ABSTRACT Investment universe means a pool of selected assets likely to be profitable. In general, assets related to a common theme or concept are selected to form an investment universe. A well organized investment universe has a strong common theme and reduces the risk by diversifying the assets included in the universe. To form or revise an investment universe, human efforts are required by using domain knowledge about the theme, but it is hard to suggest an investment universe that reflects the latest market trends. In this paper, we propose an automated investment universe selection method based on theme keywords. The theme keywords are extracted from the news articles and the business section on the companies which are included in the S&P 500 index. After that, securities are selected to form an investment universe for each theme keywords. We employ a similarity value between the security vector and the theme keyword vector to select the related securities for the certain keywords. Regarding the vector representations, word and document embeddings are carried out using both news articles and the business section on the companies. Stock price movements of the selected securities are similar which means that the investment universes are well organized to suggest the assets tightly associated with the theme. The experimental results show that the proposed method has high future returns on the investment universes with low or high historical stock price returns.

INDEX TERMS Doc2vec, investment universe, keyword search, 10-K report, theme keywords
trader model [4]. They are characterized to have low trading volume and difficulties in obtaining the financial information compared to the institutional investors. The retail investors may also use third-party sources or audit reports to obtain the financial information [5]. In this study, we propose a system that automatically constructs an investment universe by using the theme keywords. The investment assets of this system is limited to the stocks included in the S&P500 Index [6]. The 10-K reports [7] and news articles are then learned through Doc2vec. The investment universe will be set up by using the topics related to the theme keywords that correspond to the current market trends.

It is not difficult to find examples of the market trends through keywords. Google ads keyword planner [8] is a keyword search tool that anyone can use to find the right targeting keywords for their display, search, video, and application advertisements [9]. It is used to analyze the latest trends in keywords to modify the search campaigns and to maintain the relevance of contents. In addition, Google Trends [10] provides daily high-rising search keywords and real-time fast-rising search keywords by countries with Google’s search keywords and video-based big data analysis services. Each keyword shows interest in time. As you can see from the above example, the keyword search can be used to identify the market trend. In this study, we try to extract the keywords that can understand the market trends and recent issues through the keyword search approach and construct the investment universe by automatically extracting the S&P500 stocks [6] related to these keywords.

Building an investment universe for each investor requires a lot of effort. For example, categorizing the types of businesses that each company conducts and listing the problems that have recently been associated with the company are essential [11]–[13]. These type of effort to understand the companies can help build an investment universe that people want to invest in. In this research, we aim to propose a system to automatically build an investment universe based on the theme keyword search with minimal human effort, which will make a rapid response to the market trends and changes possible.

II. RELATED WORKS

One way to build an investment universe of equity stocks is to use industry classification [13]. There are many studies in the industry classification, both academia, and industry [14], [15]. One study argues that the purpose of industrial classification is creating "discrete categories, by maximizing the differences between industry groups and similarities of components within industry groups" [13]. Industry classifications have also been used to evaluate strategies for a diversified portfolio [14]. There have also been studies on industry classification that are consistent with the purpose of this study. There have been studies that utilize the 10-K report of the U.S. Securities and Exchange Commission [7] to minimize human effort and automatically classify the industry [16]. In addition, there was an attempt to construct an industry network by using news articles [17]. As an another approach, a deep learning based automated industry classification was suggested [18]. These efforts are consistent with the objectives of this study in that they minimize human efforts and enable rapid response to market changes and new problems. The main difference between previous works and the proposed method is training Doc2vec [19] using not only 10-K reports [7] but also Reuters news articles. In the proposed method, we used Reuters articles and 10-K reports in Doc2vec training to diversify keywords whereas only 10-K reports were used in Doc2vec training in case of previous works. Another difference is keyword-based clustering of securities. The purpose of [16], [17] was automatic industry classification, so that they clustered the securities with similar business lines. In this study, rather than using the arbitrarily classified industry group to construct the investment universe, the theme keyword related to the market trend is used to set up the enterprises related to these keywords as the object of the investment universe.

Examples of using keywords to identify market trends can be found easily. As an example of applying keyword search to business, there have been researches that use the keyword planner of Google ads as a marketing strategy using internet [20]. There have also been attempts to investigate emerging topics to identify trends using Twitter, a typical social network service [21]. Unlike these general trends, some researches have been carried out to find out the direction of technology development based on the technical keyword by using patent data and establish a technology development plan [22]. There has been a study that used Google Trend [10] to identify market trends in software engineering [23]. Google Trend is a service that examines news stories and search volume collected by Google’s search engine. Another research about predicting academic fame using Google Trend have been proposed [24]. There have also been cases of research using the Google Trend in investment. They used Google Trend and Yahoo finance information for stock price forecasting [25].

III. METHODOLOGY

The investment universe selection is a challenging task since it is hard to select securities that price is going to increase [26], [27]. In this work, we assumed that the price of securities related to the current market trends will increase. The keywords about current market trends are extracted before the investment universe construction. The extracted keywords are used to select the securities predicted to increase. Both keyword extraction and security selection are carried out based on the document and word embedding. We employed Doc2vec model learned on the Reuters articles and the 10-K reports from the U.S. Securities and Exchange Commission as an embedding method. The overall framework of the investment universe construction is presented in Fig. I. This framework shows the case when the keyword ‘Vehicles’ is used as an example of the theme keyword. A detailed explanation of each step in the framework is given in
the following subsections. Our method consists of three parts: III-A document and word embedding, III-B theme keywords extraction, and III-C investment universe selection.

**A. DOCUMENT AND WORD EMBEDDING**

In order to extract the theme keywords, document and word embedding must be done first. In this research, we used the Reuters articles and the 10-K reports from the U.S. Securities and Exchange Commission for document and word embedding [28]. An illustration of the Doc2vec algorithm is shown in Fig. 2. In the figure, \( s_t, w_t \) denote a security vector and a word vector for \( t \), respectively and the document can be either Reuters news article or 10-K report. According to [28], if there is a sequence of words \( \{w_t\}_{t=1}^{T} \), then the objective function of Doc2vec is defined to maximize the log-probability as follows.

\[
\text{Maximize } \frac{1}{T} \sum_{t=1}^{T-i} \log p(w_{t-i}, \cdots, w_{t+i}). \tag{1}
\]


where \( |V| \) represents the number of total unique words in 2017 Reuters news articles. Since document embeddings and word embeddings are trained concurrently, we can compute the similarity between documents and words.

**B. THEME KEYWORDS EXTRACTION**

The proposed method can provide investment universe for all trained keywords. However, examining the investment universes for all keywords is inefficient. There are too many meaningless words in the entire trained vocabulary. In this study, we assumed that business areas related to keywords...
that represent the market trend will develop and that the stock prices of companies related to that business area will increase accordingly. Therefore, this study defines the theme keywords as a set of words attracted attention in the society during the last year. These theme keywords may contain generic words such as ‘campaign’ or specific words related to the financial industry such as ‘securities’. It can also include an event such as ‘Brexit’, which recently has been focused, or a name of a new technology. The method used in this study to extract the theme keywords is as follows.

A theme keyword can be extracted in various ways. For example, we can refer to famous trend analysis services such as Google Trend [10]. In this paper, we assumed that the important keywords appear more frequently than the less important words in the 10-K reports or the Reuters news articles. This assumption is based on our text pre-processing that extracts only noun and proper noun words, and eliminates pre-defined stopwords. An example of text-preprocessing of the Reuters news article that extracts only noun and proper noun words, and eliminates pre-defined stopwords.

FIGURE 3. An example of text-preprocessing of the Reuters news article that extracts only noun and proper noun words, and eliminates pre-defined stopwords.

The above theme keywords are used to construct the investment universe for each keyword.

C. INVESTMENT UNIVERSE SELECTION

As described in the subsection III-A, embedding vectors of keywords and securities are trained by Doc2vec. After the training, we can get similarity values between each theme keywords and securities included in the S&P 500 index. We use cosine similarity between theme keyword embeddings and security embeddings as the similarity measure.

The proposed method constructing the investment universe is summarized in Algorithm 1. The number of securities in an investment universe can be different for each theme keywords. To adaptively select the number of securities, we adopted clustering to the universe selection. The companies associated with the keyword are likely to form a cluster. In this study, it is assumed that the investment universe consists of 3 to 50 stocks. Since the similarity is a scalar, clustering analysis is computationally light. To select securities, 50 companies are first culled in the order of similarities, and a pair of two clusters are separated using the $k$-means clustering method with $k = 2$. The cluster with larger similarity value is defined as the investment universe.

As mentioned earlier, this study aims to establish an investment universe only for the companies included in the S&P 500. There is no guarantee that the stock prices of the companies in this universe will rise immediately, even if you choose the right keywords to match the market trend. However, if the keyword is related to the market trend and the market trend does not change rapidly, the related businesses will develop, and the stock prices of the related companies should rise in a long term. Although it is difficult to set a proper period for examining the stock return, the long-term profit examination can be used to indirectly test the performance of the system for building the investment universe.

IV. EXPERIMENT

A. 10-K REPORTS AND NEWS ARTICLES

Every Reuters news article between 2017-01-01 and 2017-12-31 were received from the Reuters website 1, and the 10-K reports were collected from the U.S. Securities and Exchange Commission website 2, which we only used the business section of the 10-K report. In the previous work, Yang et al. (2016) has applied document embedding by applying Doc2vec to the business section of the 10-K report prior to 2018-03-01 [17], and the most recent work by Jeon et al. (2017), which improved the prior research further, extracted only nouns from the Bloomberg news articles using their own news curation algorithm [16]. In this study, document and word embedding were performed by applying Doc2vec to the Reuters articles and the business section of the 10-K report together.

1 http://www.reuters.com/resources/archive/us/
2 https://www.sec.gov/edgar/searchedgar/legacy/companysearch.html

Algorithm 1: Investment Universe Construction

Input: Theme Keywords with embeddings
and Securities with embeddings

Output: List of (Theme Keyword, Related Securities)

for word ∈ Theme Keywords do
    sim_list = cosine_similarity(word, Securities)
    sort(sim_list) as descending order of similarity
    candidates = top 50 elements of the sim_list
    cluster1, cluster2 = $k$-means(candidates)
    append (word, cluster1) to the result list

Output: List of (Theme Keyword, Related Securities)
As a data pre-processing, we used only noun words and proper noun words by using the Natural Language Toolkit (NLTK) Parts of Speech (POS) tagger [30] which is one of the most widely used language processing package for English. We used predefined stopwords list and kept the words that appeared more than or equal to five times in the entire corpus. To remove highly frequent words, we removed the words that appeared more than 80 percent of the documents. All words contain special characters except ‘-’ are removed from the remaining words list. After all the pre-processing, 114,949 tokens were used to train embedding vectors. Doc2vec is carried out with Python and Gensim package [31] because it offers various options for training and utilizing the document and word embedding. The dimension of embedding vectors, the window size, and the number of training epoch were set as 50, 2, and 10, respectively.

B. S&P500 INDEX AND INDIVIDUAL STOCK PRICES

We proposed a methodology for extracting theme keywords representing the market trend in subsection III-B. After selecting the securities for the universe of theme keywords, it is necessary to examine whether the investment universe is rightly chosen and whether the investment universe can be used to make a profit. To analyze these two possibilities, we use the S&P500 index and individual stock prices. The S&P500 index and individual stock prices were collected from Yahoo finance.

First, we utilize the stock price data of the S&P500 stocks to confirm that the investment universe has been selected appropriately. We extracted the theme keywords and constructed an investment universe for each keyword. For example, suppose we set up the ‘vehicle’ related stocks as the investment universe. The items in this investment universe are likely to have a similar impact on the vehicle-related problems such as oil price changes or electric car policies. These effects will be reflected to the prices of these stocks, and these stocks have similar trends. Using this characteristics, hedge funds may use pairs trading strategy [32]. In other words, if the returns of the stocks from the investment universe are positively correlated with each other, it is possible to indirectly confirm that the investment universe corresponding to the theme keyword is properly constituted.

Note, however, that stock prices of the S&P 500 indexes have a positive correlation as a whole because they are affected by the financial markets. Therefore, in order to examine the value of the average correlation of the investment universe, it is necessary to select an appropriate comparison target. In this study, the average correlation of all stocks in the S&P500 was used as the comparison target.

C. THE AVERAGE CORRELATION OF STOCKS

As mentioned in subsection IV-B, we examine the correlation of the stocks belonging to an investment universe in order to ascertain whether the investment universe is properly extracted. The stocks from the same investment universe will show similar movements. However, correlation is not a value representing the investment universe because it is the value that appears between the two stock returns. Thus, in this study, we use the concept of average correlation as a value representing the investment universe [33]. First, we can construct a correlation matrix of the stocks from the investment universe with their stock returns. The average correlation is defined as follows, according to [33].

\[
\rho_{avg} = \frac{2 \sum_{i=1}^{N} \sum_{j>i}^{N} w_i w_j \rho_{i,j}}{1 - \sum_{i=1}^{N} w_i^2}.
\]

where \( \rho_{i,j} \) represents the full correlation matrix and \( w_i \) is the weight in the stock portfolio. Here, we used \( w_i = \frac{1}{N} \) because we want to check not the profitability of the portfolio but the propriety of investment universe construction. Using the average correlation in (5), we can see that if the stock returns in the investment universe are positive, they exhibit similar movements. However, basically, stock prices are characterized by a positive correlation as a whole because they are affected by the financial markets. Therefore, in order to examine the value of the average correlation of the investment universe, it is necessary to select an appropriate comparison target. In this study, the average correlation of all stocks in the S&P500 was used as the comparison target.

V. EVALUATION

A. THEME KEYWORDS

We trained document embedding and word embedding on the same embedding space by using Doc2vec. There are three cases of document/word matching used for training Doc2vec embeddings: between a word and another word, between a document and another document, and between a word and a document. Since we used the matching results between keywords and related document embeddings of companies, the matching results between words and documents are the most important.

The matching result graph between words and documents based on the trained embeddings is presented in Fig. 4. Financial companies and the words most similar to the company are shown in the graph. For example, ‘Bank of..."
America Corp’ is related to the words ‘transaction’, ‘banking’, ‘financial’, ‘asset’, and so on. In the case of a word-to-company relation, financial words are related to many financial companies. As mentioned earlier, there are 114,949 unique tokens in the trained Doc2vec vocabulary. In order to extract keywords related to the market trend or important financial aspect, we extracted the 500 most frequent keywords and used them as the theme keywords. We assumed that the trendy or important keywords are frequently used in the news articles.

B. COMPOSITION OF INVESTMENT UNIVERSES
As mentioned in subsection III-C, the investment universe is composed of 3 to 50 companies, and the cosine similarity value of these companies and the theme keywords extracted from their investment universe can be obtained. A high cosine similarity implies a high relevance between the investment universe and the theme keyword. An exemplary investment universe based on the theme keywords extracted from subsection V-A is shown in Table 1. Looking back at the ‘vehicle’ example from subsection IV-B, ‘Ford Motor’, ‘Copart Inc.’, ‘General Motors’, ‘Harley-Davidson’, ‘Carmax Inc’, etc. are included in the investment universe. ‘Ford Motor’ sells automobiles and commercial vehicles. ‘Copart Inc.’ is a provider of online vehicle auction. ‘General Motors’ designs and manufactures vehicles and their parts. ‘Harley-Davidson’ is a motorcycle manufacturer. ‘Carmax Inc.’ is a used-car retailer. We extracted the investment universe with the theme keyword ‘vehicle’ and all the companies included in the investment universe are considered to be related to the word ‘vehicle’ in the business area. In most cases, the business domains were directly related to the theme keywords, but the cases where the business domains were indirectly related to the theme keywords also exist. The investment universe extracted with the theme keyword ‘oil’ includes ‘Dominion Energy’, which is not a company that drills oil like other companies in the investment universe. However, they use oil in their electricity production. Thus, when constructing an investment universe based on a given theme keyword, we can expect indirectly related companies to be extracted.

As mentioned in subsection IV-B, the theme keyword used to construct the investment universe will be associated with the companies in the investment universe and will have a similar effect on their stock price when there are related problems. As a result, the stock prices of the companies in the investment universe will move similarly. Therefore, we compare the average correlation of the stock price returns of the companies in the investment universe to the average correlation of the stock price returns of the S&P500 in order to indirectly confirm that the investment universe is correctly extracted. The average correlation of the stock returns belonging to the investment universe is shown in Fig. 5 as a histogram. The red line represents the average correlation of the stock returns of the S&P500 companies.

As in Fig. 5, the overall average correlation of the S&P500 companies is 0.2491. On the other hand, in case of the average correlation of the investment universes, the maximum, the minimum, and the mean of the average correlation are 0.6474, 0.1441, and 0.3191, respectively. The mean average correlation of the investment universes is higher than the overall average correlation of the S&P500 in the total investment universe. This implies that the stock price movements of the companies in the most of the investment universes are similar to each other. The theme keywords which showed the maximum and minimum values of average correlations, were ‘securities’ and ‘bid’, respectively. The investment universe extracted by ‘securities’ consisted of financial companies such as ‘PNC Financial Services’, ‘JPMorgan Chase & Co.’, and ‘BlackRock’ in the order of high cosine similarity. The investment universe extracted by ‘bid’ includes ‘Wynn Resorts Ltd’, ‘NRG Energy’, and ‘Dish Network’. Because the keyword ‘securities’ is a term that is often dealt with by financial companies, these companies will show similar movements in macroscopic indicators such as interest rates and exchange rates, and will show a higher average correlation. On the other hand, ‘bid’ is the keyword that is used to present price in auction. The average correlation of the firms extracted by the theme keyword ‘bid’ is low because these
The purpose of this study is to construct an investment universe related to the theme keyword and make it available for investments. An investment universe can be thought of as a pool of assets to invest. If we want to invest in the assets included in the investment universe, we will need to choose and distribute among the various assets in the investment universe. Furthermore, optimal selling and buying points should be determined. In other words, even if a specific investment universe is formed, depending on portfolio compositions and trading points, we cannot guarantee positive returns all the time. However, if we choose the theme keyword appropriately, we may think that the stock price of the companies included in the constructed investment universe will increase in the long term. To check this long term return, we compared the 3-month historical return and the 3-month future return for each investment universe. An investment universe can be treated as a stock portfolio and the 3-month return for each investment universe are calculated in the same way as the return on a stock portfolio. We assume the equally weighted portfolios. The 3-month historical return means the return over 3-month before 2018-03-01, while the 3-month future return means the return over 3-month after 2018-03-01. This comparison result between historical return and future return of investment universes are shown in Fig. 6. In Fig. 6, X-axis and Y-axis represents 3-month historical returns and the 3-month future returns of the investment universes respectively. The future returns of the investment universes were sorted by their historical returns. Furthermore, a second order polynomial is fitted to these points by using the method of least squares to see the trend and the 3-month future return of the S&P500 index regardless of the historical returns. Additionally, when the historical returns of the investment universe are very low or very high, the future returns after 3-months tends to be high. We observed the comparison result between the 3-month historical returns and the 3-month future returns for every day during March of 2018 and this tendency was continued. Therefore, in this study, we propose a strategy grouping 15 companies with low historical returns and 15 companies with high historical returns together in a large investment universe. We then compare the average historical returns of the top and bottom investment universes with the S&P500 index and the average future returns of the 30 investment universes in the middle range of the historical returns in the investment universes.

Fig. 7 shows the results comparing the average future returns of all investment universes, top & bottom, and middle of the S&P500 index. As of March 2018, we set up the investment universes for every theme keyword, every day and examined their future returns after 3-months. The average future returns of total and middle 30 investment universes are shown in Fig. 8.
move almost along with the S&P500 index, and most of the average returns in the middle investment universes are lower than the average return of the total investment universe. On the other hand, the average future returns of the top 15 and bottom 15 investment universes proposed in this study are above the S&P500 index. If, according to this strategy, the portfolio is distributed evenly in the top 15 and bottom 15 investment universes, the returns after 3-months will be from 1% to 3% higher than the S&P500 index. This result does not mean that every stock price in these 30 investment universes moves up, but this strategy can offer a good set of stocks whose 3-month average future return is higher than S&P500 index.

![Graph showing 3M Future Returns](image)

**FIGURE 7.** Return after 3-months of selected investment universe and S&P500 on March.

### VI. CONCLUSION

If someone wants to invest in a company that is related to a particular field or topic that has recently attracted a public attention, the first action to take is to find companies whose business areas are related to that field or topic. These companies can be set as an investment universe. However, it is not easy to construct an investment universe properly, because investigating many companies in detail requires a lot of human efforts. Therefore, we have defined theme keywords as a set of words frequently used in the society in the recent years and extracted 500 theme keywords. After training Doc2vec model, 50 companies are culled first in the order of similarity for each theme keyword and separated into two clusters using k-means with \( k = 2 \). Among them, companies with higher similarities are set as the investment universe. The average correlation of the investment universe extracted from the 500 theme keywords is calculated and compared with the average correlation value of all companies in the S&P500. 80.8% of the total investment universes are higher than the average correlation of all companies in the S&P500. It can be interpreted as the theme keyword have a similar effect on the companies in the investment universe extracted by the keyword and the stock price of these companies moves similarly. This indirectly shows that the companies related to the theme keyword are included in the investment universes. In addition, we suggest strategies for extracting appropriate keywords using historical stock price returns to verify that building the investment universe using theme keywords could be profitable on investment.

First, we look at the historical average stock price returns of companies included in the investment universes and their average future returns. In this case, when investment universes are sorted based on historical returns, it can be seen that the investment universes, where historical returns are high or low, tend to have higher future returns than the investment universes with intermediate historical returns. This tendency is observed in most of the investment universes extracted in March 2018. We propose a strategy to select 15 investment universes with high or low historical returns, respectively. This result can be confirmed by comparing the future returns of these 30 investment universes with the future returns of the S&P500 index and 30 investment universes with intermediate historical returns.

This study has the following contributions. First, the methodology for extracting theme keywords was introduced. These theme keywords can cover many topics that have been highlighted in 2017. These keywords are useful for understanding market trends. Second, we propose a concrete methodology for organizing companies associated with the theme keywords into the investment universes. All the processes involved in constructing the investment universe is automated. Therefore, this methodology can reduce the human effort to investigate financial information and business areas of many companies. Third, we show that the methodology of constructing the investment universe can be used for offering a good set of profitable securities. It can lead to an investment strategy by choosing the appropriate theme keywords. Furthermore, keywords that are highly relevant to a particular company might provide a new intuition for investing. Fourth, the objectivity of investment universe construction is ensured. Investment universe construction is data driven method using the 10-K reports and the Reuters news articles. Therefore, it does not depend on personal intuition or experience.

This study can be applied to the following fields. First, the investment universe building process minimizes human effort. Thus, it can be used for low-cost asset management services, including robo advisor, provided for those who cannot afford private banking or wealth management services. Secondly, it can help individuals or institutional investors who want to invest in themselves to find investment targets easier and faster.

For the future works, we will improve the theme keyword extraction. Not only important words for investment but also some meaningless words appear frequently. Therefore, we plan to improve the purity of theme keyword pool by referring to the trend analysis services such as Google Trend [10] and Facebook Analytics [34]. We expect to be able to extract theme keywords for various periods, and verify the effect of trend periods. In this paper, cosine similarity is used to compute the matching score among the keywords and se-
curiencies. Various distance measures can be applied to enhance the matching results. For the investment universe selection based on the historical returns, machine learning approach can be used to select the universe likely to be profitable. Since we have historical return and future return data for constructing the investment universes, we can use supervised learning to improve the investment universe selection.

REFERENCES