Fuzzy Systems and Data Mining VI
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA200699

News-Based Research on Forecast of International Natural Gas Price Trend

Tianxiang Li[a], Xiaosong Han[a], Aoqing Wang[a], Hui Li[a], Guosheng Liu[a] and Ying Pei[b, 1]
[a]*Key Laboratory for Symbol Computation and Knowledge Engineering of National Education Ministry, College of Computer Science and Technology, Jilin University, Changchun 130012, China*
[b]*The College of Information Engineering, Changchun University of Finance and Economics, Changchun 130122, China*

Abstract. In this paper, we build a deep learning network to predict the trends of natural gas prices. Given a time series, for each day, the gas price trend is classified as "up" and "down" according to the price compared to the last day. Meanwhile, we collect news articles as experimental materials from some natural gas related websites. Every article was then embedded into vectors by word2vec, weighted with its sentiment score, and labeled with corresponding day's price trend. A CNN and LSTM fused network was then trained to predict price trend by these news vectors. Finally, the model's predictive accuracy reached 62.3%, which outperformed most of other traditional classifiers.

Keywords. Natural Gas, Sentiment Analysis, Deep Learning, Price Trend Prediction,

1. Introduction

As the energy supply situation gets tenser, developing and using natural gas efficiently are becoming many countries' essential method to safeguard energy supply and reduce carbon dioxide emissions [1]. To address this issue, a large number of domestic and foreign scholars have made great efforts to predict the natural gas prices.

With the development of science and technology, more and more data emerged on the Internet. How to collect and manage these data efficiently and extract useful information from the data have become a research focus recently. Now, many scholars have applied big data analysis technology and text mining methods [2] to the study of the operating rules of financial markets. For example, the paper [3] shows that short-term stock price movements can be predicted using financial news. Robert et al.[4] used several kinds of textual representations to train an SVM model for stock prediction. In paper [5], Mark Dras built a character-based neural language model for the stock market of event-based trading. The model finally reached 64.74% accuracy of interday and 61.68% accuracy of intraday stock prediction. Xuerong Li et al.[6] applied deep learning techniques to oil forecasting and extracted hidden patterns within online news media using a CNN. The study also proposed a feature grouping method based on the LDA topic model for distinguishing effects from various online news topics. The

[1] Corresponding Author: Ying Pei, Changchun University of Finance and Economics, China; Email: 1076441579@qq.com

results show that the proposed topic-sentiment synthesis forecasting model performs better than the other benchmark models. Inspirited by the above work, we proposed a CNN-LSTM model to predict the natural gas price trend.

2. Data Collecting

The goal of this paper is using natural gas related news and history futures price to predict the trend of the natural gas future price. There are two types of data used in this study: news and gas history futures price data.

2.1. News articles

Among most natural gas related websites, we filtered the ones that update rapidly, contains large amounts of materials and have a reliable source of information [7]. In the end, we chose WorldOil as our news source, which has the fastest update frequency and the most articles. After developing a spider, we have collected 26490 news covering 2274 days. On average, there are about 11 news in one day.

2.2. Natural gas futures price

The natural gas future price was downloaded from Henry Hub, which is recognized as the U.S.'s largest and authoritative natural gas delivery and pricing center.

2.3. Granger causality test

Our initial hypothesis is that the sentiment of news may have an effect on the price trend. Granger causality test [8] is employed to prove that natural gas' futures price is related to the news sentiment score. Granger causality test is a method to analyze the causal relationship between economic variables. The prerequisite for the Granger causality test is that both sequences must be stationary. Thus, ADF tests are performed. The results are as Table 1.

Table 1. ADF test of sentiment score and price.

ADF	NLTK sentiment score	Price
Test Statistic	-3.965393	-3.110797
P-value	0.001603	0.025764
Lags Used	19	16
Number of Observation Used	2254	2257
Critical Value (1%)	-3.433255	-3.433251
Critical Value (5%)	-2.862823	-2.862821
Critical Value (10%)	-2.567453	-0.567452

It can be seen that, individually, two sequences' test statistic values are smaller than critical values (5%), and the p-values are also close to zero. Thus, we reject the null hypothesis and confirm that these two sequences are stationary. After that, two Grander causality tests are performed. The results are as Table 2.

Table 2. Sentiment score - price Granger causality test(Price – sentiment score Granger causality test)

	Number of lags 1	Number of lags 2	Number of lags 3	Number of lags 4
		Ssr based F test		
F	16.759(0.603)	5.919(0.575)	2.732(1.024)	2.095(0.794)

P	0.00(0.438)	0.0027(0.563)	0.424(0.381)	0.079(0.529)
Df_denom	2270	2267	2264	2261
Df_num	1	2	3	4
		Ssr based chi2 test		
Chi2	16.782(0.604)	11.864(1.153)	8.220(3.080)	8.414(3.188)
p	0.00(0.437)	0.003(0.562)	0.042(0.380)	0.078(0.527)
df	1	2	3	4
		Likelihood ratio test		
Chi2	16.720(0.604)	11.833(0.153)	8.205(3.078)	8.399(3.185)
p	0.00(0.437)	0.003(0.562)	0.042(0.380)	0.078(0.527)
df	1	2	3	4
		Parameter F test		
F	16.759(0.603)	5.919(0.575)	2.732(1.024)	2.095(0.794)
p	0.00(0.438)	0.003(0.563)	0.042(0.381)	0.079(0.529)
Df_demon	2270	2267	2264	2261
De_num	1	2	3	4

For different inspection methods, the results are examined, and we can draw the conclusion that news sentiment score has a strong Granger causality to natural gas prices. But conversely, the Granger causality of natural gas prices on the emotional value of news is not obvious. So, it can be considered that there is "Granger causality" on the news sentimental value and the price trend.

3. Methodology

There are mainly five steps in our method, data collecting, sentiment analysis, granger causality test, document representation and model building. Firstly, we obtain the news and price data from WorldOil and Henry Hub. Then, sentiment analysis is performed on the news to get the sentiment score of each article. To prove that natural gas related news is correlated to its futures price, we perform granger causality test on the sequence of news' sentiment score and futures price. After that, word2vec and TF-IDF are used to embed the news into vectors. Finally, we build various machine-learning methods to predict the price trend. The total process is shown in Figure 1.

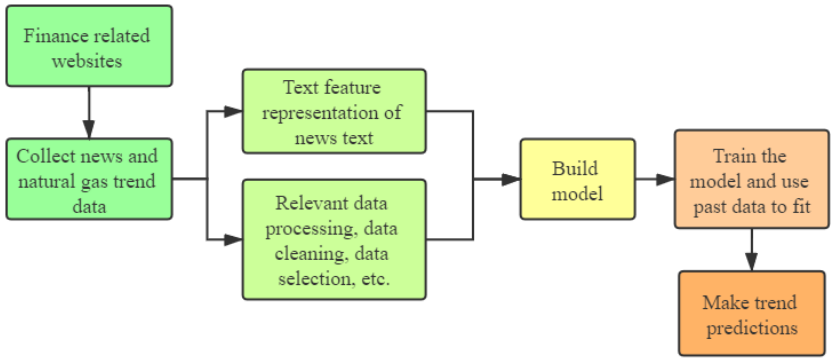

Figure 1. Main workflow

3.1. Sentiment Analysis

Sentiment analysis is a key branch of NLP (Natural Language Processing). NLTK (Natural Langrage Toolkit) [9] is the most popular NLP package. And we use NLTK's

built-in sentiment analyzer to analyze each article's sentiment score. The score is then saved for later use.

After getting the sentiment score and the price data, the two sequences are normalized and denoised, and the resulting trend graph is shown as Figure 2.

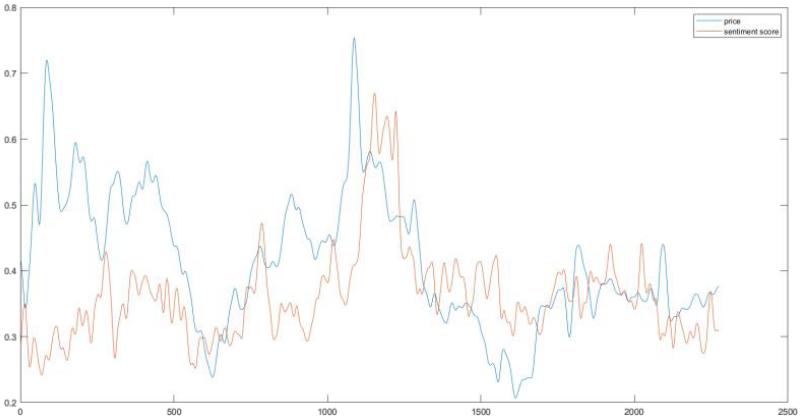

Figure 2. Sequences of Sentiment score and Price data

3.2. Document representation

We use a hybrid approach[10-11] to build article vector in this paper. Firstly, TF-IDF weight is calculated for each word in one article. Later, a Google pre-trained word2vec model is loaded to generate word vectors and the vector is then multiplied with word's corresponding TF-IDF weight. Then, the words' vectors are summed and multiplied by the article's sentiment score to form the article's vector. Since there are often more than one article in one day, the arithmetical average article vectors with the same timestamp is calculated as one day's news vector representation.

3.3. Model Building

In order to capture effective time and spatial field information from the time series sequence, a CNN and LSTM fused network is built [12]. LSTM network is needed to its extract time series features, due to its unique and powerful ability to extract features from pictures and texts, a CNN is also needed to extract features. One-dimensional convolution is performed on the daily news vectors. Then max pooling is followed. In this part, the model allows us to add various layers of different convolution and pooling layers to achieve better convolution performance. However, the optimal results from the experiment is three convolution layers and pooling layers, with convolution kernels of 2, 3, and 4. The multi-day convolution pooling results are entered a LSTM layer. After the fully connected layer, the tensor goes through a Dense layer and output a 128-dimension tensor. The CNN-LSTM network is shown as figure 3.

In order to capture the price's time series feature, another LSTM network is also used to generate a 128-dimension tensor. Then the two tensors are Concatenated and classified into two categories by a Softmax layer. The process is shown in Figure 4. And a detailed description of the network is shown in the Github (https://github.com/Wangaoqing/natural-gas-price-prediction).

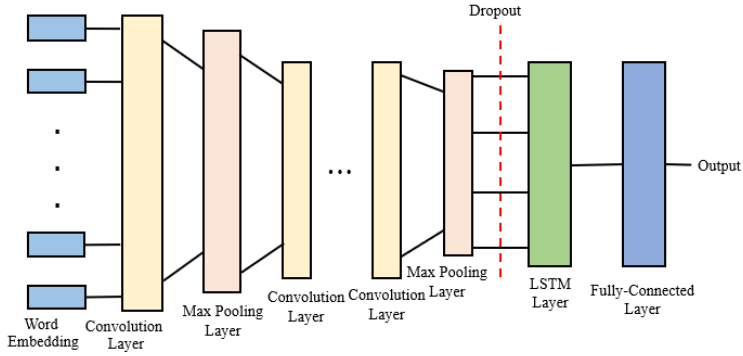

Figure 3. Brief CNN-LSTM network

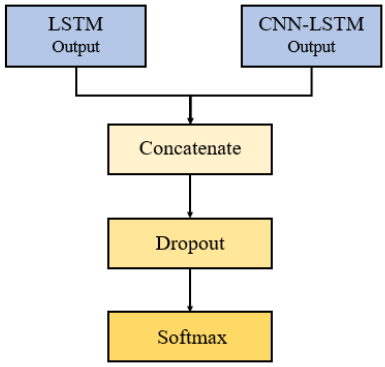

Figure 4. Final Process

4. Experiment

During the experiments, we use 30 day's news to predict next day's gas price trend. With the fused network, we achieved better results, which are shown as Figure 5.

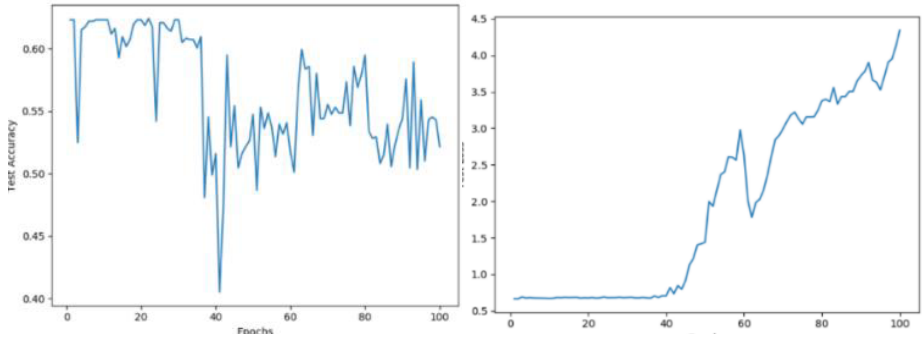

Figure 5. CNN-LSTM model predict results

When the number of iterations is less than 40, the predicted loss value does not rise, the accuracy of the test is 53%-62%, and the optimal accuracy is 62%. When the number of iteration times is greater than 50, the prediction loss rises sharply, and the model enters the over-fitting stage, which can explain the oscillation of the accuracy.

The total number of days we used in this experiment is 2274, and we split 60% of them as a training set and the other 40% as a test set. To verify our model, we also tried several supervised classifiers such as SVM (Support Vector Machines), LightGBM, RF (random forest), and NB (Naïve Bayes). The results are summed up in the Table 3.

Table 3. Results of experiments

	Accuracy
LightGBM	53%
RF	52%
NB	39%
SVM	52%
LSTM	52%
CNN-LSTM	**62%**

As a comparison, our CNN-LSTM model gives the best result with a 10% percent accuracy promotion, which means the CNN-LSTM network does extract useful features from the news and help predict the natural gas price trends.

5. Conclusion

Based on many NLP techniques and machine learning algorithms, this paper proposed a CNN-LSTM network aiming to predict the future trend of natural gas price trends. After comparison, the CNN-LSTM model performs 10% better than other compared methods. The obtained results suggest that the model works, though the accuracy still has a lot of room for improvement. In future work, we could obtain news text from a larger scale and try some state-of-art classifiers, which may have better performances.

Acknowledgement

The corresponding author is Ying Pei. The authors are grateful to the anonymous reviewers for their insightful comments which have certainly improved this paper.
This work was supported in part by the Jilin University provincial science and technology innovation project 2018B2144, and National Science Foundation of China under Grant 61972174.

References

[1] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[2] A. Tan, "Text Mining: The state of the art and the challenges," in *In Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases*, 1999, pp. 65–70..

[3] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Using Structured Events to Predict Stock Price Trend: An Empirical Investigation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1415–1425.

[4] R. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news: The AZFin text system," *ACM Trans. Inf. Syst.*, vol. 27, 2009, doi: 10.1145/1462198.1462204.
[5] L. dos Santos Pinheiro and M. Dras, "Stock Market Prediction with Deep Learning: A Character-based Neural Language Model for Event-based Trading," in *Proceedings of the Australasian Language Technology Association Workshop 2017*, Brisbane, Australia, 2017, pp. 6–15.
[6] X. Li, W. Shang, and S. Wang, "Text-based crude oil price forecasting: A deep learning approach," *Int. J. Forecast.*, vol. 35, no. 4, pp. 1548–1560, Oct. 2019.
[7] E. Junqué de Fortuny, T. De Smedt, D. Martens, and W. Daelemans, "Evaluating and understanding text-based stock price prediction models," *Inf. Process. Manag.*, vol. 50, no. 2, pp. 426–441, Mar. 2014.
[8] C. W. J. Granger, "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods," in *Essays in Econometrics: Collected Papers of Clive W. J. Granger*, USA: Harvard University Press, 2001, pp. 31–47.
[9] M. S. M. Vohra and J. Teraiya, "A COMPARATIVE STUDY OF SENTIMENT ANALYSIS TECHNIQUES 1," 2013.
[10] L.-W. Lee and S.-M. Chen, "New Methods for Text Categorization Based on a New Feature Selection Method and a New Similarity Measure Between Documents," in *Advances in Applied Artificial Intelligence*, Berlin, Heidelberg, 2006, pp. 1280–1289.
[11] A. Graves, "Generating Sequences with Recurrent Neural Networks," *ArXiv13080850 Cs*, Jun. 2014.
[12] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802–810.