the up and down labels are shown in Table 2.

3.2. Identification of relevant companies

In this study, segmentation among companies in GICS sectors is conducted to solve heterogeneity in the sector. There are many ways to implement segmentation. K-means clustering is an algorithm that divides data into k clusters. It operates in a way that minimizes the variance of distance between each cluster (MacQueen, 1967). Besides, DBSCAN performs clustering in such a way that every cluster has a specific density (Ester, Kriegel, Sander, & Xu, 1996). CLARNAS (Ng & Han, 1994) and BIRCH (Zhang, Ramakrishnan, & Livny, 1996) are also used. Among them, the widely used algorithm is K-means clustering which is simple to use and useful to process with a large amount of data. Therefore, in this paper, we use K-means clustering for segmentation

Table 2
Up and Down Label of the companies.

Company	Data point	Sector	Up label	Down label	Company	Data point	Sector	Up label	Down label
OCI	1947	material	1080	867	Daewoong Pharm.	873	pharmacy	436	437
Huchems Fine Chemical	175	material	94	81	Green Cross	1583	pharmacy	830	753
Kukdo Chemical	85	material	45	40	Yuhan	854	pharmacy	444	410
Hyundai-Steel	3741	material	1858	1883	Jeil Pharmaceutical	155	pharmacy	89	66
NamHae Chemical	179	material	104	75	Bukwang Pharm.	218	pharmacy	110	108
Hansol Chemical	110	material	71	39	Hanmi Pharm.	3051	pharmacy	1212	1839
Foosung	295	material	142	153	Dong-A ST	497	pharmacy	263	234
SKC	1184	material	658	526	Boryung Pharm.	583	pharmacy	312	271
SKChemical	1276	material	653	623	Hanall BioPharma	193	pharmacy	99	94
SeAh Steel	374	material	207	167	JW Pharm.	489	pharmacy	231	258
KISWIRE	166	material	92	74	C.K.D	1200	pharmacy	625	575
KiscoHolding	762	material	387	375	Yungjin Pharm.	171	pharmacy	84	87
Korea Zinc	619	material	378	241	Ildong Holdings	46	pharmacy	30	16
Ssangyong Cement Industrial	390	material	258	132	Hanmi Science	610	pharmacy	267	343
Lock&Lock	652	material	381	271	Dong-A Socio Holdings	309	pharmacy	126	183
Korea Petrochemical Ind.	190	material	95	95	Il-Yang Pharm	303	pharmacy	180	123
SamKwang Glass	176	material	80	96	Kwang dong Pharm.	721	pharmacy	410	311
Young Poong	970	material	441	529	CJ CheilJedang	5828	food	3019	2809
Hanwha Chemical	2088	material	1116	972	Samyang	342	food	190	152
Poongsan	1112	material	593	519	Ottogi	2087	food	1100	987
Lotte chemical	2992	material	1515	1477	Hitejinro	3764	food	2008	1756
DongKuk Steel Mill	2079	material	1081	998	Namyang	1375	food	800	575
Taekwang Ind.	327	material	150	177	Muhak	1981	food	926	1055
SeAh Besteel	362	material	200	162	KT&G	3349	food	1803	1546
POSCO	1022	material	509	513	Nonhshim	4093	food	1990	2103
Kolon Ind.	972	material	481	491	Farmsco	236	food	114	122
LG Chem	6717	material	3592	3125	Orion	2915	food	1473	1442
Lotte find Chemical Co.	178	material	108	70	Samyang Holdings	309	food	207	102
–	–	–	–	–	Dongwon F&B	1528	food	752	776
–	–	–	–	–	Lotte Chilsung	2932	food	1472	1460
–	–	–	–	–	Binggrae	1205	food	544	661
–	–	–	–	–	HiteJinro Holdings	127	food	61	66
–	–	–	–	–	Lotte Food	1737	food	893	844
–	–	–	–	–	Lotte Confectionery	3937	food	2154	1783

within the GICS sectors.

K-means clustering divides n companies in the sector into k clusters. Stock prices of companies in the same GICS sectors are referred as $x_1, x_2, \cdots x_n$. Let $\mu_1, \mu_2, \cdots, \mu_k$ be the centers of the clusters, and the clusters divided by K-means clustering be $\overrightarrow{S} = \{S_1, S_2, S_3, \cdots, S_k\}$. The algorithm is as follows (MacQueen, 1967).

$$\underset{\overrightarrow{S}}{\text{argmin}} \sum_{i=1}^{k} \sum_{S_i} \|x_i - \mu_i\|^2 \tag{6}$$

In other words, the summation of the distance among data points within a cluster through all clusters is minimized. The algorithm applies the heuristic method of finding the center point after setting the initial centers.

The initial centers are obtained by randomly selecting one center of the data, getting the distance D (x) from the data nearest to the center, and giving the weighted probability proportional to $D(x)^2$ as a new center. This process is repeated and K centers were selected. The initial point setting algorithm is called k-means++ (Arthur & Vassilvitskii, 2007). Since k-means++ solves the problem of K-means algorithm, which takes time exponentially as the amount of data increases, we used k-means++ for efficient computation.

After setting the initial centers, the data is allocated to the cluster satisfying the following conditions.

$$S_i = \{x_p : \|x_p - \mu_i\| \leq \|x_p - \mu_j\|, \forall j, i \neq j\} \tag{7}$$

After that, the center is calculated again, the data points are reallocated to the cluster, and the data points satisfying (7) are searched.

As described above, we need to select how many clusters would be used which is the most crucial issue of K-means clustering (Jain, 2010). A statistically significant K can be chosen as in Latent Class Analysis (Vermunt & Magidson, 2002). By increasing the K sequentially and measuring how similar an object is to one cluster over other clusters, we can set the optimal K (Rousseeuw, 1987). There is also a way to maximize the Bayesian Information Criterion by increasing K sequentially (Pelleg & Moore, 2000). This study, however, does not need a large number of K since it is primarily aimed at capturing heterogeneous points while being tied to similar industries. Accordingly, we apply the results of the optimal case by applying k from 2 as in the recent grid search, which is widely used in the field of data mining. As a result, they are already divided into homogeneous sectors, so there is no optimum result when k exceeded 4. Therefore, in this study, we interpret the results as K up to 4.

3.3. Text preprocessing

Text preprocessing is one of the most important parts of implementing a text-based stock price prediction system. First, we need to remove unnecessary elements from the news, so we remove all redundant information such as email and HTML tags. In order to use the Bag-of-words model, we must make all words infinitives. Therefore, we use morphological analyzer, KKMA POS Tagger, to transform all words into infinitives (Lee, Yeon, Hwang, & Lee, 2010). Each infinitive is a feature. As in Shynkevich et al. (Shynkevich et al., 2016), all infinitives mentioned less than three times are eliminated.

We use the Chi-square test for Feature selection. We select only features with the highest 10% chi-square values in the features. The number of features ranges from 500 to 1000, which is similar to 567 feature selection for bag-of-words in Hagenau et al. (Hagenau et al., 2013) and 500 in Shynkevich et al. (Shynkevich et al., 2016).

After feature selection, each feature is represented using the TF-IDF. The values of the features after TF-IDF representation are too small to efficiently perform machine learning, so a scaling process is required. The number of the features selected is multiplied in the same method of Shynkevich et al. (Shynkevich et al., 2016). Especially, if the number of features in the top 10% is k, then k*TF-IDF. In this paper, we use many features from various sources, so it is an essential scaling operation to prevent bias toward one side.

3.4. Multiple kernel learning

Multiple Kernel Learning (MKL) is a machine learning algorithm to use a predefined set of kernels and learn an optimal combination of kernels (Shynkevich et al., 2016). Multiple Kernel Learning involves the use of weak kernels $K_1, K_2, K_3, \cdots K_n$. After defining K_i, the algorithm calculates K_i through a linear combination of the variables. The following equation can be used to express MKL.

$$K = \sum \beta_i K_i, \forall \beta_i \geq 0 \tag{8}$$

However, the problem of optimizing Multiple Kernel Learning is difficult. Discovering values that are better than mere averages of weak kernels is not simple, so MKL has computational complexity (Aiolli & Donini, 2015).

Accordingly, plentiful time and memory are required to calculate MKL, and there have already been many efforts to solve the problems (Aiolli & Donini, 2015; Jain, Vishwanathan, & Varma, 2012; Sun, Ampornpunt, Varma, & Vishwanathan, 2010). In particular, EasyMKL is an optimization algorithm that maximizes the distance between positive and negative samples by adjusting the weak kernel combination vector $\overrightarrow{\beta}$ and the probability distribution $\overrightarrow{\gamma}$ in the Hilbert space (Aiolli & Donini, 2015). EasyMKL is expressed as follows.

$$\underset{|\beta|=1}{\text{maxmin}}(1-\mu)\beta^T \widehat{Y} \left(\sum_r^R \beta_i \widehat{K_i} \right) \widehat{Y} \beta + \mu |\beta|^2 \tag{9}$$

EasyMKL not only exhibits superior AUC scores in various data sets but also shows much higher memory usage efficiency than that of SPF-GMKL (A. K. Jain, 2010), which is regarded as state-of-the-art. Therefore, in this study, we use the EasyMKL algorithm (Aiolli & Donini, 2015).

In this paper, kernels are assigned to features of the target company and a homogeneous group obtained by clustering. We are unable to determine which kernel would perform best for our text data. Consequently, we utilize the linear kernel, a polynomial kernel with degree 3, Gaussian kernel and their combinations, linear and Gaussian, polynomial and Gaussian, linear and polynomial. In other words, a minimum of two and a maximum of four kernels in combination with either one or two of the linear, Gaussian, and polynomial kernels is assigned to each company and the group they belong to. To show that finding homogeneous groups with clustering gives better results, we set up comparison groups. The first comparison group is SVM with three kernels only using the dataset of the target company. Moreover, another comparison group is predicted by the Multiple Kernel Learning at the sector level of the GICS system.

3.5. Evaluation

Three years of news data and stock price data are used from January 2014 to December 2016 in this study. The training period and test period are set to 2 years and six months and six months for validation. Parameter fitting is performed within the duration of the trading period, and the quality of the result is confirmed in the actual situation during the test period. The evaluation method of the experiment is Accuracy.

Table 3
Confusion Matrix.

		Prediction	
		Up	Down
Actual	Up	TP	FN
	Down	FP	TN

When the prediction is made, if the forecast is ‘Up’ and the actual result is ‘Up,’ it is defined as TP. If the direction predicted is wrong and it is ‘Down,’ it is defined as FP. Similarly, if the direction is predicted to be ‘Down,’ and the actual result corresponds, it is predicted as TN and if the actual result is ‘Up,’ it is defined as FN. Accuracy is defined in the Confusion Matrix as follows (Table 3):

$$\text{Accuracy} \overset{\text{def}}{=} \frac{TP + TN}{TP + FP + FN + TN} \tag{10}$$

4. Results

In this section, comparative analysis is conducted based on the proposed method and existing methods. First, we show heterogeneity of the GICS sectors and the necessity of the use of the data mining technique. Second, the results of the experiment and interpretation are shown.

4.1. Heterogeneity in the sectors

As in Section 3.1.1, we measure how much stock prices in the same GICS sector move in different directions when they experience the same event (news, disclosure) (Fig. 3.).

The result of examining the heterogeneity of the seven sectors in the KOSPI 200 is as shown in Table 4.1. In Table 4.1, the higher the values, the stronger the heterogeneity. The smaller the values, the stronger the homogeneity.

We find that the heterogeneity within the sectors all differs. The Food Expenses sector shows the least amount of variance with 0.52 and is homogenous characteristic. Material sector displays high heterogeneity of 0.83, and Pharmacy sector displays middle heterogeneity of 0.61. Accordingly, we choose the Food Expenses sector, Pharmacy sector, and Material sector to identify the clustering effect depending on heterogeneity.

4.2. Evaluation of research model

The results of the experiment are presented in Table 4.2. As mentioned in Section 3, we are unable to determine which kernel is best suited for processing text data. Accordingly, we adopt a single kernel and combination of two kernels among three kernels. In Table 4.2, ’poly’ denotes a polynomial kernel, ’rbf’ denotes a Gaussian kernel, and ’lin’ denotes a linear kernel. Also, ’lp’ denotes a combination of a linear kernel and a polynomial kernel, ’rp’ denotes a combination of a Gaussian kernel and a polynomial kernel, and ’lr’ denotes a combination of a linear kernel and a Gaussian kernel.

The first line (‘Individual level’) in Table 4.2 describes the results of predicting stock movements in the three GICS sectors with the use of SVM or MKL on the dataset of the target company. Here, ’poly,’ ’rbf,’ and ’lin’ refer to the use of SVM for the different kernel functions. ’lr,’ ’rp,’ and ’lp’ refer to the use of MKL for the combination of different kernel functions on the dataset.

The second line (‘Sector’) shows the results of stock prediction in the three sectors with the use of MKL on the dataset of the target company and its corresponding sector of GICS. As in Shynkevich et al. (Shynkevich et al., 2016), ’poly,’ ’rbf,’ and ’lin’ refer to assigning the same kernel to the company and sector to determine the degree of influence of the target company and the sector. ’lr,’ ’rp,’ and ’lp’ means assigning the combination of two kernels to each dataset. For example, ’lp’ assigns a linear kernel and a polynomial kernel to the company and the sector.

The third, fourth, and fifth line (‘Segmentation’) refer to the results proposed in this study. We assign kernel in the same way as we did for the ‘Sector’ after clustering sectors. ‘Segmentation 2′ means clustering in two clusters. ‘Segmentation 3′ also means segmentation in three, and ‘Segmentation 4′ implies grouping in four.

4.3. Interpretation of experiment results

First, comparison of the ‘Individual level’ and ‘Sector’ is described as follows. Food Expenses has the average value of 0.60086 of the ‘Individual level’ and 0.60749 in the ‘Sector.’ Pharmacy has the mean value of 0.59592 of the ‘Individual level’ and 0.59784 in the ‘Sector.’ Material has the average value of 0.62235 of the ‘Individual level,’ and 0.62235 in the ‘Sector.’

In other words, prediction with the target company and the relevant data provides higher accuracy than a forecast only with the target company. It proves once again that the information of the GICS sector is related to the companies as shown by Shumaker and Chen (Schumaker & Chen, 2009). The combination has a more significant impact when considering the news of the relevant company and the news of the target company together as in Shynkevich et al. (Shynkevich et al., 2016).

A comparison of ‘Sector’ and ‘Segmentation’ shows that ‘Sector’ has the highest accuracy of 0.60749 in Food Expenses. ‘Segmentation 2′ and ‘Segmentation 3′ have the same and highest predictive power in Pharmacy. In Material, ‘Segmentation 4′ shows the best predictability overall. In the majority of cases, the results are better in the ‘Segmentation’ when we consider heterogeneity in the groups rather than in the ‘Sector.’ In other words, the results prove that heterogeneity exists within the GICS sectors. It supports the argument of this paper that simply considering the GICS sector as listing relevant companies is inefficient in predicting stock prices.

More precisely, sector level prediction, ‘Sector,’ is best in Food Expenses and forecast with clustering is best in Pharmacy and Material. As we can see in Table 4.1, Material has high heterogeneity, while Pharmacy and Food Expenses have middle and low heterogeneity respectively. Therefore, we interpret the result as follows. In the GICS sectors with a high degree of heterogeneity such as Material, predicting stock movements of the target company with homogeneous groups made through clustering analysis increases the synchronization among the companies and the predictability. In the homogeneous sectors like Food Expenses and Pharmacy, it is more advantageous not to perform

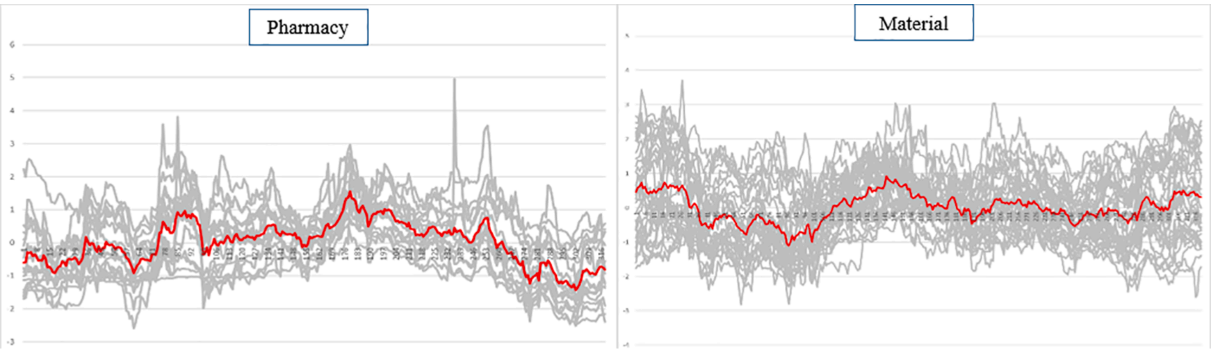

Fig. 3. Heterogeneity in GICS sectors.

Table 4.1
Heterogeneity of GICS sectors.

Pharmacy	Material	Capital goods	Automobile	Durable consumer goods	Food Expenses	Hardware
0.61	0.83	0.85	0.91	0.57	0.52	0.85

Table 4.2
Experimental results.

GICS Sector		Food Expenses		Pharmacy		Material	
Heterogeneity within a group		Low		Middle		High	
Individual level	poly	0.605981	0.60086	0.596335	0.59592	0.626100	0.62235
	rbf	0.600305		0.588095		0.624489	
	lin	0.593618		0.599443		0.615445	
	lr	0.593267		0.599443		0.615445	
	rp	0.605981		0.596335		0.626100	
	lp	0.606027		0.595877		0.626572	
Accuracy Rank		3		5		4	
Sector	poly	0.612301	**0.60749**	0.593457	0.59784	0.624119	0.62357
	rbf	0.602455		0.609666		0.629851	
	lin	0.602475		0.599103		0.619715	
	lr	0.602716		0.599033		0.619715	
	rp	0.612301		0.593457		0.624119	
	lp	0.612742		0.592376		0.623954	
Accuracy Rank		1		3		3	
Segmentation 2	poly	0.604374	0.59961	0.600576	**0.60208**	0.624879	0.61989
	rbf	0.601852		0.608041		0.622332	
	lin	0.590568		0.601509		0.611314	
	lr	0.590393		0.601509		0.611314	
	rp	0.604374		0.600576		0.624879	
	lp	0.602988		0.600229		0.624581	
Accuracy Rank		5		1		5	
Segmentation 3	poly	0.605351	0.60194	0.598782	**0.60208**	0.634248	0.62891
	rbf	0.597349		0.602652		0.627547	
	lin	0.599533		0.606641		0.621777	
	lr	0.599533		0.606641		0.621580	
	rp	0.605351		0.598782		0.634248	
	lp	0.604509		0.598948		0.633936	
Accuracy Rank		2		1		2	
Segmentation 4	poly	0.607826	0.60012	0.599132	0.59711	0.646715	**0.63479**
	rbf	0.601015		0.599204		0.624079	
	lin	0.589560		0.594029		0.622282	
	lr	0.589560		0.594098		0.622282	
	rp	0.607826		0.599132		0.646715	
	lp	0.604911		0.596980		0.646644	
Accuracy Rank		4		4		1	

clustering or clustering with a small cluster number because grouping with a large number of clusters removes the relevant companies while they already have synchronization.

In summary, if the stock prices are homogeneous within the sector, it is essential not to implement clustering analysis to use the relevant effect of the GICS sector. If the stock prices are heterogeneous in the sector, clustering firms into homogeneous groups is an excellent way to increase predictive power.

5. Implications

We have proposed a methodology for efficiently analyzing online news considering the homogeneity among the companies to predict stock movements. Previous studies reflected the uniformity in the specified GICS sector (Shynkevich et al., 2016). However, there is a limitation in that it does not reflect the heterogeneity in the sector. Therefore, this is the first paper to suggest a more sophisticated method by developing a highly predictive algorithm that assesses heterogeneity in previously physically designated sectors through data mining.

In general, we know that stock prices in the same sector have similar trends. However, there are sectors with homogeneity, and sectors with heterogeneity. In other words, the intensity of homogeneity depends on the sector. Therefore, if the analysis is conducted under the same assumption that sectors all have homogeneity, there would be a limit to the quality of results. By applying additional analysis in the sector according to the strength of homogeneity and the intensity of heterogeneity through the data mining technique, we constructed a more predictive model. At a broader level, we provide a conceptual explanation of the application of homogeneity and heterogeneity between firms, and an essential part of explaining useful points in the development of various analytical and forecasting models.

In practical terms, many financial institutions and financial service providers classify companies based on GICS. Based on this taxonomy, various analyzes are conducted (Bhojraj, Lee, & Oler, 2003). However, this study suggests that it is important to build a new taxanomy specific to the task because the taxonomy such as GICS does not fit in certain jobs such as stock price forecasting.

That is, in this study, to develop the stock price prediction model, the model was advanced by applying the homogeneous effect. The proposed model performed well, and based on these results, we confirmed that we could build a better model based on homogeneous effects. This does not end with simply applying homogeneous effects. Instead of applying existing standards and theories to the model, it is possible to find and model various effects through analysis by applying advanced data analysis methods for the big data era. That means applying it to can be of great value.

At a broader level, we provide a conceptual explanation of the application of homogeneity and heterogeneity between firms, and an essential part of explaining useful points in the development of various analytical and forecasting models.

To give practical implications, we implement backtesting with three strategies: the proposed approach, sector level, and individual level. The backtesting is only implemented in the test period of the model. We conducted backtesting as a long-short portfolio. The model receives news as input and signals buy or sell. Following this signal, 10% of the total holdings are traded. In other words, if no news comes out, there is no trading. Since trading occurs only when news comes out, trading does not appear frequently, and on average, there were only 0.138 tradings for a company per day during the test period. Therefore, the transaction cost does not impact the result significantly, and we assumed there is no transaction cost. In addition, it was assumed that the initial starting amount is 100,000. The results are shown in Figs. 4–6 and Table 5. In Figs. 4–6, the normal line is the result of the proposed method. The dashed line is the result of the sector level, and the dotted line is the result of the individual level. In Fig. 4, the material sector results showed that the proposed method yielded about 10% for six months, while the sector level and individual level yielded 8.7% and 8.1%, respectively. The difference in accuracy was about 1%, but the difference in yield was even more. This means that the proposed method is wrong when there is no difference in stock return, but in other cases, it fits more than other methods and suggests that it can be more useful than other methods when trading. In Fig. 5, the pharmacy sector results showed that the proposed method yielded about 8.26% for six months, while the sector level and individual level yielded 4.54% and 0.69%, respectively. In Fig. 6, the proposed approach yielded 5.4%, while the sector level and individual level yielded 5.7% and 4.5%.

The results is confirmed with the sharpe ratio in Table 5. We calculated the sharpe ratio with risk free rate 3%. Usually, the portfolios with a sharpe ratio of more than 3 are evaluated as excellent. In Table 5, the proposed approach showed 3 to 5 sharpe ratio, which is considered a good strategy. Especially, in pharmacy sector and material sector, the proposed approach showed better performance, but in food expenses sector, the proposed approach showed worse performance. Also, this suggest a consistent results with the Result section. In the homogeneous sectors like Pharmacy, the proposed approach may not perform well, unlike in the heterogeneous sectors. In addition, in pharmacy and material sector, this result shows that the difference in sharpe ratio is larger than the difference in accuracy, which proves that the method presented in this paper is a more stable strategy.

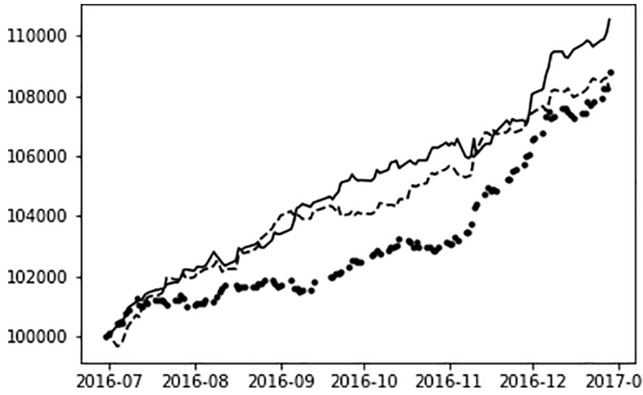

Fig. 4. Backtesting results for material sector (Normal line: the proposed method. Dashed line: sector level. Dotted line:individual level.)

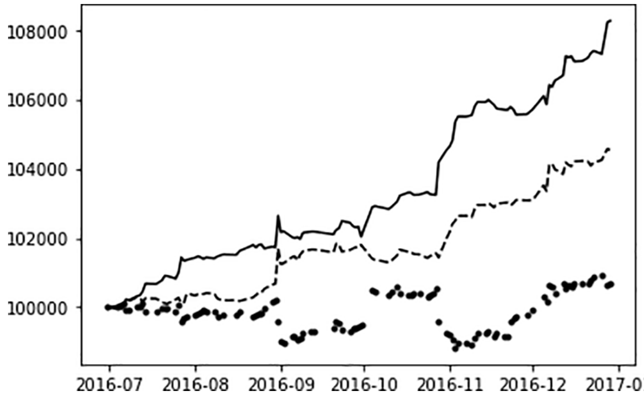

Fig. 5. Backtesting results for pharmacy sector (Normal line: the proposed method. Dashed line: sector level. Dotted line: individual level.)

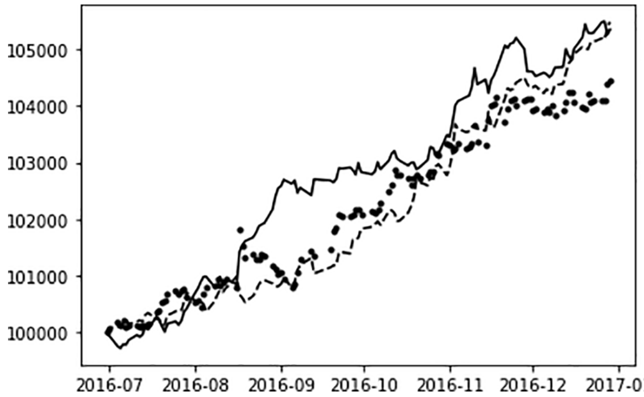

Fig. 6. Backtesting results for food expenses sector (Normal line: the proposed method. Dashed line: sector level. Dotted line: individual level.)

Table 5
Backtesting results: Sharpe Ratio.

Sharpe Ratio	Food expenses	Pharmacy	Material
Individual level	2.365	−0.4011	4.58
Sector	4.016	2.684	4.96
The proposed approach	3.379	4.661	5.10

6. Suggestions for future work

This work identifies the below as areas or aspects in need of future research and advancement:.

A. Method: the advanced methods of clustering are well worth applying. To better understand the homogeneity and heterogeneity of companies, we need to apply more effective clustering methods. Checking homogeneity is an important point in the paper, so upgrading this part and optimizing the proper method is essential. It would be worthwhile to try a variety of ways to build a high-performance stock price forecasting model that uses the latest deep learning clustering methods and how to make the most of the time series.

B. Various effects: we implemented cluster analysis on companies to form homogeneous patterns within the sectors. However, there are some firms with similar patterns based on trends in other sectors. Although we do not incorporate other sectors because of spurious correlation, we can check whether companies from other sectors could be teamed up. Further application of these effects may result in more meaningful research results. In addition to the homogeneous patterns applied in this paper for stock price prediction, various patterns to improve the prediction accuracy should be found and applied to the model. Since the homogeneous patterns in this paper suggest the direction that various factors should be considered for the stock price prediction, additional analyses should be conducted.

C. Language: even though the experiment has been done in certain situations in Korean, it would be more predictable if applied in English, which has many users worldwide. It is because text mining in Korean has lower performance in terms of sophistication and difficulty than in English. In other words, if this study is based on English, it will not only result in higher performance, but will also be able to derive more meaningful results by applying various text mining techniques specialized in English.

7. Conclusion

In the era of big data, in which the amount of information is soaring, and stock related information becomes overwhelming, it has become physically impossible for an individual to read all the news and selectively use information that affects stock prices. Accordingly, it has become a significant challenge to develop algorithms that can automatically process and predict the stock price with a large amount of text information. In particular, how to select information from a large amount of information that affects stock has become a challenge. Generally, news is perceived as providing valuable information if a company name appears or if there is a company ticker in the tag of the news.

In this study, we extend the scope of impacts to a group with homogeneous patterns with each company and study how well integrating prediction of the stock price of the company and influential companies can improve the performance. To find groups with homogeneous patterns, we find companies with similar stock price movements and cluster them to form a homogeneous group. The target sectors are Material, Food Expenses, and Pharmacy in the GICS system. The criterion is to select sectors with different levels of heterogeneity. Heterogeneity is large in Material, middle in Pharmacy, and low in Food Expenses.

As a result of the prediction using the Multiple Kernel Learning, the proposed method shows a higher prediction rate than that of the existing GICS system and that of individual companies. In addition, we find that it is beneficial to make a homogeneous group by clustering in a group with large heterogeneity in the sector.

CRediT authorship contribution statement

Nohyoon Seong: Conceptualization, Methodology, Software, Validation, Investigation, Writing - original draft, Writing - review & editing, Visualization, Supervision. **Kihwan Nam:** Conceptualization, Methodology, Software, Validation, Investigation, Writing - original draft, Writing - review & editing, Visualization, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Aiolli, F., & Donini, M. (2015). EasyMKL: A scalable multiple kernel learning algorithm. *Neurocomputing, 169*, 215–224.

Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms (pp. 1027-1035): Society for Industrial and Applied Mathematics.

Bhojraj, S., Lee, C. M. C., & Oler, D. K. (2003). What's My Line? A comparison of industry classification schemes for capital market research. *Journal of Accounting Research, 41*(5), 745–774.

Boguth, O., Carlson, M., Fisher, A., & Simutin, M. (2016). Horizon effects in average returns: The role of slow information diffusion. *The Review of Financial Studies, 29*(8), 2241–2281.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1–8.

Campbell, J. Y., & Shiller, R. J. (1987). Cointegration and tests of present value models. *Journal of Political Economy, 95*(5), 1062–1088.

Junqué de Fortuny, E., De Smedt, T., Martens, D., & Daelemans, W. (2014). Evaluating and understanding text-based stock price prediction models. *Information Processing & Management, 50*(2), 426–441.

Deng, S., Mitsubuchi, T., Shioda, K., Shimada, T., & Sakurai, A. (2011). Combining technical analysis with sentiment analysis for stock price prediction. In Dependable, Autonomic and Secure Computing (DASC), 2011 IEEE Ninth International Conference on (pp. 800-807): IEEE.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd, 96*, 226–231.

Fama, E. F. (1995). Random walks in stock market prices. *Financial Analysts Journal, 51*(1), 75–80.

Groth, S. S., & Muntermann, J. (2011). An intraday market risk management approach based on textual analysis. *Decision Support Systems, 50*(4), 680–691.

Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems, 55*(3), 685–697.

Herrera, P. J., Pajares, G., & Guijarro, M. (2011). A segmentation method using Otsu and fuzzy k-Means for stereovision matching in hemispherical images from forest environments. *Applied Soft Computing, 11*(8), 4738–4747.

Hsu, C.-F., Lin, C.-M., & Yeh, R.-G. (2013). Supervisory adaptive dynamic RBF-based neural-fuzzy control system design for unknown nonlinear systems. *Applied Soft Computing, 13*(4), 1620–1626.

Hurst, H. E. (1951). Long term storage capacity of reservoirs. *ASCE Transactions, 116*, 770–808.

Huysmans, J., Martens, D., Baesens, B., Vanthienen, J., & Van Gestel, T. (2006). Country corruption analysis with self organizing maps and support vector machines. In Intelligence and security informatics (pp. 103-114): Springer.

Jain, A., Vishwanathan, S. V., & Varma, M. (2012). SPF-GMKL: generalized multiple kernel learning with a million kernels. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 750-758): ACM.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651–666.

Kalwani, M. U., & Morrison, D. G. (1977). A parsimonious description of the hendry system. *Management Science, 23*(5), 467–477.

Karaboga, D., & Ozturk, C. (2011). A novel clustering approach: Artificial Bee Colony (ABC) algorithm. *Applied Soft Computing, 11*(1), 652–657.

Kraus, M., & Feuerriegel, S. (2017). Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems, 104*, 38–48.

Lahmiri, S. (2015). Long memory in international financial markets trends and short movements during 2008 financial crisis based on variational mode decomposition and detrended fluctuation analysis. *Physica A: Statistical Mechanics and its Applications, 437*, 130–138.

Lam, Y. K., & Tsang, P. W. M. (2012). eXploratory K-Means: A new simple and efficient algorithm for gene clustering. *Applied Soft Computing, 12*(3), 1149–1157.

LaPlaca, P. J. (1997). Contributions to marketing theory and practice from Industrial Marketing Management. *Journal of Business Research, 38*(3), 179–198.

Lee, D.-J., Yeon, J.-H., Hwang, I.-B., & Lee, S.-G. (2010). KKMA: A tool for utilizing Sejong corpus based on relational database. *Journal of KIISE: Computing Practices and Letters, 16*, 1046–1050.

Li, Q., Wang, TieJun, Li, P., Liu, L., Gong, Q., & Chen, Y. (2014). The effect of news and public mood on stock movements. *Information Sciences, 278*, 826–840.

MacKinlay, A. C. (1997). Event studies in economics and finance. *Journal of economic literature, 35*, 13–39.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, pp. 281-297): Oakland, CA, USA.

Mahajan, V., & Jain, A. K. (1978). An Approach to normative segmentation. *Journal of Marketing Research, 15*(3), 338–345.

Martínez-López, F. J., & Casillas, J. (2013). Artificial intelligence-based systems applied in industrial marketing: An historical overview, current and future insights. *Industrial Marketing Management, 42*(4), 489–495.

McDonald, M., & Dunbar, I. (2004). *Market segmentation: How to do it, how to profit from it*. Butterworth-Heinemann.

Nam, KiHwan, & Seong, NohYoon (2019). Financial news-based stock movement prediction using causality analysis of influence in the Korean stock market. *Decision Support Systems, 117*, 100–112.

Nanda, S. R., Mahanty, B., & Tiwari, M. K. (2010). Clustering Indian stock market data for portfolio management. *Expert Systems with Applications, 37*(12), 8793–8798.

Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications, 41*(16), 7653–7670.

Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., & Ngo, D. C. L. (2015). Text mining of news-headlines for FOREX market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment. *Expert Systems with Applications, 42*(1), 306–324.

Ng, R. T., & Han, J. (1994). Efficient and Effective Clustering Methods for Spatial Data Mining. In Proceedings of VLDB (pp. 144-155): Citeseer.

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications, 42*(24), 9603–9611.

Pelleg, D., & Moore, A. W. (2000). X-means: Extending k-means with efficient estimation of the number of clusters. *Icml, 1*, 727–734.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65.

Schumaker, R. P., & Chen, H. (2009). A quantitative stock prediction system based on financial news. *Information Processing & Management, 45*(5), 571–583.

Shang, F., Jiao, L. C., Shi, J., Gong, M., & Shang, R. H. (2011). Fast density-weighted low-rank approximation spectral clustering. *Data Mining and Knowledge Discovery, 23*(2), 345–378.

Shi, Z., Lee, G. M., & Whinston, A. B. (2016). Toward a better measure of business proximity: Topic modeling for industry intelligence. *MISQ, 40*(4), 1035–1056.

Shynkevich, Y., McGinnity, T. M., Coleman, S. A., & Belatreche, A. (2016). Forecasting movements of health-care stock prices based on different categories of news articles using multiple kernel learning. *Decision Support Systems, 85*, 74–83.

Singh, K. P., Gupta, S., Ojha, P., & Rai, P. (2013). Predicting adsorptive removal of chlorophenol from aqueous solution using artificial intelligence based modeling approaches. *Environmental Science and Pollution Research, 20*(4), 2271–2287.

Smith, W. R. (1956). Product Differentiation and Market Segmentation as Alternative Marketing Strategies. *Journal of Marketing, 21*(1), 3–8.

Sun, Z., Ampornpunt, N., Varma, M., & Vishwanathan, S. (2010). Multiple kernel learning and the SMO algorithm. In Advances in neural information processing systems (pp. 2361-2369).

Vermunt, J. K., & Magidson, J. (2002). In *Applied Latent Class Analysis* (pp. 89–106). Cambridge University Press. https://doi.org/10.1017/CBO9780511499531.004.

WEINSTEIN, A. (2004). Handbook of Market Segmentation: strategic targeting for business and technology firms. New York: Haworth. In: ISBN 0-7890-2156-0.

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. In ACM Sigmod Record (Vol. 25, pp. 103-114): ACM.

Zhou, T., Lu, H., Yang, D., Ma, J., & Tuo, S. (2011). Rough kernel clustering algorithm with adaptive parameters. In International Conference on Artificial Intelligence and Computational Intelligence (pp. 604-610): Springer.

Zunino, Luciano, Olivares, Felipe, Bariviera, Aurelio F., & Rosso, Osvaldo A. (2017). A simple and fast representation space for classifying complex time series. *Physics Letters A, 381*(11), 1021–1028.