International Journal of
Information
and
Management
Sciences

http://ijims.ms.tku.edu.tw/main.php

International Journal of
Information and Management Sciences

32 (2021), 121-140. DOI:10.6186/IJIMS.202106_32(2).0002

CODEN IIMSEQ
ISSN 1017-1819
DOI : 10.6186/IJIMS

International Journal
of
Information and Management Sciences

IJIMS *http://ijims.ms.tku.edu.tw*

Volume 32, Number 2 **June, 2021**

Information and Management Sciences
June, 2021 **Volume 32, Number 2**

A Hybrid Deep Learning Model for Predicting Stock Market Trend Prediction

Li-Chen Cheng, Wen-Shiu Lin and Yu-Hsin Lien
National Taipei University of Technology, Fu Jen Catholic University and Soochow University

Keywords

Deep learning
Word2Vec
Stock prediction
Text mining

Abstract.

In this work we propose a novel predictive model for improving investment capability that uses structured and unstructured data to predict stock price movements. We adopt deep learning techniques that have already been used successfully for natural language processing tasks, along with traditional data retrieval, to analyze and predict trends in the Taiwan stock market, and conduct experiments on both structured and unstructured data. Machine learning and data preprocessing techniques such as word2vec are used to train prediction models. Our experiments show that using deep learning on structured data yields improved accuracy, which attests the suitability of deep learning for structured data, especially for long short-term memory (LSTM) models. Finally, we combined structured and unstructured data using a combined approach to achieve improved accuracy with lower investment risks. The models in this work are thus suitable for real-world applications, including day trading strategy planning as well as long or short transaction strategy planning.

1. Introduction

Stock prices in the securities market reflect enterprise performance, and stock price trends can be analyzed and predicted through useful information such as market indexes, corporate financial statements, network news and financial forums. Investors can also determine price tends with technical analysis or text messages, but it is difficult to integrate the heterogeneous information collected from multiple sources for analysis.

Most traditional analyses of stock price trends are based on structured data, but sometimes the characteristics affecting stock prices are implied in unstructured data. Recent literature shows that, by using text mining to process such unstructured data, values implied in financial news can be successfully explored [12]. Ichinose and Shimada

[16] found that the development of network news can be used to predict stock market activity. They also found that the emotional characteristics of network news, and positive and negative emotions are correlated with stock price fluctuations to some extent [4, 37]. They found that financial news with more negative words usually predict lower company earnings [31]. Implied emotions will affect investor decisions, hence, the emotional state suggested in financial news is another important factor affecting stock prices [38].

The results of many studies show that the accuracy of stock trend prediction following the release of financial news rarely exceeds 58% [26]. Their exploration of the reasons for this poor prediction accuracy suggested that it could be because of the selection of scenarios or eigenvalues and the prediction technology. Some research used the traditional technology of k-nearest neighbors algorithm (k-NN) and artificial neural networks (ANNs) text mining and the other applied statistical ordinary least squares (OLS) regression analysis [27]. Bharathi et al. [3] employed sentiment analysis for stock market perdition. Vargas et al. [33] used multi-layer perceptron (MLP) neural networks for deep learning (DL) and word2vec neural networks to process word emotions, and finally obtained an accuracy above 60%.

The above studies also show that, in financial markets, the commodity prices or trends are predicted mostly based on single structured data (fundamental analysis or technical analysis) or solely based on emotion analysis of unstructured financial network news, considered in this study as one of the main reasons for poor prediction performance. A technology or method which can integrate unstructured and structured data will be important for improving the accuracy of stock price trends. In recent years, deep learning has been successfully applied for natural language processing (NLP) tasks and achieved good results [1]. Thus, deep learning techniques are used in this study along with traditional data retrieval to analyze and predict the trends of stocks that have been listed in the Taiwan 50 Index. In addition, various pre-processing methods will be applied to analyze the effects of unstructured data on stock trend prediction.

It can be difficult to select text mining and word segmentation systems. It is suggested here that it is a good strategy to compare word vectors in various models by using Bag-of-words (BoW), Latent Semantic Analysis (LSA) and Word2Vec after adopting the Jieba word segmentation system to segment words in the Chinese financial news. BoW is the most commonly used technology for analysis of word vectors in past studies, but some researchers have found that there are disadvantages to BoW such as too much noise and insufficient density. In order to solve these problems, in this study, Word2Vec will be used for article vectors. Word2Vec can convert words into vectors and project them into vector spaces to keep the words close in context to each other close in space. With the breakthrough development of deep learning, natural language processing has achieved better results than traditional statistical models. Therefore, in this study, support vector machine (SVM) and the latest deep learning technologies, including deep neural networks (DNN), LSTM and convolutional neural networks (CNN), are adopted to train technology models.

Based on the above discussion, there are three goals for this study: first, to explore whether financial news and other unstructured data have the potential to predict stock price trends. Therefore, in this study, the three word vector representations, BoW, LSA and Word2Vec, as well as various machine learning algorithms are adopted for evaluation. Next, to explore whether deep learning has the potential to predict stock price trends in structured data as compared to SVMs. Finally, some technologies to combine structured and unstructured data are proposed, such as DNN, LSTM and CNN, with the expectation of improving the stock price prediction performance.

In the experiment, information on 83 stocks listed on the FTSE TWSE Taiwan 50 Index between 2007 and 2017 is summarized (refer to Appendix 1 for the data). The technical indexes were sourced from the Taiwan Stock Exchange Corporation, including 13 pieces of daily closing price, ranging from January 3, 2017 to February 21, 2018. The news was sourced from Yahoo stock prices in Taiwan, ranging from July 17, 2017 to February 2, 2018. The Jieba word segmentation system was employed to segment the Chinese words in the financial news, while the BoW technique, LSA and Word2Vec were used for word vector processing. Finally, machine learning was used to train prediction models, with the expectation of building a security price trend prediction model based on deep learning capable of combining structured securities market and unstructured financial news data, which would make a contribution to the field of securities investment. In previous studies, Taiwan stock price data were applied for training and the stock prices were predicted through deep learning. However, analyzing Chinese news with deep learning is an innovative approach to predicting stock prices, as is combining structured and unstructured data for stock price prediction is also a first.

2. Literature Review

2.1. Text mining

Text mining refers to the exploration of implied information from un-structured or semi-structured texts and has been applied in many fields in recent years. Lee et al. [20] proposed automated systems to replace the traditional manual methods currently used to collect comments on various brands on the Internet and analyzed some technologies such as feature extraction, advice extraction and emotion analysis. Chen et al. [6] used text mining to judge and classify the meanings of news reports through SVM and Bayesian classification. Their results proved that text mining has significant effects on the interpretation of complex meanings.

Nassirtoussi et al. [25] suggested that analysis of public opinions of quality expressed online in social media and network news could be adopted to improve the predictability of financial markets, thus resulting in great benefits or losses. They achieved market prediction by using text mining and machine learning. Nassirtoussi et al. [26] divided the standard processes of text mining into 3 stages. The first stage is data collection and the second is data pre-processing. The information in texts is very important but incorrectly expressed inputs will lead to meaningless outputs, hence, how texts are pre-processed before machine learning is very important. This process can be classified into feature selection, dimensionality reduction and feature representation. The third stage is machine learning.

Past analyses of stock prices have mostly been long-term, with information considered as a lagging index, but the impact on stock prices in short-term investment has been neglected. Schumaker et al. [30] used text mining and emotion analysis on financial news. When using AZFinText, they obtained a directional accuracy of 59.0% to 50.4% and trading profits were 3.30% to 2.41%. Further emotion analysis showed that 53.5% positive emotions led to a fall in price while 52.4% negative emotions led to rise in price. They believed the reason was that the contrariness of traders, selling on good news and buying on bad news. Vu et al. [35] examined the relationship between consumer emotions and stock price fluctuations for economic analysis. They analyzed public emotions captured from Twitter which they adapted for prediction of the daily NASDAQ trends for Apple, Google, Microsoft and Amazon. Their decision-making tree (C4.5) method produced high prediction accuracies of 82.93% (Apple), 80.49% (Google), 75.61% (Microsoft) and 75.00% (Amazon).

Yu et al. [38] explored the effects of social media and traditional newspapers, TVs and magazine media on a company's short-term stocks. The social media included Twitter, Google Blogs and BoardReader, while the traditional media was Google News. They found that the overall emotions expressed in social media would have stronger effects on company stock performance than traditional media. They also found that social media and traditional media have strong interaction effects on stock performance. Ballings et al. [2] compared the performance of ensemble methods and single classifier models on stock price prediction. The ensemble methods included Random Forest, AdaBoost and Kernel Factory, and the single classifiers included ANN, Logistic Regression, SVM and KNN. The final experimental results showed that the Random Forest method performed the best.

2.2. Deep learning

Machine learning is used to recognize objects in images, convert speech into text, analyze users' preferences for products or articles and search for the most similar results, while deep learning has come to be more widely used for the above tasks in recent years [17]. Its birth has sped up the development of machine learning. The key lies in improvement in the technology of graphics processing units (GPUs), breaking through the bottleneck of hardware computing speed and greatly reducing the time needed for model training.

Deep learning can be traced back to the neural network models of McCulloch and Pitts [21]. Fukushima and Miyake [10] completed the Artificial Neural Network frame for deep learning and Rumelhart, Hinton, and Williams [29] applied backpropagation in succession. They obtained good results, which laid the foundation for the future deep learning.

At present, there are 4 main neural networks in deep learning, DNN, CNN, RNN (Recurrent Neural Network) and LSTM. DNNs or deep neural networks, are traditional shallow neural networks in which the numbers of hidden layers are increased between the input layers and output layers, and which can be used to build complex nonlinear relationships. CNNs feature shared weights and translation invariance [19]. The whole structure consists of an input layer, convolutional layer, pooling layer, full connection layer and output layer.

Williams and Zipser [36] proposed RNNs.All computations carry the memories obtained thus far (Memory) [5]. This makes RNNs perfect for sequential tasks so have been used in language models [22] and for speech recognition [11]. However, RNNs have been proven problematic, because of gradient exploding or gradient vanishing after long training of gradient backpropagation [14]. Hochreiter and Schmidhuber [15] proposed the LSTM to improve RNN with the addition of forget gates, input gates and output gates to provide clear memories.

Word representation or word embedding is a necessary process to transform natural language into language that can be understood by machine learning. The original concept is a simple one-hot representation with the length of the entire vector the same as that after the intersection of words. 1 is filled in for all words according to their lexicographically ordered vectors and then text vectors are formed correspondingly. However, there are two disadvantages to this method. First, large lexicons indicate large vector dimensions, which are difficult to handle for deep learning and storage. Secondly, all words are independent, the relationships between words cannot be shown.

We now discuss the application of deep learning to data preprocessing. Hinton [13] proposed the concept of distributed representation for training using feed-forward neural networks. The information for a word can be determined by multiple dimensions; in contrast, a dimension may support the information of multiple words. Such a representation can greatly reduce dimensional changes. The similarity between words can be computed by the cosine or Euclidean distance. Such word vector representations are also known as word embedding [6].

Word2Vec is an open source tool developed by Google engineers based on two studies, which exactly realizes the aforementioned concept of distributed representation [23, 24]. This tool includes two models: Continuous Bag-of-words (CBOW) and Skip-gram (SG). Vargas et al. [34] used Word2Vec to produce word vectors from news headlines for stock price prediction. The word vectors were averaged to obtain a unique vector for the whole headlines. Word2Vec has the advantages of capturing semantics and grammatical rules.

2.3. Applied deep learning in finance research

As the Internet matures, people are getting more and more information through multiple media sources such as mobile phones, computers, online newspapers and magazines. This means that investors can rapidly obtain more valuable and timely information about stock price news, market changes and industrial trends. Tumarkin and Whitelaw [32] found that the articles posted on financial forums can affect stock prices, suggesting that the news released on social media does affect investor determinations and does have an impact. However, on the other hand, too much information from multiple sources can confuse investor making valuable information difficult to identify. Some have used historical stock price trading data, financial news and back propagation neural networks to establish prediction models and to predict individual stock trends within days. Pagolu, Reddy, Panda, and Majhi [28] added emotion indexes and attention indexes to text mining. Their results showed that prediction accuracy could be improved. The prediction results were divided into rise, equality and fall.

Heaton, Polson, and Witte [12] mentioned that financial prediction (such as securities design and pricing, portfolio construction and risk management) usually involves complex and massive data structures which are difficult or even impossible to explain in detail in the current complete economic models. For these problems, deep learning can yield more useful results than the standard financial methods. CNN image recognition can produce outstanding classification results, and Fischer and Krauß [9] used this advantage to transform technical indexes and other information to pictures, to simulate the behavior of reading a tape with the human eye, as a substitute for the traditional structured analysis models of the past. The final selling and buying accuracy were 78.61% and 56.91%, respectively.

The trading strategy in Experiment 2 was the same as in Experiment 1, but the 5-day moving average (MA5) was used. The final selling and buying accuracy were 79.07% and 53.29% respectively. Finally, in Experiment 3, they considered the opening and closing prices and drew MA5, MA10 and MA20. The results showed that the accuracy in Experiment 3 was greatly improved and the selling and buying accuracy were 94.81% and 93.46% respectively.

Ding, Zhang, Liu, and Duan [7] used CNN to explore the effects of news events on short-term and long-term changes of stock prices, and the study results showed that more profits were made compared with previous systems. They proposed EB-CNN models and made predictions for S&P 500. The results demonstrate that convolutional neural networks can achieve the same or better classification than the traditional feedback neural networks.

Fischer and Krauß [9] argued that LSTM is very appropriate for financial applications. They used LSTM to predict different company rankings top $-k = 10$, 50, 100, 150 and 200. LSTM performed better than the random forest, deep neural network (DNN) and logistic regression methods. In previous studies, the stock price data for Taiwan were applied for training for prediction of the stock prices through deep learning.

As can be seen from the above studies, there are differences in prediction performance (accuracy or error function), differences in study implications, with values different from those obtained in this study. The main reasons for this are: (1) The study data are sourced from different countries or regions where political, economic and financial markets as well as efficient market hypothesis (EMH) are in different stages of development [8]. Hence, the effects and efficiency of the response transmission after information disclosure are different, thus resulting in different prediction results. (2) Whether the financial news selected for the study samples is from mainstream financial media, for example, Google news, Wall Street Journal (WSJ), Dow Jones News Service (DJNS), Yahoo finance, Twitter, Anue, ETtoday, etc., or not, has a big effect on the financial markets, leading to different study results. (3) Different prediction technologies were used in these studies and different study tools may cause different results. Differences caused by different prediction technologies of different ages are less significant.

In conclusion, based on a discussion of the above literature, we have recognized and learnt from their results. In this study, the words in articles from a variety of financial news media are processed and analyzed with a Jieba high performance system. Next, new deep learning technologies are adopted for prediction, including DNN, CNN and

LSTM, to improve and stabilize prediction performance. Ultimately, 13 structured stock eigenvalues and unstructured financial news data are combined to predict stock prices, quite different from the single data category used in previous studies. Prediction of stock prices based on the stock price data of Taiwan and Chinese financial news through the training of meaning analysis by deep learning is innovative and interesting.

3. The Proposed Framework

This study includes three aspects. First, we explore whether unstructured data have the potential to predict stock price trends. The three word vector representations, BoW, LSA and Word2Vec, as well as various machine learning methods are adopted for comparison. Next, we explore the potential of deep learning to predict stock price trends in structured data in comparison with support vector machines (SVMs). So far, stock price trend prediction has only been studied by using news articles or stock technical indexes, which is why we propose using a combination of structured and unstructured data for the prediction of stock prices.

The study framework consists of four major modules: a data collection module, a data pre-processing module, a machine learning module and an evaluation module. Data collection is conducted first, followed by the preprocessing of unstructured and structured data, respectively. Next is the application of a combination of different machine learning methods, and last is a comparison of the prediction effects among the different methods. Figure I shows the study framework.

3.1. Data collection and data preprocessing

In this study, crawler programs were used to capture financial news and daily closing quotation information for storage in the database for subsequent experiments.

Data preprocessing is a very important part of machine learning, and the quality of the data sets will determine the effects of prediction models. Jieba was used in this study, because it provides a customized dictionary which allowed us to add vocabularies related to finance for improvement of the word segmentation effects.

The financial news naturally includes foreign companies and stock price indexes. In this study, in order to preserve the original state of the news, Chinese stop words, newspaper offices and journalists were removed. To build the dictionary, the stocks with the most news of all types were first selected, then the word frequency was calculated with the results of word segmentation, after that the first 1,000 nouns, verbs and adjectives were selected and after arrangement in descending powers, proper nouns, verbs and adjectives related to stocks were picked out by voting by 3 experts, to ensure that the special nouns for all types of stocks were captured. Table 1 shows the collection of the first 30 financial dictionaries.

For the purpose of machine learning, due to the large values contained within, structured information should be first normalized according to the maximum and minimum values of all data of all stocks and redistributed between [0, 1]. If the maximum and minimum values of all data for all stocks are adopted, the extreme values may affect the

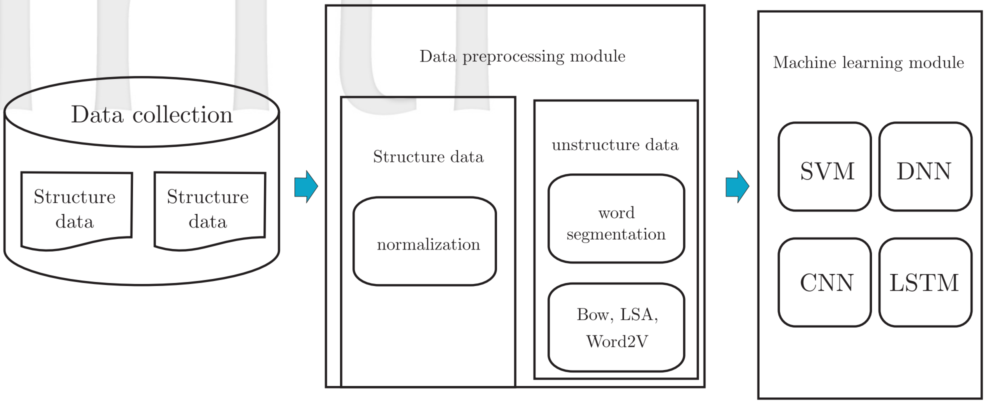

Figure 1: The proposed framework.

Table 1: Part of the dictionary.

Nouns	TSMC, UMC, WEC, etc.
Verbs and adjectives	Panic (恐慌), Explosion (爆量), Increase (揚升), net loss (淨損), Liquidation (清倉), Bear Market(熊市)···

training models. So their own maximum and minimum values can reflect the correct values for proper training.

3.1.1. Vector representation of the news

Traditional BoW and LSA as well as distributed representation based Word2Vec are used. In BoW, the bag-of-words vectors are built through consideration of the intersection of the words in all articles. Words with a bag of words frequency of less than 1 are filtered out, The bag-of-words vectors were matched with the articles to form the vectors of all articles. In this study, the first 250 words with the highest frequency were selected by BoW.

LSA is a technique which can convert documents into vector representations, and the vectors generated can then be used for comparison of document similarity, article classification and other tasks. First, term frequency inverse document frequency (TF-IDF) was used to transform all documents to matrices. Then, 10,000 words were selected for the training of TF-IDF. Singular value decomposition (SVD) was used to decompose the matrices, and finally, all pieces of news were reduced to 250 dimensions.

In older methods, individual news articles are treated as overall units, but Word2Vec convert all individual words directly into vectors. If BoW is applied, it might produce a lot of invalid information in the vector spaces. The problem with this invalid information is that it can generate many 0s. In Word Embedding, information is produced in context. As mentioned above, two models, CBOW and SG, are used. Therefore, even if the words to be predicted are not encountered in the training models, there is still a way to predict them as long as the context is the same. As a consequence, SG was used for word vector training in this study. Words with a frequency below 5 were replaced by “UNK”, generating 117,163 words and 50,724 words for word vector training, with all words being set at 250 dimensions.

3.1.2. Machine learning module

In this study, SVM, DNN, CNN and LSTM are adopted as the machine learning models to describe the structures and parameter settings of all models. The training sets were based on the data from July 17, 2017 to December 31, 2017, and the test sets were based on the data from January 2, 2018 to February 2, 2018.

SVM is a classification algorithm integrating linearity and nonlinearity. It projects the original data into a high-dimensional space, and seeks to separate the hyperplanes in classification problems, finding the optimal hyperplanes to solve classification problems. Among them, optimal hyperplanes are used to find the maximum margins of the support vectors in different categories.

Therefore, for the unstructured data in movements, in BoW and LSA, the data are directly input in fixed dimensions, while in Word2Vec, the sum of the text vectors is divided by the number of words to represent the article vectors, and then similarly, the data are directly input in fixed dimensions. In addition, the structured data are input in a fixed dimension of 13.

In DNN, the unstructured and structured data are input in the same format as in SVM. DNN has a total of 15 layers, including 1 input layer, 1 output layer, 6 full connection layers and 7 dropout layers. In addition, Relu is used as the activation function and AdamOptimizer is used as the optimizer.

Word2Vec is different from BoW and LSA and is used for processing unstructured data. The first 350 words of an article are taken and the original 250-dimensional vector is maintained to form a 2-dimensional article vector, so the input format is [None, 350, 250]. Structured data are processed the same as in the previous two methods.

CNN has a total of 11 layers, including 1 input layer, 1 output layer, 2 convolutional layers, 2 pooling layers, 1 flatten layer, 2 full connection layers and 2 output layers. Relu is used as the activation function, Adam Optimizer is used as the optimizer and VALID is used for padding.

Dynamic RNN is adopted with LSTM model used in this study, with padding conducted based on the maximum news length in each batch, allowing each input length to be different. In the LSTM experiments, only Word2Vec uses non-fixed lengths with the input format of [None, None, 250], while BoW and LSA use fixed lengths. In addition, fixed-dimensional input is adopted for the structured data for LSTM.

In the designed network structure, LSTM has a total of 3 layers, including 1 input layer, 1 output layer and 1 LSTM Cell layer. In addition, Relu is used as the activation function and AdamOptimizer is used as the optimizer.

3.1.3. Evaluation module

The proposed methods and the validity of other different combinations were evaluated; therefore, 6 indexes were used, including accuracy, Matthew's correlation coefficient (MCC), precision, sensitivity, specificity and training time (see Table 2).

Accuracy is determined by the proportion of TP and TN correctly classified, and higher accuracy more effective classification. However, the reliability of a classifier cannot only be represented by the accuracy, so here the Matthew's correlation coefficient is used to evaluate the accuracy of a classifier. If all correct predictions have a value of 1 and incorrect ones a value of -1, the value for a random guess is 0.

The method developed in this study is intended to predict stock price trends, so sensitive and specificity are included in the evaluation. Sensitivity is used to determine the accuracy of a classifier in predicting a "rise" and specificity is used to determine the accuracy of a classifier in predicting a "fall". However, not all trends can be classified 100% correctly all the time, and sometimes situations where the actual stock prices rise while the predicted ones fall or the actual stock prices fall and the predicted ones rise may happen. Under these two conditions, the situation have the greatest effect on investors is a false positive (FP). The situation where a fall has occurred but a rise was predicted could cause losses to investors and is the most influential error for investors compared a false negative (FN). Hence, precision is incorporated in the experiment to determine the proportion of FP from the classifier.

The last step is to evaluate the training time (efficiency) of all models. When the prediction effectiveness is the same, the training time is an evaluation index that is important in practical applications.

Table 2: Evaluation indexes.

Evaluation index	Formulas
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Matthew's correlation coefficient	$\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
Precision	$\frac{TP}{TP+FP}$
Sensitivity	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
Training time	Second

Table 3 shows the model combinations examined in the experiments. BoW, LSA and Word2Vec are used for preprocessing and the classification methods used are SVM, DNN, CNN and LSTM.

Table 3: Combination of models.

Data preprocessing	Classification methods
BoW	SVM
	DNN
	CNN
	LSTM
LSA	SVM
	DNN
	CNN
	LSTM
Word2Vec	SVM
	DNN
	CNN
	LSTM

4. Experimental Data

Crawler programs were used to collect structured and unstructured data. The unstructured data were sourced from Yahoo stock prices in Taiwan, ranging from July 17, 2017 to February 2, 2018, including data from 7 new sites, for a final total of 24,716 unstructured data. In addition, structured data are sourced from the Taiwan Stock Exchange Corporation, with a total of 13 pieces of daily closing quotation information, including the number of stocks traded, the number of trades, trading amount and whether prices rise or fall. The data range before intersection of news is from January 3, 2017 to February 21, 2018, for a final total of 250,738 structured data.

For correct classification of the stock price trends as affected by the news, news released after 13:31 was classified as on the next day and weekend news was classified to the following Monday . Finally, the rise and fall of prices were marked by the intersection of the TWSE database 21,526 unstructured data were processed. The structured and unstructured data were combined to find the intersect based on the news, so there will not be a day without news. Before training, the structured data must be normalized according to the maximum and minimum values of all data for all stocks. If the maximum and minimum values of all data for all stocks were adopted, the extreme values might affect the training models and thus produce errors. There were a total of 23,014 structured data in the end.

The full name of the Taiwan 50 Index is the "FTSE TWSE Taiwan 50 Index" (Taiwan 50), for the top 50 companies in Taiwan by market value and stability. The list will changes from quarter by quarter, therefore, in this study, data for 83 stocks that have been listed in the Taiwan 50 Index over the ten years are summarized as the basis for selection.

This experiment was carried out on a Ubuntu 16.04.4 LTS operating system, with an Intel i7-7700x CPU, a memory of 64GB, and a Nvidia GTX1080Ti 11GB as the GPU with Tensorflow and Tensorlayer as the development tools. In the first stage of the experiment, the words in the news articles were segmented, a process conducted in stages. The Jieba word segmentation system, with the advantage of its customized dictionary was adopted. The customized dictionary allowed us to add stock names, stock codes and Chinese, to make sure the basic word segmentation is correct and to maintain the Chinese characters only. In the second stage, the stocks with the most news of all types were selected, and then the first 1,000 nouns, verbs and adjectives selected from the word segmentation results were submitted to 3 experts for voting, to ensure that the special words for all types of stocks could be captured. Finally, the expert voting results were added to the customized dictionary for further word segmentation.

After word segmentation, Tensorlayer was used for word vector training by Word2Vec, and restricted words with a frequency below 5 were uniformly replaced by "UNK". Ultimately, 50,724 words were generated for word vector training, with all words being set at 250 dimensions. Before model training, for algorithms such as SVM and DNN, the classifiers for the input dimension must be fixed. The sum of the vectors of all words in all pieces of news were first obtained and then divided by the number of words in all pieces of news to get the article vectors.

The dynamic RNN is adopted for the LSTM in this experiment, and padding was conducted based on the maximum news length in each batch, so each input length could be different. The training sets of all models were based on the data from July 17, 2017 to December 31, 2017, and the test sets were based on the data from January 2, 2018 to February 2, 2018. The final experimental results were evaluated for accuracy, MCC, precision, sensitivity, specificity and comparison of model training time.

5. Experimental Results and Analysis

The main purpose of this study was to explore whether financial news and other unstructured data have the potential to predict stock price trends and whether deep learning can be used to predict stock price trends from structured data. Various learning algorithms were used for analysis and evaluation. Finally, some techniques for combining structured and unstructured data were proposed, such as DNN, LSTM and CNN, with the expectation of improving stock price prediction. The following experiments were designed for evaluation and verification: Experiment 1: to predict stock price trends based on unstructured data; Experiment 2: to predict stock price trends based on structured data; Experiment 3: to predict stock price trends by combining unstructured and structured data. The experiment results are discussed in more detail below.

Experiment 1: Predicting stock price trends based on unstructured data

BOW, LSA and Word2Vec were adopted for word vector representation, and SVM, DNN, Dynamic RNN and CNN were used as the classifiers. The accuracy of all combinations was verified through t-testing and the results are summarized in Table 4.

Table 4: The performance of the modules.

Preprocessing		Classification methods	Average	T value	p value
BOW	1	SVM	0.528476	174.388	2.2e-16***
		DNN	0.51425033		
	2	SVM	0.528476	-43.231	2.2e-16***
		CNN	0.6649803		
	3	SVM	0.528476	2.769	0.006713***
		LSTM	0.51586		
	4	DNN	0.51425033	-47.72	2.2e-16***
		CNN	0.6649803		
	5	DNN	0.51425033	-0.353	0.7247
		LSTM	0.51586		
	6	CNN	0.6649803	26.901	2.2e-16***
		LSTM	0.51586		
LSA	1	SVM	0.562285	-11.728	2.2e-16***
		DNN	0.67086281		
	2	SVM	0.562285	-11.309	2.2e-16***
		CNN	0.6125779		
	3	SVM	0.562285	22.06	2.2e-16***
		LSTM	0.5161134		
	4	DNN	0.67086281	5.675	7.502e-8***
		CNN	0.6125779		
	5	DNN	0.67086281	16.303	2.2e-16***
		LSTM	0.5161134		
	6	CNN	0.6125779	19.626	2.2e-16***
		LSTM	0.5161134		
W2V	1	SVM	0.562285	1028.811	2.2e-16***
		DNN	0.51437634		
	2	SVM	0.562285	-32.98	2.2e-16***
		CNN	0.8471092		
	3	SVM	0.562285	-67.452	2.2e-16***
		LSTM	0.9835159		
	4	DNN	0.51437634	-38.527	2.2e-16***
		CNN	0.8471092		
	5	DNN	0.51437634	-75.121	2.2e-16***
		LSTM	0.9835159		
	6	CNN	0.8471092	-12.799	2.2e-16***
		LSTM	0.9835159		

The results of Experiment 1 (Table 4) show the different pre-processing methods with different classifiers. On average, the accuracy of all combinations is above 0.52. When BOW is used for pre-processing and CNN is used as the classifier the accuracy is higher than with SVM and the other classifiers. When LSA is used for pre-processing, the accuracy is higher when DNN is used as the classifier than for CNN and the other classifiers. When Word2Vec is used for preprocessing and LSTM is used as the classifier the accuracy is higher than with CNN and the other classifiers.

Compared with the sentence structures in ordinary chats, on forums and social media, news articles contain regular and structured expressions, so that it is hard to see the trends with traditional BoW, LSA and Word2Vec. The training in this study is based on the news about the Taiwan 50 Index, but the stocks listed on this Index are blue chip stocks which are stable and tepid with small fluctuations, so it is not easy for the classifier to make accurate predictions.

However, the accuracy of Word2Vec+LSTM is 0.98, which indicates the algorithm has a high ability to predict a rise and can reduce the risk of investor losses, showing effective classification rise and fall classification. Therefore, compared with other preprocessing methods, in Word2Vec, word meanings can be expressed better, and the features are reduced through convolution to improve the classification effect. In other words, Experiment 1 shows that Word2Vec+LSTM has good predictive ability from stock price news.

Experiment 2: Predicting stock price trends based on structured data

There are a total of 13 structured data, as shown in Table 5. In Experiment 2, the structured data were combined to predict trends; the results are summarized in Table 6.

The study results show that the accuracy of LSTM is 0.81, and the t-test proves that LSTM is very effective at predicting stock price trends. It can be seen that machine learning is effective in making predictions from structured data, which is consistent with the conclusions of [9]. LSTM is better than DNN in term of effectiveness for only slightly, and the same problem as in Experiment 1 may exist, which is the relative stability of the Taiwan 50 Index stocks, meaning that different classifiers may demonstrate similar performance under such conditions. In addition, the amount of data in this experiment is insufficient, and under the conditions for similar results in Experiment 2, SVM has a great advantage terms of in training time. According to the results of Experiment 2, the unstructured data are combined with and LSTM is added behind the convolution layers.

Experiment 3: Prediction of stock price trends by combining unstructured and structured data

Experiment 1 and Experiment 2 prove that unstructured and structured data are all predictive. Therefore, in Experiment 3 both are combined with the expectation of improving predictive ability. Experiment 1 shows that Word2Vec+CNN produce the best effect and Experiment 2 shows that LSTM has the best performance. Hence, we propose combining unstructured and structured data and using CNN+LSTM to predict the stock prices.

The method used in Experiment 3 has an accuracy of 0.80 and an MCC of 0.60. A precision of 0.82 represents less investment risk and a specificity of 0.81 represents more

Table 5: The Daily closing market information.

	index
1	Number of Transactions
2	Trading Volume
3	Turn Over in value
4	Opening price
5	Highest price
6	Lowest price
7	Close
8	P/E ratio
9	Ups and downs spread
10	Last Best Bid Volume
11	Last Best Bid Price
12	Last Best Ask Price
13	Last Best Ask Volume

Table 6: Results of a comparison of classification models.

	Classification	Average	Tvalue	P value
1	SVM	0.725701	20.144	2.2e-16***
	DNN	0.6227718		
2	SVM	0.725701	-26.066	2.2e-16***
	CNN	0.8445315		
3	SVM	0.725701	-51.471	2.2e-16***
	LSTM	0.8130471		
4	DNN	0.6227718	-32.384	2.2e-16***
	CNN	0.8445315		
5	DNN	0.6227718	-35.34	2.2e-16***
	LSTM	0.8130471		
6	CNN	0.8445315	6.472	1.949e-09***
	LSTM	0.8130471		

effective in fall prediction than in Experiment 1 or Experiment 2 (see Table 5). Therefore, we propose using a combination of structured and unstructured data to improve the accuracy in trend prediction. The results support this proposal. However, due to the limited amount of data used in this study, the results of the three experiments were similar, although we believe that sufficient data could improve our models' predictive ability.

Table 7: Results of a comparison of classification models.

	Classification	Average	Tvalue	P value
1	SVM	0.725701	20.144	2.2e-16***
	DNN	0.6227718		
2	SVM	0.725701	-26.066	2.2e-16***
	CNN	0.8445315		
3	SVM	0.725701	-51.471	2.2e-16***
	LSTM	0.8130471		
4	DNN	0.6227718	-32.384	2.2e-16***
	CNN	0.8445315		
5	DNN	0.6227718	-35.34	2.2e-16***
	LSTM	0.8130471		
6	CNN	0.8445315	6.472	1.949e-09***
	LSTM	0.8130471		

Table 8: The comparison of different models.

preprocessing	index	models	Accuracy	MCC	Precision	Sensitive	Specificity	Training time
W2V	Daily closing market information	CNN + LSTM	0.80	0.60	0.82	0.80	0.81	558s
W2V	-	CNN	0.78	0.56	0.75	0.82	0.73	353s
-	Daily closing market information	LSTM	0.77	0.53	0.77	0.76	0.78	101s

6. Conclusion and Future Studies

Based on the complexity of stock price prediction and to overcome the limitations of past relevant studies, we proposed combining unstructured financial news data with structured stock trading data and applied various kinds of deep learning techniques to build the stock price prediction models. Additionally, in this study, different methods of preprocessing (BOW, LSA and Word2Vec) were applied to the unstructured data, with the expectation of improving the effects on stock price trend prediction, so as form an important basis of transaction strategy planning for investors in real-world applications.

The results of the three experiments show that the "pre-processing" method for unstructured data makes no obvious difference in stock trend prediction. The reasons for this may be that the structure of the news articles is regular and fixed and that the stocks listed in the Taiwan 50 Index are blue chip stocks which have more information disclosures and attentions and are stable with small fluctuations, so news has less exceptional stimulation effects on them, reflecting no obvious rise and fall. However, deep learning models - in particular Word2vec+CNN (with the accuracy of 78%) - generally outperform the traditional SVM, which suggests that unstructured data for financial prediction tasks is useful. Next, in Experiment 2, it is found that deep learning in the structured data of 13 stock eigenvalues has a good accuracy (between 73% and 77%), proving that deep learning is effective in structured data prediction and has the best performance (with the accuracy of 77%).

Finally, in Experiment 3, we combined structured and unstructured data and used CNN+LSTM algorithm which has good performance in the above experiments, proving its improved accuracy (80%). Additionally, the specificity of 81% represents its ability in predicting real fall in stock prices, to avoid the risk of price fall. The high precision of 82% represents the ability of CNN+LSTM in predicting real rise in stock prices, to assist investors in trading strategy planning, so as to increase the investment returns.

The experimental results show that the contributions of this study lies in clarifying the blind spots in structured and unstructured data processing and verifying that the combination of both will highlight the importance of behavioral finance and public opinion analysis in financial markets. In other words, structured data are added on this basis, which greatly improves the predictive ability. Therefore, the implication in trading applications is that investors can use CNN+LSTM prediction models and combine daily structured data with unstructured data to build a stock price prediction system of Taiwan. The accuracy, specificity and precision of the models in this study in stock price prediction are above 80%, showing its ability in predicting real rise and fall in stock prices. If adopting one-way swing-trading strategy, the traders go long and buy when the system predicts a rise in stock prices and sell their stocks when it predicts a fall, but stop trading when it predicts a fall. If adopting two-way swing-trading strategy, the traders enter long positions and exit short positions (security buy-back) when the system predicts a rise in stock prices, and exit long positions and enter short positions (short selling) when it predicts a fall.

In conclusion, the experiments in this study show that the academic contribution of stock price prediction based on deep learning can, in practice, assist investors in designing one-way or two-way swing-trading strategy, to improve the investment returns and reduce risks.

The main limitation of this study is insufficient data. The data range is from 2017 to 2018 and can be extended to the financial tsunami in the future, and the stocks not listed in Taiwan 50 Index can be additionally included. Other features of structured data can be added, such as KD, MA and other technical indexes, for more abilities of deep learning can be exerted with more data features. The "time lag property" always be considered in building the proposed model in traditional linear approach. Our proposed model was focused on the combination structured and unstructured data. The correlation between news and stock price also should be consider "time lag property" in the future research.

This is an exploratory study with the contribution of discussing the effects of different word vector representations combined with different classifiers. Recently, there were several novel techniques proposed in NLP area such as Fasttext, and BERT and so on. The sentiment analysis should be applied in the stock price perdition model in the future. Finally, Taiwan stocks are easily affected by American stocks, European stocks and foreign capital, so their daily rise and fall data can be added. Furthermore, the Institutional Investors in Taiwan have great effects with their strength of buying and selling, adding their data as the features can improve the effects in stock price trend prediction.

Acknowledgements

This study was supported in part by the Ministry of Science and Technology of Taiwan under grant numbers MOST 105-2410-H-031 -035 -MY3and MOST 107-2218-E-007-045.

References

[1] Akita, R., Yoshihara, A., Matsubara, T. and Uehara, K. (2016). *Deep learning for stock prediction using numerical and textual information*, Paper presented at the Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on.

[2] Ballings, M., Van den Poel, D., Hespeels, N. and Gryp, R. (2015). *Evaluating multiple classifiers for stock price direction prediction*, Expert Systems with Applications, Vol.42, No.20, 7046-7056.

[3] Bharathi, S., Geetha, A. J. I. J. o. I. E. and Systems (2017). *Sentiment analysis for effective stock market prediction*, Vol.10, No.3, 146-154.

[4] Bollen, J., Mao, H. and Zeng, X. J. J. o. c. s. (2011). *Twitter mood predicts the stock market*, Vol.2, No.1, 1-8.

[5] Britz, D. (2015). *Recurrent Neural Networks Tutorial, Part 1 - Introduction to RNNs.* Retrieved from http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/

[6] Chen, S., Chen, K.-Y., Hung, H. and Chen, B. (2015). *Exploring Word Embedding and Concept Information for Language Model Adaptation in Mandarin Large Vocabulary Continuous Speech Recognition*, Paper presented at the Proceedings of the 27th Conference on Computational Linguistics and Speech Processing (ROCLING 2015).

[7] Ding, X., Zhang, Y., Liu, T. and Duan, J. (2015). *Deep Learning for Event-Driven Stock Prediction*, Paper presented at the Ijcai.

[8] Fama, E. F. J. T. j. o. f. (1991). *Efficient capital markets*: II, Vol.46, No.5, 1575-1617.

[9] Fischer, T. and Krauß, C. (2017). *Deep learning with long short-term memory networks for financial market predictions*, European Journal of Operational Research Vol.270, No.2, 654-669.

[10] Fukushima, K. and Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition, In *Competition and cooperation in neural nets* (pp.267-285): Springer.

[11] Graves, A. and Schmidhuber, J. (2005). *Framewise phoneme classification with bidirectional LSTM and other neural network architectures*, Neural Networks, Vol.18, No.5, 602-610.

[12] Heaton, J., Polson, N. and Witte, J. (2016). *Deep learning in finance*, arXiv preprint arXiv:1602.06561.

[13] Hinton, G. E. (1986). *Learning distributed representations of concepts*, Paper presented at the Proceedings of the eighth annual conference of the cognitive science society.

[14] Hochreiter, S. (1991). Untersuchungen zu dynamischen neuronalen Netzen. Diploma, Technische Universität München, 91.

[15] Hochreiter, S. and Schmidhuber, J. (1997). *Long short-term memory*, Neural computation, Vol.9, No.8, 1735-1780.

[16] Ichinose, K. and Shimada, K. (2016). *Stock market prediction from news on the Web and a new evaluation approach in trading*, Paper presented at the 2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI).

[17] LeCun, Y., Bengio, Y. and Hinton, G. (2015). *Deep learning.* Nature, Vol.521, No.7553, 436-444.

[18] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. and Jackel, L. D. (1989). *Backpropagation applied to handwritten zip code recognition*, Neural computation, Vol.1, No.4, 541-551.

[19] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998). *Gradient-based learning applied to document recognition*, Proceedings of the IEEE, Vol.86, No.11, 2278-2324.

[20] Lee, S.-H. (2014). A Study of Using Text Mining Techniques to Word of Mouth Analysis. In.

[21] McCulloch, W. S. and Pitts, W. (1943). *A logical calculus of the ideas immanent in nervous activity*, The Bulletin of mathematical biophysics, Vol.5, No.4, 115-133.

[22] Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). *Efficient estimation of word representations in vector space*, arXiv preprint arXiv:1301.3781.

[23] Mikolov, T., Karafiát, M., Burget, L., Cernocký, J. and Khudanpur, S. (2010). *Recurrent neural network based language model*, Paper presented at the Interspeech.

[24] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013). *Distributed representations of words and phrases and their compositionality.* Paper presented at the Advances in neural information processing systems.

[25] Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y. and Ngo, D. C. L. (2014a). *Text mining for market prediction: A systematic review*, Expert Systems with Applications, Vol.41, No.16, 7653-7670.

[26] Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y. and Ngo, D. C. L. (2015). *Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment*, Expert Systems with Applications, Vol.42, No.1, 306-324.

[27] Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y. and Ngo, D. C. L. J. E. S. W. A. (2014b). *Text mining for market prediction: A systematic review*, Expert Systems with Applications, Vol.41, No.16, 7653-7670.

[28] Pagolu, V. S., Reddy, K. N., Panda, G. and Majhi, B. (2016). *Sentiment analysis of Twitter data for predicting stock market movements*, Paper presented at the 2016 international conference on signal processing, communication, power and embedded system (SCOPES).

[29] Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986). *Learning representations by back-propagating errors*, Nature, 323(6088), 533.

[30] Schumaker, R. P., Zhang, Y., Huang, C.-N. and Chen, H. (2012). *Evaluating sentiment in financial news articles*, Decision Support Systems, Vol.53, No.3, 458-464.

[31] Tetlock, P. C., Saar Tsechansky, M. and Macskassy, S. J. T. J. o. F. (2008). *More than words: Quantifying language to measure firms' fundamentals*, The Journal of Finance, Vol.63, No.3, 1437-1467.

[32] Tumarkin, R. and Whitelaw, R. F. (2001). *News or noise? Internet postings and stock prices*, Financial Analysts Journal, Vol.57, No.3, 41-51.

[33] Vargas, M. R., De Lima, B. S. and Evsukoff, A. G. (2017a). *Deep learning for stock market prediction from financial news articles*, Paper presented at the 2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA).

[34] Vargas, M. R., de Lima, B. S. and Evsukoff, A. G. (2017b). *Deep learning for stock market prediction from financial news articles*, Paper presented at the Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), 2017 IEEE International Conference on.

[35] Vu, T.-T., Chang, S., Ha, Q. T. and Collier, N. (2012). *An experiment in integrating sentiment features for tech stock prediction in twitter*. Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data, 23-38.

[36] Williams, R. J. and Zipser, D. (1989). *A learning algorithm for continually running fully recurrent neural networks*, Neural Computation, Vol.1, No.2, 270-280.

[37] Yan, D., Zhou, G., Zhao, X., Tian, Y. and Yang, F. J. C. C. (2016). *Predicting stock using microblog moods*, China Communications, Vol.13, No.8, 244-257.

[38] Yu, Y., Duan, W. and Cao, Q. (2013). *The impact of social and conventional media on firm equity value: A sentiment analysis approach*, Decision Support Systems, Vol.55, No.4, 919-926.

Department of Information and Finance Management, National Taipei University of Technology, Taipei, Taiwan, ROC.

E-mail: lijen.cheng@gmail.com
Major area(s): Text mining, machine learning, FinTech.

Department of Information Management, Fu Jen Catholic University, Taipei, Taiwan, ROC.

E-mail: wslin1949@gmail.com
Major area(s): Machine learning, FinTech, program trading system.

Department of Computer Science and Information Management, Soochow University.

E-mail: master05356015@gmail.com
Major area(s): Text mining, Machine learning, deep learning

(Received November 2020; accepted May 2021)