Semi-supervised sentiment analysis for Chinese stock texts in scarce labeled data scenario and price prediction

Ji Zhaoyan, Yan Hongfei, Ying siping, Chen Chong, Su Qi Peking University

论文摘要

The application of neural network in stock prediction is developing rapidly these years because of its excellency in series data processing. However, as most of research are conducted in English, data sources and labeled data are inadequate in Chinese. Especially for natural language processing tasks in specific domain where specialized labeled data are required to train models to adapt to terminology processing, specialized labeled Chinese in text data are very scarce, such as financial text data. To tackle this challenge, we proposed a semi-supervised learning method to generate well-labeled data and train BERT, a leading natural language processing model, to obtain a trained sentiment machine. Then we got stock-related text data sentiment score based on this machine and further combine the sentiment score and other transaction data as inputs for different neural networks to predict stock price. The experimental results on a large scale of Chinese stock data and texts showed that our proposed method successfully improved prediction accuracy compared to other established methods. Besides, we also examined our method's applicability combined with different neural networks when predicting different types of stock.

系统模型

LSTM; BERT; Semi-supervised learning

论文简介

We proposed a semi-supervised learning method to generate well-labeled data and train BERT, a leading natural language processing model, to obtain a trained sentiment machine. Then we got stock-related text data sentiment score based on this machine and further combine the sentiment score and other transaction data as inputs for different neural networks to predict stock price.

算法原理
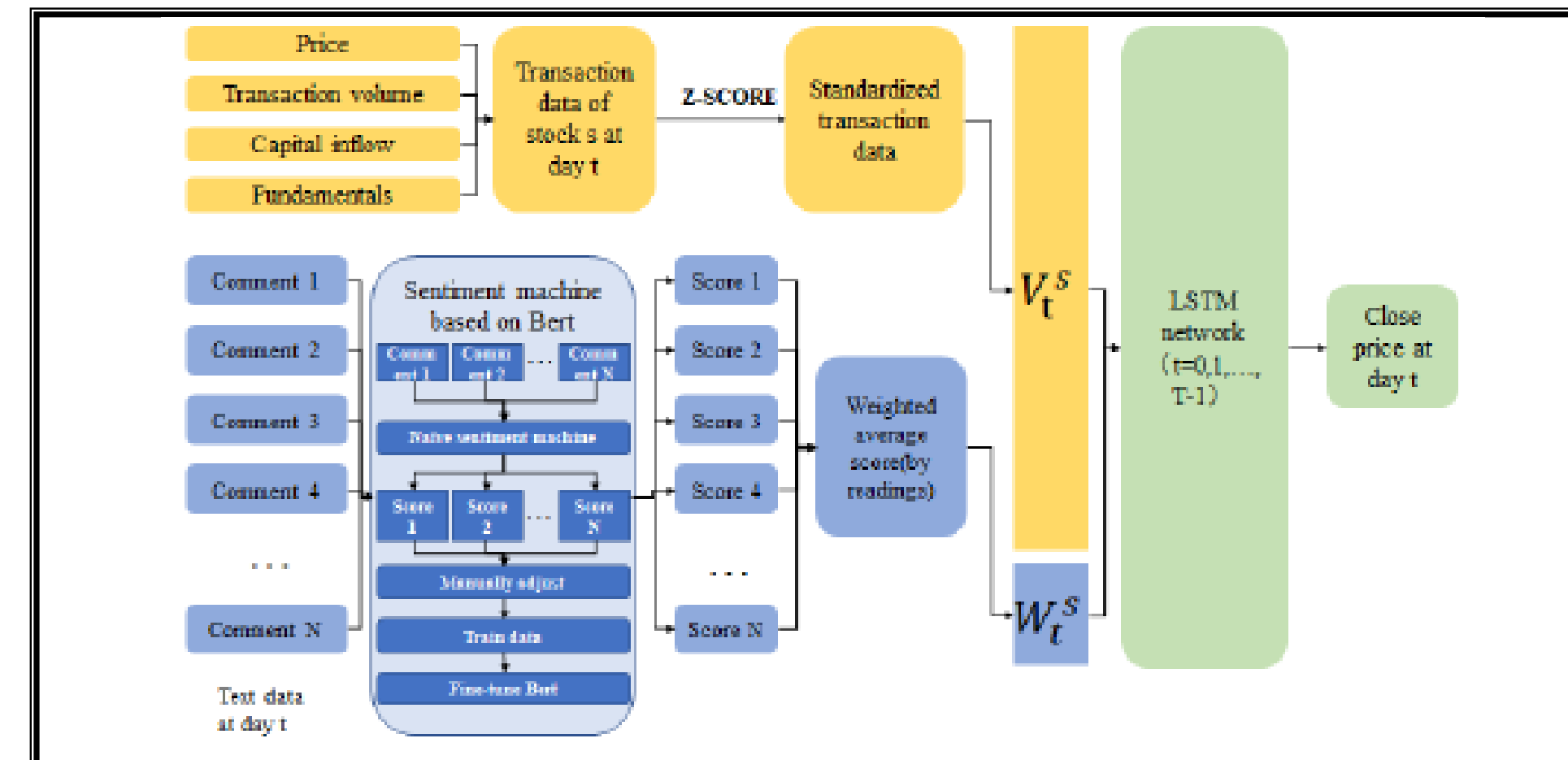
实验仿真

Model	Pingan	Yili	Wentai	Changjiang	Yangquan	(ST)Gongxin
MLP	0.146	0.259	0.748	0.116	0.061	0.039
LSTM_data	0.148	0.145	0.819	0.098	0.018	0.116
GRU_data	0.206	0.099	0.825	0.064	0.021	0.050
LSTM_text+data	0.148	0.145	0.802	0.096	0.018	0.108
GRU_text+data	0.057	0.071	0.471	0.102	0.018	0.080

(c)Transaction data

(d)Extend text data
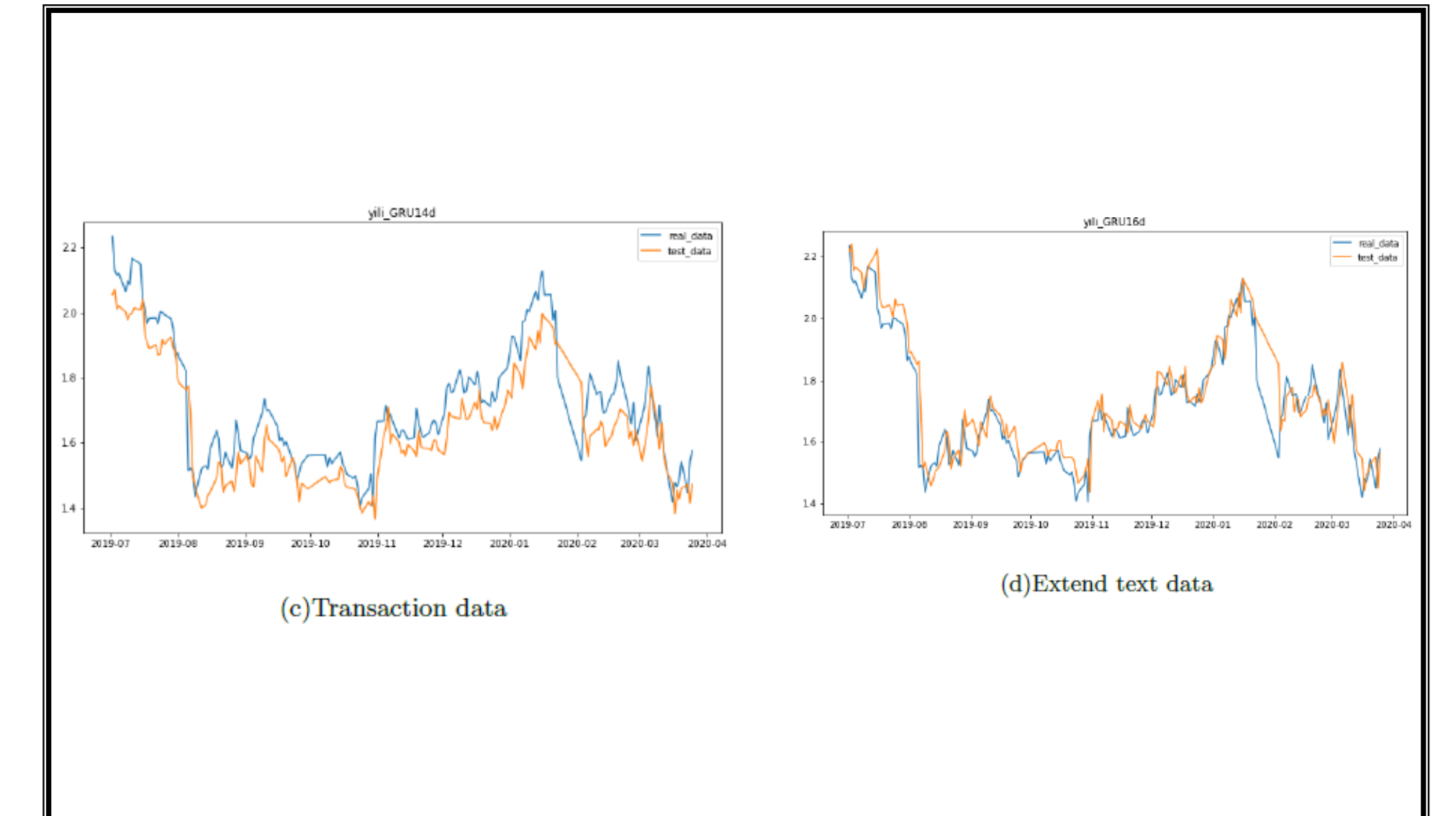
论文结论

Our paper focused on stock prediction problem on Chinese data. Specifically, we aimed to tackle a challenge lying on sentiment analysis in this problem, due to the lack of data source and labeled training data, which is a crucial part to train language processing model. We applied a semi-supervised learning method to generate training data and use the data to train BERT, a leading NLP model provided by Google. Based on this method, we successfully improved sentiment analysis efficiency on Chinese financial text data.

We further combined the sentiment data derived from our proposed method with stock transaction data and passed them into several neural network models to testify our method's efficiency when used into stock prediction problem. The experiment results successfully buttressed up our argument and also revealed its adaptability over different type of problems.

We believed our method can be widely used in this field as an effective way to supplement the insufficiency of labeled data and serve as a foundation for future works to generate more valuable insights.

XIDIAN UNIVERSITY