上 海 财 经 大 学

毕 业 论 文

题目 基于自然语言处理的投资者情绪对股市行情的影响研究

姓　　名：付舟行

学　　号：2021111662

学　　院：商学院

专　　业：工商管理（商务分析）

指导教师：田中俊

定稿日期　　年　　月

声　　明

本人郑重声明所呈交的论文是我个人在指导老师的指导下进行的研究工作及取得的研究成果，不存在任何剽窃、抄袭他人学术成果的现象。我同意（　　）/不同意（　　）本论文作为学校的信息资料使用。

论文作者（签名）_____

年　　月　　日

基于自然语言处理的投资者情绪对股市行情的影响研究

摘 要

本研究探讨了投资者情绪对股票市场短期收益率的影响，旨在回答网络平台上投资者言论所蕴含的情绪信息是否能有效预测股票价格变动的问题。研究基于行为金融学理论，认为投资者情绪作为一种非理性因素会影响投资者决策，从而对资产价格产生系统性影响；同时借鉴信息传导理论，探究情绪信息在市场中的传递机制和价格发现过程。

研究采用自然语言处理技术（NLP）和计量经济学方法。通过 Python 爬虫从东方财富网股吧获取股评文本数据，应用 StructBERT 深度学习模型对文本进行情感倾向识别与量化，构建情绪指标体系；利用面板数据固定效应模型分析情绪指标对股票收益率的影响。

实证分析覆盖了 2025 年 2 月期间，中国 A 股市场 8 个主要行业的 111 只代表性股票。研究流程具体涵盖了股评文本的预处理与情感分析，情绪指标体系的构建，个股层面的面板数据回归分析，分行业异质性分析，内生性问题讨论，以及对不同情绪测度指标和不同期限收益率的稳健性检验。

研究结果发现：投资者情绪对短期股价具有显著的预测能力，具体表现为平均情绪得分每提升一个百分点，次日股票收益率平均增加约 0.014 至 0.017 个百分点。情绪的预测效应呈现显著的行业异质性，在计算机、电子、食品饮料及医药生物等板块尤为突出。情绪的影响具有明显的时效性，主要体现在短期，随着时间延长逐渐减弱甚至出现反转，符合情绪过度反应-修正理论。在情绪指标中，平均情绪得分和正面情绪比例是预测股票收益率最有效的指标。

关键词：投资者情绪，自然语言处理，BERT，股票收益率，面板数据分析

ABSTRACT

This study investigates the impact of investor sentiment on short-term stock market returns, aiming to determine whether sentiment information embedded in online platform discussions can effectively predict stock price movements. Grounded in behavioral finance theory, the research posits that investor sentiment, as a non-rational factor, influences investor decisions and systematically affects asset prices. It also draws upon information transmission theory to examine the mechanisms through which sentiment information propagates within the market and contributes to the price discovery process.

The study employs a combination of Natural Language Processing (NLP) techniques and econometric methods. Textual data from stock comments were collected from East Money's Guba platform using Python web scraping. The StructBERT deep learning model was applied for sentiment classification and quantification, enabling the construction of a sentiment indicator system. Subsequently, a panel data fixed-effects model was utilized to analyze the impact of these sentiment indicators on stock returns.

The empirical analysis covers 111 representative stocks across 8 major industries in the Chinese A-share market during February 2025. The research procedure encompasses several stages: preprocessing and sentiment analysis of stock comments, construction of the sentiment indicator system, stock-level panel data regression analysis, examination of cross-industry heterogeneity, discussion of potential endogeneity issues, and robustness checks using alternative sentiment measures and different return horizons.

The findings reveal that: Investor sentiment significantly predicts short-term stock returns; specifically, a one percentage point increase in the average sentiment score is associated with an approximate 0.014 to 0.017 percentage point increase in the next-day stock return. The predictive power exhibits significant industry heterogeneity, being particularly prominent in sectors such as Computer, Electronics, Food & Beverage, and Pharmaceuticals & Biotechnology. The impact of sentiment is predominantly short-lived, diminishing or even reversing over longer horizons, consistent with the overreaction and correction theory. Among the constructed indicators, the average sentiment score and the proportion of positive sentiment emerge as the most effective predictors of stock returns.

KEY WORDS: Investor Sentiment, Natural Language Processing (NLP), BERT, Stock Returns, Panel Data Analysis

目 录

摘 要 I

ABSTRACT II

目 录 IV

第一章 引言 1

第一节 研究背景与意义 1

一、研究背景 1

二、研究意义 2

第二节 研究内容与方法 3

一、研究内容 3

二、研究方法 4

第三节 研究创新点 4

第四节 结构安排 4

第二章 文献综述 5

第一节 投资者情绪研究综述 5

一、投资者情绪的定义与测度 5

二、投资者情绪与资本市场表现关系研究 6

第二节 自然语言处理技术研究进展 6

一、文本情感分析方法演进 6

二、自然语言处理技术在金融领域的应用 7

第三节 研究评述 8

第三章 研究设计 9

第一节 数据来源与处理 9

一、股票评论文本数据获取与预处理 9

二、市场交易数据 12

第二节 情绪分析模型构建 14

一、基于 StructBERT 的情感分析框架 14

二、基于 RoBERTa 的情感分析框架 14

三、情感分布描述性统计 15

第三节 情绪指标体系构建 16

第四节 模型设定及变量说明 17

一、模型设定 17
二、变量说明 17
第四章 实证分析 19
第一节 描述性统计 19
第二节 单位根检验与协整检验 20
一、单位根检验 20
二、协整检验 21
第三节 相关性分析与多重共线性检验 22
一、相关性分析 22
二、多重共线性检验 23
第四节 面板数据回归分析 24
一、面板数据模型构建 24
二、Hausman 检验与模型选择 24
三、固定效应回归分析 26
第五节 异质性分析 28
一、 分行业回归分析 28
第六节 内生性处理 29
一、格兰杰因果关系检验 29
二、工具变量讨论 30
第七节 稳健性检验 31
一、不同情绪指标的比较 31
二、不同期限的收益率比较 32
第五章 总结与展望 35
第一节 研究结论 35
第二节 不足与展望 35
一、数据限制 35
二、方法局限 36
参考文献 37
附录 A 代码仓库 39
致　　谢 40

第一章 引言

第一节 研究背景与意义

一、研究背景

（一）投资者情绪对资本市场的重要影响

在经典的金融理论框架中，有效市场假说长期占据主导地位，该假说认为资产价格能够完全且迅速地反映所有可获得的信息，投资者据此进行理性决策。然而，现实中的资本市场常常展现出传统理论难以完全解释的现象，例如资产价格的过度波动、泡沫的形成与破灭以及各种市场异象的持续存在。这些现象促使学术界将目光投向了行为金融学领域，该领域承认投资者的决策过程并非完全理性，而是受到心理因素、认知偏差以及情绪波动的显著影响。在众多影响投资者行为的心理因素中，投资者情绪被认为是理解和解释市场动态，尤其是股价偏离其内在价值现象的一个核心变量，其对资本市场的运行效率、稳定性乃至资源配置功能都产生着深远且复杂的影响。

投资者情绪，通常被定义为市场参与者对于未来市场走向或特定资产价值的一种普遍性的、非完全基于基本面信息的信念或预期。这种情绪并非个体心理的简单加总，而是在群体互动和信息传播过程中形成的集体性心理倾向。当乐观情绪弥漫市场时，投资者可能倾向于高估资产价值、忽视潜在风险，从而推动股价非理性上涨，甚至催生资产泡沫；反之，当悲观情绪主导市场时，投资者可能过度反应负面信息，导致恐慌性抛售，使得股价跌破其基本面支撑，加剧市场下行压力。因此，投资者情绪的波动往往是驱动市场价格偏离其理论公允价值，引发过度反应或反应不足等市场异象的关键力量。这种偏离不仅影响了短期市场表现，长期来看也可能对资本的有效配置产生扭曲。

投资者情绪对资本市场的影响路径是多维度的。其一，它直接作用于投资者的交易决策，进而影响市场的交易量和波动性。高涨的乐观情绪往往伴随着交易活动的增加和市场波动性的放大，而低迷的情绪则可能导致市场流动性萎缩。其二，投资者情绪能够影响不同类型资产的相对吸引力，引导资金在不同板块、风格或风险等级的股票之间流动，从而影响市场的结构性特征和横截面收益差异。例如，有研究表明，在情绪高涨时期，投机性强、估值高、基本面不确定性大的股票可能更受追捧。再者，投资者情绪对于首次公开募股 IPO 的定价和短期表现也具有显著影响，高昂的市场情绪往往能推高 IPO 的发行价格和首日回报率。其三，情绪的变化还可能通过影响分析师预测、公司投资决策乃至宏观经济预期等间接渠道，进一步传导至资本市场。综上所述，投资者情绪已成为影响资本市场的重要因素，不可忽视其在价格形成和市场波动中的关键作用。

（二）自然语言处理技术在金融领域的应用

随着信息技术的迅猛发展和互联网的普及，金融市场的信息环境发生了深刻变革。海量的、非结构化的文本数据，如新闻报道、公司公告、分析师报告、社交媒体讨论、政府政策文件等，正以前所未有的速度和规模产生。这些文本数据蕴含着关于市场预期、公司基本面、宏观经济状况以及投资者情绪的丰富信息，对于理解市场动态和做出投资决策具有不可估量的价值。然而，传统的数据分析方法主要依赖于结构化的数值数据，难以有效处理和利用这些庞杂的文本信息。面对这一挑战，自然语言处理技术的崛起为金融领域带来了颠覆性的机遇。NLP 作为人工智能的一个重要分支，致力于让计算机能够理解、解释、生成和处理人类语言，其强大的文本分析能力使其成为从金融文本大数据中挖掘价值的关键工具。

在金融领域，自然语言处理技术的应用已渗透到多个层面，展现出广泛的适用性和巨大的潜力。在信息提取与分析方面，NLP 技术能够自动识别和抽取文本中的关键信息，如公司名称、财务指标、重大事件、政策变动等，将非结构化的文本转化为结构化的数据，极大地提高了信息处理的效率和准确性。这使得研究人员和投资者能够更快地捕捉市场信号，评估事件影响。在情感分析方面，通过分析文本中表达的情感倾向及其强度，NLP 技术可以量化市场参与者的情绪状态。这为直接度量投资者情绪提供了全新的、数据驱动的途径，克服了传统情绪指标在时效性、覆盖面和客观性上的诸多不足。例如，通过对财经新闻、社交媒体帖子进行大规模、实时的情感分析，可以构建高频的投资者情绪指数，为市场预测和风险管理提供重要参考。

随着大数据和人工智能技术的迅猛发展，自然语言处理技术在金融领域的应用前景日益广阔。近十年来，人工智能科技已成为全球最热门的科学技术领域。相比传统基于词典和机器学习的统计方法，基于深度学习模型的自然语言处理技术，例如 BERT、GPT 系列等，在情绪信息识别方面表现出色。通过对 BERT 模型进行金融领域特定文本的进一步预训练并调整其内部语义分布，可以显著提升舆情抽取和识别任务的准确性和可信度。这些技术进步为挖掘海量金融文本数据中的情绪信息提供了有力支持。

二、研究意义

（一）理论意义

本研究通过引入自然语言处理技术直接、动态且精细地量化投资者情绪，超越了传统间接指标的局限，深化了对情绪影响资产定价机制的理解，为行为金融学提供了来自中国市场的微观实证支持。在方法论层面，本研究验证了深度学习模型处理复杂金融文本、从另类数据中萃取市场信号的有效性，为计算金融和金融科技领域提供了重要的技术路径与方法论参考。通过系统考察 NLP 量化情绪在中国 A 股的作用，揭示了其影响股价的特定机制与行业异质性，丰富了对特定市场环境下投资者行为与资产价格动态交互的认识，为完善中国市场资产定价理论贡献了经验证据。

（二）现实意义

对投资者而言，本研究构建的情绪指标可以作为投资决策的重要参考。通过分析互联网评论中蕴含的市场情绪，投资者可以更全面地把握市场预期变化，捕捉交易机会、避免非理性跟风行为，并优化资产配置。对监管机构而言，本研究的方法有助于实时监测市场情绪波动，及时识别潜在的市场操纵、过度投机或系统性恐慌风险，为维护市场稳定和保护投资者提供重要的早期预警与决策支持。对研究机构而言，本研究提供了一种新型的市场分析工具。通过整合传统量化指标与情绪分析结果，研究人员可以构建更全面的市场预测模型，提高预测准确性。

第二节 研究内容与方法

一、研究内容

本研究旨在通过构建情绪指标体系，深入探究投资者情绪对股票市场表现的影响。本文的研究内容主要包括四个方面：

第一，数据获取与预处理。本研究利用 Python 编程技术开发网络爬虫，定向采集国内主流财经论坛—东方财富网股吧中关于特定股票的投资者评论文本数据。选取的样本覆盖 A 股市场中具有代表性的 8 个行业共计 111 只股票，时间跨度设定为 2025 年 2 月。同时，通过 Akshare 等金融数据接口获取相应的股票日度交易数据以及行业指数数据。采集到的原始文本数据经过严格的数据清洗流程，包括去除无关信息、处理特殊字符与表情符号、分词等步骤，以构建高质量的文本语料库用于后续分析。

第二，投资者情绪的量化分析。本研究采用前沿的自然语言处理技术，特别是基于 Transformer 架构的深度学习预训练语言模型，对预处理后的股评文本进行情感倾向分析。通过这些模型对每一条评论进行情感极性的判断，并输出相应的情感得分或概率，从而实现对非结构化文本中蕴含的投资者情绪进行精细化、自动化量化。

第三，情绪指标体系的构建。基于单条评论的情感分析结果，本研究在个股和行业两个层面，按日度频率构建情绪指标体系。这些指标不仅包括反映整体情绪水平的平均情绪得分、情绪强度，还涵盖情绪的波动性、情绪的一致性、不同情绪的比例以及评论热度等。此外，还构建情绪的动量指标和变化率指标，以捕捉情绪的动态特征。

第四，情绪影响的实证检验与机制探讨。本研究构建面板数据集，整合前期构建的情绪指标与股票市场交易数据。运用计量经济学方法，重点采用面板数据固定效应模型，实证检验投资者情绪指标对未来不同期限股票收益率的影响，并控制股票自身交易特征、行业效应以及时间效应等潜在混淆因素。在此基础上，进一步进行异质性分析，探究情绪影响在不同行业间的差异。同时，还将讨论并尝试处理潜在的内生性问题，并通过替换情绪指标、改变收益率期限等方式进行稳健性检验，

以确保研究结论的可靠性。

二、研究方法

（一）文献研究法。通过系统梳理国内外关于投资者情绪与股市表现关系的研究文献，总结现有研究成果，为本文的研究提供理论基础和方法借鉴。

（二）数理分析法。运用数理统计和计量经济学方法，对文本情感分析结果进行统计描述，构建情绪指标体系，并进行相关性分析初步探究变量间的关系。

（三）实证分析法。基于构建的面板数据，采用固定效应模型等计量方法，控制个体效应和时间效应，分析情绪指标对股票收益的影响。同时，为保证研究结果的稳健性，将进行一系列稳健性检验，包括采用不同的情绪指标进行对比分析，以及通过工具变量法解决可能存在的内生性问题，确保研究结论的可靠性。

第三节 研究创新点

在技术方法层面，本研究采用前沿的 StructBERT 深度学习模型对东方财富网股吧海量中文评论进行精准情感量化，显著提升了对复杂金融文本情绪识别的准确性。在情绪指标体系构建层面，本研究并未局限于单一的情绪得分或正负面评论比例，而是构建了一个多维度的情绪指标体系。在研究视角层面，本研究聚焦于中国 A 股市场这一具有独特投资者结构和市场特征的环境，系统性地检验了基于自然语言处理技术量化的高频投资者情绪对股票短期收益率的预测能力。

第四节 结构安排

本文共分五章，论文的具体组织方式如下：

第一章为引言，阐述研究背景、意义、内容、方法与创新点，并概述全文结构。

第二章为文献综述，系统梳理投资者情绪的理论、测度方法、市场影响研究，以及自然语言处理技术在金融领域的应用现状，评述现有文献并明确本研究的定位。

第三章为研究设计，详细说明数据来源、样本选择、数据预处理流程、基于 StructBERT 的情感量化模型构建、多维度情绪指标体系的设计，并设定核心的面板数据固定效应计量模型及变量。

第四章为实证分析，首先进行描述性统计、单位根与协整检验、相关性与多重共线性检验；然后运用双向固定效应模型实证检验投资者情绪对股票短期收益率的影响，深入分析行业异质性；接着讨论并处理潜在内生性问题；最后通过替换情绪指标、考察不同期限收益率等进行稳健性检验。

第五章为研究结论，总结全文主要发现，提出相应的政策建议与市场启示，指出研究存在的局限性，并对未来研究方向进行展望。

第二章 文献综述

第一节 投资者情绪研究综述

一、投资者情绪的定义与测度

投资者情绪是行为金融学领域的核心概念之一，通常被理解为投资者在缺乏明确基本面信息支撑下，对未来市场走向或资产价值形成的一种系统性偏离理性的信念或预期[5]。这种情绪并非简单的个体心理波动加总，而是在市场互动、信息传播和社会影响下形成的集体性心理倾向，可能导致资产价格偏离其内在价值。De Long 等（1990）提出的噪声交易者模型是早期解释情绪影响的重要理论，指出非理性噪声交易者的存在及其情绪波动会给市场带来风险，即使理性套利者也因面临风险而无法完全消除价格偏离[9]。Brown 和 Cliff（2005）则将投资者情绪定义为对未来收益和风险的主观信念，这种信念可能并非基于事实，而是源于启发式偏差或认知错误[7]。总体而言，学术界普遍认为投资者情绪是独立于基本面信息、能够影响投资者决策进而作用于资产价格的重要非理性因素。

对投资者情绪的测度是该领域研究的基础和难点。早期研究主要依赖间接的市场指标作为代理变量。例如，Lee, Shleifer 和 Thaler（1991）开创性地使用封闭式基金的折价率作为衡量中小投资者情绪的指标，认为折价率的变动反映了散户投资者的乐观或悲观程度[13]。Baker 和 Wurgler（2006, 2007）构建了一个更为广泛接受的综合情绪指数 BW 指数，通过主成分分析法整合了包括首次公开募股 IPO 数量、IPO 首日收益率、股权融资比例、封闭式基金折价率、股息溢价和交易换手率等多个市场层面的指标，旨在捕捉市场整体的投机热度[4,5]。此外，交易量、波动率指数如 VIX、消费者信心指数、看涨看跌期权比率 Put-Call Ratio 等也被广泛用作情绪的代理变量[3,14]。然而，这些间接指标存在明显局限，如指标选择的主观性、数据频率较低，多为月度或季度、可能混杂基本面信息、难以区分不同投资者群体的情绪等问题。

随着互联网技术和大数据分析的发展，基于文本数据的直接情绪测度方法应运而生，为情绪研究提供了新的视角和工具。研究者开始利用新闻报道、社交媒体帖子、在线论坛讨论、搜索引擎查询量等来源的文本信息，通过自然语言处理技术直接量化投资者情绪。例如，Tetlock（2007）pioneeringly 分析了《华尔街日报》特定专栏的悲观情绪词汇频率，发现其能预测市场收益[20]。Antweiler 和 Frank（2004）较早地利用雅虎财经和 Raging Bull 论坛的帖子数量和内容来构建情绪指标，并检验其与市场波动性和交易量的关系[2]。近年来，随着社交媒体的普及，Twitter、微博、StockTwits 等平台成为重要的情绪数据来源[6,22,24]。此外，谷歌趋势（Google Trends）或百度指数（Baidu Index）等搜索数据也被用来捕捉投资者的关注度和潜在情绪[8]。这些基于文本分析的直接测度方法具有高频性、实时性和直接性等优势，能够更细致地刻画情绪的动态变化，成为当前投资者情绪研究的重要方向，也为本研究采用自然语言处理技术分析股吧评论提供了方法论基础。

二、投资者情绪与资本市场表现关系研究

大量实证研究检验了投资者情绪与资本市场表现之间的关系，主要集中在情绪对股票收益率、波动性和交易量的影响。在情绪对股票收益率的影响方面，多数研究发现投资者情绪，尤其是由 BW 指数等综合指标度量的市场整体情绪，对未来股票收益率具有预测能力。Baker 和 Wurgler（2006）发现，高情绪时期，那些难以估值、主观评价空间大、投机性强的股票，如小盘股、年轻公司、高波动性股票、无股息支付股票、亏损公司股票，其当期收益率相对较高，但未来收益率则较低，反之亦然[4]。这支持了情绪驱动的价格偏离及其后的反转效应。Stambaugh, Yu 和 Yuan（2012）进一步发现，投资者情绪对市场异象如盈利公告后漂移、资产增长效应等的强度有调节作用，许多异象在情绪高涨时更为显著[18]。针对中国市场，张强等（2007）的研究区分了个人和机构投资者情绪，发现机构投资者情绪对市场有系统性影响[27]。杨永伟（2018）发现中国权威媒体新闻中的积极情绪能显著推高股价[26]。基于文本分析的情绪指标也显示出对收益率的预测力。Tetlock（2007）发现新闻悲观情绪能预测道琼斯工业平均指数的短期下跌[20]。Antweiler 和 Frank（2004）发现论坛帖子的分歧度与市场波动性正相关，而帖子数量与交易量正相关[2]。Aboody 等（2010）利用隔夜收益率作为情绪代理，发现其对短期收益有正向预测作用，但低情绪时期的股票长期表现更好，再次印证了情绪的短期推动与长期反转特征[1]。

投资者情绪也被证明是影响市场交易活动和波动性的重要因素。Baker 和 Stein（2004）的理论认为，当市场流动性高时，非理性投资者的参与度增加，高换手率本身就反映了高涨的投资者情绪[3]。实证研究也支持了情绪与交易量、波动性之间的正相关关系[5,7]。例如，通过分析社交媒体情绪，Bollen 等（2011）发现 Twitter 上的公众情绪状态可以预测道琼斯指数的涨跌[6]。段江娇等（2017）基于中国股票论坛数据的研究发现，情绪一致性对股票交易量有显著影响，而论坛关注度和情绪水平则影响股票收益[23]。李思龙等（2018）利用东方财富股吧数据发现，投资者互动如发帖量和评论量能增加股东数量、提高流动性并降低信息不对称[25]。

第二节 自然语言处理技术研究进展

一、文本情感分析方法演进

文本情感分析，又称意见挖掘，是自然语言处理 NLP 领域的一个重要分支，旨在自动识别、提取和量化文本中所表达的主观情感、评价、态度和情绪。其技术方法大致经历了三个主要发展阶段：基于情感词典的方法、基于传统机器学习的方法以及基于深度学习的方法。

早期的方法主要依赖于预先构建的情感词典。这些词典包含了大量带有情感极性如正面、负面、中性和强度评分的词语。通过对文本进行分词，匹配词典中的情感词，并根据预设规则如考虑否定词、程度副词等，聚合计算文本的整体情感得分[21]。这种方法的优点是简单直观、无需训练数据，

但在处理复杂语言现象如语境依赖、讽刺、领域特定词汇时效果有限，且词典构建和维护成本高昂。针对金融领域，研究者也构建了专门的金融情感词典，如 Loughran 和 McDonald（2011）构建的金融领域专用词典，显著提升了在财经文本上的分析效果[16]。

随着机器学习技术的发展，基于监督学习的情感分类方法成为主流。这类方法将情感分析视为一个文本分类任务，需要大量标注好的训练数据。首先，需要对文本进行预处理并提取特征，常用的特征包括词袋模型、N-grams、TF-IDF 等。然后，利用这些特征训练分类器，如朴素贝叶斯、支持向量机、逻辑回归等，来预测新文本的情感极性[17]。机器学习方法相比词典法具有更好的适应性和准确性，但其性能高度依赖于特征工程的质量和标注数据的规模与质量。

近年来，深度学习技术，尤其是基于神经网络的模型，在 NLP 领域取得了突破性进展，也极大地推动了文本情感分析技术的发展。卷积神经网络（CNN）能够有效捕捉文本的局部特征[12]，而循环神经网络（RNN）及其变种如长短期记忆网络（LSTM）和门控循环单元（GRU）则擅长处理序列信息，捕捉长距离依赖关系[19]。更具革命性的是基于 Transformer 架构的预训练语言模型，如 BERT[10]、RoBERTa[15]、GPT 系列等。这些模型通过在海量无标注文本上进行预训练，学习到了丰富的语言知识和上下文表示能力，只需在特定任务的少量标注数据上进行微调（Fine-tuning），即可在包括情感分析在内的多种 NLP 任务上达到当前最优的性能。它们能够更好地理解词语在具体语境中的含义、处理否定、转折、讽刺等复杂语言现象。本研究采用的 StructBERT 模型，正是 BERT 架构的改进版本，通过引入词序和句序预测任务增强了对文本结构信息的理解，特别适合处理结构相对松散的网络评论文本。

二、自然语言处理技术在金融领域的应用

自然语言处理技术在金融领域的应用日益广泛，深刻改变着信息获取、风险管理、投资决策和客户服务等多个方面。金融文本数据如新闻、公告、研报、社交媒体讨论、财报、监管文件等具有海量、非结构化、时效性强、专业术语多等特点，为 NLP 技术的应用提供了丰富的场景和巨大的价值潜力。

在信息提取与事件监测方面，NLP 技术可以自动从海量文本中识别和抽取关键信息，如公司实体、财务指标、并购重组、高管变动、盈利预警、政策发布等，将非结构化信息转化为结构化数据，极大提高了信息处理效率和市场反应速度[11]。

在情感分析与舆情监控方面，利用 NLP 技术分析财经新闻、分析师报告、社交媒体和论坛讨论中的情绪倾向，可以构建高频、实时的市场情绪指数或特定资产的情绪指标，为投资决策和风险预警提供重要参考。正如本研究旨在完成的，通过分析股吧评论量化投资者情绪并预测股价变动，是 NLP 在金融情绪分析中的典型应用。研究表明，基于 NLP 量化的新闻或社交媒体情绪确实包含了对未来市场走势的预测信息[6,20]。

此外，NLP 技术还应用于：自动化报告生成；智能投顾与客服；监管科技，用于自动分析监管

文件、识别合规风险；信用风险评估，通过分析借款人相关的文本信息辅助判断其信用状况；以及量化交易策略开发，将文本信息作为信号源纳入交易模型等。

尽管 NLP 在金融领域的应用取得了显著进展，但仍面临挑战，如金融术语的专业性、一词多义、隐晦表达、数据噪声、以及模型可解释性等问题。因此，针对金融领域的特点，持续优化 NLP 模型，如开发金融领域专用的预训练模型、结合金融知识图谱和数据处理方法，是未来研究的重要方向。

第三节 研究评述

综合来看，国内外关于投资者情绪及其市场影响的研究已经取得了丰硕的成果。理论层面，行为金融学为理解情绪在资产定价中的作用提供了坚实的框架；测度层面，从早期间接的市场指标到基于互联网大数据的直接文本量化，情绪测度方法不断进步，日益精细化和实时化；实证层面，大量研究证实了投资者情绪对股票收益率、波动性和交易量等市场表现具有显著影响，并揭示了其影响的复杂性，如短期效应与长期反转、不同市场环境下的差异等。同时，自然语言处理技术的飞速发展，特别是深度学习预训练模型的广泛应用，为直接、准确、大规模地从文本数据中挖掘投资者情绪提供了强大的技术武器，推动了该领域研究的深入。

现有研究仍存在一些值得进一步探讨的空间。其一，在情绪测度方面，尽管基于文本分析的方法优势明显，但如何更有效地处理金融文本的特有挑战如专业术语、数字信息、隐晦表达、噪声数据，以及如何构建更全面、更能反映情绪多维度特征如情绪强度、分歧度、传染性等的指标体系，仍需深入研究。现有研究大多集中于情绪的极性即正面负面，对情绪强度、来源、以及不同情绪成分如恐惧、贪婪、愤怒等的区分和影响研究相对较少。其二，在情绪影响机制方面，虽然情绪对市场表现的预测作用得到广泛证实，但其具体的传导路径和作用机制仍有待厘清。情绪是通过影响投资者注意力、风险偏好、交易行为，还是通过影响信息解读和预期形成来作用于市场的？不同类型投资者如个人与机构的情绪来源和市场影响有何差异？情绪在不同信息环境、不同市场制度下的作用机制是否相同？这些问题需要更微观、更细致的实证证据。其三，关于 NLP 技术在金融情绪分析中的应用，虽然深度学习模型表现优越，但模型的选择、领域适应性、以及模型输出结果的经济学意义解读仍是挑战。例如，本研究采用的 StructBERT 模型效果虽较好，但其在特定金融论坛，如东方财富股吧语境下的最优表现和解释力仍需通过实证检验。此外，如何有效融合多源数据，如文本情绪、交易数据、宏观指标，构建更强大的预测模型，也是未来的重要方向。其四，现有研究多集中于发达市场，尤其是美国市场，对新兴市场如中国 A 股市场的研究虽然逐渐增多，但考虑到中国市场散户占比较高的独特投资者结构、市场波动性以及信息环境，投资者情绪的作用机制可能存在特殊性。利用中国市场的高频、海量互联网文本数据，结合 NLP 技术，深入探究本土投资者情绪的形成、测度及其对股市行情的具体影响，具有重要的理论和现实意义。

第三章 研究设计

本研究的实验环境基于 Windows 10 操作系统，采用 Python 3.11 作为主要开发语言。数据获取与处理主要包括股票评论文本数据和市场交易数据两个方面。

第一节 数据来源与处理

一、股票评论文本数据获取与预处理

（一）数据来源与范围

作为数据来源，我们选择东方财富网股吧作为文本数据采集平台。该平台作为国内最具影响力的股票论坛之一，聚集了大量活跃的投资者，其用户发帖具有较强的时效性和代表性。采集时间跨度为 2025 年 2 月 1 日至 2025 年 2 月 28 日。

本研究依据申万行业分类标准 2021 版选取了 8 个具有代表性的行业作为研究对象，这些行业包括电子、医药生物、银行、房地产、食品饮料、电气设备、计算机和有色金属。选择这些行业的主要考虑是它们具有不同的行业特征和市场表现特点：电子行业代表科技创新导向，对市场情绪较为敏感；医药生物属于防御性行业，受政策影响较大；银行业作为蓝筹稳定型行业，估值相对较低；房地产行业具有较强的周期性，对宏观政策反应明显；食品饮料行业属于消费必需品，表现相对稳定；电气设备行业具有高成长性，政策驱动明显且市场关注度高；计算机行业作为 AI 和数字经济的核心，波动性较大；有色金属行业具有强周期性，对全球经济较为敏感。

在每个行业中，我们选择了最具代表性的股票，具体包括：电子行业 14 只股票（如京东方 A、紫光国微等），医药生物行业 16 只股票（如恒瑞医药、复星医药等），银行业 15 只股票（如工商银行、农业银行等），房地产行业 11 只股票（如保利发展、招商蛇口等），食品饮料行业 15 只股票（如贵州茅台、五粮液等），电气设备行业 11 只股票（如宁德时代、比亚迪等），计算机行业 14 只股票（如海康威视、东方财富等），有色金属行业 15 只股票（如紫金矿业、洛阳钼业等）。总计选取了 111 只具有代表性的股票作为研究样本。

这些股票的选择标准主要基于以下几个方面：第一，市值规模较大，具有较强的行业代表性；第二，交易活跃度高，能够提供足够的市场交易数据；第三，投资者关注度高，能够获取充足的评论数据；第四，上市时间较长，具有稳定的历史表现记录。

（二）爬虫技术实现

本研究采用基于 Selenium 的 Web 爬虫框架对东方财富网股吧进行数据采集。爬虫采用多线程并行处理技术，设置 3 个线程同时运行，以提高数据获取效率。对每只股票，爬虫程序抓取其股吧前

100 页的帖子内容，并以帖子为索引获取所有相关评论，总共。

在具体获取字段方面，爬虫程序分两个层次进行数据采集。第一层是帖子层面，获取的字段包括：帖子标题（post_title）、帖子浏览次数（post_view）、评论数量（comment_num）、帖子链接（post_url）、发帖日期（post_date）、发帖时刻（post_time）以及发帖作者（post_author）。第二层是评论层面，获取的字段包括：所属帖子 ID（post_id）、评论内容（comment_content）、评论点赞数（comment_like）、评论日期（comment_date）、评论时刻（comment_time）以及是否为子评论（sub_comment）。

针对数据存储结构的设计，本研究选择采用 MongoDB 作为数据库系统。MongoDB 作为一种非关系型数据库，具有良好的文档存储能力和查询效率，特别适合处理非结构化的文本数据。在 MongoDB 中，我们为帖子和评论分别建立集合，通过 post_id 字段建立关联，实现帖子与评论的一对多关系存储，便于后续的数据提取和分析处理。

（三）文本预处理

原始数据总共爬取 165985 条帖子信息，270630 条评论信息。为确保数据分析的质量和可靠性，本研究对原始文本数据进行了系统的预处理。首先，基于股票代码到行业代码的映射关系，将所有评论文件进行合并整理，构建以股票代码(stock_code)和日期(date)为索引的数据结构，包含行业代码(board_code)、来源类型(source_type)和评论内容(comment)等关键字段。

在数据清洗环节，采用多步骤处理方案：首先，删除短内容、重复内容和灌水帖子，确保数据的精确度。然后，使用 jieba 分词工具对文本进行分词处理，并通过关键词匹配方式筛除广告内容和机器人发帖。最后，对特殊字符和表情符号进行标准化处理，统一文本格式，为后续的情感分析做好准备。

预处理后帖子数量为 153242 条，评论数量为 225174 条，如表 3-1 所示。

表 3-1 股评示例

stock_code	date	board_code	source_type	comment
600549	2025/2/28	801050	post	可能走弱一段时间。一个月？！
002304	2025/2/28	801120	comment	白酒整体销量却时下滑了，现在酒席用酒比以前少了三分之二。
300223	2025/2/28	801080	comment	还是洗盘，我借助 ds 算力优化了的副图，比以前美观实用多了。只不过是从超强模式变为强势模式。
601012	2025/2/28	801730	comment	5 元不是梦，祖祖辈辈盼回本，子子孙孙盼拉升
603019	2025/2/28	801750	comment	其中一个大事
300059	2025/2/28	801750	comment	那你就等大盘 2400 以下再建仓，估计暂时没得玩了
600111	2025/2/28	801050	comment	不可能高价买自己的股票，肯定往死里压价，18 见。
601818	2025/2/28	801780	comment	防不胜防，上有政策，下有对策，没有做不到，只有想不到。
002415	2025/2/28	801750	post	前天几毛，昨天几毛，今天又是几毛，这搞得人很没有信心了。
600606	2025/2/28	801180	post	绿地死在没有核心竞争力拿地，上海北京杭州成都重庆深圳。
300059	2025/2/28	801750	comment	今天 23.8 加了 30000 股，我也吃面了

300308	2025/2/28	801080	comment	之前那个 Mr 涨涨涨也是发帖称中际即将进入暴力拉升的在我质疑他之后居然把我拉黑了，
600111	2025/2/28	801050	comment	压到 18 都算高
002463	2025/2/28	801080	post	年线可以买，和达子走势一样，没想到这玩意快杀到跌停！真狠
600340	2025/2/28	801180	comment	钢铁直男 周一开宝
600588	2025/2/28	801750	comment	cai 的好啊
601012	2025/2/28	801730	comment	最好几个剩下的最强就能到赚钱，不然股票不涨！
002439	2025/2/28	801750	post	今天时间你给我跌了 20%
600111	2025/2/28	801050	post	一般是压价后再增持的套路
600340	2025/2/28	801180	post	华子危？！？！速回
000002	2025/2/28	801180	comment	吹毛，有本事让他涨啊

（四）描述性统计分析

对预处理后的数据进行多维度统计分析，以揭示数据的基本特征和分布规律。在时间维度上，我们分析了评论的日期分布和时刻分布，发现评论集中在交易日如图 3-1，在市场交易时段评论密度较大如图 3-2，且在重要信息发布时点出现明显的评论峰值。

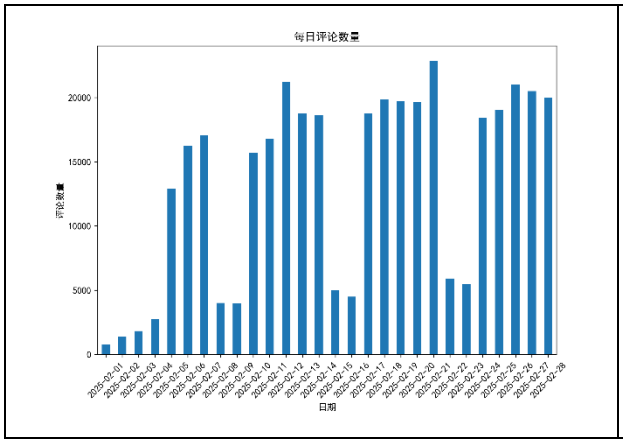


图 3-1 评论日期分布

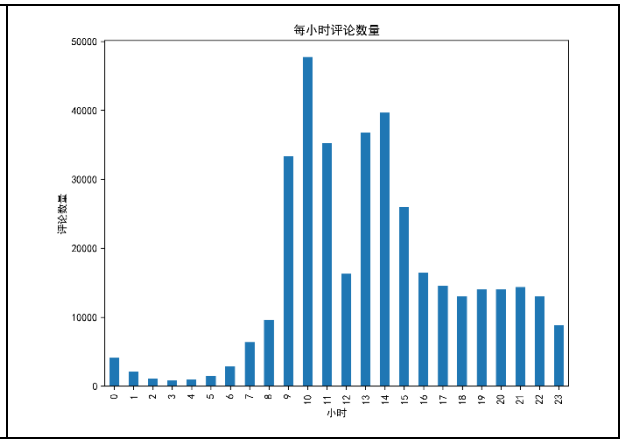


图 3-2 评论时刻分布

文本长度分布分析显示，大部分评论长度集中在 5-30 字之间如图 3-3，这反映了投资者在社交媒体平台上倾向于简明扼要地表达观点。通过高频词统计分析，我们提取了评论中出现频率最高的 50 个关键词如图 3-4，高频词例如“涨停”、“AI”、“比亚迪”；并通过词云图直观展示了整体评论关键词分布以及浪潮信息的高频词的具体表现特征如图 3-5、图 3-6 所示。

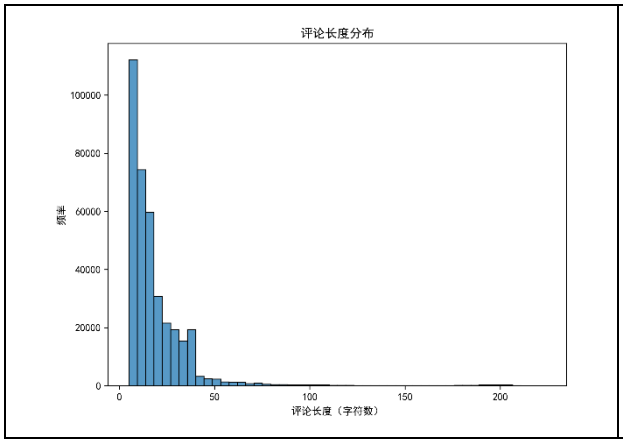


图 3-4 评论长度分布

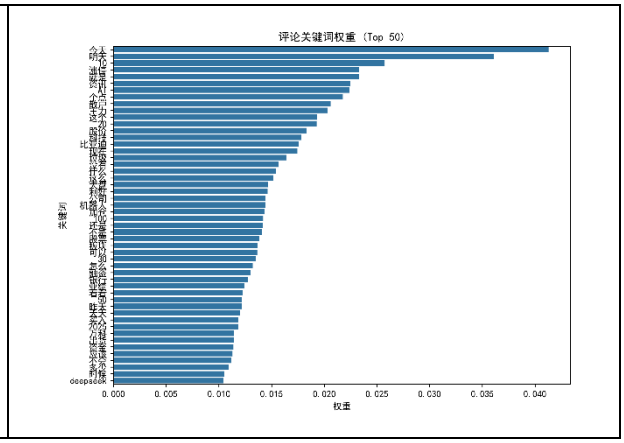


图 3-5 评论关键词权重（top50）

图 3-5 评论关键词词云

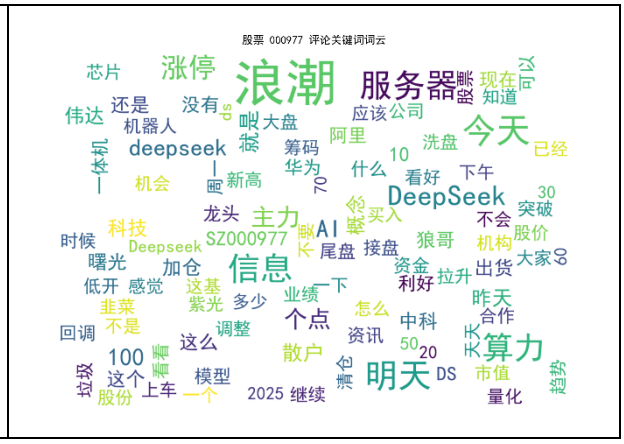


图 3-6 股票 000977 评论关键词词云

在用户活跃度方面，我们统计了评论数量最多的 20 只股票如图 3-7，并重点分析了其中评论量前 5 位股票比亚迪、浪潮信息、中际旭创、万科 A 的每日评论数量变化趋势如图 3-8，这些数据反映了市场关注度的动态变化过程。

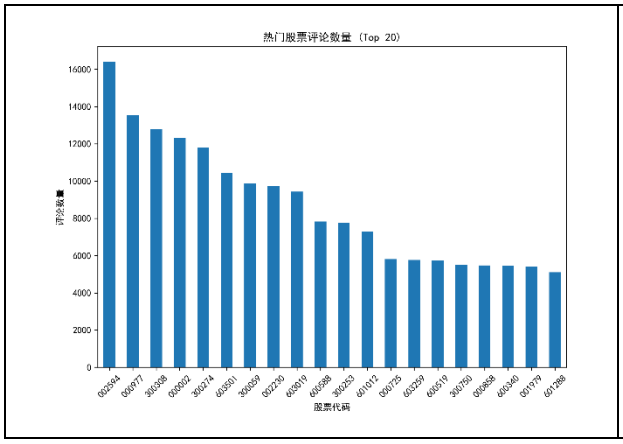


图 3-7 热门股票评论数（top20）

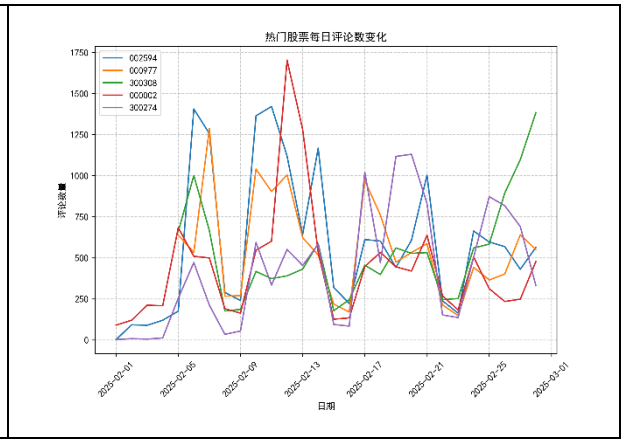


图 3-8 top5 股票每日评论数变化

二、市场交易数据

本研究采用 Python 中的 akshare 金融数据接口获取股票市场交易数据。akshare 是一个开源的财经数据接口包，可以方便地获取股票、基金、期货等金融产品的历史行情数据。考虑到研究需求和数据的可获得性，我们选择以日度频率采集数据，这既能保证数据的充分性，又能避免高频数据中可能存在的噪声问题。

（一）数据获取

在个股数据方面，我们通过 stock_zh_a_hist 接口获取了股票的日度交易数据，包括开盘价、收盘价、最高价、最低价、成交量、成交额、振幅、涨跌幅等指标;在行业指数数据方面，我们使用 index_hist_sw 接口获取了申万行业指数的日度交易数据，包括指数收盘价、开盘价、最高价、最低价、成交量和成交额等指标。考虑到研究需要分析未来收益率，我们将数据收集时间范围设定为 2025 年 2 月 5 日至 2025 年 3 月 7 日的交易日，这样可以保证在研究窗口即 2025 年 2 月 5 日至 2 月 28 日末期也能计算出未来 5 个交易日的收益率。

（二）数据处理

在数据处理环节，我们首先对个股数据进行了收益率指标的计算。通过计算 t+1 日、t+3 日和 t+5 日的股票价格相对于当日收盘价的变化率，分别得到了 forward_ret_1d、forward_ret_3d 和 forward_ret_5d 三个未来收益率指标。对于行业指数数据，我们计算了日度涨跌幅指标，即当日收盘价相对于前一交易日收盘价的变化率。随后，我们根据股票代码与行业代码的对应关系，将个股数据与行业指数数据进行合并，形成了一个包含完整市场交易信息的综合数据集。该数据集以股票代码和日期作为联合索引，包含了个股交易数据、未来收益率指标以及对应的行业指数数据等信息。

（三）描述性统计分析

为了深入理解数据特征，我们对合并后的数据进行了描述性统计分析。通过绘制行业收益率对比图，如图 3-9，揭示了不同行业在研究期间的收益表现差异。行业指数走势图和成交量趋势图，如图 3-10、图 3-11，展示了各行业在时间维度上的价格变动和交易活跃度特征。同时，行业收益率与波动率散点图帮助我们直观理解不同行业的风险-收益特征，如图 3-12。

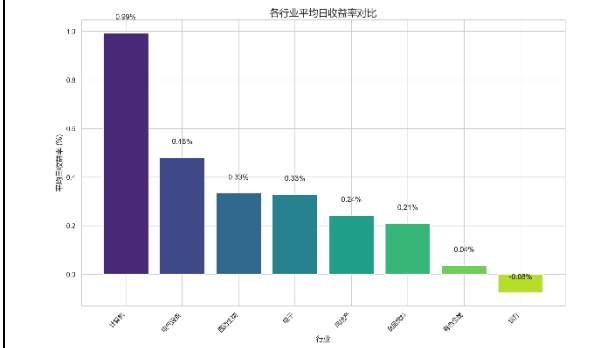


图 3-9 各行业平均日收益率

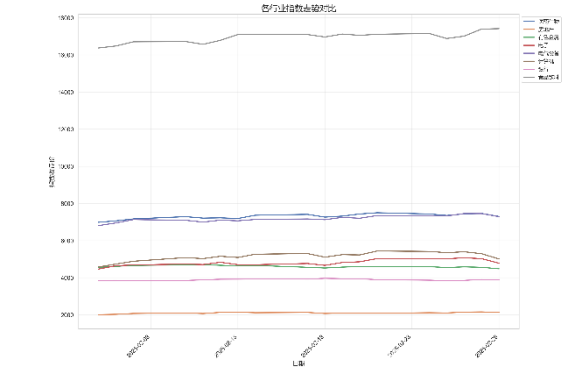


图 3-10 行业指数走势

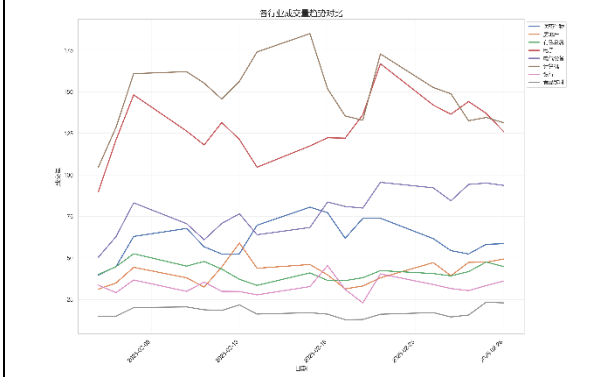


图 3-11 行业成交量趋势

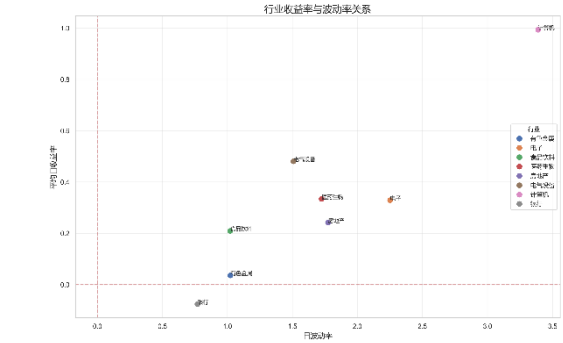


图 3-12 行业收益率与波动率散点图

第二节 情绪分析模型构建

一、基于 StructBERT 的情感分析框架

StructBERT 模型是在 BERT 架构基础上的改进版本，通过引入词序预测和句序预测两个辅助任务，增强了模型对文本结构的理解能力。本研究采用的是通义实验室提供的 structbert_sentiment-classification_chinese-large 版本，该模型基于 StructBERT-large-chinese，包含 24 层 Transformer 编码器，隐藏层维度 1024，共计 3.25 亿参数，相比 base 版本的参数规模扩大了近 3 倍。模型在训练过程中使用了多个领域的情感分析数据集进行微调，包括 BDCI、Dianping、JD Binary 和 Waimai-10k 等，总计约 11.5 万条标注数据。这些数据集涵盖了不同场景下的用户评论，有助于提高模型的泛化能力。在各个测试集上的分类准确率表现优异：BDCI2018 达到 86.26%，Dianping 达到 78.69%，JD Binary 达到 92.06%，Waimai-10k 达到 91.54%。这些性能指标表明模型具有较强的情感分析能力。

二、基于 RoBERTa 的情感分析框架

本研究同时采用了 Fengshenbang 推出的 Erlangshen-RoBERTa-330M 模型作为第二个情感分析框架。该模型基于 chinese-roberta-wwm-ext-large 架构，采用了更大的预训练语料和更优化的训练策略。模型规模为 330M 参数，在 8 个中文情感分析数据集（共计 227,347 个样本）上进行了专门的情感分析任务微调。在主要基准数据集上的测试结果显示，该模型具有出色的情感分类性能：在 ASAP-SENT 数据集上达到 97.9%的准确率，ASAP-ASPECT 数据集上达到 97.51%的准确率，ChnSentiCorp 数据集上达到 96.66%的准确率。相比 110M 参数的基础版本，330M 参数的模型在各项指标上都有显著提升，表明更大的模型规模确实带来了性能的提升。

三、情感分布描述性统计

在情感分析模型构建过程中，本研究采用零样本(Zero-shot)分类策略，直接利用预训练模型对评论文本进行情感二分类，而无需在特定的股票评论数据集上进行额外的监督训练。

表 3-2 StructBERT 股评情感分类示例

stock_code	date	board_code	source_type	comment	positive	negative
300059	2025/2/28	801750	comment	那你就等大盘 2400 以下再建仓，估计暂时没得玩了	0.4041	0.5959
600111	2025/2/28	801050	comment	不可能高价买自己的股票，肯定往死里压价，18 见。	0.0922	0.9078
601818	2025/2/28	801780	comment	防不胜防，上有政策，下有对策，没有做不到，只有想不到。	0.7517	0.2483
002415	2025/2/28	801750	post	前天几毛，昨天几毛，今天又是几毛，这搞得人很没有信心了。	0.1365	0.8635
600606	2025/2/28	801180	post	绿地死在没有核心竞争力拿地，上海北京杭州成都重庆深圳。	0.1201	0.8799
300059	2025/2/28	801750	comment	今天 23.8 加了 30000 股，我也吃面了	0.7313	0.2687
300308	2025/2/28	801080	comment	之前那个 Mr 涨涨涨也是发帖称中际即将进入暴力拉升的在我质疑他之后居然把我拉黑了，	0.0801	0.9199
600111	2025/2/28	801050	comment	压到 18 都算高	0.1940	0.8060
002463	2025/2/28	801080	post	年线可以买，和达子走势一样，没想到这玩意快杀到跌停！真狠	0.4033	0.5967
600340	2025/2/28	801180	comment	钢铁直男 周一开宝	0.9021	0.0979

表 3-3 RoBERTa 股评情感分类示例

stock_code	date	board_code	source_type	comment	positive	negative
300059	2025/2/28	801750	comment	那你就等大盘 2400 以下再建仓，估计暂时没得玩了	0.9994	0.0006
600111	2025/2/28	801050	comment	不可能高价买自己的股票，肯定往死里压价，18 见。	0.0019	0.9981
601818	2025/2/28	801780	comment	防不胜防，上有政策，下有对策，没有做不到，只有想不到。	0.1183	0.8817
002415	2025/2/28	801750	post	前天几毛，昨天几毛，今天又是几毛，这搞得人很没有信心了。	0.0000	1.0000
600606	2025/2/28	801180	post	绿地死在没有核心竞争力拿地，上海北京杭州成都重庆深圳。	0.1369	0.8631
300059	2025/2/28	801750	comment	今天 23.8 加了 30000 股，我也吃面了	1.0000	0.0000
300308	2025/2/28	801080	comment	之前那个 Mr 涨涨涨也是发帖称中际即将进入暴力拉升的在我质疑他之后居然把我拉黑了，	0.0173	0.9827
600111	2025/2/28	801050	comment	压到 18 都算高	0.0000	1.0000
002463	2025/2/28	801080	post	年线可以买，和达子走势一样，没想到这玩意快杀到跌停！真狠	0.0076	0.9924

600340	2025/2/28	801180	comment	钢铁直男 周一开宝	1.0000	0.0000

将两个模型应用于股票评论数据后，我们对情感分析结果进行了系统的整理和统计分析。处理后的数据包含以下关键字段：股票代码(stock_code)、日期(date)、行业代码(board_code)、来源类型(source_type)、评论内容(comment)、情感得分(sentiment_score)、情感极性(sentiment_polarity)、情感强度(sentiment_intensity)、正面概率(positive_prob)和负面概率(negative_prob)。通过对两个模型情感分析结果的统计分析发现，如图 3-13、图 3-14 所示：情感得分呈现两极分化，但均略有左偏，负面评论比例分别为 52.7%和 54.0%，表明投资者评论整体偏向负面；不同极性评论的情感强度分布存在明显差异，负面评论的情感强度普遍高于正面评论，表明投资者在表达负面情绪时往往更加强烈。

在综合考量了模型的理论特性、对本研究特定数据源的潜在适配性、在相关基准任务上的公开评测性能以及初步应用于本数据集的分类效果后，我们认为 StructBERT 模型在本研究场景下展现出良好的适用性。因此，本研究最终决定选取基于 StructBERT 模型的情感分类结果，将其输出的情感得分作为基础，用于构建后续实证分析所需的核心投资者情绪变量指标体系。

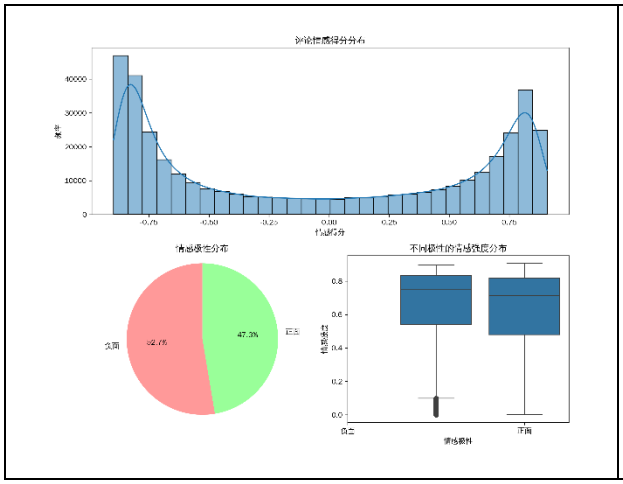


图 3-13 情感分布（StructBERT）

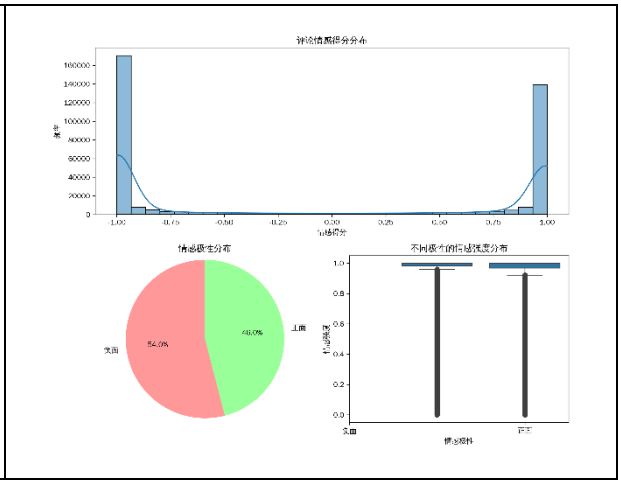


图 3-14 情感分布（RoBERTa）

第三节 情绪指标体系构建

在个股层面，我们首先基于每日评论数据计算了反映股票投资者情绪强度、波动性和一致性的多维指标。具体而言，单个评论的情感得分$sentiment_score = positive_prob - negative_prob$，使用情感得分的算术平均值(avg_sentiment)衡量整体情绪水平，标准差(sentiment_std)反映情绪的离散程度，同时计算正向和负向评论的占比(positive_ratio、negative_ratio)。单个评论的情感强度 $\text{sentiment\_intensity} = |\text{positive\_prob} - \text{negative\_prob}|$，评论的平均情感强度(avg_intensity)和情绪一致性(sentiment_consensus)，$\text{SentimentConsensus}_{it} = 1 - \frac{|\text{PosRatio}_{it} - \text{NegRatio}_{it}|}{\sqrt{2}}$，这些指标共同刻画了投资者情绪的具体特征。此外，为了捕捉情绪的动态变化特征，我们构建了基于不同时间窗口(3 日、5 日、10 日)的移动平均线指标(ma_3d、ma_5d、ma_10d)及其标准差(std_3d、std_5d、std_10d)，用于

衡量情绪的趋势性和波动性。情绪变化率指标(sentiment_change_3d、sentiment_change_5d、sentiment_change_10d)则反映了情绪的变化速度。这些情绪指标与市场交易数据(open, close, high, low等)和收益率数据(forward_ret_1d/3d/5d)整合，形成完整的个股日度面板数据。

在行业层面，我们对个股情绪指标进行加总平均，构建了行业整体情绪特征指标。包括行业平均情绪(ind_avg_sentiment)、情绪标准差(ind_sentiment_std)、行业正负面情绪比例(ind_positive_ratio、ind_negative_ratio)等，并计算了行业层面的情绪动量指标(ind_ma_3d、ind_ma_5d、ind_ma_10d)和变化率指标(ind_sentiment_change_3d、ind_sentiment_change_5d、ind_sentiment_change_10d)。这些行业层面的情绪指标与行业指数数据(idx_close, idx_open 等)整合，形成完整的行业日度面板数据。

第四节 模型设定及变量说明

一、模型设定

由于本研究采集的数据同时包含时间维度（2025 年 2 月 1 日至 2 月 28 日的交易日数据）和横截面维度（111 只不同行业的股票），形成了典型的面板数据结构。面板数据模型相比传统的横截面回归或时间序列分析具有独特的优势：一方面可以控制个体固定效应，有效处理由于股票自身特质如行业属性、公司规模等带来的异质性影响；另一方面能够处理随时间变化的宏观因素如市场整体情绪、政策环境等对所有股票的共同影响。因此，面板数据回归是研究投资者情绪与股票市场表现关系的最佳选择，故采用双向固定效应面板模型：

$$forward\_ret_{it} = \beta_0 + \beta_1 sentiment\_variables_{it} + \beta_2 control\_variables_{it} + \alpha_i + \gamma_t + \varepsilon_{it} \quad (1)$$

其中，α_i表示个体固定效应，用于控制不随时间变化的股票个体特征；γ_t表示时间固定效应，用于控制影响所有股票的时间序列特征。这种设定能够有效降低遗漏变量偏误，提高模型估计的准确性和可靠性。

二、变量说明

在变量选择方面，本研究的被解释变量为股票的未来收益率(forward_ret_1d、forward_ret_3d、forward_ret_5d)，核心解释变量为平均情绪得分(avg_sentiment)。具体而言，本研究构建了一系列情绪变量，包括平均情绪得分(avg_sentiment)、情绪标准差(sentiment_std)、平均情绪强度(avg_intensity)、评论数量(comment_count)以及情绪一致性(sentiment_consensus)等。为了更准确地估计情绪变量的净效应，本研究还加入了一系列控制变量，包括股票交易价格(close)、成交量(volume)、成交额(amount)、振幅(amplitude)、涨跌幅(pct_change)、涨跌额(price_change)、换手率(turnover_rate)等市场交易指标，以及行业指数相关指标(idx_close、idx_volume、idx_amount、idx_pct_change 等)。此外，在稳健性检验中，我们还将考虑使用正面情绪比例(positive_ratio)和移动平均(ma_3d/5d/10d)等替代指标来检验结果的稳健性。

表 3-4 实证数据变量

维度 1：投资者情绪维度				
投资者情绪	avg_sentiment	平均情感得分	sentiment_std	情感得分方差
	avg_intensity	平均情感强度	comment_count	评论数
	sentiment_consensus	情感一致性	positive_ratio	积极比率
维度 2:行情数据维度				
价格数据	close	收盘价	amplitude	振幅
	pct_change	涨跌幅	price_change	价格变动
成交数据	volume	交易量	amount	交易额
	turnover_rate	换手率		

第四章 实证分析

本章将基于前述研究设计，对投资者情绪与股票市场表现之间的关系进行实证分析。通过对样本数据进行描述性统计分析、单位根检验与协整检验、相关性分析、多重共线性检验以及面板数据回归分析，全面探究投资者情绪对股市行情的影响机制。本研究的实证分析在 STATA 统计软件 17.0 版本环境下完成，确保分析过程和结果的科学性与可靠性。

第一节 描述性统计

为了全面了解研究样本的基本特性及变量分布情况，首先对关键变量进行描述性统计分析。本研究的样本包含 111 只股票在 2025 年 2 月份交易日的日度数据，共计 1991 个有效观测值，涵盖 8 个行业板块。表 4-1 列示了主要变量的描述性统计结果，包括均值、标准差、最小值和最大值。

对于情绪指标，平均情绪得分(avg_sentiment)均值为-0.077，标准差为 0.15，最小值为-0.783，最大值为 0.512，表明总体上样本期内投资者情绪略偏负面，但不同股票之间情绪差异较大。情绪标准差(sentiment_std)均值为 0.664，表明投资者对同一股票的情绪波动相对稳定。积极情绪比例(positive_ratio)均值为 0.456，消极情绪比例(negative_ratio)均值为 0.544，进一步确认了样本期内投资者情绪整体略偏消极的特点。情绪一致性(sentiment_consensus)均值达 0.874，表明投资者对同一股票的情绪评价相对一致。评论数量(comment_count)均值为 171.956，标准差高达 189.799，最小值为 1，最大值为 1732，反映出不同股票受关注度差异显著。

对于情绪动量指标，不同窗口期(3 日、5 日、10 日)的移动平均情绪(ma_3d、ma_5d、ma_10d)均值分别为-0.077、-0.076 和-0.073，标准差分别为 0.119、0.112 和 0.107，表明随着时间窗口扩大，情绪波动趋于平稳。情绪变化率(sentiment_change_3d、sentiment_change_5d、sentiment_change_10d)均值分别为-0.752、-0.906 和-0.865，波动较大，说明短期内投资者情绪可能存在显著变化。

对于市场交易指标，样本股票的平均收盘价(close)为 59.731 元，平均成交量(volume)为 681,948.85 手，平均成交额(amount)为 15.92 亿元。振幅(amplitude)均值为 3.221%，涨跌幅(pct_change)均值为 0.308%，换手率(turnover_rate)均值为 2.168%，表明样本期内市场交易相对活跃。行业指数收盘价(idx_close)均值为 6675.254 点，行业指数成交量(idx_volume)均值为 67.779 亿股，行业指数成交额(idx_amount)均值为 1141.184 亿元，行业指数涨跌幅(idx_pct_change)均值为 0.261%。

对于被解释变量，未来一日收益率(forward_ret_1d)均值为 0.003，标准差为 0.026，最小值为-0.093，最大值为 0.2；未来三日收益率(forward_ret_3d)均值为 0.007，标准差为 0.043；未来五日收益率(forward_ret_5d)均值为 0.011，标准差为 0.055。这表明随着预测期限的延长，平均收益率和波动性均有所增加，符合金融市场的一般规律。

总体而言，描述性统计结果显示样本期内投资者情绪略微偏向负面，但不同股票和行业之间存在明显差异；市场整体表现相对活跃；短期内股票收益率虽然均值为正，但波动性随着时间窗口的扩大而增加。

表4-1 描述性统计

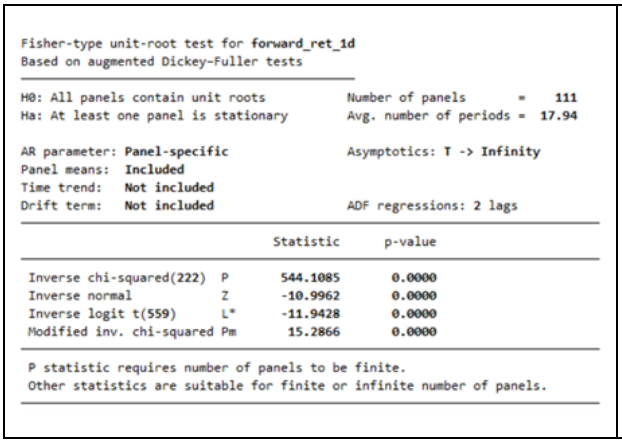
Variable	Obs	Mean	Std. Dev.	Min	Max
stock code	1991	383270.28	261941.11	2	688111
time	1991	9.522	5.184	1	18
avg sentiment	1991	-.077	.15	-.783	.512
sentiment std	1989	.664	.04	.121	.866
positive ratio	1991	.456	.103	0	1
avg intensity	1991	.638	.041	.131	.791
comment count	1991	171.956	189.799	1	1732
sentiment consensus	1991	.874	.096	.293	1
ma 3d	1991	-.077	.119	-.715	.354
std 3d	1880	.099	.068	0	.501
sentiment change 3d	1658	-.752	54.304	-1212.857	1360.312
ma 5d	1991	-.076	.112	-.715	.298
std 5d	1880	.109	.06	0	.501
sentiment change 5d	1436	-.906	53.454	-1686.305	1078.928
ma 10d	1991	-.073	.107	-.715	.298
std 10d	1880	.117	.057	0	.501
sentiment change 10d	881	-.865	11.819	-242.354	63.221
close	1991	59.731	152.232	1.82	1500.79
volume	1991	681948.85	927709.84	14744	7552709
amount	1991	1.592e+09	2.088e+09	31106281	1.819e+10
amplitude	1991	3.221	2.319	.47	20.56
pct change	1991	.308	2.661	-14.43	20.06
price change	1991	.246	3.155	-31.19	57.11
turnover rate	1991	2.168	2.626	.07	25.33
board code	1991	801343.74	310.578	801050	801780
idx close	1991	6675.254	4344.529	2013.39	17422.311
idx volume	1991	67.779	45.595	12.846	185.051
idx amount	1991	1141.184	1107.583	132.632	3839.181
idx pct change	1991	.261	1.561	-5.356	3.965
forward ret 1d	1991	.003	.026	-.093	.2
forward ret 3d	1991	.007	.043	-.167	.391
forward ret 5d	1991	.011	.055	-.177	.454

第二节 单位根检验与协整检验

一、单位根检验

为确保面板数据回归分析的结果可靠，首先对关键变量进行单位根检验，判断序列是否平稳。

本研究采用两种检验方法：ADF-Fisher 检验和 IPS(Im-Pesaran-Shin)检验。对于未来一日收益率(forward_ret_1d)和平均情绪得分(avg_sentiment)两个核心变量，Fisher 检验结果显示如图 4-1、4-2、4-3、4-4 所示，在滞后阶数为 2 的情况下，两个变量均拒绝了存在单位根的原假设，证明这些变量是平稳序列。同样，IPS 检验在最优滞后阶数(AIC 准则下为 3)的情况下，也拒绝了单位根假设，进一步确认了变量的平稳性。这表明这些变量可以直接用于后续回归分析，无需进行差分处理。



Fisher-type unit-root test for forward_ret_1d
Based on augmented Dickey-Fuller tests

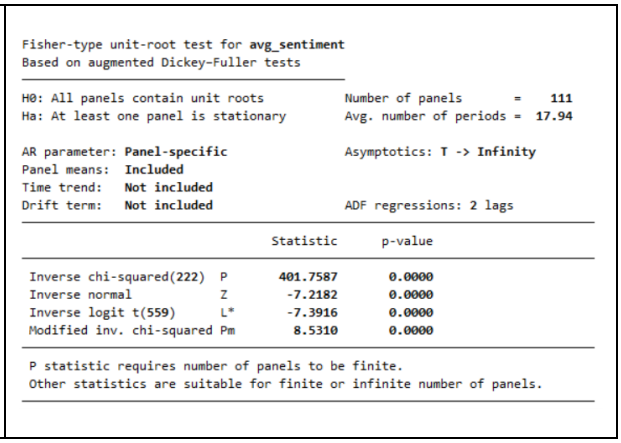
H0: All panels contain unit roots — Number of panels = 111
Ha: At least one panel is stationary — Avg. number of periods = 17.94

AR parameter: Panel-specific — Asymptotics: T -> Infinity
Panel means: Included
Time trend: Not included
Drift term: Not included — ADF regressions: 2 lags

		Statistic	p-value
Inverse chi-squared(222)	P	544.1085	0.0000
Inverse normal	Z	-10.9962	0.0000
Inverse logit t(559)	L*	-11.9428	0.0000
Modified inv. chi-squared	Pm	15.2866	0.0000

P statistic requires number of panels to be finite.
Other statistics are suitable for finite or infinite number of panels.

图 4-1 ADF-Fisher 检验（forward_ret_1d）



Fisher-type unit-root test for avg_sentiment
Based on augmented Dickey-Fuller tests

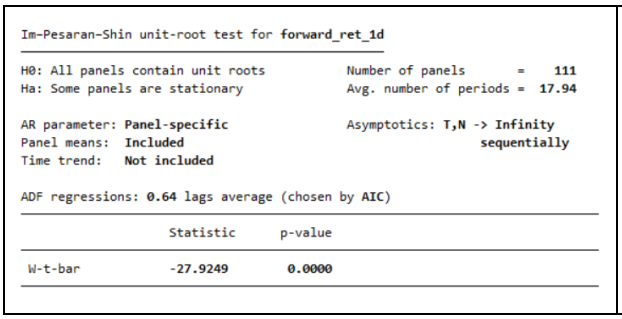
H0: All panels contain unit roots — Number of panels = 111
Ha: At least one panel is stationary — Avg. number of periods = 17.94

AR parameter: Panel-specific — Asymptotics: T -> Infinity
Panel means: Included
Time trend: Not included
Drift term: Not included — ADF regressions: 2 lags

		Statistic	p-value
Inverse chi-squared(222)	P	401.7587	0.0000
Inverse normal	Z	-7.2182	0.0000
Inverse logit t(559)	L*	-7.3916	0.0000
Modified inv. chi-squared	Pm	8.5310	0.0000

P statistic requires number of panels to be finite.
Other statistics are suitable for finite or infinite number of panels.

图 4-2 ADF-Fisher 检验（avg_sentiment）



Im-Pesaran-Shin unit-root test for forward_ret_1d

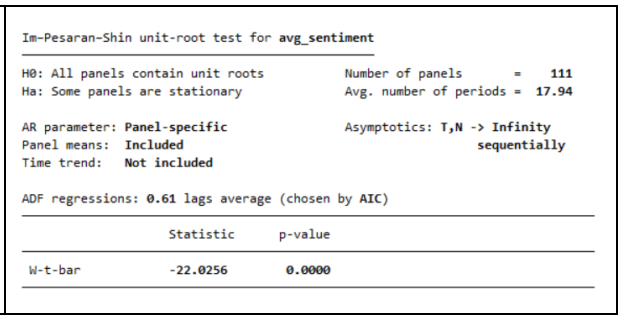
H0: All panels contain unit roots — Number of panels = 111
Ha: Some panels are stationary — Avg. number of periods = 17.94

AR parameter: Panel-specific — Asymptotics: T,N -> Infinity sequentially
Panel means: Included
Time trend: Not included

ADF regressions: 0.64 lags average (chosen by AIC)

	Statistic	p-value
W-t-bar	-27.9249	0.0000

图 4-3 IPS 检验（forward_ret_1d）



Im-Pesaran-Shin unit-root test for avg_sentiment

H0: All panels contain unit roots — Number of panels = 111
Ha: Some panels are stationary — Avg. number of periods = 17.94

AR parameter: Panel-specific — Asymptotics: T,N -> Infinity sequentially
Panel means: Included
Time trend: Not included

ADF regressions: 0.61 lags average (chosen by AIC)

	Statistic	p-value
W-t-bar	-22.0256	0.0000

图 4-4 IPS 检验（avg_sentiment）

二、协整检验

在确认变量的平稳性后，进一步采用 Kao 协整检验方法检验情绪变量与股票收益率之间是否存在长期均衡关系。检验结果如图 4-5 表明，forward_ret_1d 与情绪变量组(avg_sentiment、sentiment_std、avg_intensity、comment_count、sentiment_consensus)之间存在协整关系，拒绝了无协整关系的原假设。这一结果表明，虽然短期内投资者情绪与股票收益率之间可能存在波动，但长期内两者保持稳定的均衡关系。

Im-Pesaran-Shin unit-root test for forward_ret_1d

H0: All panels contain unit roots　　Number of panels = 111
Ha: Some panels are stationary　　Avg. number of periods = 17.94

AR parameter: Panel-specific　　Asymptotics: T,N -> Infinity
Panel means: Included　　sequentially
Time trend: Not included

ADF regressions: 0.64 lags average (chosen by AIC)

	Statistic	p-value
W-t-bar	-27.9249	0.0000

图 4-5 Kao 协整检验

第三节 相关性分析与多重共线性检验

一、相关性分析

在进行回归分析之前，对所有变量进行两两相关性分析，了解变量之间的关联程度。图 4-6 展示了主要变量之间的 Pearson 相关系数及显著性水平。

分析结果显示，未来一日收益率(forward_ret_1d)与平均情绪得分(avg_sentiment)的相关系数为 0.040，虽然在 10%水平上不完全显著(p=0.078)，但仍表明两者之间存在一定的正相关关系。此外，forward_ret_1d 与振幅(amplitude)的相关系数为 0.062，在 1%的水平上显著，与换手率(turnover_rate)的相关系数为 0.048，在 5%的水平上显著。这表明在简单相关关系中，投资者情绪、市场活跃度与未来收益率之间存在正相关关系。

情绪变量内部相关性方面，avg_sentiment 与 sentiment_std 呈显著正相关(相关系数 0.230)，与 avg_intensity 呈显著负相关(相关系数-0.214)，与 sentiment_consensus 呈显著正相关(相关系数 0.420)。情绪变量与市场交易指标之间也存在多项显著相关关系，如 avg_sentiment 与成交额(amount)的相关系数为 0.294，与振幅(amplitude)的相关系数为 0.244，与涨跌幅(pct_change)的相关系数为 0.317，均在 1%的水平上显著。这些关系表明，投资者情绪与市场交易活动密切相关。

总体而言，相关性分析初步支持了投资者情绪与股票市场表现之间存在关联的假设，但简单相关关系无法揭示真实的因果关系和净效应，需要通过回归分析进一步探究。

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
(1) forward_ret_1d	1.000												
(2) avg_sentiment	0.040	1.000											
	(0.078)												
(3) sentiment_std	0.003	0.230*	1.000										
	(0.889)	(0.000)											
(4) avg_intensity	0.017	-0.214*	0.550*	1.000									
	(0.442)	(0.000)	(0.000)										
(5) comment_count	-0.014	0.115*	0.102*	0.080*	1.000								
	(0.538)	(0.000)	(0.000)	(0.000)									
(6) sentiment_cons~s	-0.015	0.420*	0.542*	-0.105*	0.152*	1.000							
	(0.496)	(0.000)	(0.000)	(0.000)	(0.000)								
(7) close	0.007	0.163*	0.053*	0.025	0.145*	0.063*	1.000						
	(0.759)	(0.000)	(0.018)	(0.268)	(0.000)	(0.005)							
(8) volume	-0.042	0.108*	0.016	-0.072*	0.404*	0.185*	-0.176*	1.000					
	(0.059)	(0.000)	(0.475)	(0.001)	(0.000)	(0.000)	(0.000)						
(9) amount	-0.025	0.294*	0.107*	0.028	0.748*	0.192*	0.333*	0.337*	1.000				
	(0.259)	(0.000)	(0.000)	(0.205)	(0.000)	(0.000)	(0.000)	(0.000)					
(10) amplitude	0.062*	0.244*	0.051*	0.000	0.370*	0.125*	0.018	0.137*	0.358*	1.000			
	(0.006)	(0.000)	(0.022)	(0.991)	(0.000)	(0.000)	(0.418)	(0.000)	(0.000)				
(11) pct_change	0.018	0.317*	0.043	-0.050*	0.115*	0.125*	0.027	0.063*	0.141*	0.435*	1.000		
	(0.417)	(0.000)	(0.058)	(0.026)	(0.000)	(0.000)	(0.231)	(0.005)	(0.000)	(0.000)			
(12) price_change	0.008	0.180*	0.016	-0.003	0.106*	0.031	0.174*	-0.019	0.187*	0.218*	0.547*	1.000	
	(0.736)	(0.000)	(0.470)	(0.904)	(0.000)	(0.172)	(0.000)	(0.395)	(0.000)	(0.000)	(0.000)		
(13) turnover_rate	0.048*	0.176*	0.095*	0.030	0.426*	0.168*	-0.043	0.188*	0.445*	0.734*	0.250*	0.072*	1.000
	(0.032)	(0.000)	(0.000)	(0.178)	(0.000)	(0.000)	(0.055)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	

*** p<0.01, ** p<0.05, * p<0.1

图 4-6 相关性分析

二、多重共线性检验

为避免回归模型中的多重共线性问题，对模型中的解释变量进行方差膨胀因子(VIF)检验。检验结果如表 4-2 所示，所有变量的 VIF 值均小于 5，最大 VIF 值为情绪标准差(sentiment_std)的 4.087，平均 VIF 值为 2.396。这表明模型中不存在严重的多重共线性问题，回归估计结果将是可靠的。

表4-2 方差膨胀因子VIF

	VIF	1/VIF
sentiment std	4.087	.245
amount	3.103	.322
avg intensity	3.088	.324
sentiment consensus	3.014	.332
comment count	2.625	.381
amplitude	2.612	.383
turnover rate	2.56	.391
pct change	1.805	.554
avg sentiment	1.596	.627
price change	1.529	.654
volume	1.373	.729
close	1.357	.737
Mean VIF	2.396	.

第四节 面板数据回归分析

一、面板数据模型构建

(1) 混合 OLS 模型 (Pooled OLS Model)

混合 OLS 模型是最基础的面板数据分析方法，它将所有观测视为独立，忽略了面板数据的时间和个体结构。其基本形式为：

$$forward\_ret_{it} = \beta_0 + \beta_1 sentiment\_variables_{it} + \beta_2 control\_variables_{it} + \varepsilon_{it} \quad (2)$$

其中，i表示第i个股票，t表示第t个交易日，ε_{it}是随机误差项。该模型假设所有观测是独立同分布的，且误差项与解释变量不相关。

(2) 随机效应模型 (Random Effects Model)

随机效应模型考虑了个体异质性，将误差项分解为两部分：一个是股票特定的随机变量，另一个是纯随机误差。其基本形式为：

$$forward\_ret_{it} = \beta_0 + \beta_1 sentiment\_variables_{it} + \beta_2 control\_variables_{it} + \mu_i + \varepsilon_{it} \quad (3)$$

其中，μ_i是随机个体效应，假设$\mu_i \sim IID(0, \sigma_\mu^2)$，且与$\varepsilon_{it}$不相关。随机效应模型假设个体效应与解释变量不相关，因此可以将个体效应视为误差项的一部分。

(3) 固定效应模型 (Fixed Effects Model)

固定效应模型允许每个股票有其特定的截距项，从而控制不随时间变化的个体特征。其基本形式为：

$$forward\_ret_{it} = \beta_0 + \beta_1 sentiment\_variables_{it} + \beta_2 control\_variables_{it} + \alpha_i + \varepsilon_{it} \quad (4)$$

其中，α_i是股票i的固定效应，可以通过引入股票虚拟变量来估计。与随机效应模型不同，固定效应模型允许个体效应与解释变量相关，因此能更好地控制不可观测的个体异质性。

(4) 双向固定效应模型 (Two-way Fixed Effects Model)

双向固定效应模型同时控制了个体固定效应和时间固定效应，是本研究的核心模型。其基本形式为：

$$forward\_ret_{it} = \beta_0 + \beta_1 sentiment\_variables_{it} + \beta_2 control\_variables_{it} + \alpha_i + \gamma_t + \varepsilon_{it} \quad (5)$$

其中，α_i是股票i的固定效应，γ_t是时间t的固定效应。这一模型能够同时控制不随时间变化的个体特征和不随个体变化的时间特征，如宏观经济环境或市场整体情绪的变化。

二、Hausman 检验与模型选择

为确定最适合本研究数据的模型，我们进行了 F 检验和 Hausman 检验。这些检验可以帮助我们在混合 OLS、随机效应和固定效应模型之间做出科学选择。

F 检验用于比较固定效应模型与混合 OLS 模型，其原假设为：所有个体的截距项相同，即无个

体固定效应。F 统计量的计算公式为：

$$F = \frac{(RSS_{pooled} - RSS_{FE})/(N-1)}{RSS_{FE}/(NT-N-K)} \tag{6}$$

根据估计结果如表 4-3，F 检验的 p 值接近于 0，强烈拒绝了原假设，表明数据中存在显著的个体异质性，混合 OLS 模型不适合本研究。这一结果意味着不同股票之间确实存在系统性差异，需要采用固定效应或随机效应模型来控制这种异质性。

表4-3 固定效应模型

forward_ret_1d	Coef.	St.Err.	t-value	p-value	[95%Conf	Interval]	Sig
avg_sentiment	.016	.005	2.93	.003	.005	.026	***
sentiment_std	-.02	.029	-0.68	.496	-.076	.037	
avg_intensity	.017	.027	0.62	.537	-.036	.069	
comment_count	0	0	-1.92	.055	0	0	*
sentimentcons~s	-.012	.011	-1.08	.282	-.033	.01	
close	-.001	0	-10.54	0	-.001	-.001	***
volume	0	0	-1.61	.108	0	0	
amount	0	0	-1.67	.094	0	0	*
amplitude	0	.001	0.14	.889	-.001	.001	
pct_change	0	0	-1.34	.181	-.001	0	
price_change	0	0	1.62	.106	0	.001	
turnover_rate	0	.001	-0.44	.659	-.002	.001	
Constant	.093	.015	6.27	0	.064	.122	***

Mean dependent var	0.003	SD dependent var	0.026
R-squared	0.084	Number of obs	1989
F-test	14.186	Prob > F	0.000
Akaike crit. (AIC)	-9199.589	Bayesian crit. (BIC)	-9132.445

*** *p*<.01, ** *p*<.05, * *p*<.1

Hausman 检验用于比较固定效应模型与随机效应模型，其原假设为：随机效应与解释变量不相关。Hausman 统计量的计算公式为：

$$H = (\widehat{\beta_{FE}} - \widehat{\beta_{RE}})' [Var(\widehat{\beta_{FE}}) - Var(\widehat{\beta_{RE}})]^{-1} (\widehat{\beta_{FE}} - \widehat{\beta_{RE}}) \tag{7}$$

Hausman 检验的卡方统计量如表 4-4 为 144.455，p 值为 0，明确拒绝了原假设，表明随机效应估计量不一致，个体效应与解释变量相关。因此，固定效应模型是更为合适的选择。这一结果在经济学意义上表明，股票的固有特性，如所属行业、公司规模、治理结构等，可能与投资者情绪和其他解释变量相关，采用固定效应模型可以有效控制这种相关性带来的内生性问题。

表4-4 Hausman检验

	Coef.
Chi-square test value	144.455
P-value	0

基于上述检验结果，我们确定采用固定效应模型作为主要分析工具，并考虑加入时间固定效应构建双向固定效应模型，以进一步控制随时间变化的共同因素。

三、固定效应回归分析

表 4-5 展示了四种不同设定的面板回归模型结果。模型 1 为仅控制个体固定效应的模型，模型 2 为双向固定效应模型，即同时控制了个体效应和时间效应，模型 3 为加入了行业指数变量的仅控制个体固定效应的模型，模型 4 为加入了行业指数变量的双向固定效应模型。

从核心解释变量 avg_sentiment 的系数来看，四个模型中该系数均为正且在 1%水平上显著，数值分别为 0.0158、0.0144、0.0171 和 0.0170。这表明，在控制其他因素后，投资者情绪每提高一个单位，未来一日股票收益率平均提高约 1.4-1.7 个百分点。这一结果有力地支持了投资者情绪对股票短期收益具有预测作用的假设。在加入时间固定效应后，模型 2 中 avg_sentiment 的系数略有下降但显著性保持不变，表明控制时间异质性后情绪的影响仍然稳健。

其他情绪变量方面，sentiment_std、avg_intensity 和 sentiment_consensus 在四个模型中均不显著，表明情绪的波动性、强度和一致性对股票收益率的影响不如平均情绪得分显著。评论数量(comment_count)在模型 1 中呈现负相关且在 10%水平上显著，但在加入固定效应后变得不显著，表明评论数量的影响可能被股票固有特性所解释。

在控制变量中，股票收盘价(close)在所有模型中均显示显著的负相关关系，系数在-0.0011 至-0.0007 之间，表明高价股在样本期内平均收益率较低。交易量(volume)在加入行业指数变量的模型中变得显著为负，表明交易活跃度与未来收益可能存在负相关关系。振幅(amplitude)在加入固定效应后变得显著为负，系数为-0.0013，表明价格波动较大的股票在短期内可能面临收益率下降。

对比不同模型的拟合优度(R^2)，可以发现模型 4 的 R^2 值最高，达到 0.276，明显高于其他模型，表明同时控制个体和时间效应和加入行业指数变量可以更好地解释股票收益率的变化。F 统计量在所有模型中均显著，表明模型整体上具有良好的解释力。

此外，在加入行业指数变量的模型中，idx_close 呈现显著的负相关关系，idx_volume 呈现显著的正相关关系，idx_amount 呈现显著的负相关关系，这表明行业层面的行情对个股收益率也有显著影响。在双向固定效应模型中，idx_pct_change 转为显著为正，表明在控制个体和时间效应后，行业指数涨幅与个股未来收益率呈现正相关关系。

固定效应回归分析结果表明，投资者情绪，特别是平均情绪得分，对股票未来一日收益率具有显著的正向预测作用，且这一效应在控制个体异质性和时间效应后依然稳健。这一发现与行为金融

学理论一致，支持了投资者情绪作为一种信息传导机制影响股票价格的观点。

表 4-5 固定效应回归

	(1)	(2)	(3)	(4)
	Model 1	Model 2	Model 3	Model 4
avg_sentiment	0.0158***	0.0144***	0.0171***	0.0170***
	(2.9259)	(2.9017)	(3.2643)	(3.4418)
sentiment_std	-0.0195	-0.0048	-0.0147	-0.0054
	(-0.6814)	(-0.1843)	(-0.5300)	(-0.2114)
avg_intensity	0.0166	0.0087	0.0180	0.0127
	(0.6178)	(0.3571)	(0.6943)	(0.5278)
comment_count	-0.0000*	-0.0000	-0.0000	-0.0000
	(-1.9189)	(-1.2387)	(-0.9530)	(-1.3723)
sentiment_consensus	-0.0116	-0.0082	-0.0113	-0.0110
	(-1.0754)	(-0.8420)	(-1.0878)	(-1.1312)
close	-0.0011***	-0.0008***	-0.0007***	-0.0007***
	(-10.5442)	(-8.0542)	(-6.0112)	(-6.6778)
volume	-0.0000	-0.0000	-0.0000*	-0.0000*
	(-1.6062)	(-1.1581)	(-1.7362)	(-1.8386)
amount	-0.0000*	-0.0000*	-0.0000	-0.0000
	(-1.6741)	(-1.7720)	(-1.2033)	(-1.1417)
amplitude	0.0001	-0.0013***	-0.0008	-0.0011**
	(0.1401)	(-2.6281)	(-1.5461)	(-2.3186)
pct_change	-0.0004	0.0001	-0.0002	-0.0002
	(-1.3371)	(0.4898)	(-0.6441)	(-0.7530)
price_change	0.0004	0.0003	0.0003	0.0002
	(1.6172)	(1.3935)	(1.2789)	(1.2292)
turnover_rate	-0.0004	0.0008	0.0005	0.0012
	(-0.4419)	(1.0590)	(0.6059)	(1.5605)
idx_close			-0.0000***	-0.0000***
			(-8.6649)	(-4.2493)
idx_volume			0.0003**	0.0004***
			(2.1592)	(2.7012)
idx_amount			-0.0000**	-0.0000***
			(-2.4856)	(-3.3486)
idx_pct_change			-0.0005	0.0012**
			(-1.0412)	(2.0890)
_cons	0.0928***	0.0778***	0.3048***	0.2067***
	(6.2669)	(5.7002)	(10.4279)	(6.1150)
Entity Effects	Yes	Yes	Yes	Yes
Time Effects		Yes		Yes
N	1989	1989	1989	1989

R^2	0.084	0.257	0.149	0.276
F	14.186	22.052	20.416	21.353

***p<0.01，**p<0.05，*p<0.10

第五节 异质性分析

一、分行业回归分析

为探究投资者情绪对不同行业股票收益率的影响是否存在差异性，我们对八个行业分别进行了固定效应回归分析。表 4-6 展示了分行业回归的详细结果，第 1 至 8 列分别对应不同行业。

从核心解释变量 avg_sentiment 的系数来看，行业间存在明显差异。计算机行业的系数最大，为 0.056，且在 5%水平上显著；电子、食品饮料和医药生物行业的系数分别为 0.034、0.016 和 0.021，在 10%水平上显著；而电子、食品饮料、电气设备和有色金属行业的系数则不显著。结果表明，投资者情绪对不同行业股票收益率的影响存在明显的异质性，对于计算机、电子、食品饮料和医药生物等行业，情绪的预测作用更为显著。

表 4-6 分行业回归

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	有色金属	电子	食品饮料	医药生物	房地产	电力设备	计算机	银行
avg_sentiment	0.013	0.034*	0.016*	0.021*	-0.006	0.009	0.056**	-0.004
	(1.298)	(1.665)	(1.913)	(1.814)	(-0.406)	(0.415)	(2.389)	(-1.366)
sentiment_std	-0.029	-0.124	-0.005	0.016	-0.052	-0.184	0.208	0.046**
	(-0.318)	(-0.814)	(-0.091)	(0.514)	(-0.895)	(-1.092)	(0.940)	(2.225)
avg_intensity	-0.014	0.165	0.011	-0.043	0.055	0.125	-0.131	-0.040**
	(-0.176)	(1.263)	(0.237)	(-1.278)	(1.024)	(0.892)	(-0.697)	(-2.190)
comment_count	0.000	0.000	0.000	0.000	-0.000	-0.000	-0.000	0.000
	(0.004)	(1.158)	(0.765)	(0.061)	(-0.081)	(-0.183)	(-0.941)	(0.907)
sentiment_consensus	-0.021	-0.045	-0.004	-0.027	0.064**	0.057	-0.081	-0.015**
	(-0.901)	(-1.030)	(-0.213)	(-1.513)	(2.577)	(1.278)	(-1.412)	(-2.095)
close	-0.008***	-0.001**	-0.000	-0.003***	-0.025**	-0.001***	-0.001***	-0.006***
	(-2.886)	(-2.536)	(-0.377)	(-5.636)	(-2.508)	(-3.281)	(-3.856)	(-3.126)
volume	-0.000	0.000	-0.000	-0.000**	0.000	-0.000	-0.000	0.000
	(-1.310)	(1.073)	(-0.594)	(-2.212)	(1.187)	(-1.126)	(-0.742)	(0.201)
amount	0.000**	-0.000*	-0.000	0.000*	-0.000	-0.000	0.000	-0.000
	(2.075)	(-1.829)	(-0.310)	(1.893)	(-1.176)	(-1.514)	(0.141)	(-0.489)
amplitude	-0.000	-0.001	0.000	0.002	-0.000	-0.003*	-0.001	-0.003***
	(-0.265)	(-0.801)	(0.190)	(1.084)	(-0.105)	(-1.781)	(-0.449)	(-2.809)
pct_change	0.001	0.001	-0.001	0.000	-0.002*	-0.001	-0.001	-0.001
	(0.850)	(1.219)	(-1.020)	(0.162)	(-1.846)	(-0.570)	(-1.032)	(-1.105)
price_change	-0.019***	-0.001	-0.000	-0.002	0.007	0.000	0.000	0.004

	(-3.349)	(-0.931)	(-0.005)	(-1.510)	(0.528)	(0.686)	(0.639)	(0.933)
turnover_rate	0.001	0.003	-0.005*	-0.007***	-0.004	0.012***	-0.000	0.007*
	(0.479)	(1.207)	(-1.689)	(-3.204)	(-1.640)	(7.153)	(-0.207)	(1.954)
_cons	0.202***	0.100**	0.017	0.235***	0.118	0.073	0.159**	0.067***
	(4.000)	(2.117)	(0.632)	(5.748)	(1.572)	(1.258)	(2.099)	(3.571)
N	270	252	270	286	198	198	246	269
R^2	0.544	0.552	0.413	0.568	0.583	0.596	0.591	0.672
F	9.307	8.871	5.493	10.926	7.617	8.035	10.124	15.923

***p<0.01，**p<0.05，*p<0.10

第六节 内生性处理

在考察投资者情绪对未来股票收益率的影响时，必须审慎考虑潜在的内生性问题，以确保估计结果的可靠性和因果解释的有效性。内生性偏误可能源于多种因素，在本研究的背景下，主要关注遗漏变量偏误、测量误差以及变量间的联立性或反向因果关系。

对于遗漏变量偏误(Omitted Variable Bias)，可能存在一些未被模型观测到的因素，它们既影响了投资者在股吧中表达的情绪(avg_sentiment)，也影响了股票未来的收益率(forward_ret_1d)。例如，未公开的重大利好或利空消息的提前泄露、宏观经济突发事件的预期等。本研究采用双向固定效应模型，通过控制个体固定效应和时间固定效应，在很大程度上缓解了由不随时间变化的个体异质性和影响所有样本的共同时间冲击所导致的遗漏变量问题。然而，随时间变化的、特定于个股的遗漏变量仍可能存在。

对于测量误差(Measurement Error)，基于自然语言处理技术量化的投资者情绪(avg_sentiment)是对真实、潜在投资者情绪的一种度量。尽管本研究采用了先进的 StructBERT 模型，但文本情感分析本身无法完全避免误差，例如模型对复杂语境、反讽、隐晦表达的理解偏差，以及股吧评论本身可能存在的噪声如水军评论未完全清除。测量误差通常会导致解释变量的系数被低估。

对于反向因果关系(Reverse Causality)，本研究的核心模型设定是用 t 期的情绪 avg_sentiment 预测 t+1 期的收益率 forward_ret_1d。从时间顺序逻辑上看，t+1 期的收益率结果发生在 t 期情绪形成之后，因此未来收益率无法在时间上“导致”过去的投资者情绪。模型设计本身在时间维度上避免了严格的反向因果关系。

一、格兰杰因果关系检验

为进一步探究投资者情绪与股票收益率之间因果关系的方向性，我们运用了格兰杰因果关系检验方法，该方法的核心在于评估一个变量的历史信息是否有助于预测另一个变量的未来值。基于信息准则选定的 2 阶滞后，我们构建了包含未来一日收益率（forward_ret_1d）和平均情绪指数（avg_sentiment）的面板向量自回归模型（PVAR）。根据 Dumitrescu & Hurlin (2012) 的面板格兰杰非因果检验结果，我们发现平均情绪指数（avg_sentiment）在统计上显著地格兰杰导致未来一日收益

率（forward_ret_1d），如图 4-7，Z-bar 统计量的 p 值为 0.0010，小于 5%的显著性水平，这表明情绪变量的历史信息确实有助于预测未来的股票收益率。检验结果同样显示，未来一日收益率（forward_ret_1d）也显著地格兰杰导致平均情绪指数（avg_sentiment），如图 4-8，p 值<0.0001，这意味着过去的收益率信息对于预测未来的投资者情绪同样具有显著贡献。综合来看，格兰杰因果检验揭示了投资者情绪与股票收益率之间可能存在双向的预测关系，而非单一方向的影响。

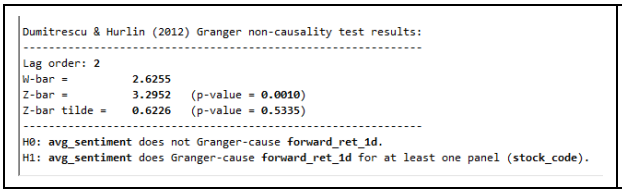


图 4-7 情绪对收益率格兰杰检验

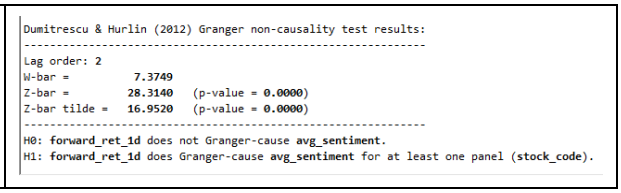


图 4-8 收益率对情绪格兰杰检验

二、工具变量讨论

理想情况下，采用工具变量（Instrumental Variable, IV）方法是处理内生性问题的有效途径。一个合格的工具变量需满足相关性（与内生解释变量 avg_sentiment 显著相关）和外生性（除通过 avg_sentiment 外，不直接影响被解释变量 forward_ret_1d）两个核心条件。为了识别合适的工具变量，本研究进行了多方面的探索和尝试。

首先，我们考虑了情绪指标自身的滞后项，例如滞后一期的平均情绪得分（L.avg_sentiment）。理论上，情绪具有一定的惯性，昨日情绪可能影响今日情绪，满足相关性条件。然而，实证检验中，在使用 L.avg_sentiment 作为工具变量后，核心解释变量 avg_sentiment 的系数变为-0.0471，且其对应的 t 统计量仅为-0.6429，远未达到常规的统计显著性水平。这意味着，在尝试通过工具变量法控制潜在内生性后，投资者情绪对未来一日股票收益率的影响不再显著。同时，对于外生性条件也存在顾虑，即滞后情绪可能因信息传递滞后或情绪效应持续而直接影响未来一日的收益率，而非仅仅通过当期情绪传导。

其次，我们探讨了外部非经济因素作为工具变量的可能性，例如可能影响广泛投资者情绪的天气状况、重大社会事件或特定节假日效应。这类变量理论上可能满足外生性，因为它们通常不直接关联于特定股票的基本面或未来收益。但是，在实践中，将这些宏观或区域性变量与本研究中基于特定股票论坛的、匿名的、高频的个体评论情绪进行有效匹配存在巨大困难。例如，难以获取评论发布者的精确地理位置以关联天气数据，而节假日或社会事件的影响可能过于弥散，与特定股票日度情绪的相关性可能较弱且不稳定。

此外，我们还考虑了市场层面的宏观指标，如市场的投机氛围指标，如商品期货市场的交易活跃度。这些指标可能在一定程度上反映了整体投资环境的情绪或风险偏好，可能与个股论坛的情绪存在关联。然而，这些市场层面的变量很可能直接影响个股的收益率，从而严重违反外生性要求。

经过上述多方面的尝试与审慎评估，本研究未能在现有数据和研究框架内找到既满足强相关性、又满足严格外生性条件的理想工具变量。尽管我们认识到内生性问题的潜在影响，但由于缺乏合适的工具变量，无法通过两阶段最小二乘法（2SLS）对主要结果进行进一步的内生性修正。

第七节 稳健性检验

为验证主要研究结论的稳健性，我们从两个方面进行了稳健性检验：使用不同的情绪指标和考察不同期限的收益率。

一、不同情绪指标的比较

在模型 2 的基础上，我们用正面情绪比例（positive_ratio）替代平均情绪得分（avg_sentiment）作为核心解释变量，检验结果是否稳健。如表 4-7 第二列所示，positive_ratio 的系数为 0.0179，在 5% 水平上显著，表明正面情绪比例越高，未来一日收益率越高。这与使用 avg_sentiment 的结果方向一致，支持了投资者情绪对收益率具有预测作用的结论。

同时，我们考察了情绪动量指标对收益率的影响。表 4-7 第三、四列显示，3 日和 5 日移动平均情绪（ma_3d、ma_5d）的系数分别为 0.0037 和-0.0019，但均不显著。这表明短期情绪的累积效应对收益率的预测作用不如即时情绪显著。同时，情绪波动性（std_3d、std_5d）和情绪变化率（sentiment_change_3d、sentiment_change_5d）也均不显著，表明情绪的稳定性和变化速度对收益率的影响有限。

这些结果表明，在各种情绪指标中，平均情绪得分（avg_sentiment）和正面情绪比例（positive_ratio）是预测股票收益率最有效的指标，而情绪动量指标的预测作用则不显著。投资者和分析师在利用情绪信息时，应更关注当前的平均情绪水平和正面情绪比例。

表 4-7 不同情绪指标的比较

	(1)	(2)	(3)	(4)
	Model 2	Model 5	Model 6	Model 7
avg_sentiment	0.0144***			
	(2.9017)			
positive_ratio		0.0179**		
		(2.5560)		
sentiment_std	-0.0048	0.0020		
	(-0.1843)	(0.0776)		
ma_3d			0.0037	
			(0.4312)	
std_3d			0.0065	
			(0.5971)	
sentiment_change_3d			-0.0000	
			(-0.0037)	

ma_5d				-0.0019
				(-0.1586)
std_5d				-0.0129
				(-0.7436)
sentiment_change_5d				0.0000
				(0.9047)
avg_intensity	0.0087	0.0019	-0.0026	-0.0014
	(0.3571)	(0.0803)	(-0.1536)	(-0.0752)
comment_count	-0.0000	-0.0000	-0.0000**	-0.0000**
	(-1.2387)	(-1.2892)	(-2.1811)	(-2.1987)
sentiment_consensus	-0.0082	-0.0098	-0.0055	-0.0034
	(-0.8420)	(-0.9871)	(-0.7378)	(-0.4349)
close	-0.0008***	-0.0008***	-0.0010***	-0.0013***
	(-8.0542)	(-8.0088)	(-7.0860)	(-7.3584)
volume	-0.0000	-0.0000	-0.0000	-0.0000
	(-1.1581)	(-1.1059)	(-0.5960)	(-0.3245)
amount	-0.0000*	-0.0000*	-0.0000	-0.0000
	(-1.7720)	(-1.7207)	(-0.7761)	(-1.0563)
amplitude	-0.0013***	-0.0013***	-0.0013**	-0.0014**
	(-2.6281)	(-2.6075)	(-2.1711)	(-2.1087)
pct_change	0.0001	0.0002	-0.0005	-0.0001
	(0.4898)	(0.6459)	(-1.5558)	(-0.2752)
price_change	0.0003	0.0003	0.0008***	0.0006**
	(1.3935)	(1.3574)	(3.0585)	(2.1630)
turnover_rate	0.0008	0.0008	0.0015*	0.0018*
	(1.0590)	(1.0532)	(1.7338)	(1.9596)
_cons	0.0778***	0.0693***	0.0634***	0.0899***
	(5.7002)	(5.0906)	(4.0098)	(4.9578)
N	1989	1989	1658	1436
R^2	0.257	0.256	0.237	0.236
F	22.052	21.965	17.493	16.097

***p<0.01，**p<0.05，*p<0.10

二、不同期限的收益率比较

为检验情绪对不同期限收益率的预测作用，我们将被解释变量分别设定为一日、三日和五日未来收益率（forward_ret_1d、forward_ret_3d、forward_ret_5d）。

如表 4-8 所示，avg_sentiment 对 forward_ret_1d 的影响显著为正（系数为 0.0144，1%水平显著），但对 forward_ret_3d 的影响变得不显著（系数为 0.0017），对 forward_ret_5d 的影响甚至转为负向且接近显著（系数为-0.0135）。这表明投资者情绪对股票收益率的预测作用主要集中在短期（一日），

随着时间延长，这种预测作用迅速减弱，甚至可能出现反转。

这一发现与行为金融学中的"情绪过度反应-修正"理论一致：投资者初始情绪可能导致短期内股价过度反应，随后市场会逐渐修正这种偏离，导致中长期内出现收益率反转。这表明情绪信息在短期交易策略中可能更有价值，而在中长期投资中则需谨慎使用。

控制变量方面，随着预测期限延长，某些变量的影响发生了显著变化。例如，收盘价（close）对不同期限收益率的负向影响逐渐增强，系数从-0.0008 增至-0.0029；换手率（turnover_rate）对一日收益率不显著，但对五日收益率显著为负（系数为-0.0071，1%水平显著）；涨跌幅（pct_change）对一日收益率不显著，但对三日和五日收益率显著为正。这些变化表明，不同因素对短期和中期收益率的影响机制可能不同。

表 4-8 不同期限收益率比较

	(1)	(2)	(3)
	Model 8	Model 9	Model 10
avg_sentiment	0.0144***	0.0017	-0.0135
	(2.9017)	(0.2199)	(-1.4119)
sentiment_std	-0.0048	-0.0100	0.0006
	(-0.1843)	(-0.2410)	(0.0122)
avg_intensity	0.0087	-0.0131	-0.0226
	(0.3571)	(-0.3372)	(-0.4843)
comment_count	-0.0000	-0.0000	0.0000
	(-1.2387)	(-0.4005)	(1.5844)
sentiment_consensus	-0.0082	-0.0181	-0.0148
	(-0.8420)	(-1.1536)	(-0.7904)
close	-0.0008***	-0.0020***	-0.0029***
	(-8.0542)	(-12.4343)	(-15.1349)
volume	-0.0000	-0.0000**	-0.0000***
	(-1.1581)	(-2.0477)	(-2.8384)
amount	-0.0000*	-0.0000	-0.0000**
	(-1.7720)	(-0.4204)	(-1.9850)
amplitude	-0.0013***	-0.0037***	0.0001
	(-2.6281)	(-4.6466)	(0.1150)
pct_change	0.0001	0.0015***	0.0019***
	(0.4898)	(3.1297)	(3.3201)
price_change	0.0003	0.0000	0.0003
	(1.3935)	(0.1092)	(0.7864)
turnover_rate	0.0008	-0.0009	-0.0071***
	(1.0590)	(-0.7617)	(-4.8466)
_cons	0.0778***	0.2101***	0.2692***
	(5.7002)	(9.6355)	(10.2804)
N	1989	1989	1989

R^2	0.257	0.256	0.295
F	22.052	21.895	26.710

***p<0.01，**p<0.05，*p<0.10

综上所述，稳健性检验结果支持了本研究的主要结论：投资者情绪，特别是平均情绪得分和正面情绪比例，对股票短期收益率具有显著的预测作用。同时，这种预测作用具有时效性，主要体现在一日内，随着时间延长迅速减弱。

第五章 总结与展望

第一节 研究结论

本研究基于行为金融学理论与信息传导理论，借助自然语言处理这一前沿技术手段，深入探究了中国 A 股市场中，源自在线投资者社区（东方财富网股吧）的投资者情绪如何影响股票的短期收益表现。通过运用 Python 爬虫技术获取海量股评文本数据，并采用先进的 StructBERT 深度学习模型进行情感量化，我们构建了多维度的情绪指标体系，并利用面板数据双向固定效应模型进行了严谨的实证分析。研究时段覆盖了 2025 年 2 月的交易日，样本囊括了 8 个主要行业的 111 只代表性股票。本研究的实证分析结果揭示了基于自然语言处理技术量化的投资者情绪对中国 A 股市场股票短期收益率具有显著的预测能力。

（1）本研究发现投资者情绪，特别是以日度平均情绪得分和正面情绪评论比例衡量的指标，能够显著正向预测股票次日的收益率。具体而言，在控制了股票自身交易特征、行业效应以及时间固定效应后，平均情绪得分每提升一个百分点，预计将带来次日股票收益率约 0.014 至 0.017 个百分点的增长。这一发现为投资者情绪影响资产定价提供了来自中国市场的直接微观证据。

（2）本研究证实了投资者情绪的预测效应存在显著的行业异质性。在所考察的八个行业中，情绪对股票收益率的预测能力在计算机、电子、食品饮料以及医药生物等行业表现得尤为突出且统计上显著。这可能与这些行业的特性有关，例如较高的成长预期、更依赖信息驱动或拥有更活跃的散户投资者基础，使得它们对市场情绪的变化更为敏感。

（3）研究结果强调了投资者情绪影响的时效性特征。情绪指标对股票收益率的预测能力主要集中在短期，即未来一日。随着预测期限延长至三日或五日，情绪指标的预测能力迅速减弱，甚至出现统计上不显著或反转的迹象。这与行为金融学中关于情绪驱动的市场过度反应及其后续修正的理论预期相符，表明由情绪引发的价格偏离往往是短暂的。

第二节 不足与展望

一、数据限制

（1）本研究的样本期相对较短，仅覆盖了 2025 年 2 月这一个月的交易数据。较短的时间跨度可能无法完全捕捉不同市场周期如牛市、熊市、震荡市下投资者情绪作用的差异性，研究结论的普适性可能受到一定限制。

（2）本研究使用的是相对低频的日度数据。虽然日度数据在金融研究中较为常用，但投资者情绪和市场反应在日内可能发生快速变化。更高频的数据如小时甚至分钟级别或许能更精细地刻画情绪的瞬时冲击与市场的即时反应机制，但相应的数据获取和处理难度也会显著增加。

（3）本研究在处理股评数据时，未充分考虑评论的权重或质量差异。现实中，不同用户的评论

影响力可能不同，例如意见领袖与普通散户的差异，评论本身的质量也参差不齐。高质量、有深度的分析评论可能被大量低质量、情绪化的“水帖”所淹没。目前采用简单平均或比例计算情绪指标的方法，未能区分这些差异，可能导致情绪信号的精确度受到影响。未来研究可以探索引入用户影响力评分、评论点赞数、回复数或基于内容复杂度的质量评估等因素，对评论进行加权处理，以构建更精准的情绪指标。

二、方法局限

（1）金融文本信息的处理本身具有高度的复杂性和专业性。自然语言处理模型在理解特定金融语境下的微妙语义、反讽、双关以及隐含情绪时仍面临挑战。例如，一句表面中性的话在特定市场背景下可能蕴含强烈的看涨或看跌预期。要实现更精准的情感识别，理想情况下需要金融领域专家的深度参与，进行大量的人工标注以构建高质量的领域专用训练数据集，从而优化模型性能。同时，数据清洗、停用词选择、无用信息过滤等预处理环节也需要专业知识指导，以避免误删有用信息或引入偏差。

（2）本研究采用了基于 BERT 和 RoBERTa 微调的 StructBERT 和 Erlangshen 模型，这些模型在当时已属先进。然而，自然语言处理技术，特别是大规模预训练语言模型正处在飞速发展之中。近年来，国内外研究机构不断推出训练规模更大、结构更复杂、参数量更多、训练成本也更高的模型，如 GPT 系列、更强大的 BERT 变种等。这些新一代模型在理解语言的深度和广度上可能具有更优越的表现。

（3）情绪指标的构建方式仍有优化空间。本研究主要关注了情绪的平均水平、比例和波动性等维度。然而，投资者情绪是一个多维度的复杂概念，还可以从更多角度进行刻画，例如情绪的分歧度、情绪的传染性、特定主题相关的情绪等。

（4）本研究在处理潜在的内生性问题时，虽然采用了双向固定效应模型控制了部分遗漏变量，并通过格兰杰因果检验初步探讨了变量间的动态关系，但未能找到合适的工具变量来更彻底地解决可能存在的反向因果或更复杂的遗漏变量问题。

参考文献

[1] Aboody D, Lehavy R, Trueman B. Limited attention and the earnings announcement returns of past stock market winners[J]. Review of Accounting Studies, 2010, 15(2): 317-344.

[2] Antweiler W, Frank M Z. Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards[J]. The Journal of Finance, 2004, 59(3): 1259-1294.

[3] Baker M, Stein J C. Market liquidity as a sentiment indicator[J]. Journal of Financial Markets, 2004, 7(3): 271-299.

[4] Baker M, Wurgler J. Investor Sentiment and the Cross-Section of Stock Returns[J]. The Journal of Finance, 2006, 61(4): 1645-1680.

[5] Baker M, Wurgler J. Investor sentiment in the stock market[C]. Journal of Economic Perspectives, 2007: 129-151.

[6] Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market[J]. Journal of Computational Science, 2011, 2(1): 1-8.

[7] Brown G W, Cliff M T. Investor sentiment and the near-term stock market[J]. Journal of Empirical Finance, 2004, 11(1): 1-27.

[8] Da Z, Engelberg J, Gao P. In Search of Attention[J]. The Journal of Finance, 2011, 66(5): 1461-1499.

[9] De Long J B, Shleifer A, Summers L H, et al. Noise Trader Risk in Financial Markets[J]. Journal of Political Economy, 1990, 98(4): 703-738.

[10] Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C], 2019: 4171-4186.

[11] Ding X, Zhang Y, Liu T, et al. Using Structured Events to Predict Stock Price Movement: An Empirical Investigation[C], 2014: 1415-1425.

[12] Kim Y. Convolutional Neural Networks for Sentence Classification[C], 2014: 1746-1751.

[13] Lee C M C, Shleifer A, Thaler R H. Investor Sentiment and the Closed-End Fund Puzzle[J]. The Journal of Finance, 1991, 46(1): 75-109.

[14] Lemmon M, Portniaguina E. Consumer Confidence and Asset Prices: Some Empirical Evidence[J]. The Review of Financial Studies, 2006, 19(4): 1499-1529.

[15] Liu Y, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach[M]. 2019.

[16] Loughran T, Mcdonald B. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks[J]. The Journal of Finance, 2011, 66(1): 35-65.

[17] Pang B, Lee L, Vaithyanathan S. Thumbs up? sentiment classification using machine learning techniques[C]. Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10, 2002: 79–86.

[18] Stambaugh R F, Yu J, Yuan Y. The short of it: Investor sentiment and anomalies[J]. Journal of Financial Economics, 2012, 104(2): 288-302.

[19] Tang D, Qin B, Liu T. Document Modeling with Gated Recurrent Neural Network for Sentiment Classification[C]. Conference on Empirical Methods in Natural Language Processing, 2015.

[20] Tetlock P C. Giving Content to Investor Sentiment: The Role of Media in the Stock Market[J]. The Journal of Finance, 2007, 62(3): 1139-1168.

[21] Turney P D. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews[C]. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002: 417–424.

[22] Zhang X, Fuehres H, Gloor P A. Predicting Stock Market Indicators Through Twitter “I hope it is not as bad as I fear”[J]. Procedia - Social and Behavioral Sciences, 2011, 26: 55-62.

[23] 段江娇, 刘红忠, 曾剑平. 中国股票网络论坛的信息含量分析[J]. 金融研究, 2017(10): 178-192.

[24] 黄润鹏, 左文明, 毕凌燕. 基于微博情绪信息的股票市场预测[J]. 管理工程学

报, 2015, 29(01): 47-52+215.
[25] 李思龙, 金德环, 李岩. 网络社交媒体提升了股票市场流动性吗?——基于投资者互动视角的研究[J]. 金融论坛, 2018, 23(07): 35-49+63.
[26] 杨永伟. 基于财经新闻的股票收益方向预测[D]. 2018.
[27] 张强, 杨淑娥, 杨红. 中国股市投资者情绪与股票收益的实证研究[J]. 系统工程, 2007, 25(7): 5.

附录 A 代码

本附录提供代码仓库地址：https://github.com/K1ndredzzz/Paper-demo

致　　谢

恭敬在心，不在虚文。