

深圳大学

本科毕业论文（设计）

题目： 基于趋势信号的事件驱动型
期货交易策略研究

姓名： 邹杰

专业： 数学与应用数学

辅修学士学位专业： 金融学

学院： 数学与统计学院

学号： 2019193009

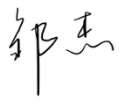
指导教师： 孙正佳

职称： 助理教授

2023 年 3 月 30 日

深圳大学本科毕业论文（设计）诚信声明

本人郑重声明：所呈交的毕业论文（设计），题目《基于趋势信号的事件驱动型期货交易策略研究》是本人在指导教师的指导下，独立进行研究工作所取得的成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式注明。除此之外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。本人完全意识到本声明的法律结果。

毕业论文（设计）作者签名： 
日期： 2023 年 4 月 12 日

目录

1 绪论	2
1.1 研究背景与意义	2
1.2 国内外相关研究	2
1.3 本文主要研究工作	4
2 基础理论知识	6
2.1 爬虫技术	6
2.1.1 爬虫技术介绍	6
2.1.2 Selenium: 模拟点击技术	6
2.2 自然语言处理	7
2.2.1 Jieba 中文分词工具	7
2.2.2 Word2vec: 词嵌入技术	7
2.3 基于金融事件的情感词典	8
2.4 Logistic 回归	8
2.5 Backtrader 量化平台	9
3 模型构建	10
3.1 数据选取	10
3.2 数据预处理	10
3.2.1 数据清洗	10
3.2.2 Jieba 分词	11
3.3 基于金融事件的情感词典构建	11
3.4 Logistic 回归生成事件情绪信号	12
3.5 Backtrader 量化平台回测	13
4 实验结果与分析	15

4.1 实验设计	15
4.2 评价指标	15
4.3 RB2305 描述性统计分析	16
4.4 滑动窗口值调优	17
4.5 交易策略比较	18
5 总结与展望	21
5.1 总结	21
5.2 不足与展望	21
参考文献	22
致谢	24
Abstract	24

基于趋势信号的事件驱动型期货交易策略研究

数学与统计学院 数学与应用数学 邹杰

学号：2019193009

【摘要】本文主要对基于趋势信号的事件驱动型期货交易策略进行研究，以上海期货交易所的螺纹钢期货 RB2305 为交易标的，爬取了东方财富网中螺纹钢期货板块新闻数据，接着基于金融事件的情感词典、词嵌入技术设计并构建了事件情绪信号，然后通过 Logistic 回归训练出事件情绪信号的分类器，最后使用 Backtrader 量化平台构建了单均线和双均线两种回测环境，以普通的趋势交易策略为基准模型来比较本文提出的事件驱动型交易策略。实验结果表明，在应用了事件情绪信号后，能够过滤大部分错误趋势信号，使得交易策略表现更好更稳健，证实了本文提出的基于趋势信号的事件驱动型策略的有效性。本文构建的基于趋势信号的事件驱动型期货交易策略最重要的价值在于，它提供了一种可移植的有效的策略构建方向，因为受到交易标的的波动小、交易时间较短的限制，所以实验的结果受到了一定的约束。若更换为价格波动大、交易量多的期货标的，则可以期待本文构建的事件情绪信号能够发挥更大的价值。

【关键词】量化交易；螺纹钢期货；Logistic 回归；词嵌入；金融事件情感词典

1 绪论

1.1 研究背景与意义

量化投资是一种使用数学模型、算法和计算机程序来确定投资组合和交易决策的投资策略。量化投资的目的是在把人类的主观判断和偏见自动化的同时，通过分析大量的数据和市场信息来优化投资决策。20 世纪初，第一批量化投资策略出世。该领域的先驱之一本杰明·格雷厄姆提出了价值投资的方法，即通过基本面分析买入那些被低估的股票。在 20 世纪 50 和 60 年代，在哈里·马科维茨和威廉·夏普等学者的带领下，涌现出了新一批量化投资者。马科维茨提出了一种同时考虑了风险和收益的投资组合理论，即使用均值和方差分别代表投资标的的预期收益和风险。夏普提出了资本资产定价模型（CAPM），该模型通过无风险收益率、市场的预期收益率以及投资组合对市场的敏感度这几个参数来估计投资组合的预期收益率。20 世纪 70、80 年代，基于算法和计算机程序的交易策略涌现。随着信息技术的发展，投资者可以使用电脑编写算法来处理大量数据，这也逐渐成为量化投资者交易的主流方式。因此在这段时间，出现了统计套利策略和趋势跟踪策略。20 世纪 90 年代、21 世纪初，量化投资进一步发展，出现了高频交易策略和基于机器学习的交易策略。高频交易善于捕捉 tick 级数据中的交易信号，通常在当天买入当天清仓，而机器学习方法则被用来开发预测模型，进而预测金融数据的趋势。如今，量化投资策略是金融业不可或缺的一部分，许多对冲基金和资产管理公司依靠量化技术来创造回报。同时，大数据和人工智能的发展可能会继续推动新的量化策略的发展。

目前，工业界使用的主要策略包括：股票策略、市场中性策略、管理期货策略、CTA 趋势策略、CTA 套利策略、债券策略、指数增强策略、高频策略、事件驱动策略等。根据《2021 年中国量化投资白皮书》，所调研的私募或公募机构中，超八成已使用人工智能来挖掘因子及构建模型，其中，事件驱动型策略只占比 1%。事件驱动型策略一般用于微观上选股，从而获取个股的 Alpha 收益。事件驱动型策略将事件可以分为三类：第一类是公司的公告，比如公司并购、股票回购、业绩预告等；第二类是公司披露的财务报表，投资者可能从企业盈利性、资金流动性、财务杠杆等方面衡量投资标的，通过财务报表中披露的数据，计算出一些财务指标比如市盈率、市净率、流动比率、资产负债率等；第三类是新闻，可以是该公司的新闻，也可以是行业新闻或者宏观经济新闻，但三种新闻与金融资产的相关性不同，普通投资者较难直接从新闻中判断趋势，这类研究做的相对最少。因此，事件驱动交易是通过利用市场的低效率来赚取利润，因为总体来说新闻或事件出现的速度是快于市场反应的，此时新闻或事件的价值没有被完全计入股票的价值，特殊情况就是小部分机构投资者能提前掌握市场内幕消息，从而能够在官方新闻发布之前就率先采取行动。所以，这种方法通常由对冲基金和其他机构投资者使用，他们拥有资源和专业知识来研究和分析可能影响特定股票、期货或期权的潜在事件，从而判断是买入、卖出还是持有。

从理论角度来看，对事件驱动策略的研究可以为金融市场的行为以及特定事件对股票价格的影响提供有价值的见解。这项研究可以通过帮助开发新的市场行为理论和模型，以及更好地理解股票价格运动的基本驱动因素，为金融领域作出贡献。从实践的角度来看，对事件驱动策略的研究可以为那些有兴趣实施这些策略的投资者提供宝贵的信息，包括更好地理解驱动市场价格变化的因素，选择和执行交易的方法。事件驱动研究的另一个重要的实际意义是开发新的和改进投资工具和技术。例如，大数据分析和机器学习算法的使用在事件驱动策略中已经变得越来越普遍，使投资者能够分析大量的数据，并更快、更有效地识别潜在机会。这些工具还可以帮助投资者更好地管理风险，对市场走势进行更准确的预测，并帮助发现潜在的市场异常情况。

1.2 国内外相关研究

在量化投资领域，目前学术界主要的研究的方向是机器学习、深度学习方法在该领域的适用性。Wu（2022）提出了一个基于模糊分析方法的系统，它可以对适合动量或逆向策略的股票进行分类，在台湾 50 数据集上，它的盈利能力提高了 1.5 倍。Syu（2023）介绍了 TripleS，一个利用模糊集理论建立股票和投资策略之间联系的选股系统。除了利用 LSTM 提取股票价格序列的时间序列特征外，一些学者还将股票及其领先指标（期货/期权）价格序列表示为图形数据，然后利用 CNN 提取特征。Wu（2020）提出了一种用于训练 CNN 网络预测股市的二维张量输入数据和特

征提取方法，该方法在避免噪声和过拟合方面优于以往的算法。Wu (2021) 提出了一个基于 CNN 和 LSTM 的新框架，通过聚合多个变量，通过 CNN 自动提取特征，并将其输入 LSTM 来预测股票市场的走向。HIST 是微软亚洲研究院的 Xu (2021) 等在 2022 年开发的一个高频交易模拟器。它的目的是为开发和测试高频交易算法提供一个真实的环境。Qlib 是微软亚洲研究院的 Yang (2020) 等在 2020 年开发的开源 Python 库，支持各种深度学习、强化学习和传统机器学习模型。由于股票数据是时间序列，学者们通常将时间序列分解为不同的频谱来提取特征，并从中得出经验模态分解 (EMD) 和完全集成的经验模态分解 (CEEMD) 算法。Rezaei, Hadi (2021) 在此基础上建立了 CEEMD-CNN-LSTM 和 EMD-CNN-LSTM 混合算法，并与 LSTM 模型相结合，实验表明混合模型的性能优于其单独的对对应物。Milon Biswas (2021) 使用长短期记忆、XGBoost、线性回归、移动平均和终值模型等算法，对超过 12 个月的历史股票数据建立预测模型，观察到 LSTM 方法的表现优于其他所有方法，五个模型中错误率最高的是移动平均模型。KHALED A. ALTHELAYA (2021) 将深度学习技术与多分辨率分析相结合来预测股票，该模型基于经验小波变换，所提出的模型被用于 S&P500 指数和 McKee-Glass 时间序列，证明比其他模型有效得多。Jing, Nan (2021) 首先使用卷积神经网络对股票投资者的情绪进行分类，并使用 LSTM 算法对股票的技术指标进行分析，在上海证券交易所的六个主要板块上进行了实验验证，结果显示混合算法的表现优于单一模型以及没有情绪分析的模型。Sinha, Siddhant (2022) 提出了一种基于 3D 卷积神经网络的方法来对股票价格的方向趋势进行分类。为此，将一个行业的五家公司组合在一起，同时预测每家公司的总体趋势。对于每家公司，选择多个技术指标，并将股票价格转换为大小为 3D 的图像。结果表明在某些股票实现了高达 35% 的年回报率，平均为 9.19%。

在事件驱动领域，已经有研究集中在特定类型的公司事件，例如，研究人员调查了并购活动如何影响目标公司的股票价格和收购方公司的异常回报。其他研究则考察了收益公告对股票价格的影响，研究了市场对正面或负面消息的反应。现在，对事件驱动的股票交易策略的研究重点更加广泛，包含了更多的事件，包括收益公告、并购活动、破产公告和内幕交易。这些研究还探索了分析事件驱动数据的不同方法，包括统计分析、机器学习和人工智能。Fang (2019)、Zhang (2010)、Chen (2020) 和 Yang (2020) 都在研究中提及来自财务报表的基本面数据和来自商业新闻的其他数据等与机器学习算法相结合，可以获得投资信号或对公司的前景做出预测，从而筛选出好的投资目标。研究人员试图找出事件和股票价格变动之间的模式和关系，目的是开发出能够产生异常收益的交易策略。

在国内，Ding Xiao, Liu Ting 团队对事件驱动研究作出了很大贡献。2015 年，Ding Xiao (2015) 提出了使用深度卷积神经网络的方法建模新闻标题中事件的词向量分别在短期和长期上对 S&P 股价运动的影响。他们的方法比当时最新的基线模型在 S&P500 指数和个股上收益率高出 6%，证明了事件驱动策略的有效性。2019 年，Ding Xiao 团队 (2019) 提出了融合了外部信息的事件表征方法，这种方法能够更好地预测股市的波动性。微软中国的 Xu Wentao 团队 (2021) 在 2021 年提出了 REST 模型，一种基于关系事件的股票趋势预测框架，他们对股票背景进行建模，学习事件信息对不同背景下股票的影响。同时，构建了一个股票图，并设计了一个新的传播层来传播相关股票的事件信息的影响。模拟投资结果表明，他们的框架可以实现比基线更高的投资回报。黄亮点等 (2021) 深度挖掘了财经新闻与期货市场的相关性，通过引入基于新闻文本情绪分析的决策支持系统，将其与传统交易策略相结合，形成了事件驱动程序化交易策略。李备友等 (2021) 通过对事件驱动下的股票价格走势分析发现，事件对股票价格的波动率影响与该事件的作用强度、影响幅度和作用时间相关，并用 S-曲线定量描述了事件驱动的股票价格波动特性。祝清麟 (2022) 提出一种基于金融领域情感词典的注意力机制来为不同实体获取重要的情感信息，并与基准模型相比取得了更好的效果，构建了一个较为通用的情感分析模型。与本文的新闻情绪信号不同，林培光 (2022) 则是通过股民评论提取出今日情感权重，再将今日情感权重及其若干权重均线值与股价一起放入卷积神经网络和长短期记忆网络进行训练，发现该模型能够在小样本下有效预测一些蓝筹股如阿里巴巴和格力电器的股价走势。由于卷积神经网络的情感分析方法存在标注数据分布不平衡问题，刘玉玲 (2022) 则利用生成对抗网络来预测股市的波动，其中生成器生成股票序列数据，判别器则采用卷积神经网络对生成数据和真实数据进行区分。结果显示，该方法能动态地更新股价预测结果并且误差较小。黄进 (2015) 认为用户情感倾向与市场波动之间存在联系，对金融市场监管和股价异常处理具有重要作用。他充分考虑了金融领域情感词的特征、

单个句子中词语的位置权重以及情感词相互间的修饰关系，提出了结合 Stanford 句法依存分析方法的基于 SVM 分类的文档情感值计算方法，从而实现对整篇文章进行情感分析。通过实验证明该方法相较于其他方法更为精准。刘薇（2022）则通过集成了金融 Bert（Financial Bert）模型与卷积神经网络模型，提升了对股吧评论的情感分类准确率，并通过关联性分析验证了投资者情绪与股市走势之间存在相关性。许雪晨（2021）研究了如何对媒体报道、公司新闻等非结构化数据进行文本分析，并将其应用于股票价格波动预测。她提出了一种基于金融文本情感分析的指数预测模型，命名为 SA-BERT-LSTM，该模型用于对沪深 300 指数的涨跌进行预测，并在实验中有效提高了股指趋势预测的准确率。

在国外，Rubi Gupta（2021）分析了 StockTwits 的推文内容，并使用一套文本特征化和机器学习算法提取金融情绪。然后研究汇总的每日情绪和每日股票价格变动之间的相关性。最后，在过去的股票时间序列数据的基础上，情绪信息被用来提高股票价格变动预测的准确性。通过对五家公司（苹果、亚马逊、通用电气、微软和塔吉特）的实验，使用九个月的 StockTwits 数据和每日股票数据，证明了所提出的工作对股票价格预测的有效性。Zhou Zhihan（2021）构建了基于新闻稿件的事件驱动策略，核心是一个双层次的事件检测模型。低层次的事件检测器从每一节中识别出事件的存在，而高层次的事件检测器则将整个文章的表述和低层次的检测结果结合起来，在文章层面上发现事件。他们还开发了一个精心注释的数据集 EDT，用来进行事件检测和得到基于新闻的股票预判基准。对 EDT 的实验表明，该策略在获胜率、相对于市场的超额收益以及每笔交易的平均收益方面都优于所有的基准。Narayana Darapaneni（2022）利用基于 LSTM 和随机森林的组合模型，通过历史价格和情感数据来预测股票的未来走势，LSTM 以历史价格为输入量，使用强度分析器捕获的情绪分析指数被用作随机森林模型的主要参数，一些宏观参数如黄金、石油价格、美元汇率和印度政府证券收益率也被添加到模型中以提高模型的准确性。最后，使用上述两个模型预测了 4 只股票的价格，并使用 RMSE 指标对结果进行了评估，发现使用 LSTM 的预测误差小于 ARIMA 以及线性回归。Padmanayana（2021）的目标是利用历史股票数据与新闻头条和 Twitter 帖子的情感分析相结合，来预测感兴趣的股票的未来价格。他们首先使用 API 来获取特定公司的实时 Twitter 推文，并把所有的停顿词、特殊字符都从数据集中提取出来。过滤后的数据使用 Naïve bayes 分类器进行情感分析。因此，推文被分为正面、负面和中性推文。为了预测股票价格，股票数据集从雅虎金融 API 中获取。股票数据与推文数据一起被作为机器学习模型的输入，以获得结果。XGBoost 分类器被用来作为预测股票市场价格的模型。获得的预测值与实际的股票市场价值进行比较。通过对苹果、亚马逊、微软等几家公司的实验，使用实时 Twitter 数据和每日股票数据，证明了模型对股票价格预测的有效性。

1.3 本文主要研究工作

为了挖掘金融领域内新闻文本的情绪价值对投资产品未来涨幅预期的影响，本文旨在构建一个基于趋势信号的事件驱动型期货交易策略。选取上海期货交易所中 2023 年 5 月执行的螺纹钢期货（RB2305）作为标的资产，研究东方财富网中期货频道内钢材资讯板块的新闻对螺纹钢期货未来预期涨幅的影响程度。选取基础的趋势策略，比如单均线策略和双均线策略作为基准模型，对 RB2305 这一品种进行回测得到基准策略收益，再同本文的基于趋势信号的事件驱动型策略的收益进行比较。比较两者的累计收益率、交易次数、交易胜率等指标能够直观地反映出事件驱动型策略的价值。

首先使用 Selenium 搭建基于模拟点击的爬虫方法爬取东方财富网中钢材资讯内的所有新闻，共 1000 条，时间跨度从 2022 年 3 月 23 日到 2023 年 2 月 17 日。因此，为了充分利用新闻数据，选取的期货的交易时间最好能够充分涵盖存在新闻的时间。爬取的新闻经过数据清洗、排序、分词等预处理后，原有的字符串类型的新闻标题转化为一个个列表，列表内的元素为字符串类型的新闻切分词语。接着，因为金融新闻带来的情绪是通过经济规律、金融世界运行法则体现的，因此我们有必要从宏观、微观经济学出发，基于新闻中常见的事件类型，构建一个专属于钢材期货的金融情感词典，包括了 5 类：供给需求、市场价格、资金技术、经济政策、宏观环境。以螺纹钢这一品种专属的期货情感词典融合了经济学的底层逻辑，意味着形成的情绪信号具有较强的理论依据作为支撑。同时，每个大类包含正向情感和负向情感两个子类，因此一共 10 类事件。针对影响的不同，包括直接影响和间接影响，可以将这 10 类事件赋予不同的分值。对于供给需求和市场价格这两大类，事件的发生我们认为会对期货价格造成直接的影响，因此对其中的正向情感事件赋予 2 分、对其中的负

向情感事件赋予-2 分。对于资金技术、经济政策和宏观环境这三大类，事件的发生我们认为会对期货价格造成间接的影响，意味着它们的发生的作用时间会相对更长，并且随着时间的推移，间接作用的影响可能会逐渐被其他事件所淡化。因此对其中的正向情感事件赋予 1 分、对其中的负向情感事件赋予-1 分。基于 word2vec 模型中的计算两段文本的相似度方法，我们能够实现对爬取的 1000 条新闻数据中每一条新闻进行分类，并赋予情绪分值。单从某一天的情绪分值的高低去预测未来的资产价格变化是不具有说服力的，特别是当基准模型是一个趋势策略时。对于预测任务，可认为是一个预测未来一天价格涨跌的二分类模型。本文建立了 Logistic 模型来拟合情绪分值与未来一天的价格涨跌之间的关系，训练好后进行预测时，会根据输入的情绪分值输出 0 或 1，0 代表预测明天价格会跌，1 代表预测明天价格会涨。RB2305 的历史交易数据来源于东方财富网，时间跨度是 2022 年 5 月 17 日到 2023 年 2 月 17 日。具体建模时，本文计算了以一段时间 T 作为滑动时间窗口，在当前时间 t 往过去看时间步长 T 天，这段时间内的情绪总分值 S_t^T ，并以情绪总分值 S_t^T 作为自变量 X ，以当前时间 t 的未来一天价格 P_{t+1} 相较于当前时间的价格 P_t 的涨跌情况 ΔP_t 作为因变量 y ，若涨则令 $\Delta P_t = 1$ ，反之则为0。对于训练好的 Logistic 模型，输入时间步长为 T 的情绪总分值 S_t^T ，将输出的预测信号 ΔP_t^{Pred} 作为事件情绪信号。使用 Backtrader 进行回测时，将单均线、双均线策略作为基准模型，与添加了事件驱动的情绪信号的趋势策略进行对比。事件驱动型策略运作时，只有当趋势信号和事件情绪信号 ΔP_t^{Pred} 相同时才进行买入和卖出。最后，通过可视化展示事件驱动型策略的价值。

2 基础理论知识

2.1 爬虫技术

2.1.1 爬虫技术介绍

爬虫 (Web Scraping) 是指通过程序自动化地访问互联网上的信息, 从中提取并整理数据的过程。通常爬虫程序通过网络请求访问网站, 并解析网页源代码中的特定信息, 例如文本、图片、链接等。爬虫技术在数据采集、信息监控、网络分析等领域都有着广泛的应用。

爬虫的起源可以追溯到 20 世纪 90 年代初期, 当时互联网的规模还很小, 网页的数量和种类都非常有限。最早的爬虫是由学术机构和搜索引擎开发的, 用于索引互联网上的文献和网页。后来, 随着互联网的普及和发展, 爬虫逐渐被用于商业领域, 如数据挖掘、网络营销等。随着时间的推移, 爬虫技术得到了不断的发展和完善。

如今, 爬虫已经成为一种非常成熟的技术, 能够在不同的操作系统和开发环境中进行应用。比如, 爬虫可以帮助搜索引擎抓取互联网上的信息, 构建搜索引擎的索引库, 使用户可以更快地找到他们需要的信息。此外, 爬虫还被广泛应用于数据采集、市场调研、舆情监测、商业情报等领域。除此之外, 爬虫还被用于一些互联网产品的开发中。例如, 社交媒体平台可以通过爬虫收集用户发布的内容, 进行数据分析和推荐算法的优化。电子商务网站可以通过爬虫收集竞争对手的价格和促销信息, 进行商品定价和营销策略的调整。当然, 爬虫技术也存在一些潜在的问题和风险。例如, 某些爬虫可能会侵犯网站的版权和隐私, 或者被用于进行网络钓鱼、网络欺诈等非法活动。因此, 针对这些问题, 相关部门和组织也会采取一系列的技术和法律措施来加以规范和管理。

2.1.2 Selenium: 模拟点击技术

Selenium 是一个流行的自动化测试工具, 主要用于 Web 应用程序的功能测试和自动化测试。Selenium 可以模拟用户在浏览器中的行为, 包括点击、输入、选择等, 从而实现对 Web 应用程序的自动化测试。除此之外, Selenium 还可以用于数据采集和网站爬取。如果没有模拟点击事件, 那么就很难进行这些操作, 也无法获取到相关的页面数据。

Selenium 可以与多种编程语言进行集成, 如 Java、Python、C# 等。在使用 Selenium 时, 首先需要安装相应的浏览器驱动, 例如 Chrome Driver 或 Firefox Driver 等, 这些驱动程序可以与不同的浏览器进行通信, 以模拟用户在浏览器中的行为。安装完成后, 使用相应的编程语言编写脚本, 通过 Selenium API 调用浏览器驱动程序执行模拟操作, 最终获取数据或进行测试。

Selenium 主要包括三个组件: Selenium WebDriver、Selenium IDE 和 Selenium Grid。

Selenium WebDriver 是 Selenium 最核心的组件, 它提供了一系列的 API 接口, 可以模拟浏览器的行为, 如打开网页、填写表单、点击按钮等, 同时可以获取网页的源码、截图和网页元素等信息。

Selenium IDE 是一个浏览器插件, 可以用于记录和回放用户在浏览器中的行为, 从而生成测试脚本。Selenium IDE 可以自动生成测试代码, 并且可以导出为多种语言的代码。

Selenium Grid 是一个分布式测试工具, 可以在多台机器上同时执行测试, 从而加速测试速度和提高效率。Selenium Grid 可以自动将测试用例分发到不同的节点上执行, 而且还可以同时测试多个不同的浏览器和操作系统组合。

总的来说, Selenium 具有强大的功能和灵活的扩展性, 可以应用于多种场景, 如 Web 应用程序的自动化测试、数据采集、网站爬取等。在网页中有很多交互操作需要用户点击才能触发, 例如链接跳转、按钮点击、输入框输入等。

相比于 Requests 和 Beautiful Soup 这两个库, Selenium 的优点在于可以完全模拟用户操作, 使得爬虫更接近真实用户的行为, 从而可以更精确地抓取网页数据。另外, Selenium 还可以解决动态网页的渲染问题, 而 Requests 和 Beautiful Soup 则只能处理静态网页。但是,

Selenium 也存在一些缺点，例如执行速度较慢，资源消耗较大，需要启动浏览器等。此外，对于一些特殊的网站，Selenium 可能无法正常工作，需要进行一些额外的设置和配置。

Requests 和 BeautifulSoup 的优点在于执行速度较快，资源消耗较少，可以处理大量的静态网页。此外，它们也比较简单易用，无需启动浏览器，适用于一些简单的数据抓取任务。但是，对于动态网页的处理能力相对较弱，无法模拟用户操作，也无法处理一些需要 JavaScript 支持的网页。

2.2 自然语言处理

自然语言处理（Natural Language Processing, NLP）是人工智能领域的一个分支，旨在训练一个人工智能，让其具备理解、生成、处理人类语言的能力。通过自然语言处理技术，计算机可以理解人类语言，从而进行自然语言交互、信息提取、情感分析、语音识别等任务。在情感分析任务中，自然语言处理技术可以通过对文本情感倾向的识别和分析，帮助企业了解消费者的情感态度，从而调整产品和营销策略。在命名实体识别任务中，自然语言处理技术可以帮助识别出文本中的人名、地名、组织机构名等信息，为后续的信息提取和分析提供基础。在关键词提取任务中，自然语言处理技术可以帮助识别出文本中的关键词，从而帮助用户快速了解文本的主题和内容。在自然语言生成任务中，自然语言处理技术可以帮助计算机生成自然语言的文本，如机器翻译、文本摘要、对话系统等。

本文主要使用了 Jieba 中文文本分词工具和词嵌入（word embedding）技术中的 word2vec 模型，对爬取的新闻标题进行分词、清洗以及转化为词向量。

2.2.1 Jieba 中文分词工具

Jieba 是一种基于 Python 的中文分词工具，它能够将一段中文文本切分成一个个独立的词语，从而方便后续对文本的处理和分析。Jieba 使用了基于规则的算法和基于统计的算法相结合的方法进行分词，同时提供了用户自定义词典、停用词、删除标点符号等功能，可以更加灵活地适应不同的文本分析任务。同时，Jieba 还支持多种分词模式和多种输出格式，方便用户根据需求自定义使用。

基于规则的算法是一种基于人工规则的分词算法，它通过预先定义好的规则来进行分词。这种算法的优点是分词准确度高，但是需要大量的人工规则来支持。在 Jieba 分词中，基于规则的算法主要是通过正则表达式来进行匹配，例如可以通过正则表达式匹配中文词语的开头和结尾来进行分词。

基于统计的算法是一种基于概率模型的分词算法，它通过统计大量的语料库来建立概率模型，从而进行分词。这种算法的优点是可以自动学习，不需要人工规则，但是准确度可能会受到语料库的影响。在 Jieba 分词中，基于统计的算法主要是通过隐马尔可夫模型（HMM）和最大匹配算法来进行分词。其中，HMM 模型是一种基于概率的序列模型，可以用来对中文分词进行建模，而最大匹配算法则是一种贪心算法，它会尽可能地匹配最长的词语。

2.2.2 Word2vec：词嵌入技术

词嵌入（Word Embedding）是一种将自然语言中的单词映射到低维向量空间的技术，它可以将单词之间的语义关系转化为向量空间中的几何关系，从而更好地进行自然语言处理任务。常见的词嵌入方法包括 One-hot 编码、基于计数的方法（如 LSA、pLSA）和基于预测的方法（如 word2vec、GloVe）等。

本文使用的是基于预测方法中的 word2vec 模型，它是一种基于神经网络的词嵌入模型，通过训练一个浅层神经网络来学习单词的向量表示。word2vec 模型有两种训练方式：CBOW（Continuous Bag-of-Words）和 Skip-gram。

CBOW 模型的训练目标是从上下文单词的平均向量中预测中心单词。具体地，对于给定的一个中心单词，CBOW 模型会从上下文单词的向量中取平均值作为输入，然后通过一个隐藏层将其

转化为一个低维向量表示，最后通过 softmax 函数输出该中心单词的概率分布。CBOW 模型的优点是训练速度快，但是可能无法捕捉到单词之间的复杂关系。

Skip-gram 模型的训练目标是从中心单词预测上下文单词的概率分布。具体地，对于给定的一个中心单词，Skip-gram 模型会通过一个隐藏层将其转化为一个低维向量表示，然后通过 softmax 函数输出该中心单词的上下文单词的概率分布。Skip-gram 模型的优点是可以捕捉到单词之间的复杂关系，但是训练速度相对较慢。

2.3 基于金融事件的情感词典

情感词典是一种用于情感分析的工具，它包含了一系列单词或短语以及它们所对应的情感极性，通常包括积极、消极和中性三种情感。情感词典的构建通常有三种方法：1. 人工标注法：由人工对文本进行标注，确定每个单词或短语的情感极性。这种方法的优点是准确度高，但是需要大量的人力和时间成本。2. 自动挖掘法：通过自然语言处理技术，如文本挖掘、机器学习等，自动从大规模文本语料库中挖掘出情感词汇。这种方法的优点是效率高，但是准确度可能会受到语料库的影响。3. 人工标注与自动挖掘的结合方法：结合人工标注法和自动挖掘法，既保证了准确度，又提高了效率。

金融情感词典是一种针对金融领域的情感词典，它包含了一系列与金融相关的单词或短语以及它们所对应的情感极性。金融情感词典的构建需要考虑到金融领域的特殊性，如金融产品、金融市场、金融政策等。常见的金融情感词典包括 Loughran-McDonald Financial Sentiment Dictionary、SentiWordNet 等。以 Loughran-McDonald Financial Sentiment Dictionary 为例，它是一种基于人工标注法的金融情感词典，包含了超过 2300 个金融相关的单词或短语以及它们所对应的情感极性。其中，积极情感词包括“increase”、“opportunity”等，消极情感词包括“decline”、“loss”等，中性情感词包括“asset”、“market”等。这种金融情感词典可以用于金融领域的情感分析任务，如预测股票价格、分析投资者情绪等。

但是应用金融情感词典时，需要结合上下文来判断，也就是需要清楚分析的是一个什么事件的情感。比如说，我们无法只从“增加”这一普遍认为是积极的情感极性出发，然后对所有包含“增加”的新闻标题都判定为积极情感。举具体的例子，“增加螺纹空单”与“铁矿石涨”这两个事件中都含有“增加”或与这一单词相似度很高的词“涨”，但背后的经济学含义完全不同，从而对期货价格造成的影响也不同。“增加螺纹空单”意味着供给需求端受到冲击，短期内不看好螺纹期货的上涨；“铁矿石涨”意味着市场价格端得到直接利好，短期内利于螺纹期货的上涨。

因此，本文从宏观、微观经济学出发，基于新闻中常见的事件类型，构建一个专属于钢材期货的金融情感词典，包括了 5 类：供给需求、市场价格、资金技术、经济政策、宏观环境。以螺纹钢这一品种专属的期货情感词典融合了经济学的底层逻辑，意味着形成的情绪信号具有较强的理论依据作为支撑。同时，每个大类包含正向情感和负向情感两个子类，因此一共 10 类事件。针对影响的不同，包括直接影响和间接影响，可以将这 10 类事件赋予不同的分值。对于供给需求和市场价格这两大类，事件的发生我们认为会对期货价格造成直接的影响，因此对其中的正向情感事件赋予 2 分、对其中的负向情感事件赋予-2 分。对于资金技术、经济政策和宏观环境这三大类，事件的发生我们认为会对期货价格造成间接的影响，意味着它们的作用时间会相对更长，并且随着时间的推移，间接作用的影响可能会逐渐被其他事件所淡化。因此对其中的正向情感事件赋予 1 分、对其中的负向情感事件赋予-1 分。

2.4 Logistic 回归

Logistic 回归是一种广泛应用于分类问题的统计学习方法，它可以用于二分类和多分类问题，并且可以处理离散型和连续型特征。在 Logistic 回归中，我们通过对样本数据进行拟合，得到一个分类模型，然后用该模型对新的样本进行分类。

Logistic 回归模型的基本形式为：

$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}. \quad (1)$$

其中, $h_{\theta}(x)$ 是预测值, x 是输入特征向量, θ 是模型参数。

Logistic 回归模型的训练目标是最大化似然函数, 即:

$$\max_{\theta} \prod_{i=1}^m h_{\theta}(x^{(i)})^{y^{(i)}} (1 - h_{\theta}(x^{(i)}))^{1-y^{(i)}}. \quad (2)$$

其中, m 是训练样本数量, $x^{(i)}$ 是第 i 个样本的特征向量, $y^{(i)}$ 是第 i 个样本的标签。

为了避免过拟合, 我们通常会加入正则化项, 得到正则化的 Logistic 回归模型:

$$\max_{\theta} \prod_{i=1}^m h_{\theta}(x^{(i)})^{y^{(i)}} (1 - h_{\theta}(x^{(i)}))^{1-y^{(i)}} - \frac{\lambda}{2} \sum_{j=1}^n \theta_j^2. \quad (3)$$

其中, λ 是正则化参数, n 是特征数量。

Logistic 回归模型的训练通常采用梯度下降法或牛顿法等优化算法进行求解。在预测时, 我们将输入特征向量带入模型中, 得到预测值, 然后根据阈值将预测值转化为类别标签。对于本文中的预测任务, 可认为是一个预测未来一天价格涨跌的二分类模型。因此, 本文建立了 Logistic 模型来拟合情绪分值与未来一天的价格涨跌之间的关系, 训练好后进行预测时, 会根据输入的情绪分值输出 0 或 1, 0 代表预测明天价格会跌, 1 代表预测明天价格会涨。具体建模时, 本文计算了以一段时间 T 作为滑动时间窗口, 在当前时间 t 往过去看时间步长 T 天, 这段时间内的情绪总分值 S_t^T , 并以情绪总分值 S_t^T 作为自变量 X , 以当前时间 t 的未来一天价格 P_{t+1} 相较于当前时间的价格 P_t 的涨跌情况 ΔP_t 作为因变量 y , 若涨则令 $\Delta P_t = 1$, 反之则为0。对于训练好的 Logistic 模型, 输入时间步长为 T 的情绪总分值 S_t^T , 将输出的预测信号 ΔP_t^{Pred} 作为事件情绪信号。

2.5 Backtrader 量化平台

Backtrader 是一款基于 Python 的开源量化交易平台, 它提供了丰富的功能和工具, 可以帮助用户进行策略回测、优化和实盘交易等操作。Backtrader 包含了以下主要功能:

1. 数据处理。Backtrader 支持多种数据源, 包括 CSV 文件、Pandas DataFrame、Yahoo Finance 等, 可以方便地加载和处理历史数据。Backtrader 还提供了多种数据预处理工具, 如数据对齐、数据合并、数据重采样等, 可以帮助用户更好地处理历史数据。

2. 策略回测。Backtrader 提供了灵活的策略回测功能, 用户可以自定义交易策略并进行回测。Backtrader 还支持多种交易手续费计算方式、多种资产类别、多种时间周期等, 可以满足不同用户的需求。

3. 策略优化。Backtrader 提供了多种优化工具, 如参数扫描、遗传算法等, 可以帮助用户寻找最优的交易策略和参数。它还提供了多种评估指标, 如夏普比率、最大回撤等, 可以帮助用户评估交易策略的效果。

4. 实盘交易。Backtrader 支持多种实盘交易接口, 如 IB、Oanda、Alpaca 等, 可以方便地进行实盘交易。Backtrader 还提供了多种交易订单类型、多种交易执行方式等, 可以满足不同用户的需求。

5. 可视化。Backtrader 提供了丰富的可视化工具, 如 K 线图、交易信号图等, 可以帮助用户更好地理解交易策略和回测结果。它还支持多种输出格式, 如 PDF、CSV 等, 方便用户进行结果分析和报告生成。

本文利用 Backtrader 构建了两种趋势策略模型, 分别是单均线 and 双均线策略。同时, 使用 GenericCSV_extend 方法, 加上事件情绪信号 ΔP_t^{Pred} 后得到基于趋势信号的事件驱动型策略。

3 模型构建

3.1 数据选取

1. 期货新闻标题数据

以上海期货交易所的 2023 年 5 月交割的螺纹钢期货（RB2305）为标的资产，使用 Selenium 爬取东方财富网期货频道钢材资讯内从 2022 年 3 月 23 日到 2023 年 2 月 17 日的所有新闻标题，共 1000 条。

上海期货交易所的 RB2305 期货信息如表 3-1。

表 3-1 RB2305 期货介绍

交易代码	上市交易所	合约交割月份	交易单位	涨跌停幅度	最后交易日
RB	上海期货交易所	1~12 月	10 吨/手	上一交易日结算价 ±3%	合约月份的 15 日

爬取后的钢材新闻标题的首尾数据如表 3-2。

表 3-2 爬取的钢材期货新闻节选

Date	Headline
2023 年 2 月 17 日	螺纹钢需求环比大增超六成 多地市场价上涨
2023 年 2 月 17 日	黑色持仓龙虎榜：铁矿石涨 2.01% 中信期货增持超 7 千手螺纹多单
2023 年 2 月 17 日	【期市收评】商品综合指数收涨 铁矿石主力合约连续四日资金净流入排行前三
2023 年 2 月 17 日	黑色：牛市第二阶段开启
2022 年 3 月 23 日	大面积飘红！成本上涨产量下滑 钢企去年业绩仍让市场“松口气”
2022 年 3 月 23 日	美国将放宽对英国钢铝产品进口关税

2. 期货收盘价数据

以 RB2305 为标的资产，从东方财富网中爬取从 2022 年 5 月 17 日（起始交易日）到 2023 年 2 月 17 日的历史交易数据，只使用收盘价一系列数据。

3.2 数据预处理

3.2.1 数据清洗

对爬取的 1000 条钢材期货新闻标题数据，为了降低无关因素对后续分析的干扰，因此有必要对数据进行清洗。

清洗的过程包括以下几个步骤：

1. 删除所有包含“|”这一符号的行。因为包含“|”的新闻标题没有实际意义，比如“一图看懂 | 2 月 16 日黑色系期货机构观点汇总”。
2. 删除所有新闻标题中含有“：”的前边的内容，即对于“商务部：中澳贸易一直正常开展去年货物贸易额超 2200 亿美元”，删除标题中的“商务部：”。
3. 按时间顺序升序排序。回测时要求数据按时间顺序升序排序。

3.2.2 Jieba 分词

对清洗完后的新闻标题，使用 Jieba 中文分词工具进行分词。具体来说，则是使用 Jieba.lcut(str)这一函数，遍历 dataframe 中每一行新闻标题，再将分完词后的列表中心存储在新的列表中。最后生成新的一列贴到原有 dataframe 的末尾，如表 3-3 所示。

表 3-3 爬取的钢材期货新闻节选并分词

Date	Headline	Cut
2023 年 2 月 17 日	螺纹钢需求环比大增超六成 多地市场价上涨	['螺纹钢', '需求', '环比', '大增', '超', '六成', ' ', ' ', '多地', '市场价', '上涨']
2023 年 2 月 17 日	铁矿石涨 2.01% 中信期货增持超 7 千手螺纹多单	['铁矿石', '涨', '2.01%', ' ', ' ', '中信', '期货', '增持', '超', '7', '千手', '螺纹', '多单']
2023 年 2 月 17 日	牛市第二阶段开启	['牛市', '第二阶段', '开启']

3.3 基于金融事件的情感词典构建

本文从宏观、微观经济学出发，基于新闻中常见的事件类型，构建一个专属于钢材期货的金融情感词典，包括了 5 类：供给需求、市场价格、资金技术、经济政策、宏观环境。以螺纹钢这一品种专属的期货情感词典融合了经济学的底层逻辑，意味着形成的情绪信号具有较强的理论依据作为支撑。同时，每个大类包含正向情感和负向情感两个子类，因此一共 10 类事件。针对影响的不同，包括直接影响和间接影响，可以将这 10 类事件赋予不同的分值。对于供给需求和市场价格这两大类，事件的发生我们认为会对期货价格造成直接的影响，因此对其中的正向情感事件赋予 2 分、对其中的负向情感事件赋予-2 分。对于资金技术、经济政策和宏观环境这三大类，事件的发生我们认为会对期货价格造成间接的影响，意味着它们的作用时间会相对更长，并且随着时间的推移，间接作用的影响可能会逐渐被其他事件所淡化。因此对其中的正向情感事件赋予 1 分、对其中的负向情感事件赋予-1 分。

通过快速浏览所有钢材新闻标题，能够初步概括出一些高频事件，并将它们分为以上 10 类，表 3-4 中展示了一些子类及其部分代表事件。

表 3-4 金融事件的情感词典构建

Class	Event
供给需求_pos	需求大增 增加螺纹多单 减少螺纹空单 产量增加
市场价格_neg	铁矿石跌 成交一般 市场价下跌 钢价回落
资金技术_pos	资金流入 预增 节约钢材 利润增加
宏观环境_neg	熊市
经济政策_pos	项目投产 低碳 碳达峰 暂停征税

基于 word2vec 模型中的计算两段文本的相似度方法，我们能够实现对爬取的 1000 条新闻数据中每一条新闻进行分类，并赋予情绪分值。Word2vec 模型使用的是用 1.5 亿搜狐新闻训练出来的模型，包含了政治、金融、体育、娱乐等新闻。具体来说，首先先提取出一条分词后的新闻标题，然后依次计算它在这 10 个子类中的跟所有代表事件的相似度，并选取最高的一个相似度作为子类相似度，最后从 10 个候选相似度中选取最大的作为最终相似度，从而确定其标签和分数。在 3.1 中表加上标签和分数后，如表 3-5 所示。

表 3-5 使用金融事件情感词典对标题分类与打分

Date	Headline	Label	Score
2022 年 3 月 23 日	美国将放宽对英国钢铝产品进口关税	policy_pos	1
2022 年 3 月 23 日	大面积飘红！成本上涨产量下滑 钢企去年业绩仍让市场“松口气”	capital_tech_neg	-1
2023 年 2 月 17 日	【期市收评】商品综合指数收涨 铁矿石主力合约连续四日资金净流入排行前三	price_neg	-2
2023 年 2 月 17 日	铁矿石涨 2.01% 中信期货增持超 7 千手螺纹多单	price_neg	-2
2023 年 2 月 17 日	螺纹钢需求环比大增超六成 多地市场价上涨	supply_demand_pos	2

最后，将同一天的分数相加，只保留日期和分数，作为新闻标题的基于事件的情绪分值。这里的分数分为正向和负向。正向分数越高，代表着资产价格未来上涨的概率越大；负向分数越高，代表着资产价格未来下跌的概率越大。

3.4 Logistic 回归生成事件情绪信号

本文使用的是 `sklearn.linear_model` 里的 Logistic Regression 类。定义了以下函数：

1. 函数 1: 导入原始数据。导入 RB2305 的收盘价数据和每一天的事件情绪分值数据。
2. 函数 2: 计算真实价格信号。通过 RB2305 的收盘价数据，通过当前时间和未来一天时间的价格变化，计算真实的价格信号。若未来一天价格上涨，则信号为 1；反之则为 0。
3. 函数 3: 计算过去 N 天的事件情绪总分。定义一个时间窗口 N，则对当前时间的事件情绪总分定义为过去 N 天的事件情绪分值的和。
4. 函数 4: 合并真实价格信号列和滑动时间窗口的事件情绪总分列。使用 `pandas` 中的 `merge` 函数，将真实价格信号列 `real_price_signal` 和滑动时间窗口的事件情绪总分列 `past_n_day_score_list` 合并起来，得到一个新的 `dataframe`。
5. 函数 5: 训练 Logistic 回归。将 `past_n_day_score_list` 作为自变量 X，为了提高训练准确率，先将 X 进行归一化，即使用公式：

$$X = \frac{X + X_{max} - X_{mean}}{X_{max} - X_{min}}. \quad (4)$$

归一化后，使用 `reshape(-1, 1)` 将其转换为列向量。将 `real_price_signal` 作为因变量 y，然后使用 `model.fit()` 函数，输入 X 和 y，即可得到训练好的模型。以滑动时间窗口 N=15 为例，当前时间的事件情绪总值用过去 15 天的事件情绪值的总分代替。价格训练数据如图 3-1，图中蓝色散点代表真实交易信号 y，橙色的时序图代表 N=15 时的事件情绪总分序列 X。



图 3-1 Logistic 回归训练数据

6. 函数 6:预测价格信号。将拟合出的 Logistic 回归模型对原数据进行预测，得到事件情绪信号预测序列，并筛选出若干天的事件情绪总分和预测的事件情绪信号，如表 3-6 所示。

表 3-6 使用 Logistic 回归预测

n_day_score	pred_signal
-23	0
-14	0
-2	0
3	1
11	1
12	1

从表中可以看出，训练出来的 Logistic 只是做了一件事：只要当事件情绪总分小于 0，则预测未来价格会跌；只要当事件情绪总分大于 0，则预测未来价格会涨。这符合事件情绪信号的设计与逻辑。

3.5 Backtrader 量化平台回测

在使用 Backtrader 量化平台构建回测策略时，首先先使用定义一个类，类名为 MABacktrader，并继承 bt.Strategy 这一个类。类 bt.Strategy 用于定义交易策略。在这个类中，我们可以定义交易信号，止损和止盈策略，以及其他与交易相关的逻辑。以下是构建趋势回测策略时需要用到的函数及其参数的设置：

1. 初始函数。初始函数中最重要的是定义当前时刻价格 `self.dataclose=self.datas[0].close`，意味着永远是当前时刻的收盘价。其次设计几个变量 `self.order`, `self.buy_price` 都为 None。然后是趋势策略中用到的移动平均指标，以单均线策略为例，定义 `self.sma10`，使用 Backtrader 内 indicators 自带的方法计算。

2. Next 函数：下一次的执行动作。在 next 函数中设计买入和卖出条件，对于单均线策略来说：当前一天收盘价低于 sma10，并且当前收盘价高于 sma10 时，买入价为当前收盘价，买入，并修改当前 order 状态为 buy；当前一天收盘价高于 sma10，并且当前收盘价低于 sma10 时，卖出价为当前收盘价，卖出，并修改当前 order 状态为 sell。

3. Notify_order 函数：订单状态处理。订单一共有 5 种状态，分别为：已提交、已接收、已完成、已取消、已拒绝。当订单处于已提交、已接收状态时，不需要做任何处理；当订单处于已完成状态时，判断是买单还是卖单，再进入到相应的状态中记录买单或卖单的价格、总花费以及佣金税费（在本文中默认为 0）。

4. Notify_trade 函数：交易状态处理。主要作用是记录订单成交后目前持仓金额，订单导致的变动金额。

5. Stop 函数：交易终止函数。主要作用是当来到最后一个交易日时，停止交易，并显示最终金额。

回测策略主要包含以上五个模块，创建完回测策略后，在主运行模块中，先使用 `bt.Cerebro()` 创建一个交易员实例，然后设置这个交易员有一百万初始资金，并规定它每次买入以 100 手为单位。接着，使用 `bt.feeds.PandasData` 导入标的资产交易数据。同时，把定义好的 Strategy 传递给交易员。最后，使用 Cerebro 实例中的 `run` 函数即可开启回测。

若要融入事件情绪信号，只需要使用 `bt.feeds.GenericCSV` 导入数据，它能够扩展原来规定好的数据输入格式，添加自定义交易信号。

4 实验结果与分析

4.1 实验设计

本文主要对基于趋势信号的事件驱动型期货交易策略进行研究。实验内容主要为以下部分：

1. 对计算事件情绪总分的滑动窗口值进行参数调优，寻找通过 Logistic 回归建模后准确率最高的滑动窗口值 N ，并使用该值作为唯一滑动窗口进行后续实验。

2. 对 RB2305 构建回测交易策略，先构建 2 种基准策略：单均线策略和双均线策略，考虑到一个合适的交易频率，单均线策略的单均线设为 10 天滑动平均值，双均线策略的两条均线分别是 5 天滑动平均值和 10 天滑动平均值。接着，在 2 种基准策略的基础上融入事件情绪信号作为实验组。按照单均线和双均线，可将实验分成两组。最后，记录这两组每次交易的持仓金额变化，并通过评价指标表格对比事件驱动型交易策略的价值。

4.2 评价指标

1. 累积收益率（Cumulative Return）

累积收益率的计算方式是通过将投资组合的最终价值减去其初始价值，然后除以初始价值来计算。它反映了一个投资组合在交易阶段结束时的总回报。

$$CR = \frac{P_{\text{end}} - P_0}{P_0}. \quad (5)$$

2. 最大回撤率（Maximum Pullback）

最大回撤率是交易期间的最大损失百分比，它衡量了一个模型的稳健性。

$$MPB = \frac{\max(A_x - A_y)}{A_y}. \quad (6)$$

其中 A_x, A_y 是策略的总资产， $x > y, A_y > A_x$ 。

3. 夏普比率（Sharpe Ratio）

夏普比率通过从年化收益中减去年化无风险利率，再除以年化波动率来计算。它综合考虑了收益和风险，反映了单位系统风险的超额收益。

$$SR = \frac{E(R_P) - R_f}{\sigma_P}. \quad (7)$$

4. 胜率（Win Rate）

胜率是指交易中盈利交易的比例。它通常用百分比表示，并且是评估交易策略性能的重要指标之一。

胜率的计算公式如下：

$$\text{Win Rate} = \frac{\text{Number of Winning Trades}}{\text{Total Number of Trades}} \times 100\%. \quad (8)$$

其中，Number of Winning Trades 表示盈利交易的数量，Total Number of Trades 表示总交易数量。

5. 过滤胜率（Filter Win Rate）

过滤胜率是本文针对事件情绪信号提出的性能衡量指标，它的分子是与基准模型相比成功过滤失败交易的次数，分母是过滤总次数，即总交易次数。

过滤胜率的计算公式如下：

$$\text{Filter Win Rate} = \frac{\text{Number of Winning Trades to Filter}}{\text{Total Number of Trades}} \times 100\%. \quad (9)$$

其中，Number of Winning Trades to Filter 表示跟基准模型相比交易中过滤失败交易成功的数量，Total Number of Trades 表示总交易数量。

4.3 RB2305 描述性统计分析

在进行实验前，先对 RB2305 的价格序列进行描述性统计分析。

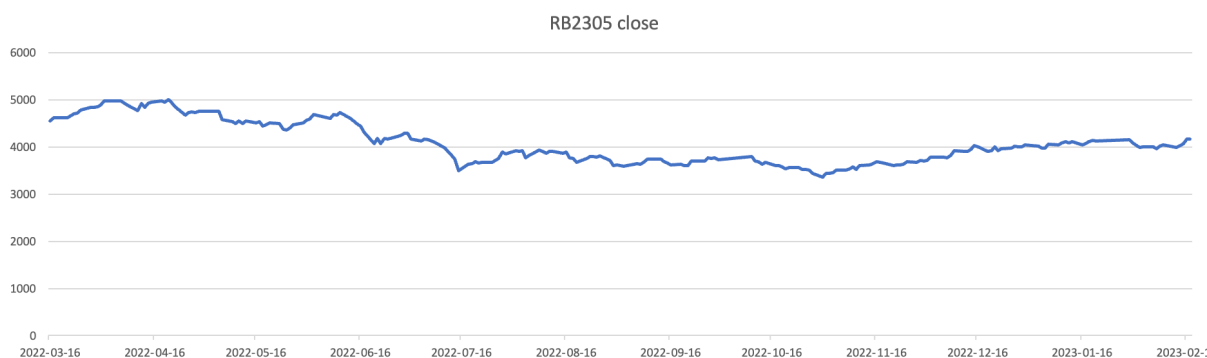


图 4-1 RB2305 收盘价时序图

对于 RB2305 螺纹钢期货的收盘价的时间序列，总体上还是比较平稳的，没有特别大的起伏。从 2022 年 3 月 16 日开始到 2022 年 7 月 16 日，这段时间 RB2305 的整体趋势是波动向下的，之后 3 个月则横盘震荡，并在 2022 年 10 月 16 日后开始缓慢攀升。

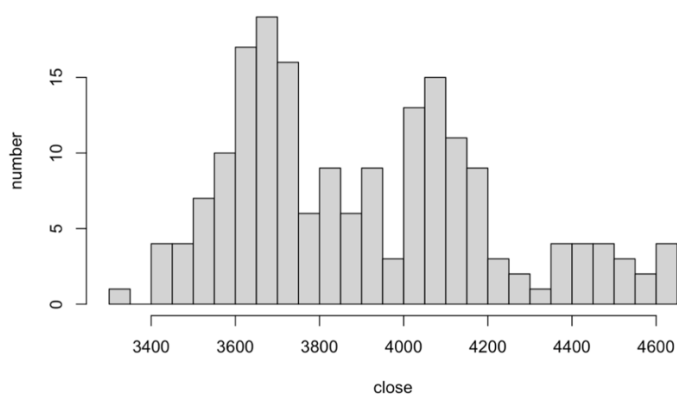


图 4-2 RB2305 收盘价直方图

如图所示，从收盘价的直方图看，收盘价主要集中在 3700 左右、4100 左右以及两者之间的值，更直观地体现出了该时间序列的较高的集中程度。

表 4-1 RB2305 收盘价分位表

min.	1st Qu.	median	mean	3rd Qu.	max.
3336	3666	3854	3906	4102	4628

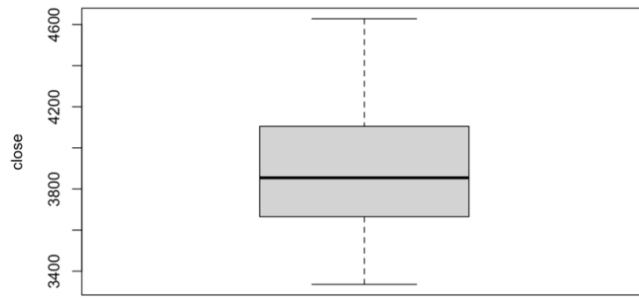


图 4-3 RB2305 收盘价箱线图

从 RB2305 收盘价的分位表及箱线图来看，中位数和平均数差距不大，平均值只比中位数高了约 1.3%，第一分位数和第三分位数距离中位数也不大，进一步地说明了 RB2305 螺纹钢期货的收盘价总体上是较为集中的。

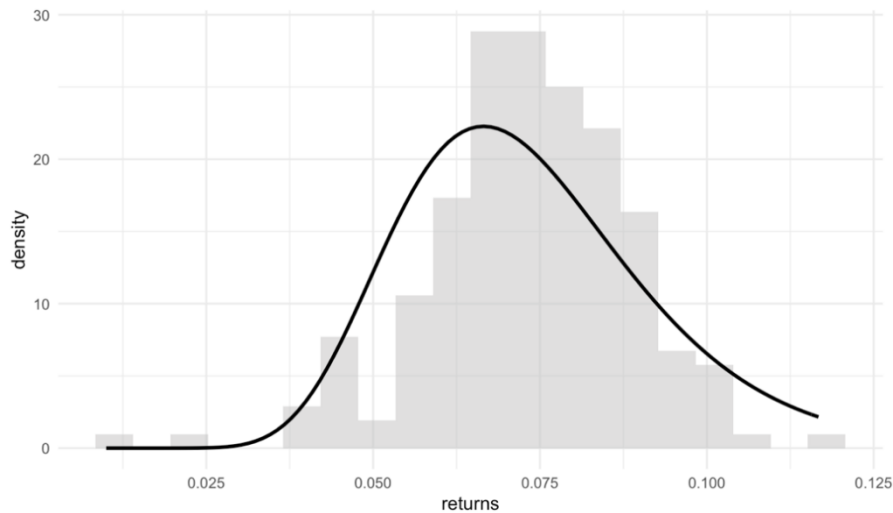


图 4-4 RB2305 的对数收益率直方图及拟合图

进一步地，画出 RB2305 的对数收益率的直方图，并用对数正态分布进行拟合，如图所示。需要说明的是，为了拟合对数正态分布，需要对对数收益率进行平移：

$$return_i = return_i - \min(return_i) + 0.01. \quad (10)$$

因此，实际上在图中 $returns=0.07$ 的左边都是对数收益率小于 0 的部分。由图可知，实际上 RB2305 的对数收益率服从的分布比对数正态分布更具尖峰厚尾的特点，并且负收益率的偏离值显著多于正收益率，分布呈现左偏的特点。因此，想要抓住盈利的机会还是具有较大挑战的。

4.4 滑动窗口值调优

首先，本文对计算事件情绪总分的滑动窗口值进行参数调优，寻找通过 Logistic 回归建模后准确率最高的滑动窗口值 N ，并使用该值作为唯一滑动窗口进行后续实验。本文对滑动时间窗口选取了 $Time\ Window = [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30]$ ，并依次进行 Logistic 回归建模，得到准确率折线图，如图 4-1 所示。

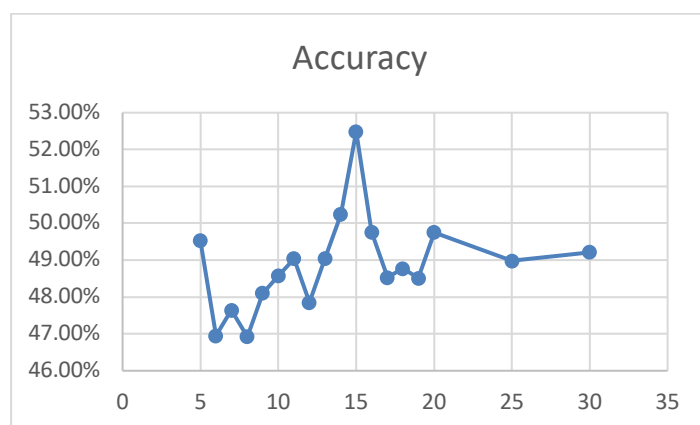


图 4-5 Logistic 回归滑动窗口调优

由图 4-1 可知，在滑动时间窗口 Time Window=15 时拟合准确率达到最高，达到 52.49%，其余窗口值则在 47%到 50%之间浮动。因为考虑到使用的是趋势策略，因此本文选取的时间窗口值测试起点是 5，能够代表过去 5 天整个事件情绪的情况，一直测试到 30 天。从预测结果来看，当事件情绪总分序列的分布方差较大时，Logistic 可以很好地做到将事件情绪总分小于 0 的则判定事件情绪信号为 0，反之则为 1。图 4-2 展示了当时窗口为 15 天时情绪总分序列的直方图。

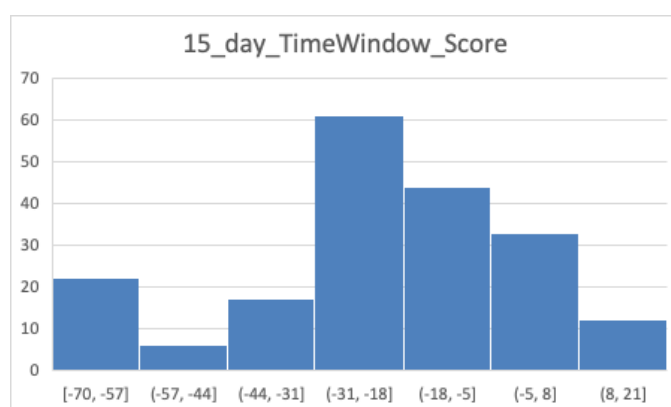


图 4-6 滑动窗口为 15 天时的情绪总分分布

可以看出，分值的分布整体是比较分散的，但总分<0 的值占了大多数，从这一点上看本文开发出的事件情绪信号能够更严格地对买入信号进行过滤。

4.5 交易策略比较

首先，我们先比较了单均线策略环境下的两种策略：MA10 的单均线基准策略与加入了事件情绪信号的 MA10 单均线策略。通过 Backtrader 量化平台进行回测，记录策略运行过程中的持仓金额，如图 4-3 所示。

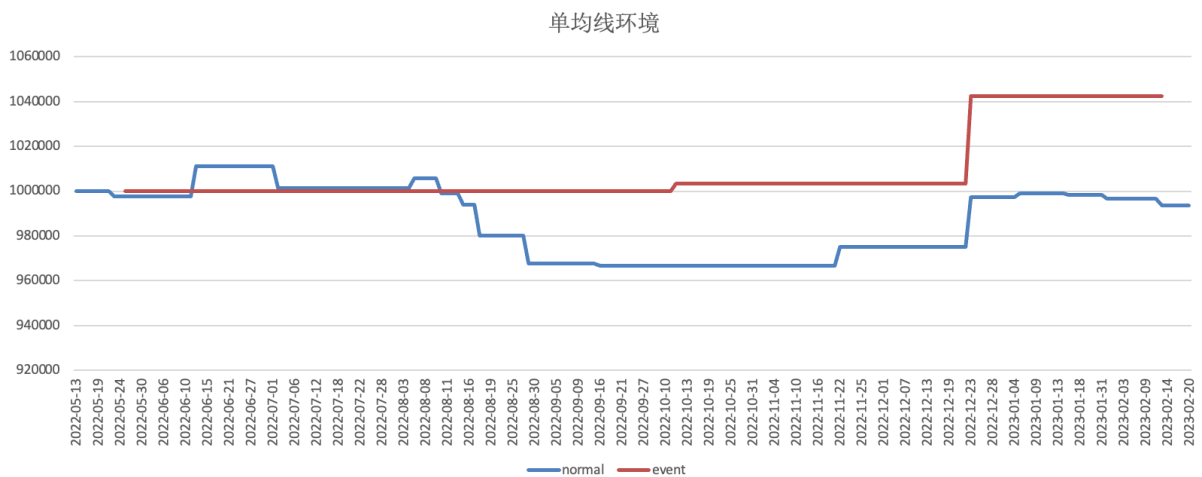


图 4-7 单均线环境的两种策略持仓金额图

图中每一处跳动的地方代表一次交易，折线上升代表交易盈利，下降代表交易亏损，融入事件情绪信号后，策略的交易频率大大减少，只进行了两次交易，并且胜率达到了 100%，说明事件情绪信号有效地过滤了趋势策略中的无效信号。同时，通过计算出两种策略的评价指标如表 4-1 所示，事件驱动的交易策略在累计收益率、最大回撤率、胜率和夏普比率上都战胜了普通的趋势策略。

表 4-2 单均线环境的两种策略评价指标对比

(22/5/13-23/2/13)	累计收益率	最大回撤率	胜率	夏普比率	过滤胜率
normal	-0.65%	-1.38%	37.5%	-0.039	NAN
event	4.26%	0	100%	0.842	75%

接着，比较双均线策略环境下的两种策略：MA5 与 MA10 的双均线基准策略与加入了事件情绪信号的 MA5 与 MA10 双均线策略。通过 Backtrader 量化平台进行回测，记录策略运行过程中的持仓金额，如图 4-4 所示。

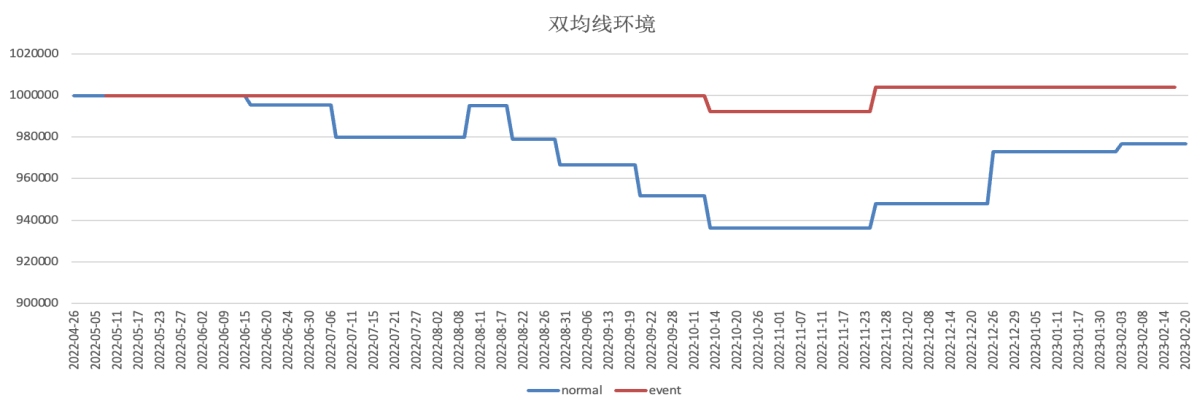


图 4-8 双均线环境的两种策略持仓金额图

双均线环境中的交易频率更少，因为需要满足的要求相较单均线更高。同时，单从基准策略对比看，双均线的表现更差。引入事件情绪信号后，可以发现仍然过滤了大部分错误信号，达到

了非常好的止损效果，体现出了较好的提升作用，如表 4-2 所示。但就最终结果而言，双均线下的事件驱动表现没有单均线的事件驱动策略好。

表 4-3 双均线环境的两种策略评价指标对比

22/4/26- 23/2/20	累计收益率	最大回撤率	胜率	夏普比率	过滤胜率
normal	-2.34%	-0.016	40%	-0.143	NAN
event	0.40%	-0.008	50%	0.148	60%

从以上对两组交易策略的对比实验结果中可以得出结论：在应用了事件情绪信号来过滤趋势信号后，两组实验中的基于趋势信号的事件驱动型策略在累积收益率、最大回撤率、胜率和夏普比率上都比趋势基准策略表现得更好和更稳定。这说明该事件情绪信号能够有效过滤错误的趋势信号，使得投资策略更加稳健，也证实了事件驱动型策略的有效性。

为了进一步探究本文开发的事件情绪信号有效性，对趋势信号的时间窗口进行调整从而进行更多的实验。对于单均线策略，调整时间窗口值的范围从 2 到 10，以及 15 和 20，并统计每个单均线策略下基于事件驱动的交易策略的累计收益率与普通趋势策略的累计收益率的差额，如表所示。

表 4-4 单均线策略下的扩展实验

time window	event-normal
ma2	2.85%
ma3	4.61%
ma4	3.65%
ma5	2.74%
ma6	6.21%
ma7	6.05%
ma8	5.52%
ma9	5.45%
ma10	4.91%
ma15	2.05%
ma20	3.76%

在实验过程中，所有的普通趋势策略的累计收益率都是负收益率，而所有加上本文的事件情绪信号后的交易策略都取得了比原来策略普遍高好几倍的收益。因为考虑到 RB2305 的波动范围小、价格集中程度较高的特点，在这样窄幅波动的空间内获取收益是较为困难的，但是本文的事件驱动型策略却达到了 100%胜率，以及累计收益率能够是基准策略的若干倍。这充分说明了本文的事件情绪信号的有效性。

5 总结与展望

5.1 总结

本文主要对基于趋势信号的事件驱动型期货交易策略进行研究，对新闻数据进行爬取、事件情绪信号进行了设计与构建，并构建了回测策略。实验结果表明，在应用了事件情绪信号后，能够过滤大部分错误趋势信号，使得交易策略表现更好更稳健。下面简单总结本文的主要工作。

首先，对本文的研究背景和国内外的相关研究进行了充足的调研。然后，介绍了本文研究的相关基础理论知识，包括 Selenium 爬虫、词嵌入技术中的 word2vec 模型、Logistic 回归与 Backtrader 量化平台。接着，对新闻数据进行爬取，对事件情绪信号进行了设计与构建，并训练了 Logistic 回归。最后，使用 Backtrader 构建两组对比试验的回测策略，取得不错的实验结果，说明该事件情绪信号能够有效过滤错误的趋势信号，使得投资策略更加稳健，也证实了事件驱动型策略的有效性。

本文构建的基于趋势信号的事件驱动型期货交易策略最重要的价值在于，它提供了一种可移植的有效的策略构建方向，因为受到交易标的的波动小、交易时间较短的限制，所以实验的结果受到了一定的约束。若更换为价格波动大、交易量多的期货标的，则可以期待本文构建的事件情绪信号能够发挥更大的价值。

5.2 不足与展望

本文的研究存在一些不足与未来可改进的方向。

第一，实验数据集比较有限。本文仅能获得近一年的新闻标题数据，并且数据的来源渠道少，而且对能够建立有效信号的新闻的质量要求较高。

第二，实验样本有限。因为可爬取的、高质量的期货新闻标题来源渠道少，本文只采用了 RB2305 这一种期货进行实验，若有更为专业、丰富的数据渠道，可考虑扩充实验样本，充分验证该事件情绪信号的有效性。

第三，使用基于金融事件的情感词典分类问题。本文构建的基于金融事件的情感词典来源于近一年的新闻标题，代表事件只挖掘了最高频的部分。分类是基于 word2vec 模型，可以考虑使用金融数据训练的更加专业的模型进行分类。

参考文献

- [1] 黄进, 阮彤, 蒋锐权. 基于 SVM 结合依存句法的金融领域舆情分析[J]. 计算机工程与应用, 2015, 0(23): 230-235
- [2] 黄亮点, 高梓耕, 张行健, 蔡凯莉, 罗小可, 李岩. 新闻事件驱动的海龟交易策略优化研究: 文本情绪方法[J]. 全国流通经济, 2021(19): 153-156
- [3] 李备友, 张桂艳, 路英, 李守伟. 基于事件驱动的证券市场波动生成机制分析[J]. 华东经济管理, 2012(3): 93-98
- [4] 林培光, 周佳倩, 温玉莲. SCONV: 一种基于情感分析的金融市场趋势预测方法[J]. 计算机研究与发展, 2020, 57(8): 1769-1778
- [5] 刘薇, 姜青山, 蒋泓毅, 胡金帅, 曲强. 基于 FinBERT-CNN 的股吧评论情感分析方法[J]. 集成技术, 2022, 11(1): 27-39
- [6] 刘玉玲, 赵国龙, 邹自然, 吴升婷. 基于情感分析和 GAN 的股票价格预测方法[J]. 湖南大学学报: 自然科学版, 2022, 49(10): 111-118
- [7] 许雪晨, 田侃. 一种基于金融文本情感分析的股票指数预测新方法[J]. 数量经济技术经济研究, 2021, 38(12): 124-145
- [8] 祝清麟, 梁斌, 徐睿峰, 刘宇瀚, 陈奕, 毛瑞彬. 结合金融领域情感词典和注意力机制的细粒度情感分析[J]. 中文信息学报, 2022, 36(8): 109-117
- [9] Althelaya K A, Mohammed S A, El-Alfy E-S M. Combining Deep Learning and Multiresolution Analysis for Stock Market Forecasting[J]. IEEE Access, 2021, 9:13099-13111.
- [10] Biswas M, Shome A, Islam M A, et al. Predicting Stock Market Price: A Logical Strategy using Deep Learning[C]//Proceedings of the 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE). Penang, Malaysia: IEEE, 2021:218-223.
- [11] Chen Q, Liu X-Y. Quantifying ESG alpha using scholar big data: An automated machine learning approach[C]//ACM International Conference on AI in Finance, ICAIF 2020, 2020.
- [12] Darapaneni N, Paduri A R, Sharma H, Manjrekar M, Hindlekar N, Bhagat P, Aiyer U, Agarwal Y. Stock Price Prediction using Sentiment Analysis and Deep Learning for Indian Markets[J]. ArXiv, 2022, abs/2204.05783.
- [13] Fang Y, Liu X-Y, Yang H. Practical machine learning approach to capture the scholar data driven alpha in AI industry[C]//2019 IEEE International Conference on Big Data (Big Data) Special Session on Intelligent Data Mining. 2019:2230-2239.
- [14] Jing N, Wu Z, Wang H. A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction[J]. Expert Systems with Applications, 2021, 178:115019.
- [15] Padmanayana V, Bhavya K. Stock Market Prediction Using Twitter Sentiment Analysis[J]. International Journal of Scientific Research in Science and Technology, 2021.
- [16] Gupta R, Chen M. Sentiment Analysis for Stock Price Prediction[C]//2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). Shenzhen, China: IEEE, 2020:213-218. doi: 10.1109/MIPR49039.2020.00051.
- [17] Rezaei H, Faaljou H, Mansourfar G. Stock price prediction using deep learning and frequency decomposition[J]. Expert Systems with Applications, 2021, 169:114332.
- [18] SINHA S, MISHRA S, MISHRA V, et al. Sector influence aware stock trend prediction using 3D convolutional neural network[J]. Journal of King Saud University-Computer and Information Sciences, 2022, 34(4):1511-1522.
- [19] Syu J-H, Lin J C-W, Wu C-J, Ho J-M. Stock Selection System Through Suitability Index and Fuzzy-Based Quantitative Characteristics[J]. IEEE Transactions on Fuzzy Systems, 2023, 31(1):322-334.
- [20] Wu J M-T, Li Z, Herencsar N, Vo B, Lin J C-W. A graph-based CNN-LSTM stock price prediction algorithm with leading indicators[J]. Multimedia Systems, 2021.

- [21]Wu J M-T, Syu J C-W, Lin J C-W, Ho J-M. A graph-based convolutional neural network stock price prediction with leading indicators[J]. Software: Practice and Experience, 2021, 51:628-644.
- [22]Wu M-E, Syu J-H, Lin J C-W, Ho J-M. Effective Fuzzy System for Qualifying the Characteristics of Stocks by Random Trading[J]. IEEE Transactions on Fuzzy Systems, 2022, 30(8):3152-3165.
- [23]Xiao Ding, Kuo Liao, Ting Liu, Zhongyang Li, Junwen Duan. Event Representation Learning Enhanced with External Commonsense Knowledge[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019:4894-4903.
- [24]Xiao Ding, Yue Zhang, Ting Liu, Junwen Duan. Deep learning for event-driven stock prediction[C]//Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15). AAAI Press, 2015:2327-2333.
- [25]Xu W, Chen L, Xie W, Li X, Li Z, Liu Y. HIST: A graph-based framework for stock trend forecasting via mining concept-oriented shared information[J]. arXiv preprint arXiv:2110.13716, 2021.
- [26]Xu W, Liu W-Q, Xu C, Bian J, Yin J, Liu T-Y. REST: Relational Event-driven Stock Trend Forecasting[C]//Proceedings of the 14th ACM Conference on Recommender Systems. 2020:1-10. doi: 10.1145/3442381.3450032.
- [27]Yang H, Liu X-Y, Zhong S, Walid A. Deep reinforcement learning for automated stock trading: An ensemble strategy[C]//Proceedings of the First ACM International Conference on AI in Finance. 2020:1-8.
- [28]Yang Xiao, et al. Qlib: An ai-oriented quantitative investment platform[J]. arXiv preprint arXiv:2009.11189, 2020.
- [29]Zhang W, Skiena S. Trading strategies to exploit blog and news sentiment[C]//Fourth International AAAI Conference on Weblogs and Social Media. 2010.
- [30]Zhihan Zhou, Liqian Ma, Han Liu. Trade the Event: Corporate Events Detection for News-Based Event-Driven Trading[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021:2114-2124.

致谢

大学的本科四年时光转瞬即逝，一路走过来需要感谢许多人。首先我想感谢我的家人，谢谢他们一直以来对我的鼓励与支持，谢谢我的父母让我无需担心自己的衣食住行，谢谢我的奶奶从小到大对我的抚养与掌声。其次，感谢我的女朋友和好朋友。谢谢她这两年多的互相陪伴与支持，谢谢我的朋友们一直给予的力量。

感谢深圳大学，给了我一个宽阔的平台，自由地探索自己的兴趣，希望深大以后能够越来越好。感谢我的毕业论文导师孙正佳老师，孙老师耐心、仔细地指导我的写作过程，并给予了許多宝贵的修改建议。

最后，我想感谢自己。这四年带给我的不仅仅是一个较为满意的成绩单和一个满意的研究生 offer，我觉得让我收获更大的是真正感受到了成长带来的心智上的变化。经过四年的数学和金融课程的学习，我收获了大量的知识，掌握了许多重要的学习技能，我也踏上了四年前自己希望前行的路，希望自己能够继续脚踏实地，坚持不懈努力实现自己的目标。

Event-Driven Futures Trading Strategy Based on Trend Signals

【 Abstract 】 This paper focuses on an event-driven futures trading strategy based on trend signals. The rebar futures RB2305 in Shanghai Futures Exchange is used as the trading target, and the news data of the rebar futures in the East Wealth is crawled, then the event sentiment signal is designed and constructed based on the sentiment dictionary of financial events and word embedding. Then, the classifier of event sentiment signals is trained by Logistic Regression. Finally, two backtesting environments, single-average and double-average, are constructed using the Backtrader quantitative platform to compare the event-driven trading strategy proposed in this paper with a common trend trading strategy as the benchmark model. The experimental results show that after applying the event sentiment signal, it can filter most of the wrong trend signals and make the trading strategy perform better and more robust, which confirms the effectiveness of the event-driven strategy based on trend signals proposed in this paper. The most important value of the event-driven futures trading strategy based on trend signals constructed in this paper is that it provides a portable and effective direction for strategy construction, because the results of the experiment are constrained by the limitations of small volatility and short trading time of the target. If replaced with a futures target with high price volatility and trading volume, one can expect the event sentiment signal to be of greater value.

【Key words】 Quantitative trading; Rebar futures; Logistic regression; Word embedding; Sentiment dictionary of financial events