Decision Support Systems 117 (2019) 100–112

Contents lists available at ScienceDirect

Decision Support Systems

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

Financial news-based stock movement prediction using causality analysis of influence in the Korean stock market

KiHwan Nam, NohYoon Seong*

College of Business Management Engineering Department, KAIST, Seoul, South Korea

ARTICLE INFO

Keywords:
Stock movement prediction
Transfer entropy
Causal relationship
Multiple kernel learning
Text mining

ABSTRACT

With the advent of the Big Data era and the development of machine learning technologies, predicting stock movements by analyzing news articles, which are unstructured data, has been studied actively. However, so far no attempts have been made to utilize the asymmetric relationship of firms. Thus far, most papers focus on only the target firm, and few papers focus on the target firm and relevant firms together. In this article, we propose a novel machine learning model to forecast stock price movement based on the financial news considering causality. Specifically, our method analyzes the causal relationship between companies, and it accounts for the directional impact within the Global Industry Classification Standard sectors. In our proposed method, transfer entropy is used to find causality, and multiple kernel learning is used to combine features of target firm and causal firms. Based on a Korean market dataset and out-of-sample test, our experimental results reveal that the proposed causal analytic-based framework outperforms two traditional state-of-the-art algorithms. Furthermore, the experimental results show that the proposed method can predict the stock price directional movements even when there is no financial news on the target firm, but financial news is published on causal firms. Our findings reveal that identifying causal relationship is important in prediction problems, and we suggest that it is important to develop machine learning algorithms and it is also important to find connections with well-established theories such as the complex system theory.

1. Introduction

The price of a stock increases and decreases in response to the dealings of the sellers and buyers in the market to reach a reasonable price. Thus, stock prices are determined by the law of supply and demand [1]. The stock price demand will increase if investors believe that the company performs well; in contrast, if investors think that the company does not perform well, the supply will increase. In other words, theoretically, the stock price is a measure of the company's future performance. However, many factors can affect the expected value of a company's future performance. In the academic literature, there are various theories that are related to stock prices. The most representative theory is the Efficient Market Hypothesis (EMH) [2]. The EMH states that the price of the market reflects the value accurately and responds only to new information which consists of historical prices, public information and private information. EMH is divided into three categories according to how much information is reflected:

I. Weak-form efficient market. In a weak-form efficient market, the historical prices of all financial assets, such as stocks, bonds, and tangible assets that can be traded in the current market are already reflected in the current stock price through all available historical information. Analysis using historical prices cannot yield excess returns so that one needs to predict the stock price with public information and private information.
II. Semi-strong-form efficient market. In a semi-strong-form efficient market, all past public information, such as the past stock price, disclosures, and news, is already reflected in the financial assets, and forecasting using historical prices and public information cannot generate excess returns in the market. So one needs to predict the stock price with private information.
III. Strong-form efficient market. In a strong-form efficient market, non-public information, as well as past prices and public information, is reflected in market prices. Thus, forecasting using all information cannot yield excess returns in the market.

Since a majority of the mature stock markets, like those in the U.S.A, Korea, U.K., and France, are weak-form efficient markets [3], many

* Corresponding author at: College of Business at Korea Advanced Institute of Science and Technology (KAIST), SUPEX Building, 85 Hoegiro Dongdaemoon-gu, Seoul 130-722, South Korea.
E-mail addresses: namkh@kaist.ac.kr (K. Nam), nyseong@kaist.ac.kr (N. Seong).

https://doi.org/10.1016/j.dss.2018.11.004
Received 29 April 2018; Received in revised form 4 November 2018; Accepted 25 November 2018
Available online 30 November 2018

studies have been conducted on predicting stock prices using financial news. To quantitatively measure market states, various measurement methods have been applied. One method, the Hurst Exponent, is a concept used in econophysics to measure market states [3]. The Hurst Exponent measures long-term memory quantitatively [4]. If the Hurst Exponent is < 0.5, stock prices can be interpreted as mean-reverting. If it equals 0.5, it means that stock prices follow a random walk, and if it is > 0.5, stock prices can be interpreted as following a trend. In other words, the larger the Hurst Exponent, the greater the effect of past prices on current prices. If the Hurst Exponent is high, the stock market is not a weak-form efficient market because it can be predicted with historical prices. According to Eom et al. [3], the Korean stock market has a Hurst Exponent of approximately 0.5 so that the Korean stock market is the weak-form efficient market. Therefore, when predicting stock prices in the Korean market, it is meaningful to use public information for analysis, such as financial news, rather than historical prices.

In research on stock price forecasting through financial news, it is common to build a keyword dictionary for each company. This dictionary provides keywords that influence the fluctuations of stocks of individual companies, and they should be used to predict future stock prices. Recently, studies on identifying relevant firms [5], and studies on reflecting the effects of relevant firms based on the Global Industry Classification Standard (GICS) sector [6,7] emerge. Especially, Shynkevich et al. [7] constructed an individual firm dictionary, sub-industry dictionary, industry dictionary, group industry dictionary, and sector dictionary in the S & P 500 healthcare sector and predicted stock movements with the combination of them. The results of this study are as follows: The prediction accuracy of integrating them is higher than that of news of individual firms only. In other words, the dictionaries that include higher-level concepts, e.g. sector dictionaries, reflect information that affects industry characteristics or industries that are not covered by the concepts of the subordinate individual firms.

Although research has progressed gradually on the basis of the influence within the GICS sector, it has been conducted based on the assumption that every firm influences other firms, and the influence between firms is bidirectional. However, companies in the same GICS sector may not influence each other, and there is a structure in which a company affects other companies but not inversely [8]. In this study, we overcome the limitations of the existing research by applying the transfer entropy technique, which has been actively studied in the complex system theory. We find the causal relationships of the firms within the GICS sectors and predict the stock price based on causal relationships. Especially, we integrate the effect of the target firm and the effects of the causal firms by employing Multiple Kernel Learning method [9].

The results show that our approach improves the prediction performance in comparison with approaches that are based on news on target firms [10] and on the GICS Sector-based integration approach [7], which are two state-of-the-art algorithms. Furthermore, the experimental results show that the proposed method can predict the stock price directional movements even when there is no financial news on the target firm, but financial news is published on causal firms. In addition, we find that the results change by setting the statistical significance of transfer entropy. Therefore, it is important to set the threshold of statistical significance through a grid search.

In this study, we make three main contributions: First, in solving socioeconomic problems, we were able to achieve higher performance by successfully combining physics theory with machine learning. To the best of our knowledge, this paper is the first paper to combine complex system methodology with machine learning. Second, previous studies have predicted stock prices at the individual level and searched for relevant companies to consider their impact. This study is the first to predict stock prices while considering the causality between the companies. Finally, existing studies were able to predict stock movements only when the news on the company was released. In this paper, we propose a method for predicting stock movements with a causal relationship through causality detection, even when no news is published directly.

We organize the remainder of this paper as follows: Section 2 provides an overview of the relevant literature on complex networks and text mining. Section 3 describes news and stock datasets, transfer entropy analysis, text pre-processing techniques, machine learning approaches and evaluation metrics. Section 4 describes the experimental results. Section 5 presents the study's conclusions and outlines directions for future work.

2. Literature review and hypothesis development

As text mining techniques are gradually evolving, research on predicting the stock prices of companies using the textual data of company-related financial news articles, company disclosures, and social network service (SNSs) is increasing. Research has been in full swing since the 1990s [11] and has been more active since 2000, especially with the advancement of machine learning. This chapter summarizes the existing studies in the flow of research on predicting stock prices with financial news articles.

2.1. Key related research

Text analysis procedures are broadly divided into (1) text preparation, (2) text mining, and (3) model learning and prediction [10].

(1) The text preparation stage refers to the collection of textual data that are related to the finances of a company through various methods, such as online news crawling. (2) The text mining step is the generation of stock impact features through text mining techniques. The text mining step consists of feature extraction, feature selection, and feature representation. Finally, (3) in the model learning and prediction stage, we predict stock price movements with machine learning on the generated stock impact features. The model learning and prediction stage consists of machine learning algorithms, forecast type, and combining them with other effects. Each step is summarized in Table 1.

Hagenau et al. [10] designed a procedure that receives corporate announcements and financial news automatically, that implements feature engineering, and that uses machine learning techniques to forecast stock prices. The authors implemented many feature extraction methods and compared Bag-of-words, Noun phrases, and 2-Gram and 2-word combinations. In addition, the authors implemented two feature selection methods and compared chi-square feature selection and bi-normal-separation feature selection. High accuracy rates were obtained as a result. In this paper, we use the Chi-square feature selection and TF-IDF weighting for the Bag-of-words model, as used in Hagenau et al. [10].

Schumaker and Chen [6] suggested the Arizona Financial Text System (AZFinText). AZFinText constructs a textual dataset (i.e., financial news, trading experts and stock quotes) and predicts stock prices. In addition, the authors divided the financial news that affects the stock prices into several groups of datasets: sector-based, sub-industry-based, industry-based, group-based, and stock-specific news, based on the GICS (Global Industry Classification Standard), which is an industry taxonomy that was developed by MSCI and S&P. The authors predicted the stock prices with each dataset of news articles. Stock-specific news affected the prediction of the stock price, just like the previous studies. Surprisingly, the sector-based news was also effective in predicting stock prices. However, there is a limitation that various levels of datasets were not applied simultaneously. Therefore, in this paper, multiple kernel learning is used to integrate various levels of features simultaneously.

By assimilating news groups of various levels of relevance, Shynkevich et al. [7] built a system for estimating stock movements.

Table 1
Key related research.

–	(1) Preparation	(2) Text mining			(3) Model learning and forecasting		
Reference	Data type	Feature extraction	Feature selection	Feature representation	Machine learning algorithm	Forecast type	Other effects
Hagenau et al. [10]	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	–
Schmaker and Chen [6]	Financial news	Bag-of-words Noun phrases Named entities Proper Nouns	Minimum occurrence per document	Binary	SVR	Price value	Relevant firms' financial news based on GICS
Shynkevich et al. [7]	Financial News	Bag-of-words	Chi-square	TF-IDF	Multiple kernel learning	Up and down	Relevant firms' financial news based on GICS
Li et al. [12]	Web news and financial discussion board	Noun phrases Sentiment terms	–	Modified TF-IDF	SVM, SVR	Up and down, price value	–
Groth and Muntermann [13]	German ad-hoc announcements	Bag-of-words	Chi-square Information gain	TF-IDF	SVM	Market risk	–
The proposed Approach	Financial News	Bag-of-words	Chi-square	TF-IDF	Multiple kernel learning	Up and down	Causal firms' financial news based on GICS

The groups that the authors created are as follows: sector-based, sub-industry-based, industry-based, group-based, and stock-specific news as in [6]. The authors considered and compared them by utilizing multiple kernel learning with all of the companies. The authors found that forecasting the stock price while taking into account the significance of various levels yields better results than anticipating the stock price alone. However, regardless of whether they are in the same industry, the relevance will not be high [5]. Even if the relevance between firms is high, they do not share the same effect but have an asymmetric influence. Therefore, we utilize an information theory approach, as opposed to utilizing only the GICS system, to determine the causality between firms.

Fig. 1 shows a general concept of stock price forecasting through online news. In terms of the media effect analysis, online news includes the overall situation of a company, such as the financial situation and economic activities [1]. News on the same information differs substantially according to investor psychology. According to research in behavioral finance and investment psychology, investor behavior can be determined by whether investors feel optimistic or pessimistic about the future market value [12,14]. In addition, investor sentiment can impact the individual firms and the industrial sector. Combining and analyzing these sectors of industry improves the performance of stock forecasting [6,7]. In other words, stock price forecasting is analyzed based on individual corporate media effect, and the industry-level media effect. Previous studies on media effects have analyzed the influence based on the assumption that every firm in the same industry affects each other, and that magnitude of impacts are all the same. However, these studies have several limitations. Companies have structures that exchange asymmetric influences, but these factors were not considered in the existing methods. Therefore, in this study, we propose a methodology for predicting stock prices based on a more explicative and effective analysis by analyzing the causal relationships among these influences.

2.2. Transfer entropy

In this paper, when predicting stock prices, the aim is to grasp only the factors that affect companies in the same group based on information theory and to predict the stock price by only considering the causal effect. We must check the causality between companies to determine who affects whom. In this paper, we use transfer entropy, which quantitatively measures the asymmetric information flow and is mainly used to measure the causality in complex systems [8,15–19].

2.2.1. Transfer entropy (TE)

TE is a concept that was developed by Schreiber [20] and designed to measure the flow of information asymmetrically between two systems in a complex system. When there are two processes I and J, Entropy is defined as (1) [21].

$$\mathrm{H(I)} = -\sum p_I(i)\log p_I(i) \tag{1}$$

Entropy is a measure of the average uncertainty. In other words, entropy is the average amount of information that is needed to predict a process. In the same vein, joint entropy and conditional entropy are defined as follows:

$$\mathrm{H(I,J)} = -\sum\sum p_{IJ}(i,j)\log p_{IJ}(i,j) \tag{2}$$

$$\mathrm{H(I \mid J)} = -\sum\sum p_{IJ}(i,j)\log p_{I|J}(i \mid j) \tag{3}$$

The mutual information that is shared by two processes is defined as follows [20,22]:

$$\mathrm{M(I,J)} = H(I) + H(J) - H(I,J) \tag{4}$$

$$\mathrm{M(I,J)} = \sum p_{IJ}(i,j)\log\frac{p(i,j)}{p_I(i)p_J(j)} \tag{5}$$

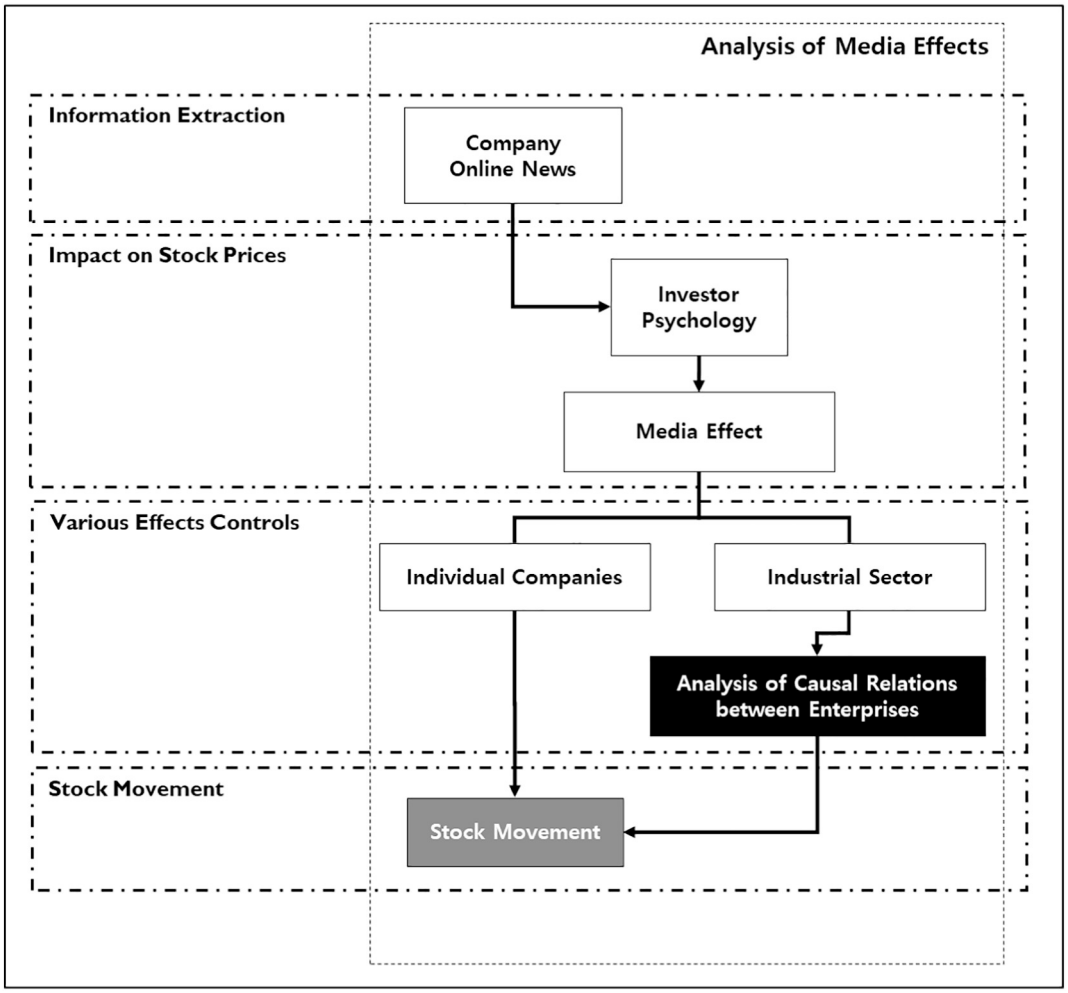

Fig. 1. Concept of stock price prediction through news.

Since transfer entropy involves directional mutual information, TE for k sample processes of I and l sample processes of J is defined as follows [20]:

$$TE_{J\to I} \stackrel{def}{\equiv} \sum p(i_{t+1}, i_t^k, j_t^l)\, log \frac{p(i_{t+1} \mid i_t^k, j_t^l)}{p(i_{t+1} \mid i_t^k)} \quad (6)$$

where i_t and j_t denote data points at time t of I and J processes, and i_t^k and j_t^l are k-dimensional and l-dimensional delay vectors that are one-dimensional delay vectors of sequences I and J, respectively. In addition, $p(i_{t+1}, i_t^k, j_t^l)$ is the joint probability of i_{t+1}, i_t^k and j_t^l. TE is mainly estimated through the KSG Estimator [23,24].

$TE_{J\to I}$ process J measures the transition probability of process I and vice versa because of the differences between the joint probability and the conditional probability. In other words, TE provides information about the direction of interaction between two systems [19], which can be expressed as shown in Fig. 2.

2.2.2. Statistical significance: p-value

Theoretically, TE between two processes without information flow should be zero. However, there are nonzero cases because the empirically measured TE at a finite number of data points has a bias [21]. Therefore, it is necessary to know how statistically significant the TE is.

Permutation testing is mainly used to measure the confidence intervals of transfer entropy and statistical significance [17,21,24]. In this paper, we measure the p-value as statistical significance. To perform the test, a null hypothesis must be established. The null hypothesis is as follows:

$$H_0: \textit{There is no directed relationship from } J \textit{ to } I;\ TE_{J\to I} = 0 \quad (7)$$

In addition, we need to know the distribution of the TE measurements. To obtain the distribution of TE, we can identify a surrogate process p^s that satisfies $p^s(i_t|i_{t-1}^k) = p(i_t|i_{t-1}^k, j_{t-1}^l)$ by sub-sampling. Superscript s denotes a surrogate process. In other words, a surrogate process of J, J^s, has the same statistical properties as J but should not have a direct relationship with I [17,21]. Transfer Entropy is asymptotically distributed as follows [21,25], where dJ and dI are the dimensionalities of processes J and I:

$$TE_{J^s\to I} \sim \frac{\aleph^2}{2N}(\textit{in nats}),\ \textit{degree of freedom} = ldJdI \quad (8)$$

Following (8), we can determine the statistical significance of the transfer entropy that is obtained by null hypothesis testing and we can obtain the p-value.

2.2.3. Causality detection of transfer entropy

Generally, causality refers to a 'cause-effect relationship.' When something causes a problem and something occurs as a result, it is popular to say that the cause has causality in the result. However, it is very difficult to precisely define the concept of causality [26]. The concept of causality was first quantitatively defined by Wiener [27]. According to the definition of Wiener [27], when there are two signals

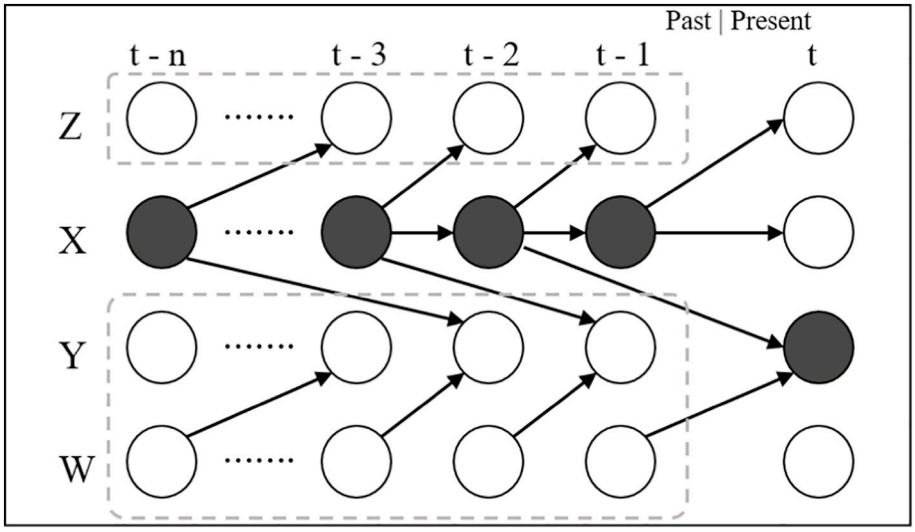

Fig. 2. Conceptual representation of Transfer Entropy [19].

X and Y that are measured at the same time, when estimating Y using X, we say that X causes Y if it can predict better than Y alone. In this paper, causality is based on the Granger definition [28].

The causality detection with Wiener-Granger Causality is used in stock market. Výrost et al. [29] examined return spillovers among stock indices by constructing Granger causality financial network. However, since Wiener-Granger Causality is not appropriate for finding causal relationships in nonlinear systems [26,30], it has limitations in cases of collective behavior, such as stock herding behavior [31], and complex systems with a critical point [32]. To overcome these limitations, a nonlinear extension that is based on the information-theoretic formulation is often used for causality detection [21]. The representative extension is transfer entropy [8,15–18].

Transfer entropy quantitatively measures the asymmetric information flow within a complex system. In particular, causality detection using transfer entropy is also widely used in financial markets. In Marschinski and Kantz [16], the relationship between the US Dow Jones Industrial Average (DJIA) and the German DAX Xetra Stock Index (DAX) was measured using transfer entropy. They calculated the effective transfer entropy by shuffling the time series to remove the random effects and found that the effect of DAX on DJIA was approximately three times greater than that of DJIA on DAX. However, they only compared the effects of the stock indices and did not confirmed that the values are statistically significant. Therefore, we analyze the transfer entropy considering the statistical significance.

Kwon and Yang [18] conducted transfer entropy analysis using the daily data of 25 market indices in the global financial market. Through the analysis, they found that the US stock market has the greatest impact on the global stock market and confirmed that the market that receives the most information is the Asia/Pacific market. Sensoy et al. [34] investigated the strength and direction of nonlinear causality between exchange rates and stock prices with effective transfer entropy and compared the changes between analysis before the 2008 crisis and analysis after the 2008 crisis. However, Kwon and Yang [18] and Sensoy et al. [34] had the limitation that they only showed relationships in aggregate level, where relationships can vary by GICS sectors or companies.

Kwon and Oh [8] measured information flow using transfer entropy between a market index and individual stocks in several markets around the world. The authors show that the information from the index to the individual stock is higher than the information from the individual stock to the index. This finding implies that the market index affects the future prices of individual stocks. However, the market index is simply a sum of various stocks and not all of them affect a particular stock. Therefore, in this paper, we calculate transfer entropy at the individual stock level rather than at the market index level, determine which stock price affects each stock price, and use transfer entropy to measure the causality at the company level in a complex system, namely, the financial market.

Oh et al. [33] measured an information flow among industry sectors and compared them for three periods: before, during, and after the subprime crisis. They measured the degree of asymmetric information inflow, which is the difference between transfer entropy from X- > Y and transfer entropy from Y- > X. However, authors assumed that causality exists in only one way, where 'X' causes 'Y' and 'Y' causes 'X' simultaneously. This is inappropriate for prediction so that we consider simultaneous causality.

The existing papers constructed financial networks using causality detection. However, all of these studies were limited to only assessing the nonlinear causal relationship and building a financial network, and they did not predict the actual movement of stocks. To fill the research gap, we suggest a novel method that predicts stock movements with a machine learning technique based on causality detection. We calculate nonlinear causality, which is a similar approach in [8]. And we select the financial news that has influence based on nonlinear causality and predict stock movement with machine learning. As far as we know, this is the first paper that implements a stock prediction system based on nonlinear causal relationships (Table 2).

2.3. Text pre-processing

Text pre-processing is the process of finding features that can be used for machine learning from unstructured textual data. There are many studies on using textual data to determine its effect on stock prices. Bollen and Mao [14] showed that the public mood on Twitter can be used to predict the stock price using sentiment analysis. However, Li et al. [12] found that a new set of keywords that consists of words that specifically affect the stock price is needed to ensure that the effect is not simply a sentiment analysis. In addition, it is necessary to consider firm-specific words for each company since the keywords that affect each company are different. Therefore, research on corporate-specific news rather than general news has been actively pursued, and this has been gradually increasing with the increase of computing capacity [35]. In this paper, context-aware text mining based on the company-specific financial news is applied.

The text-mining stage can be divided into three steps: feature extraction, feature selection, and feature representation [13].

Feature extraction starts from unstructured textual data and builds

Table 2
Summary of key papers on causality detection in stock market.

Reference	Methodology	Summary	Limitation	Dataset	
				Causality detection	Prediction
Výrost et al. [29]	Granger causality	Examines return spillovers among stock indices.	Inappropriate for nonlinear complex system. Only accounts for the aggregate level impact.	Stock index price	No
Marschinski and Kantz [16]	Transfer entropy	Measures the information flow between the Dow Jones and DAX stock index.	Does not test statistical significance.	Stock index price	No
Kwon and Yang [18]	Transfer entropy	Observes the strength and direction of information flow between stock indices.	Only for stock indices. Does not find the individual level causality.	Stock index price	No
Sensoy et al. [34]	Transfer entropy	Examines the strength and direction of nonlinear causality between exchange rates and stock markets.	Only accounts for the aggregate level impact.	Stock index price and exchange rates	No
Kwon and Oh [8]	Transfer entropy	Observes asymmetric information flow between the stock market index and their component stocks.	Does not find the causal relationship between component stocks.	Stock index price and individual stock prices	No
Oh et al. [33]	Transfer entropy	Measures asymmetric information flow between the GICS sectors.	Does not test statistical significance.	Sector level stock prices	No
The proposed approach	Transfer entropy	Finds causal relationships between the companies in the same GICS sector, and predicts stock movements with machine learning.	Does not account for impact across GICS sectors.	Individual stock prices	Financial news

features that are expected to be informative, thereby simplifying the subsequent steps. In previous literature, 2-g [10], noun phrases [6], sentiment words [12], topic modeling [36] and Bag-of-words [7,13] have been used. The most popular and basic approach is Bag-of-words in the field of stock prediction with financial news [1]. Therefore, we used Bag-of-words for feature extraction in this paper.

The subsequent step is feature selection. Feature selection is the selection of a subset of relevant features for simplification and avoiding the curse of dimensionality. Feature selection based on pre-defined dictionaries has been used [37]. However, there is a disadvantage that it is difficult to generalize because the set of words changes with time. Therefore, methods of feature selection in that are based on statistical methods are that affect the stock movement mainly used. These methods are Minimum occurrence per document [6], Information Gain [13] and Chi-square [7,10]. Among these methods, Chi-square analysis is chosen for this paper, as it performed well in other studies [10] and Chi-square methods are structurally different in that it reflects the external market feedback [10].

The final step is feature representation. Feature representation is a step that represents every feature by a numeric value so that it can be input into machine learning algorithms. The most basic is a binary representation that indicates the absence or presence of a feature. However, this method has limitations in that it does not reflect the frequency of the word even though it is important. In this paper, Term Frequency-Inverse Document Frequency (TF-IDF) is used to overcome this limitation, as in Hagenau et al. [10] and Shynkevich et al. [7].

2.4. Machine learning prediction techniques

In the model learning and prediction phase, various machine learning algorithms are used. For example, Support Vector Machine (SVM) [6,10,35], Naïve Bayes [38], Nearest Neighbor [11] and Neural Network [39] have been widely used. Among them, SVM showed especially outstanding performance [13].

Since SVM is based on a single kernel, there is a disadvantage in that it can only use a single data source. To use an ensemble technique that combines and predicts data from various sources, Multiple Kernel Learning (MKL) has been widely used recently. MKL can be used to learn various functions by combining several kernels. Yeh et al. [38] developed MKL for solving the Support Vector Regression (SVR) problem, where the hyperparameter had to be manually adjusted to be able to combine the advantages of various hyperparameter settings. Luss and d'Aspremont [40] used the MKL approach to simultaneously learn separate kernels that are assigned to a text dictionary and a time series of absolute returns. The results were compared with the results of MKL using only textual data and MKL using only stock return data. Combining the two data sources yielded higher accuracy and a higher Sharpe ratio than any single data source. Therefore, the main discovery of this paper is to combine information such as news articles and stock returns to predict abnormal returns, yielding better results and improving performance compared to predictions based on a single data source.

This study evaluates the accuracy of the best-performing method according to many research studies at each stage - text preparation, text mining, model learning, and prediction. The objective of this study is to confirm the structure of the stock market and apply the model to improve the accuracy of the stock forecast for each company.

3. Proposed approach

This chapter describes the construction of a system that can forecast stock movements by examining the causal relationships among influential companies through online news. First, the data is described. Second, transfer entropy analysis, which is utilized to examine the causal relationship between firms, is described. Thereafter, text preprocessing, machine learning algorithms, evaluation, and evaluation

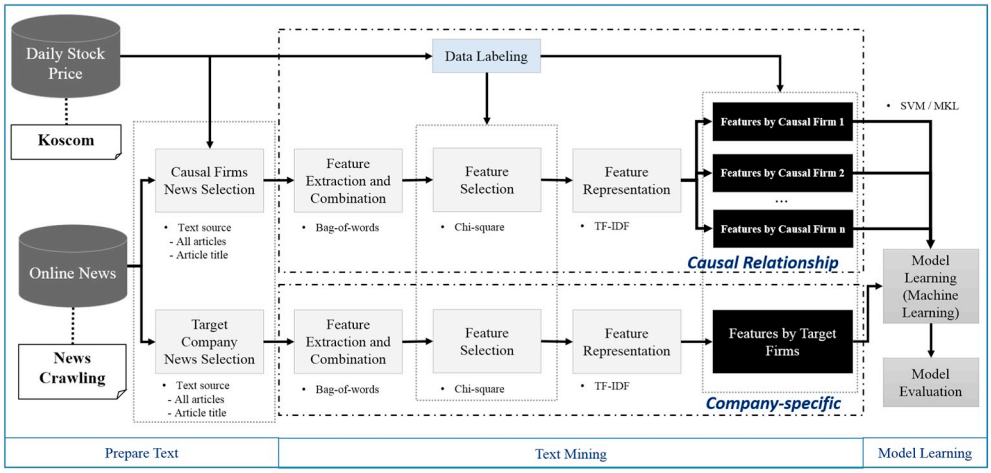

Fig. 3. Proposed approach.

metrics are explained. This terminology was employed in subsequent works [7]. The overall process is illustrated in Fig. 3.

3.1. Data

3.1.1. Stock history data

Information on stock prices is used in causal relationship inference, feature selection and data labeling. The data was obtained from KOSCOM,[1] which is a company that manufactures IT infrastructure for the financial industry. The data lists daily stock prices, including the open price and close price for every company from January 2014 to December 2016. The most expressive features are selected based on the market reaction to the publication of news articles. The reaction is derived from a movement of a stock price defined as the ratio the open price to close price on the day of publication. This is defined as return and it is shown below as (9). Data instances are classified into two classes in this paper. Labels 'Up' or 'Down' that correspond to return of the target stock are given to each data point.

$$\text{Return}_t \stackrel{\text{def}}{=} \frac{Close_t}{Open_t} \tag{9}$$

In the literature, it was shown that stock movements can be predicted with daily data, since the market state changes slowly with new public information [7].

3.1.2. News articles data

We experimented with real data to implement the proposed method in this paper. The data consist of financial news and stock price data from January 1, 2014 to December 31, 2016. We crawled all the financial news articles and economic news articles that are registered in Naver, which is Korea's largest portal site. This information includes most of the financial news articles that are available to the public in Korea. During this period, a total of 1,397,800 articles were crawled, excluding duplicates. Also, there are cases where a news article might quote sources like Facebook or Twitter, and it may hurt the effect of financial news. We checked if news articles are from Twitter or Facebook by searching words including 'Twitter,' 'Facebook,' and 'Social Networking Service.' As a result, we confirmed that our news dataset does not have an information from Twitter or Facebook.

To find news that is relevant to a company, we identified news that includes the name of the company in the news body. Among all news articles, we used 80,741 news articles which were assigned to previously selected companies. We will describe how to select companies later in Section 4. Table 3 gives details about the number of articles retrieved per news provider.

The format of the news data includes title, publisher, post time, content, and category (i.e., economy, finance and politics). The category is predefined in Naver. According to the influence of the news on the stock price, news article is categorized and each news item is labeled according to the return of the target company. For example, a news article on a company that was published at 11:00 on Monday is labeled 'Up' if the return is greater than or equal to 1; if it is < 1, the return is labeled 'Down.' However, since the Korean stock market opens at 9:00 and closes at 16:00, and Korea has one time zone (UTC +09:00), news that is published after 16:00 is interpreted as affecting the next day [12]. The list of all companies that were used in the experiment and the up and down labels are shown in Table 4.

3.2. Transfer entropy: Finding causal relationships

Since Transfer Entropy is difficult to apply precisely, it is important to use an appropriate method [21]. In this paper, we followed the steps that were suggested by Vicente and Wibral [23] and Wibral et al. [24].

I. Measure TE in all pairs of variables in the system.
II. For each source-target pair, find the null distribution for TE and obtain the p-value.
III. Determine the threshold of the p-value and select only causal relationships that have a lower p-value.

In this paper, we propose that predicting stock prices with causal relationship firms yields better results than forecasting with each company or including other unnecessary relationships. To determine the causal relation, the transfer entropy of the stock price is calculated. Since price is a market value that reflects existing past records and information, asymmetric information flow between stock prices can be interpreted as the company's information records having asymmetric information flow. Therefore, we find the causal relationship using stock prices.

However, two time series often leads to spurious causality [41]. To prevent spurious causality, we selected the subjects to be analyzed initially based on the GICS sector, which is an industry classification standard that is based on economic theory [42]. We performed transfer entropy analysis on all pairs of companies within the sector.

The variables are stock prices of companies in the GICS sectors of KOSPI 200 in this study. We perform a log return transformation on the stock price to scale the stock price so that it is not affected by the absolute value of the price [16,18,43]. To make a proper prediction, we use training datasets only.

[1] www.koscom.co.kr.

Table 3
News articles sources.

News article sources	# of news articles	% of news articles (%)	Type
Dong-A Ilbo	3361	4.162693	Comprehensive newspaper
Kukmin Ilbo	2568	3.18054	Comprehensive newspaper
Chosun Ilbo	2387	2.956367	Comprehensive newspaper
Seoul Shinmun	1882	2.33091	Comprehensive newspaper
Segye Times	1695	2.099305	Comprehensive newspaper
Munhwa Ilbo	1529	1.89371	Comprehensive newspaper
Kyunghyang Shinmun	1416	1.753756	Comprehensive newspaper
Hankook Ilbo	1125	1.393344	Comprehensive newspaper
Hankyoreh	785	0.972245	Comprehensive newspaper
JoongAng Ilbo	761	0.94252	Comprehensive newspaper
Digital Times	2568	3.18054	Internet newspaper
Dailian	2478	3.069073	Internet newspaper
Money S	1615	2.000223	Internet newspaper
Yonhap News Agency	14,831	18.36861	Broadcast newspaper
Newsis	13,284	16.45261	Broadcast newspaper
YTN	1096	1.357427	Broadcast newspaper
KBS News	1001	1.239767	Broadcast newspaper
MBC News	717	0.888025	Broadcast newspaper
Korea Economic Daily	5739	7.107913	Economic Newspaper
Financial News	3769	4.668013	Economic Newspaper
Jose Ilbo	3219	3.986822	Economic Newspaper
Maeil Business Newspaper	2705	3.350219	Economic Newspaper
Herald Business	1122	1.389629	Economic Newspaper
Edaily	902	1.117152	Economic Newspaper
Money Today	854	1.057703	Economic Newspaper
Etc.	7332	9.080888	Etc.
Total	80,741	100%	

Table 4
Up and Down Labels of the companies.

Company	# of data points	Sector	% of up labels	% of down labels	Company	# of data points	Sector	% of up labels	% of down labels
OCI	1947	Material	55.47	44.53	Daewoong Pharm.	873	pharmacy	49.94	50.06
Huchems Fine Chemical	175	Material	53.71	46.29	Green Cross	1583	Pharmacy	52.43	47.57
Kukdo Chemical	85	Material	52.94	47.06	Yuhan	854	Pharmacy	51.99	48.01
Hyundai-Steel	3741	Material	49.67	50.33	Jeil Pharmaceutical	155	Pharmacy	57.42	42.58
NamHae Chemical	179	Material	58.10	41.90	Bukwang Pharm.	218	Pharmacy	50.46	49.54
Hansol Chemical	110	Material	64.55	35.45	Hanmi Pharm.	3051	Pharmacy	39.72	60.28
Foosung	295	Material	48.14	51.86	Dong-A ST	497	Pharmacy	52.92	47.08
SKC	1184	Material	55.57	44.43	Boryung Pharm.	583	Pharmacy	53.52	46.48
SKChemical	1276	Material	51.18	48.82	Hanall BioPharma	193	Pharmacy	51.30	48.70
SeAh Steel	374	Material	55.35	44.65	JW Pharm.	489	Pharmacy	47.24	52.76
KISWIRE	166	Material	55.42	44.58	C.K.D	1200	Pharmacy	52.08	47.92
KiscoHolding	762	Material	50.79	49.21	Yungjin Pharm.	171	Pharmacy	49.12	50.88
Korea Zinc	619	Material	61.07	38.93	Ildong Holdings	46	Pharmacy	65.22	34.78
Ssangyong Cement Industrial	390	Material	66.15	33.85	Hanmi Science	610	Pharmacy	43.77	56.23
Lock&Lock	652	Material	58.44	41.56	Dong-A Socio Holdings	309	Pharmacy	40.78	59.22
Korea Petrochemical Ind.	190	Material	50.00	50.00	Il-Yang Pharm	303	Pharmacy	59.41	40.59
SamKwang Glass	176	Material	45.45	54.55	Kwang dong Pharm.	721	Pharmacy	56.87	43.13
Young Poong	970	Material	45.46	54.54	CJ CheilJedang	5828	Food expenses	51.80	48.20
Hanwha Chemical	2088	Material	53.45	46.55	Samyang	342	Food expenses	55.56	44.44
Poongsan	1112	Material	53.33	46.67	Ottogi	2087	Food expenses	52.71	47.29
Lotte chemical	2992	material	50.64	49.36	Hitejinro	3764	Food expenses	53.35	46.65
DongKuk Steel Mill	2079	Material	52.00	48.00	Namyang	1375	Food expenses	58.18	41.82
Taekwang Ind.	327	Material	45.87	54.13	Muhak	1981	Food expenses	46.74	53.26
SeAh Besteel	362	Material	55.25	44.75	KT&G	3349	Food expenses	53.84	46.16
POSCO	1022	Material	49.80	50.20	Nonhshim	4093	Food expenses	48.62	51.38
Kolon Ind.	972	material	49.49	50.51	Farmsco	236	Food expenses	48.31	51.69
LG Chem	6717	material	53.48	46.52	Orion	2915	Food expenses	50.53	49.47
Lotte find Chemical Co.	178	material	60.67	39.33	Samyang Holdings	309	Food expenses	66.99	33.01
–	–	–	–	–	Dongwon F&B	1528	Food expenses	49.21	50.79
–	–	–	–	–	Lotte Chilsung	2932	Food expenses	50.20	49.80
–	–	–	–	–	Binggrae	1205	Food expenses	45.15	54.85
–	–	–	–	–	HiteJinro Holdings	127	Food expenses	48.03	51.97
–	–	–	–	–	Lotte Food	1737	Food expenses	51.41	48.59
–	–	–	–	–	Lotte Confectionery	3937	Food expenses	54.71	45.29

Therefore, the variable of the stock price y is defined as follows:

$$y_t \stackrel{def}{=} \log \frac{Close_t}{Open_t} \tag{10}$$

Transfer entropy is measured in pairs in the sectors. The KSG Estimator is used in this study. Additionally, as suggested by Lizier [44], in the non-Markov process, $k \to \infty$ becomes the optimal choice where k is the history length of the process. However, to calculate $k \to \infty$, substantial calculation time is required and the target history length shows only a small difference in the results [45]. Therefore, in this paper, since the number of time series data points in the period is approximately 500, the target history length is defined as k = 128. The source history length of l = k is selected as in Schreiber [20]. The p-value is also obtained by the method that is presented in Section 2.2.2 (Statistical Significance: p-value). In this paper, we use JIDT [46] to measure the transfer entropy using the KSG Estimator and the statistical significance. JIDT is a Java package that implements analytical computation of Information Theory and has been used in many studies [47,48].

Finally, it is necessary to estimate the stock price by selecting a threshold p-value and a meaningful causal relationship. Generally, we test the hypothesis with the significance level of 0.05 in the context of economics and social science [49]. However, the level of significance needs to change depending on context and purpose [49]. In the context of transfer entropy and the purpose of forecasting, we do not know which p-value threshold yields the best result because this study is the first to consider the causal relationship to predict stock movements with financial news. Therefore, we change the thresholds among 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5 and find the threshold that yields the best predictability for each sector by grid search.

3.3. Text pre-processing

Text pre-processing is one of the many highly significant parts of the financial-news-based stock price prediction system. This approach consists of building datasets, feature extraction, feature selection and feature representation. To start off, we must eliminate the unnecessary parts of the news. This can be done by discarding all the unnecessary information, such as HTML tags and emails.

For building datasets, we have two categories: news assigned for target firm and news assigned for causal firms.

For feature extraction, a Bag-of-words model is used. Since all words must be infinitives to utilize the Bag-of-words model, KKMA POS Tagger is used to create infinitives where every infinitive implies a component [50]. We use the Bag-of-words model on every dataset with only their training sets. If there are n causal firms, we do Bag-of-words n + 1 times. In addition, if there is more than one news article for a company on one day, we aggregate the news. As in [7], the infinitives of all words that appear in less than three articles are removed.

To select the components that have an effect on each company, Chi-square testing is performed. The Chi-square test is used to determine whether the observed frequency is significantly different from the expected frequency, and the higher the value, the more different the frequencies. In other words, features with higher value in the Chi-square test have a larger effect on the stock price. Especially, we used Chi-square testing on the target firm with target firm stock price, and on the causal firms with target firm stock price. Only the top 10% of features with the highest discriminatory power are selected in the Chi-square test. The number of features in this paper ranges from 500 to 1000. It is consistent with 567 of Hagenau et al. [10] and 500 of Shynkevich et al. [7].

Each feature is weighted after feature selection with the TF-IDF method. However, the features after TF-IDF weighting have a different scale depending on the number of the features. Thus, a scaling process is needed to appropriately execute the machine learning process because we simultaneously consider many datasets on the target firm and causal firms. At this stage, the number of features that are selected is multiplied by the same method as was used by Shynkevich et al. [7] — if the number of features in the top 10% is k, then k is multiplied by the TF-IDF (k * TF-IDF).

3.4. Machine learning techniques

Multiple Kernel Learning (MKL) is a combination of sub-kernels with positive and linear combination parameters. It is possible for a kernel to differ with a different set of parameters or have separate datasets for the same labels [51]. That is, MKL is expressed as Eq. (11), where K_s is a predefined kernel matrix, and η_s is a weight of the kernel K_s.

$$K = \sum_{s=0}^{S} \eta_s K_s \; s.\,t.\; \eta_s \geq 0 \tag{11}$$

The weight parameters of kernels are tuned during the training process. There are many MKL methods to find the optimal weight parameters [9,52]. Among them, we choose the EasyMKL method [9] to optimize the weights of kernel matrices, which is considered one of the state-of-the-art algorithms. Assuming that the training dataset is $G_{tr} = \{(x_{s,\ 1}, y_1), \ldots, (x_{s,\ i}, y_i)\}$ and the test dataset is $G_{te} = \{(x_{s,\ i+1}, y_{i+1}), \ldots, (x_{s,\ l}, y_l)\}$, where $x_{s,\ i} \in R^m$ means input data to K_s belonging to an input space X, and $y_i \in \{-1\ (down), +1\ (up)\}$ is the desired target value for the pattern $x_{s,\ i}$. We use a hat, e.g., $\widehat{Y}$, to denote the submatrices obtained considering training examples only. EasyMKL optimization is as follows.

$$\max_{\|\eta\|=1} \min_{\gamma \in \Gamma} (1-\Lambda)\gamma^T \widehat{Y}\left(\sum_{s=0}^{S} \eta_s \widehat{K}_s\right)\widehat{Y}\gamma + \Lambda \|\gamma\|^2$$

$$s.\,t.\quad \Gamma = \left\{\gamma \in R_+^L \;\middle|\; \sum_{y_i=+1} \gamma_i = 1, \sum_{y_i=-1} \gamma_i = 1\right\} \tag{12}$$

In this paper, we use the radial basis function kernel (rbf kernel). The rbf kernel is the most commonly used kernel and is applied to a variety of data since it can handle nonlinear relations [53]. In addition, because the rbf kernel can include a linear kernel [54] and a sigmoid kernel [55], depending on the range of parameters, and it has low numerical complexity [53], the rbf kernel is a reasonable choice. The rbf kernel is expressed as follows:

$$K(z_i, z_j) = e^{-\frac{\|z_i - z_j\|^2}{2\sigma}} \; s.\,t.\; \sigma > 0 \tag{13}$$

The parameter σ is a width of rbf kernel. Depending on σ, the properties of the rbf kernel change substantially so that it is important to find appropriate parameters [53]. In this paper, the grid search method is performed at σ = 0.01, 0.1, 0.25, 0.5, 1, 2, 5, 10, 15, 20, 30, 50, and 100 in a range that is similar to the one that was proposed by Hsu et al. [53].

The proposed model utilize categories of news simultaneously, news of the target firm and news of causal relation firms within GICS sector. If there are n causal firms to a target firm, we have one kernel for a target firm and n kernels for causal firms. To select the best parameter for each kernel, we implement a grid search with each category on the SVM algorithm. After finding optimal parameters, we combine kernels and implement MKL.

Since the focus of this paper is the development of a machine learning model to predict stock movement, we compare the proposed model with two state-of-the-art algorithms. The first baseline model is the work of Hagenau et al. [10]. The first baseline model utilize financial news of the target firm only, and predicts with SVM. The second baseline model is the work of Shynkevich et al. [7]. The second baseline model utilizes two categories of news simultaneously, news of target firm and aggregate news of firms within GICS sector. We assign two separate rbf kernels to the target firm and GICS-sector firms, and predict with MKL.

Table 5
Confusion matrix.

		Prediction	
		Up	Down
Actual	Up	TP	FN
	Down	FP	TN

In addition, we propose an additional algorithm (Proposed Method2) to predict stock movements even if there is no news on the target firm. Training procedures are similar to the proposed method. We build a training dataset if there is a news item on the target firm or causal firms. We use the Bag-of-words model in the same way, and use Chi-square based on the stock movements of the target firm. MKL is also used in the same way. However, if there is no news on the target firm or causal firms, the assigned kernel is not trained in the training set on that day. Since previous methods can't predict stock movements even if news on the target company did not appear, we can't compare our Proposed Method2 with other methods. So, we implement two algorithms when only news of the direct company appears.

3.5. Evaluation

We used three years of news data and stock price data from January 2014 to December 2016. To perform out-of-sample test, we set the training set to contain data collected from January 1, 2014 to December 31, 2015, the validation set to contain data collected from January 1, 2016 to June 30, 2016, and the test set to contain data collected from July 1, 2016 to December 31, 2016. The training set is required to initially fit the models. The validation set is required to tune hyper-parameters. Tuning of the parameter σ, the width of the rbf kernel, is required for both MKL and SVM. Optimal parameters are determined using a grid search. During the validation, the performance of the model with different parameter settings is measured by classification accuracy. Finally, we obtain the experiment results during the test period.

3.6. Evaluation metrics

The evaluation metrics are Accuracy and F1-score. When the experiment has been finished, if the prediction is 'Up' and the actual result is 'Up,' it is defined as True Positive (TP). If the prediction is incorrect and the actual result is 'Down,' it is defined as False Positive (FP). Similarly, if the number is predicted to be 'Down' and agrees with the actual result, it is predicted as True Negative (TN), and if the actual result is 'Up,' it is defined as False Negative (FN). It is summarized in Table 5 below.

In the Table 5, Accuracy and F1-score are defined as follows.

$$\text{Accuracy} \stackrel{\text{def}}{=} \frac{TP + TN}{TP + FP + FN + TN} \tag{14}$$

$$\text{F1} - \text{score} \stackrel{\text{def}}{=} 2 * \frac{precision * recall}{precision + recall} \text{ where precision} \stackrel{\text{def}}{=} \frac{TP}{TP + FP} \& recall \stackrel{\text{def}}{=} \frac{TP}{TP + FN} \tag{15}$$

4. Results

This chapter discusses the results from the proposed news-based stock movement forecasting system. Before analyzing the results of the research, we analyze the causal relationship between the companies through transfer entropy within the GICS sectors and find the indicators that reflect the level of causality.

The results are shown in Table 6 below. 'Average' means the average of p-value of all relationships, and 'Ratio (threshold)' means the ratio of the number of all causal relationships with p-value under 'threshold' to the number of all relationships. The lower the 'Average,' the higher the number of causal relationship within a sector. The higher the 'Ratio,' the higher the number of causal relationships within a sector. Accordingly, the Food Expenses sector and Durable goods sectors are sectors with a high level of causality. The Material and Capital goods are sectors with a middle level of causality. Pharmacy, Car, and Hardware are sectors with a low level of causality.

To show the robustness of our proposed method, we do not choose only one sector, but choose three sectors for evaluation of the proposed method. Because the Hardware sector has a small number of component companies, we remove Hardware sector to avoid sample size bias for evaluation. To prove that the results do not change by the ratio of the causal relationships, we select the Pharmacy, Material, and Food Expenses sectors, which have a low level of causality, a middle level of causality, and a high level of causality respectively, for further analysis.

The following Fig. 4 depicts the stock price-based causal relationships among the three sectors: Food Expense, Pharmacy, and Materials. Each node represents a company within a sector and links and arrows represent causal relationships that affect firms. Causality that has a p-value of < 0.2 was selected in Fig. 4.

In this section, a comparative analysis of the proposed method and existing methods is conducted. We compare the accuracy and F1-score of stock movement prediction using the causal relationship analysis method (Proposed approach) with two state-of-the-art algorithms. One method applies the influence of the GICS sector (Shynkevich et al. [7]), and the other is at the individual level, which can be used as a basis for comparison (Hagenau et al. [10]) which are.

4.1. Results of the research model

Table 7 displays the experimental results obtained from a comparison between the proposed model and two state-of-the-art algorithms. The first column of Table 7 shows the results produced when the proposed method is used. The second column represents the results with the state-of-the-art method considering the influence of the GICS sector. The third column shows the results with the state-of-the-art method of the individual level. The rows reflect in which GICS sector the results

Table 6
Causality statistical verification.

GICS Sector	Average	Ratio (0.05)	Ratio (0.1)	Ratio (0.2)	Ratio (0.3)	Ratio (0.4)	Ratio (0.5)
Hardware	0.5860	0.0156	0.0625	0.1094	0.1250	0.2031	0.2188
Car	0.5549	0.0622	0.0933	0.1555	0.2222	0.3288	0.3822
Pharmacy	0.5501	0.0553	0.0969	0.1592	0.2595	0.3599	0.4429
Capital goods	0.5224	0.0539	0.0914	0.1714	0.2620	0.3510	0.4490
Material	0.5124	0.0476	0.0942	0.1972	0.3001	0.3901	0.4922
Durable goods	0.4834	0.0347	0.055	0.1875	0.3056	0.3889	0.4930
Food expenses	0.4974	0.0415	0.1142	0.2284	0.3114	0.4430	0.5225

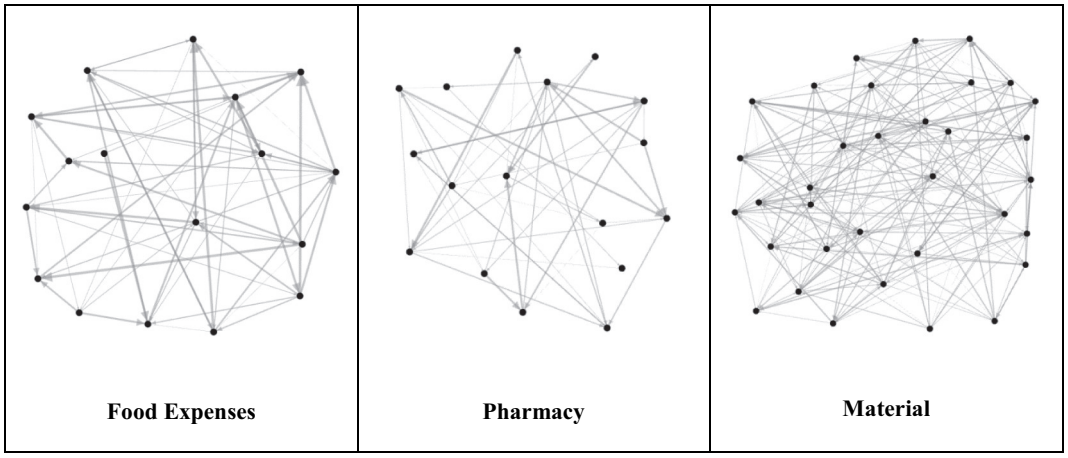

Fig. 4. Causal relationship graphs of companies within the sectors.

Table 7
Experimental results.

GICS Sector	Results	Proposed Method	Shynkevich et al. [7]	Hagenau et al. [10]
Pharmacy	Average	**0.584381**	0.551828	0.534255
	Standard deviation	0.073236	0.078205	0.085140
Material	Average	**0.623473**	0.588867	0.585316
	Standard deviation	0.072492	0.062896	0.061214
Food expenses	Average	**0.558975**	0.546884	0.542623
	Standard deviation	0.084652	0.049542	0.050500

Table 8
Experimental results by p-value.

p-Value		Pharmacy	Material	Food expenses
		Results	Results	Results
0.05	Average	0.571615	0.585959	0.53658903
	Standard deviation	0.072737	0.071329	0.082699234
0.1	Average	0.57212	0.5872	0.540213
	Standard deviation	0.075738	0.065488	0.080919
0.2	Average	0.570321	0.594735	0.53771
	Standard deviation	0.071073	0.074038	0.083431
0.3	Average	0.564806	0.594902	**0.542576**
	Standard deviation	0.076117	0.079983	0.083491
0.4	Average	**0.575014**	**0.597599**	0.541202
	Standard deviation	0.078479	0.085422	0.083596
0.5	Average	0.571514	0.588698	0.537125
	Standard deviation	0.081611	0.089186	0.081719

are obtained. 'Average' is the average of results obtained using all three methods. 'Standard Deviation' represents the standard deviation of the results. The highest forecasting accuracy is marked in bold and underlined. In all GICS sectors, the proposed method showed the best predictive power. Detailed results are presented in Table 7 below.

In this paper. we use transfer entropy to predict the stock price by analyzing the causal relationship between firms. We used the p-value to determine the statistical significance of the causal relationship. We experimented by changing the statistical significance level as stated in Section 3.2. Table 8 shows the average forecasts and standard deviations for stock movement accuracies for each sector according to the p-value thresholds. The Pharmacy sector and Material sector show the highest accuracy when the p-value equals 0.4, while the Food expenses sector shows the highest accuracy when the p-value equals 0.3. The prediction accuracy significantly changes for each sector according to the change of the p-value. Accordingly, we suggest that controlling the statistical significance differently affects the prediction and we need to perform a grid search for every company to find the best p-value for appropriate prediction.

4.2. Additional analysis

One of the most important factors in predicting stock prices based on the news is whether the articles on the company are published at the relevant time. In general, it is difficult to perform an analysis at the individual company level when there are no articles that are related to the company. However, there has been no attempt to overcome this limitation, to the best of our knowledge. By analyzing the causal relationship, we overcome these problems. In other words, even if direct corporate news does not appear, we can predict the stock movements with news on the causal companies. Although there are differences depending on the sector, this research method shows higher accuracy than other methods for two out of three sectors. Detailed results are shown in Table 8 below. In Table 9, the Proposed Method2 is a result

Table 9
Experimental results for the Proposed Method2.

GICS Sector	Results	Proposed Method2	Shynkevich et al. [7]	Hagenau et al. [10]
Pharmacy	Average	**0.566711**	0.551828	0.534255
	Standard deviation	0.041933	0.078205	0.085140
Material	Average	0.578514	**0.588867**	0.585316
	Standard deviation	0.033816	0.062896	0.061214
Food expenses	Average	**0.602333**	0.546884	0.542623
	Standard deviation	0.036026	0.049543	0.050500

Table 10
Experimental results – F1-score.

GICS sector	Results	Proposed method		Shynkevich et al. [7]	Hagenau et al. [10]
		Proposed Method	Proposed Method2		
Pharmacy	Average	**0.603790**	0.596246	0.563747	0.563615
	Standard Deviation	0.121985	0.068114	0.123071	0.118995
Material	Average	**0.689825**	0.623628	0.677068	0.677142
	Standard Deviation	0.074463	0.058707	0.087689	0.081261
Food expenses	Average	0.601027	0.574451	**0.630857**	0.630411
	Standard Deviation	0.085216	0.068521	0.073959	0.073526

Table 11
Benchmark against two state-of-the-art causality detection algorithms.

GICS Sector	Results	Proposed method	Oh et al. [33]	Výrost et al. [29]
Pharmacy	Average	**0.584381**	0.564935	0.573040
	Standard deviation	0.073236	0.071273	0.075190

from prediction where analysis was implemented even if news on the target company did not appear.

To confirm the robustness of stock movement forecasting, we check the stability of the model by examining it from various angles. We not only examine the accuracy but also examine F1-score. Overall, the proposed method was dominant. Details of the results are shown in Table 10 below.

To test our causality detection is valid, we benchmark with two state-of-the-art causality detection papers [29,33]. Oh et al. [33] means considering only uni-directional causality, and Výrost et al. [29] state causality detection using Granger causality. We implemented all the procedures in the Pharmacy sector including text preprocessing, MKL and grid search, except the causality detection. Results show that the proposed method shows better results than two state-of-the-art methods. Especially, results suggest that considering bi-directional causality is important and TE shows better results than Granger causality in the Pharmacy sector. Details of the results are shown in Table 11.

We also implemented sensitivity analysis in the Pharmacy sector to check if prediction power is different by news article sources. Results from sensitivity analysis showed that there is no statistically significant difference between news article sources.

5. Conclusions and future work

Entering the era of Big Data, stock price forecasting using machine learning has been actively studied, and research on predicting stock prices based on unstructured data has attracted considerable attention. Since it is impossible for investors to read all of the news about stocks, investors can gain potential benefits by using automated systems that can identify information from multiple sources and accurately predict changes in market prices. We propose a machine learning algorithm that predicts stock movements by analyzing financial news.

In this study, we analyze the causal relationships between firms that have not been controlled by the analysis of the individual companies and within a sector to improve the accuracy of the model. This method improves predictability by overcoming the main limitation of previous research that bidirectional influence within the GICS sectors is assumed. We apply the transfer entropy method, which is actively used in the field of physics, and we analyze the causal relationship clearly and apply it to the prediction model. The analysis was conducted at the sector level to identify causal relationships that had an impact on firms. The reason for analyzing the causal relationships at the sector level without analyzing all the companies is that this approach may show a spurious causality [41] because it finds the causality with time series. To prevent this, we perform our analysis in the economically classified GICS sector.

The results of this study show that high prediction accuracy and F1-score are obtained with the causal relationship between firms. Our results propose that the directional relationship between companies should be analyzed and it needs to be reflected in the prediction phase. In addition, results show that the Proposed Method2 forecasts the stock movements well even when there is no financial news on the target firm, but financial news is published on causal firms. We also find that the results vary with the p-value threshold of transfer entropy. Therefore, it is important to find proper the threshold through a grid search.

We have a limitation in that we searched only three sectors so that we did not find a correlation between characteristics of the sector and the proposed method. In future work, we should find the relationship between the level of causality of the sector and the proposed approach. Additionally, we have a limitation in that the threshold of the p-value needs to be determined by a grid search. In the future work, we need to find a way to set the threshold by analyzing the characteristics of the GICS sector, or the characteristics of firms. Another possible direction for future work is to analyze the causal relationship not within the GICS

sector but within the KOSPI 200. We can perform a more sophisticated analysis by examining various relationships and not being limited by the assumption that we have a relationship only within the sector.

References

[1] A.K. Nassirtoussi, S. Aghabozorgi, T.Y. Wah, D.C.L. Ngo, Text mining for market prediction: a systematic review, Expert Syst. Appl. 41 (2014) 7653–7670.
[2] E.F. Fama, Random walks in stock market prices, Financ. Anal. J. 51 (1995) 75–80.
[3] C. Eom, S. Choi, G. Oh, W.-S. Jung, Hurst exponent and prediction based on weak-form efficient market hypothesis of stock markets, Physica A Stat. Mech. Appl. 387 (2008) 4630–4636.
[4] H.E. Hurst, Long term storage capacity of reservoirs, ASCE Trans. 116 (1951) 770–808.
[5] Z.M. Shi, G. Lee, A.B. Whinston, Toward a better measure of business proximity: topic modeling for industry intelligence, Manag. Inf. Syst. Q. 40 (2016) 1035–1056.
[6] R.P. Schumaker, H. Chen, A quantitative stock prediction system based on financial news, Inf. Process. Manag. 45 (2009) 571–583.
[7] Y. Shynkevich, T.M. McGinnity, S.A. Coleman, A. Belatreche, Forecasting movements of health-care stock prices based on different categories of news articles using multiple kernel learning, Decis. Support. Syst. 85 (2016) 74–83.
[8] O. Kwon, G. Oh, Asymmetric information flow between market index and individual stocks in several stock markets, EPL 97 (2012) 28007.
[9] F. Aiolli, M. Donini, EasyMKL: a scalable multiple kernel learning algorithm, Neurocomputing 169 (2015) 215–224.
[10] M. Hagenau, M. Liebmann, D. Neumann, Automated news reading: stock price prediction based on financial news using context-capturing features, Decis. Support. Syst. 55 (2013) 685–697.
[11] B. Wuthrich, V. Cho, S. Leung, D. Permunetilleke, K. Sankaran, J. Zhang, Daily stock market forecast from textual web data, IEEE International Conference on Systems, Man, and Cybernetics, IEEE, 1998, pp. 2720–2725.
[12] Q. Li, T. Wang, P. Li, L. Liu, Q. Gong, Y. Chen, The effect of news and public mood on stock movements, Inf. Sci. 278 (2014) 826–840.
[13] S.S. Groth, J. Muntermann, An intraday market risk management approach based on textual analysis, Decis. Support. Syst. 50 (2011) 680–691.
[14] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, J. Comput. Sci. 2 (2011) 1–8.
[15] M. Paluš, V. Komárek, Z. Hrnčíř, K. Štěrbová, Synchronization as adjustment of information rates: detection from bivariate time series, Phys. Rev. E 63 (2001) 046211.
[16] R. Marschinski, H. Kantz, Analysing the information flow between financial time series, Eur. Phys. J. B 30 (2002) 275–281.
[17] P. Verdes, Assessing causality from multivariate time series, Phys. Rev. E 72 (2005) 026222.
[18] O. Kwon, J.-S. Yang, Information flow between stock indices, EPL 82 (2008) 68003.
[19] J. Runge, J. Heitzig, N. Marwan, J. Kurths, Quantifying causal coupling strength: a lag-specific measure for multivariate time series related to transfer entropy, Phys. Rev. E 86 (2012) 061121.
[20] T. Schreiber, Measuring information transfer, Phys. Rev. Lett. 85 (2000) 461.
[21] T. Bossomaier, L. Barnett, M. Harré, J.T. Lizier, An Introduction to Transfer Entropy, Springer, 2016.
[22] T.M. Cover, J.A. Thomas, Elements of Information Theory, John Wiley & Sons, 2012.
[23] R. Vicente, M. Wibral, Efficient estimation of information transfer, Directed Information Measures in Neuroscience, Springer, 2014, pp. 37–58.
[24] M. Wibral, R. Vicente, M. Lindner, Transfer entropy in neuroscience, Directed Information Measures in Neuroscience, Springer, 2014, pp. 3–36.
[25] L. Barnett, T. Bossomaier, Transfer entropy as a log-likelihood ratio, Phys. Rev. Lett. 109 (2012) 138105.
[26] K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, J. Bhattacharya, Causality detection based on information-theoretic approaches in time series analysis, Phys. Rep. 441 (2007) 1–46.
[27] N. Wiener, The theory of prediction, Modern Mathematics for Engineers, 1956, pp. 165–190 DOI.
[28] C.W. Granger, Investigating causal relations by econometric models and cross-spectral methods, Econometrica J. Econ. Soc. (1969) 424–438 DOI.
[29] T. Výrost, Š. Lyócsa, E. Baumöhl, Granger causality stock market networks: temporal proximity and preferential attachment, Physica A Stat. Mech. Appl. 427 (2015) 262–276.
[30] F.A. Razak, H.J. Jensen, Quantifying 'causality' in complex systems: understanding transfer entropy, PLoS One 9 (2014) e99462.
[31] T.D.M. Peron, F.A. Rodrigues, Collective behavior in financial markets, EPL 96 (2011) 48004.
[32] T. Lux, M. Marchesi, Scaling and criticality in a stochastic multi-agent model of a financial market, Nature 397 (1999) 498.
[33] G. Oh, T. Oh, H. Kim, O. Kwon, An information flow among industry sectors in the Korean stock market, J. Korean Phys. Soc. 65 (2014) 2140–2146.
[34] A. Sensoy, C. Sobaci, S. Sensoy, F. Alali, Effective transfer entropy approach to information flow between exchange rates and stock markets, Chaos, Solitons Fractals 68 (2014) 180–185.
[35] E.J. De Fortuny, T. De Smedt, D. Martens, W. Daelemans, Evaluating and understanding text-based stock price prediction models, Inf. Process. Manag. 50 (2014) 426–441.
[36] T.H. Nguyen, K. Shirai, J. Velcin, Sentiment analysis on social media for stock movement prediction, Expert Syst. Appl. 42 (2015) 9603–9611.
[37] P.C. Tetlock, All the news that's fit to reprint: do investors react to stale information? Rev. Financ. Stud. 24 (2011) 1481–1512.
[38] Y. Yu, W. Duan, Q. Cao, The impact of social and conventional media on firm equity value: a sentiment analysis approach, Decis. Support. Syst. 55 (2013) 919–926.
[39] M. Kraus, S. Feuerriegel, Decision support from financial disclosures with deep neural networks and transfer learning, Decis. Support. Syst. 104 (2017) 38–48.
[40] R. Luss, A. d'Aspremont, Predicting abnormal returns from news using text classification, Quant. Finan. 15 (2015) 999–1012.
[41] Z. He, K. Maekawa, On spurious Granger causality, Econ. Lett. 73 (2001) 307–313.
[42] S. Bhojraj, C. Lee, D.K. Oler, What's my line? A comparison of industry classification schemes for capital market research, J. Account. Res. 41 (2003) 745–774.
[43] L. Sandoval, Structure of a global network of financial companies based on transfer entropy, Entropy 16 (2014) 4443–4482.
[44] J.T. Lizier, M. Prokopenko, A.Y. Zomaya, Local information transfer as a spatio-temporal filter for complex systems, Phys. Rev. E 77 (2008) 026110.
[45] J.T. Lizier, Measuring the dynamics of information processing on a local scale in time and space, Directed Information Measures in Neuroscience, Springer, 2014, pp. 161–193.
[46] J.T. Lizier, JIDT: an information-theoretic toolkit for studying the dynamics of complex systems, Front. Robot. AI 1 (2014) 11.
[47] J. Garland, R.G. James, E. Bradley, Leveraging information storage to select forecast-optimal parameters for delay-coordinate reconstructions, Phys. Rev. E 93 (2016) 022221.
[48] D. Darmon, P.E. Rapp, Specific transfer entropy and other state-dependent transfer entropies for continuous-state input-output systems, Phys. Rev. E 96 (2017) 022121.
[49] R.L. Wasserstein, N.A. Lazar, The ASA's statement on p-values: context, process, and purpose, Am. Stat. 70 (2016) 129–133.
[50] D.-J. Lee, J.-H. Yeon, I.-B. Hwang, S.-G. Lee, KKMA: a tool for utilizing Sejong corpus based on relational database, J. KIISE Comput. Pract. Lett. 16 (2010) 1046–1050.
[51] Y. Gu, C. Wang, D. You, Y. Zhang, S. Wang, Y. Zhang, Representative multiple kernel learning for classification in hyperspectral imagery, IEEE Trans. Geosci. Remote Sens. 50 (2012) 2852–2865.
[52] A. Jain, S.V. Vishwanathan, M. Varma, SPF-GMKL: generalized multiple kernel learning with a million kernels, Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2012, pp. 750–758.
[53] C. Hsu, C. Chang, C. Lin, A Practical Guide to Support Vector Classification, Department of Computer Science National Taiwan University, Taipei, 2010.
[54] S.S. Keerthi, C.-J. Lin, Asymptotic behaviors of support vector machines with Gaussian kernel, Neural Comput. 15 (2003) 1667–1689.
[55] H.-T. Lin, C.-J. Lin, A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods, submitted to, Neural Comput. 3 (2003) 1–32.

Kihwan Nam received a Ph.D. in information systems from the College of Business, Korea Advanced Institute of Science and Technology (KAIST). He is currently Adjunct Professor, College of Business, Ulsan National Institute of Science and Technology (UNIST). His research interests are focused on Quantitative Marketing, Recommender System, Big Data Analytics, Data Mining, Statistical Analysis, Applying Machine Running and Deep Learning in Business Analytics and Econometric model. He is also a data scientist. In addition to his academic research, he is making a positive contribution to both academia and industry by successfully carrying out various projects in a big international company based on theory.

NohYoon Seong is a Ph.D. Candidate in information systems from the College of Business, Korea Advanced Institute of Science and Technology (KAIST). He graduated from KAIST with a Bachelor degree in physics. His research interests include Natural Language Processing, Big Data Analytics, Machine Learning, Econophysics, and Applying Machine Learning and Econometric model in Business Analytics. He also participated in many industrial projects with big international companies as a data scientist.