

基于文本信息的碳价预测

孙梦敏

贵州大学经济学院, 贵州 贵阳

收稿日期: 2024年9月19日; 录用日期: 2024年10月16日; 发布日期: 2024年11月21日

摘要

为探究文本信息对我国碳价预测的影响, 本研究搜集了2014年1月1日至2021年12月31日期间与碳价相关的新闻及政策资料, 并通过相似度分析将文本信息进行量化处理。随后, 本研究将量化后的文本信息与期货结算价、WTI原油价格、欧元兑人民币汇率、期货成交量以及核证减排量、空气质量指数等指标相结合, 以验证文本信息在预测国内碳价方面的有效性。经过分析, 得出以下结论: 1) ARIMA模型和LSTM模型在捕捉序列信息方面表现更为出色, 其效果超越了其他模型。2) 文本中蕴含了对碳价预测有帮助的有效信息, 这些信息能够提高模型的预测准确性。总的来说, 对碳价进行更精确的预测, 能帮助企业及时掌握市场动态, 提前制定投资决策; 同时也有利于政府适时实施宏观调控。

关键词

深度学习, 文本信息, 碳价预测

Carbon Price Prediction Based on Text Information

Mengmin Sun

School of Economics, Guizhou University, Guiyang Guizhou

Received: Sep. 19th, 2024; accepted: Oct. 16th, 2024; published: Nov. 21st, 2024

Abstract

In order to explore the impact of text information on China's carbon price forecast, this study collected news and policy data related to carbon price from January 1, 2014 to December 31, 2021, and quantitatively processed the text information through similarity analysis. Subsequently, this study combined the quantified text information with futures settlement price, WTI crude oil price, euro-RMB exchange rate, futures trading volume, certified emission reduction, air quality index and other indicators to verify the validity of text information in predicting domestic carbon price. After

analysis, the following conclusions are drawn: 1) ARIMA model and LSTM model perform better in capturing sequence information than other models. 2) The text contains effective information helpful for carbon price prediction, which can improve the prediction accuracy of the model. In general, a more accurate forecast of carbon prices can help enterprises grasp market dynamics in time and make investment decisions in advance. At the same time, it will also help the government carry out macro-control in a timely manner.

Keywords

Deep Learning, Text Information, Carbon Price Prediction

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

2005年2月16日,《京都议定书》正式生效,其规定旨在将大气中的温室气体浓度稳定在适宜水平,以避免极端气候变化对人类造成损害。自《京都议定书》生效以来,中国积极推动碳金融的发展,设立湖北、广东、上海、天津、北京、深圳、重庆七个试点,进行碳排放交易。经过七个试点几年以来的经验总结,2021年7月16日,全国碳排放交易体系正式启动,意味着我国碳排放交易体系的更进一步发展。

在碳排放领域,对碳排放权的明确界定正是基于科斯定理的原理,旨在使那些产生碳排放负外部性的企业承担相应的社会成本,从而在社会整体层面上实现减排目标。在政府对碳排放权进行初始分配之后,排放量较大的企业往往需要额外的碳排放权,而排放量较小的企业则可能拥有剩余的碳排放权。这些企业之间的碳排放权交易便形成了所谓的碳排放权交易价格,简称碳价。碳排放权的合理定价能够准确及时地反映市场供求状况、交易者的成交意愿以及交易风险等关键信息。恰当的定价机制能够有效地引导资金流向,助力实现降低二氧化碳排放的目标。此外,合理的碳价有助于企业更有效地控制生产成本,并作出更明智的经营决策。Daskalakis 和 Markellos 的研究表明,碳价通过提高电力成本,可以削弱能源密集型产业的竞争力,进而影响公众福利[1]。碳价过高可能会在短期内对产业和社会造成不利影响,而碳价过低则可能无法有效激励企业减少排放。因此,市场投资者和监管部门都迫切需要全面且及时地了解碳价市场的信息,并据此做出准确的碳价预测,以便及时调整相应的策略。

然而,碳价的非线性、非平稳性和高噪声等特点,导致对碳价的预测工作难以取得满意的效果。为应对这一难题,相关研究采用了擅长处理非线性、非平稳数据的机器学习模型,并取得了一定的进展。但是,由于传统机器学习模型固有的局限性,过拟合和仅能找到局部最优解的问题仍然存在,这成为该方法的主要障碍。此外,依据半强有效市场假说,文本中蕴含着对价格预测有帮助的重要信息,但这些信息往往难以被传统机器学习模型所利用[2]。针对这些问题,人工智能的最新发展——深度学习,在解决过拟合和寻找全局最优解方面展现出了良好的效果,并且在有效提取文本信息方面也取得了显著的成就。

根据目前的学术研究进展,诸多文献已经运用深度学习技术对碳价进行预测分析。然而,审视这些研究中所使用的数据,可以发现大多数研究依赖于汇率、空气质量指数等结构化数据[3]。尽管有部分研究者尝试引入文本数据,但这些文本数据主要限于百度搜索指数,这可能无法充分反映政策文件或新闻

报道中所蕴含的关键信息[4]。鉴于此,本研究选取了2014年1月1日至2021年12月31日的深圳碳价数据作为研究样本,将政策文本和新闻文本数据纳入考量,并结合深度学习模型对碳价的波动进行预测。本研究旨在为碳价预测领域提供一种创新的视角和方法论,同时为政策制定者和投资者提供更为全面和精确的碳价预测数据支持。

2. 文献综述

目前,国内外碳排放权价格的研究从两方面展开:一是对碳价驱动因素的研究,二是对碳价进行预测。

在碳价驱动因素的研究方面,针对于国际碳市场,Mansanet-Bataller等通过多个计量模型表明,布伦特原油和天然气价格变化对欧盟碳价有显著影响[5]。Mansanet-Bataller和Pardo通过实证发现有关碳价的新闻报道对欧盟碳价有显著影响[6]。朱帮助等通过EMD分解认为国际碳价主要由其内在的长期趋势,重大事件影响和市场短期波动三方面构成[7],Lovcha等通过实证分析表明高达90%(平均65%)的碳价格波动由基本市场变量的波动解释[8]。对于国内碳市场,王丹舟和杨德天通过对广东地区碳价的研究,发现对碳价产生显著正向影响的有天然气与原油价格,工业产业、极端天气,而煤炭价格、清洁点火差价对碳价有显著的负向影响[9]。张云研究表明,交易所层面关于碳排放配额信息显著影响碳价,政府层面相关信息并不显著[10]。吕靖烨等通过对湖北和深圳的数据进行研究认为宏观经济变量对碳价波动影响较大[11]。魏宇等等通过动态模型选择和动态模型平均揭示了经济形势和国内外金融市场和碳市场的走势对我国碳交易价格有重要影响[3]。

针对碳价预测,主要分为两部分,一部分是在EMD或EEMD分解基础上构建模型,以提升模型的预测精度。张晨和胡贝贝在EEMD分解基础上用BP神经网络分别对高中低频的数据进行预测,取得了较好的预测效果[12]。Sun等将欧洲气候交易所(ECX)和韩国交易所(KRX)碳价数据进行EEMD分解,得到不同频次的IMF分量,同时采用改进的bat算法对极限学习机进行增强,取得了不错的预测效果[13]。还有学者在MEEMD基础上,利用改进的鲸鱼优化算法(IWOA)优化的LSTM网络对北京、福建和上海碳价数据进行预测,有效地提高了模型的准确性[14]。另外,Zhou等首先使用ICEEMDAN分解原始数据,然后利用Lempel-Ziv复杂度算法将分解后的子序列重构为一个新序列,最后利用Sparrow搜索算法对极端学习机模型优化后进行碳价预测,在湖北,北京和广东试点取得了不错的效果,并认为分解重构预测集成策略是一种有效的预测碳价的方法[15]。Chen等扩展了集成经验模态分解(EEMD)工作,使用模糊熵和极限学习机构造新模型,使得对欧盟碳期货合约每日价格的预测精度提升了40% [16]。

另一部分则是直接利用非线性模型,对碳价进行预测。如Jiang和Wu将国际碳价分解为线性和非线性两部分,分别使用ARIMA和随机向前集拟合线性和非线性子序列,取得了不错的预测效果[17]。王娜实证证明加入百度指数和媒体指数的网络结构ADL模型有助于提高碳价预测效果[4]。蒋锋和彭紫君提出CPSO-BP模型对碳交易价格进行预测,得到了较好的结果[18]。Xu等利用时间序列复杂网络分析技术将碳价数据映射到一个碳价网络,后利用极限学习机对欧盟碳价进行预测,模型精度有所提升,且发现对模型进行随机数据、不同频率数据或结构变化的数据进行测试时,模型具有较好的稳定性[19]。Wang等基于机器学习模型的综合特征筛选技术(CFS)和概率估计方法相结合,提出了一种新的碳价格预测模型,在上海碳交易所取得了较好的效果[20]。

由上述分析可知,碳价的驱动因素主要包括化石能源价格、重大事件、极端气候现象、宏观经济状况以及新闻报道等。在我国碳市场中,尤其受极端气候现象、宏观经济状况、国内外碳市场动态以及交易所层面的碳排放配额信息影响显著。在碳价预测方面,一种方法是采用经验模态分解(EMD)等技术对碳价序列进行分解,分别预测不同频率的子序列,再将这些子序列综合以形成最终的预测结果,这种方法已显示出良好的预测性能。另一种方法是直接应用非线性模型进行碳价预测,同样取得了令人满意的

效果。然而，目前大多数学者在进行碳价预测时主要依赖结构化数据，而忽略了新闻报道等非结构化数据。尽管非结构化数据被认为是碳价的重要驱动因素，但其内部信息的提取和应用却极具挑战性。部分学者注意到这样的问题，并尝试将非结构化数据融入模型，取得了较好的预测效果，但其所采用的百度指数和媒体指数存在一定的局限性，可能无法完全捕捉政策或新闻文本中的关键信息。因此，本研究选取 2014 年 1 月 1 日至 2021 年 12 月 31 日的深圳碳价数据作为研究对象，结合深度学习领域的自然语言处理技术，将政策文本和新闻文本数据整合进模型，以预测碳价的涨跌趋势。

3. 模型介绍

对于碳价预测，本文采取了三个深度学习模型，分别是 CNN，RNN 和 LSTM 模型，三个模型原理如下：

3.1. CNN 简介

CNN，即卷积神经网络，是深度学习中用于图像识别的模型。它是一种前馈神经网络，特别设计来处理多维数据，主要由交替的卷积层和池化层构成。其通过感受野来获取一定范围内的上下文信息，对处理时间序列问题有一定能力，CNN 卷积层和池化层原理如下：

卷积层：

$$O_i = f(W_i * O_{i-1} + B_i) \quad (1)$$

其中 O_i 为第 i 层特征子图； f 为激活函数； W_i 为第 i 层卷积核权重系数； O_{i-1} 为第 $i-1$ 层特征子图； B_i 为第 i 层偏置系数。池化层：

$$S_i(Z_j) = \max(0, O_i | Z_j \times Z_j) \quad (2)$$

式中， Z_j 为第 j 个池化层的尺寸， $O_i | Z_j \times Z_j$ 为第 i 层特征子图 O_i 在第 $Z_j \times Z_j$ 区域中的最大值。

总之，作为一种深度学习模型，其核心思想是利用卷积操作自动提取特征，从而减少手动特征工程的需求，该模型能够处理大量数据，有效捕捉空间特征，适合大规模数据集的训练。

3.2. LSTM 简介

LSTM 又称长短期记忆，是一种特殊的循环神经网络，相对于简单的 RNN，其更擅长于处理长序列问题。LSTM 的原理示意图如图 1 所示：

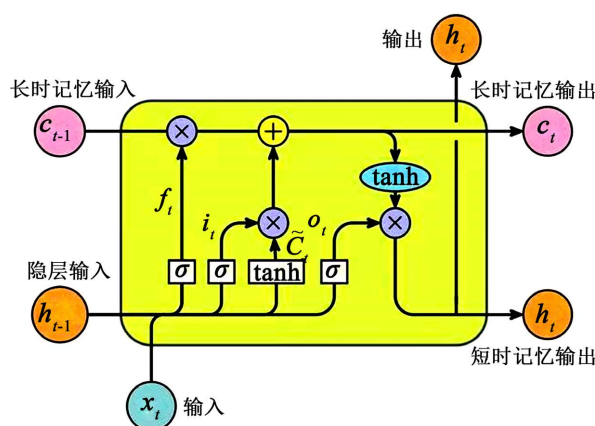


Figure 1. LSTM schematic diagram

图 1. LSTM 示意图

其中，遗忘门用来决定从历史中丢弃和保留什么信息，其数学表达式是：

$$f_t = \sigma(w_f \cdot [h_{t-1}, x_t] + b_f) \quad (3)$$

输入门一是确定哪些记忆需要被更新，一是将更新后的记忆重新写入记忆 中。其数学表达式为：

$$i_t = \sigma(w_i \cdot [h_{t-1}, x_t] + b_i) \quad (4)$$

$$\tilde{C}_t = \tanh(w_c \cdot [h_{t-1}, x_t] + b_c) \quad (5)$$

将上面两个公式做元素级别的相乘便可以得到当前 t 时刻需要记忆的信息，下面将此刻记忆的信息更新到旧的记忆中，其数学表达式为：

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \quad (6)$$

C_t 作为当前 t 时刻的最新记忆将一直传递到下一刻。

输出门：和 Sigmoid 函数相乘表示输出的概率。输出门的数学表达式为：

$$o_t = \sigma(w_o \cdot [h_{t-1}, x_t] + b_o) \quad (7)$$

$$h_t = o_t \cdot \tanh(C_t) \quad (8)$$

仍然使用 $\tanh$ 激活函数对当前记忆进行重整，然后和输出概率向量 o_t 做元素级别相乘，得到输出向量 h_t 。

基于数据处理和预测方面的优势，LSTM 被广泛应用于自然语言处理和时间序列预测等，不仅能够捕捉长时间的依赖关系，适用于长序列数据，还能通过门控机制有效控制信息流，提高模型的表现。但相比于传统 RNN，LSTM 的计算复杂度较高，训练时间较长。

3.3. TF-IDF 简介

TF-IDF 算法被广泛应用于评估词汇在特定文档集合中的重要性。该算法的优势在于能够有效过滤掉语气词等高频但信息价值较低的词汇，从而准确地提取出文档中的关键信息。其原理如下：

词频(term frequency, TF)指的是某一个给定的词语在该文件中出现的频率。TF 计算方式：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \quad (9)$$

式中， $n_{i,j}$ 是该词在文件 d_j 中出现的次数，而分母则是在文件 d_j 中所有字词出现次数之和。

逆向文件频率(inverse document frequency, IDF)是用来度量一个词语的普遍性，其是消除高频语气词的关键。某一特定词语的 IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到：

$$idf_i = \log \frac{|D|}{1 + |\{j : t_i \in d_j\}|} \quad (10)$$

其中， $|D|$ 为语料库中的文件总数， $|\{j : t_i \in d_j\}|$ 为包含词语 t_i 的文件数目。

$$tfidf_{i,j} = tf_{i,j} \times idf_i \quad (11)$$

基于在文本特征提取方面的优势，TF-IDF 模型被广泛应用于信息检索、文本挖掘和自然语言处理等领域，其优点在于可以有效地降低常见词对文本表示的影响，且容易实现。然而，对于长文档，TF-IDF 模型可能会偏高，导致结果失真，而且可能无法很好处理同义词或多义词问题。

3.4. 评价指标

基于以上对 CNN, RNN 和 LSTM 等模型的分析, 结合碳价预测的特点, 本文选取 MSE 为评价指标, 其计算方式如下:

$$Mse = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \tag{12}$$

其中 n 为预测样本个数, y_i 为样本实际值, $\hat{y}_i$ 为样本预测值。

本文选取 MAE 为评价指标, 其计算方式如下:

$$Mae = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \tag{13}$$

其中 n 为预测样本个数, y_i 为样本实际值, $\hat{y}_i$ 为样本预测值。

作为评估回归模型性能的一种常用指标, MSE 定义清晰, 易于计算和解释。它通过计算预测值与实际值之间的平方差的平均值来量化模型的预测误差。MSE 越小, 说明模型的预测效果越好。

4. 数据来源及特征筛选

本文选取深圳市 2014 年 1 月 1 日~2021 年 12 月 12 日数据, 分为文本数据和非文本数据。本文数据集包括两部分: 文本数据和非文本数据。文本数据来自 Wind, 非文本数据来自碳排放交易网每日新闻数据和政策文件。本文以 2014 年 1 月 1 日~2021 年 1 月 6 日共 890 个有效数据作为训练集, 同时以 2021 年 1 月 6 日~2021 年 3 月 22 日共 20 个有效数据作为测试集。数据集的统计信息见表 1 所示:

Table 1. Data set statistics
表 1. 数据集统计信息

	训练集	测试集
有效数据	890	20

4.1. 非文本数据筛选

本文根据决策树, 和 XGBoost, LightGBM, GDBT, Adaboost 五个树类模型所输出的特征重要性进行排序, 根据特征在每个模型中的重要性次序求取平均次序, 选取平均次序低于 10 的特征, 选出如下六个特征, 见表 2:

Table 2. Basic information of structured data
表 2. 结构化数据基本信息

指标	单位	特征重要性	频率	数据来源
期货结算价欧盟排放配额	美元/桶	2.6	日	Wind
期货结算价焦炭	元	3.6	日	Wind
中间价美元兑人民币		4.0	日	Wind
期货结算价核证减排量	吨二氧化碳当量	4.6	日	Wind
期货结算价动力煤	元/吨	4.6	日	Wind
深圳平均气温	摄氏度	9.2	日	Wind

对于我国碳市场，经济形势、金融市场走势、国际碳价和大气环境对我国碳价的影响较大，国际化石能源价格对我国碳价的影响力逐步下降[3]。结合上述筛选结果，欧盟排放配额和核证减排量的是国际碳市场重要的碳资产，其期货结算价可以代表国际碳价水平。焦炭期货结算价和动力煤期货结算价反映了国际化石能源价格，美元兑人民币汇率反映了金融市场走势，深圳平均气温反映了大气环境。因此，本文认为基于大数据筛选出的特征具有一定的稳健性和准确性。

4.2. 文本数据处理

文本数据中又分为新闻关注度和政策力度两方面，其构造方式如下：

- (1) 根据 TF-IDF 算法为每则文本数据中提取出 20 个关键词，设该关键词结合为 $D(t)$ ；
- (2) 将 2014 年 1 月 1 日~2021 年 12 月 31 日所有关键词汇总，提取出词频超过 100 的关键词，作为预标准词；
- (3) 预标准词除去一些与碳无关的词汇后，构成标准词；
- (4) 将每则文本的关键词进行整合，通过简单的拼接，得到每一日的关键词；
- (5) 通过算法计算每日关键词与标准词之间的相似度，本文认为相似度即体现该日文本关注度，相似度高者文本关注度高，相似度低者文本关注度低。

通过上述构造方式分别对新闻类文本和政策类文本进行打分。另外，如果某一日新闻或政策发布文本次数较多，则该日的关注度也应越大。对于这一问题，本文认为，在第四步中，如果一日发布文本次数较多，则该日关键词数量较多，例如发布两次，该日关键词有 40 个。该日关键词越多，则与标准词相似的概率越大，因此认为考虑了文本次数问题。

5. 实证结果

5.1. 基准模型

在以上分析的基础上，为了更好地评价文本信息对碳价预测的作用，本文选择了时间序列模型 ARIMAX 和普通机器学习模型包括 AdaBoost 模型和随机森林模型进行对比，验证文本信息在碳价预测的有效性。ARIMAX 模型具有简单高效，可以处理高维特征等优点，AdaBoost 模型采用迭代思想，每次迭代只训练一个弱分类器，训练好的弱分类器将参与下一次迭代使用；同时，该模型具有计算复杂度低，运行速度快、准确度高等特点。此外，随机森林模型采用“民主投票”方式，即每个基础模型都有一票，最终结果通过所有基础模型投票，少数服从多数的原则产生预测结果，具有训练速度快，能够处理高维数据，抗干扰能力强的特点。

5.2. 实验结果与分析

在上述模型的基础上，实验结果见表 3 所示，可以看出：总体而言，LSTM 模型和 ARIMAX 模型在预测结果上显著优于其他模型。这一现象可能源于 LSTM 模型作为循环神经网络的特性，使其具备较强的长期记忆能力，从而能够更加有效地处理时间序列数据。相比之下，CNN 模型在处理此类问题时则表现相对逊色。此外，ARIMAX 模型作为时间序列分析中的经典代表，其设计也使其在捕捉长期趋势和季节性变化方面具有明显优势，因此与其他基准模型比较中，ARIMAX 展现出了更好的长期记忆能力。另一方面，无论是深度学习模型还是传统的基准模型，在引入文本信息后，预测结果在均方误差(MSE)和平均绝对误差(MAE)方面均有明显降低。这一结果充分表明，文本信息在碳价格预测中发挥了重要作用，增强了模型对复杂市场动态的理解与预测能力。这进一步支持了将文本数据与时间序列模型结合应用的有效性，为未来的研究和应用提供了新的思路与方向。

Table 3. Empirical results of Shenzhen data**表 3.** 深圳数据实证结果

	CNN 加入文本模型	CNN 未加入文本模型
Mse	0.2948	0.6022
Mae	0.5294	0.7711
	LSTM 加入文本模型	LSTM 未加入文本模型
Mse	0.0692	0.0751
Mae	0.2350	0.2468
	ARIMAX 加入文本模型	ARIMAX 未加入文本模型
Mse	0.1093	3.0910
Mae	0.2833	1.7468
	Adaboost 加入文本模型	Adaboost 未加入文本模型
Mse	0.7769	0.8413
Mae	0.8338	0.8850
	随机森林加入文本模型	随机森林未加入文本模型
Mse	0.2592	0.2610
Mae	0.4121	0.4122

5.3. 稳健性检验

为了验证上述结论的稳健性,本文选取上海和广东碳交易所 2014 年 1 月 1 日~2021 年 12 月 31 日的碳价数据进行稳健性检验,结果如下:

Table 4. Empirical results of Shanghai data**表 4.** 上海数据实证结果

	CNN 加入文本模型	CNN 未加入文本模型
Mse	0.3275	0.3786
Mae	0.5624	0.6108
	LSTM 加入文本模型	LSTM 未加入文本模型
Mse	0.0970	0.1231
Mae	0.2791	0.3313
	ARIMAX 加入文本模型	ARIMAX 未加入文本模型
Mse	0.0107	1.1778
Mae	0.0812	1.0816
	Adaboost 加入文本模型	Adaboost 未加入文本模型
Mse	0.2850	0.2906
Mae	0.5261	0.5303

续表

	随机森林加入文本模型	随机森林未加入文本模型
Mse	0.1432	0.1966
Mae	0.3204	0.3570

Table 5. Empirical results of Guangdong data
表 5. 广东数据实证结果

	CNN 加入文本模型	CNN 未加入文本模型
Mse	0.4733	0.5607
Mae	0.6470	0.7149
	LSTM 加入文本模型	LSTM 未加入文本模型
Mse	0.4020	0.4661
Mae	0.5871	0.6349
	ARIMAX 加入文本模型	ARIMAX 未加入文本模型
Mse	0.2915	4.2496
Mae	0.5075	2.0202
	Adaboost 加入文本模型	Adaboost 未加入文本模型
Mse	1.5570	1.9693
Mae	1.1781	1.3336
	随机森林加入文本模型	随机森林未加入文本模型
Mse	0.7774	0.8754
Mae	0.8086	0.8706

从表 4 和表 5 的分析结果可以看出，无论是上海碳交易所的数据还是广东碳交易所的数据，LSTM 模型和 ARIMAX 模型在预测表现上均明显优于 CNN、XGBoost 以及随机森林模型。这一结果充分证明了 LSTM 模型和 ARIMAX 模型在碳价格预测领域的稳健性与有效性。具体来说，这两种模型具备更强的时间序列处理能力，能够更好地捕捉碳价的波动模式和趋势，从而提供更为准确的预测结果。另一方面，当将文本信息纳入模型时，不论是使用深度学习模型还是传统的基准模型，均在均方误差(MSE)和平均绝对误差(MAE)等指标上表现出了显著的提升。这一现象表明，文本数据的引入不仅增强了模型的预测能力，还为碳价的动态变化提供了更多的背景信息和市场情绪洞察。因此，结合文本信息进行碳价格预测的方法显示出其潜在的有效性，为更全面地理解和预测碳市场的变动提供了有力支持。

6. 结论

碳达峰和碳中和的目标将有助于我国在经济社会进步与生态环境保护之间实现平衡，推动可持续发展的模式。而精确的碳价预测对于全面、准确、及时地反映碳排放权交易市场中的各类信息至关重要，这将引导资金在价格信号的指引下迅速而合理地流动，以实现低成本减排的目标。此外，碳金融融入经济体系将促进我国碳市场的进一步发展。

然而, 长期以来, 文本信息的利用受到技术缺乏的限制。随着自然语言处理技术的迅猛发展, 文本信息的利用已变得可行。基于此, 本研究从政策文件和新闻报道中提取了文本信息, 并对其在碳价预测方面的有效性进行了验证。本文以深圳碳交易所为研究对象, 采用 TF-IDF 算法对碳排放交易网络中的政策文本信息和新闻文本信息进行提取。随后, 通过深度学习模型、传统机器学习模型以及时间序列模型对碳价的波动进行预测, 得出以下核心结论: 1) ARIMA 模型和 LSTM 模型在捕捉序列信息方面表现更佳, 其预测效果超越其他模型。2) 文本信息中蕴含着对碳价预测有益的关键数据, 这些数据有助于提高模型的预测准确性。

总的来说, 本文的碳价预测方法具有十分显著的实际应用价值。精准的碳价预测不仅对企业的长远发展至关重要, 也在推动国家节能减排目标的实现方面发挥着关键作用。对于企业而言, 准确的碳价预测能够帮助它们及时掌握市场动态, 从而在竞争激烈的环境中保持领先地位。通过对未来碳价格走势的科学预测, 企业可以提前制定相应的投资策略, 合理安排资源配置, 甚至在必要时采取风险防范措施。这种前瞻性的决策能力, 不仅能够降低运营成本, 还能够在政策变化或市场波动时保持企业的可持续发展。而对于国家层面而言, 碳价的精准预测则有助于全面评估碳市场的整体发展趋势。这些信息能够为政府决策提供科学依据, 使其能够及时实施宏观调控政策, 以应对可能出现的市场不确定性和波动。通过有效的政策干预, 国家能够更好地引导市场, 促进低碳经济的发展, 实现既定的节能减排目标。因此, 碳价预测不仅是企业战略的重要组成部分, 也是国家环保政策制定与实施的基础, 二者相辅相成, 共同推动了可持续发展的进程。

本文对文本信息在碳价预测中的作用进行了深入的探讨, 并验证了文本信息在增强碳价模型预测能力方面的有效性。对于未来的进一步研究, 主要可以从两方面来考虑, 一是进一步提升文本信息的提取效率。可以采用 BERT 等基于注意力机制的模型来深度挖掘文本信息的价值。二是获取更广泛和全面的文本信息。除了新闻文本和政策文本之外, 还应包括市场情绪、突发事件等多维度信息。

参考文献

- [1] Daskalakis, G. and Markellos, R.N. (2009) Are Electricity Risk Premia Affected by Emission Allowance Prices? Evidence from the EEX, Nord Pool and Powernext. *Energy Policy*, 37, 2594-2604. <https://doi.org/10.1016/j.enpol.2009.02.010>
- [2] 谢治菊, 彭智邦. 嵌入式有为与适应性有效: 东西部消费协作中的政府与市场[J]. 中州学刊, 2022(11): 9-19.
- [3] 魏宇, 张佳豪, 陈晓丹. 基于 DMS 和 DMA 的我国碳排放权交易价格预测方法——来自湖北碳市场的经验证据[J]. 系统工程, 2022, 40(4): 1-16.
- [4] 王娜. 基于大数据的碳价预测[J]. 统计研究, 2016, 33(11): 56-62.
- [5] Mansanet-Bataller, M., Pardo, A. and Valor, E. (2007) CO₂ Prices, Energy and Weather. *The Energy Journal*, 28, 73-92. <https://doi.org/10.5547/issn0195-6574-ej-vol28-no3-5>
- [6] Mansanet-Bataller, M. and Pardo, A. (2009) Impacts of Regulatory Announcements on CO₂ Prices. *The Journal of Energy Markets*, 2, 75-107. <https://doi.org/10.21314/jem.2009.019>
- [7] 朱帮助, 王平, 魏一鸣. 基于 EMD 的碳市场价格影响因素多尺度分析[J]. 经济学动态, 2012(6): 92-97.
- [8] Lovcha, Y., Perez-Laborda, A. and Sikora, I. (2022) The Determinants of CO₂ Prices in the EU Emission Trading System. *Applied Energy*, 305, Article 117903. <https://doi.org/10.1016/j.apenergy.2021.117903>
- [9] 王丹舟, 杨德天. 中国碳排放权交易价格的驱动因素[J]. 首都经济贸易大学学报, 2018, 20(5): 87-95.
- [10] 张云. 中国碳交易价格驱动因素研究——基于市场基本面与政策信息的双重视角[J]. 社会科学辑刊, 2018(1): 111-120.
- [11] 吕靖烨, 杨华, 郭泽. 基于 GA-RS 的中国碳排放权价格影响因素的分解研究[J]. 生态经济, 2019, 35(11): 42-47+130.
- [12] 张晨, 胡贝贝. 基于误差校正的多因素 BP 国际碳市场价格预测[J]. 价格月刊, 2017(1): 11-18.

-
- [13] Sun, W. and Zhang, J. (2020) Carbon Price Prediction Based on Ensemble Empirical Mode Decomposition and Extreme Learning Machine Optimized by Improved Bat Algorithm Considering Energy Price Factors. *Energies*, **13**, Article 3471. <https://doi.org/10.3390/en13133471>
 - [14] Yang, S., Chen, D., Li, S. and Wang, W. (2020) Carbon Price Forecasting Based on Modified Ensemble Empirical Mode Decomposition and Long Short-Term Memory Optimized by Improved Whale Optimization Algorithm. *Science of the Total Environment*, **716**, Article 137117. <https://doi.org/10.1016/j.scitotenv.2020.137117>
 - [15] Zhou, J. and Chen, D. (2021) Carbon Price Forecasting Based on Improved CEEMDAN and Extreme Learning Machine Optimized by Sparrow Search Algorithm. *Sustainability*, **13**, Article 4896. <https://doi.org/10.3390/su13094896>
 - [16] Chen, P., Vivian, A. and Ye, C. (2021) Forecasting Carbon Futures Price: A Hybrid Method Incorporating Fuzzy Entropy and Extreme Learning Machine. *Annals of Operations Research*, **313**, 559-601. <https://doi.org/10.1007/s10479-021-04406-4>
 - [17] Jiang, L. and Wu, P. (2015) International Carbon Market Price Forecasting Using an Integration Model Based on SVR. *Proceedings of the 2015 International conference on Engineering Management, Engineering Education and Information Technology*, Chengdu, 6-9 December 2015, 303-308. <https://doi.org/10.2991/emeeit-15.2015.61>
 - [18] 蒋锋, 彭紫君. 基于混沌 PSO 优化 BP 神经网络的碳价预测[J]. 统计与信息论坛, 2018, 33(5): 93-98.
 - [19] Xu, H., Wang, M., Jiang, S. and Yang, W. (2020) Carbon Price Forecasting with Complex Network and Extreme Learning Machine. *Physica A: Statistical Mechanics and Its Applications*, **545**, Article 122830.
 - [20] Wang, J., Cui, Q. and Sun, X. (2021) A Novel Framework for Carbon Price Prediction Using Comprehensive Feature Screening, Bidirectional Gate Recurrent Unit and Gaussian Process Regression. *Journal of Cleaner Production*, **314**, Article 128024.