Sentiments classification in stock network public opinion space based on long-short memory convolution neural network

Gaowei Zhang[1,*], *Lingyu* Xu[2] , and *Lei* Wang[1]

[1]Computer Sciences, Shanghai University,200444 , China
[2]East Sea Information Centre, SOA China, 200444, China

Abstract. Deep learning is used to deal with natural language processing problems. Some are based on phrases and some are based on words. This article is inspired by the pixel level in the CV world and therefore retrains the neural network from a character perspective. Neural networks do not need to know about word lookup table or word2vec in advance, and the knowledge of these words is often high-dimensional and it is difficult to apply to convolutional neural networks. In addition, our long-short term memory convolutional neural networks no longer need to know the syntax and semantics in advance. The purpose of this paper is to analyse the investor's psychological characteristics and investment decision-making behaviour characteristics, to study the investor sentiment in the network public opinion space.

1 Introduction

The stock market can be regarded as a group decision-making system, which is restricted by external (online public opinion) and internal (stock itself), so the whole system can be divided into network public opinion space and real stock transaction space corresponding to stock trading. At present, academic circles have not formed a unified consensus on the connotation of investor sentiment and failed to give a normative definition. Investor sentiment is first and foremost an investor's overall perception of market information, which in turn creates a subjective judgment of the market and the company.

The process of emotional judgment is divided into two steps: first, we judge the importance of this matter to ourselves, and then we judge whether it is good or bad. The sources of information that investors pay attention to are mainly divided into two categories. One is objective information, including objective descriptions directly related to the market and companies, such as macro policy environment, market rumours, company performance reports, and corporate strategic planning. The other is subjective information, which mainly refers to the spread of emotions formed by other investors, including institutional investors' judgment of different information.

Most of the information in the network public opinion space is presented in the form of Chinese. The information published by different investors can be regarded as a text. Then

* Corresponding author: yiqigo0215@163.com

the information released by different investors in the network public opinion space constitutes a large number of texts. The extraction of emotional sentiment in network public opinion space is transformed into the classification of texts formed by groups of published information in the network public opinion space. Text classification is a classical theme of natural language processing, in which the predefined categories of texts need to be assigned. In the previous research work on text classification, it included the best feature extraction for text and designed the best classifier for text classification. However, before the work on the best classifiers, many models based on deep learning used higher-level grammar or semantic units to model texts or languages, such as words, phrases, sentence levels or semantics. In this paper, we use characters as the atomic unit of text, and apply convolutional neural networks and long- and short-term memory neural networks to process it. Existing research has shown that convolutional neural networks can be applied directly to distributed or discrete word vectors [1,2,3] without any knowledge of syntax or semantic structures, and this type of approach has been demonstrated to have the same result with traditional models. .

2 Related work

There are mainly two kinds of research on the classification of traditional emotional texts: one is the method based on the dictionary[4-6], and the other is based on the method of statistical machine learning[7-11]. The dictionary-based approach represents works are Lu [12] and Turney [13]. Lu [12] use common sentiment lexicons, such as synonyms in WordNet, antonym information, and grammatical rules to determine the sentiment tendency. Its drawback is that it relies too much on external dictionaries. Turney [13] used the PMI-IR method to calculate the emotional tendency of phrases that fit the rules in texts and judge the polarity of the texts based on the average of these emotional tendencies. Dictionary-based methods are overly dependent on the support of relevant knowledge bases. However, these rules are difficult to describe the uncertainty events, and the compatibility between the rules are difficult to be effectively controlled. The representative works based on statistical machine learning methods are Zou [14] and Mullen [15]. Zou [14] introduced statistical machine learning methods into the commentary classification tasks of movie reviews. In the article he used a number of features including monism, binary words, part-of-speech tagging and other selected Naive Bayes, maximum entropy, support vector machine training model. The experimental results shown that the SVM is the best one, and the one-word feature is chosen. Mullen [15] used the SVM classifier and integrated various characteristics of different sources of information to enhance the classification results.Cui [16] proved that the effect of Uni-gram is the best when the training corpus is small, but the effect of N-gram (n> 3) becomes more and more obvious with the expansion of training corpus. Tan [17] used N-gram as well as nouns, verbs, adjectives and adverbs respectively as text features in sentiment classification of Chinese texts. Their experimental results shown when the training set is large enough and the appropriate number of features are selected, the emotion classification can achieve good results.

3 Character long-short memory convolutional neural network

In this section, we introduce the design of character-level Long-short memory Convolution network for sentiment classification. The model takes an encoded sequence of characters as input. To encode each character, we need to pre-define a character table in the size of the input language and then use one-hot encoding to quantify each character. In this way, the sequence of characters is converted into a sequence of vectors in which the length of each

vector is fixed to the size of the alphabet, and the characters not in the alphabet are converted to an all-zero vector. The network is a 11-layer neural network that contains 6 layers of convolutional layers 2 layers long-short memory layers and 3 layers of fully connected layers. Figure 1 shows the structure of the network used in this paper.

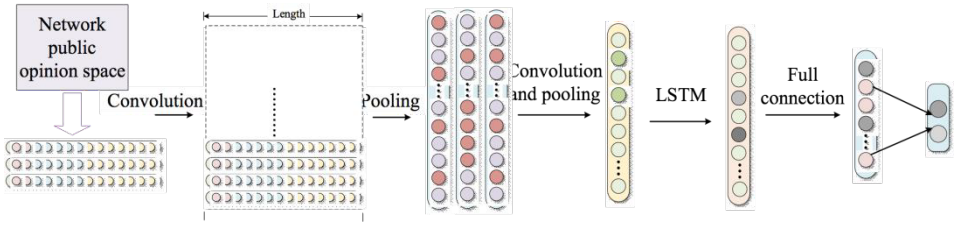

Fig. 1. Character Convolution Neural Network Structure

The core of the model is a $1 \times D$ time convolutional module. Suppose $g(x) \in [1,l] \rightarrow$ [?] is a discrete input function and $f(x) \in [1,k] \rightarrow$ [?] is a discrete kernel function. The convolution function of $f(x)$ and $g(x)$ with step d is defined as $h(y) \in [1,(l-k)/d+1] \rightarrow$ [?],

$$h(y) = \sum_{x=1}^{k} f(x) g(y[?] - x + c) \tag{1}$$

where $c = k - d + 1$ is an offset constant, the model uses the max pooling. The maximum pooling function is defined as $h(y) \in [1,(l-k)/d+1] \rightarrow$ [?],

$$h(y) = \max_{x=1}^{k} g(y[?] - x + c) \tag{2}.$$

4 Experiment

In this article, we set the length of the alphabet to 2000, so the dimension of the input feature is equal to 2000. We study the headline of investor postings in network public opinion space, most of the titles are less than 256 characters, so the features length is set to 256. 2 dropout modules are inserted between 3 fully connected layers for normalization [10], their drop probability is set to 0.5. The convolution layer has a step size of 1. The convolution and pooling layer use "VALID" method, that is, no zero padding is performed. For example, in this paper, the network public opinion information is divided into two categories, then the number of the final output unit of the network is 2.

In order to verify the performance and effectiveness of the emotion classification algorithm designed in this paper, We manually tagged 100,000 posts and randomly divide the entire data set into 50 equally-sized parts, and then use the first one to train the initial classifier and the remaining 49 as a input flow, put into the trained classifier one by one. Select the entry training set accounting for u to update the initial classifier, as shown in figure 2, abscissa represents the number of copies, p represents the accuracy rate of the emotion tendency classification algorithm, r represents the recall rate of the emotional tendency classification algorithm, and f represents the f1 value of the emotion classification algorithm. With the increase in the number of training sets, the p-value is always above 0.84, the r-value is above 0.69, and the f-value is above 0.76. As the number of tagged posts

entering the training set increases, the classifier's accuracy rate, recall rate, and f1 value will continue to increase. Figure 3 shows the results of different algorithms, CLCNN represents the character long-short term memory convolution network, LR represents the logistic regression, RF represents the random forest. We can see that our model is better in p, r, f1.

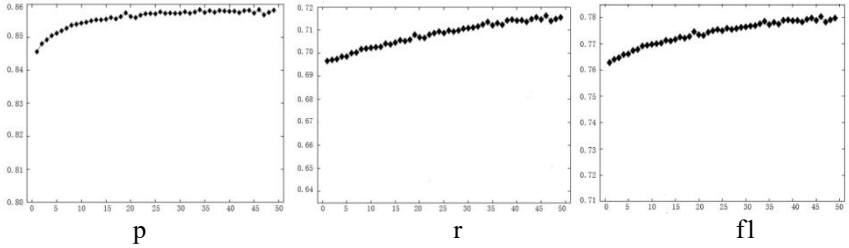

Fig. 2. Network public opinion space information emotion classification effect.

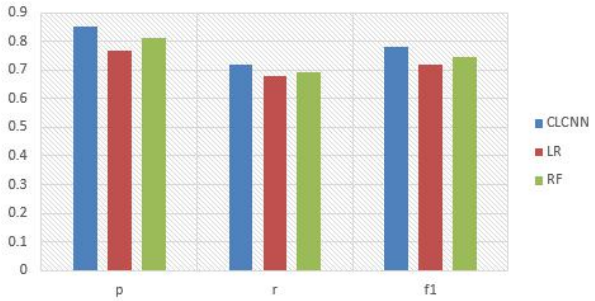

Fig. 3. Comparison of Classification Results of Different Algorithms.

For the 600,000 posts collected in the network public opinion space, we randomly select 12 stocks to extract the emotional sentiment, part of the statistical results shown in figure 4.

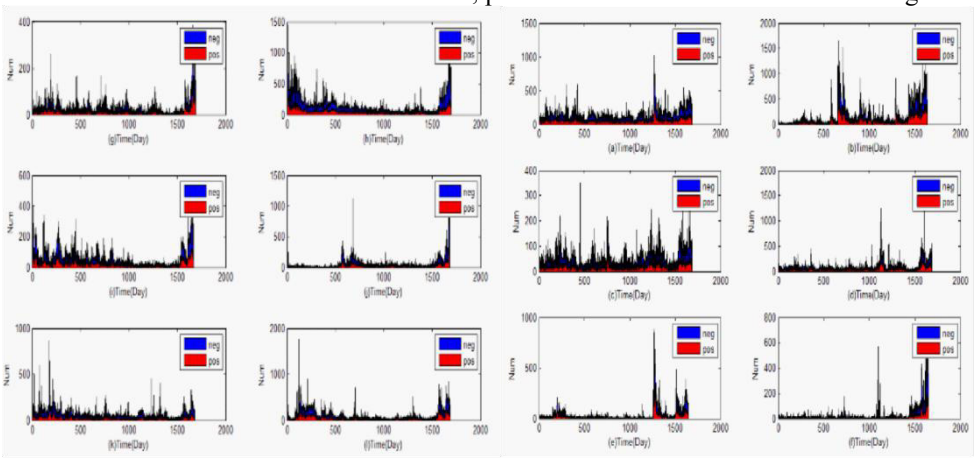

Fig. 4. Network public opinion sentiment tendency results.

We find that negative sentiment is always above the positive side of emotions during the actual trading hours due to the "silent spiral" in sociology. Compared with the real world, the phenomenon of "silent spiral" in the online world is more Prominent: supportive people are often silent, people with strong desire to express are often strongly dissatisfied people. However, this is an illusion in many cases. It cannot be effectively verified in the real trading world, and many times there is a phenomenon of reverse operations. Therefore, the

tendency of real trading cannot be determined simply by the tendency of the media to discuss spatial emotions.

This is due to the resonance effect produced by the similarity of the content of most media reports; The cumulative effect of continuity and repetitiveness of similar information dissemination; The ubiquitous effect of the breadth of information reach. These three characteristics often create an opinion climate for the group, and people fear to be isolated, and they will adopt similar actions on the dominant climate. The result is a spiral process in which one side grows more and more, and the other side goes silent. This explains why the negative emotions in the network public opinion space have always accounted for the majority

5 Conclusion

We first propose a long-short term memory convolutional neural network model based on word vectors, and analyse the emotional orientation of information in the network public opinion space. This method constructs the characteristics of information based on the word vector, constructs the characteristics of information based on the convolutional neural network and classifies the emotional tendency. However, since a lot of information currently contains some pictures to express emotions, the next step is to combine the pictures and texts to improve the accuracy of classification of emotional sentiment in network public opinion space, we also need to consider the information contained in Weibo and WeChat.

Acknowledgement

This paper is supported by Project supported by the National Research Program of China(Grant No. 2016YFC1401902).

References

1. C. dos Santos and M. Gatti. Deep convolutional neural networks for sentiment analysis of short texts. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 69–78(2014).
2. Y. Kim. Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1746–1751(2014).
3. R. Johnson and T. Zhang. Effective use of word order for text categorization with convolutional neural networks. CoRR, abs/1412.1058, 2014.
4. Das A, Bandyopadhyay S. Dr Sentiment knows everything![C]// Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations. Association for Computational Linguistics, 50-55(2011).
5. Pawar K K, Shrishrimal P, Deshmukh R R. Twitter Sentiment Analysis: A Review[J]. 6(2015).
6. Saif H, Fernandez M, He Y, et al. Evaluation Datasets for Twitter Sentiment Analysis. A survey and a new dataset, the STS-Gold[C]// Workshop: Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives From Ai(2013).

7. Ghiassi M, Skinner J, Zimbra D. Twitter brand sentiment analysis: A hybrid system using n -gram analysis and dynamic artificial neural network[J]. Expert Systems with Applications, 40,6266-6282(2013).
8. Zimbra D, Ghiassi M, Lee S. Brand-Related Twitter Sentiment Analysis Using Feature Engineering and the Dynamic Architecture for Artificial Neural Networks[C]// Hawaii International Conference on System Sciences. IEEE, 1930-1938(2016).
9. Pandarachalil R, Sendhilkumar S, Mahalakshmi G S. Twitter Sentiment Analysis for Large-Scale Data: An Unsupervised Approach[J]. Cognitive Computation,7,254-262(2015).
10. Selvi C, Ahuja C, Sivasankar E. A Comparative Study of Feature Selection and Machine Learning Methods for Sentiment Classification on Movie Data Set[M]// Intelligent Computing and Applications. Springer India, 367-379(2015).
11. Socher R, Perelygin A, Wu J Y, et al. Recursive deep models for semantic compositionality over a sentiment treebank[J](2013).
12. Lu Y, Castellanos M, Dayal U, et al. Automatic construction of a context-aware sentiment lexicon:an optimization approach[C]// International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April. DBLP, 347-356(2011).
13. Turney P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews[C]// Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 417-424(2002).
14. Zou H, Tang X, Xie B, et al. Sentiment Classification Using Machine Learning Techniques with Syntax Features[C]// International Conference on Computational Science and Computational Intelligence. IEEE, 175-179(2016).
15. Mullen T, Collier N. Sentiment Analysis using Support Vector Machines with Diverse Information Sources[C]// Conference on Empirical Methods in Natural Language Processing, EMNLP 2004, A Meeting of Sigdat, A Special Interest Group of the Acl, Held in Conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain. DBLP, 412-418(2004).
16. Cui H, Mittal V, Datar M. Comparative experiments on sentiment classification for online product reviews[C]// National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, Usa. DBLP, 61–80(2006).
17. Huifeng, Tan Songbo, Cheng Xueqi . Comparative Study on Chinese Emotion Classification Based on Supervised Learning[J]. Chinese Journal of Information. 21,88-94(2007).