aiai at the NTCIR-14 FinNum Task: Financial Numeral Tweets Fine-Grained Classification Using Deep Word and Character Embedding-Based Attention Model

Ke Tian[1] and Zi Jun Peng[2]

[1] OPT, Inc., Tokyo Bancho Bldg., 6 Yonbancho, Chiyoda-ku, Tokyo, Japan
[2] Harbin Institute of Technology(Weihai), No.2, Wenhua(W) Road Weihai, Shandong Prov, P.R.China
tianke0711@gmail.com[1],986320586@qq.com[2]

Abstract. This paper describes how we tackle the Fine-Grained Numeral Understanding in Financial Tweet (FinNum) task of NTCIR14. The deep word and character embedding-based attention model is proposed to fine-grained classify the financial numeral tweets. The experiment is shown that the model has good performance and the ensemble result achieved F1-micro and F1 macro of task 1 are 87.41%, 78.04% respectively, and the final F1-micro and F1 macro of task 2 is 80.64%, 73.43 respectively

Team Name. aiai

SubTasks. 7-Category classification (English), 17-(Sub)Category classification (English)

Keywords: NTCIR14, Financial Numeral Tweets, Attention, LSTM, Word Embedding, Character Embedding.

1 Introduction

Finance decisions, such as stock market trading is difficult for investors to be made since the stock market is influenced by a lot of financial indicators. Recently, with the development of financial technologies, stock prediction has attracted a lot of attention from academia as well as financial business, which used the finance technologies, such as natural language processing (NLP) [1], deep learning [2][3][4] et al., to predict the stock market. One of the popular indicators for predicting the stock market movement is online social media (microblog, tweets et al.,). Vilas et al. [5] have checked and measured that twitter could be used as a social sensor for the financial and stock market. Bollen et al. [1] have shown that a predictive correlation between the public mood which are measured by daily twitter feeds and Dow Jones Industrial Average values. Deng et al. [6] have presented that the influence of microblog sentiments on stock returns is both statistically and economically significant. Ranco et al. [7] have shown that the sentiment polarity of Twitter peaks implies the direction of stock returns. Xu et al. [8] used tweets and historical stock price data to predict the stock movement and achieved good performance using the deep model; however, most of the research mainly makes use of tweet sentiment analysis about stock prediction. Chen et al. [9] focus on understanding the meaning of numerals in financial tweet data since they considered that numerals contain crucial information in financial documents, and they are quite important when analyzing the financial instruments as they provide investors with opinion information. Moreover, Chen et al. have opened the financial numeral tweet data with a challenging task named FinNum, which is co-located with NTCIR-14 [10].

FinNum is a task for fine-grained numeral understanding in financial tweet data to identify the category of a numeral in a tweet. In this task, the numerals in financial tweets are classified into seven categories; four of them are further categorized into several subcategories. The details about these two tasks can be obtained in Table 1.

Table 1. Two tasks of FinNum.

Category (Task1)	Subcategory (Task2)
Monetary	money, quote, change, buy price, sell price, forecast, stop loss, support or resistance
Percentage	relative, absolute
Option	exercise price, maturity
Indictor	indicator
Temporal	date, time
Quantify	quantify
Product	product

For example, as the tweet T1, there are 20 and .52 numerical words in the tweet. The categories of these two numbers in the tweet are temporal and monetary, respectively. Moreover, the subcategories of these two numbers are date and quote, respectively.

(T1) $INPX since october 20th when we hit .52 we been going down ever since wtf

We observed that the critical words in tweets influence the category of the tweet. For example, in the above example (T1), the "October" keyword is a date, and it is the category of Temporal. The attention mechanism is useful to detect the weights of words in NLP tasks, the deep word and character embedding-based attention model is proposed to classify the financial numeral tweets of the FinNum task.

Section 2 explains the details of our methods. Section 3 shows experimental configurations and discusses the result. And we conclude this paper in Section 4.

2 Methods

In order to solve the task of FinNum, the deep word and character embedding-based attention model is proposed to classify the finance numeral tweets. The structure of the proposed method is shown in Fig.1. The word embedding and character embedding of the tweet are firstly described in section 2.1, the attention long short-term memory (LSTM) model is described in section 2.2, and the ensemble result is presented in section 2.3.

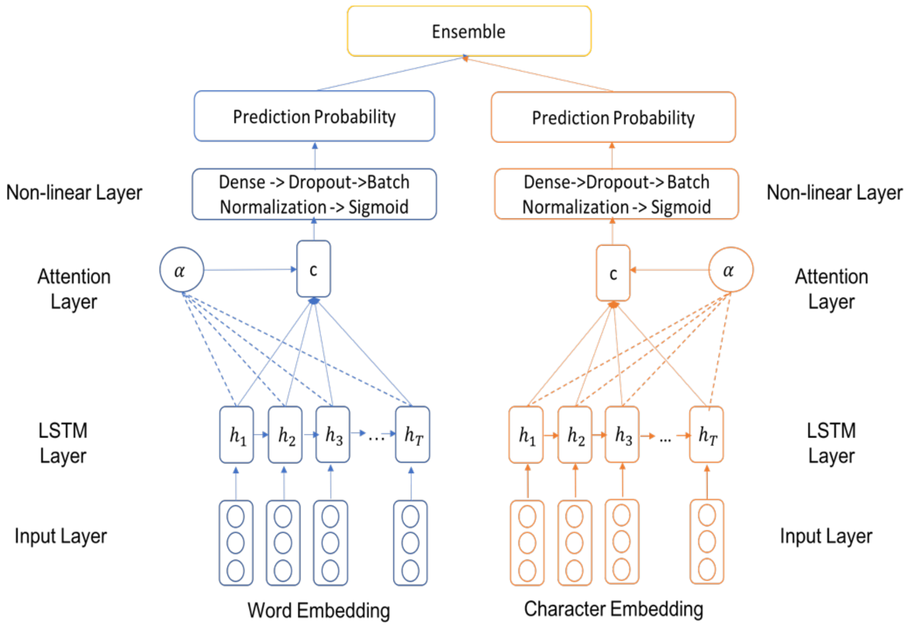

Fig. 1. Deep word embedding and character embedding-based attention model

2.1 Word embedding and Character embedding

There are about 5315 tweets with 11,603 unique words from the training, valid, test tweet data. The CBOW model [11] is taken to train word vectors for the numerical financial domain, and the dimension of word2vec is set to 100. Moreover, we convert the digits (0-9) to the character "D". For example, the date 5/16 will be D/DD and the time 11:00 will be DD:DD, and these two patterns are more likely to be a Temporal category. The word2vec can detect semantic pattern information, such as Percentage, Temporal, etc.

Character embedding is also used to detect the char-based hints for classifying the category or subcategory. For example, as mentioned in [9], char-based vectors can overcome the issue of abbreviations in the tweets. In this paper, character embedding is trained by using the long short-term memory (LSTM) model as shown in Fig. 2. First, there are all the tweet texts provided by the FinNum task to create the character and id dictionary named char2id. Second, the character-based training data [X, Y] is created by the char2id, as the X is the previous three characters before the Y character. Taking the sentence "Hello world" for example, the ["Hel," "l"], ["ell," "o"], ["llo," ‘ w'] et al. character training data can be created. Third, we set the length of the embedding matrix as the length of the char2id, which is named len_char2id, and set all the values as 0. The pretrained character embedding named glove.840B.300d-char.txt [12] is used to be the initialization vector of the embedding matrix, and other characters that are not in glove.840B.300d-char.txt are set to be 0. Fourthly, the embedding dimension is set to be 300; then the [X, Y] training data is taken to train the neural network model. In the neural model, there are two outputs; one is trained by just two layers, and the other is trained by adding hidden layers. Finally, the character embedding is trained by the two neural models.

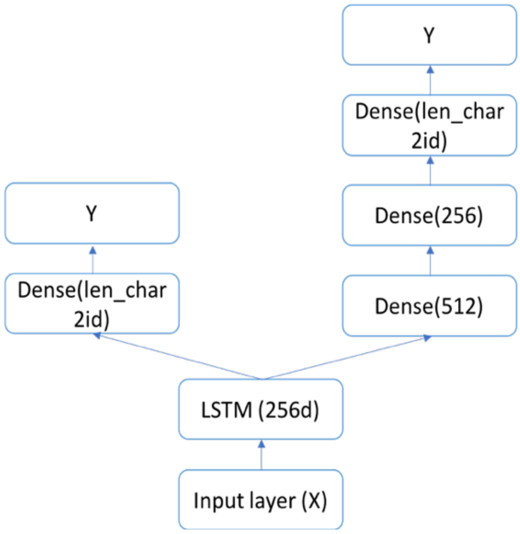

Fig. 2. The processing of training character embedding

2.2 Attention-based LSTM model

Through the numerical financial tweets of train data, we observe that some keywords could help decide the category and subcategory of tweets. For example, "day," "week," "month," "year" indicate the category of temporal category and date subcategory. The percentage category is indicated by the keywords "percent" or "pc." Moreover, "quantity" and "iphone7" to help indicate the category of quantity and product/version respectively. So, some keywords in the tweet have more importance to classify the category or subcategory of tweets. Since the attention mechanism can enable the neural model to focus on the relevant part of your input, such as the words of input tweets, the attention mechanism is used to solve the task. In this paper, we mainly use the feed-forward attention mechanism [13]. The attention mechanism can be formulated with the following mathematical formulation:

$$e_t = a(h_t),\ \alpha_t = \frac{exp(e_t)}{\sum_{k=1}^{T} exp(e_k)},\ \ c = \sum_{t=1}^{T} \alpha_t h_t$$

In the above mathematical formulation, a is a learnable function and only depend on h_t. The fixed-length embedding c of the input sequence computes with an adaptive weighted average of the state sequence h to produce the attention value.

In the structure of the proposed model, as the LSTM layer, the embedding dimension and max word length of word embedding are set to be 100 and 140 respectively as the embedding dimension and max character length of character embedding are set to 300 and 400 respectively. The embedding layer of the word embedding and character embedding matrix as an input layer of LSTM and the size of the output dimension is 300. We used the feed-forward attention mechanism as the attention layer. As the non-linear layer, the activation function is to dense the output of the attention layer to be 256 dimensions, and by using the dropout rate of 0.25, the output result after the dropout rate will be batch normalization. Finally, the sigmoid activation function that will dense the dimension of batch normalization input will be the length of the label as the final output layer.

2.3 Ensemble Result

In the model training stage, the 10-fold cross validation for split training data will be the training mode and will predict the test data. We sum 10 folds of predict probability and get the mean value of 10 folds for the final predict probability result. In the FinNum task, two results are submitted: one result is based on the word embedding while the other result uses the mean value of two to predict probability based on word embedding and character embedding.

3 Experiments

3.1 Tweet preprocessing

The first step is to preprocess tweet data. Since one tweet has several numerals, it also has several categories and subcategories. For example, take the following T10 tweet text: The tweet has two numbers, 10 and 2.29, which are in the Temporal and Monetary categories respectively. To classify the tweet into different categories, the target number will be replaced by the "UNK_NUM," and the label of the tweet is the category of the target number. So the T2 tweet can be converted into two tweets (T3 and T4) and the labels are Temporal and Monetary respectively.

(T2) $DPW calm before the storm. My pT. For 10 am est is 2.29 check back tomorrow
(T3) $DPW calm before the storm. My pT. For UNK_NUM am est is 2.29 check back tomorrow
(T4) $DPW calm before the storm. My pT. For 10 am est is UNK_NUM check back tomorrow

Moreover, the noise words in the tweet have been removed and preprocessed. For example, the URLs such as "http://," emotional signs, @username pattern, #hashtag pattern are replaced with "URL," "EMO_POS," "USER_MENTION," "hashtag" words respectively. Moreover, the number except the target number is replaced with "D." In addition, the punctuation, space, and repeated characters are removed. Finally, the text words are converted into lowercase.

3.2 Experiment Implementation

There are two tasks, task1 is the category classification, and task 2 is the subcategory classification. The attention-LSTM based on word embedding and character embedding respectively have been used to predict the test data. The 10-fold cross-validation to predict the test data is used in the model. The deep model in our research implemented with Keras [14]. Based on the evaluation requirements of the FinNum Task, the micro- and macro-averaged F-scores are taken to evaluate the performance of the proposed model in the paper.

3.3 Result and Discussion

The result of attention-LSTM based on word embedding and character embedding is shown as Table 2. From Table 2, we can obtain that the performance of word embedding is better than the character embedding in task 1 and task 2. We have mentioned that word and character could attention different points in the tweet text, one of submitted result is the mean value of the word embedding and character embedding result. As the ensemble result, the F1-micro and F1 macro of task 1 is 87.41%, and 78.04% respectively. Moreover, the final F1-micro and F1 macro of task 2 is 80.64%, and 73.43 respectively.

Table 1. Experiment Results: micro- and macro-averaged F-score of Two tasks

Embedding	Task 1		Task 2	
	Micro	Macro	Micro	Macro
Word Embedding	86.45%	78.09%	80.24%	74.11%
Character embedding	83.35%	73.53%	67.72%	49.61%
Ensemble result	87.41%	78.04%	80.64%	73.43%

Based on the final report about FinNum task [15], as task1, the ranking of our team is 6 and task2 is 3. The result showed that the proposed model could have good performance. Moreover, combing the word embedding and character embedding may achieve better performance than just using the word embedding and character embedding.

4 Conclusion

This paper mainly discusses how we tackle the FinNum task of NTCIR14. There are two tasks of FinNum that classify the financial numerical tweet text into seven categories and 17 subcategories. Since word and character, embedding can detect the word and char-based hints for classifying the category or subcategory, the deep word, and character-embedding-based attention is proposed to classify the finance numerical tweets. The experimented result showed that combining word embedding and character embedding could effectively classify the financial numeral tweet text into categories and subcategories in the FinNum task.

References

1. Bollen, J.; Mao, H.; Zeng, X. J.: Twitter mood predicts the stock market. Journal of Computational Science 2(1), pp.1–8 (2010).
2. Ding, X., Zhang Y., Liu T., Duan, J.: Deep learning for event-driven stock prediction. International Conference on Artificial Intelligence; pp. 2327-2333, Buenos Aires, Argentina (2015).
3. Hu, Z. N., Liu, W. Q., Bian, J., Liu, X. Z., Liu. T. Y.: 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, Los Angeles, California, USA, pp. 261–269 (2018).
4. Feng, F., Chen, H. M., He, X. N., Ding, J., Sun, M. S., Chua. T. S.: Improving stock movement prediction with adversarial training. ArXiv: 1810.09936 (2018).
5. Vilas, F. A., Redondo, R. P. D., Crockett, K., Owda, M., Evans, L.: Twitter permeability to financial events: an experiment towards a model for sensing irregularities. Multimedia Tools Applications. https://doi.org/10.1007/s11042-018-6388-4 (2018)
6. Deng, S., Huang, Z. J., Sinha, A. P., Zhao, H.: The Interaction Between Microblog Sentiment and Stock Returns: An Empirical Examination. MIS Quarterly 42(3), pp. 895–918 (2018).
7. Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., Mozetič, I.: The Effects of Twitter Sentiment on Stock Price Returns, arXiv Available: http://arxiv.org/abs/1506.02431 (2015).
8. Xu, Y., Cohen, S. B.: Stock movement prediction from tweets and historical prices, Proc. 56th Annual Meeting of the Association for Computational Linguistics, pp. 1970-1979, Melbourne, Australia (2018).
9. Chen, C.C., Huang, H. H., Shiue, Y.T., Chen, H. H.: 2018. Numeral Understanding in Financial Tweets for Fine-grained Crowd-based Forecasting. Proceedings of the 2018 IEEE/WIC/ACM International Conference on Web Intelligence, Santiago, Chile (2018).
10. NTCIR14. http://research.nii.ac.jp/ntcir/ntcir-14/index.html (2019)
11. T, Mikolov., I, Sutskever., Chen, K., G, Corrado., Jeffrey Dean., T, Mikolov., I, Sutskever: Distributed representations of words and phrases and their compositionality. https://arxiv.org/abs/1310.4546
12. Char-embeddings, https://github.com/minimaxir/char-embeddings
13. Raffel, C., Ellis, D. P.: Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems, https://arxiv.org/abs/1512.08756 (2015)
14. Keras: The Python Deep Learning library, https://keras.io
15. Chen, C.C., Huang, H. H., Takamura. H., Chen, H.H.: Overview of the NTCIR-14 FinNum Task: Fine-Grained Numeral Understanding in Financial Social Media Data.

Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo, Japan (2019)