分类号： F830　　学校代码：10697

密　级：公开　　学　　号：202031536

西北大学
Northwest University

# 专业学位硕士学位论文

Dissertation for the Professional Degree of Master

## 基于文本挖掘的分析师情绪指数量化策略研究

专业学位类别：金融硕士

领 域　名 称：金融

作　　者：董刚　　指导老师：王莉　副教授

西北大学学位评定委员会

二〇二二年

# Research on Quantitative Strategy of Analyst Sentiment Index Based on Text Mining

A thesis submitted to

Northwest University

in partial fulfillment of the requirements

for the degree of Master

in Finance

By

Dong Gang

Supervisor: Wang Li Associate Professor

May 2022

# 摘要

分析师在专业技能和信息渠道方面具有普通投资者不可比拟的优势，而且分析师的观点时常被普通投资者采纳并反应在投资决策中，本文以分析师研报作为传递分析师情绪的媒介，以东方财富作为研报数据来源，采用基于筛选和话题向量的情绪提取模型 SESTM 提取研报所蕴含的情绪，并以模型开仓信号为依据构建量化策略，分析策略绩效。

本研究主要从以下几方面展开：第一，基于中证 500 全行业训练单话题模型，得到最优参数。第二，等参数条件下分行业构建话题模型，将分行业模型收益等权汇总，并与全行业模型做对比。第三，使用不同数据源以及不同文本挖掘模型验证本文所使用模型的稳健性，然后将优化后的参数在聚宽平台构建策略并进行回测。

本文研究结论如下：（1）以研报为训练数据的 SESTM 模型具备一定的选股能力。（2）研报数据的市场有效性很短，在一天内即可被市场消化。（3）模型对于研报的积极情绪具有放大作用，但对消极情绪几乎不具备识别能力。（4）分行业模型会显著提高收益，但不可使用单行业模型进行投资决策。（5）分行业模型具备更强的进攻性，而全行业模型具备更好的回撤控制能力。

**关键词：**文本挖掘，分析师研报，量化策略

# ABSTRACT

Analysts have incomparable advantages over ordinary investors in terms of professional skills and information channels, and their opinions are often adopted by ordinary investors and reflected in their investment decisions. In this paper, analyst research newspaper is used as the medium to convey analyst sentiment, and oriental fortune website is used as the data source of the research report. The Supervised Sentiment Extraction via Screening and Topic Modeling was used to extract the emotions contained in the research report, and the quantitative strategy was constructed based on the model opening signal and the model performance was analyzed.

This study is mainly carried out from the following aspects: First, the optimal parameters are obtained based on the whole industry model of CSI500 with single topic. Second, under the condition of equal parameters, we constructed another model by industry, and the benefit of each industry has equal weight, and we compared the difference of two models. Thirdly, We use different data sources and different text mining models to verify the robustness of the model, then we use the optimized parameters to construct strategy which is backtested on joinquant.

The research conclusions of this paper are as follows :(1) SESTM model with research reports as training data has certain stock selection ability. (2) The market validity of research report data is very short and can be digested by the market in one day. (3) The model has a magnifying effect on positive emotions, but almost no ability to identify negative emotions. (4) The model constructed with different topic vectors can significantly improve returns, but the single industry model cannot be used to make investment decisions. (5) The model constructed with different topic vectors has stronger aggressiveness, while the whole industry model has better drawdown control ability.

**Keywords:** Text Mining, Analyst Research Report, Quantitative Investment

# 插图索引

# 表格索引

# 附图索引

# 附表索引

# 目录

# 第一章 绪论

## 1.1 研究背景与意义

### 1.1.1 研究背景

互联网的渗透使得传统的信息交换方式发生了翻天覆地的改变，传统的机械式、近距式信息传播所面临的桎梏迅速被网络打破，作为当今社会至关重要的交流空间，互联网所承担的更多是个人思想的绽放，更多是价值观念的传播，更成为了民众接收信息的第一途径。互联网在金融市场的角色更是不言而喻，交易执行、市场动态、企业资讯均通过互联网实现，极大提高了市场的运行效率。金融从业者的研究对象逐渐从传统的量价基本面数据转向文本数据，文本所蕴含的信息逐渐开始被市场接受并加以挖掘，随之诞生的便是各种各样的非结构化数据研究方法。

文本挖掘的兴起必然离不开计算机科学的进一步发展，因为文本数据在不同语种下都有不同的处理方法，庞杂的数据如果靠人力处理必然耗费大量的时间与精力，最终处理的准确程度不见得会超过计算机。近两年随着编程语言铺天盖地的普及，数据处理手段不断丰富，数据接口不断简化，社会整体的数据处理能力得到提升，成为推动文本挖掘进一步发展的动力。

在金融领域，情绪是从业人员时刻关注的一个指标，情绪影响着市场走向，无论是对基本面投资还是技术面都是关键变量。然而衡量市场情绪的指标有效性参差不齐，有众多经济学者试图构建过市场情绪指标来表征或预测市场走向，但得出的效果往往差强人意，学者转而研究文本中所蕴涵的情绪，也正是在网络信息相对透明且效率极高的背景下，研究投资者情绪才成为现实，而分析师作为集专业技能和一手信息于一身的专业人士，他们定期发布的研究报告中包含了自己大部分的市场观点，通过挖掘分析师发布的研报所蕴含的情绪对于学界而言具备一定的意义。

量化投资指的是利用理论模型分析资本市场各种可投资标的，并采用计算机自动下单交易从而挖掘市场非有效所带来的收益的投资模式。量化投资最大的优点在于不受人主观意愿的干扰，能够依据既定的模型信号进行交易，此外量化交易还具有下单速度快，避免错过市场中转瞬即逝的交易机会等优点。但市场中同时也存在一些反对量化交易的声音，他们认为量化投资依靠机器执行速度快的优势，挤掉众多主观经理

的超额，尤其批评趋势类策略放大了市场波动率。尽管存在众多批评量化的观点，但中国的量化交易规模动能不减，众多前沿理论方法诸如机器学习、深度学习、文本挖掘以及算法交易均在量化交易中得以广泛运用，量化类投顾在近两年取得了较好的表现。在此背景下，本文将结合分析师情绪构建量化策略，为投资者的投资决策提供参考。

### 1.1.2 研究意义

目前学术界针对市场情绪分析的文献多以间接指标构建，利用文本挖掘技术构建情绪指标在近几年才逐渐发展起来，然而众多文献中分析的文本对象主要以股吧评论、市场热点新闻为主，利用分析师出具的研究报告对进行市场情绪测度的文献较少。卖方分析师在技能和信息上所享有的优势是散户投资者所不能比拟的，此外，无论是机构投资者还是散户或多或少都会受到分析师所出具研究报告的影响，最终会体现在投资者的买卖行为中，因此本文选择从分析师研报出发，通过分析大数据下研报所隐含的分析师情绪观点来构建分析师情绪指数，从而在理论层面丰富针对研报进行文本挖掘的相关研究。

本文将利用构建的分析师情绪指数来预测市场走向并根据情绪指标的结果进行个股买卖，指标的有效性同时也会体现在回测结果中，如果利用分析师情绪指标交易的回测曲线具备明显的超额，说明我们所构建的情绪指标对收益具备一定的解释力度。同时，本文利用聚宽回测平台构建策略，希望获得比较稳定的收益曲线，从而对投资者形成一定的借鉴意义。

## 1.2 研究思路与方法

### 1.2.1 研究思路

金融市场从来都是不同的观点、行为作用在一起的结果，至少目前并不存在完全有效的市场，这也是行为金融学研究的基石，人们的心理究竟如何影响他们的行为，从而传导到对市场走势的影响中。本文的研究对象为分析师出具的研究报告，但本质上还是在研究分析师的市场观点，为何不采用散户投资者的观点做研究呢，为了回答这一问题，本文首先对行为金融学相关理论做系统的梳理，研究分析师相对于普通投资者的优势，从而为研究分析师研报所带来的市场反应提供理论支撑；其次，本文梳

理了文本挖掘理论的基本流程，说明了传统文本挖掘方法存在的一些瑕疵，并简要介绍了本文所使用的 SESTM (Sentiment Extraction via Screening and Topic Modeling) 基于筛选和话题模型进行情绪提取的优点和步骤。然后对国内外相关文本挖掘和情绪指数文献进行了系统的梳理。

在搭建好理论基础之后，本文将从以下几个方面展开研究：（1）首先利用训练集数据训练全行业话题模型，全行业话题模型是我们搭建模型的第一步，该小节会对文本的数据预处理和 SESTM 模型建模做详细介绍。（2）由于全行业话题模型使用的是同一个话题矩阵，而不同行业话题必然存在差异，所以本文在搭建好全行业话题模型之后又单独对每一行业分别训练以此构建分行业模型并对比与全行业模型之间的差距。（3）在构建量化策略前，我们采用不同数据源以及不同的文本挖掘模型进行测试，在证得模型稳健的前提下，构建量化策略并进行回测以观察策略表现，从而综合评价该模型效果。

### 1.2.2 研究方法

（1）文本挖掘法

文本挖掘方法是大数据时代下利用非结构化数据进行研究的方法，本质在于如何使文本数据转化为可以量化处理的结构化数据，而这一转化过程就是文本挖掘技术的核心。文本挖掘流程包括对文本数据分词、去停用词、去标点符号等操作，得到较为干净的词典是文本挖掘的第一步，而后确定特征项的权重也是文本挖掘的重要过程，最终目的在于从文本中提取出包含主要信息的情绪矩阵。

（2）潜在狄利克雷分布话题模型算法

潜在狄利克雷模型引入参数的先验分布，并结合贝叶斯估计方法从文本语料中发现隐藏在词汇表面之下的潜在语义，弥补了有限数据统计存在的缺陷，提高模型的泛化性能。本文在判断合适的话题个数之时，采用潜在狄利克雷分布模型，并利用余弦相似度检验每一个话题个数下语料间的平均相似程度，以此挑选对语料最具区分能力的话题个数。

（3）对比分析法

对比分析法在文献中经常被用来衡量文章选用模型的效果，本文在对全行业语料训练建模之后，为了突出模型优越性，将相同的数据源放至传统的词典模型之中进行

建模，并对两个模型的效果进行比较，以检验模型是否能有较好的表现。

## 1.3 研究内容与框架

### 1.3.1 研究内容

本文一共包含六章，各章节的主要研究内容如下：

第一章为绪论。绪论部分主要表达本文研究方向的现实背景，大数据时代丰富的文本信息叠加先进的数据处理与计算能力促进了文本挖掘技术的迅猛发展。同时目前对投资者情绪的研究多停留在散户投资者的股评言论上，较少以分析师研报作为文本对象进行研究的文献，从而引出本文研究的理论意义与现实意义。绪论部分也介绍了本文所使用的主要研究方法，梳理了本文各章节的研究内容与框架，并阐明了本文的创新点。

第二章为相关理论与文献综述。这一章系统的介绍了本文理论基础——行为金融理论，与所使用的文本挖掘方法。行为金融理论的介绍主要从信息不对称、投资者认知限制和“交易诱导”这三个方面展开，为本文研究分析师情绪提取理论支撑。文本挖掘方法则按照将文本转为向量矩阵与从向量矩阵提取有用信息两个步骤展开，紧接着介绍了本文所要使用的 SESTM 模型（基于筛选和话题模型的情绪提取）的建模过程。在介绍完理论和方法之后，本文对两方面的国内外文献做了系统的梳理，主要包括文本挖掘相关文献以及投资者情绪指标如何构建的相关文献，最后对该部分做出文献评述。

第三章为分析师情绪指数构建。本章首先介绍数据来源与数据预处理流程，并说明了词典的生成过程，紧接着构建 SESTM 模型，包括筛选情绪敏感词汇，为研报文章打标签，训练出两话题向量，并利用话题向量预测新研报标题的情绪，将预测情绪值用于个股买卖。通过控制情绪阈值变量，研究发现基于研报的策略市场持续性比较短，而且一天持仓周期表现比较好，通过多空组合和纯多头策略对比发现该模型对于积极词的识别能力要远超消极词。

第四章为分行业检验分析师情绪指数。本章在上一章构建出全行业模型的基础上，通过对不同行业分别构建话题向量，并依据不同行业的话题向量进行情绪预测和个股买卖，最终汇总模拟出分行业模型净值走势。为了与全行业模型进行对比，本章考虑

到选股池规模差异，重新构建了仅用于本章对比的“全行业”模型，以此分析分行业模型是否有更好的模型表现。经对比发现分行业模型的收益和回撤均比“全行业”模型要高，具备更强的进攻性，但是从夏普和卡玛比率两个指标来看，“全行业”模型更为稳定。

第五章为模型稳健性检验与分析师情绪指数策略构建。首先为了检验 SESTM 模型在不同数据源中的适应能力，将该模型放在中证 1000 的研报数据中进行分析；然后为了检验 SESTM 模型相对于传统模型的优越性，利用传统的词典方法来预测研报情绪走向，结果表明本文所使用的 SESTM 模型在其他数据中也有良好表现，同时相较于传统词典模型表现更优。在模型稳健的前提下，基于前几章节所构建的 SESTM 模型得出的最优参数，设定一系列符合现实交易环境的成本在聚宽平台进行回测，然后对回测结果进行绩效评价，回测结果表现出很高的收益，但与此同时作者指出了本研究不可避免的部分“未来函数”会对模型产生一定影响。

第六章为结论。本章主要对本文所使用模型以及得出的结论进行概括，同时指出文章的不足之处，以及下一步的研究方向与展望。

### 1.3.2 研究框架

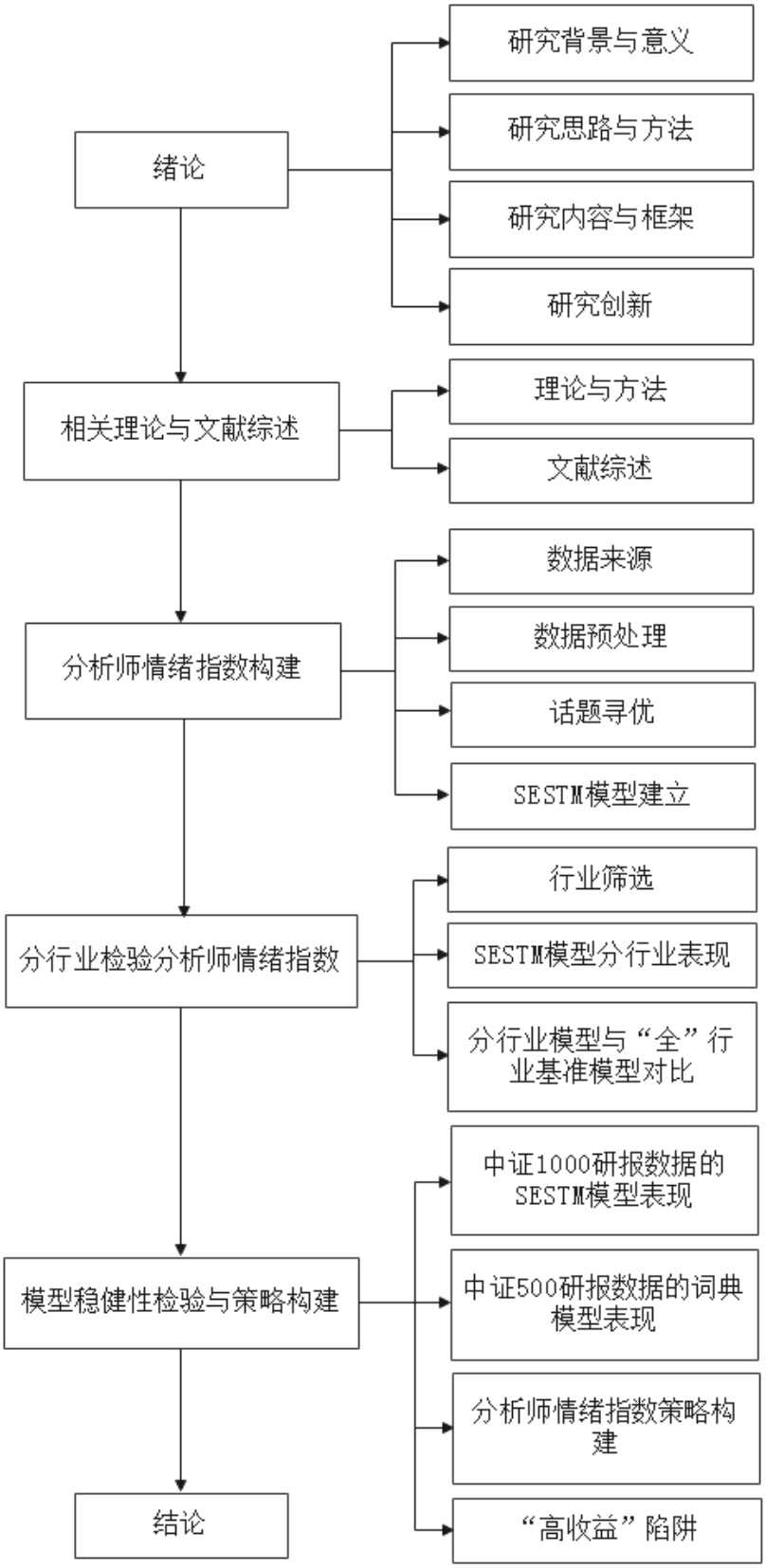



**图 1 研究框架**

## 1.4 研究创新

本文创新之处在于运用文本挖掘对分析师研报进行情绪提取从而构建情绪指数，并将其运用到个股预测中。以往针对分析师研报的研究更多是对数量指标的分析，或者是对分析师研报预测的诱导性或者独立性进行研究，本文在考虑这些因素的基础上进行更深一层的研究，即使存在这些非理性的因素，探究利用研报所构建的策略是否存在策略价值。另一方面本文所使用的模型完全透明化，读者能够密切跟踪预测输出过程，不会出现传统机器学习的“黑箱”，有助于读者根据自己需求对模型参数进行调整和优化。

# 第二章 相关理论与文献综述

## 2.1 理论与方法

### 2.1.1 行为金融理论

长久以来，学界都按照有效市场假说来研究整个市场的有效性，但现实中的股价走势却和有效市场假说的理论大相径庭，在假设条件逐渐苛刻，假设无法成立的现实世界中，学界不得不开辟新的研究方向进而解释资产价格的变化，行为金融学应运而生。与传统金融学不同，行为金融学并没有理性人假设，它认为人的理性都是相对有限的，无论是专业投资人还是散户，都会受到环境带来的影响，这种行为上的偏差导致资产价格偏离内在价值，也就是资产误定价现象。

行为金融学研究认为投资者情绪是影响金融市场运行的重要因素，而在传统金融学范畴内，理性人的假设隐含了投资者的决策合理性，与现实存在巨大差异。Barberies et al., (1998)[1]提出了投资者情绪理论模型，该模型认为人们进行投资决策时存在两种错误范式：一种是个体由于锚定效应存在导致新信息出现时对原有信念修正不足所产生的保守性偏差，表现为股价对新信息反应不足，另一种则是投资者存在代表性启发导致股价对信息反应过度。

（1）信息不对称

信息不对称表现为利益双方可获得的信息存在质与量上的差异，该现象在金融市场中尤其明显。投资者大致可以分为散户和机构投资者，机构投资者背后往往有分析师意见作为参考，而分析师与散户相比，无论是在信息优势还是专业技能上都更胜一筹。分析师背后资金实力雄厚，获得信息的渠道广，通常能获得一手信息，并结合自己的专业知识给出投资意见。

（2）认知限制

认知学研究表明人的大脑处理信息的能力是有限的，会存在诸多限制以至于无法及时处理全部信息，人们只能在约束下做出次优解，而无法实现完全理性下的最优解。

认知限制很重要的一方面是有限注意力理论。传统金融学理论认为人们进行投资决策时会使用全部可以得到的信息，其投资行为理论上应该是对所有信息加工处理后的结果，但实际上却无法做到，尤其是在大数据时代，海量信息更加放大了投资者的

认知局限性，人们会偏好去应对最显著、最重要的信息。

散户是有限注意力理论很好的体现，大多数散户投资者都不是职业股民，只不过将投资作为业余理财的一种方式，而且散户只会对自己所关注的信息进行分析与投资决策，而分析师的工作就是每天监控市场动态、热点新闻、分析事件冲击对市场的影响程度并提供投资意见，因此单从认知限制角度来看，分析师可信度高于散户，这也体现了“将专业的事情交给专业的人去做”这一思想。

（3）交易诱导

阿克洛夫和席勒(2016)[2]将欺骗划分为信息型和心理型两种，前者是指利用人们缺乏信息或容易受到错误信息的误导而实施的欺骗，后者是指利用人们在认知上的障碍或心理上的偏误而实施的欺骗。在资本市场中，我们可以将分析师看作是拥有可靠信息资源且经验丰富的一方，而散户投资者就属于信息相对匮乏且经验不足的一方，分析师具备实施“欺骗”的条件。有心理学研究表明，人们的思想很容易受到叙事性思维的影响，即想法很容易在与他人的交流中改变，胡昌生和高玉森(2020)[3]认为分析师具备对投资者进行思维移植的能力，通过利用植入故事情节的研报放大分析师自身的情绪并诱导投资者进行交易。

在上述理论支撑下，本文认为无论是出于信息优势还是认知上的突破，又或者说是分析师“诱导”交易的行为，最终的结果是卖方分析师情绪的确会影响散户投资者决策，而传递分析师情绪最直观的体现是研报，机构投资者往往有自己的买方分析师提供投资建议，因此本文研究重点主要是卖方分析师的公开研报数据。

### 2.1.2 文本挖掘理论

随着信息的爆炸式增长，人们获取信息的数据来源也呈现出多样性，从结构化数据到非结构化数据的演变，文本数据的重要性在互联网衬托下愈加显著，面对多样、庞大且更新速度极快的文本数据，文本挖掘技术为我们分析这些文本提供了极大的便利。沈艳(2019)[4]将从原始文本数据出发解释因变量的过程分解为三个步骤：第一，将文本库所有文本转化成数据矩阵；第二，通过计量或者统计方法将数据矩阵转为目标信息序列，如情绪、关注度等指数；第三，使用从文本中提取出来的信息序列来解释或预测因变量。本文将借鉴这一思路梳理文本挖掘过程。

（1）非结构化数据到结构化数据的转变

目前学界的定量分析体系大多是建立在结构化数据的基础之上，文本挖掘技术的本质是将不规范的非结构化数据转变为可计量、可操作的的结构化数据。从这一本质出发，当我们拿到一堆文本数据时，首先要将文本格式规范化，将其处理为最小语义单元，然后结合不同语言的不同特点运用合适的计量方法将文本表达成目标数据矩阵，并最大程度保留原有数据结构的信息。

常规的数据预处理要包括分词、去停用词、词形规范化。分词是指将给定的文本切割为词汇单词的过程，这一过程对不同的语言处理方法不同，对于英文而言，英文天然以空格为分隔符，只需要利用空格或者标点就能实现分词效果，但对于中文之类的连续序列而言，就必须按照一定的规范将汉字序列切分成词语，关于汉语自动分词的方法，国内外有大量研究工作，从早期基于词典分词法发展到基于n元语法的统计切分方法，再发展到后来的由字构词的汉语分词方法，极大提升了汉语分词方法的丰富性。

目前国内最流行的中文分词工具包是Jieba，考虑到Jieba分词的接口使用简便以及算法的优越性，本文也将采用这一工具包进行分词处理工作。之所以要去停用词，是因为文档中频繁出现的、附带极少文本信息的助词、介词、连词和语气词等高频词对于文本区分没有太大的实质性意义，为了减少文本挖掘系统的存储空间，提高文本运行效率，我们需要在分词之后将这些停用词从词典中去除。词性规范化更多是针对西方语言而言，以英文为例，为了提高文本处理的效率，减缓离散特征可能导致的数据稀疏问题，我们需要将词形还原，比如将复数统一变为单数，同时还需要进行词干提取，去除词缀得到词根，从而降低单词表示的复杂度与稀疏性，增强文本数据分析的稳健性。

得到分词后的词典之后，下一步工作是将文本向量化。这一步工作的本质是为不同的特征项赋予不同的权重从而突出不同的特征项在文本中的重要性。常见的特征项权重包括以下几种：

1）布尔权重

布尔权重即简单的将特征项是否出现在文本中作为该特征项的权重，表示如下：

$$bool_i = \begin{cases} 1, & \text{如果特征项} t_i \text{出现在文本中} \\ 0, & \text{否则} \end{cases} \tag{2.1}$$

2） 特征频率-倒文档频率(TF-IDF)权重

特征频率 (TF) 表示该特征项在当前文本中出现的次数，该权重假设高频特征项所包含的信息量高于低频特征项的信息量，也就是说特征项在文本中出现的次数越高，其重要性越大。公式如下：

$$TF_i = N(t_i, d) \tag{2.2}$$

公式中的$t_i$代表第 i 个特征项，d 代表整个文本。如果考虑到某些高频特征项的绝对词频远高于平均权重，我们还可以采用对数处理的方式以简便文本表示。

$$f_i = \log(tf_i + 1) \tag{2.3}$$

在解释倒文档频率 (IDF) 之前，我们首先要了解文档频率的定义，文档频率表示语料中包含特征项的文档个数，特征项的文档频率越高，其包含的有效信息量往往越低，倒文档频率的定义如下：

$$idf_i = \log\left(\frac{N}{df_i}\right) \tag{2.4}$$

式中的 N 代表语料中的文档总数，$df_i$代表特征项$t_i$的文档频率。将特征频率与倒文档频率相乘，便得到了特征频率-倒文档频率权重：

$$tf_{idf_i} = tf_i * idf_i \tag{2.5}$$

特征频率-倒文档频率法认为对区别文本中最有意义的特征项应该是那些在当前文本中出现频率足够高，而在文本集合的其他文本中出现频率足够小的词语。

经过以上处理我们能够将文本变成最小基本语义单元，并应用不同的赋权方法对特征项赋予权重以更好的表示该文本语料，同时最大程度保留文本信息，至此我们能够得到该语料的向量矩阵表示，一维代表所有文档，另一维代表特征项集合。目前国内常用的文本向量表示技术是 Word2Vec 模型，该模型在经济金融领域的的应用越来越广泛，Li et al(2019)[5]对比了独热表示法和 Word2Vec 两种方法，发现相较于独热表示法，使用 Word2Vec 表征文本会显著提高文本情绪的分类准确性。

（2）数据矩阵的信息提取

在我们将非结构化的文本数据转化为数据矩阵后，接下来要考虑的工作就是如何从数据矩阵中获得我们想要的信息。在金融领域，应用最广泛的方法就是词典法。词典法的有效性建立在一个合适的词典之上，通过统计文档中不同类别的词语在词典中出现的次数，再结合不同的加权方法来提取文本信息。国外比较有代表性的英文词典

包括 GI 英文情感词典、WordNet 情感词典[6]；国内比较权威的情感词典包括大连理工大学情感本体库，知网 Hownet 情感分析词典[7]等。但是在实际应用中，直接拿公开词典进行分析的文献并不多，学者往往会根据已有语料库进行情感词典的扩充。在确定词典后，接下里要处理的问题就是如何确定情感词的权重，常见的加权方法在上一小节已经介绍过。

在情感分析框架下，我们需要考虑的一个问题是如何对情感进行分类，直接区分为“积极”和“消极”两类还是区分为“积极”,“消极”和“中性”，这对于分析不同的问题会得到不同的结论，在本文中，我们将采用潜在狄利克雷分布模型(Latent Dirichlet Allocation, LDA)[8]来检验合适的话题个数，该模型能有效提取大规模文档集合和语料库中的隐含主题，起到了很好的降维与信息提取效果，有关具体 LDA 模型的研究和应用可参考何伟林等(2018)[9]的研究综述。

### 2.1.3 SESTM 模型

本文将借鉴 Zheng Tracy Ke, et(2019)[10]提出的 SESTM 模型来提取文本信息，之所以选择该模型，是因为该模型具备一个很大的优势——透明度高。目前大部分深度学习方法对用户而言就是个黑箱，从数据的输入到输出，我们无法得知其中的处理过程，但 SESTM 模型使用的有监督学习方法可以称之为“白箱”，用户可以直观的看到数据的生成过程，下面对该模型做一个详细的描述。

SESTM 模型主要包括三个步骤：

第一、从大量术语词汇中分离出最能代表文档信息的特征。我们可以采用一定的降维方法获取最具信息提示意义的金融市场相关词汇，在面对多变量回归时，比较常用的降维方法是主成分分析法，但由于在文本数据分析中最小语义单元是词语，庞大的词维度造成了主成分分析所不可避免的稀疏性问题。因此在 SESTM 模型中，筛选特征词汇的思想是找到那些与股票收益相同信号且共现最频繁的词汇。为了从降维后的情绪术语列表中尽可能快的得出估计，挑选出通过相关性筛选的词语是必要的过程。

第二、基于已有语料库训练出上一步筛选出来的特征词汇的情绪权重。虽然目前的词语赋权方法已经考虑到了词频巨大差异导致的权重错配问题，但是 SESTM 模型使用的是基于概率的方法考虑词频的偏度。

第三、使用语料库训练集训练出的话题向量为样本外的研报情绪打分。计算出新

文章中特征词汇的词频向量，并结合训练出的话题向量，使用最大似然概率估计法来综合估计每一篇研报的情绪值。SESTM 模型巧妙的一点是设计了一个带有未知参数的最大似然估计量来估计文章情感，施加这一惩罚的贝叶斯解释是对集中在 1/2 的情感分值上施加一个$\beta$分布的先验，即我们的估计从认为一篇文章是情绪中立开始。

由于笔者能力有限，下面仅对 SESTM 模型的数学部分做一个理论上的描述，至于具体模型推导过程详见 Zheng Tracy Ke(2019)一文的附录。

假设现在有 n 研报章以及 m 个单词的字典，将第 i 篇研报的词频记为向量形式$d_i$, 于是我们可以将$d_{i,j}$作为单词 j 出现在研报 i 中的次数。将这 n 篇研报均表达为同一个词典的向量形式，即可得到一个形状为 n*m 的矩阵 D, D=[d1, d2, …, dn]'。由于词典中并非所有词都和该研报表达的情绪相关，所以我们需要从该词典中识别出最为情绪敏感词的集合 S，并用集合 S 来向量化表示 n 篇研报，记相应的矩阵为$D_{\cdot,[S]}$, 用$d_{i,[S]}$表示新矩阵$D_{\cdot,[S]}$的行向量。每篇研报都有自己对应的股票名称，所以我们记研报 i 在发布日的对应股票收益为$y_i$。

假设每一篇研报都有一个情绪分值$p_i$位于[0, 1]的闭区间内，p 为 1 代表研报积极情绪最强，p 为 0 代表消极情绪最强；假设$p_i$是代表研报对股票收益影响的有效统计量，即在$p_i$给定的情况下，$d_i$和$y_i$之间是相互独立的。除该条件独立假设之外，引入两个额外成分解释数据生成过程，一个支配着给定$p_i$条件下股票收益$y_i$的分布，一个支配着给定$p_i$条件下研报词向量$d_i$的分布。对于条件收益分布，假设

$$P(sgn(y_i) = 1) = g(p_i), g(\cdot)\text{为单调递增函数} \tag{2.7}$$

公式中的 sgn 为符号函数，即如果股票收益为正则等于 1，如果股票收益为负则等于 0，直观上，这一假设表明研报情绪值$p_i$越高，实现正收益的可能性越大。

对于研报词向量$d_i$，假设 m 个词的词典：

$$\{1,2,\dots,m\} = S \cup N \tag{2.8}$$

公式中的 S 代表情绪敏感词集合，N 代表情绪中性词集合，二者并集为整个词典。假设 S 的维度为丨S丨，则 N 的维度为 m-丨S丨，那么$d_{i,[S]}$就代表由情绪敏感词 S 所表示的研报词频向量，同样，$d_{i,[N]}$ 就代表由情绪中性词 N 所表示的研报词频向量。模型

假设$d_{i,[S]}$和$d_{i,[N]}$之间相互独立，由于$d_{i,[N]}$对于判断研报的情绪来说是额外的噪音，所以模型不对$d_{i,[N]}$建模。

模型假设情绪敏感词词频向量$d_{i,[S]}$服从混合二项分布：

$$d_{i,[S]} \sim Multinomial(s_i, p_i O_+ + (1-p_i)O_-) \tag{2.9}$$

公式中的$si$代表研报 i 中情绪敏感词的总个数。接下来，用两话题模型为单个情绪敏感词概率进行建模，O+代表了当研报情绪分值 $p_i$ 等于 1 的时候情绪敏感词集合 S 的词频概率分布，O-则代表当研报情绪分值 $p_i$ 等于 0 的时候情绪敏感词集合 S 的词频概率分布。在绝大多数情况下，研报情绪分值 $p_i$ 介于 0-1 之间，且词频向量为两个话题向量的概率结合。模型整体结构如**图 2** 所示：

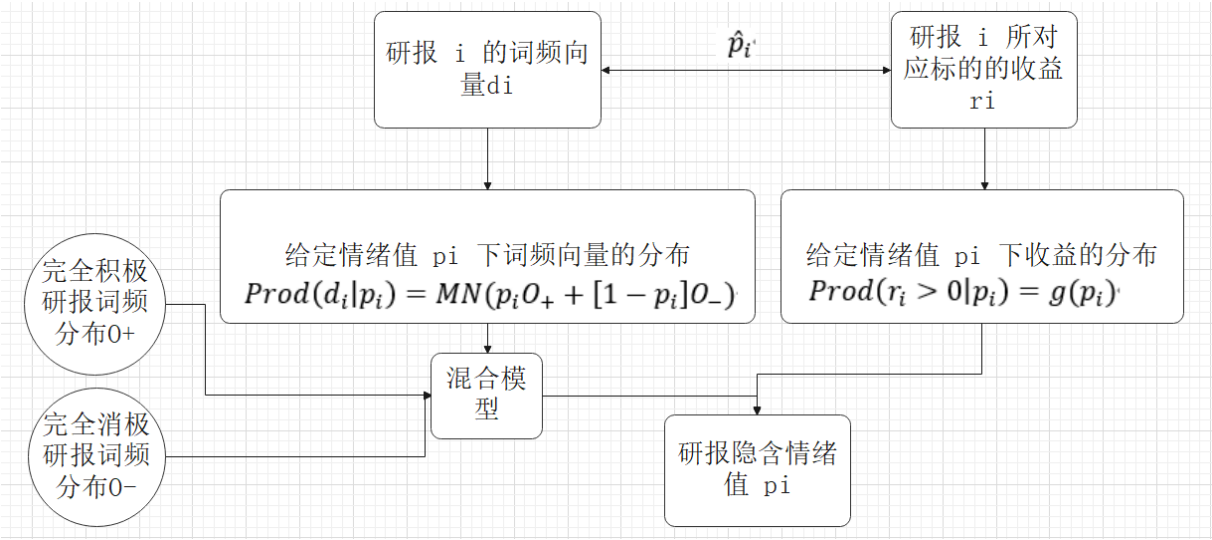



**图 2 SESTM 模型结构**

在搞懂模型原理后，最主要的任务就是学习模型参数，$p_i, O_+$和$O_-$，关于这一参数的学习将在建模过程中逐步探讨。主要就是三个步骤：1、分离出情绪敏感词集合 S；2、确定话题向量$O_+$和$O_-$；3、估计研报层面情绪分值 $p_i$。

首先第一步是分离出情绪敏感词集合，该步骤需要用到特征提取技术，SESTM 模型使用的方法是采用有监督方法，利用研报对应标的的股票收益从中提取研报情绪。如果研报中的某个词的出现经常伴随着标的的正向收益，则认为这篇研报的情绪是正向的。为了提取与正收益共现频率最高的单词，首先需要计算每个词的的频率。公式如下：

$$f_j = \frac{\text{包含词语}j\text{且标的为正收益的研报数量}}{\text{包含词语}j\text{的研报数量}} \tag{2.10}$$

在得到每个词语的正收益共现频率后，还需要设置共现频率的上下阈值，因为常规的 0.5 阈值可能不符合文本分析的情感特征，另外，还需要设置词语在研报中出现的次数的阈值，因为可能有些词在研报中出现的频率很低，但是恰巧出现的几次标的收益都是正的，这类携带噪音的词语需要删除掉。设定了$\alpha_+$、$\alpha_-$和 k 之后，至此我们就得到了情绪敏感词集合 S，公式如下：

$$\hat{S} = \left\{j: f_j \geq \frac{1}{2} + \alpha_+, \text{或者} f_j \leq \frac{1}{2} - \alpha_-\right\} \cap \{j: k_j \geq k\} \tag{2.11}$$

在得到情绪敏感词集合 S 后，现在我们的目的是训练话题向量 O。$O = [O_+, O_-]$，这一话题向量决定了每一篇研报中的敏感词的生成过程。O 即能捕捉到词频又能捕捉到词语的情绪。

在 SESTM 模型中，每篇研报的情绪分值用参数$p_i$代表，$p_i$描述了研报对于积极词分布的依赖程度。假如说我们已经观察到所有研报文章的情绪分值，使用$\widetilde{d_{\iota,[S]}} = {d_{i,[S]}}/{s_i}$代表词频向量，则有

$$E\widetilde{d_{\iota,[S]}} = E\frac{d_{i,[S]}}{s_i} = p_i O_+ + (1 - p_i)O_- \tag{2.12}$$

用矩阵形式表达，就是

$$E\widetilde{D'} = OW \tag{2.13}$$

$$W = \begin{bmatrix} p_1 \cdots & p_n \\ 1 - p_1 \cdots & 1 - p_n \end{bmatrix}, \quad \widetilde{D} = [\widetilde{d_1}, \widetilde{d_2}, \ldots, \widetilde{d_n}]'$$

总的看来，D 代表由情绪敏感词集合构造的语料库的词频向量，O 代表两话题向量，W 代表语料库文章的情绪矩阵，我们的目的是通过 D 对 W 回归来估计话题向量，但是 D 和 W 都无法直接观察到，我们首先要得到两个参数的估计值，为了估计 W，我们使用训练集中所有研报对应标的的收益标准排序来估计 W，公式如下：

$$\widehat{p_\iota} = \frac{rank\ of\ y_i\ in\ \{y_l\}_{l=1}^{n}}{n} \tag{2.14}$$

使用研报情绪的估计值$\widehat{p_\iota}$来填充情绪矩阵 W。

在利用训练集数据训练好话题向量 O 后，接下来就要讨论如何估计测试集中的新研报的情绪分值。假设$d_i$是研报的词频向量，$s_i$是该篇研报情绪敏感词总个数，则有

$$d_{i,[S]} \sim Multinomial(s_i, p_i O_+ + (1-p_i)O_-) \tag{2.15}$$

给定$\hat{S}$和$\hat{O}$之后，SESTM 模型中使用最大似然估计法来估计研报情绪值$p_i$。同时还在估计方程中添加了一项惩罚项，惩罚项存在的意义是可以解决研报所含情绪敏感词不足以及低信噪比的问题。施加这一惩罚项将预测出的情绪值倾向于接近 0.5，这一缩减程度取决于惩罚系数$\lambda$的大小，主要是考虑到大部分研报标题情绪趋于中性。研报预测情绪值的公式如下：

$$\hat{p} = \arg max \left\{ \hat{s}^{-1} \sum_{j=1}^{\hat{s}} d_j \log\left(p\hat{O}_{+,j} + (1-p)\hat{O}_{-,j}\right) + \lambda \log\left(p(1-p)\right) \right\} \tag{2.16}$$

这里的$d_j$，$\hat{O}_{+,j}$和$\hat{O}_{-,j}$分别是对应向量的第 j 个元素。

## 2.2 文献综述

### 2.2.1 文本挖掘文献综述

文本挖掘技术可行性得到认可并逐渐运用在各大领域，包括旅游、餐饮、医学和金融等各大领域，文本所富含的信息得到有效利用。目前大多数文本挖掘方面开展的研究以评论为分析对象，Reyes(2012)[11]通过建立代表“讽刺”意味的关键词数据集来自动识别客户评论的“讽刺”倾向，董爽(2017)[12]、刘敏(2018)[13]和张红霞(2019)[14]分别以 B2C 购物网站评论、网络商品评论、天猫生鲜频道评论为对象展开评论信息挖掘，Pu(2019)[15]利用 SVM 对客户意见句子进行编码并进行文档情感分类以证实模型有效性，范宁(2019)[16]和何立峰(2019)[17]均基于情感词典法提取酒店评论所包含的情绪信息，并将其运用到客户需求挖掘模型中，李薇和杨东山(2021)[18]以美团餐饮评论为研究对象，进行词义网络图和系统聚类分析，并提取出影响客户满意度的关键指标。

随着文本挖掘技术的普及，文本挖掘的分析对象也在逐渐多样化。杨秀璋、武帅、夏换等(2020)[19]以微信公众号有关贵州三大战略行动网页文本为分析对象，从而研究民众关注的热点话题，为国家政策分析和舆情挖掘提供帮助。晁筱雯(2021)[20]运用文档语义结构挖掘工具 gensim 对近三十年以来的传染病类期刊文章进行处理，并绘制出传染病相关科学主题的演进模式。姜坤、刘苗(2021)[21]以印度主流媒体关于中美关系的新闻语料作为分析对象从而进行印媒新闻话语的情感立场分析。许光、任明、宋城宇(2021)[22]通过分析达沃斯论坛期间的新闻文本得出从主题、观点和倾向三个角度提取国

家形象的结论。陈聪聪, 赵怡晴, 姜琳婧等(2021)[23]采用 Apriori 关联算法分析尾矿事故因素关联性，卢玉昆和唐文(2021)[24]对古代描写南京的诗词文本做情感分析从而探究城市景观保护与延续策略，刘赛红, 黄馨锋和余意(2021)[25]从论文中提取农业主体经营风险相关信息，韩天园, 田顺, 吕凯光等(2021)[26]通过对 254 份特大交通事故调查报告进行信息提取，发现了安全运行系统失稳的根本原因。

除了利用文本挖掘技术进行定性分析之外，还有众多文献将该技术应用到预测研究中，司法领域也不缺文本挖掘的身影，舒洪水(2020)[27]以中国裁判文书网判决书为分析对象，对其进行自动化分类以及数据转换，并利用线性回归实时分析判决数据，从而构建量刑预测模型。石勇, 安文录, 曲艺(2021)[28]则运用文本挖掘技术建立起一套检察起诉决策支持系统。当然文本挖掘的应用更多还是出现在金融领域，杨超、姜昊和雷峥嵘(2019)[29]运用百度搜索数据和美元兑人民币汇率中间价日数据建立包含网络搜索指数的多变量模型进行汇率预测并证实了短期内人民币汇率的可预测性。戴德宝、兰玉森、范体军等(2019)[30]通过构建上证投资者综合情绪指数预测股市价格变化，发现该情绪指数能够提高股指走势预测的精度。张杰、张永卿和翟东升(2021)[31]通过引入互联网财经新闻对日汇率波动趋势进行预测，发现该方法能提高汇率波动趋势预测的准确率并能获得较高的投资收益。

### 2.2.2 投资者情绪文献综述

投资者情绪的研究历经几次发展，从最初采用市场调查得到的直接计量指标到采用股票市场可观测到的经济变量，再发展到现在的利用文本大数据挖掘投资者情绪，度量方法逐步更新迭代，对情绪的测量也逐渐清晰。

（1）直接计量法

直接计量法采取市场调查的方式获得最原始的情绪指标，国内以前常用好淡指数、央视看盘指数和消费者信心指数来反映市场情绪，不同之处在于前两个指数的调查对象均为机构投资者，而消费者信心指数调查对象是个人投资者。Cheng(2005)[32]研究发现好淡指数对牛熊市具有一定的区分度，Xue(2005)[33]基于封闭式基金折价率研究发现消费者信心指数能有效解释投资者情绪。

（2）间接计量法

股市所反映出来的信息是投资者对未来预期最即时的刻画，因此学界通过构造股

市指标间接反应投资者情绪，比如股市换手率、IPO 溢价率、封闭式基金折价率等。

封闭式基金折价是指在基金封闭期间，单位份额市价低于单位份额净值，反映了该基金在市场上的热度。封闭式基金折价率最初由 Delong, et(1990)[34]引入，而后张超(2014)[35]利用封闭式基金周振幅作为噪音交易的替代指标对折价问题进行研究，研究发现我国封闭式基金折价主要受噪音交易的影响。汪丽雯(2019)[36]则发现不同市场下封闭式基金单位净值收益率与折价率之间呈现不同关系，牛市两者呈正相关，熊市则为负相关。

换手率被定义为成交量与流通股数的比率，反映了市场流动性。Statman et al.(2006)[37]从行为金融学角度出发，研究指出换手率能够反映投资者非理性情绪，比如盲目乐观、过度自信，因而换手率更像是一个情绪指标。Wen(2015)[38]综合考虑封闭式基金折价率、IPO 数量、IPO 首日收益率和换手率，利用主成分分析法构建投资者情绪指标，发现正向投资者情绪对股票收益影响较为显著，而负向投资者情绪影响不显著。

（3）基于文本数据的投资者情绪文献

1）非分析师角度情绪文献

随着互联网的深度普及，网络成为投资者发表观点的首要渠道。与其从股市中找寻反应投资者预期的既定事实，提前从投资者发表观点中挖掘其情绪走向或许是个更好的选择。随着文本挖掘和情绪提取技术的崛起，提取并分析实时文本信息成为学者研究手段之一。

由于网络论坛与新闻数据信息量大而且公开易得，最初的文本研究多基于此类数据，Antweiler 和 Frank(2004)[39]基于网络论坛帖子积极与消极分类构建投资者情绪指标，Bollen(2010)[40]和 Oh(2011)[41]均通过股票推文或评论研究文本包含的情绪与股市收益率之间的关系。Bollen 和 Mao(2011)[42]通过推特内容研究公众对总统选举的反应能力所产生的情绪时间序列来预测道琼斯指数收盘价的变化，并得出包含公众情绪的指标能够将道指预测准确程度提升至 86.7%的结论。Oliveira(2017)[43]通过合成包含推特情绪的单指标证实推特数据对股市预测的有效性。雨婷, 宋泽芳和李元(2021)[44]则是基于情感词典，采用 SVM 对股评文本分类并构建文本情绪指数，研究表明投资者情绪对股票收益率具有短期正向预测作用和长期负向预测作用。

高频预测也是众多学者应用文本挖掘的研究领域之一，Renault(2017)[45]利用推特

数据构建股市涨跌观点的词汇库，以半小时为周期动态监测投资者情绪，并得出投资者情绪的前半小时的变化可以预测标普500ETF后半小时的回报的结论。Yin(2019)[46]同样以半小时为节点，以发帖量为指标，得出日内高频投资者情绪正向影响股市运行的结论。徐维军, 付志能, 李茂昌(2021)[47]利用BERT模型构造基于新闻数据的股指期货高频预测模型，回测发现该模型能够取得较高准确率和收益率。

2）分析师情绪文献

分析师发布的研究报告是其观点表达的主要途径，可以通过处理研报文本反映分析师市场情绪。早期学者研究分析师研报能否显著提升投资者收益，这一论题有不同的研究结论，Frey, Herbst(2014)[48]认为卖方分析师的研报作为公开信息，已经反映在股票价格中，因而得出卖方分析师研报不具明显价值。但 Cheng(2005)[49]、Costello 和 Hall(2011)[50]则发现美国资本市场机构持股方向和卖方分析师的预测呈现显著的正相关性。Franck 和 Kerl(2013)[51]研究发现与分析师预测一致的基金投资行为会明显提高组合收益。

随着以传统论坛新闻数据识别投资者情绪的研究逐渐丰富，学者逐渐开始研究以研报为分析对象所反映出的分析师情绪对市场的影响。目前研报处理主要包括定性和定量两种方式，定量研究主要利用研报发布密集程度来反映分析师情绪强烈，伊志宏等(2015)[52]采用分析师每年对每家公司发布研报数量的均值衡量其努力程度，发现单位时间内关注个股的分析师越多，发布的研报数量越多，则市场情绪越高涨。胡昌生和高玉森(2018)[53]用分析师研报数量代理分析师情绪，发现代理变量和股价之间呈现“跷跷板”效应。定性研究主要是运用文本挖掘将研报内容进行量化，目前针对研报的研究多数针对研报标题或具体内容展开。蔡庆丰和杨侃(2013)[54]利用分析师研报对于个股的评级以及评级变动来度量分析师情绪，戴方贤和尹力博(2016)[55]则以分析师的价格预测信息作为衡量指标。

有学者对分析师发布研报的独立性进行了相关研究，张化侨(2010)[56]研究发现分析师会为了短期利益而避免向市场传递有效信息，蔡庆丰和杨侃(2013)[57]认为A股市场下分析师出于自身利益考量为了迎合噪音交易者会发布有偏研报，吴超鹏等(2013)[58]认为分析师为迎合机构投资者倾向于提供乐观的评级。张宗新和杨万成(2016)[59]从证券分析师的声誉和挖掘信息能力两个角度出发研究分析师的影响路径与机制，得出结论：分

析师能通过声誉模式和信息模式直接影响市场；从长期看，信息模式能够提高多空双方的投资收益，但根据新财富分析师称号进行组合构建仅在卖空组合上绩效较好。丁方飞等(2019)[60]运用向量自回归和结构向量自回归模型对分析师的市场影响进行研究，发现分析师无法做到完全理性引导，有些情形下即使证券分析师也会随波逐流。胡昌生, 高玉森(2020)基于沪深二级市场 2012 年 1 月 1 日至 2016 年 12 月 31 日的数据进行横截面分析，研究得出 A 股分析师发布的研报多数是“虚情假意”的，也即分析师存在一定程度的诱导和欺骗行为。

在上述“交易诱导”理论支撑下，即使分析师存在有限理性甚至是诱导性行为，但不可避免的事实是大多数投资者仍会参考卖方分析师的公开研报，因此对卖方分析师研报进行情绪提取仍是具有研究价值和研究意义的。

## 2.3 文献评述

在系统梳理相关文献之后，可以发现文本挖掘技术逐渐成熟并在各大领域得到应用，无论是文本分类还是指标预测任务，文本挖掘的表现都不亚于传统定量分析表现。在投资者情绪度量方面，多数研究采用的文本数据为新闻和评论数据，基于此类文本进行价格预测的研究相对较多，而涉及到分析师情绪的研究多数停留在指标层面，如研报数量、分析师评级、分析师价格预测等，另外一些分析师情绪文献则比较关注分析师发布研报的独立性与可信度，很少有将分析师情绪这一指标真正用来做个股预测。尽管对于分析师研报可信度有不同意见，本文决定在相关理论支撑以及前人研究基础上进行挖掘，无论是“顺藤摸瓜”还是“将错就错”，本文选择利用文本挖掘从研报中提取分析师情绪信息，并以此构建相应的量化策略，从而补充学术界在该方面的研究。

# 第三章 分析师情绪指数构建

## 3.1 数据来源

本文以分析师研报作为研究对象，在数据来源的选取上，考虑到东方财富网站近年用户使用度提升，成为全国用户最为活跃，涵盖范围最广的权威网站，故本文的研报数据以东方财富研报页面提供的个股研报为分析对象，在数据期限长短方面，由于东方财富提供的期限最长的研报数据为两年内，所以限定了本文数据区间为两年。以平安银行为例，其研报页面如**图 3** 所示。

研报明细 | 银行研报 | 银行盈利预测　　时间:两年内

| 序号 | 报告名称 | 东财评级 | 评级变动 | 作者 | 机构 | 日期 |
|---|---|---|---|---|---|---|
| 1 | 平安银行年度业绩发布会点评：基本面反转的零售龙头 | 买入 | 维持 | 梁凤洁 邱冠华 | 浙商证券 | 2022-03-14 |
| 2 | 点评报告：盈利增速维持高位 | 增持 | 维持 | 郭懿 | 万联证券 | 2022-03-11 |
| 3 | 2021年年报详解：财富管理突破，资产质量向好 | 买入 | 维持 | 余金鑫 | 民生证券 | 2022-03-11 |
| 4 | 业绩持续高增，战略推进积极 | 增持 | 调低 | 林媛媛 林颖颖 | 中银证券 | 2022-03-11 |
| 5 | 零售突破战略有效推进，风险认定依然审慎 | 买入 | 维持 | 刘志平 李晴阳 | 华西证券 | 2022-03-10 |
| 6 | 不良包袱有序出清，零售高阶转型成长性凸显 | 买入 | 维持 | 郭其伟 刘斐然 范清林 | 天风证券 | 2022-03-10 |
| 7 | 营收增速环比提升，资产质量持续改善 | 增持 | 维持 | 刘丽 孙田田 | 山西证券 | 2022-03-10 |
| 8 | 业绩保持稳健高增，资产质量持续改善 | 买入 | 维持 | 林瑾璐 田馨宇 | 东兴证券 | 2022-03-10 |
| 9 | 2021年报点评：业绩增速稳中有升，资产质量持续改善 | 增持 | 维持 | 陈俊良 王剑 田维韦 | 国信证券 | 2022-03-10 |
| 10 | 归母净利润同比+25.6%，拨备覆盖率大幅增长 | 买入 | 维持 | 崔晓雁 | 华金证券 | 2022-03-10 |

**图 3 东方财富个股研报页面**

本文通过 python 第三方库 requests 和自动化测试库 selenium 来实现研报数据的爬取，具体思路如下：首先将问题拆分为针对个股数据的爬取，对于研报数量不超 50 篇，即无需翻页的的个股，利用正则表达式从响应数据中获取，这样可以提高抓取效率；对于研报数量较多的个股，使用自动化测试工具 selenium 模拟实现浏览器翻页操作，并利用 xpath 解释库定位元素所在节点，由于 selenium 模拟操作需要等待网页动态加载出的数据，所以在爬取过程中使用 sleep 操作避免访问过快导致程序退出。由于本文所研究的指数为中证 500，通过遍历中证 500 个股代码从而爬取所有个股的研报数据，再解析表单结构仅提取我们想要的数据，例如研报标题和日期是本文最关心的两个要素，至于为什么没有爬取研报标题所链接的二级页面，是因为作者观察到研报详情页面多为数值型数据，而在处理文本结构数据时我们一般不分析数值型数据，数值繁杂且在不同公司之间缺乏可对比性，包含的信息量太小，从分析师的角度出发，在为研报命

名时，无论是出于吸引读者的目的还是为券商覆盖公司扩大声望的目的，分析师的情绪大都已经体现在研报标题中，所以这里我们仅仅以研报标题为分析对象而不深入到研报具体内容中，500 只股票的研报爬取共计耗时半小时。

## 3.2 数据预处理

### 3.2.1 标题清洗与数据分布

本文所有数据分析工作均使用 python 完成。在爬取到中证 500 指数各成分股最近两年内的研报后，我们初步分析了研报数量情况。将文件读取到 jupyter 中进行分析，数据帧的格式如**图 4** 所示：

| 日期 | 600008 | 600021 | 600022 | 600026 | 600027 | 600037 | 600038 | 600039 | 600056 | 600060 | ... | 300376 | 300383 | 300418 | 300463 | 300474 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2020-04-29 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN | NaN | NaN |
| 2020-04-29 | NaN | NaN | NaN | NaN | 一季报超预期，看好火电龙头业绩高弹性 | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN | NaN | NaN |
| 2020-04-29 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN | NaN | NaN |
| 2020-04-29 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN | NaN | NaN |
| 2020-04-29 | NaN | NaN | NaN | NaN | 2020年一季报点评：成本改善显著，业绩增速亮眼 | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN | NaN | NaN |

**图 4 个股研报数据帧**

可以看到，由于某些标的会在同一天有多篇研报发布，所以行索引会出现重复日期，导致研报数量较少标的的单元格出现大量空值。这里我们做的初步处理是删除掉在这两年内没有发布研报的公司，经筛选后发现 500 家公司里面有 82 家在这两年内没有发布研报，所以最终我们的分析数据为 2019 年 10 月 7 日至 2021 年 9 月 30 日两年内的 418 家公司。经计算，数据集中一共有 12603 篇研报，在此过程中发现比较有意思的一点是不同标的的研报标题之间会出现重合，考虑到出现重合的研报标题所对应的标的之间必然存在高关联度，为了简化处理，这里我们进行去重，去重后数据集中还剩余 12537 篇研报，66 篇的重复研报标题对整体的情绪不会造成影响。接下来我们分析了研报数量的日度分布情况，如**图 5** 所示：

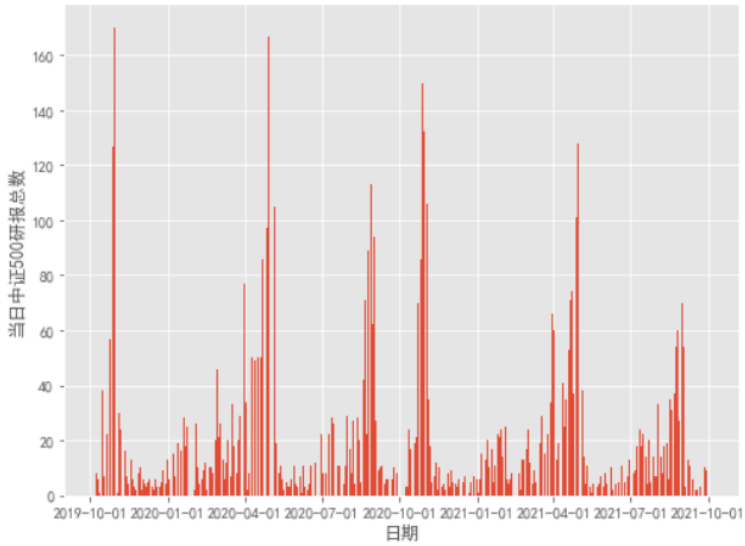



**图 5 中证 500 研报数量近两年分布**

可以发现研报数量在每年的 4 月、8 月和 10 月达到高峰，在该区间内机构发布研报数量最高时一天可达 170 篇，而大部分时间研报数量低于 20 篇，这主要是因为 4 月底是年报披露的截止时间，8 月和 10 月则是半年报和季报的披露时间，因此这一时间段内成为机构根据财务披露信息对公司出具研报的高峰期。

将数据划分为训练集和测试集，考虑到数据区间有限，同时又要保证模型在测试集的回测区间足够，因此取一年半作为训练集，剩余的半年时间作为测试集。

### 3.2.2 标题分词及词云展示

得到干净的样本内研报标题后，下一步操作是文本分词，本文使用当前国内比较流行的中文分词词库 Jieba 分词，示例如**表 1** 所示：

**表 1 Jieba 分词实示例**

| | |
|---|---|
| 原文本 | Q3 归母盈利同比增长 149%，水电铝产能增长强劲 |
| Jieba 分词 | 'Q3'，'归母'，'盈利'，'同比'，'增长'，'149%'，'，'，'水电'，'铝'，'产能'，'增长'，'强劲' |
| 加载用户词典后的分词效果 | '归母'，'盈利'，'同比'，'增长'，'水电铝'，'产能'，'增长'，'强劲' |

由于研报标题中经常会出现行业专有术语，因此本文首先浏览了研报标题，将其中的专业术语记录在用户词典中。在分词之前，使用 Jieba 的 load_userdict 属性将预先定义好的用户词典加载到 Jieba 词库中，而后对每个标题进行分词。由于研报标题中经常出现一些高频词汇，例如“简评”，“报告”，“公司”，“点评”等，而这些高频词汇

对于分析研报情感是毫无意义的，所以在分词之后对其做去除停用词处理；此外，研报标题中还会出现一些特殊标点符号，例如：“《”，“》”，“ II ”等也需要进行删除，这两步操作之后，我们最终会得到一个记录文档中所有词的词语列表。但是分词后发现列表中会出现一些单字词，而这些单字词总体上对标题情绪无法起到很好的区分作用，因此本文进行了去除单字词操作，目前为止列表一共有 63701 个词语，其中包含大量重复词语，实际上去重后词典中一共有 5905 个不同的词语。

接下来根根据列表中词语的频率绘制词云图，如**图 6** 所示：


**图 6 中证 500 全行业词云图**

可以发现“增长”一词在研报标题中出现的频率极高，其次是“持续”、“预期”、“盈利”、“提升”等词语。

接下来需要对研报标题打标签，根据研报对应标的在研报发布后的行情走势进行来表征研报情绪，考虑到市场存在时滞，如果以研报发布当天的股价走势作为研报情绪表现很可能存在偏误，因此本文在大量观察研报发布后股价走势之后发现 3-5 天是一个能够较好展现股价变动的时间段，同时考虑到某些机构发布研报的时间在非交易日，所以我们将研报发布后接下来三个交易日的标的涨跌幅作为判断研报情绪的指标。即如果某只标的在某一天有大量研报发布，而该标的在接下来三个交易日是涨的，我们则认为这些研报中的词语均为积极情感词。

## 3.3 话题寻优

为了从文本词语的情绪分层中更好的挖掘出信息，我们使用话题寻优模型来检测较为合适的话题个数。输入的情感词典为上一小节我们以标的涨跌幅作为判断指标所得出的词典，同时利用两向量之间的余弦相似度作为判断两个向量相似程度的指标。

以话题个数作为循环对象，提取 LDA 模型输出的主题词，构造主题词的词频向量，然后计算出每个话题个数情况下的平均余弦相似度，得到的研报 LDA 主题寻优结果如**图 7** 所示：

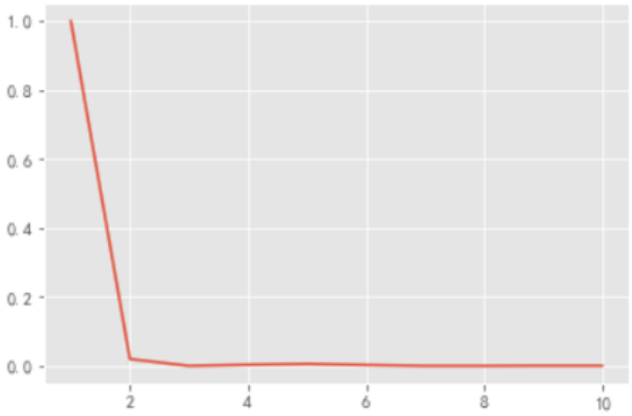



**图 7 中证 500 全行业研报 LDA 主题数寻优**

可以看出，当话题个数达到 3 的时候，文本向量之间的余弦相似度已经接近 0，此时对文本的区分效果比较好，因此下面我们以三话题向量作为文本情感话题个数进行分析。

## 3.4 SESTM 模型建立

### 3.4.1 情绪敏感词的筛选

本文第二章已经介绍了提取情绪敏感词的方法，我们的目的是找出词典中最能代表研报情绪的积极词和消极词，对于包含较多噪音的情感中性词我们则不考虑对其进行建模，这也与上一小节我们分析所得到的三话题个数相对应，因此在本节中主要任务就是分离出情绪敏感积极词和消极词。

通过遍历词典中的每一个词语，计算出训练集中包含该词语的研报数，包含该词语且标的收益为正的研报数目，从而得到该词语与正收益共现的频率。结果如**表 2** 所示：

**表 2 词语正收益共现频率**

| 词语 J | 包含 J 的文章 | 包含 J 且收益为正 | J 和正收益共现频率 |
|---|---|---|---|
| 性价比 | 9 | 6 | 0.666667 |
| 尖端 | 1 | 1 | 1.000000 |
| … | … | … | … |
| 粤东 | 11 | 7 | 0.636364 |
| 小荷 | 1 | 0 | 0.000000 |

接下来我们需要对包含词语的研报总数寻找一个合适的阈值，以避免出现频率过低的词语带来的干扰，经反复实验，本文决定使用 30 作为阈值，即词语在研报标题中出现的次数不能低于 30 次，第一步筛选后，训练集词典中还剩下 440 个词语。接下来通过循环共现频率从而得到积极词和消极词各 200 个的目标词典，之所以要设定 400 个词语，是因为如果词典中词语太少，会导致测试集中研报标题出现大量的稀疏矩阵，特别是会出现大量零向量，这不利于模型的情感方向预测。将经初步筛选后的词汇按照与正收益共现频率排序，取出共现频率最高的前 200 个词语作为积极情绪敏感词，共现频率最低的后 200 个词语作为消极情绪敏感词。需要注意的是，即使是消极情绪敏感词，其中也存在一部分与正收益共现频率超过 0.5 的词语，这是由于研报数据自身的乐观倾向所导致的，这种乐观倾向会自动推升情绪分界点。

目前为止我们得到了积极情绪敏感词和消极情绪敏感词各 200 个，**表 3** 列出了各自前 10 个词语：

**表 3 中证 500 全行业前十情绪敏感词**

| 情绪倾向 | 词语集合 |
|---|---|
| 积极敏感词 | 自主、光伏、好转、安全、新基建、市占率、激励、成效、开拓、扩产 |
| 消极敏感词 | 银行、供应、低估值、研发、拖累、兑现、新能源、质量、推荐、价格 |

观察上表，我们不禁会产生疑问，通过这种方法区分开的积极词和消极词似乎并没有什么区别，即使是消极敏感词在平时看来更偏向中性，并没有明显的消极特征。关于这一点，我们应时刻记得研报标题自身的乐观倾向，可能对肉眼而言看不出差距，但是统计得到的结果发现这是区分度比较好的词语，有理由在此基础上进一步分析。

### 3.4.2 拟合两话题向量

这一小节中我们对每篇研报的情绪矩阵进行估计，估计方法如第二章所述，采用研报对应标的涨跌幅的标准排序作为估计研报情绪的指标，计算结果如**表 4** 所示：

**表 4 中证 500 研报收益标准排序**

| 标题 | senti | 股票收益 | 收益排序 | 情绪 | 1-情绪 |
|---|---|---|---|---|---|
| 水务龙头业绩优异，十四五发力水环境+固废全产业链 | 0 | -0.009646 | 3319.0 | 0.362613 | 0.637387 |
| 三季度业绩大增，水固全产业链协 | 1 | 0.027875 | 6463.0 | 0.706107 | 0.293893 |

**续表 4　中证 500 研报收益标准排序**

| 标题 | senti | 股票收益 | 收益排序 | 情绪 | 1-情绪 |
|---|---|---|---|---|---|
| 同发展<br>成本改善助盈利回升，清洁能源增厚业绩 | 0 | -0.013908 | 2958.0 | 0.323173 | 0.676827 |
| 2019 年年报点评：清洁能源持续加码，核心业务盈利可观 | 1 | 0.029455 | 6580.0 | 0.718890 | 0.281110 |
| … | … | … | … | … | … |
| 受疫情影响上半年业绩略有下降，关注 HNB 落地机遇 | 0 | -0.037867 | 1407.5 | -0.153775 | 0.846225 |
| 业绩受下游需求变化小幅降低，壁垒优势助力中长期稳健增长 | 1 | 0.020693 | 5969.0 | 0.652136 | 0.347864 |

在训练集的 9153 篇研报中，一共有 4870 篇研报对应标的在随后三个交易日内涨跌幅为正，即 53.21%的比例，我门可以发现，研报发布后的股票涨跌幅正负比例接近，但是多出来的 3.21%的正收益比例能否由研报情绪来解释则还需要进一步的分析。至此我们得到了回归方程 2.13 中的情绪矩阵 W，下一步是获取训练集中所有研报的词频矩阵。

词频矩阵构建的词典基础是上面所得到的 400 个词语，我们将以这 400 个情绪敏感词语来构造每一篇研报标题的词频向量，方法就是通过循环遍历每一个情绪词，得到词典中每个词语在一篇研报标题中的词频，最终可以得到整个语料库的词频矩阵 D，词频数据帧的行索引是情绪敏感词，列索引是每篇研报的标题。

至此，我们得到了语料库的情绪矩阵 W 和词频矩阵 D，便可以通过公式（2.13）计算出两话题向量 O，结果如**表 5** 所示：

**表 5　中证 500 全行业两话题向量**

| | pos | neg |
|---|---|---|
| 电解液 | 0.010396 | 0.002888 |
| 石化 | 0.005413 | 0.002775 |
| 转债 | 0.005727 | 0.002021 |
| 内销 | 0.006944 | 0.000588 |
| 成效 | 0.010101 | 0.001858 |
| … | … | … |

**续表 5　中证 500 全行业两话题向量**

| | pos | neg |
|---|---|---|
| 资源 | 0.002549 | 0.008277 |
| 上市 | 0.002278 | 0.014065 |
| 加强 | 0.000936 | 0.006794 |
| 外销 | 0.000430 | 0.005336 |
| 优异 | 0.001309 | 0.007084 |

### 3.4.3 预测研报情绪

通过训练集得到话题向量 O 后，接下来便可以运用公式预测新研报标题所蕴含的情绪值。以一篇样本外的研报标题作为示例， 该研报标题为“年报业绩符合预期，新冠检测有望持续放量，新平台增量可期“。首先计算情绪敏感词在该研报标题中的词频向量，结合标题中出现的词语在话题向量中的数值大小，利用极大似然估计方法可以得到研报标题的情绪分值。其词频与话题向量对应数值如**表 6** 所示：

**表 6　研报标题词频与对应话题向量示例**

| | fre | pos | neg |
|---|---|---|---|
| 放量 | 1 | 0.029184 | 0.021257 |
| 可期 | 1 | 0.044906 | 0.037382 |
| 新冠 | 1 | 0.007029 | 0.004476 |
| 增量 | 1 | 0.005869 | 0.006294 |
| 平台 | 1 | 0.010621 | 0.010611 |
| 持续 | 1 | 0.129965 | 0.089799 |
| 符合 | 1 | 0.065978 | 0.068265 |
| 有望 | 1 | 0.065978 | 0.068265 |

对于惩罚项的选择，Zheng Tracy Ke, et(2019)的文章中给出了 3 个值，分别是 1、5、10，随着惩罚系数的增大，得到的研报情绪值越接近 0.5，如**图 8** 所示：

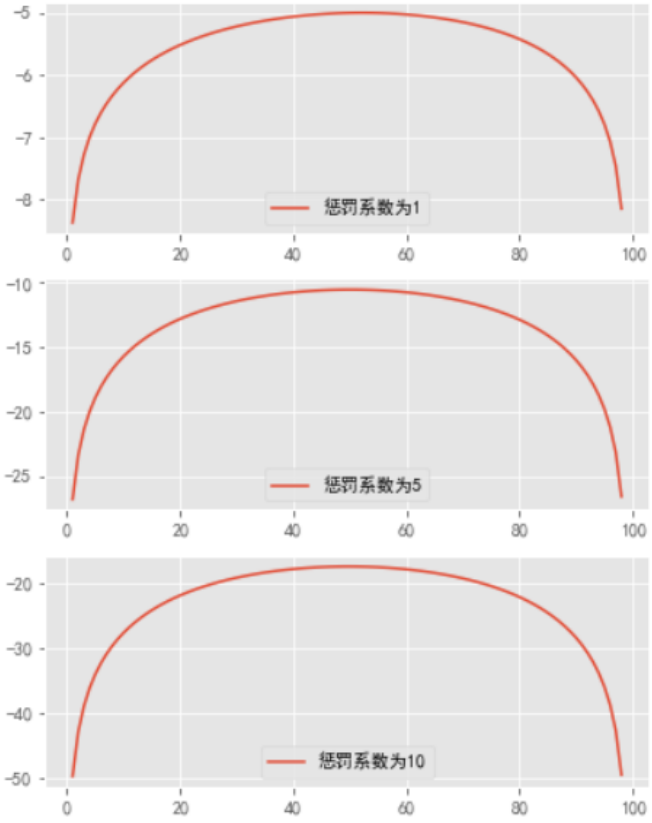



**图 8 不同惩罚系数下研报预测情绪值**

横坐标是将情绪值 0-1 分成了 100 份，我们可以看到，随着惩罚系数的增大，曲线的最大值越来越靠近中间，即研报标题情绪越来越接近 0.5。本文考虑到为了使情绪区间的范围易于区分，决定使用最小的惩罚系数 1。

另外需要说明的一点是，本文所使用的东方财富个股研报并没有关于研报发布详细时间的说明，爬取下来的时间均为零时零分，我们无法得知研报是在当日收盘前还是收盘后发布的，出于谨慎性考虑，本文在计算持仓收益时，均采用研报发布随后交易日的收益作为开仓收益，而不包括研报发布当日个股的收益状况。

### 3.4.4 多空组合下研报情绪策略净值分析

本小节对测试集中所有研报标题进行情感分析，将情感值大于某一阈值作为买入开仓条件，小于某一阈值作为卖出开仓条件，不考虑手续费、滑点等现实交易因素，分析净值走势。随机选择 3 天作为持仓周期，然后检验在 1 倍惩罚系数不同情绪阈值的净值走势，如**图 9** 所示：


**图 9 1 倍惩罚系数下持仓 3 天的不同情绪阈值净值走势**

从图中可以看出在 0.01 的情绪颗粒度下，情绪阈值较低时净值走势比较高，当情绪阈值超过 0.5 时，净值一路下跌。直观上看，随着情绪阈值变大，买入标的数量递减，而开空标的数量递增，导致净值下跌，说明情绪阈值增大导致策略开错方向次数增加。同时我们发现自九月初开始，持仓三天的情况下 0.48 的情绪阈值开始大幅下跌，而 0.49 和 0.5 的情绪阈值表现尚可，因此我们要进一步研究不同持仓周期产生的影响。

考虑到市场对研报的解读可能存在一定的时滞，但互联网时代下研报的更新速度又比较快，所以对于持仓周期的设置不能太长，以避免在持仓周期内其他研报带来的交叉影响，本文将持仓周期设置为 1-7 天进行研究，**图 10** **图 11** 和**图 12** 分别展示了不同情绪阈值下不同持仓周期的策略净值走势：

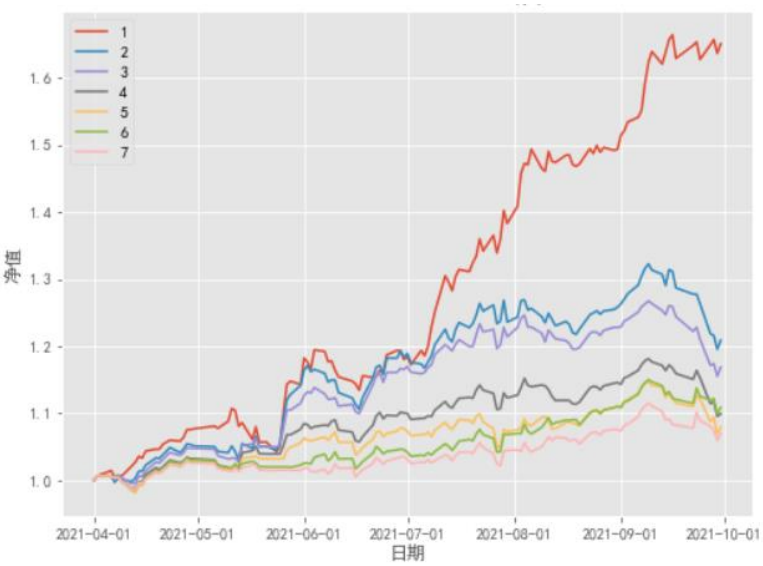



**图 10 0.48 情绪阈值下不同持仓周期净值走势**


**图 11 0.49 情绪阈值下不同持仓周期净值走势**


**图 12 0.50 情绪阈值下不同持仓周期净值走势**

从上图我们可以看出在 0.48 和 0.49 情绪阈值下，净值曲线都是随着持仓周期的增加而逐渐平缓，而且净值之间的分层效果比较明显，但当情绪阈值达到 0.5 时，净值走势变得杂乱无章；另一方面我们会观察到在 0.48 情绪阈值下，除一天持仓周期之外的其他净值曲线从 9 月开始便集体下行，而 0.49 情绪阈值则表现较为稳定；同时我们会发现无论是 0.48 还是 0.49 的情绪阈值，一天持仓周期的净值走势与其他持仓周期之间有较大差异，且收益更高，本文认为这是研报市场反应短暂的表现，也即研报包含的信息能够在较快时间内被市场消化并反应在股价中，因此本文综合考虑选择使用一天的持仓周期和 0.48 的情绪阈值进行分析。

业界在分析股票类策略表现时，会比较关注模型相对于基准的表现，即使模型在

一段时间收益为负，但如果基准在相应时间区间内下跌幅度更大，仍然认为模型取得了“正”收益，真正体现出模型好坏的应该是模型的超额走势。**图 13** 为中证 500 指数在测试区间内的走势：

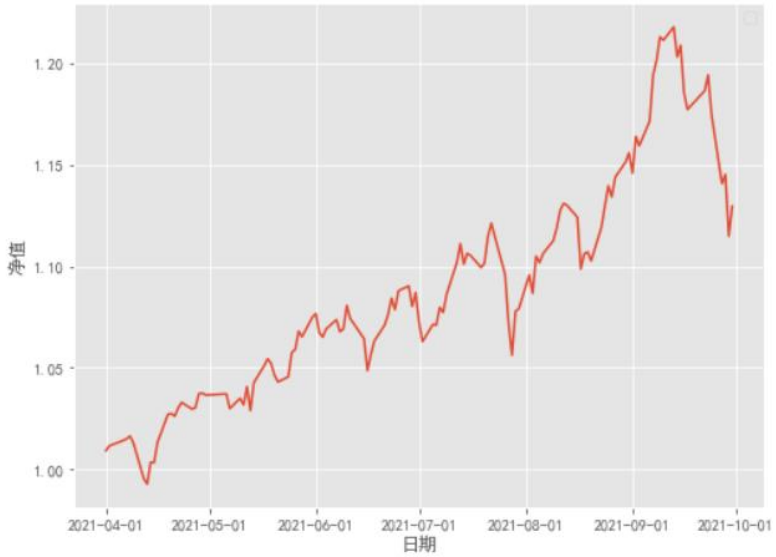



**图 13 中证 500 在回测区间内走势**

从上图可以看出尽管中证 500 指数在 2021 年 7 月下旬以及同年 9 月下旬发生了较大回撤，但在对应时间区间内指数最终获得了 12.74%的累计收益。为了更好的检验模型超越基准指数之外的收益能力，我们将模型每日的加权收益减去中证 500 指数当天的收益从而计算模型当天的超额收益水平，并拟合为超额净值曲线，**图 14** 展示了 0.48 情绪阈值下模型不同持仓周期的超额走势，并在**表 7** 列出了不同持仓周期下超额收益的绩效表现：

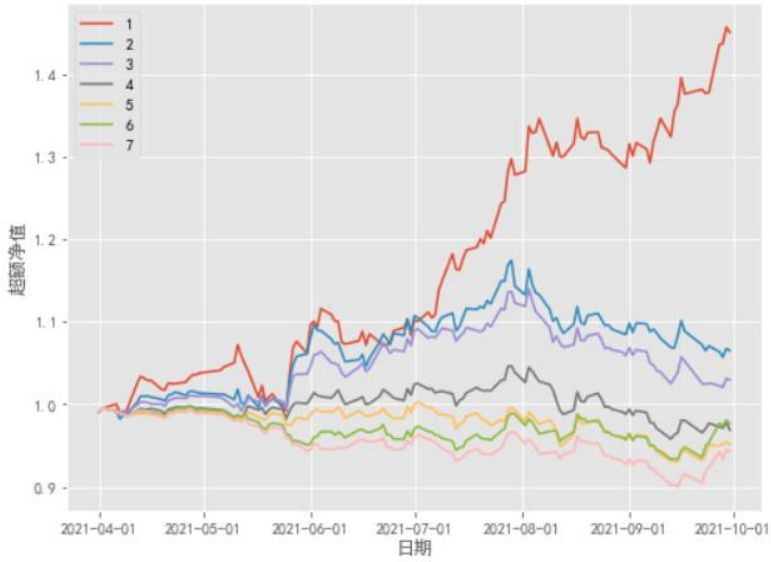



**图 14 0.48 情绪阈值下不同持仓周期的超额净值走势**

**表 7 0.48 情绪阈值下不同持仓周期的超额绩效表现**

| 持仓周期 | 区间超额收益% | 最大回撤% | 最大回撤发生时间 | 夏普比率（不减无风险收益） | 卡玛比率 |
|---|---|---|---|---|---|
| 1 天 | 46.53% | -7.45% | 2021-05-25 | 4.19 | 6.24 |
| 2 天 | 7.53% | -10.01% | 2021-09-28 | 0.93 | 0.75 |
| 3 天 | 3.94% | -10.48% | 2021-09-28 | 0.59 | 0.38 |

从上图我们发现模型在五月底的超额出现最大回撤，此时模型收益和基准收益保持一致，而此后的超额表现保持稳定增长。最为明显的是从九月上旬开始市场风格切换，中证 500 指数一路下跌，但是模型并没有太大的波动，超额反而是向上走的，这种规避指数下跌的属性也是衡量模型表现的关键因素。当模型在五月出现最大回撤时，我们观察到中证 500 指数是向上走的，这里我们给出的解释是由于模型是基于研报发布才会进行预测开仓，但研报分布显示五月份研报数量较少，导致模型开仓频率较低，同时由于基准具有一定的涨幅综合导致超额出现大幅回撤。另外从绩效表现表中我们会发现随着持仓周期增加，区间超额收益呈现快速下降趋势，回撤逐渐加大，夏普和卡玛比率均快速减小，反映了研报策略短暂的市场有效性。

### 3.4.5 纯多头研报情绪策略净值分析

本小节将按照纯多头组合构建策略，毕竟中国的市场环境不支持空头交易，同样以 0.48 为情绪阈值，其净值走势与超额走势分别如**图 15** 和**图 16** 所示：

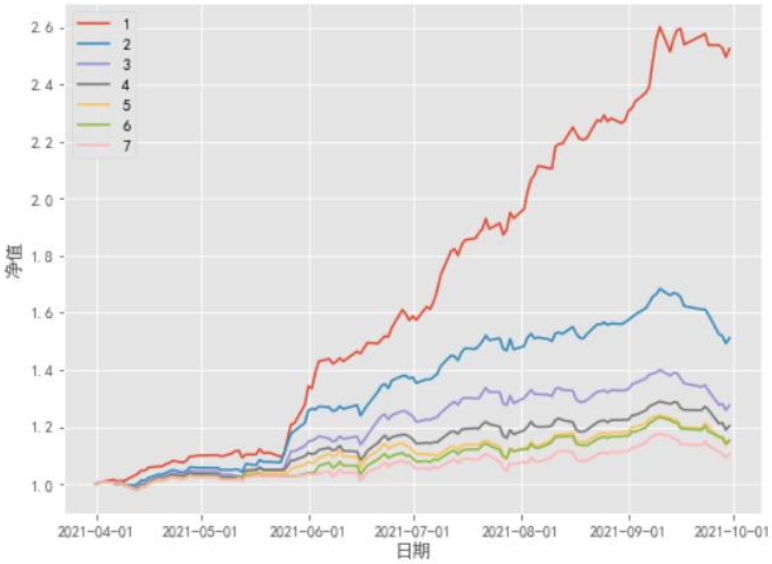



**图 15 纯多头在 0.48 情绪阈值下不同持仓周期净值表现**

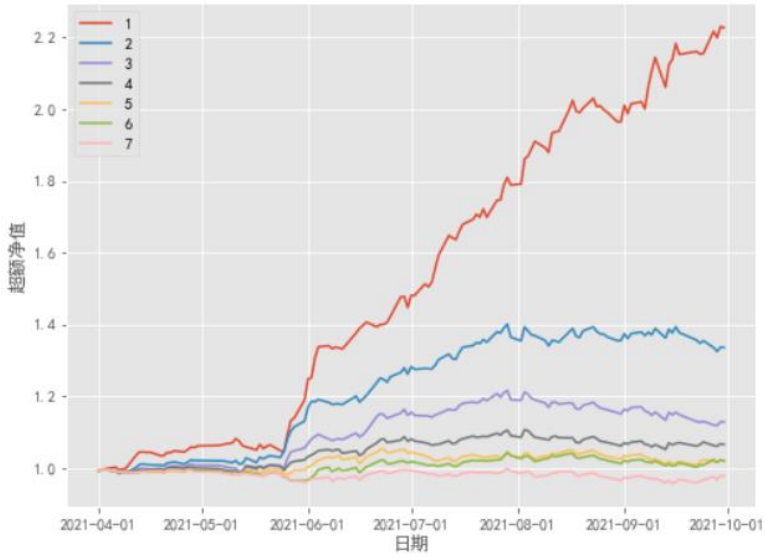



**图 16 纯多头在 0.48 情绪阈值下不同持仓周期超额净值表现**

在 0.48 情绪阈值纯多头下，我们惊奇的发现纯多头表现要远远优于多空组合，而且纯多头持仓一天的表现要远远超过其他持仓周期的表现。为了更好的比较一天的持仓周期下多空组合和纯多头组合的超额表现，我们将其绘入一张图中，如**图 17** 所示：

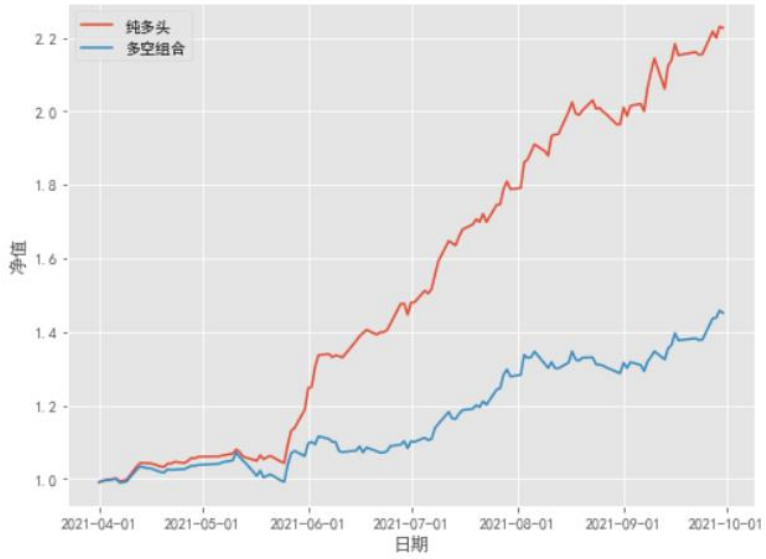



**图 17 0.48 情绪阈值下持仓一天的纯多头和多空组合超额净值对比**

上图较为直观的反应出纯多头和多空组合在该时间区间内的走势差异，可以看出，在 4 月初至 5 月下旬之前，同时做多做空标的的策略和纯做多策略的收益没有太大差异， 但从 5 月下旬起，纯多头组合净值迅速上涨，远超多空组合，而且 8 月份的回撤以及修复时间都明显小于多空组合，可见做空情绪分值低的标的的行为对整个组合的收益产生了负贡献度。为进一步量化二者差异，我们在**表 8** 种列出了纯多头和多空组合各自的绩效表现：

**表 8 0.48 情绪阈值下持仓一天的纯多头和多空组合的超额绩效分析**

| 策略 | 区间超额收益% | 最大回撤% | 最大回撤发生时间 | 夏普比率（不减无风险收益） | 卡玛比率 |
|---|---|---|---|---|---|
| 多空组合 | 46.53% | -7.45% | 2021-05-25 | 4.19 | 6.24 |
| 纯多头 | 124.78% | -3.88% | 2021-09-13 | 10.30 | 32.16 |

我们可以发现无论是在抓涨能力还是回撤控制方面纯多头的表现都要优于多空组合，也就是说，在中证500范围内的模型确实具备一定的选股能力，但多空开仓效果还不如纯多头开仓，空头开仓反而拉低了总体收益，即对于研报此类具有乐观倾向型数据，模型对积极词语的识别能力要超过对于消极词语的识别能力。对样本外研报数据进行预测时，模型可以准确识别出强烈的积极信号，但面对偏中性甚至带有一点乐观倾向的研报标题词汇时，模型的识别和预测能力变差，以至于无法识别出开空仓的标的情感，而这些标的的涨跌具有不确定性，导致空头开仓表现不尽人意。

## 3.5 本章小结

本章详细说明了模型的数据来源以及文本数据的预处理流程，并运用 LDA 主题模型判断合适的情绪区分度，在此基础上运用 SESTM 模型预测研报标题的情绪走向，利用情绪预测结果构建净值曲线，分析了不同情绪阈值和不同持仓周期组合下的策略绩效表现，为了剔除基准影响，以策略的超额表现来评价模型，另外对比分析了多空组合和纯多头的绩效表现，本章得出的结论有两点：（1）研报策略的市场持续性较短，在一天的持仓周期下表现最好；（2）模型对于研报消极情绪不具识别能力，导致纯多头表现远远优于多空组合。

# 第四章 分行业检验分析师情绪指数

上一章我们采用指数所有成分股研报的话题向量来构建分析师情绪指数，但是不得不承认的是不同行业之间研报标题的差异比较明显，这样得到的情绪敏感词典很可能会忽略掉行业因素，因此本章我们将按照不同行业分别构建行业词典，拟合行业话题向量，最终合成一条新的净值曲线。

## 4.1 行业筛选

首先本文统计了 500 标的成分股排除未发布研报的 82 家公司后剩余的 418 家企业的行业分布情况，综合考虑了行业内公司数量以及该行业研报数量，使用每一行业的研报总数除以该行业的公司总数作为衡量该行业公司研报情况的指标，最终决定挑选出行业公司总数超过 20 家并且平均研报数量超过 20 篇的行业作为研究对象，最终符合条件的一共有五个行业，分别是“化工”、“医药生物”、“电子”、“计算机”和“传媒”，每个行业的研报与公司数量如**表 9** 所示：

**表 9　经筛选出的五个行业的研报与公司数量**

| 行业 | 公司数量 | 研报总数 | 平均研报数量 |
|---|---|---|---|
| 化工 I | 30 | 1615 | 53.833333 |
| 电子 I | 31 | 990 | 32.225806 |
| 医药生物 I | 38 | 1213 | 31.921053 |
| 计算机 I | 23 | 696 | 30.260870 |
| 传媒 I | 24 | 520 | 21.666667 |

## 4.2 SESTM 模型分行业表现

与全行业分析的思路一样，针对每一行业，我们首先做情感话题个数的检验，实验结果表明每一行业的主题寻优数在主题数达到 3 时区分度最高，所以均可采用 SESTM 模型进行建模。

在构建中证 500 全行业模型时构建的积极消极词典中一共包含了 400 个词语，但对于单行业词典的构建，过多的情感词汇对于预测样本外研报标题情绪时会带来更多的噪音，所以本文在构建单行业词典时选择积极消极词语各 30 个。以电子行业为例，统计电子行业情绪敏感词词频并画出词云图，如**图 18** 所示：


**图 18 电子行业研报标题词云图**

我们会发现电子行业的词云图和全行业的词云图有一些相似性，在一些高频词汇上几乎没有明显的差别，比如“增长”，“预期”，“符合”，“盈利”和“持续”等词语，同时也有一些电子领域专业词语，比如“芯片”，“电感”和“光学”等词。这对于我们的分行业预测是一个好现象，因为对于情感预测而言，高频的情感倾向词语通常具有共性，而唯一的区别点就在于不同行业所具备的专业词汇可能有不同的情感倾向，但影响程度如何还要具体分析。

通过计算电子行业内研报的情感向量和词频矩阵，我们得到了该行业的两话题向量，部分数据如**表 10** 所示：

**表 10 电子行业两话题向量**

| | pos | neg |
|---|---|---|
| 影响 | 0.046110 | 0.024234 |
| 放量 | 0.047365 | 0.036303 |
| 助力 | 0.047832 | 0.055807 |
| 空间 | 0.020869 | 0.055878 |
| 覆盖 | 0.021622 | 0.035166 |
| 加码 | 0.008100 | 0.045225 |
| 替代 | 0.037509 | 0.066026 |
| 国内 | 0.048097 | 0.022268 |
| 订单 | 0.046904 | 0.006810 |
| 市场 | 0.032378 | 0.044583 |

为了与全行业模型结果更具对比性，在分行业模型的构建时我们均采用和全行业模型一致的参数，即一倍的预测惩罚系数，0.48 的情绪阈值和一天的持仓周期。为了直观对比，这里我们仅给出五个行业各自的纯多头不同持仓周期净值表现，五个行业

各自的纯多头与超额绩效对比分析见附录。

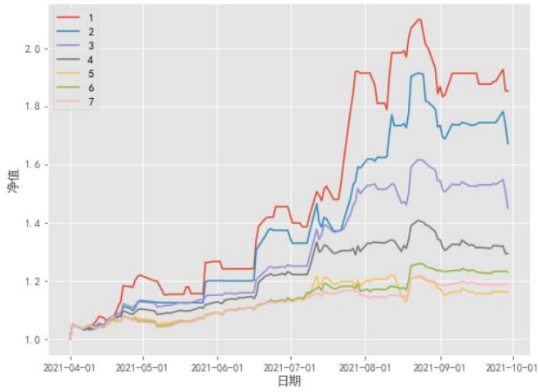



**图 19 电子行业纯多头不同持仓周期净值走势**

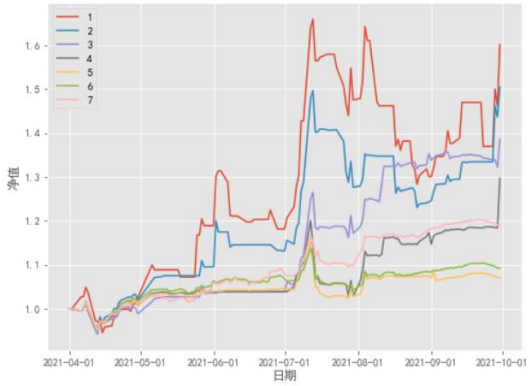



**图 20 化工行业纯多头不同持仓周期净值走势**

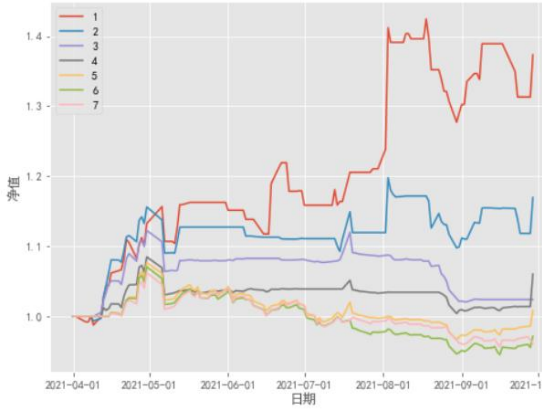



**图 21 电子行业纯多头不同持仓周期净值走势**

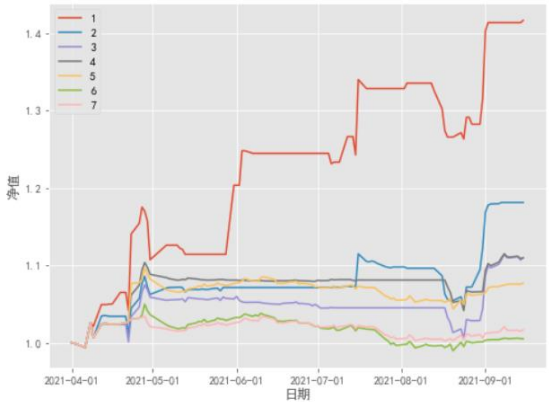



**图 22 化工行业纯多头不同持仓周期净值走势**

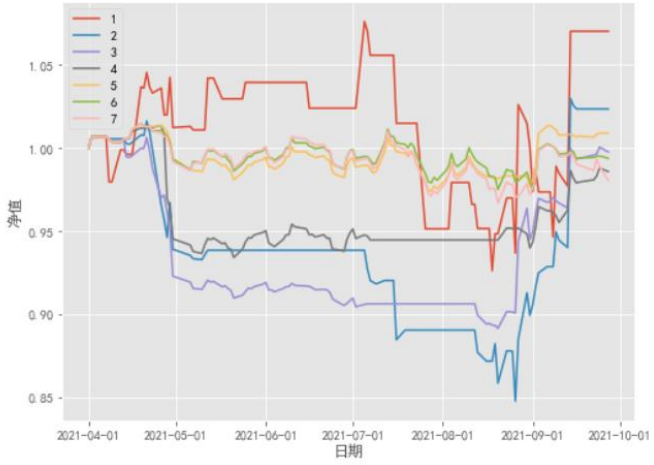



**图 23 计算机行业纯多头不同持仓周期净值走势**

通过观察这五张图我们会发现，电子、化工、医药生物和传媒四个行业的纯多头

策略在持仓一天时是表现最好的，总体上的走势还算稳定，但是计算机行业的走势却不尽人意，为了分析计算机行业走势的原因，本文构建了中证500指数内计算机行业股票收益的等权指数，构建方法为：获得计算机行业个股每日的对数收益，在时间截面上对所有个股的收益等权求均值，获得的净值曲线如**图 24** 所示：

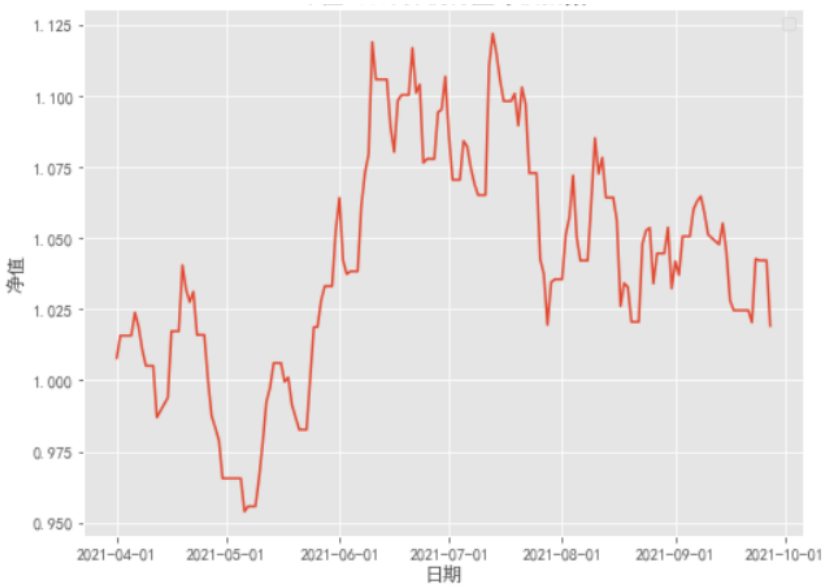



**图 24 中证 500 计算机行业等权指数**

通过观察计算机行业等权指数我们发现在样本外的时间区间内，计算机行业最终取得 2%的累计收益，但是在五月初至六月上旬等权指数上涨将近 15 个百分点，此时模型并没有什么反应，因此我们又画出了计算机行业在样本外区间的研报日度分布情况，如**图 25** 所示：

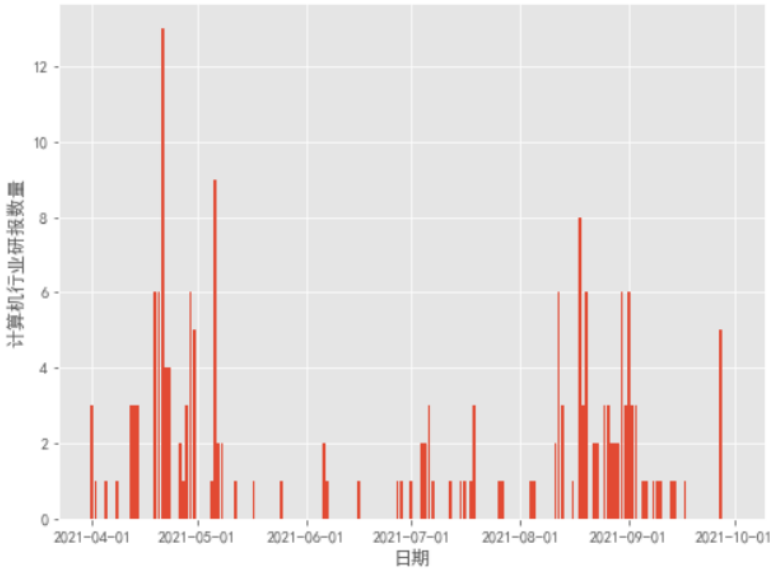



**图 25 中证 500 计算机行业样本外研报分布情况**

我们会发现在五月初至六月上旬计算机行业大幅上涨之时，市场上几乎没有关于该行业的研报发布，模型自然无法开平仓，这就造成等权指数大幅上涨之时，模型表现平平，而自七月中旬至八月底等权指数一路下跌，此时模型的开仓又造成了净值的下跌，综合导致计算机行业的模型表现较差，至此我们解释了分行业模型中计算机行业模型失效的原因。

## 4.3 分行业模型与“全”行业基准模型对比

为了将分行业模型与全行业模型做对比，下面我们将每一行业各自模拟出来的收益率求均值，但由于分行业模型我们仅挑选了五个行业，而全行业模型则包揽了中证500 所有的行业，如果直接对比将会产生较大偏差。分行业与全行业进行对比的本质是区分每个行业独自构建话题向量与全行业仅使用一个话题向量之间的差异，因此这里我们使用五个行业所有的研报数据共同构建一个话题向量进行预测，得出的模型便很好的起到了同全行业模型一致的基准对比效果，将分行业模型汇总收益和基准进行对比，其净值表现效果**图 26** 如所示：


**图 26 分行业模型与“全”行业基准净值对比**

从上图可以看出分行业模型净值波动要比自建的全行业模型波动大，但同时也取得了更高收益。在该回测期间内，分行业模型取得将近 46%的收益，而全行业模型则取得 28%的收益。单从净值走势上看，分行业模型期间内有四次大的上涨，然而全行业模型并不能抓住每一次上涨的机会，此外上涨幅度相对分行业模型而言较小，我们于**表 11** 列出了分行业模型与基准的绩效表现。

**表 11 分行业模型与“全”行业基准绩效分析**

| 策略 | 区间累计收益% | 最大回撤% | 最大回撤发生时间 | 夏普比率（不减无风险收益） | 卡玛比率 |
|---|---|---|---|---|---|
| 分行业模型 | 45.97% | -6.01% | 2021-08-31 | 6.05 | 7.65 |
| 五个行业构建的全行业基准模型 | 28.51% | -1.99% | 2021-08-10 | 7.01 | 14.29 |

可以发现，分行业模型在在样本外区间内取得较高的收益，但同时回撤也相对较大，尽管五个行业构建的基准模型仅取得了 28.51%的区间收益，但回撤只有 1.99%，从夏普和卡玛比率两方面来看，全行业基准模型要优于分行业模型。如果投资者具备一定的回撤忍受度，分行业模型则更好一些。

## 4.4 本章小结

本章我们筛选出五个行业并分别构建了各个行业的话题向量进行预测，画出每个行业各自的纯多头策略净值走势，并针对模型表现效果较差的计算机行业给出合理解释。计算机行业的模型表现提醒我们，避免将单一行业模型结果用于投资，模型对于研报数据的依赖性是模型自身存在的一大缺陷。

通过自行构建与分行业模型具有可比性的五行业基准模型，与分行业模型对比之后发现分行业模型在收益和回撤上要明显高于基准模型，具有更高的进攻性，但从夏普和卡玛比率来看，本文构建的基准模型走势更加稳定，这和中证 500 全行业模型走势稳定相符合。

# 第五章 模型稳健性检验与策略构建

在第三章和第四章中我们分别对中证500全市场模型和分行业模型做了分析，并得出了分行业构建话题模型进行预测的表现要好于全行业仅构建单个话题模型的结论。为了检验模型是否稳健，本章将选用不同的数据源以及不同的情感分析模型分别进行测试，在模型稳健的基础上构建相应的分析师情绪指数量化策略。

## 5.1 中证1000研报数据的SESTM模型表现

我们使用中证1000指数在过去两年内发布的研报数据作为研究对象，并构建1000指数的全市场话题模型从而检验模型对新数据的适应能力。同样使用一倍的惩罚系数，积极消极情感词各选取200个，模型在持仓周期固定的条件下不同情绪阈值净值表现如**图27**所示：

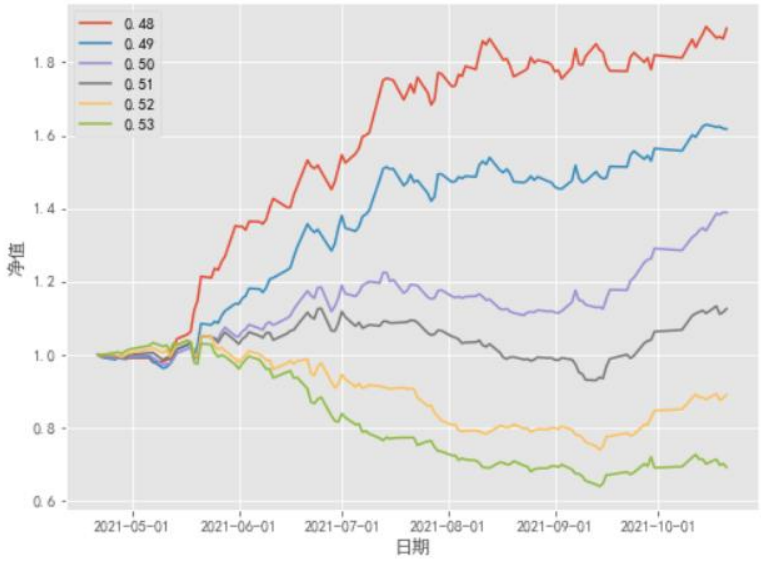



**图27 中证1000研报数据在持仓一天时不同情绪阈值净值走势**

可以看出持仓一天的不同情绪阈值净值走势分层明显，且 0.48 情绪阈值表现最好，在 0.48 情绪阈值下多空组合和纯多头各自在不同持仓周期下净值以及超额收益净值如**图28**所示：

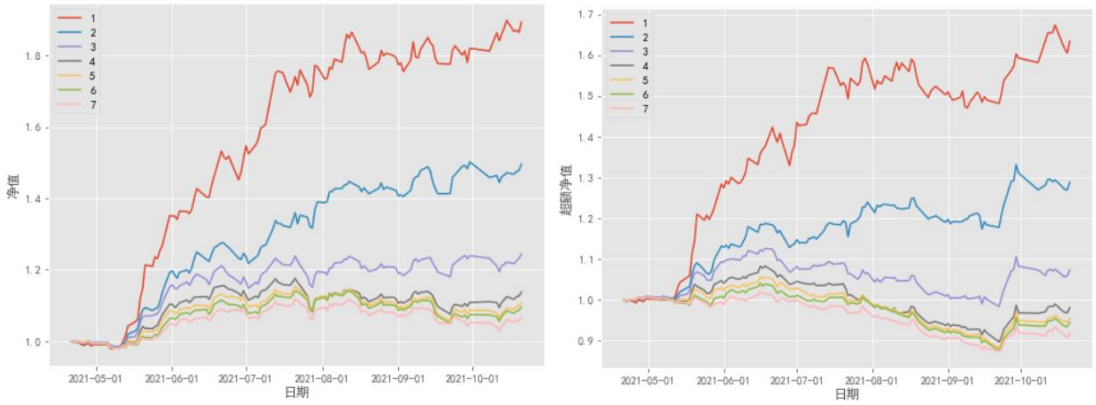



**图 28 0.48 情绪阈值下多空组合不同持仓周期净值和超额收益净值**

多空组合同样在持仓一天的情况下表现最好，且表现明显优于更长持仓周期的净值；接下来对纯多头在 0.48 情绪阈值下不同持仓周期的净值及超额收益净值进行检验，结果如**图 29** 所示。

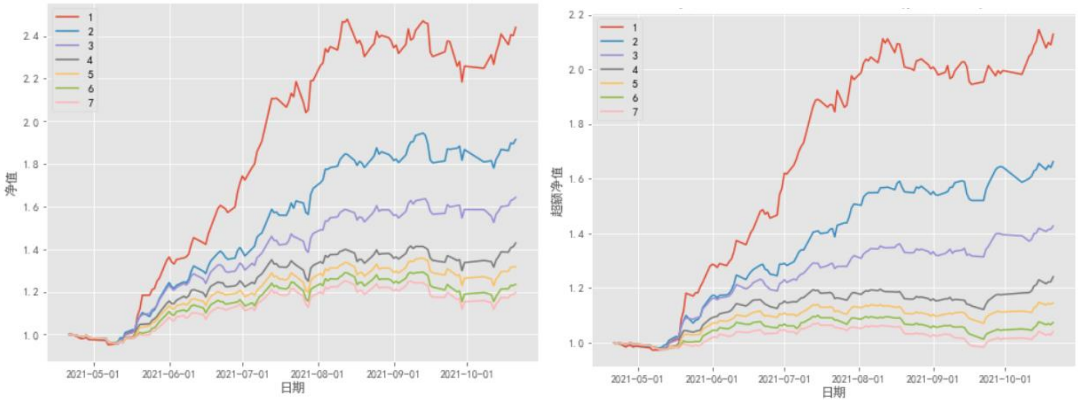



**图 29 0.48 情绪阈值下纯多头不同持仓周期净值和超额收益净值**

上图可以看出纯多头策略收益远超多空组合，**表 12** 中列出了纯多头和多空组合的绩效表现。

**表 12 0.48 情绪阈值以及持仓一天时纯多头和多空组合绩效表现**

| 策略（超额收益） | 区间累计收益% | 最大回撤% | 最大回撤发生时间 | 夏普比率（不减无风险收益） | 卡玛比率 |
|---|---|---|---|---|---|
| 多空组合 | 62.4% | -7.66% | 2021-09-09 | 3.97 | 6.64 |
| 纯多头 | 114% | -7.91% | 2021-09-17 | 5.69 | 9.82 |

值得注意的是本文所使用的夏普比率均为不减去无风险收益的方式，因此可以充当 t 统计量的角色，夏普比率越高，模型的收益越显著。

通过观察净值走势我们可以明显的发现，模型相对于中证 1000 指数有明显的超额收益，而且纯多头的表现要优于多空组合，与以中证 500 研报作为数据源拟合出的模型

所得出的结论保持一致。

## 5.2 中证 500 研报数据的词典模型表现

下面我们尝试使用舆情分析传统的词典法来构建策略，并检验 SESTM 模型相对于传统词典法的表现优异程度。

本文所使用的情感词典来自于姜富伟等(2021)[61]所构建的中文金融情感词典，该词典综合了众多情感词典，不仅包含了英文 LM 金融词典翻译过来的中文词汇，同时对国内应用程度较为广泛的词典如知网 Hownet 情感词典、台湾大学 NTUSD 简体中文情感词典和清华大学李军中文褒贬义词典等进行合并去重，作为通用情感词典。此外，为了扩充金融情感词典，他们采用了 word2vec 算法寻找关联度较高且具有明显情感倾向的词语，最终将三种方法得到的词语合并去重从而得到本文所使用的中文金融词典，经扩充后的词典中积极词语有 3338 个，消极词语有 5890 个。

由于分析对象是研报标题，而分析师在命名标题时几乎不会掺杂消极词汇，所以在利用词典法计算研报情绪值时，可以忽略标题中出现否定词对预测造成的偏差。本文将积极词权重设为 1，消极词权重设为-1，从而得到一篇研报标题的情绪：

$$SENTIMENT = \frac{\sum_{i=1}^{N} W_i}{N} \tag{5.1}$$

$$W_i = \begin{cases} 1, & positive \\ -1, & negative \end{cases}$$

公式中的 Wi 代表情感词的权重，N 代表一篇研报的标题中词语出现在情感词典中的次数，SENTIMENT 代表一篇研报标题的情绪分值。

研报数据的预处理操作同上，唯一不同的是预测研报情绪分值的方法，运用词典法构建策略时，仍然区分多空组合和纯多头，得到每日发布研报数据的公司的情绪分值之后，买入情绪分值在前 50%的股票。多空组合的处理则稍有不同，在 SESTM 模型中，我们买入情绪阈值之上的股票卖出情绪阈值之下的股票，因为情绪阈值在其中起到分界作用，但是在词典法中得到的情绪分值却不具备区分积极消极词性的作用，因此不能在买入情绪分值前 50%的股票的同时卖出剩下 50%的股票，因为我们无法得知这后 50%的股票的研报到底是消极还是积极情感，本文在进行词典法多空组合构建时采取买入前 50%股票，卖出后 20%股票的方式。

运用词典法得到的多空组合不同持仓周期下的净值走势如**图 30** 所示：

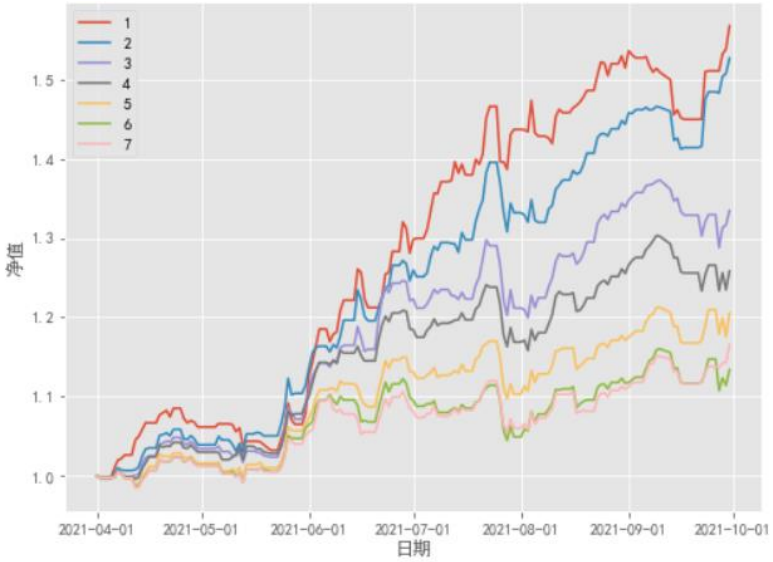



**图 30 词典法多空组合不同持仓周期净值走势**

从图中我们可以看出，词典法下的多空组合在持仓周期为一或两天时区间收益十分接近，但总体看来持仓两天净值走势波动较持仓一天的小，最终一天持仓周期的策略获得了 56.75%的区间收益，最大回撤为 5.63%。随持仓周期的增加，净值表现逐渐平缓。

下面本文同样检验词典法下纯多头不同持仓周期的净值走势，如**图 31** 所示：

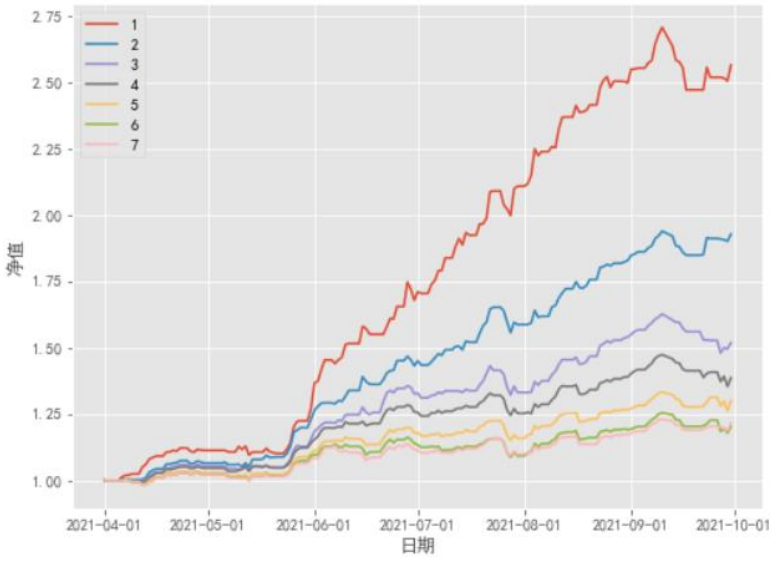



**图 31 词典法纯多头不同持仓周期净值走势**

从上图可以明显看出纯多头净值走势总体上要优于多空组合，同上文我们的结论保持一致。无论是词典法下的纯多头组合还是多空组合，不同持仓周期所展现出来的净值曲线都具有明显的分层效果，而且持仓周期与组合表现呈现负相关特征，持仓周期越短，组合效果更好。

无论是本文所采用的 SESTM 模型还是传统的词典模型，在同一份数据集下都得到了相同的结论，即研报标题所蕴含的积极情绪可以被模型挖掘并放大，但是研报自身的乐观倾向却掩盖了消极情绪，使得两种模型都无法辨别出研报标题包含的消极情绪；本文得到的另一个结论是基于研报标题所释放的乐观情绪进行组合构建仅仅适用于较短的持仓周期，也就是说研报一经发布，所蕴含的信息便可以在较短的时间内被市场消化，无法根据研报标题获得较为持久的组合收益。

在总结了两种模型所得结论一致的情况下，本文又对模型间差异进行了比较。**图 32** 和**图 33** 分别展示了两种模型在纯多头和多空组合下的表现。


**图 32 两种模型各自多空组合的净值走势**

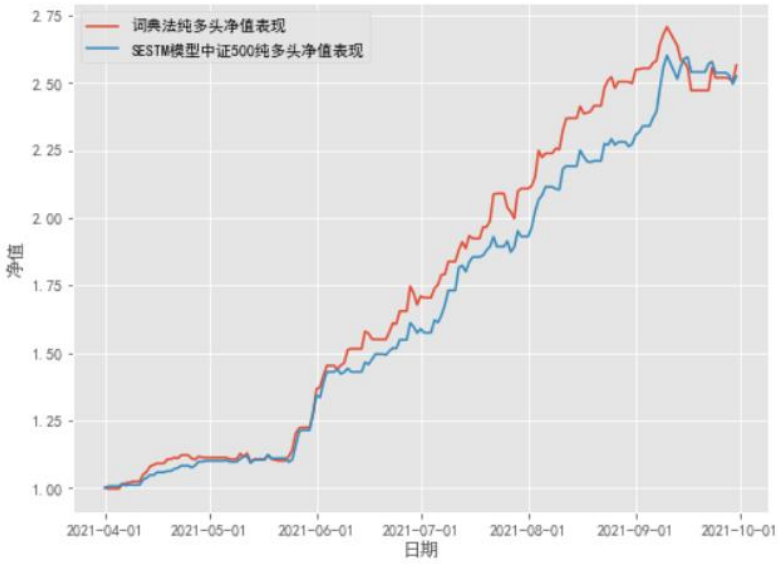



**图 33 两种模型各自纯多头的净值走势**

上图可以直观看出 SESTM 模型走势更平稳，**表 13** 展示了两个模型各自的多空组合

和纯多头绩效表现。

**表 13 词典模型和 SESTM 模型绩效分析**

| 模型 | 策略 | 累计收益 | 最大回撤 | 最大回撤发生时间 | 夏普比率 | 卡玛比率 |
|---|---|---|---|---|---|---|
| SESTM 模型 | 多空组合 | 65.07% | -5.85% | 2021-05-24 | 6.24 | 11.12 |
| | 纯多头 | 152.45% | -4.07% | 2021-09-29 | 12.09 | 37.45 |
| 词典法 | 多空组合 | 56.75% | -5.63% | 2021-09-17 | 4.70 | 10.08 |
| | 纯多头 | 156.51% | -8.70% | 2021-09-17 | 9.80 | 17.98 |

可以看出，在多空组合下，本文使用的SESTM模型收益超出词典法10个百分点，一方面是由于设定的情绪阈值起到了合适的分界作用，另一方面则是因为 SESTM 模型进行情感预测时采用的话题向量经过了训练集的训练，而词典法进行情绪预测则是依赖外部词典，所以在判断是否要开空时，经过训练的模型对样本外数据的区分能力要强于未经训练的词典法。而在纯多头情况下，两个模型表现相似，因为此时已经不存在判断消极情绪能力的影响了，两个模型对于积极情绪均有一定的识别与放大能力。如果从收益上看，本文采用的 SESTM 模型在样本外区间收益略微低于词典法，但从回撤上看，SESTM 模型回撤是词典法回撤的二分之一，此外 SESTM 模型无论是多空组合还是纯多头，其夏普比率和卡玛比率都高于词典法的两个指标。因此本文认为，SESTM 模型除了具有较好的抓涨能力外，其回撤控制能力较传统词典模型更为突出。

## 5.3 分析师情绪指数策略构建

本文所使用的 SESTM 模型的稳健性在前面两小节已经得证，本节我们将基于该模型进行量化策略构建。构建策略第一步应该是检验所用因子的有效性，但由于研究报告数据在时间和标的两个维度上的极度稀疏性，我们无法做到在一个完整的时间序列上分析研报情绪值与对应标的收益之间的相关性，回顾前两章，我们分别对全行业和分行业模型净值走势做了拟合，发现尽管研报比较稀缺，但通过研报进行标的筛选的方法行之有效，情绪阈值对净值的分层效果明显，而且 t 统计量较为显著，一定程度上对该情绪因子起到了检验有效性的作用。

### 5.3.1 初始参数设置

由于研报数据区间长度限制，本文选择模拟回测的区间为 2021 年 4 月 1 日至 2021 年 9 月 30 日，初始资金为 10 万元。交易费用等均通过聚宽平台的接口实现，本文设置的交易成本为每笔交易买入佣金为万分之三，卖出时佣金为万分之三加千分之一的印

花税，每笔交易佣金最低扣 5 元，交易滑点设置为固定值 0.02 元，交易时自动加减 0.01 元，选择的比较基准为中证 500 指数。

**表 14 回测初始信息**

| 初始资金 | 比较基准 | 滑点 | 佣金 | 印花税 | 调仓周期 |
|---|---|---|---|---|---|
| 10 万元 | 000905 | 0.02 | 0.03% | 0.1% | 1 天 |

### 5.3.2 策略构建

策略的持仓周期与情绪阈值参数均使用在前几章中所得到的最优结果，即 0.48 的情绪阈值与一天的持仓周期。该模型的选股逻辑为对前一日发布研报的标的，由 SESTM 模型进行情绪打分，凡是超过情绪阈值的股票均进行买入。由于持仓周期较短，本文的回测将不考虑择时因素，如果施加择时条件，会导致原本就不富裕的研报数据的开仓频率更低，因此本文模型主要体现为选股能力，择时更像是由研报数据发布时点所体现。另外，由于模型可根据昨日发布的研报判断出研报情绪值并决定次日的开仓，所以本文会在本地根据每日研报标题生成持仓矩阵，将此持仓文件上传至聚宽研究环境，并在回测时读取模型持仓矩阵，以东方财富的研报公示日收盘价买入并于持仓日以收盘价卖出，资本分配上将以每日可用资金平均分配至每日的开仓标的数量上。对于已经持仓的标的，如果次日持仓矩阵仍显示买入信号，模型将继续持有而不卖出，从而避免同一标的连续开平仓所带来的手续费。

### 5.3.3 绩效分析

对上述策略逻辑我们在聚宽平台进行测试，模型回测结果如**图 34** 所示：

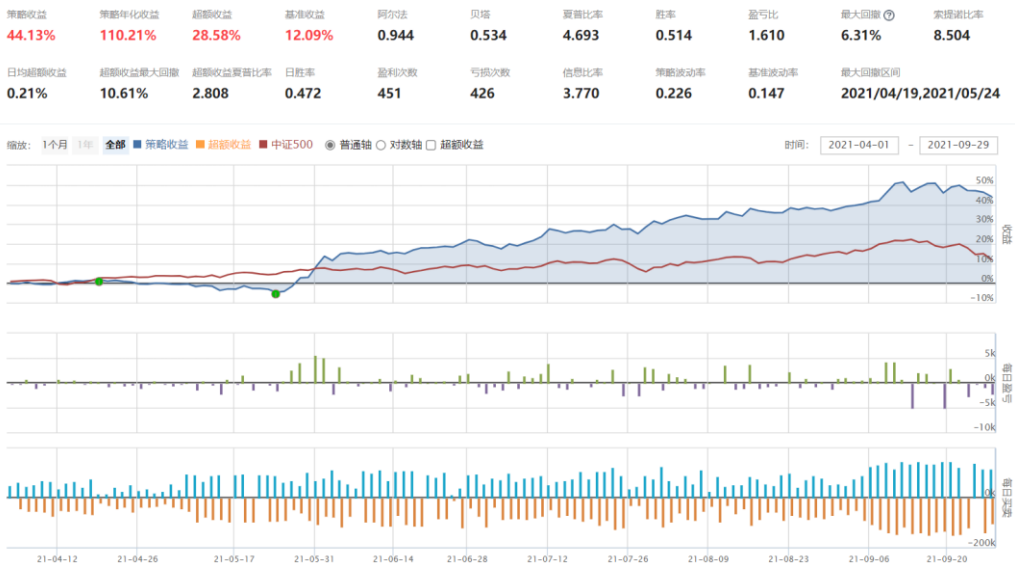



**图 34 SESTM 模型半年回测区间表现**

我们发现在考虑实际交易因素后，模型的表现会比理想情况大打折扣，但也更加真实，模型在该区间内获得 44.13%的收益，超额为 28.58%，模型的胜率为 0.514，虽然模型在胜率上存在略微的优势，但相较于胜率，盈亏比才是模型真正获得收益的主要源泉，从上图中间的每日盈亏柱形图中可以形象的看出代表收益的绿色柱子与代表亏损的紫色柱子的数量上并没有明显的差别，但绿色柱子普遍要比紫色柱子高，因此高赔率是该模型最终获得较高收益的原因。

上图回测曲线的两个绿点为最大回撤发生的起止时间，该时间段包含了四月中旬至五月中旬，出现了 6.31%的回撤，观察最下方的每日交易柱形图，发现该时间段内交易尤其少，而且策略买入的股票一直在亏损，自 5 月 24 日开始策略才开始真正盈利。

下图显示了该模型在每天的持仓数量分布情况，可以看出除了年报和半年报发布节点上模型持仓标的较为分散之外，其余大部分时间的持仓标的数量均位于 10 以下，而本文研报数据覆盖公司有 418 家，充分反映出模型具备优异的选股能力。

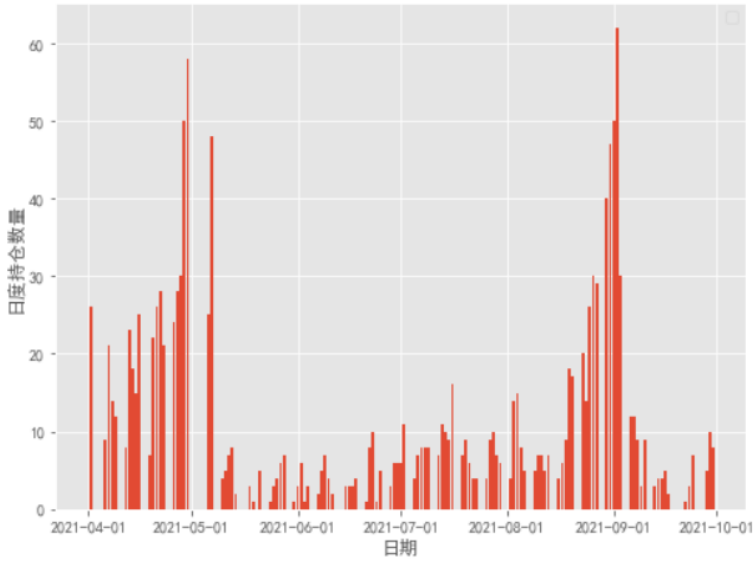



**图 35 SESTM 模型回测区间日度持仓数量**

本文在考虑到样本外区间过短的问题后，重新划分训练集和测试集，将训练集和测试集均设置为一年并进行回测，回测初始参数同上文保持一致，回测结果如下所示：

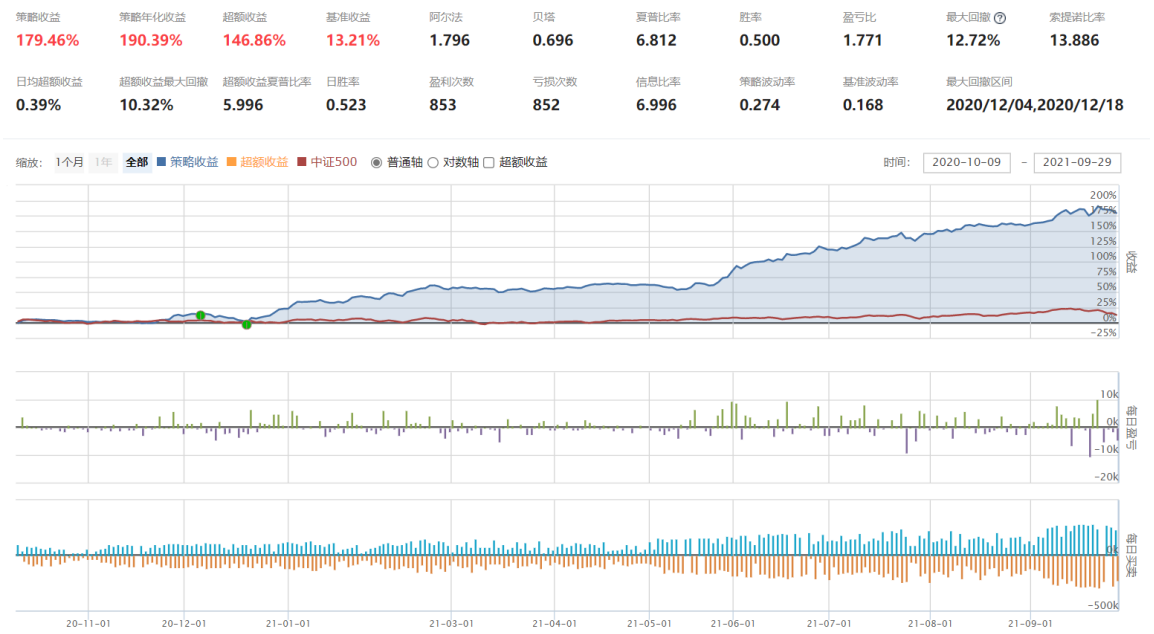



**图 36 SESTM 模型一年样本外区间回测表现**

当以一年作为样本外区间进行回测后发现模型年化收益高达 190.39%，相较于中证 500 指数模型的超额收益高达 146.86%。在胜率上模型并没有展现出明显的优势，但是模型的盈亏比高达 1.771，说明以研报数据作为研究对象，SESTM 模型策略是一个高赔率的策略。策略的最大回撤为 12.72%，回撤区间为 2020 年 12 月 4 日至 2020 年 12 月 18 日，统计后发现该区间内中证 500 跌幅仅为 2.7%，反而是在 2021 年基准指数大幅下跌的三月上旬和九月下旬，模型并没有走出大幅回撤，这也反映出以研报作为数据来源的策略与大盘下跌之间的关联较小，但策略的上涨终究还要依赖指数内股票的上涨。

## 5.4 “高收益”陷阱

如果与市场上大多数包含舆情因子的实盘策略相比，即使是我们的半年回测区间，年化收益也远超市场，但是不得不提的是本文所构建的策略在“一定程度”上使用到了未来函数。如第三章的**图 4** 所示，本文从东方财富上爬取的研报数据日期仅包含年月日信息，而并没有指明该篇研报公布的时分秒信息，但是本文所构建的策略是以该日期下的个股收盘价买入并于次日以收盘价卖出，所以存在的问题是如果研报是在当天收盘后发布的，模型是无法买入的，也就无法在次日以收盘价卖出。

为了解决这一问题，本文尝试从网上找到个股研报发布的具体时间，在查找诸多汇聚研报数据的网站如三个皮匠报告网、发现报告网、萝卜投研和慧博投研等网站后，仅在慧博投研平台上发现携带有研报发布的具体时间，试图通过爬虫的方式获取到每

一篇研报发布的详细时间，但是遗憾的是东方财富一部分研报标题无法在慧博投研中查找到，即使是能够找到对应研报标题，由于慧博平台根据大数据筛选出与搜索信息相似的所有标题，导致无法精准定位到该标的发布的该研报，所以不得不人工筛选对应标题并记录其发布时间。考虑到研报数据量庞大，作者无法利用有限的时间在慧博平台逐篇查找所有研报信息，但同时为了反映出“未来函数”对模型的影响，本文采用逆向处理的方式，挑选一个月的回测区间，将模型在该月份发出交易信号的标的研报数据找出，并手动查找生成开仓信号的研报发布具体时间，以 2021 年 6 月为例，经统计，该月共有 314 篇研报，发出开仓交易信号的有 97 篇，其中有 34 篇研报是在当天收盘后发布的；从持仓标的上看，6 月交易的 64 只股票中有 19 只无法买到，剩余的 45 只股票可以产生信号并在当天以收盘价买入，将考虑研报详细发布时间前后的 2021 年 6 月收益画出，如**图 37** 所示：

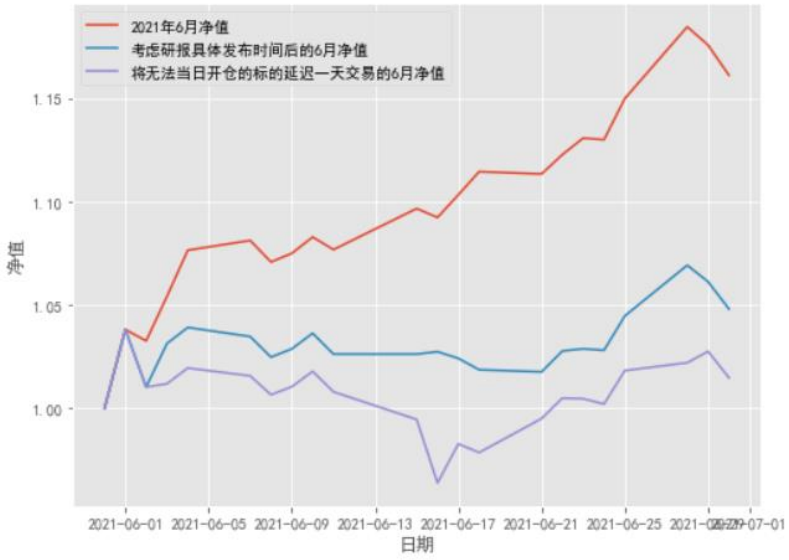



**图 37 考虑研报于收盘前或收盘后发布的 6 月净值数据**

可以明显的看到，在剔除掉 6 月份当天收盘后发布的研报之后，模型的收益从 15 个点降低到 5 个点，两者的巨大差异反映出大多数在收盘后发布的研报对应标的在次日涨幅明显，但由于无法在当日开仓买入使得总体净值下降。本文考虑到研报所带来的刺激效应可能具有一定的持续性，尝试将原本无法在当日开仓的标的延迟一天开仓并于开仓后下一交易日卖出，但最终的净值曲线却更糟糕，回撤幅度更大，充分反映出研报所带来的刺激效应不具备持续性，市场在一个交易日内便可充分吸收信息。胡昌生和高玉森(2020)认为利用分析师情绪对个体投资者进行交易诱导是实现“韭菜”收割的有效途径。从该图中我们发现研报发布后下一交易日收益率与第二天的收益率呈反

向变动，体现出短期高涨的情绪和隐约存在的机构“高抛”的影子。不过我们同时应该承认本文所做研究具备一定的意义，即使某些研报在收盘后才产生信号，但“未来函数”对于策略收益的负面影响是有限的，不足以推翻本文的研究结论，更细化的影响系数还需要进一步的研究工作来证实。

## 5.5 本章小结

本章首先检验模型的稳健性，一方面采用不同的数据来源测试模型对新市场的适应能力，另一方面为了使模型更具对比性，采用传统的词典法与 SESTM 模型进行对比分析。实验结果表明，无论是对于中证 500 还是中证 1000 的研报数据，模型均具有良好的超额，而通过与词典法的比较则突出了 SESTM 模型除选股能力外较为优秀的回撤控制能力，验证了 SESTM 模型的有效性。

接着本文根据该模型简单构建了相应的量化策略在聚宽平台分别对半年和一年的样本外区间进行回测，并对日度持仓数量分布进行可视化，发现无论是半年还是一年的回测区间均取得较好的收益。此外，从日度持仓分布上可以看出多数时间持股数量少于 10 只，侧面反映出以研报作为数据源的模型具备良好的选股能力。

最后，本章指出了模型存在部分“未来函数”的影响，这是本文一个缺陷，但该影响并未推翻本文结论，模型仍能获得一定的超额收益。

# 第六章 结论

本文运用基于话题向量的有监督学习模型 SESTM 对中证 500 的个股分析师研报进行文本挖掘，一定程度上丰富了学界对研报情绪领域的研究，并运用该模型对个股情绪走向进行预测，从而构建相应的量化策略，文章主要从以下几个方面进行分析。首先，基于爬取的研报数据，进行大量预处理和样本内建模工作，通过所有成分股构建的话题向量对样本外研报情绪进行预测。其次，考虑到同一时期各行业具备不同特征，对每一行业分别构建话题向量并综合对比分行业和全行业模型的绩效差异。然后，在考虑交易费用、成本等因素影响后，使用全行业模型最优参数来构建量化策略，并对持仓和交易特征进行分析。最后，为了检验模型稳健性，本文采用不同的数据源和词典模型分别进行绩效分析，从而检验模型对新数据的适应性，以及相较于传统模型的优越性。

综合分析研究后，本文得出以下五点结论：

第一，基于 SESTM 的研报文本挖掘模型具备一定的选股能力，从某种意义上讲，分析师分析的标的在短期内确实有不错的表现。

第二，研报数据的市场有效性较短，在一天时间内即可充分被市场消化吸收，因此跟踪分析师研报做长期投资决策是站不住脚的。

第三，不同行业具备不同的话题特征，因此分行业模型效果要优于全行业模型，但不能单独使用单行业话题模型进行投资决策；

第四，分行业模型具备更高的进攻性，而全行业模型则更加稳定，这也说明对单行业建模确实显著提升了组合收益，而全行业模型更具普适性。

第五，研报与生俱来的乐观倾向导致模型对于消极情绪的识别能力较差。这一点与经验相符，毕竟大多数的分析师研报更倾向于发布乐观观点，抬高了积极消极情绪的分界线。

本研究的不足之处有两点，一方面在于数据区间过短，不具备足够的说服力，另一方面是未能对未来函数的影响系数进行量化，本文仅仅是初步检验了一个月的于收盘之后发布的研报无法开仓对组合收益的影响，但拉长时间区间去看这个影响会是什么样的还需要进一步的探究。

在大数据潮流趋势下，信息会愈加透明化，研究报告作为卖方分析师发表观点以及获取收益的途径，其规模和形式必然呈现出丰富化和多样性特点，研报所蕴含的海

量信息还需要投资者进一步挖掘。

# 参考文献

[1] Barberis, Nicholas, Andrei Shleifer, and Robert Vishny. 1998. “A Model of Investor Sentiment”, Journal of Financial Economics, 49(3): 307~343.

[2] [美]乔治·阿克洛夫, 罗伯特·席勒. 钓愚: 操纵与欺骗的经济学, 张军译[M]. 北京: 中信出版集团, 2016.

[3] 胡昌生, 高玉森. 分析师情绪与交易诱导: A 股分析师是“虚情假意”的吗[J]. 金融经济学研究, 2020, 35(01): 131-145.

[4] 沈艳, 陈赟, 黄卓. 文本大数据分析在经济学和金融学中的应用:一个文献综述[J]. 经济学(季刊), 2019, 18(04): 1153-1186.

[5] Li, F, Y. Chen, Y. Wang, Z. Huang, “Measuring China’s Stock Market Sentiment”[J], Working Paper, 2019.

[6] 王科, 夏睿. 情感词典自动构建方法综述[J]. 自动化学报, 2016(04): 495-511.

[7] Zhen D, Qiang D. HowNet and the computation of meaning[J]. World Scientific, 2006.

[8] Blei D M. Probabilisitic topic models[J]. Communication of the ACM, 2012, 55(4): 77-84.

[9] 何伟林, 谢红玲, 奉国和. 潜在狄利克雷分布模型研究综述[J]. 信息资源管理学报, 2018, 8(01): 55-64.

[10] Zheng Tracy Ke, Brain Kelly, Dacheng Xiu. Predicting Returns with Text Data[J]. Working Paper. No. 2019-69.

[11] Antonio Reyes, Paolo Rosso, Making objective decisions from subjective data: Detecting irony in customer reviews[J]. Decision Support Systems, Volume 53, Issue 4, 2012, Pages 754-760, ISSN 0167-9236.

[12] 董爽, 王晓红, 葛争红. 基于文本挖掘的 B2C 购物网站在线评论内容特征分析[J]. 图书馆理论与实践, 2017(06): 54-58.

[13] 刘敏, 王向前, 李慧宗, 张宝隆. 基于文本挖掘的网络商品评论情感分析[J]. 辽宁工业大学学报(自然科学版), 2018, 38(05): 330-335.

[14] 张红霞. 生鲜农产品电子商务消费者满意度影响因素——基于在线评论的探索分析[J]. 江苏农业科学, 2019, 47(17): 4-8.

[15] PU X J, WU G S, YUAN C F. Exploring Overall Opinion for Document Level Sentiment Classification with Structural SVM[J]. Multimedia Systems, 2019(1): 21-33.

[16] 范宁. 基于文本挖掘在民宿满意度中的研究[D]. 广西师范大学, 2019.

[17] 何立峰. 基于在线评论的酒店顾客需求分析研究[D]. 青岛大学, 2019.

[18] 李薇, 杨东山. 基于回头客在线评论的餐饮消费满意度影响因素分析[J]. 重庆邮电大学学报(社会科学版), 2021, 33(02): 125-134.

[19] 杨秀璋, 武帅, 夏换, 于小民, 范郁锋, 丛楠, 张懿源. 面向贵州省三大战略行动的文本挖掘及 LDA 模型分析研究[J]. 现代计算机, 2020(25): 9-14.

[20] 晁筱雯, 周京生, 李育平, 刘雨婷, 李疏影, 陈麒, 卢光玉. 基于文本挖掘的我国传染病研究主题与方法演进分析[J]. 预防医学情报杂志, 2021, 37(06): 865-871.

[21] 姜坤, 刘苗. 基于文本挖掘技术的印媒中美关系报道情感立场分析[J]. 对外传播, 2021(02): 77-80.

[22] 许光, 任明, 宋城宇. 西方媒体新闻中的中国经济形象提取[J]. 数据分析与知识发现, 2021, 5(05): 30-40.

[23] 陈聪聪, 赵怡晴, 姜琳婧, 唐舟, 田欣然. 基于文本挖掘的尾矿库隐患因素关联分析[J]. 矿业研究与开发, 2021, 41(11): 26-33.

[24] 卢玉坤, 唐文. 基于古代诗词文本挖掘与分析的南京城市诗意景观研究[J]. 园林, 2021, 38(12): 52-57.

[25] 刘赛红, 黄馨锋, 余意. 新型农业经营主体生产性消费金融风险识别——基于文本挖掘及问卷调查研究[J]. 系统工程, 2022, 40(01): 121-132.

[26] 韩天园, 田顺, 吕凯光, 李旋, 张佳涛, 魏朗. 基于文本挖掘的重特大交通事故成因网络分析[J]. 中国安全科学学报, 2021, 31(09): 150-156.

[27] 舒洪水. 司法大数据文本挖掘与量刑预测模型的研究[J]. 法学, 2020(07): 113-129.

[28] 石勇, 安文录, 曲艺. 基于文本挖掘的检察起诉决策支持与案卷分类管理系统[J/OL]. 管理评论: 1-10[2021-12-18]. https://doi.org/10.14120/j.cnki.cn11-5057/f.20211217.002.

[29] 杨超, 姜昊, 雷峥嵘. 基于文本挖掘和百度指数的汇率预测[J]. 统计与决策, 2019, 35(13): 85-87.

[30] 戴德宝, 兰玉森, 范体军, 赵敏. 基于文本挖掘和机器学习的股指预测与决策研究[J]. 中国软科学, 2019(04): 166-175.

[31] 张杰, 张永卿, 翟东升. 融合财经新闻信息的汇率波动预测[J]. 系统工程, 2021, 39(03): 121-131.

[32] Cheng K, Liu R H. Analysis on Interaction of Investor Sentiment and Stock Market[J]. Shanghai Economic Review, 2005(11): 88-95.

[33] Xue F. Empirical Test of Investor Sentiment Index Selection in China[J]. World Economic Outlook, 2005(14): 14-17.

[34] Delong J.B, Shleifer A, Summers L. and Waldmam R J. Noise trader risk in financial markets[J]. The Journal of Political Economy, 1990, 98(4): 703-738.

[35] 张超. 基于噪声交易理论的中国封闭式基金折价研究[D]. 西北大学, 2014.

[36] 汪丽雯. 中国封闭式基金折价问题的研究[D]. 安徽农业大学, 2017.

[37] Statman, M., S. Thorley, and K. Vorkink(2006). Investor overconfidence and trading volume[J]. Review of Financial Studies 19(4), 1531-1565.

[38] WEN F H, YANG X, GONG X. The research on investor sentiment contagion between China and U.S. based on the background of financial crisis[J]. Social Science Electronic Publishing, 2015, 28(10): 1-32.

[39] ANTWEILER W, FRANK M Z. Is all that talk just noise? The information content of Internet stock message boards[J]. Journal of Finance, 2004, 59(3): 1259-1294.

[40] BOLLEN J, MAO H, ZENG X. Twitter mood predicts the stockmarket[J]. Computer Science, 2010, 2(1): 1-8.

[41] Oh C, SHENG O. Investigating Predictive Power of Stock Price Directional Movement[C]//Proceedings of the International Conference on Information Systems(ICIS 2011), Shanghai China, 2011: 2860-2877.

[42] Johan Bollen, Huina Mao, Xiaojun Zeng. Twitter mood predicts the stock market[J]. Journal of Computational Science. Volume 2, Issue 1. 2011. Pages 1-8.

[43] Nuno Oliveira, Paulo Cortez, Nelson Areal. The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices[J]. Expert Systems with Applications. Volume 73. 2017. Pages 125-144.

[44] 黄雨婷, 宋泽芳, 李元. 基于文本挖掘的股评情绪效应分析[J/OL]. 数理统计与管理: 1-14[2021-12-18]. https://doi.org/10.13860/j.cnki.sltj.20211130-010.

[45] Thomas Renault. Intraday online investor sentiment and return patterns in the U.S. stock market. Journal of Banking & Finance. Volume 84. 2017. Page 25-40.

[46] Yin H Y, WU X Y. Predictive Effect of High-frequency Investor Sentiment on the Intraday Stocks Return[J]. China Industrial Economics, 2019(8): 80-98.

[47] 徐维军, 付志能, 李茂昌, 张卫国. 基于新闻文本挖掘的股指期货高频预测研究[J]. 系统科学与数学, 2021, 41(07): 1856-1875.

[48] Frey S. and P.Herbst, 2014, “The Influence of Buy-side Analysts on Mutual Fund

Trading", Journal of Banking & Finance, 49, pp.442-458.

[49] Chen X. and Cheng, 2005, "Institutional Holdings and Analysts' Stock Recommendations", Journal of Accounting, Auditing and Finance, 21(4), pp.399-440.

[50] Costello D. and J. Hall, 2014,"The Impact of Security Analyst Recommendations upon the Trading of Mutual Funds", Journal of Asset Management, 15(2), pp.92-109.

[51] Franck A. and A. Kerl, 2013, "Analyst Forecasts and European Mutual Fund Trading", Journal of Banking & Finance, 37(8) ,pp.2677-2692.

[52] 伊志宏, 李颖, 江轩宇. 女性分析师关注与股价同步性[J]. 金融研究, 2015(11).

[53] 胡昌生, 高玉森. "分析师情绪会影响股票价格吗?". 《投资研究》, 2018 年第 1 期, 第 99-113 页.

[54] 蔡庆丰, 杨侃, 2013: 《是谁在"捕风捉影": 机构投资者 VS 证券分析师》,《金融研究》第 6 期.

[55] 戴方贤, 尹力博. 分析师目标价预测是否引导了基金集中持股行为[J]. 投资研究, 2016(11).

[56] 张化侨, 2010《一个证券分析师的醒悟》, 中信出版社.

[57] 蔡庆丰, 杨侃, 2013: 《是谁在"捕风捉影": 机构投资者 VS 证券分析师》,《金融研究》第 6 期.

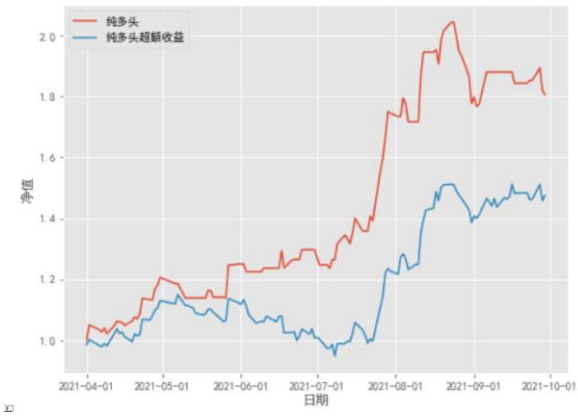
[58] 吴超鹏, 郑方镳, 杨世杰, 2013: 《证券分析师的盈余预测和股票评级是否具有独立性?》,《经济学(季刊)》第 3 期.

[59] 张宗新, 杨万成. 声誉模式抑或信息模式: 中国证券分析师如何影响市场?[J]. 经济研究, 2016, 51(09): 104-117.

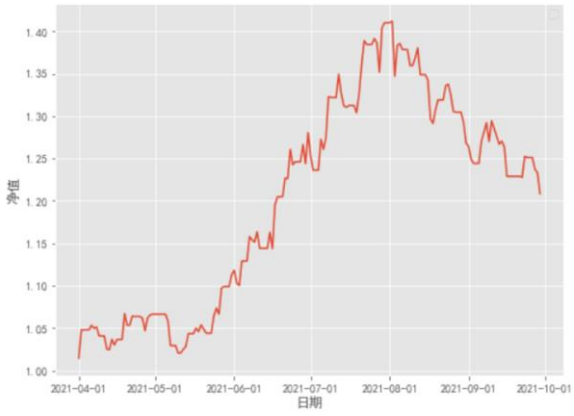
[60] 丁方飞, 肖晓乐, 陈智宇, 乔紫薇. 证券市场上的分析师:理性引领抑或随波逐流?——一个文献综述[J]. 湖南财政经济学院学报, 2020, 36(01): 90-98.

[61] 姜富伟, 孟令超, 唐国豪. 媒体文本情绪与股票回报预测[J]. 经济学(季刊), 2021, 21(04): 1323-1344.

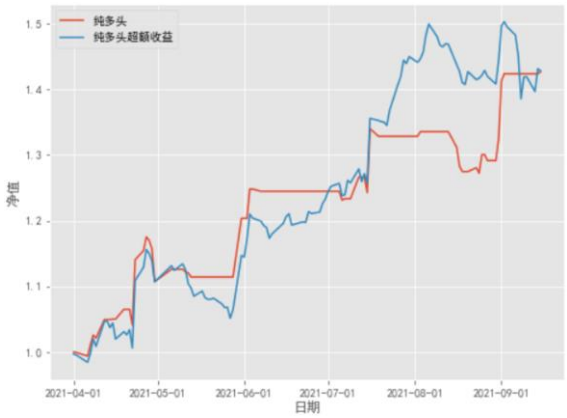
# 附图



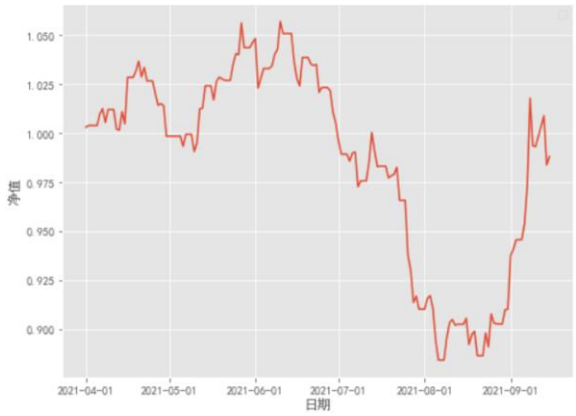

附图 1　电子行业 0.48 情绪阈值下持仓 1 天纯多头




附图 2　中证 500 电子行业等权指数

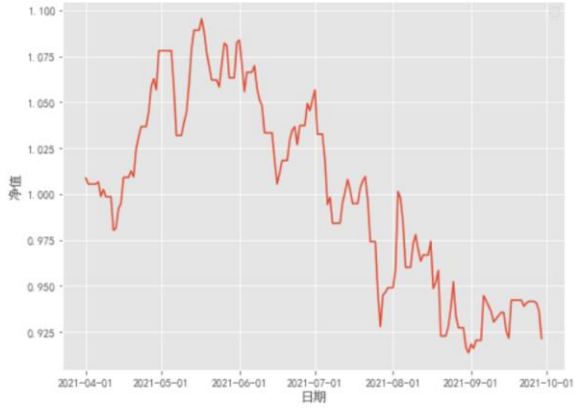



附图 3　传媒行业 0.48 情绪阈值下持仓 1 天纯多头




附图 4　中证 500 传媒行业等权指数

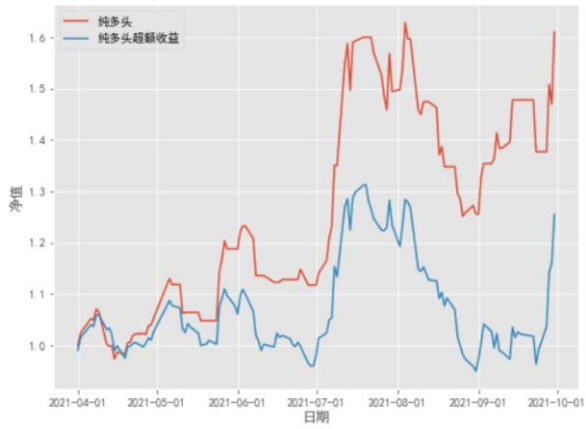

附图 5　医药生物行业 0.48 情绪阈值持仓 1 天纯多头



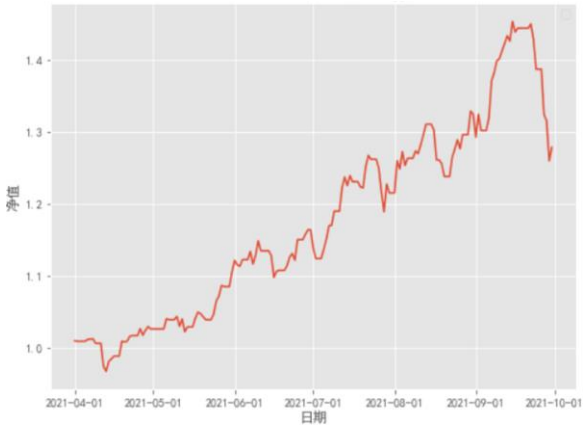

附图 6　中证 500 医药生物行业等权指数


附图 7　计算机行业 0.48 情绪阈值持仓 1 天纯多头


附图 8　中证 500 计算机行业等权指数




附图 9　化工行业 0.48 情绪阈值持仓 1 天纯多头




附图 10　中证 500 化工行业等权指数

# 附表

**附表 1 电子行业纯多头及其超额收益绩效对比**

| 净值 | 区间累计收益% | 最大回撤% | 最大回撤发生时间 | 夏普比率 | 卡玛比率 |
|---|---|---|---|---|---|
| 纯多头 | 64.83% | -14.02% | 2021-08-31 | 3.44 | 4.62 |
| 纯多头超额收益 | 36.30% | -10.91% | 2021-07-22 | 1.92 | 3.33 |

**附表 2 传媒行业纯多头及其超额收益绩效对比**

| 净值 | 区间累计收益% | 最大回撤% | 最大回撤发生时间 | 夏普比率 | 卡玛比率 |
|---|---|---|---|---|---|
| 纯多头 | 50.55% | -5.74% | 2021-08-24 | 4.10 | 8.81 |
| 纯多头超额收益 | 51.07% | -7.83% | 2021-09-8 | 3.65 | 6.53 |

**附表 3 医药生物行业纯多头及其超额收益绩效对比**

| 净值 | 区间累计收益% | 最大回撤% | 最大回撤发生时间 | 夏普比率 | 卡玛比率 |
|---|---|---|---|---|---|
| 纯多头 | 34.57% | -10.10% | 2021-08-30 | 2.36 | 3.42 |
| 纯多头超额收益 | 46.09% | -7.87% | 2021-07-01 | 3.29 | 5.86 |

**附表 4 计算机行业纯多头及其超额收益绩效对比**

| 净值 | 区间累计收益% | 最大回撤% | 最大回撤发生时间 | 夏普比率 | 卡玛比率 |
|---|---|---|---|---|---|
| 纯多头 | 1.94% | -13.67% | 2021-08-18 | 0.14 | 0.14 |
| 纯多头超额收益 | -0.60% | -19.57% | 2021-09-08 | -0.04 | -0.03 |

**附表 5 化工行业纯多头及其超额收益绩效对比**

| 净值 | 区间累计收益% | 最大回撤% | 最大回撤发生时间 | 夏普比率 | 卡玛比率 |
|---|---|---|---|---|---|
| 纯多头 | 60.97% | -23.17% | 2021-08-26 | 2.36 | 2.63 |
| 纯多头超额收益 | 26.66% | -27.63% | 2021-08-31 | 1.12 | 0.96 |

# 攻读硕士学位期间取得的科研成果

无

# 致谢

2020 年，难以忘记的一年，全国团结抗疫，众志成城，全校封闭管理，战战兢兢，朋友互帮互助，其乐融融。不平淡的一年，带我走入不平凡的研究生生活。

在这里要由衷的感谢王莉老师的指导，王莉老师带领学生不断探索新领域，不断参加国内比赛，在取得优异成绩的同时，还经常为学生的工作忙的焦头烂额，她对待学术研究的激情和专注深深的感染了实验室每一位同学，也为我们共同前进指明了方向。

感谢实验室的伙伴们，你们的坚持不懈和互帮互助是我前进的动力，怀念我们一起度过的那些即艰难又幸福的日子，怀念我们共同 debug 的日子，更怀念我们一起考证和分享快乐的时光，这些脑海中时常闪过的画面，是我一生的珍藏，谢谢你们。

感谢实习遇到的伙伴们，相遇很短，成长很多，尤其感谢有旗姐和田田姐的指导与帮助。

同时也要向我的父母说声谢谢，感谢你们在背后默默的支持，感谢你们对我的信任，更要感谢你们一生的教诲。

最后我想说一声，学海无涯，感谢自己，希望你能怀揣梦想，拥抱希望，在生活中奔向阳光。