Hierarchical Complementary Attention Network for Predicting Stock Price Movements with News

Qikai Liu, Xiang Cheng, Sen Su, Shuguang Zhu
State Key Laboratory of Networking and Switching Technology
Beijing University of Posts and Telecommunications, Beijing, China
{liuqikai,chengxiang,susen,zsg1990ok}@bupt.edu.cn

ABSTRACT

It has been shown that stock price movements are influenced by news. To predict stock movements with news, many existing works rely only on the news title since the news content may contain irrelevancies which seriously degrade the prediction accuracy. However, we observe that there is still useful information in the content which is not reflected in the title, and simply ignoring the content will result in poor performance. In this paper, taking advantage of neural representation learning, we propose a hierarchical complementary attention network (HCAN) to capture valuable complementary information in news title and content for stock movement prediction. In HCAN, we adopt a two-level attention mechanism to quantify the importances of the words and sentences in a given news. Moreover, we design a novel measurement for calculating the attention weights to avoid capturing redundant information in the news title and content. Experimental results on news datasets show that our proposed model outperforms the state-of-the-art techniques.

CCS CONCEPTS

• **Information systems → Decision support systems**; • **Computing methodologies → Natural language processing**;

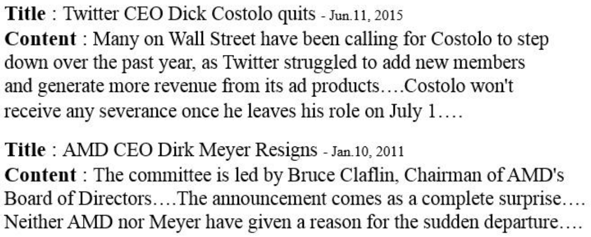
KEYWORDS

Stock prediction; News representation; Attention mechanism; Neural network

ACM Reference Format:
Qikai Liu, Xiang Cheng, Sen Su, Shuguang Zhu. 2018. Hierarchical Complementary Attention Network for Predicting Stock Price Movements with News. In *The 27th ACM International Conference on Information and Knowledge Management (CIKM '18), October 22–26, 2018, Torino, Italy.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3269206.3269286

1 INTRODUCTION

According to the Efficient Market Hypothesis (EMH) [5], stock price movements are thought to be related to news. The reason lies in that news information can affect stock investors' decisions which lead to stock price changing. Meanwhile, news are freely available in a large amount with the fast development of the Internet. Utilizing news articles to predict stock price movements has drawn much attention in recent years.


Title : Twitter CEO Dick Costolo quits - Jun.11, 2015
Content : Many on Wall Street have been calling for Costolo to step down over the past year, as Twitter struggled to add new members and generate more revenue from its ad products….Costolo won't receive any severance once he leaves his role on July 1….

Title : AMD CEO Dirk Meyer Resigns - Jan.10, 2011
Content : The committee is led by Bruce Claflin, Chairman of AMD's Board of Directors….The announcement comes as a complete surprise…. Neither AMD nor Meyer have given a reason for the sudden departure….

Figure 1: Two news about the CEO's departures from *twitter Inc.* and *Advanced Micro Devices Inc.*

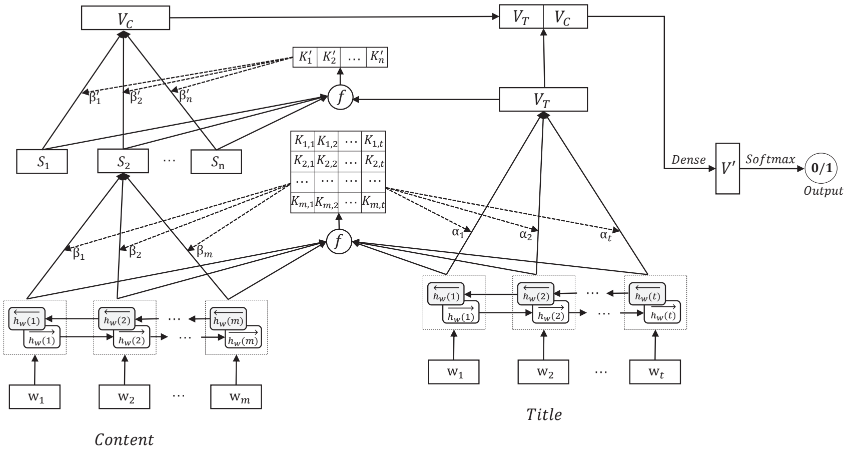
A number of works have been proposed, which use news to predict stock price movements. In particular, Xie et al. [13] use semantic frames to help train predictive models for detecting the (positive and negative) roles of specific companies. Based on ensemble learning, Akhtar et al. [1] combine deep learning and classical feature based models for financial sentiment analysis. Li et al. [10] propose to combine news representation and public moods. To capture the key event in the news title, Ding et al. [4] adopt Open Information Extraction (Open IE) techniques and employ nonlinear models for event-based stock price movement prediction. Hu et al. [6] utilize the attention mechanism to capture information in news for news-oriented stock trend prediction.

When reading news, people focus on the valuable information rather than the whole news. Therefore, many existing works [1, 4] which utilize news to predict stock movements rely only on the news title, since the title is the core of news and there are many irrelevancies in the content which seriously degrade the prediction accuracy. However, we found that there is still useful information in the content which is not reflected in the title. Figure 1 shows two news about the CEO's departures from *twitter Inc.* and *Advanced Micro Devices Inc.*, respectively. If we consider only the title, there is no difference between these two news. However, if we take the reasons and influences of the CEO's departures in the contents into consideration, the two news' attitudes are quite different. In particular, the first news' attitude is positive while the second news' attitude is negative. Due to the influences of these two news, twitter's stock rose, whereas AMD's stock fell. Therefore, utilizing only the title of news for stock movement prediction is insufficient and results in poor performance.

In this paper, we attempt to utilize both the title and content of news to accomplish stock movement prediction. Although there are some works [6, 10, 13] which take both the news title and content into consideration, they treat every word in the news content equally, and thus cannot reduce the noise in the content. Taking

CIKM '18, October 22–26, 2018, Torino, Italy

ACM ISBN 978-1-4503-6014-2/18/10…$15.00
https://doi.org/10.1145/3269206.3269286



V_C
V_T V_C
K'_1 K'_2 ⋯ K'_n
β'_1 β'_2 β'_n
f
V_T
S_1 S_2 ⋯ S_n
$K_{1,1}$ $K_{1,2}$ ⋯ $K_{1,t}$
$K_{2,1}$ $K_{2,2}$ ⋯ $K_{2,t}$
$K_{m,1}$ $K_{m,2}$ ⋯ $K_{m,t}$
Dense
V'
Softmax
0/1
Output
β_1 β_2 β_m
α_1 α_2 α_t
f
$h_w(1)$ $h_w(2)$ $h_w(m)$
$h_w(1)$ $h_w(2)$ $h_w(t)$
w_1 w_2 ⋯ w_m
w_1 w_2 ⋯ w_t
Title
Content

Figure 2: Architecture of Hierarchical Complementary Attention Network (HCAN)

the two news in Figure 1 as examples, the two sentences *"Costolo won't receive any severance once he leaves his role on July 1"* and *"The committee is led by Bruce Claflin, Chairman of AMD's Board of Directors"* have little impact on stock price changes. To address this problem, the crux is how to capture the valuable information in the news content which is not provided by the news title.

To this end, taking advantage of neural representation learning [2, 9, 14], we propose a new neural network architecture, namely Hierarchical Complementary Attention Network (HCAN). In particular, HCAN adopts a two-level attention mechanism, which includes a word-level attention and a sentence-level attention, to quantify the importances of the words and sentences in a given news. To ensure that the information captured in the title and content is complementary, we introduce a new measurement called score-inverse similarity (*S-IS*) for calculating the attention weights.

We evaluate the performance of our HCAN on news datasets. Experimental results show that HCAN outperforms the state-of-the-art techniques.

2 HIERARCHICAL COMPLEMENTARY ATTENTION NETWORK

Figure 2 shows the architecture of the proposed model, namely Hierarchical Complementary Attention Network (HCAN). In HCAN, we first encode each word in a news by Bi-GRU. After getting the word embeddings, we then use the word-level attention to obtain the weights of each word in the news title and content, and use the weighted average values of word vectors to derive the title vector and the sentence vectors of the content. Next, we use the sentence-level attention to obtain the weights of each sentence in the content, and use the weighted average values of sentence vectors to derive the content vector. Finally, we combine the vectors of the news title and content to obtain the representation of the news for stock movement prediction.

2.1 Bi-GRU for Word Embedding

For a given news article, suppose its title contains t words, while its content contains n sentences and each sentence contains m words. Obviously, the news content contains $m * n$ words, and we use q to denote $m * n$ for convenience. The gated recurrent unit (GRU) [3] is a popular type of recurrent neural network. We use a bidirectional GRU to finetune the pre-trained word embeddings. In particular, given the pre-trained word embedding of the i-th word w_i, we concatenate the hidden states of the forward-GRU and backward-GRU to obtain the word embeddings.

$$\overrightarrow{h_w(i)} = \overrightarrow{GRU}(w_i), \quad \overleftarrow{h_w(i)} = \overleftarrow{GRU}(w_i) \tag{1}$$

$$h_w(i) = [\overrightarrow{h_w(i)}, \overleftarrow{h_w(i)}] \tag{2}$$

where $h_t(j) \in R^{2d}$ ($j \in [1,t]$) and $h_c(i) \in R^{2d}$ ($i \in [1,q]$) represent the words in the title and content, respectively, and d is the dimension of the unidirectional GRU.

2.2 Word-Level Attention

After getting the word embeddings of the news title and content, we adopt the attention mechanism in word-level to quantify the importance of each word in the news title and content. Specifically, we design a measurement called score-inverse similarity (*S-IS*) for calculating the attention matrix $K \in R^{q*t}$,

$$K_{i,j} = f(h_c(i), h_t(j)), i \in [1,q], j \in [1,t] \tag{3}$$

Given the i-th word $h_c(i)$ in the content and the j-th word $h_t(j)$ in the title, we use Eq. (4) to calculate $f(h_c(i), h_t(j))$.

$$f(h_c(i), h_t(j)) = \frac{Score(h_c(i), h_t(j))}{Sim(h_c(i), h_t(j))} \tag{4}$$

In the above equation, the function *Score* measures the correlations between different words in the content and title, which is defined as follows.

$$Score(h_c(i), h_t(j)) = h_c(i) \cdot W_1 \cdot h_t(j)^T \tag{5}$$

where $W_1 \in R^{2d*2d}$ denotes the internal weight matrix, learned in the training process, and $h_t(j)^T$ denotes the transpose of $h_t(j)$.

In addition, the function *Sim* measures the degree of similarity between any two words, which is defined as follows.

$$Sim(h_c(i), h_t(j)) = \frac{h_c(i) \cdot h_t(j)^T}{||h_c(i)|| \cdot ||h_t(j)||} \tag{6}$$

where $||x||$ is the L2-norm of vector x.

By introducing the function Sim, we achieve our aim to focus on the complementary information in the title and content.

Based on *S-IS*, we get the attention matrix K. We utilize K to generate the attention vectors α and β for words in the title and content respectively and use them to derive the representations of the title and each sentence in the content.

For the title, it can be seen as a sentence, which contains t words. To get the word attention weights of the title, we calculate the word attention vector $\alpha \in R^t$ as follows.

$$\alpha = softmax(\frac{1}{q}\sum_{i=1}^{q} K_{i,1}, ..., \frac{1}{q}\sum_{i=1}^{q} K_{i,t}) \tag{7}$$

Then, we can get the representation V_T of the title by:

$$V_T = \sum_{j=1}^{t} \alpha_j \cdot h_t(j) \tag{8}$$

For the content, there are n sentences, each of which contains m words. To get the word attention weights of the content, we calculate the word attention vectors β as follows.

$$\beta_l = softmax(\frac{1}{t}\sum_{j=1}^{t} K_{lm-m+1,j}, ..., \frac{1}{t}\sum_{j=1}^{t} K_{lm,j}) \tag{9}$$

where $\beta_l \in R^m$ ($l \in [1, n]$) denotes the word attention vector of the l-th sentence in the content.

After getting the word attention vector of the l-th sentence, we can use it to derive the the representation S_l of the l-th sentence by:

$$S_l = \sum_{i=1}^{m} \beta_{li} \cdot h_{cl}(i), l \in [1, n] \tag{10}$$

where $h_{cl}(i)$ denotes the i-th word in the l-th sentence.

2.3 Sentence-Level Attention

Through the word-level attention, we get the news title representation V_T and the content sentence representation S_l ($l \in [1, n]$). Obviously, in the news content, different sentences also have different contributions. Thus, we further quantify the importance of each sentence in the content by the sentence-level attention.

Similar to the word-level attention, we use the news title and sentences in the content to calculate the attention matrix $K^{'} \in R^{n*1}$:

$$K_l^{'} = f(S_l, V_T), l \in [1, n] \tag{11}$$

where S_l denotes the representation of the l-th sentence in the content and V_T denotes the representation of the title. In particular, the function f in Eq. (11) is the same as the function shown in Eq. (4). The difference is that we use $W_2 \in R^{2d*2d}$ as the internal weight matrix of the function *score*.

Given $K^{'}$, we calculate the sentence attention vector of each sentence in the content by Eq. (12), and derive the representation V_C of the content by Eq. (13).

$$\beta^{'} = softmax(K_1^{'}, ..., K_n^{'}) \tag{12}$$

$$V_C = \sum_{l=1}^{n} \beta_l^{'} \cdot S_l \tag{13}$$

		Train	Test
Time Interval		01/2007 - 12/2011	01/2012 - 12/2012
Number	Positive	18863	4237
	Negative	20435	3952

Table 1: Statistics of Financial News Dataset

2.4 Stock Movement Prediction

After getting the title representation V_T and the content representation V_C, we use them to predict stock price movements.

Specifically, we first concatenate V_T and V_C as a new vector V. Then, we use V to get a vector $V^{'}$ through the *dense* layer of our HCAN. The rationale that adding such a dense layer in HCAN can be explained as follows. Since relations between news and stock price movements are nonlinear [4], adding the dense layer in HCAN can express such relations better. Finally, we use a *softmax* layer to get the *output*, which represents the rise and fall of stock price.

3 EXPERIMENTS

3.1 Datasets

We conduct experiments on real financial news and stock data.

Financial news: We use the financial news dataset obtained from Reuters during January 2007 to December 2012, which contains 47487 news after removing duplicates and empties. In addition, we use the news from year 2007 to 2011 as the training data, and year 2012 as the test data. Table 1 shows the detailed statistics of training and test news numbers. It can be seen that the numbers of positive and negative labels are balanced.

Stock price data: We get the stock price data from Yahoo-Finance. In particular, we collect the stock data of *S&P 500 Index* (Standard & Poor's 500), which is a classical stock market index based on the market capitalizations of 500 large companies.

3.2 Comparison Methods

To evaluate the effectiveness of our HCAN, we compare it against the following methods.

BoW: Bag-of-words [7] is a classical representation method in NLP (Natural Language Processing). We use it to represent news and use SVM as the classifier for prediction.

FastText^{t}: FastText [8] is a open source tool of text classification developed by *facebook* , which can been seen as a one-hidden-layer neural network. FastText^{t} uses only the news title for prediction.

FastText^{t+c}: This method is similar to FastText^{t}, but the difference is that it uses both the news title and content to represent the news for prediction.

Structured-Event: This method [4] captures the structured event in the title, and represents an event tuple as a vector by combining all event elements for prediction.

IAN: This method is extended from [11]. In IAN, it first utilizes the interaction between the news title and content to get the sentence-level representations. Then it simply concatenates all sentences vector as the representation of the news for prediction.

HCAN$^{\#}$: It is a variant of HCAN, where only the function *score* is used to calculate the attention matrices. This comparison method is aimed to test the effectiveness of the proposed measurement *S-IS* for calculating attention weights.

Method	Accuracy
BoW	54.32%
FastText$^{\mathbf{t}}$	58.60%
FastText$^{\mathbf{t+c}}$	57.14%
Structured-Event	58.73%
IAN	59.15%
HCAN$^{\#}$	60.11%
HCAN	**61.38%**

Table 2: Experimental Results

3.3 Experimental Setups

In our experiment, the initial word embeddings are obtained by *word2vec* [12]. We set the dimension of all word embeddings and hidden states of one-way GRU to be 300.

We use the rise (1) and fall (0) of stock price as *output*, which are calculated by closing prices (one day apart). In addition, we use the accuracy of prediction as the evaluation criterion to measure the percentage of correct predictions.

The internal weights in our model are initialized by sampling from the uniform distribution and tuned in the training process. Mini-batch is adopted during the training process, with a batch size of 128.

3.4 Results and Analysis

Table 2 shows our experimental results. Clearly, our HCAN achieves the best performance, boosting the accuracy of prediction compared to other methods.

From Table 2, we can see that the performance of BoW is the worst. This is because it uses only the simple statistics of words in the news and cannot capture the semantic information of the news. In addition, it encounters the curse of dimensionality problem when the vocabulary size is too large.

Both the FastText based methods outperform BoW, since their ability of representing textual information is more powerful than BoW. We notice that FastText$^{\mathbf{t}}$ obtains better performance than FastText$^{\mathbf{t+c}}$. This is because that there is much irrelevant information in the content, causing a bad result. Structured-Event captures the key event in the title, in a sense, it can be seen as a method of reducing noise. However, as mentioned previously, the information contained in the title is very limited. Therefore, compared to FastText$^{\mathbf{t}}$, its performance has only slightly improved.

Compared to our HCAN, IAN implements only the word-level attention and ignores different contributions of sentences. Moreover, it considers only the correlations of the title and content when calculating the attention matrices. Compared to IAN, HCAN$^{\#}$ implements a hierarchical attention mechanism (i.e., the two-level attention mechanism). However, the information captured in the title and content may be redundant by using HCAN$^{\#}$. The information redundancy leads to negative impacts on forecasting. Even so, its ability of capturing the valuable information in the content is better than IAN.

Our HCAN takes a further step to capture the valuable information in the news through a two-level attention mechanism, which effectively reduces the noise in the content. In addition, it can guarantee that the information captured in the title and content is complementary by considering the function Sim when calculating the attention matrices. Therefore, HCAN obtains better vector representations for news, which is the reason why it obtains the best performance.

4 CONCLUSION

In this paper, we propose the Hierarchical Complementary Attention Network (HCAN) to get news representations for predicting stock market movements. In HCAN, we use the two-level attention mechanism to measure the importance of each word and each sentence in the news title and content. Additionally, we capture the complementary information in the news title and content by designing a new measurement, namely score-inverse similarity (*S-IS*), for calculating attention weights. Experimental results show that HCAN improves the quality of news representation and boosts the performance of stock price prediction. In the future, we plan to improve the performance of HCAN by incorporating other type of information, such as news sentiment and social media opinions.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China under Grant 61502047. The corresponding author of this paper is Prof. Xiang Cheng.

REFERENCES

[1] Md Shad Akhtar, Abhishek Kumar, Deepanway Ghosal, Asif Ekbal, and Pushpak Bhattacharyya. 2017. A Multilayer Perceptron based Ensemble Technique for Fine-grained Financial Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 540–546.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[3] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[4] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2014. Using Structured Events to Predict Stock Price Movement: An Empirical Investigation.. In *EMNLP*. 1415–1425.

[5] Eugene F Fama, Lawrence Fisher, Michael C Jensen, and Richard Roll. 1969. The adjustment of stock prices to new information. *International economic review* 10, 1 (1969), 1–21.

[6] Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. 2018. 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*. ACM, 261–269.

[7] Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*. Springer, 137–142.

[8] Armand Joulin, Edouard Grave, and Piotr Bojanowski Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. *EACL 2017* (2017), 427.

[9] Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. *arXiv preprint arXiv:1603.01547* (2016).

[10] Qing Li, TieJun Wang, Ping Li, Ling Liu, Qixu Gong, and Yuanzhu Chen. 2014. The effect of news and public mood on stock movements. *Information Sciences* 278 (2014), 826–840.

[11] Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive Attention Networks for Aspect-Level Sentiment Classification. *arXiv preprint arXiv:1709.00893* (2017).

[12] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[13] Boyi Xie, Rebecca J Passonneau, Leon Wu, GermÃąn G Creamer, Boyi Xie, Rebecca J Passonneau, Leon Wu, and GermÃąn G Creamer. 2013. Semantic Frames to Predict Stock Price Movement. In *Meeting of the Association for Computational Linguistics*. 873–883.

[14] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J Smola, and Eduard H Hovy. 2016. Hierarchical Attention Networks for Document Classification.. In *HLT-NAACL*. 1480–1489.