股市预测中的 NLP

赵婉羽/SDU

摘　　要

股票市场是资本市场的重要组成部分，股市的波动与市场经济息息相关。股票预测问题伴随股票市场的建立而一直存在。无论是在理论还是实践上，人们希望找到有效预测股票走势的方法。然而股票预测问题极具挑战。一方面因为股票预测需要运用经济学、计算机学与数学等多学科知识，另一方面因为股票市场受随机事件的影响而波动不稳定。近几年，随着计算机技术在大数据与人工智能方面的进展，尤其是深度学习在自然语言处理领域取得突破，计算机收集和分析股票市场公开信息的效果越来越好，人们有望在股票预测方面取得新的进展。本文的目的是就 NLP 在股票预测方面的研究提供一个概览，综合现阶段学术和工业界的主流解决方法，对基于深度学习（NLP）的股票预测的流程和指标等进行简要的介绍和总结，以方便后续更有针对性工作的开展。

关键词: 股票预测；NLP；特征抽取

1　引言

整个美国股票市场中有 70%的交易量由算法产生。这 70%的自动交易中很大一部分是高频交易算法，仅试图预测未来几毫秒的结果。它们通常使用非常简单的方法，如一连串的硬编码规则或简单的线性回归模型。

深度学习模型可以学习更复杂的数据模式。但没有人确切地知道是否可以使用深度学习来预测市场中的长期价格走势。许多大型金融机构正以高薪聘用数据科学家、机器学习工程师和深度学习专家，这表明了一种投资策略的趋势。机构投资不同于个人投资。个人买卖的一般是极少量的股票，几乎不会对价格产生任何影响。但如果大量买卖，如何执行交易就会造成很大的不同。机器学习模型可以帮助决定如何随着时间的推移拆分销售，以避免引起大的价格波动。

要实际预测价格走势有很多方法。从非常简单的，如以历史价格训练 LSTM 或时间卷积网络，到极其复杂的模型，如以卫星图像训练卷积 LSTM 来预测宏观经济走势。可能建立的任何预测模型实际上都在试图发现市场上的低效率。例如，模型可以分析来自各种来源（如金融新闻网站和社交媒体）的文本，以确定特定股票的涨跌；可以在字符级别的文本或音频上进行情感分析；不仅可以分析收益表中的内容，还可以分析其公布方式。

在完全高效市场中，这些模型都不会起作用。有效市场假设指出，股票价格会立即反映所有可用的相关信息[1]。当模型可以从网络上解析价格时，影响价格的所有新信息都已包含在价格中。如果认为这个假设是正确的，那么使用价格以外的任何数据都是多余的。

通过查找过去市场数据中的模式来尝试预测价格的方向称为技术分析。技术分析是日间交易员的工作。他们查看图表并命名其所看到的图案，例如头肩顶、杯托、流星线等。神经网络非

常擅长查找数据模式。如果确实存在这样的模式，那么具有足够容量的神经网络将很容易就能选择可能导致获利的任何模式。给一个神经网络一个价格图表，它将尽可能接近该图表的功能。

问题在于，能够找到过去数据中的模式并不意味着这些模式也将在未来能够泛化和保持。神经网络甚至可以在完全随机的数据中找到“模式”。这种模型实际学习的内容不比查找表有用。表中仅包含过去数据的价格信息，但没有预测力，称为过度拟合。并非只有机器学习模型才能在数据中找到不存在的模式。人类也是如此。我们看到云中的面孔和无生命的物体。类似地，在价格图表中看到模式，并假设价格将回归均值。

因此，应对技术分析保持谨慎。关于如何通过技术分析使一些作家致富的书籍和成功故事有很多，但是他们大多数都是通过写如何致富的书或者只是碰巧而变得富有。这种运气的可能性不是很高，但因为失败者几乎没有写过关于失败故事的书，因此存在很多生存偏差。如果在一轮中聚集了 1000 个人并要求他们预测硬币翻转，至少一个人预测连续 8 次翻转的结果的机率超过 98%。只需要一个人的房间就可以找到千里眼。市场上有更多的人在交易。每一千个人就可以找出一个连续 10 年投资回报比大盘好的人；在中国，至少有 100 个人连续 20 年可以做到投资回报比大盘好。

但是，并非所有市场都是有效的或理性的。例如，像对数周期幂定律这样简单的模型就能够在 8 天前预测到 2018 年比特币泡沫。但这并不意味着它将预测下一个，或者是否还会存在另一个加密货币泡沫。行为金融理论认为，证券的市场价格并不只由证券内在价值所决定，还在很大程度上受到投资者主体行为的影响，即投资者心理与行为对证券市场的价格决定及其变动具有重大影响[2] 。这为使用 NLP 预测股市提供了理论基础。

要想用机器学习模型来预测股市，可以先定义模型的目标并选择相应的损失函数。例如，如果希望模型在某些选择中选择最好的股票，则可以将其视为 n 向分类问题，并使用 softmax 交叉熵作为损失函数；如果希望模型对任何给定的股票给出 0 到 1 之间的评分，则可以使用 S 型交叉熵。

接下来继续设计模型架构。不一定非要是复杂的模型。可以堆叠一层或两层 LSTM 或门控循环单元。甚至可以使用简单且易于训练的时间卷积网络。在此基础上，可以进一步使用深度学习模型进行 NLP 训练，为模型增加特征。

关于硬件，当数据量较小时，通常不需要任何特殊的硬件。否则，因模型过于复杂而无法在 CPU 上进行训练很可能会导致过度拟合。若计划使用一些特别大的数据，则可以使用 GPU/FPGA 加速，也可使用云服务等。

在深层模型中使用股票市场数据存在的问题之一是，没有足够的数据来训练大型模型而避免过度拟合。为了减少过度拟合的风险，可以进行各种数据预处理和扩充。例如，为数据添加少量噪音。在每个时期的每个时间间隔随机选择一组股票。并通过对现有库存进行随机线性组合来生成新样本。这些样本的行为本质上应类似于随机管理的共同基金或 ETF。

除了使用价格和数量信息，还可以使用 Google/Baidu 趋势查询来查看有多少人在搜索特定的关键字。但是，这种信号可能会滞后。

需要注意的是，这些也可能都不起作用。

2 价格预测

对给定的任何股票进行价格预测是研究最多的经济应用。在 DL 实现中也有相同的趋势[3]。取决于预测时间范围，从高频交易（HFT）和当日价格变动到每日、每周甚至每月的股票收盘价中选择不同的输入参数。此外，技术分析、基础分析、社交媒体供稿、情感等也用于预测模型的参数。

对股价预测的论文，可以按照研究使用的特征集合进行分类。有的文章仅使用原始时间序列数据（价格、开盘、收盘、最高点、最低点、交易量（OCHLV））；有的研究使用各种其他数据；还有的论文使用了文本挖掘技术。

2.1 原始数据

第一种模型直接使用原始时间序列进行价格预测。表 1 列出了仅使用原始时间序列数据股票价格预测的文章。还根据四个子类别列出了不同的方法/模型：DNN（深度网络，但没有任何给定的拓扑详细信息）和 LSTM 模型；多模型；混合模型；新方法。

DNN 和 LSTM 模型在 3 篇论文中使用。一篇文章中，DNN 和滞后股票被用来预测韩国综合股价指数（KOSPI）中的股价。另一篇文章中将原始价格数据作为输入应用于 LSTM 模型。

同时，有一些研究仅使用原始价格数据（OCHLV）实现**多个 DL 模型**以比较性能。值得关注的研究包括比较了 RNN 、堆叠递归神经网络（SRNN）、LSTM 和 GRU，LSTM、RNN、CNN 和 MLP。在一篇文章中 RNN、LSTM、CNN 和自回归综合移动平均值（ARIMA）被认为更好。另一篇文章比较了 3 种 RNN 模型（SRNN、LSTM、GRU）的股票价格预测，然后根据预测选择了投资组合，并基于此构建了一个阈值。还有一篇文章实现了 DBN 。最后，一位作者比较了 4 种不同的 ML 模型（1 个 DL 模型-AE 和 RBM ）、MLP、径向基函数神经网络（RBF）和极限学习机（ELM）用于预测下 1 分钟后的价格数据，还将结果与不同大小的数据集进行了比较。另一作者使用了价格数据以及 DNN、梯度增强树（GBT）、随机森林（RF）和标准普尔 500 指数（S & P500）中的股票预测方法。在另一篇文章中，合作神经进化、RNN（Elman 网络）和 DFNN 被用于全国证券交易商协会的股价预测自动报价（NASDAQ）中。

同时，在某些论文中使用了**混合模型**。一个作者在研究中使用了 CNN+LSTM。另一些人使用 AE 实现智能索引。一位作者通过预测每只股票的每月对数回报，仅选择预期优于中位数股票表现的股票，将 DBN 和 MLP 结合起来构建了一个股票投资组合。

此外，一些研究还采用了一些**新方法**。一个作者提出了新的深度和广域神经网络（DWNN），它是 RNN 和 CNN 的结合。一些作者在他们的研究中实现了状态频率记忆（SFM）循环学习网络。

2.2 其他指标

在另一组研究中，一些研究人员再次关注基于 LSTM 的模型。但是，它们的输入参数来自各种来源，包括原始价格数据、技术和/或基本面分析、宏观经济数据、财务报表、新闻、投资者情绪等。表 2 列出了在文章中使用了各种数据，如原始价格数据、技术和/或基本面分析、宏

表 1: 仅使用原始时间序列数据的股价预测

数据集	时期	特征集	滞后	时长	方法	性能评价标准	环境
38 stocks in KOSPI	2010-2014	Lagged stock returns	50min	5min	DNN	NMSE, RMSE, MAE, MI	-
China stock market, 3049 Stocks	1990-2015	OCHLV	30d	3d	LSTM	Accuracy	Theano, Keras
Daily returns of 'BRD' stock in Romanian Market	2001-2016	OCHLV	-	1d	LSTM	RMSE, MAE	Python, Theano
297 listed companies of CSE	2012-2013	OCHLV	2d	1d	LSTM, SRNN, GRU	MAD, MAPE	Keras
5 stock in NSE	1997-2016	OCHLV, Price data, turnover and number of trades.	200d	1..10d	LSTM, RNN, CNN, MLP	MAPE	-
Stocks of Infosys, TCS and CIPLA from NSE	2014	Price data	-	-	RNN, LSTM and CNN	Accuracy	-
10 stocks in S&P500	1997-2016	OCHLV, Price data	36m	1m	RNN, LSTM, GRU	Accuracy, Monthly return	Keras, Tensorflow
Stocks data from S&P500	2011-2016	OCHLV	1d	1d	DBN	MSE, norm-RMSE, MAE	-
High-frequency transaction data of the CSI300 futures	2017	Price data	-	1min	DNN, ELM, RBF	RMSE, MAPE, Accuracy	Matlab
Stocks in the S&P500	1990-2015	Price data	240d	1d	DNN, GBT, RF	Mean return, MDD, Calmar ratio	H2O
ACI Worldwide, Staples, and Sea-gate in NASDAQ	2006-2010	Daily closing prices	17d	1d	RNN, ANN	RMSE	-
Chinese Stocks	2007-2017	OCHLV	30d	1..5d	CNN + LSTM	Annualized Return, Mxm Retracement	Python
20 stocks in S&P500	2010-2015	Price data	-	-	AE + LSTM	Weekly Returns	-
S&P500	1985-2006	Monthly and daily log-returns	*	1d	DBN+MLP	Validation, Test Error	Theano, Python, Matlab
12 stocks from SSE Composite Index	2000-2017	OCHLV	60d	1..7d	DWNN	MSE	Tensorflow
50 stocks from NYSE	2007-2016	Price data	-	1d, 3d, 5d	SFM	MSE	-

观经济数据等的文章。在表 2 中，根据五个子类别列出了不同的方法/模型：DNN 模型；LSTM 和 RNN 模型；多模型和混合模型；CNN 模型；新方法。

DNN 模型在一些股价预测论文中使用。一篇文章使用 DNN 模型和 25 个基本特征预测日本指数成分。一篇文章还使用了基本特征和 DNN 模型进行预测。DNN 模型、宏观经济数据（例如 GDP、失业率、库存等）在另一文章中被用于预测美国低水平分类的宏观经济时间序列。

一些研究选择了 **LSTM 和 RNN 模型**。一些作者使用财经新闻和股票市场数的文本挖掘进行迁移学习实现了 LSTM。类似地，另一些作者使用公司行动事件和宏观经济指数和 LSTM 预测股票的下一日价格。另两位作者实现了 DeepStockRanker，一种使用 11 种技术指标基于 LSTM 的股票排名模型。在另一项研究中，作者使用价格时间序列和文本帖子的情感数据，用 LSTM

表 2: 使用各种数据的股价预测

数据集	时期	特征集	滞后	时长	方法	性能评价标准	环境
Japan Index constituents from WorldScope	1990-2016	25 Fundamental Features	10d	1d	DNN	Correlation, Accuracy, MSE	Tensorflow
Return of S&P500	1926-2016	Fundamental Features	-	1s	DNN	MSPE	Tensorflow
U.S. low-level disaggregated macroeconomic time series	1959-2008	GDP, Unemployment rate, Inventories, etc.	-	-	DNN	R^2	-
CDAX stock market data	2010-2013	Financial news, stock market data	20d	1d	LSTM	MSE, RMSE, MAE, Accuracy, AUC	TensorFlow, Theano, Python, ScikitLearn
Stock of Tsugami Corporation	2013	Price data	-	-	LSTM	RMSE	Keras, Tensorflow
Stocks in China' s A-share	2006-2007	11 technical indicators	-	1d	LSTM	AR, IR, IC	-
SCI prices	2008-2015	OCHL of change rate, price	7d	-	EmotionalAnalysis + LSTM	MSE	-
10 stocks in Nikkei 225 and news	2001-2008	Textual information and Stock prices	10d	-	Paragraph Vector + LSTM	Profit	-
TKC stock in NYSE and QQQQ ETF	1999-2006	Technical indicators, Price	50d	1d	RNN (Jordan-Elman)	Profit, MSE	Java
10 Stocks in NYSE	-	Price data, Technical indicators	20min	1min	LSTM, MLP	RMSE	-
42 stocks in China' s SSE	2016	OCHLV, Technical indicators	242min	1min	GAN (LSTM, CNN)	RMSRE, DPA, GAN-F, GAN-D	-
Google' s daily stock data	2004-2015	OCHLV, Technical indicators	20d	1d	$(2D)^2$ PCA + DNN	SMAPE, PCD, MAPE, RMSE, HR, TR, R^2	R, Matlab
GarantiBank in BIST, Turkey	2016	OCHLV, Volatility, etc.	-	-	PLR, Graves LSTM	MSE, RMSE, MAE, RSE, R2	Spark
Stocks in NYSE, AMEX, NASDAQ, TAQ intraday trade	1993-2017	Price, 15 firm characteristics	80d	1d	LSTM+MLP	Monthly return, SR	Python,Keras, Tensorflow inAWS
Private brokerage company' s real data of risky transactions	-	250 features: order details, etc.	-	-	CNN, LSTM	F1-Score	Keras, Tensorflow
Fundamental and Technical Data, Economic Data	-	Fundamental, technical and market information	-	-	CNN	-	-
The LOB of 5 stocks of Finnish Stock Market	2010	FI-2010 dataset: bid/ask and volume	-	*	WMTR, MDA	Accuracy, Precision, Recall, F1-Score	-
Returns in NYSE, AMEX, NASDAQ	1975-2017	57 firm characteristics	*	-	Fama-French n-factor model DL	R^2, RMSE	Tensorflow

预测第二天的股票开盘价。还有一些人使用文字信息和股价通过 Paragraph Vector + LSTM 预测价格，他们还并比较了不同的分类器。还有一项研究使用了技术指标和 JordanElman 网络的股票数据进行价格预测。

也有**多模型和混合模型**主要使用技术分析特性作为 DL 模型的输入。一篇文章中，在 LSTM 和 MLP 网络中加入了一些技术指标，用于进行日内价格预测。一些人使用最小化预测误差损失和方向预测损失的 GAN（GAN-FD）模型进行股票价格预测，并将其模型性能与 ARIMA、ANN 和支持向量机（SVM）进行比较。另一篇文章作者使用了主成分分析（PCA）几种技术指标特征和时间序列数据来降维，并与 DNN（2 层 FFNN）级联，用于股票价格预测。在一篇文章中，作者使用基于市场微观结构的贸易指标作为 RNN 的输入，Graves LSTM 检测伊斯坦布尔证券交易所指标（BIST），以便执行智能股票交易的价格预测。在另一篇文章中，能够预测下个月的回报，并建立业绩最佳的投资组合。LSTM 和 LSTM-MLP 模型获得了良好的每月回报。

同时，在某些论文中，**CNN 模型**更受欢迎。有一些作者使用了 250 个特征：订单明细等，用于预测私人经纪公司的风险交易的真实数据。他们使用 CNN 和 LSTM 进行股价预测。另一位作者使用 CNN 模型、基本面和技术分析，以及市场数据进行预测。

在一些研究中还开发了**新方法**。在一篇文章中，FI-2010 数据集：出价/要价和交易量用作预测的特征集。在这个研究中，他们提出了加权多通道时间序列回归（WMTR）、多线性判别分析（MDA）。另一篇文章的作者使用了 57 个特征，如市场份额、市场 Beta、行业发展势头、资产增长等，作为 Fama-French n 因子 DL 模型的输入，预测纽约证券交易所（NYSE）、美国证券交易所（AMEX）或纳斯达克的美国月度股票回报率。

2.3 文本挖掘

有许多研究论文也使用文本挖掘技术来进行特征提取，但使用非 LSTM 模型进行股价预测。表 3 列出了使用文本挖掘技术预测股票价格的文献。在表 3 中，不同的方法/模型分为三个子类别：CNN 和 LSTM 模型；GRU、LSTM 和 RNN 模型；新方法。

某些论文中使用了 **CNN 和 LSTM 模型**。在一篇文章中，事件是通过文本挖掘从路透社和彭博新闻中检测到的，并且该信息通过 CNN 模型用于价格预测和股票交易。另一篇论文中，作者通过 LSTM+CNN 混合模型对路透社的 S & P500 指数新闻进行文本挖掘，一起进行价格预测和日内方向移动估计。还有一篇文章的作者使用金融新闻数据，并通过 Word2vec 以及 MA 和随机振荡器实现了词嵌入，来为用于股价预测的循环 CNN（RCNN）创建输入。还有一文章的作者也通过词嵌入和文本挖掘对分析报告进行了情感分析，并将情感特征用作 DFNN 的输入来预测股价。然后根据预计的股票收益选择不同的投资组合。

在这组论文中，**GRU、LSTM 和 RNN 模型**更受欢迎。一些人在 Twitter 帖子上进行了情感分析，并使用 RNN 预测股票数据价格。类似地，另一些作者通过各种 LSTM 模型使用情感分类（中立、正、负）进行股票开盘价或收盘价预测。他们将其结果与 SVM 进行了比较，并获得了更高的整体性能。在另一篇文章中，文本和价格数据用于预测上证综合指数（SCI）价格。

在某些论文中还出现了一些**新方法**。一些作者使用了词嵌入从网页中提取信息，然后与股价数据结合预测股价。他们比较了自回归（AR）模型和带有/不带有新闻的 RF 。结果表明，嵌入的新闻信息可以改善性能。在另一篇文章中，使用了财经新闻和 ACE2005 中文语料库。基于一

表 3: 使用文本挖掘技术进行特征提取的股价预测

数据集	时期	特征集	滞后	时长	方法	性能评价标准	环境
S&P500 Index, 15 stocks in S&P500	2006-2013	News from Reuters and Bloomberg	-	-	CNN	Accuracy, MCC	-
S&P500 index news from Reuters	2006-2013	Financial news titles, Technical indicators	1d	1d	RCNN	Accuracy	-
TWSE index, 4 stocks in TWSE	2001-2017	Technical indicators, Price data, News	15d	-	CNN + LSTM	RMSE, Profit	Keras, Python, TALIB
Analyst reports on the TSE and Osaka Exchange	2016-2018	Text	-	-	LSTM, CNN, Bi-LSTM	Accuracy, R-squared	R, Python, MeCab
Stocks of Google, Microsoft and Apple	2016-2017	Twitter sentiment and stock prices	-	-	RNN	-	Spark, Flume, Twitter API
Stocks of CSI300 index, OCHLV of CSI300 index	2009-2014	Sentiment Posts, Price data	1d	1d	Naive Bayes + LSTM	Precision, Recall, F1-score, Accuracy	Python, Keras
SCI prices	2013-2016	Text data and Price data	7d	1d	LSTM	Accuracy, F1-Measure	Python, Keras
Stocks data from S&P500	2006-2013	Text data and Price data	7d	1d	LAR+News, RF+News	MAPE, RMSE	-
News from Sina.com, ACE2005 Chinese corpus	2012-2016	A set of news text	-	-	Their unique algorithm	Precision, Recall, F1-score	-

种新的事件类型模式分类算法对不同的中国公司的事件类型进行了分类，使用额外的输入，第二天的股价变化也得到了预测。

3 指数预测

一些研究人员尝试预测股市指数，而不是预测单个股票的价格。一般而言，由于指数由来自不同部门的多个股票组成，其波动性要低于单个股票，因此在总体动力和总体经济状况上更具指示性。在文献中，不同的股市指数数据被用于实验。最常用的指数数据如下：S & P500、中国证券指数（CSI）300、印度国家证券交易所（NIFTY）、东京日经指数（NIKKEI）225、道琼斯工业平均指数（DJIA）、上海证券交易所（SSE）180、香港恒生指数（HSI）、深交所综合指数（SZSE）、伦敦金融《金融时报》证券交易所指数（FTSE）100、台湾市值加权股票指数（TAIEX）、BIST、纳斯达克、道琼斯工业平均指数 30（DOW30）、KOSPI、标普 500 波动率指数（VIX）、纳斯达克 100 波动率指数（VXN）、巴西股票前变更（Bovespa）、斯德哥尔摩证券交易所（OMX）、纽约证券交易所。

此外，在文献中，有多种预测指数数据的方法。一些研究仅使用原始时间序列数据，而另一些研究则使用其他各种数据，例如技术指标、索引数据、社交媒体供稿、路透社、彭博社的新闻、数据的统计特征（标准差，偏度，峰度，ω 比率，基金 α）。对指数预测文章根据其特征集进行分组，可以分为仅使用原始时间序列数据（价格/指数数据，OCHLV）进行的研究，和使用

各种其他数据的研究。

4 特征选择

不管潜在的预测问题如何，原始时间序列数据几乎总是直接或间接地嵌入到特征向量中，这尤其对基于 RNN 的模型有效。但是，在大多数其他模型类型中，其他特征也包括在内。基本面分析和技术分析特征是股票/指数预测研究中最多的选择。

同时，近年来，金融文本挖掘尤其受到关注，主要用于提取投资者/交易者的情感。金融新闻的传输流、推文、声明和博客可让研究人员整合数字和文本数据，建立更好、更通用的预测和评估模型。通用的方法包括通过文本挖掘来提取经济情感分析，并将其和基本/技术分析数据结合起来，以实现更好的整体性能。可以合理地认为，这种趋势将随着更先进的文本和 NLP 技术的综合而继续。

4.1 历史

最后一个被发现的 “特别灵” 的股票技术指标（信息）是著名投资人 Peter Lynch 发明的 PEG（市盈率相对盈利增长比率）。Peter Lynch 闷声发大财，从来没有告诉别人，因此创造了股市上不败的神话。但是他四十多岁就收手了，因此很多人认为如果他再炒下去，难免失败。

可惜的是，很难再像 Peter Lynch 那样找到一个新的技术指标。这种特别管用的信号在早期股市上是存在的，在今天几乎已经找不到了，因为容易找的早被人找光了。对于那些影响力不是很大的技术指标 (信号)，可能还存在，但这不是散户可以找到的，甚至不是绝大多数专业人士可以找到的，因为它们本身太复杂，而且作用也不是很明显，因此很难找到。即便找到了，由于单个信号实在对股价预测的影响太小，基本上没有用。如果能找到一大堆，或许组合起来管用，但这是一个非常非常难的机器学习问题，连 Google 和微软这样的公司也不一定做得到。

事实上，全世界可能只有著名的对冲基金文艺复兴公司敢说找到了很多有用的信号，而且能够组合得很好。公司创始人 James Simons 在 2011 年的 TED 采访中曾透露，公司的预测方案会考虑各种事物，各种对工作有价值的东西——天气、年报、季报、历史数据、成交量... 应有尽有。一天内接收 TB 级的数据，储存、处理、分析，寻找异常现象。

5 未来工作[3]

最有可能的是，在可预见的将来，金融时间序列预测将会像以往一样，与算法交易、投资组合管理等其他金融应用领域紧密合作。但是，可用数据中的特性变化和新资产类别的引入可能不仅会改变开发人员的预测策略，也迫使开发人员寻找新的或替代的方法技术，以更好地适应这些新的挑战性的工作条件。此外，像连续排名概率评分（CRPS）这样的用于评估概率分布的指标可能会被包含在内，以进行更全面的分析。

人机交互和 NLP 研究不仅在财务时间序列预测，而且在所有智能决策支持系统中都呈一种上升趋势。其中，文本挖掘和金融情感分析对金融时间序列预测尤为重要。行为金融可能会从这些领域的新进展中受益。

为了利用文本挖掘的力量，研究人员开始开发像 Stock2Vec 这样新的数据表示，可用于组合文本和数字数据以获得更好的预测模型[4] 。此外，集成了数据语义和时间序列数据的基于 NLP 的集成模型可能会提高现有模型的准确性。

可从相互联系的金融市场中受益匪浅的一个领域是自动统计套利交易模型的发展。它之前已用于外汇和商品市场。此外，由于在各种市场上存在大量的可用货币，目前有许多从业者在加密货币市场中寻求套利机会[5]。价格中断，高波动性，买卖价差差异导致不同平台的套利机会。一些机会主义者开发软件模块，跟踪这些价格异常以立即获取利润。此外，还有可使用适当的模型构建跨不同资产类别的交易资产组合。DL 模型比传统的基于规则的系统更快更有效地学习（或预测）这些机会是可能的。这也将有益于 HFT 研究，不断寻找延迟最小的、更快、更有效的交易算法和嵌入式系统。为此，可以使用嵌入 DL 模型中的图形处理单元（GPU）或现场可编程逻辑门阵列（FPGA）的硬件解决方案。在金融时间序列预测和算法交易方面缺少硬件研究。因为回报很高，只要有足够的计算能力，研究更好的算法是值得的。

参考文献

[1] FAMA E F. The behavior of stock-market prices[J/OL]. The Journal of Business, 1965, 38(1):34-105. http://www.jstor.org/stable/2350752.

[2] 行为金融学[EB/OL]. 2021. https://zh.wikipedia.org/wiki/%E8%A1%8C%E4%B8%BA%E9%87%91%E8%9E%8D%E5%AD%A6.

[3] SEZER O B, GUDELEK M U, OZBAYOGLU A. Financial time series forecasting with deep learning : A systematic literature review: 2005-2019[J]. ArXiv, 2020, abs/1911.13288.

[4] Lien Minh D, Sadeghi-Niaraki A, Huy H D, et al. Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network[J/OL]. IEEE Access, 2018, 6:55392-55404. DOI: 10.1109/ACCESS.2018.2868970.

[5] FISCHER T G, KRAUSS C, DEINERT A. Statistical arbitrage in cryptocurrency markets[J]. Journal of Risk and Financial Management, 2019, 12(1):31.