11-2021

Stock return prediction using financial news: A unified sequence model based on hierarchical attention and long-short term memory networks

Haoling CHEN

Peng LIU
Singapore Management University, liupeng@smu.edu.sg

2021 International Conference on Signal Processing and Machine Learning (CONF-SPML) | 978-1-6654-1734-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/CONF-SPML54095.2021.00034

Stock Return Prediction using Financial News: A Unified Sequence Model based on Hierarchical Attention and Long-Short Term Memory Networks

Haoling Chen
Department of Mathematics
King's College London
London, UK
haoling.chen@kcl.ac.uk

Peng Liu
Department of Statistics and Data Science
National University of Singapore
Singapore
liu.peng@u.nus.edu

***Abstract*—Stock return prediction has been a hot topic in both research and industry given its potential for large financial gain. The return signal, apart from its inherent volatility and complexity, is often accompanied by a multitude of noises, such as other stocks' performance, macroeconomic factors and financial news, etc. To better characterize these factors, we propose a new model that consists of two levels of sequence: an NLP-based module to capture the sequential nature of words and sentences in the financial news, and a time-series-based module to exploit the sequential nature of adjacent observations in the stock price. In this proposed framework, we employ Hierarchical Attention Networks (HAN) in the text mining module, which could effectively model the financial news and extract important signals at both word and sentence level. For the time series module, the established Long-Short Term Memory (LSTM) network is used to model the complex serial dependence in the time series data. We compare with benchmark models using either module alone, as well as other alternatives using the traditional Bag of Words (BOW) approach, based on the Dow Jones Industrial Average (DJIA) dataset. Experiment results show that our proposal method performs better in several classification metrics for both positive and negative stock returns.**

Keywords— stock price prediction, text classification, natural language processing, hierarchical attention networks (HAN), long short-term memory (LSTM)

I. Introduction

In the current financial market, regular news reports such as corporate release and public announcements influence the trending of stock prices in a volatile and unpredictable manner. An accurate forecast on future stock movements is thus essential for investors to formulate optimal investment strategies that maximise profits while minimising potential risks. However, due to its inherent inconsistency, it is quite challenging for researchers and industry practitioners to come up with an accurate prediction.

Early research on this front is established upon different time series models using historical stock prices. Vishwanath et al. (2013) proposed a method called Approximation and Prediction of Stock Time-series data (APST) using similarity of Pattern Sequence [1]. It performed data approximation by Multilevel Segment Mean (MSM) and prediction using Euclidean distance and Nearest-neighbour. Ariyo et al. (2014) presented a stock price prediction model using autoregressive integrated moving average (ARIMA) model [2]. Angadi and Kulkarni (2015) revealed that the ARIMA model has a strong potential for short-term prediction of stock market trends [3]. Meanwhile, as textual information such as Yahoo Finance and Bloomberg becomes more readily available, it is important to analyse the financial news and announcements from relevant industries or companies. Developing complex machine learning and deep learning models, which show promising performances in capturing structural components in the volatile financial data, has been gaining popularity among researchers and industry practitioners.

In recent years, a significant contribution has been made in the development and application of text mining methods in the finance sector, analysing market behaviour and predicting the future fluctuations for a specific or the industry average stock price. Khedr et al. (2017) and Kalra and Prasad (2019) proposed a news sentiment analysis method that uses naïve Bayes algorithm to get the text polarity and improve prediction accuracy [4]-[5]. Derakhshan et al. (2019) introduced a part-of-speech graphical model to process opinionated text, which achieved a higher prediction accuracy than using explicit sentiment labels for comments [6]. Elagamy et al. (2018) explored how text mining combined with the Random Forest algorithm could be used to extract critical indicators and classify financial news [7]. Pagolu et al. (2016) indicated that a strong correlation exists between the rise and fall of stock prices and the public sentiments in tweets [8]. Batra and Daudpota (2018) performed sentiment analysis on tweets to predict the mood of people, which has an impact on stock prices [9]. Islam et al. (2018) published a literature review about text mining techniques and methods such as principle component analysis, for stock market prediction [10]. Nabipour et al. (2020) highlighted that for the continuous data, Recurrent Neural Network (RNN) and Long short-term memory (LSTM) networks outperform other prediction models by a considerable margin when compared with nine machine learning models including Decision Tree, Random Forest, Adaptive Boosting (Adaboost), eXtreme Gradient Boosting (XGBoost), Support Vector Classifier (SVC), Naïve Bayes, K-Nearest Neighbors (KNN), Logistic Regression, Artificial Neural Network (ANN)), RNN and LSTM [11].

Compared with traditional approaches, attention-based model using word embedding has become a popular and powerful alternative in recent years. Gao et al. (2018) proposed an accurate and fast approach to train Hierarchical Convolutional Attention Networks model, which combines self-attention mechanism and convolutional filters using a hierarchical structure to support document classification [13]. Pappas and Popescu-Belis (2021) explored multilingual hierarchical attention networks for learning document

Some contents of this paper have been uploaded to King's College London as schoolwork for 2021 King's Experience Research Award.

978-1-6654-1734-1/21/$31.00 ©2021 IEEE
DOI 10.1109/CONF-SPML54095.2021.00034

structures, with shared encoders or attention mechanisms across languages [14]. Huang et al. (2020) divided their stock prediction model into article attention network and time series attention network, resulting in a higher accuracy through stock encoding [15]. Liu et al. (2018) investigated a hierarchical complementary attention network (HCAN) to capture valuable complementary information in news titles and contents, in addition to a new measurement for attention weights [16].

This research is an attempt to build an effective model that reflects how daily news impact the industry average stock movements. We use both financial news and historical stock prices to build sequence models for stock return prediction, based on advanced text mining and forecasting methods such as Hierarchical Attention Networks and LSTM. The proposed model has two architectural highlights. First, we adopt Hierarchical Attention Networks from Kränkel and Lee (2019) to find the most important words and sentences in the daily news [17]. Attention mechanisms are applied at both word and sentence level to improve the prediction performance, providing better contextual modelling for words or sentences that convey a different meaning based on a specific context. Second, the proposed hybrid model combines news encoding and time series forecasting to jointly predict future stock movement, which differentiates from traditional time series forecasting methods that only consider previous stock volatility, or deep learning models only based on financial news.

The remainder of the paper is organized as follows. We introduce the architecture of the proposed model and its sub-components in Section 2. Section 3 details the experiment process and supporting data used. Section 4 concludes the paper.

II. METHODOLOGY

The proposed model consists of two types of deep neural networks: hierarchical attention networks for word and sentence level sequence modelling, and long short term memory networks for sequential modelling of time series. We first introduce the two type of networks, followed by a detailed introduction of their composition in the proposed network architecture.

A. Hierarchical Attention Networks (HAN)

In order to analyse the news data to predict stock price movement, we use HAN, an advanced document classification method proposed by Yang et al. (2016), due to its advantage in modelling complex dependence structure embedded in textual data [12]. While most text classification methods assess importance weights from previous word features, HAN deploys a hierarchical structure that mirrors the composition of documents, using two layers of attention mechanisms to capture the difference of importance at both word and sentence level. This is based on the fact that sentences consist of words and roll up to documents. Both words and sentences convey meaning at different levels, jointly shaping the meaning of the whole document. The general architecture of HAN contains a word sequence encoder, a word-level attention layer, a sentence encoder and a sentence-level attention layer.

- Word Encoder

Given a sentence with words $w_{it}, t \in [0, T]$, where w_{it} denotes word i in sentence t, a pre-trained encoder GloVe is used to convert each word feature to corresponding word embedding $x_{it} = W_e w_{it}, t \in [1, T]$, using an embedding matrix W_e with 100 dimensions (i.e., each word is represented by 100 dimensions in a matrix). This is then followed by a bidirectional Gated Recurrent Network (GRU), which serves as an encoding mechanism to get annotations of words by combining contextual information in forward and backward hidden states $\overrightarrow{h_{it}}$ and $\overleftarrow{h_{it}}$. To be specific,

$$\overrightarrow{h_{it}} = \overrightarrow{GRU(x_{it})}, t \in [1, T],$$
$$\overleftarrow{h_{it}} = \overleftarrow{GRU(x_{it})}, t \in [T, 1],$$
$$h_{it} = \left[\overrightarrow{h_{it}}, \overleftarrow{h_{it}}\right].$$

- Word Attention

We use attention mechanism to extract important word representations when rolling up to a sentence. Denote u_{it} as the hidden representation of h_{it} through a one-layer perceptron, and α_{it} as the normalised importance weight per word by a softmax function. The corresponding weights to each word and sentence vector s_i could then be derived by summarising the importance weights of word annotations:

$$u_{it} = tanh(W_w h_{it} + b_w),$$
$$\alpha_{it} = \frac{exp(\boldsymbol{u_{it}}^{\top} u_w)}{\sum_t exp(\boldsymbol{u_{it}}^{\top} u_w)},$$
$$s_i = \sum_t \alpha_{it}\, h_{it}.$$

- Sentence Encoder

Similarly, given a sentence vector s_i, a bidirectional GRU is used to calculate the annotations of sentences by combining forward and backward hidden states $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$. To be specific,

$$\overrightarrow{h_i} = \overrightarrow{GRU(s_i)}, s \in [1, L],$$
$$\overleftarrow{h_i} = \overleftarrow{GRU(s_i)}, s \in [L, 1],$$
$$h_i = \left[\overrightarrow{h_i}, \overleftarrow{h_i}\right].$$

- Sentence Attention

We again use attention mechanism to extract important sentence level representations and derive the document vector v by summarising the importance weights of sentence annotations. Specifically,

$$u_i = tanh(W_s h_i + b_s),$$
$$\alpha_i = \frac{exp(\boldsymbol{u_i}^{\top} u_s)}{\sum_t exp(\boldsymbol{u_i}^{\top} u_s)},$$
$$v = \sum_t \alpha_i\, h_i.$$

B. Long Short-Term Memory (LSTM)

The historical stock prices contain complex serial dependence and are usually subject to external factors such as news release. To capture the inherent pattern in the time series, we resort to the widely used LSTM network. LSTM is a RNN based architecture that is widely used in time series forecasting. There are three gates in a LSTM model: input gate, forget gate and output gate, all of which use the sigmoid activation functions σ to form a probabilistic output.

Input gate: $i_t = (w_i[h_{t-1}, x_t] + b_i)$
Forget gate: $f_t = \left(w_f[h_{t-1}, x_t] + b_f\right)$

Output gate: $o_t = (w_o[h_{t-1}, x_t] + b_o)$
where w_x is the weight for respective gate(x) neurons, h_{t-1} is the output of the previous LSTM blocks, x_t is the input at current timestamp and b_x is the biases for the respective gate(x).

C. Proposed Model

The overall architecture of our model is shown in Fig. 1. The model consists of two modules: a text mining module based on HAN network, and a time series module using LSTM network. The first module is used to effectively model the financial news and extract important signals at both word and sentence level module, and the second module is used to model the complex serial dependence in the time series data. Both are designed to follow a parallel structure, merging at the end of the network before reaching the final output layer. Source code on the proposed model will be released once the paper is accepted.

D. Alternatives Comparison

- Naive Bayes

Naive Bayes is the most straightforward and fast classification algorithm. It is a classification technique based on Bayes' theorem for binary (two-class) and multi-class classification problems. And it assumes the effect of a particular feature in a class is independent of other features

Bayes' theorem: $P(B|A) = \frac{P(B|A)*P(A)}{P(B)}$

- Logistic Regression

Logistics regression is a common and useful regression method based on Machine Learning algorithms for solving the binary classification problem. And it describes and estimates the relationship between one dependent binary variable and independent variables.

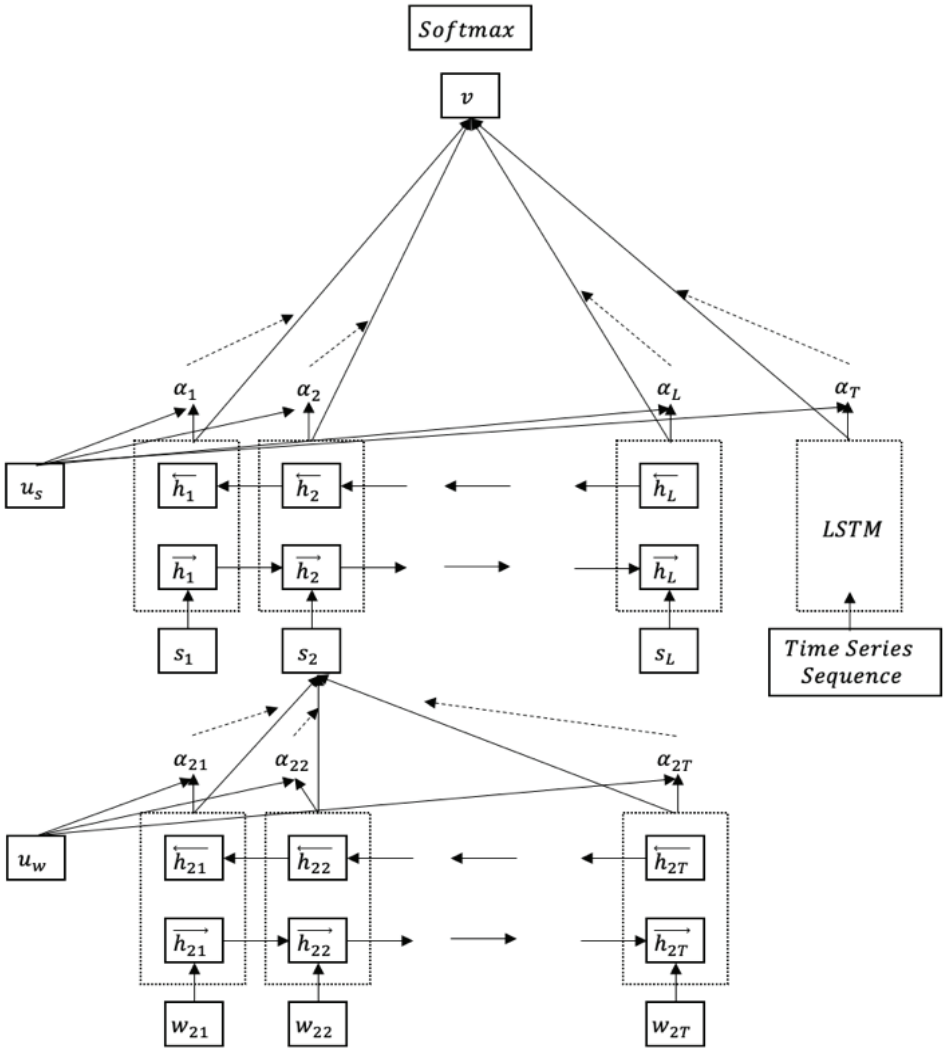

Fig. 1. Architecture of proposed model.

- Support Vector Machine (SVM)

SVM is a classification approach known for its kernel trick to handle nonlinear input spaces. It constructs a hyperplane in multidimensional space to separate different classes; and generates optimal hyperplane in an iterative manner, which is used to minimize an error. The core idea of SVM is to find a maximum marginal hyperplane (MMH) that best divides the dataset into classes.

III. EXPERIMENTS

A. Data Collection

In this research, there are two channels of data provided in this dataset. Our news data is extracted from historical news headlines from Reddit WorldNews Channel and news are ranked based on their popularity. Our stock data is extracted from Dow Jones Industrial Average (DJIA) from Yahoo Finance for the period of 15/ 08/ 2008 to 01/ 07/ 2016. There are 1984 rows in total. Table 1 is the head of news data.

B. Data Pre-processing

- Labelling Techniques

In order to identify the correlation between the news data and corresponding stock price movement, we use the percentage change of the previous day to create a target label consisting of three categories— negative, neutral and positive. Percentage change = $\frac{Close_{currentday} - Close_{previousday}}{Close_{currentday}}$.

If the percentage change is lower than −0.4%, the news data is labelled negative (0); if the percentage change is between −0.4% to +0.5% the news data is labelled neutral (1); if the percentage change is greater than 0.5% the news data is labelled positive (2). Table II is the sample data labelling result.

- Data Cleaning

We clean the raw data by lemmatising and lower-casing each word, followed by removing stop words and non-alpha numeric words, where English stop words from the NLTK library are used. And we use integer encoding for the labels. Table III is the head of sample data after cleaning.

- Data Set Splitting

We split our data to set a train, validation and test data set. The training set contains 1190 rows of data, which accounts for 60% of the whole data. The test set contains 397 rows of data, which accounts for 20% of the whole data.

TABLE I. SAMPLE NEWS DATA

Date	News	Closing price change
01/07/2016	A 117-year-old woman in Mexico City finally re...	0.001081
30/06/2016	Jamaica proposes marijuana dispensers for tour...	0.013298
29/06/2017	Explosion At Airport In Istanbul Yemeni former...	0.016368
28/06/2018	2,500 Scientists To Australia: If You Want To ...	0.015722
27/06/2019	Barclays and RBS shares suspended from trading...	-0.014971

TABLE II. SAMPLE DATA LABELLING RESULT

Date	Percentage Change	label
01/07/2016	0.001081	1
30/06/2016	0.013298	2
29/06/2017	0.016368	2
28/06/2018	0.015722	2
27/06/2019	-0.014971	0

TABLE III. DATA AFTER CLEANING

News	Category	Code
woman mexico city finally received birth certi...	1	1
jamaica proposes marijuana dispenser tourist a...	2	2
explosion airport istanbul yemeni former presi...	2	2
scientist australia want save great barrier re...	2	2
barclays rb share suspended trading tanking 8 ...	0	0

C. Structure of Proposed Model

By using the model proposed in methodology and putting in our training set and test set, the detailed architecture of our model is shown in Fig. 2.

D. Model Comparison

We compared the precision, recall and F1-score of training set and test set of different models.

Precision is the ability of a classifier not to label an instance positive that is actually negative. For each class, it is defined as accuracy of positive predictions, i.e., the ratio of true positives to the sum of a true positive and false positive.

Recall is the ability of a classifier to find all positive instances. For each class it is defined as fraction of positives that were correctly identified, i.e., the ratio of true positives to the sum of true positives and false negatives.

Table IV contains comparisons of precisions of different models. We can see that our proposed model (HAN-LSTM) achieves the highest precision overall and in negative, middle, and positive classes in test set.

Table V contains comparisons of recall of different models. We can see that our proposed model (HAN-LSTM) achieves the highest recall overall and in positive class, and a comparable high recall in negative and middle class in test set.

Table VI contains comparisons of F1-score of different models. We can see that our proposed model (HAN-LSTM) achieves the highest F1-score overall and in positive class, and a comparable high F1-score in negative and middle class in test set.

Our proposed model can deliver better classification results in overall, positive and negative categories, which are more relevant to practitioners when it comes to financial text mining.

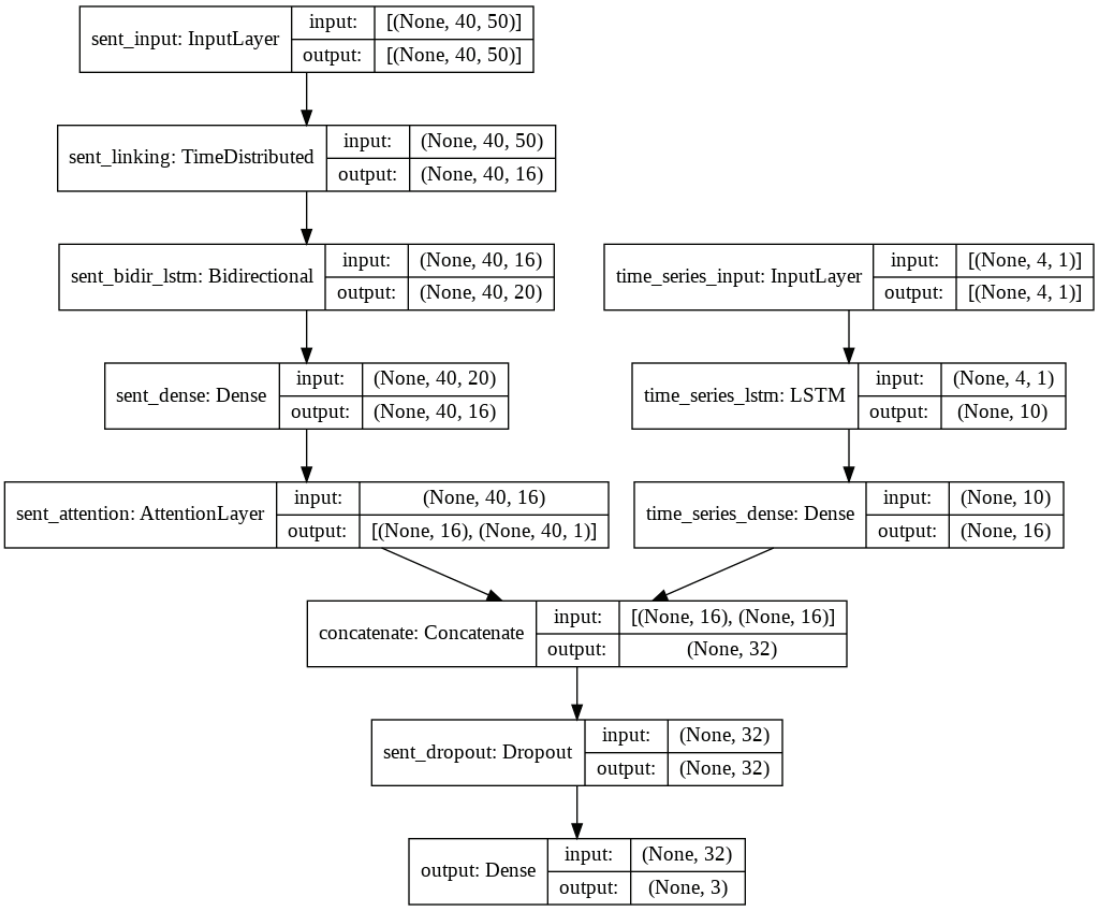

Fig. 2. Detailed structure of proposed model after putting in data set.

TABLE IV. PRECISION

Type	Model	Precision[a]			
		Overall	Negative	Middle	Positive
Training	NB	0.99	0.99	1.00	0.99
	LR	0.84	0.95	0.62	0.93
	SVM	0.74	0.82	0.58	0.83
	HAN	0.15	0	0.46	0
	LSTM	0.15	0	0.46	0
	HAN-LSTM	0.77	0.83	1	0.47
Test	NB	0.32	0.20	0.48	0.27
	LR	0.34	0.27	0.47	0.27
	SVM	0.36	0.33	0.48	0.26
	HAN	0.15	0	0.46	0
	LSTM	0.15	0	0.46	0
	HAN-LSTM	**0.42**	**0.50**	**0.49**	**0.27**

[a] $\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Positive}}$

TABLE V. RECALL

Type	Model	Recall[b]			
		Overall	Negative	Middle	Positive
Training	NB	0.94	0.92	1	0.91
	LR	0.64	0.48	0.99	0.47
	SVM	0.94	0.92	1	0.91
	HAN	0.33	0.36	0.97	0.37
	LSTM	0.33	0	1	0
	HAN-LSTM	0.64	0.03	0.88	0.99
Test	NB	0.31	0.22	0.46	0.26
	LR	0.34	0.06	0.90	0.08
	SVM	0.35	0.10	0.88	0.08
	HAN	0.33	0	1	0
	LSTM	0.33	0	1	0
	HAN-LSTM	**0.35**	**0.08**	**0.54**	**0.45**

[b] $\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}}$

TABLE VI. F1-SCORE

Type	Model	F1-Score[c]			
		Overall	Negative	Middle	Positive
Training	NB	0.99	0.99	0.99	0.99
	LR	0.68	0.64	0.76	0.63
	SVM	0.58	0.50	0.73	0.51
	HAN	0.21	0	0.63	0
	LSTM	0.21	0	0.63	0
	HAN-LSTM	0.54	0.06	0.94	0.64
Test	NB	0.32	0.21	0.47	0.26
	LR	0.28	0.09	0.62	0.13
	SVM	0.30	0.16	0.62	0.12
	HAN	0.21	0	0.63	0
	LSTM	0.21	0	0.63	0
	HAN-LSTM	**0.33**	**0.13**	**0.51**	**0.34**

[c] F1-Score = $\frac{2*Precision*Recall}{Precision+Recall}$

IV. CONCLUSION

In this paper, we propose a new network architecture that models two levels of sequence: financial news data that highlight the sequential nature of words and sentences in the texts, and a historical stock price which has complex serial dependence structure. In this proposed framework, we use HAN in the text mining module to effectively extract and utilize important signals at both word and sentence level. For the time series module, the LSTM network is used to systematically capture the structural components in the time series data. By running multiple tests using different benchmark models, the results show that the proposed method performs better in all out-of-sample metrics for both positive and negative stock returns, demonstrating the potential of complex composite models when analyzing both structured and unstructured financial data.

REFERENCES

[1] Vishwanath, R., eena, S.V., Srikantaiah, K.C., Kumar, K., Shenoy, P., Venugopal, K., and Patnaik, L., 2013. APST: Approximation and Prediction of Stock Time-Series Data using Pattern Sequence.

[2] Ariyo, A., Adewumi, A. and Ayo, C., 2014. Stock Price Prediction Using the ARIMA Model. 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation,.

[3] Angadi, M.C., and Kulkarni, A., 2015. Time Series Data Analysis for Stock Market Prediction using Data Mining Techniques with R. International Journal of Advanced Research in Computer Science, 6, 104-108.

[4] Khedr, A., S.E.Salama and Yaseen, N., 2017. Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis. International Journal of Intelligent Systems and Applications, 9(7), pp.22-30.

[5] Kalra, S. and Prasad, J., 2019. Efficacy of News Sentiment for Stock Market Prediction. 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon),.

[6] Derakhshan, A. and Beigy, H., 2019. Sentiment analysis on stock social media for stock price movement prediction. Engineering Applications of Artificial Intelligence, 85, pp.569-578.

[7] Elagamy, M., Stanier, C. and Sharp, B., 2018. Stock market random forest-text mining system mining critical indicators of stock market movements. 2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP),.

[8] Pagolu, V., Reddy, K., Panda, G. and Majhi, B., 2016. Sentiment analysis of Twitter data for predicting stock market movements. 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES),.

[9] Batra, R. and Daudpota, S., 2018. Integrating StockTwits with sentiment analysis for better prediction of stock price movement. 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET),.

[10] Islam, M., Al-Shaikhli, I., Mohd Nor, R. and Varadarajan, V., 2018. Technical Approach in Text Mining for Stock Market Prediction: A Systematic Review. Indonesian Journal of Electrical Engineering and Computer Science, 10(2), p.770.

[11] Nabipour, M., Nayyeri, P., Jabani, H., S., S. and Mosavi, A., 2020. Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; a Comparative Analysis. IEEE Access, 8, pp.150199-150212.

[12] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. and Hovy, E., 2016. Hierarchical Attention Networks for Document Classification. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,.

[13] Gao, S., Ramanathan, A. and Tourassi, G., 2018. Hierarchical Convolutional Attention Networks for Text Classification. Proceedings of The Third Workshop on Representation Learning for NLP,.

[14] Pappas, N. and Popescu-Belis, A., 2021. Multilingual Hierarchical Attention Networks for Document Classification. [online] ACL Anthology. Available at: <https://www.aclweb.org/anthology/I17-1102> [Accessed 14 April 2021].

[15] Huang, L., Yan, H., Ying, S., Li, Y., Miao, R., Chen, C. and Su, Q., 2020. Hierarchical Attention Network in Stock Prediction. Lecture Notes in Computer Science, pp.124-136.

[16] Liu, Q., Cheng, X., Su, S. and Zhu, S., 2018. Hierarchical Complementary Attention Network for Predicting Stock Price Movements with News. Proceedings of the 27th ACM International Conference on Information and Knowledge Management,.

[17] Kränkel, M. and Lee, H., 2019. Text Classification with Hierarchical Attention Network. [online] Humboldt-wi.github.io. Available at: <https://humboldt-wi.github.io/blog/research/information_systems_1819/group5_han/> [Accessed 14 April 2021].