VISUALISATION OF TEXT-BASED DEEP STOCK PREDICTION

Mr.G.DILIP KUMAR

Assistant Professor, Department of CSE, Sri Mittapalli College of Engineering,

Tummalapalem, NH-16, Guntur, Andhra Pradesh, India.

1) A. V. LEELAVATHI, 2) N. SAILAJA, 3) B. SAI TEJASWINI, 4) M. MEGHANA

B. Tech Student, Department of CSE, Sri Mittapalli College of Engineering,

Tummalapalem, NH-16, Guntur, Andhra Pradesh, India.

***Abstract*:** The recent advance of deep learning has enabled trading algorithms to predict stock price movements more accurately. Unfortunately, there is a significant gap in the real-world deployment of this breakthrough. For example, professional traders in their long-term careers have accumulated numerous trading rules, deep learning models have been hardly interpretable. This paper presents DeepClue, a system built to bridge text-based deep learning models and end users through visually interpreting the key factors learned in the stock price prediction model. We make three contributions in DeepClue. First, by designing the deep neural network architecture for interpretation and applying an algorithm to extract relevant predictive factors, we provide a useful case on what can be interpreted out of the prediction model for end users. Second, by exploring hierarchies over the extracted factors and displaying these factors in an interactive, hierarchical visualization interface, we shed light on how to effectively communicate the interpreted model to end users. Specially, the interpretation separates the predictable from the unpredictable for stock prediction through the use of intercept model parameters and a risk visualization design. Third, we evaluate the integrated visualization system through two case studies in predicting the stock price with online financial news and company-related tweets from social media. Quantitative experiments comparing the proposed neural network architecture with state-of-the-art models and the human baseline are conducted and reported. Feedbacks from an informal user study with domain experts are summarized and discussed in details. All the study results demonstrate the effectiveness of DeepClue in helping to complete stock market investment and analysis tasks.

1. INTRODUCTION

Deep learning techniques are reshaping the landscape of predictive analysis in the big data research area and have made major breakthroughs in image and speech recognition , question answering , machine translation and many other application domains. In this paper, we focus on the financial analytics domain. It has been shown that texts such as financial news and tweets on stock markets are useful in predicting stock price movements. For example, financial news such as “Amazon profit beats forecasts” was accompanied with a surge of Amazon’s stock price, while “Oil price hits a record high” triggered worries on the auto industry and weakened their performance in the stock market. Previous work has demonstrated an over 60% accuracy in predicting the daily stock price movement using deep neural networks over a large collection of financial news. For most other domains, there is still little clue on how deep learning models work. In our scenario, the use of text input introduces an additional word embedding stage to map text collections onto the feature space, which makes it more difficult to interpret the prediction model.

In this paper, we target the research problem of how to interpret text-based deep stock prediction model for end users, so that they can make up their stock trading decisions as well as improve the prediction model based on the interpretation. In particular, we investigate research questions including what kind of information can be efficiently extracted from the prediction model as interpretations, and how to communicate such information in an effective way to end users. Throughout this work, we depend on an interactive visualization interface to bridge the

Prediction model and end users, which turns out a natural and straightforward choice. Yet, designing and prototyping such a visualization system can be quite challenging. First, traditional patterns discovered from data can be presented by visually distinct channels in the same data view, while in this case, the information extracted from the model lies in a higher order than the data pattern. Multiple coordinated views should be designed elaborately to illustrate the relationship among data, model, and interpretation. Second, the deep learning model is designed in a bottom-up structure to take advantage of the machine's capability in processing huge amount of data, while the visual information-seeking mantra is "overview first, details on demand". Third, it is commonly accepted that the stock market is information efficient, but not all stock price movements are predictable or reflected in text information. Ingenuity is required to separate predictable and unpredictable price changes. In the literature, there is a recent surge on the topic of visualizing deep neural networks (DNN) for model interpretation. A large portion of these methods focused on the display of neural network architecture to help users understand the functionality of individual neurons and features, interpret the mechanism of both small-scale neural networks and large-scale multi-layer DNNs. Another thread of research proposed to visualize the model output (e.g. the image class model) or their correspondence to the input data through algorithms similar to back propagation. While our study aligns with these successful methods on DNN model interpretation, the goal is fundamentally different. Instead of visually illustrating DNN structures, we target at extracting useful information from the prediction model, and incorporating this interpretation with domain expertise to improve the performance of stock trading and modeling. In addition, existing literature mostly studied model interpretation for image recognition and object detection tasks, while to our knowledge, we are the first to visually interpret the hidden linkage between public text collections and stock prices through deep learning models.

II. METHODLGY

A. Text-Based Stock Price Prediction

Data Collection: We consider S&P 500 stocks in the US stock market. Their historical prices are acquired from Yahoo Finance. We crawled financial news from Reuters and Bloomberg, obtaining in total 341,310 news articles. For each news, we extracted the title, textual content, and timestamp from their raw HTML file. To map each news to the corresponding stocks, we maintained a list of keywords for each firm (e.g., Apple: AAPL, AAPL.O, APPLE, AAPL.N, Apple Inc, etc.). The stock-related tweets were collected through Twitter API in a period from April 2015 to November 2015, by matching the firm's cashtags in the tweet content. Cashtag is a new way of sharing financial information on social media developed by Twitter and other providers. The firm's stock ticker symbols are prefixed with a dollar sign to compose the cashtag, e.g., Apple=$AAPL, Google=$GOOG. In total, we obtained 6,869,771 stock-related tweets. For each tweet, we recorded the create time, textual content, source, user, location and related firms.

B. Deep Neural Network Architecture

We take news data as an example to introduce the architecture of the neural network model adopted in this work. The model is built for each particular S&P 500 firm. The goal of the model is to predict a stock price ˆ y that is close to the real stock price y of the firm. The raw input of each model is the set of financial news titles collected on the target firm. Our proposed deep regression model organized in a hierarchical neural network structure. The network consists of four layers: a word representation layer, a bigram representation layer, a title representation layer, and a feed-forward regression layer. The word representation layer accepts all the news titles as input and turns each word in the title into a real-valued word embedding vector . The bigram representation layer constructs representation vectors for word bigrams based on the representation vector of individual words. The title representation layer summarizes representations of word bigrams and encodes each title into a title vector. The feed-forward regression layer receives the output of the title encoder and maps the output to a real-valued prediction through a feed-forward neural network with residual connections . In addition to the prediction, the proposed model is also optimized for the interpretation purpose by three key designs. First, we explicitly extract representation vectors (i.e., features) from the input news titles in a hierarchical, interpretable way (word →bigram→title), which provides the opportunity to efficiently visualize a large amount of contributing factors. Second, we make use of a combination of techniques to prevent overfitting, e.g. the dropout mechanism. Third, as the hierarchical method lengthens the backward propagation path, we introduce residual connections to ease the burden of training a deep neural network. Note that the proposed deep stock prediction model can be upgraded by introducing state-of-the-art deep neural network structure, such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). We will describe two such alternative designs by replacing the bigram-based title representation method with convolution layers and RNN with LSTM cells.

1) Word Representation Layer

In the word representation layer, we adopt the distributed method proposed by Mikolov et al., which embeds each word into the high-dimensional space as a real-valued dense vector. If two words' representation vectors are close to each other in the high dimensional space, their corresponding words will have similar semantic meanings.

2) Bigram Representation Layer

Over the word representation layer, we capture the information of phrases using bigram, which is the sequence of two adjacent words in a sentence, e.g., "hit record", "draw investors". Consider a bigram B, in which the representation vector of its two words are vi and vj. The representation vector of B, denoted by uB, is computed by uB = tanh(vi +vj). With this design, the bigram representation captures both the semantic of phrases and their composing words, effectively going beyond the original bag-of-words model.

3) Title Representation Layer

The title representation layer maps each news title into a vector and then aggregates all the vectors on the same day into a single representation to summarize the overall input data of that day.

#Days	#Titles	#Words	Training Time (s)
1312	8122	70331	96.33
656	2043	16800	55.33
328	663	4854	44.33
164	311	2253	24.67

TABLE 1: The training time of Deep Clue prediction model (Apple Inc.'s financial news data set).

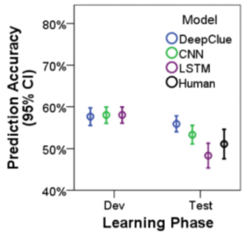
III. RESULTS AND DISCUSSION

Visualization: The DeepClue interface is composed of four coordinated views, as shown in Figure 1(a)(b)(c)(d). We follow two principles in the visualization design. First, the visualization interface should help users complete three key tasks in the scenario of stock price prediction and analysis. Understanding stock market: The base line task is to examine the underlying stock data, including price movements, trading volume, historical rise&fall trends, and the potential temporal patterns. Visualizing prediction result: Over the stock data, users should get access to the result produced by the model, i.e., whether a certain stock is predicted to rise or fall on the next day. S/he also needs to navigate the input data to the prediction model, i.e., the news/tweets collection in our scenario. Interpreting prediction model: Finally, users are expected to unveil the myth of the model by learning why and how each rise&fall prediction is decided. In DeepClue, this is achieved by visualizing the key textual factors that jointly make up the decision. Second, we design DeepClue for financial domain users, i.e., stock traders and investors. These users are mostly accustomed to classical financial visualization interfaces (e.g., Yahoo Finance), especially for the first two tasks in presenting stock data and their predictions. The classical visualization depends heavily on statistical charts. Therefore, to reduce the user's learning cost, we build DeepClue from commodity statistical charts, both in the stock data and prediction visualization (Figure 1(a)(c)) and in illustrating their relevant predictive factors (Figure 1(b)(d)).

Fig. 1: DeepClue interface: (a) stock timeline view showing the stock price history in an overview+detail design; (b) factor hierarchy view displaying the relevant textual factors to the stock prediction in a hierarchical structure; (c) document list view showing the related documents with the selected factor; (d) keyword map view depicting the relationship of relevant keywords.

Model	Phase	Days	Accuracy (%)									
			AAPL	GOOG	BAC	WMT	GM	XOM	BA	GSPC	T	Average
DeepClue	train	1425	83.8	87.8	**93.8**	52.5	83.9	90.3	91.2	87.6	46.2	79.7
	dev	178	56.1	58.7	55.9	59.2	**62.8**	59.4	57.4	53.7	55.6	57.6
	test	177	56.4	58.1	52.4	55.1	56.8	52.4	**60**	56.4	55.8	55.9
LSTM	train	1425	80.8	67.3	74.2	88.8	**95.3**	87.7	76.3	54.9	73.5	77.6
	dev	178	56.1	57.3	57.3	58.7	**63**	56.1	60.3	58.6	54.8	58.0
	test	177	50.6	42	51.1	44.6	50.1	**52.8**	52	47.6	44.2	48.3
CNN	train	1425	90.8	50.6	50.1	49.4	**91.4**	46.5	51	54.8	51.3	59.5
	dev	178	53	58	55.6	60.3	58	57.9	59.8	58.1	**61.3**	58.0
	test	177	54.3	55.3	52	50.8	53.8	**58.5**	54	48	53.4	53.3

(a)



(b)

Fig. 2: The prediction accuracy over the stock price movement of nine companies in S&P 500. The news data is used. (a) The comparison of three alternative models in the training, development and test phases; (b) The corresponding chart of prediction accuracy distribution (95% CI), the result of human prediction is also presented as a baseline.

RESULT

The result is summarized in Figure 2(a), which lists the performance of three alternative models in predicting the stock price movement during the training, development and test phases. On each row of Figure 2(a), nine representative S&P 500 firms are predicted under the same setting, and each cell indicates the ACC of one predictive model over a particular firm. The highest ACC in each row is depicted in bold to high light the top performance among different firms, i.e., 63% in the development phase(LSTM on GM) and 60% in the test phase(DeepClue on BA). Note that the training set accuracy is less useful in representing the model performance, so we do not discuss it in details. In Figure 2(b), the ACCs on each row are treated as a probability distribution and further depicted with error bars (95% confidence interval), grouped by models and learning phases. On the test set, we also depict the distribution of ACCs by human prediction obtained in the controlled user experiment. In general, we have two findings on the experiment result. First, in all cases, DeepClue leads to comparable or even higher prediction accuracy compared with the other two models (i.e., CNN and LSTM). By the analysis of variance (ANOVA) test, there are a few cases that significant differences are observed among these models.

Discussion: Overall, under the current stock price prediction scenario and the financial news data set, DeepClue performs the best among alternative models that have been implemented. CNN and LSTM do not necessarily perform as well as what we have expected. There can be certain deficiencies of these models. For example, CNN introduces additional parameters in the convolution layer that is hard to train with only ~1000 training samples. LSTM that considers longer context than the bigram seems to be inappropriate in predicting the stock price movement, as the stock price is mostly affected by the local feature of the news title, e.g., key event or concept, but not the whole title. We believe that the main reason for the failure of advanced models lies in data noise. First, there are many financial news in the data set which is irrelevant to the price change of the next day. For example, the background report of a firm can appear in any day and includes historical events of the firm that may affect its previous but not the current stock price. Second, there are a lot of cases that the root cause of the price change is not disclosed in the public reports. Therefore, one stock with all positive news in one day can have its stock price drop in the next day. Fitting these noisy cases will potentially downgrade the prediction accuracy, in addition to the overfitting issue on a particular data set. The interactive model analysis by DeepClue is a potential way to address this data noise issue. We have detected several cases that the training data is irrelevant or even misleading with respect to the stock price movement. Third, the delayed release of news, irrationality, and insider trading can also cause noise in the data set. For conceptual simplicity in visualization, we choose the current model structure.

V. CONCLUSION AND FUTURE WORK

We present DeepClue, a system that visually interprets text-based deep learning models in predicting stock price movements. DeepClue integrates three key designs from the cutting-edge deep learning technology:

A hierarchical neural network model that embeds semantics in intermediate processing layers for interpretation; a back propagation-like algorithm that effectively distributes the decision of prediction back to individual documents, bigrams and words; and an interactive visualization

ISSN: 2278-4632
Vol-10 Issue-7 No. 11 July 2020

interface that allows users to navigate and analyze stock price timelines, textual factors, and their correlations.

DeepClue has been deployed to predict S&P 500 stocks using mainstream financial news and firm specific tweets. Both case studies, quantitative experiments, and the informal user study with domain experts demonstrate the usefulness of the proposed system in learning from, evaluating and improving the text-based deep stock prediction models. In future all the study results demonstrate the effectiveness of the deep stock prediction is used in helping to complete stock market investment and analysis tasks.

VI. REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in NIPS'12, 2012, pp. 1097–1105.

[3] A. Bordes, S. Chopra, and J. Weston, "Question answering with subgraph embeddings," in EMNLP'14, 2014, pp. 615–620.

[4] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in ICLR'15, 2015.

[5]X.Ding,Y.Zhang, T.Liu, and J.Duan, "Using structured events to predict stock price movement: An empirical investigation," in EMNLP'14, 2014, pp. 1415–1425.

[6] ——, "Deep learning for event-driven stock prediction," in IJCAI'15, 2015, pp. 2327–2333.

[7] B. G. Malkiel and E. F. Fama, "Efficient capital markets: A review of theory and empirical work," The journal of Finance, vol. 25, no. 2, pp. 383–417, 1970.

[8]"Tensorflow playground," <http://playground.tensorflow.org/>.

[9] M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, and S. Liu, "Towards better analysis of deep convolutional neural networks," IEEE Transactions on Visualization and Computer Graphics, vol. 23, no. 1, pp. 91–100, 2017.

[10] "Twitter cashtag," <http://money.cnn.com/2012/07/31/technology/twitter-cashtag/>.

[11]T.Mikolov, I.Sutskever, K.Chen, G.S.Corrado, and J.Dean, "Distributed representations of words and phrases and their compositionality," in NIPS'13, 2013, pp. 3111–3119.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," arXiv:1512.03385, 2015.

[13] H. Chen, P. De, Y. J. Hu, and B.-H. Hwang, "customers as advisors: the role of social media in financial markets," in BFC'13, 2013.

[14] H. Ziegler, M. Jenny, T. Gruse, and D. A. Keim, "Visual market sector analysis for financial time series data," in VAST'10, 2010, pp. 83–90.

[15] D. A. Keim, T. Nietzschmann, N. Schelwies, J. Schneidewind, T. Schreck, and H. Ziegler, "A spectral visualization system for analyzing financial time series data," in EuroVis'06, 2006, pp. 195–202.

[16] J. Hullman, N. Diakopoulos, and E. Adar, "Contextifier: Automatic generation of annotated stock visualizations," in CHI'13, 2013, pp. 2707–2716.

[17] E. Sorenson and R. Brath, "Financial visualization case study: Correlating financial timeseries and discrete events to support investment decisions," in IV'13, 2013, pp. 232–238.

[18] Wikipedia, "Cross-correlation," <https://en.wikipedia.org/wiki/Cross-correlation>.

[19] K. Cho, B. Van Merri¨enboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.

[20] B. Xie, R. J. Passonneau, L. Wu, and G. G. Creamer, "Semantic frames to predict stock price movement," in ACL'13, 2013, pp. 873–883.

[21] E. F. Fama and K. R. French, "Common risk factors in the returns on stocks and bonds," Journal of financial economics, vol. 33, no. 1, pp. 3–56, 1993.