计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201000056

基于新闻的国际天然气价格趋势预测方法

裴　莹[1]　李天祥[2,3]　王鏖清[4]　付加胜[5]　韩霄松[4]
1 长春财经学院信息工程学院　长春 130122
2 中国科学院大学　北京 100049
3 中国科学院新疆理化技术研究所　乌鲁木齐 830011
4 吉林大学计算机科学与技术学院符号计算与知识工程教育部重点实验室　长春 130012
5 中国石油集团工程技术研究院有限公司　北京 102206
(pei_ying_ok@126.com)

摘　要　天然气作为新型清洁能源,不仅有着重要的能源意义,作为期货交易的大宗商品之一,也有着重要的经济意义,是国家经济和国际贸易的重要组成。但是由于天然气价格受经济因素、政治因素、自然因素甚至人为因素等多种因素的影响,准确预测其价格十分困难。因此,文中设计了一种基于新闻的天然气价格趋势预测方法,该方法首先利用爬虫获取大量天然气相关新闻,并针对新闻进行嵌入表示和情感分析,运用格兰杰因果检验方法证明了天然气价格与相关新闻的情感倾向具有因果关系,并将新闻情感作为新闻向量的权值,将其相乘作为模型输入,然后构建了一个 CNN-LSTM 融合模型,CNN 用于提取新闻特征,LSTM 用于捕捉新闻和天然气价格时间序列信息,从而得到了 62% 的准确率,优于绝大多数机器学习算法。

关键词:自然语言处理;天然气价格;深度学习;因果检验;新闻
中图法分类号　TP391.1

Prediction Method of International Natural Gas Price Trends Based on News

PEI Ying[1],LI Tian-xiang[2,3],WANG Ao-qing[4],FU Jia-sheng[5] and HAN Xiao-song[4]
1 College of Information Engineering,Changchun University of Finance and Economics,Changchun 130122,China
2 University of Chinese Academy of Sciences,Beijing 100049,China
3 Xinjiang Technical Institute of Physics and Chemistry,Chinese Academy of Sciences,Urumqi 830011,China
4 Key Laboratory for Symbol Computation and Knowledge Engineering of National Education Ministry,College of Computer Science and Technology,Jilin University,Changchun 130012,China
5 CNPC Engineering Technology R&D Company Limited,Beijing 102206,China

Abstract　As a new type of clean and important energy,natural gas is one of the bulk commodities of futures trading. As an important component of the national economy and international transactions,it has important economic significance. However,due to the influence of economic,political,natural and even human factors on the price of natural gas,it is very difficult to predict the price accurately. Therefore,a news-based prediction model of natural gas price trends is proposed in this paper. In this model,text embedding and sentiment analysis are conducted on natural gas-related news. The Granger causality test is employed to prove the causality between the price of natural gas and the emotional tendency of relevant news. The news sentiment is multiplied as the weight of the news vector,and the weighted vectors are the input of CNN and LSTM fused model. CNN is used to extract news features,LSTM is used to capture time series information of news and natural gas price trends. Finally,the network achieves an accuracy as 62%. The accuracy is still better than most traditional machine learning algorithms.

Keywords　Nature language process,Gas price,Deep learning,Causality test,News

1　引言

21 世纪以来,由于石油危机的出现和天然气产业化的成熟,天然气作为一种优质清洁能源,成为了目前世界能源供应的重要组成。目前,天然气产业被赋予了更多更重要的意义,合理地使用天然气资源,可以减少温室效应的进程,也能推动

基金项目:国家自然科学基金(61972174);吉林省科技发展计划(20190302107GX);吉林省产业技术专项研究与开发(2019C053-7);广东省应用基础研究重点项目(2018KZDXM076);广东省重点学科建设计划(2016GDYSZDXK036)
This work was supported by the National Natural Science Foundation of China(61972174),Science Technology Development Project of Jilin Province(20190302107GX),Special Research and Development of Industrial Technology of Jilin Province(2019C053-7),Guangdong Key Project for Applied Fundamental Research(2018KZDXM076) and Guangdong Premier Key-Discipline Enhancement Scheme(2016GDYSZDXK036).
通信作者:韩霄松(hanxiaosong@jlu.edu.cn)

社会经济的发展;天然气贸易也逐渐成为世界能源贸易的重要组成,合理地进出口天然气也会为国家带来巨大的收益,推动全球经济发展。当前国内国外的很多学者对天然气的价格预测进行了深入研究,从对天然气价格变化行为的理论研究到对天然气价格和经济指标之间关系的研究,学者们对天然气价格进行了各种预测。在各种优秀的理论研究中,天然气价格变化行为理论旨在研究天然气价格变化的影响因素,这对于了解天然气价格变化机制具有重要意义。有研究表明,在世界范围内,需求情况、攻击情况、可替代产品的价格等条件均对天然气价格波动有影响作用,而新闻可以在一定程度上反映各种形式的变化,因此理论上来讲,使用新闻文本可以进行天然气价格走势的预测。

天然气价格走势预测问题可以采用与石油价格以及股票价格走势相同的方法,早期价格预测研究主要采用基于时间序列分析的方法。Jammazi 等[1]提出了集成了 Harr 小波变换的多层反向传播神经网络方法,提高了原油价格预测精度。Godarzi 等[2]研究发现在石油价格平稳期和振荡期,融合异质数据源的输入的 ARIMA 算法效果更好。Xie 等[3]通过特征工程使得 SVM 的原油价格预测性能优于 ARIMA 模型和 BP 神经网络。Han 等[4]引入原油日度和周度的投资者关注指数数据,使得 EEMD-PSO-LSSVM-GARCH 方法和 WN[5-6]方法在短期石油价格预测方面表现更好[7];Souzae 等[8]的研究表明,在 93%的模拟实验中,集成小波技术和隐马尔可夫模型的方法可获得更高额的利润,并实现 57%的平均趋势预测成功率。得益于计算机算力的提升和深度学习理论的兴起,神经网络模型可以采用更深更复杂的结构,从而能够进行更复杂的特征表示。在金融市场尤其在股票市场预测方面,递归神经网络(Recurrent Neural Networks,RNN)凭借其能够拟合复杂时序数据的特点,已经成为另一个股票预测研究领域的主流模型之一。Balaji 等[9]利用多种递归神经网络进行股价预测,并对标准普尔 BSE-BANKEX 指数中的所有股票进行了实证评估。Chen 等[10]使用文本挖掘技术,分类了股票相关新闻的情感倾向。最近,受国外研究的启发,国内学者逐渐将微博等社交媒体的文本数据应用于股票预测。Huang 等[11]提出了基于微博情绪信息的股票价格预测模型,该模型爬取微博文本,根据情感字典生成情感时间序列数据,然后使用 SVM 模型预测库存趋势,相较于不使用情感序列的模型,其准确性获得了大大提高。Chen 等[12]从新浪微博中精选官方账户,通过提取情感特征和潜在 Dirichlet 分布(LDA)特征,分析从这些账户中获取的新闻内容,然后将这些特征与技术指标一起输入到一个新的混合模型 RNN-boost 中,以预测中国股市的波动性,取得了良好的效果。

纵观以上工作,当前工作主要将价格趋势预测问题对应成时间序列分析问题,在此基础上进行特征工程,目前少有将新闻语义结合到天然气价格趋势预测中来的工作,且目前研究的预测精度较低。本文提出了一种融合了新闻语义的天然气价格趋势预测方法,该方法在深度神经网络的基础上充分融合了新闻语义,有效提高了天然气预测精度。本文的技术框架如图 1 所示。

图 1 技术框架图

Fig. 1 Framework

2 数据获取

关于天然气的新闻网站较少,并且一般与石油相关联,考虑到天然气是石油采集时的副产品,因此石油的产量与天然气产量息息相关,并且有着相似的价格走势。在众多的天然气相关网站中,我们选取了 oilgasdaily,worldoil,cnbc 等较为权威且新闻发布也较为及时的网站,应用爬虫技术,对新闻标题、新闻文本,以及发布时间进行内容抓取,并对新闻的时间跨度、每天发布的文章数、文章内容长度平均值进行了统计,如表 1 所列。

表 1 爬虫新闻统计

Table 1 Statistics crawled News

网站名/属性值	oilgasdaily	worldoil	cnbc
新闻开始时间	2012/2/14	2009/7/9	2014/8/11
新闻截止时间	2018/10/5	2018/10/5	2018/10/5
题目单词数	17403	258119	19447
文章正文单词数	987621	8101257	716232
平均每篇文章词数	471	299	428
文章总数	2136	27992	1720
平均每日文章数	0.88	8.30	1.13
总天数	2425	3373	1514

本文的天然气价格数据使用 henry hub 期货数据,它是纽约商品交易所天然气交割地和定价中心,在天然气行业起着至关重要的作用。henry hub 是一个地点,作为亨利中心天然气期货的交付点,位于路易斯安那州埃拉特的亨利中心,它是数个天然气互连的纽带,而它的商业意义也是其战略地位和基础设施综合作用的结果。在所选的时间跨度内,天然气价格走势如图 2 所示。

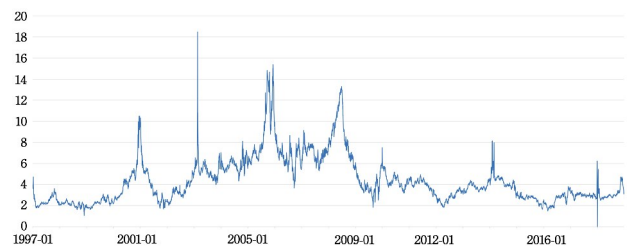


图 2 天然气价格曲线

Fig. 2 Trends of gas price

3 数据预处理

3.1 新闻预处理流程

我们利用爬虫得到了 3 个主流石油咨询网站的新闻数

据，但是新闻文本中存在着大量的标点、符号、特殊字符等，因此需要对新闻文本进行预处理，从而保存最纯粹的语义信息。本文采用宾夕法尼亚大学研发的 NLTK 文本预处理包进行预处理，包含分词、词干提取、词形还原、去除停用词、去除特殊字符、大小写统一等步骤。经过预处理后的文本较为整洁，进行文本嵌入时可有效减少噪声，便于后续的工作。我们还对预处理后的每篇新闻利用 NLTK 进行了情感分析，得到每篇新闻属于积极或者消极的概率，本文将积极概率减去消极概率作为每篇新闻的情感打分(sentiscore)。预处理的流程如图 3 所示。

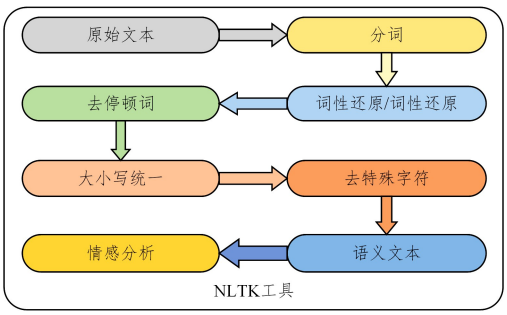


图 3　新闻数据预处理

Fig. 3　Preprocessing of news

3.2　新闻情感和天然气价格的因果分析

我们将天然气价格进行归一化，与每天的新闻情感均值放在一起，并进行差分，所得结果如图 4 所示。可以看到情感与价格之间的走势具有一定的相关性，并且很多时候情感的走势比价格走势提前，由此我们假定天然气新闻情感对天然气价格走势具有预测意义。接下来使用格兰杰因果检验，进行证明。

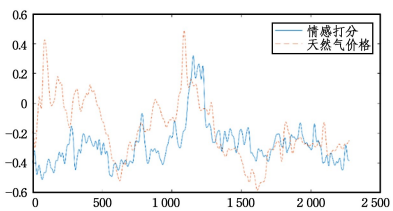


图 4　天然气价格和每日新闻情感对比

Fig. 4　Gas price and daily news sentiscore

格兰杰因果检验于 1969 年提出，是一种统计学检验假设检验方法，一般用于确定一个时间序列是否可以用于预测另一个序列。通常，回归反射出“纯粹”的相关性，但格兰杰认为因果关系在经济学中，可以通过测量一种时间序列对另一个时间序列预测未来值的能力所度量。格兰杰因果关系不是检验 Y 是否导致 X，而是检验 Y 是否预测 X。通常通过一系列的 t 检验和 F 检验上 X 的滞后期(Y 的滞后期也包含在内)来说明 X 是 Y 的格兰杰原因，这些 X 可以提供一些对 Y 未来统计值很重要的信息。格兰杰基于两个原则定义了因果关系：

(1)原因发生在其影响之前；

(2)某种原因对于未来值的影响是独特的。

鉴于这两个因果关系的假设，格兰杰建议检验以下假设，以确定是否存在因果关系：

$$\mathbb{P}[Y(t+1)\in A|\mathcal{J}(t)]\neq\mathbb{P}[Y(t+1)\in A|\mathcal{J}_{-X}(t)] \quad (1)$$

其中，$\mathbb{P}$指概率，A 指任意一个非空集合，$\mathcal{J}(t)$和$\mathcal{J}_{-X}(t)$分别表示截止时间 t 的可用信息以及表达不包括 X 的修改后的空间内的可用信息。如果上述假设被接受，我们就可以说 X 是 Y 的格兰杰原因。

在实际使用格兰杰因果检验时，如果我们所用的时间序列是平稳序列，则使用两个或多个变量的水平值进行测试；如果变量是非平稳的，则使用一阶差分或者更高阶的差分进行检验。我们首先对新闻情感时间序列以及天然气价格时间序列进行平稳性检验，即 ADF 检验，检验结果如表 2 所列。

表 2　平稳性检验

Table 2　ADF test

ADF	NLTK sentiment score	price
Test Statistic	−3.965 393	−3.110 797
p-value	0.001 603	0.025 764
Lags Used	19	16
Number of Observation Used	2 254	2 257
Critical Value(1%)	−3.433 255	−3.433 251
Critical Value(5%)	−2.862 823	−2.862 821
Critical Value(10%)	−2.567 453	−2.567 452

从结果可以看到两个序列的 test statistic 值均小于 Critical value (5%)，p-value 接近于 0。因此我们拒绝 0 假设，认为两个序列都是平稳的。接下来进行格兰杰因果检验，我们检验了情感-天然气价格因果关系，以及天然气价格-情感因果关系，检验结果如表 3 和表 4 所列。由结果可见，新闻情感对于天然气价格具有很强的格兰杰因果关系，然而天然气价格对于新闻情感的因果关系却不是很明显，因此我们认为在新闻情感和天然气价格之间具有格兰杰因果关系。

表 3　情感-价格格兰杰因果检验

Table 3　Sentiment score-price Granger causality test

Number of lags		1	2	3	4
Ssr based F test	F	16.759 4	5.918 8	2.731 5	2.095 2
	p	0.00	0.002 7	0.424 0	0.079 0
Ssr based chi2 test	$Chi2$	16.781 5	11.863 8	8.219 8	8.414 0
	p	0.00	0.002 7	0.041 7	0.077 5
Likelihood ratio test	$Chi2$	16.719 9	11.832 9	8.205 0	8.398 5
	p	0.00	0.002 7	0.042 0	0.078 0
Parameter F test	F	16.759 4	5.918 8	2.731 5	2.095 2
	p	0.00	0.002 7	0.042 4	0.079 0

表 4　价格-情感格兰杰因果检验

Table 4　Price-sentiment score granger causality test

Number of lags		1	2	3	4
Ssr based F test	F	0.603 2	0.575 3	1.023 5	0.793 7
	p	0.437 5	0.562 6	0.381 1	0.529 1
Ssr based chi2 test	$Chi2$	0.604 0	1.153 2	3.080 0	3.187 6
	p	0.437 1	0.561 8	0.379 5	0.526 9
Likelihood ratio test	$Chi2$	0.603 9	0.152 9	3.077 9	3.185 4
	p	0.437 1	0.561 9	0.379 8	0.527 3
Parameter F test	F	0.603 2	0.575 3	1.023 5	0.793 7
	p	0.437 5	0.562 6	0.381 1	0.529 1

4　数据嵌入

本文使用 Word2Vec 进行新闻文本的特征表示，并且使用谷歌预训练的 Word2Vec 模型，并在此基础之上利用我们获取的新闻数据进行了微调，预训练模型通过大量数据训练得到，有助于提升预测效果。Word2Vec 是谷歌于 2013 年提出的一种两层神经网络的深度学习方法。Word2Vec 神经网络使用大量的数据进行训练，并把文本映射到向量空间，此技

术可得到词的语义向量，使得语义相近的词的向量距离接近。

本文除了使用 Word2Vec 词向量外，还使用了 TF-IDF 权值，由于每篇新闻的主题不同，因此对一篇新闻中的词计算每个词的 TF-IDF 权重，选出权重前 20 的词进行向量化表示，并乘以每个词对应的 TF-IDF 权重，相加得到一篇新闻的向量表示。TF-IDF 的计算方式如式(2)和式(3)所示。

$$TF=\frac{\text{某个单词在文档中的出现次数}}{\text{文档的总词数}} \tag{2}$$

$$IDF=\log_{10}\frac{\text{文档数}}{\text{包含某一单词的文档数}} \tag{3}$$

最后，我们利用新闻预处理阶段得到的情感值作为权值与新闻向量相乘，得到融合了情感信息的新闻向量。文本嵌入的流程如图 5 所示。

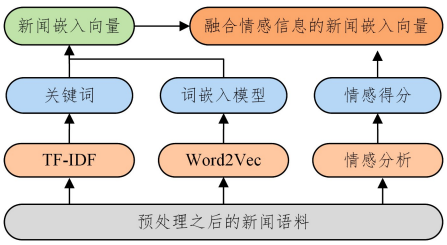


图 5 新闻嵌入流程

Fig. 5 Pipeline of News embedding

5 基于深度学习的天然气价格趋势预测模型

本文首先分析天然气价格的历史数据，这是一个典型的时间序列分析问题。LSTM(Long Short-Term Memory)是当下最流行用于时间序列分析的人工神经网络，LSTM 是 Hochreiter 等[13]于 1997 年为了解决循环神经网络梯度弥散问题提出的复合的记忆单元网络结构，典型的 LSTM 结构如图 6 所示。

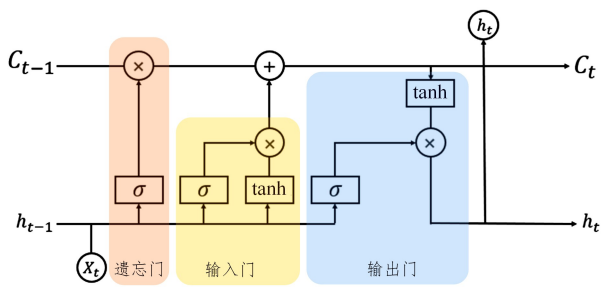


图 6 LSTM 结构

Fig. 6 Structure of LSTM

遗忘门用于决定从内容中丢弃哪些细节部分，由 *Sigmoid* 函数决定。它查看前一时刻的状态 h_{t-1} 和内容输入 x_t，并为单元状态 C_{t-1} 中的每个数字都计算出一个介于 0,1 之间的值，并输出。输入门是用于发现输入中的哪个值被用来修改存储器，*Sigmoid* 激活函数将一个数调整到 0,1 之间，之后 tanh 函数对所传递的值进行加权，并确定它们的重要性级别(−1 到 1 之间)。输出门与输入门类似，当前输入和块(block)的存储器用于确定输出。*Sigmoid* 函数决定让哪些值通过，之后 tanh 函数对所传递的值进行加权，并确定它们的重要性级别(−1 到 1 之间)，再乘以 *Sigmoid* 的输出。

本文尝试使用 LSTM 网络对天然气价格的时间序列信息进行学习并预测，为了提取更深层的数据规律，使用了双层 LSTM 网络，输入为 30 天的价格序列。实验机器配置为 i5-8500CPU，16 GB 内存，GTX1080 显卡。LSTM 的结果如图 7 所示。

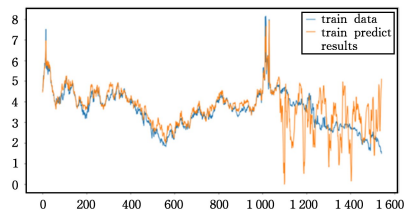


图 7 基于价格序列的 LSTM 结果

Fig. 7 LSTM result based on gas price

图 7 表明在模型在预测阶段有过拟合，把回归问题改成分类问题(价格相对于前一天是升还是降)，测试准确率在 52%左右波动。这是由于简单的 LSTM 网络仅考虑了时间维度，而单纯地依据价格时间序列维度不能获取到价格波动的规律信息。因此，尝试使用新闻语义向量的进行预测，结果如图 8 所示。

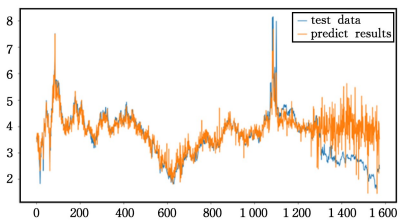


图 8 基于新闻向量的 LSTM 结果

Fig. 8 LSTM result based on news vector

图 8 表明在模型预测阶段仍有过拟合现象，转换成趋势分类问题，准确率随着迭代次数的增长在 57%左右徘徊。使用新闻向量进行回归预测的效果优于使用价格序列进行预测，说明价格的波动规律不能单纯地从历史价格中学习，同时说明新闻向量能够对价格走势进行一定的预测作用，但是其仍然不能很好地拟合测试集的波动情况。因此本文尝试综合价格的时间序列信息和新闻文本的特征对价格的影响，构建了一个 CNN 和 LSTM 相融合的网络。

我们首先对多天多篇新闻数据进行一维卷积，然后将卷积数据合并。在这一部分中，我们添加了不同层数的卷积层和池化层，以实现更好的卷积性能和特征学习效果。多次实验表明 3 个卷积层和池化层效果最好，卷积核分别为 2,3 和 4。多天的卷积池化结果经过一个 TimeDistributed 层，使用了许多连续的高层输出结果作为 LSTM 的输入，网络结构如图 9 所示。

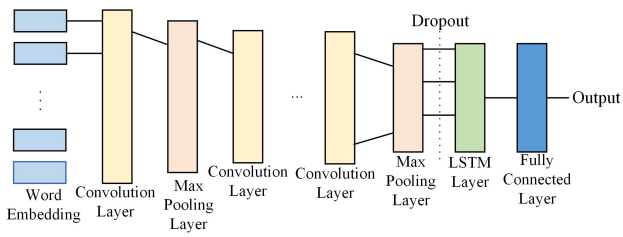


图 9 CNN-LSTM 融合网络

Fig. 9 CNN-LSTM network

CNN-LSTM 网络有效地在时间和空间上对新闻向量进行了特征学习，我们在此基础之上在网络全连接层拼接了仅

利用天然气价格进行预测的 LSMT 网络的全连接层，从而将新闻的语义信息和天然气历史价格信息融合起来进行预测。网络结构如图 10 所示。

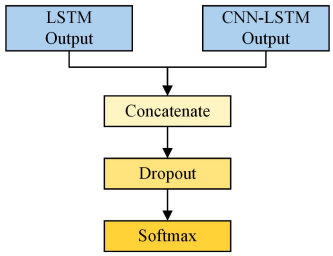
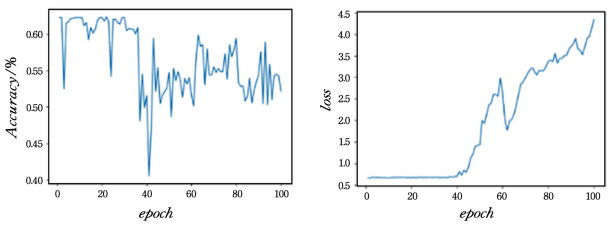


图 10　总体网络构架

Fig. 10　Overall network architecture

网络预测结果如图 11 所示，迭代次数小于 40 时，预测的 *loss* 不再增加，测试的 *acc* 值为 53%～62%，最佳准确率为 62%。当迭代次数大于 50 时，随着迭代次数的增加，*loss* 成倍增加，模型进入过度拟合阶段，因此可以解释准确率的波动情况。



(a) 准确率　(b)损失

图 11　总体网络构架的结果

Fig. 11　Results of overall network architecture

本文将构建的网络和其他机器学习算法进行对比，结果如表 5 所列。

表 5　结果对比

Table 5　Results comparison

Method	*acc*/%
LightGBM	53
Random Forest	52
Native Bayes	39
SVM	52
LSTM	52
CNN-LSTM(our method)	62

从表 5 可见，传统的机器学习方法无法对大量新闻向量进行有效的特征学习，从而导致性能不佳；而 LSTM 模型仅使用价格数据或者新闻数据无法做出精确的预测。相比之下，本文提出的 CNN-LSTM 模型可提供更好的结果，准确率提高了 10%左右，意味着 CNN-LSTM 网络确实从新闻向量中学习到了有效特征并辅助预测天然气价格的走势。

结束语　本文从 henryhub 获取了近年来的天然气价格，并利用爬虫从 3 个主流的石油网站获取了近年来的相关新闻，利用格兰杰分析证明了新闻情感和天然气价格具有因果关系。然后利用 Word2Vec 工具将获取的新闻进行了向量化，并利用情感得分进行了加权。最后，本文利用 CNN-LSTM 网络有效地对新闻数据进行了特征学习，并融合了基于天然气价格的 LSTM 网络。实验结果表明，本文提出的方法优于传统的机器学习算法，最高准确率可以达到 62%。本文的模型虽然提高了天然气价格趋势预测准确率，但准确率依然偏低，我们将在后续工作中尝试更多的方法来进一步提高模型的准确率。

参 考 文 献

[1] JAMMAZI R，ALOUI C. Crude oil price forecasting：experimental evidence from wavelet decomposition and neural network modeling[J]. Energy Economics，2012，34(3)：828-841.

[2] GODARZI A A，AMIRI R M，TALAEI A，et al. Predicting oil price movements：a dynamic artificial neural network approach [J]. Energa-Policgy，2014，68：371-382.

[3] XIE W，YU L A，XU S Y，et al. A New Method for Crude Oil Price Forecasting Based on Support Vector Machines[C]// International Conference on Computational Science. Springer，Berlin，Heidelberg，2006：444-451.

[4] ZHANG J L，ZHANG Y J，ZHANG L. A novel hybrid method for crude oil price forecasting[J]. Energy Economics，2015，49：649-659.

[5] WESTERLUND J，NARAYAN P K. Does the choice of estimator matter when forecasting returns? [J]. Journal of Banking & Finance，2012，36(9)：2632-2640.

[6] WESTERLUND J，NARAYAN P. Testing for predictability in conditionally heteroskedastic stock returns[J]. Journal of Financial Econometrics，2015，13(2)：342-375.

[7] HAN L Y，LV Q N，YIN L B. Can investor attention predict oil prices? [J]. Energy Economics，2017，66：547-558.

[8] DE SOUZAE S，LEGEY L F L，DE SOUZA E. Forecasting oil price trends using wavelets and hidden Markov models[J]. Energy Economics，2010，32(6)：1507-1519.

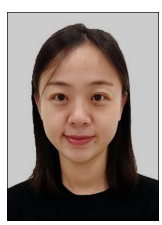
[9] BALAJIA J，DS H R，NAIR B B. Applicability of Deep Learning Models for Stock Price Forecasting：An Empirical Study on BANKEX Data[J]. Procedia Computer Science，2018(143)：947-953.

[10] CHEN Q，LIAN W L. Sentiment Classification of Stock News from Internet based on Text Mining[J]. China Market，2015 (24)：234-235.

[11] HUANG R P，ZUO W M，BI L Y. Predicting the Stock Market Based on Microblog Mood[J]. Journal of Industrial Engineering/ Engineering Management，2015(01 vo 29)：47-52，215.

[12] CHEN W L，YEO C，LAU C T，et al. Leveraging social media news to predict stock index movement using RNN-boost[J]. Data & Knowledge Engineering，2018，118(NOV.)：14-24.

[13] HOCHREITER S，SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation，1997，9(8)：1735-1780.



PEI Ying，born in 1990，Ph.D student，assistant professor. Her main research interests include financial big data analysis and machine learning.

HAN Xiao-song，born in 1984，Ph.D，associate professor，is a member of China Computer Federation. His main research interests include machine learning and optimization algorithm.