MULTIPLE INSTANCE LEARNING NETWORKS FOR STOCK MOVEMENTS PREDICTION WITH FINANCIAL NEWS

Yiqi Deng and Siu Ming Yiu

Department of Computer Science, The University of Hong Kong, Hong Kong, China

ABSTRACT

A major source of information can be taken from financial news articles, which have some correlations about the fluctuation of stock trends. In this paper, we investigate the influences of financial news on the stock trends, from a multi-instance view. The intuition behind this is based on the news uncertainty in random news occurrences and the lack of annotation for every single financial news. Under the scenario of Multiple Instance Learning (MIL) where training instances are arranged in bags, and a label is assigned for the entire bag instead of instances, we develop a flexible and adaptive multi-instance learning model and evaluate its ability in directional movement forecast of Standard & Poor's 500 index on financial news dataset. Specifically, we treat each trading day as one bag, with certain amounts of news happening on each trading day as instances in each bag. Experiment results demonstrate that our proposed multi-instance-based framework gains outstanding results in terms of the accuracy of trend prediction, compared with other state-of-art approaches and baselines.

KEYWORDS

Multiple Instance Learning, Natural language Processing, Stock Trend Forecasting, Financial News, Text Classification

1. INTRODUCTION

Stock trend prediction has always been a hotspot for both investors and researchers to facilitate making useful investment decisions, conducting investment, and gaining profits. Normal trend prediction tasks mainly take direct views on the stock prices. Based on stock prices, fundamental analysis [1], technical analysis [2, 3], and historical price time series analysis [4-6] have been used to aid in previous stock analysis. In addition to the direct quantitative information the numeric price brings on stock trends, financial news implies qualitative relations between daily events and their effect on the stock prices. Intuitively, people intend to buy stocks on hearing positive news and sell on negative news. Literature in [7-9] has also indicated that events reported in financial news play important roles concerning the stock trends in the financial market.

Using financial news to predict stock trends can be regarded as one text binary classification task. Take one trading day as an example, the trend of the stock is up if the closing price of the day is higher than the previous day, otherwise, it will have a downward trend. However, the uncertainty in daily financial news presents challenges to our normal financial text analysis. It comes from two sources: uncertainty in occurrences of news each day and uncertainty in the number/distribution of positive and negative news each day. For the uncertainty in daily news occurrence, financial news appears randomly most of the time. As it can be seen from Figure 1, sometimes there is no relevant news in one day while sometimes there can be more than ten to

DOI: 10.5121/ijaia.2022.13402

hundreds of news in one day. In view of the uncertainty in the number/distribution of positive and negative news each day, stock trend is generally related to certain amounts of financial news instead of one piece. On a day with stock trends going down, it is quite unrealistic to consider that all the news within this day is conveying negative sentiments, in that there can be some good news, as well as neutral or even unrelated news (see Figure 2).

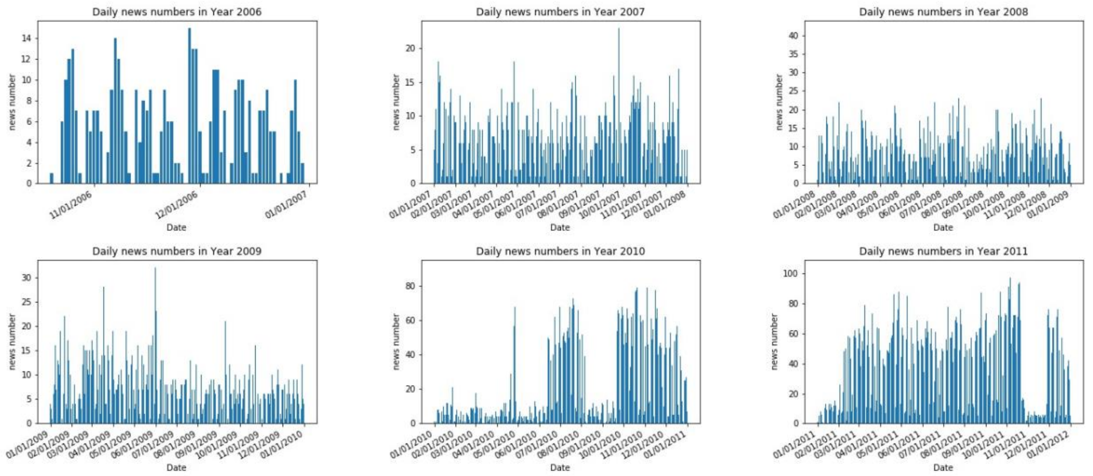


Figure 1. Random amounts of news appearing in each year.

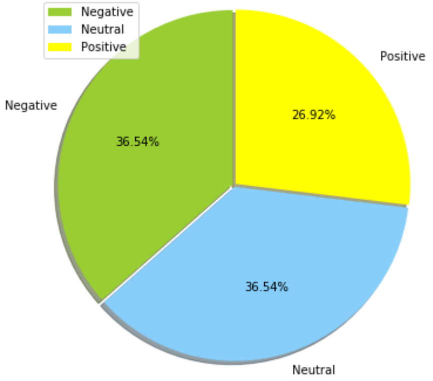


Figure 2. Sentiment values of news items on September 20, 2011, calculated by Natural Language Toolkit (NLTK)

In the scenario of stock trends forecasting, we mostly obtain labels for groups of news (bag-level labels) instead of each piece of news (instance-level labels). Labels for groups of news (baglabels) are clearly reflected in stock trends, which are estimated through the change of two consecutive stock closing prices. With complete daily group class labels, supervised learning has dominated in previous literature [10-15] pertaining to stock trend prediction tasks through financial news. In contrast to group labels, directly gaining single news labels is quite difficult. Considering the random and rapid changes in news amounts, labeling single news (instance labels) is quite expensive and impractical. Besides investors with different risk preferences treat and mark differently on each news, which also hinders the labeling of single news. However, the unknown news labels, as well as their weak relationships each day are indirect yet hardly negligible. Sometimes, a sudden piece of good news could alter investors' previously bearish decisions when investors do trading. What's more, opposed decisions are likely to be made by investors when they look at single news. An example is illustrated through the news published on

September 20, 2011. The final stock is ended with a price that lower than its previous day, which we considered this day with a downward trend. For news on September 20, 2011 shown in Table 1, if investors only look at the first two news that convey positive signals, say 'boost', 'increase’, they may think the stock will go up. Or investors will get puzzled when they see negative aspects ('cost', 'falter', 'low'), that indicate a downward trend in the following news. Therefore, inference for news labels is needed and should be considered in a precise stock prediction.

Table 1. News posing on Sep. 20, 2011

VADER Sentiment[1]	**News**
++ (pos)	China’s Stocks Rise From 14-Month Low; Commodity Producers Gain.
++ (pos)	Obama’s Home State Illinois Turns to China for Economic Boost.
++ (pos)	U.S. Gulf Crude-Oil Premiums Increase as Brent-WTI Gap Widens.
++ (pos)	U.S. Natural Gas Fund Premium at 0.31% on Sept. 19
…	
- - (neg)	China Endorsing Tobacco in Schools Adds to $10 Trillion Cost.
- - (neg)	China Stock-Market Sentiment at Historic Low, Citigroup Says.
- - (neg)	Hurricane Irene Cost NYC at Least $55M: Official.
- - (neg)	Oil Slides in New York on Speculation Demand to Falter; Brent Erases Drop.
…	
+ - (neu)	Oil Trades Near a Three-Week Low in New York; Brent Crude Climbs in London.
+ - (neu)	Short-Term Stimulus Won’t Help U.S. in Long Run: Glenn Hubbard.
+ - (neu)	U.S. August Building Permits by Type and Region.
+ - (neu)	U.S. Solar Power Rises 69 Percent, Led by Commercial Projects.
+ - (neu)	China Jan.-Aug. Average Export Prices Rise 10.3%.
…	
Trend	DOWN TREND

Actually, as one type of weakly supervised learning algorithm, multiple instance learning (MIL) can be utilized to infer unknown news (instance) labels and the weak correlation between them. It helps ameliorate the limitation on the uncertainty of financial news mentioned above and construct models on stock trends prediction at the instance, bag levels. Therefore, in this work, we aim to adopt Multiple Instance Learning (MIL) [17] and consider the effects of financial news on stock trends from the perspective of Multiple Instance Learning. Related to the earlier work, in this paper we make the following contributions:

- A summarization of MIL principles used in scenarios of stock trend prediction using the financial news.

- Build up a novel MIL model to alleviate the problems of finance news uncertainty and intact single news labels when predicting the stock trends.

- Empirical evidence that our proposed MIL model can achieve impressive results on the S&P 500 stock index prediction, competing with other conventional neural architectures and previous MIL methods.

[1] We use the VADER sentiment analysis tools in [16] to estimate sentiment values for all news items on September 20, 2011.

The paper is organized as follows. We review related work and previous mainstream approaches in news stock trend prediction in Section 2. In Section 3, we introduce our proposed multiple instance learning (MIL) framework, with a description of how to represent each news, how to infer the possibility of constituent news within each day, as well as the day vectors construction and day(bag)-level supervision. Section 4 presents our experiments conducted on specific financial news datasets. We give our experimental results as well as result analysis by comparing our method against previous approaches in this part. Finally in Section 5, we conclude and summarize the paper.

2. Related Work

2.1. Multiple Instance Learning

Multiple instance learning was originally introduced by Dietterich et al., [18] in drug activity prediction. In multiple instance learning, the training set consists of labeled "bags". Each bag is a collection of instances. The exact labels are known for the entire bag, whereas labels for every instance keep unknown. There are several classification situations where class labels are incomplete at the instance level but only available for groups of examples due to time-consuming manual annotation and restricted label sources acquired. Alleviating the burden of obtaining limited-labeled datasets, Multi-instance learning has been successfully put into practice in areas of image classification [19, 20, 21], document modelling [22], event extraction [23], sound event detection (SED) [24], etc. The multi-instance learning approach also shows its feasibility in the application of text mining tasks. He Wei et at., [25] treat each document as a bag, the sentences in the document as each instance, to investigate text classification problems. Dimitrios et al., [26] adopt multi-instance learning on the problem of predicting labels for sentences given labels for reviews. Based on instance-level similarity and group-level labels, an objective function 'Group Instance Cost Function'(GICF) is proposed to encourage smoothness of inferred instance-level labels. Nikolaos et al., [22, 27] introduce a weighted multiple-instance regression (MIR) framework for document modeling. Stefanos et al., [28] present a neural network model for fine-grained sentiment analysis within the framework of multiple instance learning. Without the need for segment-level labels, their model learns to predict the sentiment of text segments. From above literature, it can be indicated that multiple instance learning can help develop strong predictive models albeit imperfect instance labels are employed. In the original assumption of multiple-instance binary classification, a bag is classified as negative if every instance inside the bag is negative. If there is at least one positive instance, the bag is classified as positive. However, in field of financial news analysis, original assumption becomes less plausible. One, as was already said, is that it’s quite unrealistic to believe that all the news on a day with downturn convey negative sentiments. Another holds that if there is at least one positive news, then all predictions will be positive. This will lead to severe label imbalance issues. To deal with that, our work proposes to extend original assumption of MIL. We directly infer the probability that one financial news belonging to a certain label and find weak correlation between news. Then we develop MIL, the challenging yet potentially powerful variant of incomplete supervision learning, to the task of news-based trend prediction.

2.2. Financial News for Stock Trend Prediction

Financial news plays an important role with respect to the stock trends in the financial market. By means of deep learning and natural language processing (NLP), existing methods on stock market prediction by analyzing financial news have proven to be quite effective. Financial news contains useful information in unstructured textual form. When representing each news title, it is non-trivial to extract semantic information and context information within each news title. The vector representation of words [29, 30] facilitates feature extraction from not only words but also sentences and documents. Classical methods such as averaging word vectors [31], training paragraph vectors [32] can be efficient, yet they have been indicated incapable of preserving semantics and gaining interpretation of linguistic aspects such as word order, synonyms, co-reference in the original news. To overcome this limitation, some improved representation techniques have been advanced in the following studies. Ding et al. [33] use open information extraction (Open IE) to obtain structured events representations in news. Later in [34], he put forward a novel neural tensor network to extract events in financial news. Get inspired by work [35], works such as [36, 37] adopt hierarchical structures to perform the classification: Hu et al. [36] adopt a hierarchical structure called Hybrid Attention Networks (HAN) to catch more features and help address the challenge of low-quality, chaotic online news. Liu et al. [37] extract news text features and context information through Bidirectional-LSTM. A self-attention mechanism is applied to distribute attention to most relative words, news and days. Ma et al. [38] develop a novel Distributed Representation of news (DRNews) through creating news vectors that describe both the semantic information and potential linkages among news events in an attributed news network. News vector representation has achieved state-of-art performances on various financial text classification tasks. A better text representation on news titles is vital in financial news analysis to capture features related to stock trends forecasting.

As predictive methods, deep learning models present high performances in traditional natural language processing tasks, namely, Convolution Neural Network (CNN) [39-41], Recurrent Neural Network (RNN) [42, 43], etc. In recent studies, authors in [31] propose a recurrent convolutional neural network (RCNN) model on stock price predictive tasks. Word embedding and sentence embedding are made as better embedding vectors for each piece of news. Huy et al. in [42] utilize a new Bidirectional Gated Recurrent Unit (BGRU) model for the stock price movement classification. Xu Y et al. [45] propose a stock price prediction model with the aid of news event detection and sentiment orientation analysis, through introducing Convolutional Neural Network (CNN) and Bi-directional Long Short-Term Memory (Bi-LSTM) in their predictive model. Most recently, a recurrent state transition model, integrating the influence of news events and random noises over a fundamental stock value state, is constructed in [46] for the task of news-driven stock movement prediction. A tensor-based information framework for predicting stock movements in response to new information is also introduced in [47]. From neural-network-based approaches to hierarchical structures-based models, to tensor-based networks, these methods have evolved into the mainstream and cutting-edge methodologies in the field of stock trend prediction from financial news texts.

3. METHODOLOGY

In this section, we outline the structure of our suggested multiple instance learning model in this section. To further, we relate how to obtain news (instance) representations to better extract keywords and context information within the news, as well as how to apply multiple instance learning to address some of the pitfalls mentioned in previous parts, which are: the uncertainty of news relating to the randomness occurrences and the unknown annotation for each piece of news. The model design has 4 stages: word embedding, news(instance) encoding, news(instance)-level classifiers, and bag-level representation and final classification.

3.1. Definitions & Formulation

Given an input dataset D, the dataset D contains a set of labeled bags $B = \{B_1, B_2, \dots, B_M\}$, each of which is made up of a group of unlabeled instances. In our multiple instance learning (MIL) framework, we regard news as instances and all the news that appears on that day as a bag.

Consider prediction of stock trend over M trading days, each trading day k represents the bag $B_k, k = 1,2,\dots,M$, with n_k news titles (instances). The i-th news in bag B_k, $n_{i,k} \in R^d, i = 1,\dots,n_k$, is a d-dimensional vector learning from neural networks. With numerical labels Y_k derived from the daily stock close price, we are given bag labels for the stock trends each day. Then we have:

$$D = \{(B_k, Y_k)\}, k = 1,2,\dots,M$$

where $B_k \in B$ and Y_k is a bag label assigned to day k.

We assume binary classification in this paper, in this case,

$$\mathrm{Y_k} = \begin{cases} 0, & Close_{day\,k} < Close_{day\,k-1} \\ 1, & Close_{day\,k} > Close_{day\,k-1} \end{cases}$$

where 0 represents a downward stock trend and 1 shows the upward stock trend for the day k.

Previous studies have mostly focused on the relevance of news. They divide news into related or unrelated parts. Indeed, each news item contains a portion of the information that influences the direction of the stock price trend. Therefore, our model makes an effort to predict how likely each piece of news is to move the stock upwards or downwards. The philosophy of multi-instance learning is to build classifiers to predict the labels of unknown bags by analyzing the label-known bags and its multiple instances. Based on that, in our work, we promote the relevance of news to the inference of individual news probability of being up and down.

3.2. Proposed Model

3.2.1. Word Embedding

To obtain vector representation of each news text (instance), one key step is the use of embedding techniques. The embedding techniques map words into numerical vector spaces through an embedding matrix. Through the mapping, richer numerical representations of text input are created, enabling the deep multi-instance models to rely on these vector representations and improve performances in specific tasks. In our paper, the embedding takes a sequence as input, corresponding to a set of news titles. Assume that the i-th news title in bag B_k comprises T_i words. $w_{it,k}, t \in [1, T_i]$ stands for the t-th words in news n_i in bag B_k. We first embed the individual words $w_{it,k}$ to vectors through word embedding matrix:

$$L_w \in R^{d\times|V|}$$

where d is the dimension of word vector and $|V|$ is vocabulary size. Then the embedded vectors for word $e_{it,k}, \in R^d$ is gained through

$$e_{it,k} = L_w \times w_{it,k}$$

The word vectors can be either randomly initialized or be pre-trained with embedding learning algorithms such as Glove and Word2Vec. Here, we adopt Glove [48] for better use of semantic and grammatical associations of words. In details, the Glove file that pre-trained 100-dimension word vectors on 6 billion tokens, 400K vocabulary, has covered most of the words in our news texts.

3.2.2. News(instance) encoding

Drawing inspiration from [35, 37], we exploit a Bidirectional-LSTM after word embedding to incorporate the contextual information from both directions for words. The recurrent structure in LSTM promotes the capture of context information. Compared with standard recurrent neural network (RNN), the gated mechanism in LSTM prevents the unbounded cell state and tackles the problem of vanishing/exploding gradient, which makes it more applicable in modeling semantics of long texts. Hence, we have the following computation of LSTM cells:

$$f_t = \sigma\left(W_f\left[h_{i(t-1),k}, e_{it,k}\right] + b_f\right)$$

$$i_t = \sigma\left(W_i\left[h_{i(t-1),k}, e_{it,k}\right] + b_i\right)$$

$$\widetilde{C_t} = \tanh\left(W_C\left[h_{i(t-1),k}, e_{it,k}\right] + b_C\right)$$

$$C_t = f_t \otimes C_{t-1} + i_t \otimes \widetilde{C_t}$$

$$o_t = \sigma\left(W_o\left[h_{i(t-1),k}, e_{it,k}\right] + b_o\right)$$

$$h_{it,k} = o_t \otimes \tanh(C_t)$$

In LSTM, there are three gates, i.e., input gate i_t, forget gate f_t, and output gate o_t. For current input $e_{it,k}$ at time t and previous hidden state $h_{i(t-1),k}$ at time $t-1$, the calculation in forget gate f_t indicates the ability to forget old information. This gate decides what information should be forgotten or kept. Input gate i_t is derived from input data $e_{it,k}$ and previous hidden node $h_{i(t-1),k}$ through a neural network layer. $\widetilde{C_t}$ represents the cell state update value. Through the forget gate and the input gate, the cell state C_t is gained, with information of C_{t-1} and $\widetilde{C_t}$. The output gate o_t decides what the next hidden state $h_{it,k}$ should be. $h_{it,k}$ is obtained from the output gate o_t and cell state C_t, where o_t is calculated in the same way as f_t and i_t. σ represents the sigmoid activation function.

The bidirectional LSTM contains the past and future context of the word. Through two hidden states $\overrightarrow{LSTM}$, $\overleftarrow{LSTM}$, information can be preserved, at any point in time, from both past and future. The forward $\overrightarrow{LSTM}$makes news be read from the first word to the last word, and the backward $\overleftarrow{LSTM}$ allows information in news to flow from $w_{iT_i,k}$ to $w_{i1,k}$. Therefore,

$$\overrightarrow{h_{\imath t,k}} = \overrightarrow{LSTM} e_{it,k}, \qquad t \in [1, T_i]$$

$$\overleftarrow{h_{\imath t,k}} = \overleftarrow{LSTM} e_{it,k}, \qquad t \in [1, T_i]$$

We concatenated two hidden vectors $\overrightarrow{h_{\imath t,k}}$and $\overleftarrow{h_{\imath t,k}}$ into

$$h_{it,k} \;=\; \left[\overrightarrow{h_{\imath t,k}}, \overleftarrow{h_{\imath t,k}}\right] \in R^{2\times u},$$

which represents ith news title in the kth day. n_k refers to the total amount of news items on kth day, and u is hidden units of LSTM.

Words within the news are not equally informative to investors. Investors usually pay more attention to keywords whenever they see a news story. Hence, we introduce an attention mechanism on top of the Bi-LSTM layer, so that it can reward the words offering critical information in our news(instance) representation. In details:

$$u_{it,k} = \tanh\left(W_w h_{it,k} + b_w\right)$$

$$\alpha_{it,k} = \frac{\exp\left(u_{it,k}\right)}{\sum_t \exp\left(u_{it,k}\right)}$$

$$n_{ik} = \sum_t \alpha_{it,k} \times h_{it,k} \in R^{2u}$$

We output the news(instance) vector as a weighted sum of the encoder hidden states. Then we compute the attention scores $\alpha_{it,k}$ and take softmax to get attention scores into a probability distribution. Finally, we take a weighted sum of values using attention distribution, obtaining the attention output as our news(instance) vector. The attention output mostly contains information from the hidden states that received high attention. Thus, the news(instance) vector is beneficial to aggregate the representation for informative words and better focus on keywords within the news.

3.2.3. News (Instance)-Level Classifiers

After we obtain the dense representations n_{ik} for each piece of news(instance) in day k, the news-level classifiers, albeit labels are unobserved in the training set, are constructed to make predictions at the news(instance) level and infer the probability of each unseen individual news driving the stock up or down. For the classifiers of news(instances), we feed the news vectors n_{ik} into a one-layer MLP with sigmoid activation:

$$\widehat{p_{ik}} = sigmoid(W_{news} n_{ik} + b_{news})$$

Where $\widehat{p_{ik}}$ represents a real-valued score, demonstrating the predicted probability that an instance, i.e., one piece of news belongs to a particular class label. W_{news}, b_{news} are the parameters of new-level classifiers.

3.2.4. Bag-Level Vector Representation and Classifiers

In classical MIL problems, once instance-level classifiers are set up to get inferred instance labels in the bag, bag labels can be obtained by combining its individual instances labels using *The aggregation functions*. Maximum operation, mean, or weighted averaging have been selected for frequently used aggregation functions in earlier works of literature [23, 24, 26, 27]. Vectors for instances don't need to be processed any further. However, for financial news text classification problems, additional steps for the bag feature extraction and bag-level representation are still necessary.

To make multi-instance learning more suitable in the classification problems of financial news text, we take a novel approach to learn vector representations of days(bag). To be specific, after we deduce the news(instance) probabilities from news(instance)-level classifiers, we do not directly aggregate them into the stock trend probabilities of a day(bag) and get bag-level predictions. Instead, we encode the possibility of its composed news(instance) into the day(bag) vector representation, and then build the day(bag)-level classifiers on top of that. In this case, a class of transformations from instances to bags can be parameterized by neural networks, making MIL

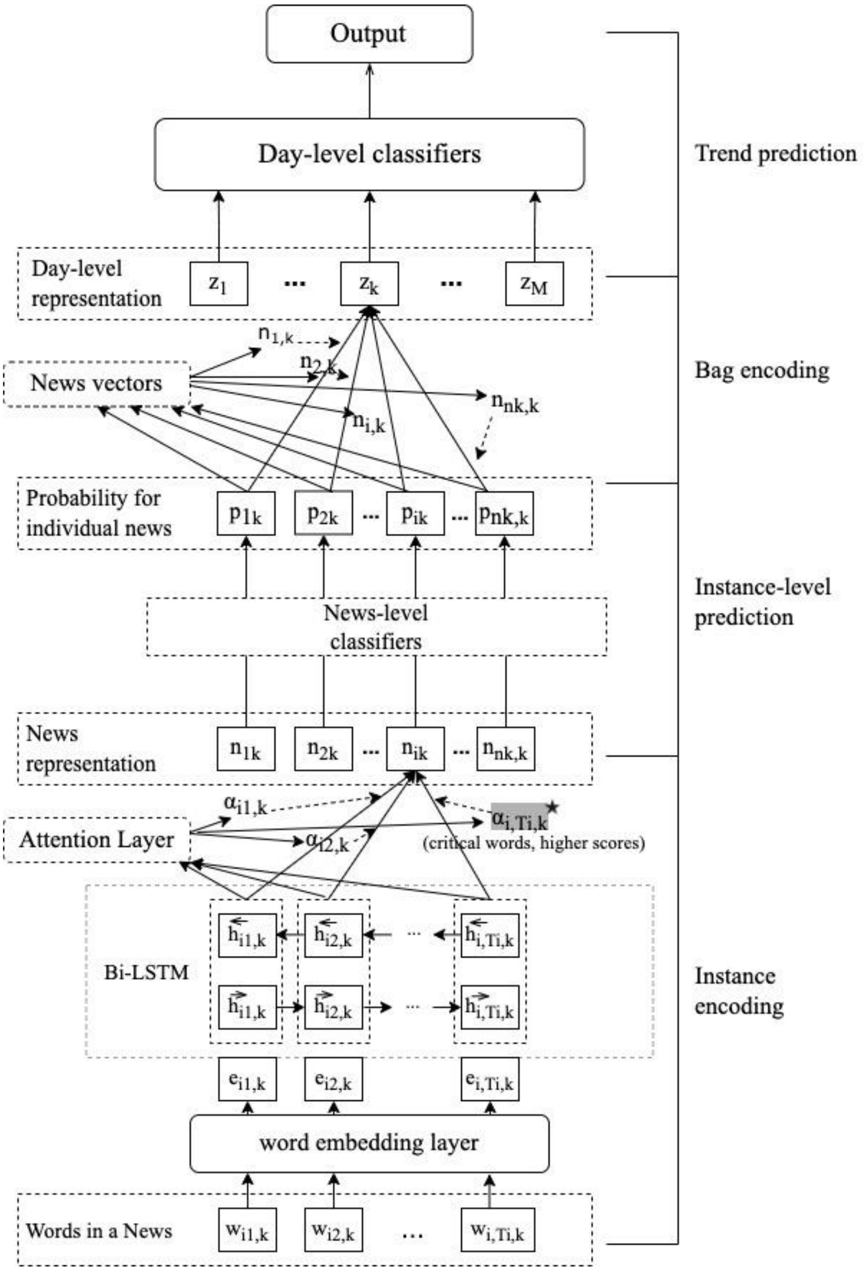


Figure 3. The overall architectures of our proposed Multiple Instance Learning Network

more flexible and being trained end-to-end. Hence, in the following steps we have vector-based representations for day(bag) k:

$$z_k = \frac{1}{n_k}\sum_{i=1}^{n_k} \widehat{p_{ik}}\, n_{ik}$$

Bag representation $z_k, k = 1,2,\dots,M$ is based on the prediction probability of its component new(instance). Using these vector representations, we can get the predicted day-level probabilities through day-level classifiers. By comparing the predicted day-level probabilities against the actual day labels, we can compute a cost function, and the network is then trained to minimize the cost. Detailed design of our proposed multiple instance learning model is displayed in Figure 3.

4. EXPERIMENTS

4.1. Datasets

To conduct our experiments, we use financial news datasets released by Ding et al. [33], between October 2006 and November 2013 in daily frequency. This dataset contains 106,521 news from Reuters and 447,145 news from Bloomberg. According to [33, 37], news titles alone are more predictive than adding news contents for trend forecasting tasks. Therefore, we extract the publication timestamps, the title for each piece of news from this dataset for our experiment. To catch the time period of the financial news, the historical stock price data for shares in Standard & Poor's 500 (S&P 500) index at the same period are also collected from Yahoo Finance to conduct our experiment on forecasting tasks.

4.1.1. Data Pre-Processing

In the following experiments, we conduct a series of basic text pre-processing in Figure 4. The basic text pre-processing steps include lowercasing, tokenizing each news item, removing stop words and infrequent words that appear less than 5 times. Then, we filter out news without any correlation to stocks, and make sure that all relevant symbols and company names are included. Following filtering, we obtained a total of 63403 news items. We divided the dataset into a training, validation, and test set. Table 2 contains summary statistics for the training, validation, and test sets.

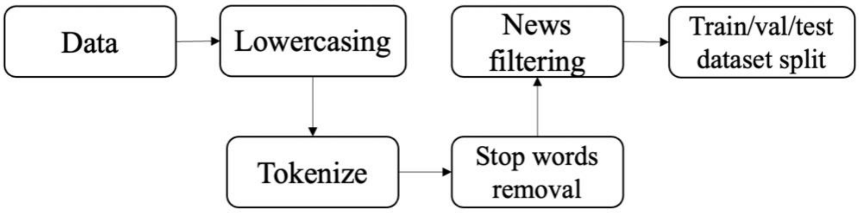
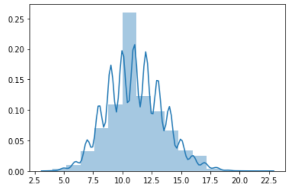
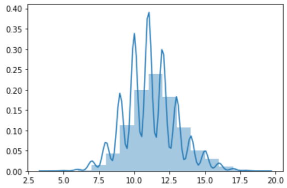
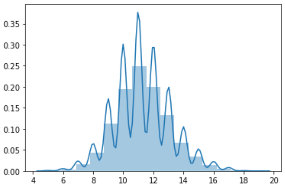


Figure 4. Pre-processing Steps on Datasets

4.1.2. Experiment setup

To train our model, we adopt Adadelta algorithm as our optimization algorithm. Adadelta can converge faster and be used when training deeper and more complex networks using the adaptive learning rate. To further, we set the initial learning rate α as 0.1. Mini-batches of 32 is organized through the training process. As the pre-trained word embeddings for news (instance) representation, we use GloVe embeddings [48], where |e| =100 is the vector size of the word embedding. The attention vector dimensionality is set to 100 and the LSTM hidden vector dimensions are set to 50 for each direction. In the stage of the day-level (bag-level) prediction, the model convolves its input with news(instance) embeddings. For hidden layers within the news (instance) classifiers, the hidden units are set to 150. Additionally, we use a dropout rate of 0.5 in the instance level and the bag level to avoid over fitting.

Table 2. Statistics Details of Datasets

Datasets	**Training**	**Validation**	**Test**
Time period	20/10/2006-27/06/2012	28/06/2012-13/03/2013	14/03/2013-20/11/2013
News amounts	38454	13237	11712
Mean	11.078795	11.127219	11.261783
Std	2.369729	1.834530	1.885941
Min	4.000000	4.000000	5.000000
Max	22.000000	19.000000	19.000000
Distribution			

4.2. Model Comparison

To evaluate our proposed model, in this section we set up a few baselines in contrast to our hybrid model. Baseline models include preceding mainstream models and previous MIL models with differences in aggregation approaches for instances and bags (mean operation and encoding with instance-level results). Each model is identified by the following notation for the sake of simplicity:

- BW-SVM: bag-of-words and support vector machines (SVMs) prediction model
- E-NN: structure events tuple input and NN prediction model, originally put forwards in Ding et al. [33]
- EB-NN: Event embedding input and NN prediction model in Ding et al. [34]
- S-NN: Following models in [31], a mean operator is performed on word embedding vectors within news as news vectors. In one day, we average all news vectors. A standard neural network is used as the prediction model.
- S-LSTM: The same vector representations as S-NN model, except to use LSTM as the prediction model instead of NN.
- Hier-NN: Leverage a hierarchical attention network (HAN) analogous with the one in Yang [35]. We consider each day as each document, news collection in one day as sentences in each document. We first build representations of news, then aggregate them into a day representation. Word-level attention layer and a new-level attention are added. A standard neural network is used to make predictions.
- Hier-LSTM: The same encoding and HAN's structure embeddings as Hier-NN model except to use LSTM as the prediction model
- S-GICF: The multiple instance learning framework proposed in [26]. According to [26], an objective function named Group-Instance Cost Function (GICF) is formulated. GICF includes inference of instance-level labels based on instance similarity and bag label constraints. News(instance) is represented through performing a mean operator on word vectors within the news. The parameter λ in GICF is set the same in [26] as 0.04.

- ATT-GICF: Same multiple instance learning setup and GICF function as in S-GICF except to represent news (instance) by encoding news titles through Bi-LSTM and attention mechanism.

- MIL-S: News (instance) is represented through performing a mean operator on word vectors within the news. On top of that, the multiple instance learning model is exploited. News (instance)-classifiers are set to infer class possibilities of each single news. Vector representation and the inferred trend probability of its component news are used to construct the day (bag) representation.

- MIL-Encode: Our proposed model. Encoding each piece of news (instance) using Bi-LSTM and attention mechanism. Additionally, news (instance)-classifiers are set to infer class possibilities of each individual news. Vector representation and the inferred trend probability of its component news are used to construct the day (bag) representation.

To demonstrate if our approach can compete with the best state-of-the-art methods on the benchmark dataset, we utilize the classification accuracy as the evaluation. Table 3 shows the results when compared to the aforementioned baselines in the previous literature.

Table 3. Final results on the test dataset

Models	**Test Accuracy (Max)**
BW-SVM	56.38%
E-NN	58.83%
EB-NN	62.84%
S-NN	57.92%
S-LSTM	60.79%
Hier-NN	57.59%
Hier-LSTM	61.93%
S-GICF	58.52%
ATT-GICF	59.09%
MIL-S	60.23%
MIL-Encode	**63.06%**

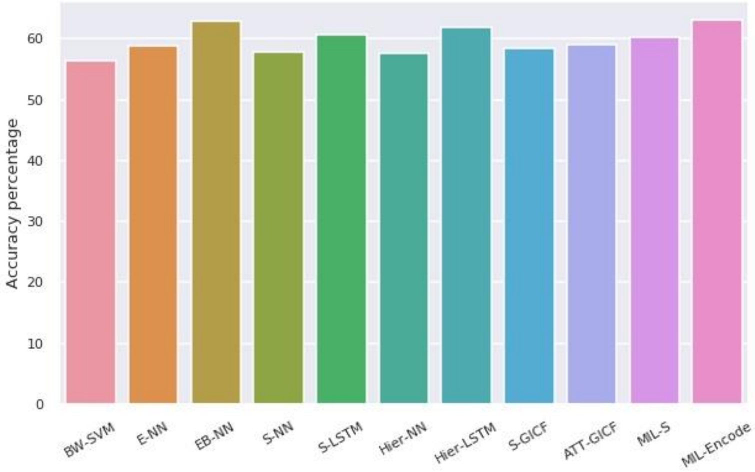


Figure 5. Performances in different models

4.3. Discussion

Through all the experiments above, our proposed framework, MIL-Encode, outperforms all baseline models with a predominant accuracy of 63.06% (see in Table 3 and Figure 5). Although event embedding (EB) in EB-NN is competitive, our proposed multiple instance learning framework is still powerful in the extraction of financial news for stock trend forecasting. By making the comparison, we conclude the following discussions with respect to the following aspects:

Comparison of different news representations: For the representation of news, we use simple averaging embedding (S) and encoding with Bi-LSTM, attention mechanism (ATT) to get news vectors. For MIL models (S-GICF, ATT-GICF, MIL-S, MIL-Encode), the comparison of the results among S-GICF vs ATT-GICF, MIL-S vs MIL-Encode, leads to the conclusion that better performances can be derived from the news(instance) representation of Bi-LSTM and attention (ATT) encoding, than the simple average embedding. In terms of the non-MIL method, ATT encoding performs marginally better than simple averaging as input of the model, by comparing results of S-LSTM vs Hier-LSTM. Although something special in Hier-NN for case S-NN vs Hier-NN, we will discuss it later. It is not hard to see that the ATT representation for news is conducive to predicting the stock trend, especially in MIL models. This can be explained by that the input to the model is organized in sequential contexts through Bi-LSTM. Keywords that show important trend signals in the news title are greatly extracted by the attention mechanism.

Comparison of different multi-instance learning frameworks: There are two kinds of multiple instance frameworks we adopt in the proposed tasks. One is MIL model that aggregates instance labels with Group-Instance Cost Function (GICF) [26], another is our proposed MIL framework that encodes instance labels. The difference between ours and MIL with GICF function is whether to process instance vectors in bag prediction and the manner we handle inferred instance labels to bag prediction.

The MIL methods with GICF cost function are reflected in S-GICF and ATT-GICF models. News(instance) similarity is used in GICF function to read the combination possibility and assign news(instance) labels in a day(bag), based on the intuition that news(instances) pairs with higher similarity are more likely to assign the same labels. Next, an aggregation function is selected to gather instance labels as the anticipated day(bag) labels. We represent instances and calculate pair similarity through both averaging and ATT-encoding approaches. However, the results in S-GICF and ATT-GICF are modest, with less than 60% accuracy. When compared to non-MIL methods, it also performances not as good as the LSTM structures in S-LSTM, Hier-LSTM. For the reasons of modest performances in MIL with GICF function, this method is instance-label-oriented and does not perform further vector operations at the bag level. The averaging aggregation function has a limited generalization capacity. There is a chance that the model was not properly trained. The averaging aggregation of inferred instance labels may omit some potential information, resulting in the failure to identify complex patterns in the proposed task.

In Group-Instance Cost Function (GICF), bag label is obtained by aggregating instances labels like averaging, maximum operation, and the instances vectors are not further processed anymore. In our proposed MIL construction, we adopt a joint representation of a bag instead of gathering labels of instances. The probability of the class to which each instance(news) belongs, as predicted by instance-level classifiers, is used to encode into the bag vectors. We encode the possibility of instance and still process instance vectors into the bag representation, then build the bag-level classifiers to get bag labels. As a result, our proposed model can not only establish bag-level classifiers, allowing for full, end-to-end training, but also include instance features of news uncertainty. According to the results of MIL-S and MIL-Encode, this embedding manner improves the fitting effects of financial news. The learning and generalization capabilities of our MIL models (MIL-S and MIL-Encode) have been enhanced, surpassing not only the S-GICF and ATT-GICF models but also other non-MIL baselines.

Discussion on hierarchical attention in Hier-NN: Hier-NN model has hierarchical structures, embed vectors in common with ours. To distinguish our model from the Hier-NN model, we encode the predicted class probability of news(instance) into the day(bag) representations. We put up the news-level classifiers in order to obtain the expected class probability for each news text. In contrast to our method, Hier-NN keeps encoding the day vectors via GRU and attention mechanisms on news vectors, without creating classifiers at the news-level. In compliance with HAN’s structure, we presume that there exists a significant association between each news item published on the same day. All daily news items are highly context dependent. However, the assumption can be too strong for news, resulting in inferior outcomes during model training. Since in real life, the news is relatively independent of each other. There is no semantic, context relationship between news pairs on the same day as there is between sentences in one document. Therefore, HAN’s structure works better in application of document modeling than financial news analysis. In our model, we lessen this sensitivity and treat the news each day as relatively independent variables, with their effects being related and synergistic to the day’s prediction performances. The experiment results show that our approach has the merits of high learning efficiency, high classification accuracy, and high generalization capability.

5. CONCLUSION AND FUTURE WORK

In this paper, we present the challenges associated with predicting stock movement in light of financial news, particularly, the uncertainty of financial news in terms of random occurrences and unknown individual labels composition. To address these issues, we propose to adopt the principle of multi-instance learning and solve the problem of news-oriented stock trend forecasting from the perspective of multiple instance learning. With this end in view, an adaptive, end-to-end MIL framework is developed in this paper with the goal of improving results. Experimental results on S&P 500 index demonstrate that our proposed model is powerful and effectively increases performance. Nowadays, multi-instance learning is receiving moderate popularity for its applicability in learning problems with label ambiguity. In a variety of application scenarios, some supervised learning methods have also been included or extended to the MIL setting. Nevertheless, there is still much need to explore in the field of financial forecasting for multi-instance learning, which could be a direction for our future research. In addition, in this paper, we only focus on the impact of news on stock trends. Taking more online data resources from social media into account in the trend prediction model is also a potential research site in the future.

REFERENCES

[1] Tsai, M.-C., Cheng, C.-H., Tsai, M.-I., & Shiu, H.-Y. (2018). Forecasting leading industry stock prices based on a hybrid time-series forecast model. PloS one, 13(12), e0209922.

[2] Gao, T., & Chai, Y. (2018). Improving stock closing price prediction using recurrent neural network and technical indicators. Neural computation, 30(10), 2833-2854.

[3] Nti, I. K., Adekoya, A. F., & Weyori, B. A. (2020). A systematic review of fundamental and technical analysis of stock market predictions. Artificial Intelligence Review, 53(4), 3007-3057.

[4] Hiransha, M., Gopalakrishnan, E. A., Menon, V. K., & Soman, K. P. (2018). NSE stock market prediction using deep-learning models. Procedia computer science, 132, 1351-1362.

[5] Luo, Z., Guo, W., Liu, Q., & Zhang, Z. (2021). A hybrid model for financial time -series forecasting based on mixed methodologies. Expert Systems, 38(2), e12633.

[6] Qiu, J., Wang, B., & Zhou, C. (2020). Forecasting stock prices with long-short term memory neural network based on attention mechanism. PloS one, 15(1), e0227222.

[7] Antoniou, A., Holmes, P., & Priestley, R. (1998). The effects of stock index futures trading on stock index volatility: An analysis of the asymmetric response of volatility to news. The Journal of Futures Markets (1986-1998), 18(2), 151.

[8] Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. The Journal of finance, 62(3), 1139-1168.

[9] Hussain, S. M., & Omrane, W. B. (2021). The effect of US macroeconomic news announcements on the Canadian stock market: Evidence using high-frequency data. Finance Research Letters, 38, 101450.

[10] Kumar, B. S., & Ravi, V. (2016). A survey of the applications of text mining in financial domain. Knowledge-Based Systems, 114, 128-147.

[11] Yoshihara, A., Seki, K., & Uehara, K. (2016). Leveraging temporal properties of news events for stock market prediction. Artif. Intell. Res., 5(1), 103-110.

[12] Xing, F. Z., Cambria, E., & Welsch, R. E. (2018). Natural language based financial forecasting: a survey. Artificial Intelligence Review, 50(1), 49-73.

[13] Velay, M., & Daniel, F. (2018). Using NLP on news headlines to predict index trends. arXiv preprint arXiv:1806.09533.

[14] Minh, D. L., Sadeghi-Niaraki, A., Huy, H. D., Min, K., & Moon, H. (2018). Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. Ieee Access, 6, 55392-55404.

[15] Eck, M., Germani, J., Sharma, N., Seitz, J., & Ramdasi, P. P. (2021). Prediction of Stock Market Performance Based on Financial News Articles and Their Classification. In Data Management, Analytics and Innovation (pp. 35-44). Springer, Singapore.

[16] Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the international AAAI conference on web and social media (Vol. 8, No. 1, pp. 216-225).

[17] Keeler, J. D., Rumelhart, D. E., & Leow, W. K. (1991). Integrated segmentation and recognition of hand-printed numerals (pp. 557-563). Microelectronics and Computer Technology Corporation.

[18] Dietterich, T. G., Lathrop, R. H., & Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence, 89(1-2), 31-71.

[19] Sudharshan, P. J., Petitjean, C., Spanhol, F., Oliveira, L. E., Heutte, L., & Honeine, P. (2019). Multiple instance learning for histopathological breast cancer image classification. Expert Systems with Applications, 117, 103-111.

[20] Khatibi, T., Shahsavari, A., & Farahani, A. (2021). Proposing a novel multi-instance learning model for tuberculosis recognition from chest X-ray images based on CNNs, complex networks and stacked ensemble. Physical and Engineering Sciences in Medicine, 44(1), 291-311.

[21] Zhu, W., Sun, L., Huang, J., Han, L., & Zhang, D. (2021). Dual Attention Multi-Instance Deep Learning for Alzheimer’s Disease Diagnosis with Structural MRI. IEEE Transactions on Medical Imaging.

[22] Pappas, N., & Popescu-Belis, A. (2017). Explicit document modeling through weighted multiple-instance learning. Journal of Artificial Intelligence Research, 58, 591-626.

[23] Wang, W., Ning, Y., Rangwala, H., & Ramakrishnan, N. (2016, October). A multiple instance learning framework for identifying key sentences and detecting events. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (pp. 509-518).

[24] Wang, Y., Li, J., & Metze, F. (2019, May). A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 31-35). IEEE.

[25] He, W., & Wang, Y. (2009, September). Text representation and classification based on multi-instance learning. In 2009 International Conference on Management Science and Engineering (pp. 34-39). IEEE.

[26] Kotzias, D., Denil, M., De Freitas, N., & Smyth, P. (2015, August). From group to individual labels using deep features. In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 597-606).

[27] Pappas, N., & Popescu-Belis, A. (2014, October). Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis. In Proceedings of the 2014 Conference on Empirical Methods In Natural Language Processing (EMNLP) (pp. 455-466).

[28] Angelidis, S., & Lapata, M. (2018). Multiple instance learning networks for fine-grained sentiment analysis. Transactions of the Association for Computational Linguistics, 6, 17-31.

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

[30] Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-thought vectors. In Advances in neural information processing systems (pp. 3294-3302).

[31] Vargas, M. R., De Lima, B. S., & Evsukoff, A. G. (2017, June). Deep learning for stock market prediction from financial news articles. In 2017 IEEE international conference on computational intelligence and virtual environments for measurement systems and applications (CIVEMSA) (pp. 60-65). IEEE.

[32] Akita, R., Yoshihara, A., Matsubara, T., & Uehara, K. (2016, June). Deep learning for stock prediction using numerical and textual information. In 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS) (pp. 1-6). IEEE.

[33] Ding, X., Zhang, Y., Liu, T., & Duan, J. (2014, October). Using structured events to predict stock price movement: An empirical investigation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1415-1425).

[34] Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015, June). Deep learning for event-driven stock prediction. In Twenty-fourth international joint conference on artificial intelligence.

[35] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016, June). Hierarchical attention networks for document classification. In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies (pp. 1480-1489).

[36] Hu, Z., Liu, W., Bian, J., Liu, X., & Liu, T. Y. (2018, February). Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In Proceedings of the eleventh ACM international conference on web search and data mining (pp. 261-269).

[37] Liu, H. (2018). Leveraging financial news for stock trend prediction with attention-based recurrent neural network. arXiv preprint arXiv:1811.06173.

[38] Ma, Y., Zong, L., & Wang, P. (2020). A novel distributed representation of news (drnews) for stock market predictions. arXiv preprint arXiv:2005.11706.

[39] Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. EMNLP.

[40] Lu, W., Duan, Y., & Song, Y. (2020, December). Self-Attention-Based Convolutional Neural Networks for Sentence Classification. In 2020 IEEE 6th International Conference on Computer and Communications (ICCC) (pp. 2065-2069). IEEE.

[41] Guo, B., Zhang, C., Liu, J., & Ma, X. (2019). Improving text classification with weighted word embeddings via a multi-channel TextCNN model. Neurocomputing, 363, 366-374.

[42] Lai, S., Xu, L., Liu, K., & Zhao, J. (2015, February). Recurrent convolutional neural networks for text classification. In Twenty-ninth AAAI conference on artificial intelligence. [42] Kotzias, D., Denil, M.,

[43] Li, S., Zhang, Y., & Pan, R. (2020). Bi-directional recurrent attentional topic model. ACM Transactions on Knowledge Discovery from Data (TKDD), 14(6), 1-30.

[44] Huynh, H. D., Dang, L. M., & Duong, D. (2017, December). A new model for stock price movements prediction using deep neural network. In Proceedings of the Eighth International Symposium on Information and Communication Technology (pp. 57-62).

[45] Xu, Y., Lin, W., & Hu, Y. (2020, December). Stock Trend Prediction using Historical Data and Financial Online News. In 2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom) (pp. 1507-1512). IEEE.

[46] Liu, X., Huang, H., Zhang, Y., & Yuan, C. (2020). News-driven stock prediction with attention-based noisy recurrent state transition. arXiv preprint arXiv:2004.01878.

[47] Li, Q., Tan, J., Wang, J., & Chen, H. (2020). A multimodal event-driven lstm model for stock prediction using online news. IEEE Transactions on Knowledge and Data Engineering.

[48] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).