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

上海财经大学毕业（设计）论文独创性声明

本人声明所呈交的毕业（设计）论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得上海财经大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

论文作者签名：_____ 日期：_____年_____月_____日

上海财经大学毕业（设计）论文使用授权声明

上海财经大学有权保留本人所送交毕业（设计）论文的复印件和电子文档，可以采用影印、缩印或其他复制手段保存论文。本人电子文档的内容和纸质论文的内容相一致。除在保密期内的保密论文外，允许论文被查阅和借阅，可以公布（包括刊登）论文的全部或部分内容。论文的公布（包括刊登）授权上海财经大学教务处办理。

论文作者签名：_____ 导师签名：_____

日期：_____年_____月_____日 日期：_____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预测作用，尝试解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 深度学习模型进行情感分析，构建多维情绪指标体系；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，并通过工具变量法和格兰杰因果关系检验处理潜在的内生性问题。

研究内容涵盖样本期内(2025 年 2 月)8 个行业 111 只股票的评论文本分析和股价表现研究。具体包括：股评文本的预处理与情感分析，情绪指标体系的构建，个股和行业层面的面板数据回归分析，分行业异质性分析，以及对不同情绪指标和不同期限收益率的稳健性检验。

研究发现：平均情绪得分每提高一个百分点，未来一日收益率平均提高约 0.014-0.017 个百分点；情绪的预测效果在不同行业间存在显著差异，对计算机、银行、医药生物和房地产等行业的影响更为显著；情绪的影响具有明显的时效性，主要体现在短期（一日），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；在多种情绪指标中，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，股票收益率，情绪指标，面板数据分析

ABSTRACT

This study examines the predictive power of investor sentiment on short-term stock returns, addressing whether emotional information embedded in investors' online comments can effectively forecast stock price movements. Based on behavioral finance theory, the research posits that investor sentiment, as an irrational factor, influences investment decisions and systematically affects asset prices; it also draws on information transmission theory to explore the mechanisms of sentiment information diffusion and price discovery in markets.

The research employs natural language processing technology and econometric methods. Using Python web crawlers to collect comment data from East Money Stock Forum, it applies StructBERT and RoBERTa deep learning models for sentiment analysis and constructs a multi-dimensional sentiment indicator system. Panel data fixed-effects models are utilized to analyze the impact of sentiment indicators on stock returns, while potential endogeneity issues are addressed through instrumental variable methods and Granger causality tests.

The research content covers text analysis and stock performance for 111 stocks across 8 industries during the sample period (February 2025). Specifically, it includes preprocessing and sentiment analysis of stock comments, construction of a sentiment indicator system, panel data regression analysis at both individual stock and industry levels, heterogeneity analysis across industries, and robustness tests using different sentiment indicators and different time horizons for returns.

The findings reveal that for each percentage point increase in average sentiment score, next-day stock returns increase by approximately 0.014-0.017 percentage points; the predictive effect varies significantly across industries, with more pronounced impacts on computer, banking, pharmaceutical, and real estate sectors; the influence of sentiment is notably time-sensitive, primarily manifested in the short term (one day) and gradually weakening or even reversing over time, consistent with the sentiment overreaction-correction theory; among various sentiment indicators, average sentiment score and positive sentiment ratio are the most effective predictors of stock returns.

KEY WORDS: Investor Sentiment, Natural Language Processing, Stock Returns, Sentiment

Indicators, Panel Data Analysis

目 录

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

投资者情绪作为一种难以量化但影响深远的因素，长期以来在资本市场中扮演着重要角色。早在 1636 年荷兰郁金香狂潮事件中，人们就已发现投资者情绪对资产价格的显著影响，投资者心理中存在的从众和恐慌心理往往导致市场非理性波动。在金融危机来临时，金融市场的急剧变化与投资者情绪密切相关，体现了情绪因素在市场运行机制中的重要地位。

近年来，学术界对投资者情绪的关注度不断提高。研究表明，互联网投资者情绪对于股票市场的波动具有指导意义。吴杰胜等（2019）基于多部情感词典结合规则集，对微博情感进行分析，实证结果显示该方法在微博情感分析领域效果较好。戴德宝等（2019）使用机器学习方法对文本进行分析，同样发现投资者情绪指标对股价走势具有一定的预测作用。

陈其安等（2017）的研究进一步证实，中国股票市场价格波动性与投资者情绪呈正相关关系，与市场利率呈负相关关系，且投资者情绪会弱化货币政策的调控作用。这些研究共同表明，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前景

随着大数据和人工智能技术的迅猛发展，自然语言处理（NLP）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，特别是模拟人工神经元网络的机器学习流派深度学习的兴起，极大地推动了通用人工智能方法的普及。

在金融文本分析方面，情感分析是从股票文本中获取投资者情绪的主要手段，其准确性直接影响后续股票价格波动分析的基础。不同的情感分析方法准确率各异，因此选择较高准确率的情感分析方法对研究结果至关重要。然而，需要注意的是，直接套用现成的情感分析软件和外部词典资源在金融领域难以取得很好的效果，需要针对金融特定领域进行模型调整和优化。

相比传统基于词典和机器学习的统计方法，基于深度学习的自然语言处理技术，如 BERT 模型，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。这种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工具，为投资者情绪研究打开了新的可能性。

二、研究意义

（一）理论意义

本研究从行为金融学的角度，探讨了互联网平台上投资者言论所蕴含的情绪信息与股票市场表现之间的关系，丰富了行为金融学的理论内涵。通过构建情绪指标体系，本研究有助于深化对投资者情绪如何传导并影响市场的理解，为非理性投资行为研究提供了新的分析框架。本研究还将 NLP 技术与金融数据分析相结合，探索了深度学习模型在金融文本分析中的应用效果，拓展了自然语言处理技术在专业领域中的应用边界。通过比较不同情感词典和数据集在股票预测领域的表现，本研究为金融领域的文本分析方法提供了重要的方法论参考。

（二）现实意义

首先，互联网评论作为投资者情绪的直接表达，对市场预期具有重要的指示作用。通过分析和理解网络评论中所包含的情绪信息，可以为投资决策提供参考，帮助投资者更好地把握市场动向。其次，随着"互联网+"政策的实施，网络媒体在信息传播过程中发挥的作用越来越大。正确识别新闻报道和网络评论中的舆情信息，并理解其与股票收益之间的关系，不仅有利于投资决策，还能为政府监管部门制定相关政策提供参考，维护市场秩序和稳定。

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，避免盲目跟风，优化投资策略。对监管机构而言，本研究有助于构建金融市场情绪监测系统。互联网评论中的情绪波动往往能提前反映市场异常，通过实时监测投资者情绪，监管部门可以更早识别市场风险，及时采取干预措施，维护市场稳定。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。研究表明，情感分析能有效反映股票市场状况，并对股价走势产生强烈影响。

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：首先，利用 Python 编程语言开发网络爬虫，从东方财富股吧获取 8 个行业 111 只股票在 2025 年 2 月的评论数据。通过数据清洗与预处理，构建高质量的文本分析样本。其次，基于先进的自然语言处理技术对股评文本进行情感分析，构建包括情感得分、情感极性、情感强度等在内的多维情绪指标体系。这些指标能够从不同角度刻画投资者情绪的特征。第三，通过 akshare 接口获取样本股票的日度交易数据，将情绪指标与股票行情数据进行匹配，构建包含情绪、交易等信息的面板数据集。第四，运用计量经济学方法，重点探究不同维度情绪指标对未来股票收益的预

测能力。通过构建面板回归模型，分析投资者情绪对股市的影响机制。

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，并通过相关性分析、回归分析等方法研究情绪指标与股票收益的关系。

（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 研究创新点与贡献

在技术方法层面，本研究采用先进的自然语言处理模型进行情感分析。具体而言，通过应用预训练语言模型对金融文本进行深度语义理解，提高了对投资者评论情感的识别准确性。该模型能够有效捕捉文本中的细微情感差异，为后续构建情绪指标提供了可靠的技术支持。

在研究视角层面，本文创新性地构建了多维情绪指标体系。通过整合情感得分、情感极性、情感强度等多个维度的指标，全面刻画了投资者情绪的不同特征。这种多维度的情绪指标体系不仅能够更准确地反映投资者情绪状态，还可以深入分析不同维度情绪指标对股票市场的差异化影响。

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。该部分主要介绍研究背景与意义，阐述研究内容与方法，并说明本文的创新点与贡献。

第二章为文献综述。该部分回顾投资者情绪研究现状，包括情绪的定义与测度方法，以及投资者情绪对股市的影响机制。同时对自然语言处理技术的研究进展进行梳理。

第三章为研究设计。该部分详细说明数据来源与处理方法，构建情绪分析模型框架，设计情绪指标体系，并对模型设定和变量进行说明。

第四章为实证分析。该部分从个股层面展开研究。通过描述性统计、单位根检验、协整检验等方法验证数据特征，运用面板数据回归分析投资者情绪对股市的影响，并进行异质性分析和稳健性检验。

第五章为研究结论。该部分总结主要研究发现，提出政策建议与启示，同时讨论研究局限，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中的核心概念，反映了投资者对未来市场走势的预期和认知偏差。从理论界定角度看，Brown 和 Cliff（2006）认为投资者情绪本质上反映了市场参与者的期望，是导致市场波动的重要因素。Peng 和 Xiong（2005）则从认知资源角度提出，投资者情绪是一种稀缺的认知资源，有限的投资者关注会导致存在类别学习的现象，从而影响投资决策。Baker 和 Stein（2003）从市场流动性视角出发，指出在流动性强的市场中，价格主要由非理性的投资者主导，高流动性反映了投资者的积极情绪。

传统的投资者情绪测度方法主要依赖于间接的金融市场指标作为代理变量。Lee 等（1991）率先使用封闭式基金的折现率作为投资者情绪的代理变量，发现该代理变量有效，基金的风险直接受到投资者情绪影响，从而影响资产组合的价格。Baker 和 Wurgler（2006）通过构建了包括 IPO 数量、股权、换手率、首日平均收益率和股息溢价等在内的复合指标来度量投资者情绪，发现投资者情绪像跷跷板一样决定了投资者如何选择投资标的。Huang 等（2015）指出了利用主成分分析方法存在的缺陷，并使用偏最小二乘法来估计投资者情绪。Aboody 等（2018）则创新性地使用隔夜收益率来衡量投资者情绪，发现投资者情绪对市场短期收益率有着显著的正向作用，同时低隔夜收益率的股票长期表现更好。

随着互联网技术和大数据分析方法的快速发展，基于互联网数据的新测度方法逐渐兴起，主要包括社交媒体数据、金融论坛数据和互联网搜索数据三大类。在社交媒体方面，Zhang（2011）等人通过收集六个月的 Twitter 信息并分析其情绪倾向，发现情绪化的 Twitter 与道琼斯、纳斯达克和标普 500 指数显著负相关。黄润鹏等（2015）利用新浪微博平台的接口分析微博情绪与上证综指的相关性，发现微博情绪指标的加入可以有效提升模型的准确率。在金融论坛数据方面，段江娇等（2017）构建了股票论坛关注度、投资者情绪一致性等指标，结果表明股票收益受到网络论坛关注和情绪影响，而股票交易量受到情绪一致性影响。在互联网搜索数据方面，郑瑶（2016）将投资者情绪与百度指数构建回归模型，证明了搜索指数有助于提高股票市场变量预测的精度。赵妍妍等（2017）则利用点互信息（PMI）衡量对应情感倾向的相关性分数，并借此构建大规模情感词典，提高了情绪测度的准确性。

二、投资者情绪与资本市场表现关系研究

投资者情绪对股票收益的影响是学术界关注的焦点。来自中国市场的研究表明，张强等（2007）

发现相比于几乎不造成影响的个人投资者，机构投资者情绪是影响中国股市的系统因素。杨永伟（2018）研究中国权威媒体发布的新闻后发现，当新闻中蕴含强烈的积极情绪时，会使得股价大幅度上涨。国际市场研究方面，Lin 等（2018）通过研究投资者情绪与美国股指期货市场及其对应的现货标的价格的关系，发现投资者情绪与期货市场表现呈现明显的负相关关系。Cedric 等（2019）发现投资者关注度对投资者情绪有因果关系，并且投资者情绪对于大盘股而言较为短暂，对于小盘股而言情绪效应较为持久。Aboody 等（2018）的研究表明投资者情绪对市场短期收益率有着显著的正向作用，而低投资者情绪时期的股票在长期表现更好，这反映了投资者情绪的短期推动与长期反转特征。

投资者情绪也对市场交易量产生显著影响。Baker 和 Stein（2003）的研究表明，高流动性反映了投资者的积极情绪，在流动性强的市场中，价格主要由非理性的投资者主导。段江娇等（2017）通过对股票论坛数据的分析发现，股票交易量主要受到情绪一致性的影响，而非单纯的情绪水平。李思龙等（2018）采用东方财富股吧论坛帖子考察投资者情绪，发现投资者的互动可以增加企业股东数量并且提高流动性，同时降低信息不对称性。这些研究表明投资者情绪不仅直接影响价格，还通过交易行为影响市场微观结构。

在牛熊市中，投资者情绪表现出显著的差异化影响。项辛怡（2018）的研究显示，投资者情绪对股票收益产生负向影响，并在牛市阶段和熊市阶段各具特点：投资者情绪对股票收益的影响在牛市比熊市更为强烈。这可能是因为牛市中投资者乐观情绪更容易形成羊群效应，导致市场反应更为剧烈。王聪等（2018）的研究结果表明上证指数涨跌幅与投资者情感之间同时存在线性与非线性关系，且日区间联动幅度较大；在构建模型时应综合考虑两者关系和投资者情感的异方差性。此外，网络新闻舆情也表现出对投资者情绪的差异化影响。Yang 和 Yan（2011）发现新闻舆情导致的高情绪在超过某个临界值后会带来负的超额收益，低于该临界值则超额收益为正。这种非线性关系表明，理解投资者情绪与市场表现的关系需要考虑市场环境和情绪临界点的综合影响。

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

自然语言处理（Natural Language Processing, NLP）作为人工智能领域的重要分支，主要研究通过计算机对自然语言进行分析、理解和识别，广泛应用于机器翻译、文本分类等多个场景。在文本情感分析领域，技术方法经历了从基于词典到机器学习再到深度学习的演进过程。

基于词典的传统方法是最早的文本情感分析方法，其核心是通过构建情感词典，对文本进行分词后进行情感匹配和分类。这种方法的主要步骤包括词干提取、分词、词形还原和词性标注等基础处理。然而，基于词典的方法存在局限性，主要体现在需要大量人力维护词典，且难以准确处理网络新词和缩略词。

随着机器学习技术的发展，研究者开始采用更为智能的文本分析方法。这类方法首先需要将文本数据按比例划分为训练集和测试集，通过观察损失函数的变化来优化模型性能。相比词典方法，机器学习方法具有更强的自适应能力，但仍需要大量人工标注的训练数据。

近年来，深度学习方法在文本情感分析领域取得了突破性进展。以卷积神经网络(CNN)、循环神经网络(RNN)、长短期记忆网络(LSTM)和 Transformer 为代表的深度学习算法，在文本情感分类任务中展现出强大的性能。目前多款专用于深度学习的程序框架，如 Google 公司的 TensorFlow 和 Facebook 公司的 PyTorch，这些工具为深度学习研究提供了极大便利。

二、自然语言处理技术在金融领域的应用

金融文本处理具有其特殊性，主要体现在专业性强、词义模糊度高等方面。传统情感分析方法在金融领域的应用面临着专业词汇识别困难、上下文语义理解不准确等挑战。为解决这些问题，研究者开始构建专门的金融领域词典，如 CFSD 中文金融情感词典，并通过持续更新和扩充来提高分析效果。

在不同模型的应用效果比较方面，实证研究表明深度学习模型表现最为优异。这主要得益于深度学习模型能够自动学习特征，适应金融文本的复杂性。特别是在处理时间序列信息方面，LSTM 等模型展现出独特优势。然而，金融领域的文本分析仍需要结合专业知识，单纯依靠算法难以获得理想效果。

第三节 研究评述

近年来,投资者情绪与股市表现的关系研究取得了显著进展。现有研究从理论和实证两个层面较为系统地探讨了投资者情绪的本质内涵、测度方法及其对资本市场的影响机制。在情绪测度方面,研究方法经历了从传统金融指标到互联网大数据的转变,特别是基于社交媒体、金融论坛和搜索数据的新型测度方法极大地丰富了研究手段。在影响机制方面,研究发现投资者情绪不仅直接影响股票收益和交易量,还在不同市场环境下呈现出显著的差异化特征。与此同时,自然语言处理技术的快速发展为投资者情绪研究提供了有力的技术支持,从基于词典的传统方法到深度学习算法的应用,极大地提升了文本情感分析的准确性和效率。

现有研究仍存在若干不足之处：首先，现有的情绪指标在时效性和准确性方面还有待提高，难以完全捕捉市场情绪的快速变化。其次,多数研究关注投资者情绪的整体水平,而对情绪的结构性特征研究不足,如情绪的分布特征、波动特征等缺乏深入探讨。其次,在采用互联网数据测度情绪时,往往忽视了不同信息源的可靠性差异,且未充分考虑信息传播过程中的时滞效应。再次,现有的自然语言处理模型在处理金融专业文本时仍面临着专业词汇识别困难、上下文语义理解不准确等技术瓶颈。此外,研究多集中于探讨投资者情绪对市场的单向影响,对二者之间可能存在的动态反馈机制关注不足。最后，对情绪影响市场的传导机制研究不够深入，缺乏对微观层面的细致分析。

未来研究进行改进和深化的方向：一是构建更加实时和精准的情绪指标体系，充分利用高频数据捕捉市场情绪变化；二是构建多源异构数据融合框架,综合考虑不同信息源的特点和权重，建立更为完善的情绪测度体系；三是针对金融领域特点开发专门的自然语言处理模型,提高文本分析的准确性；四是建立投资者情绪与市场表现的双向互动模型,更好地揭示二者的动态关系。五是加强对市场微观结构的研究，深入分析投资者情绪对个股和不同类型投资者的差异化影响。这些改进将有助于更全面地理解投资者情绪对股市的影响机制。

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 3.11.9 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。

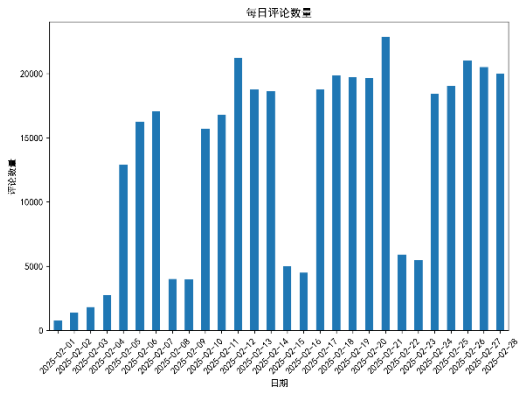


图 3-1 评论日期分布

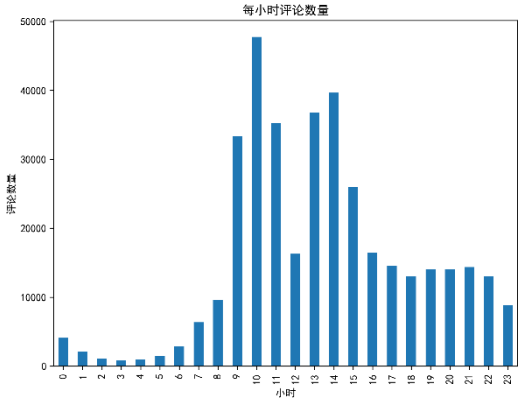


图 3-2 评论时刻分布

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。

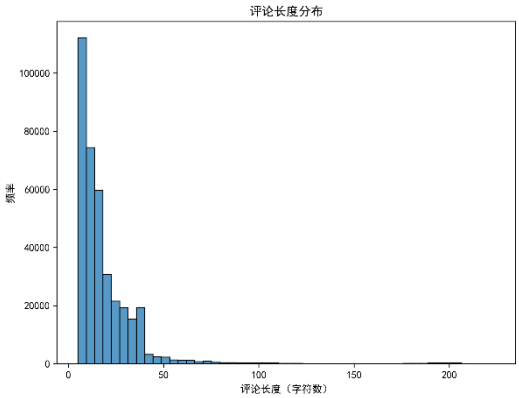


图 3-4 评论长度分布

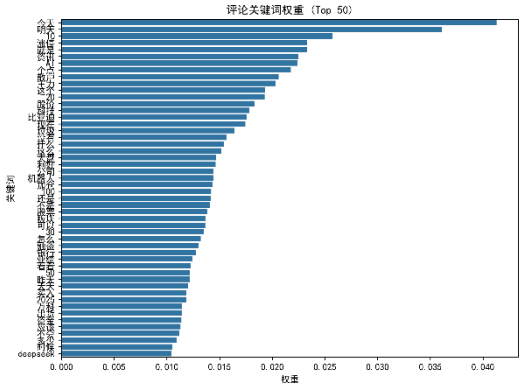


图 3-5 评论关键词权重（top50）

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。

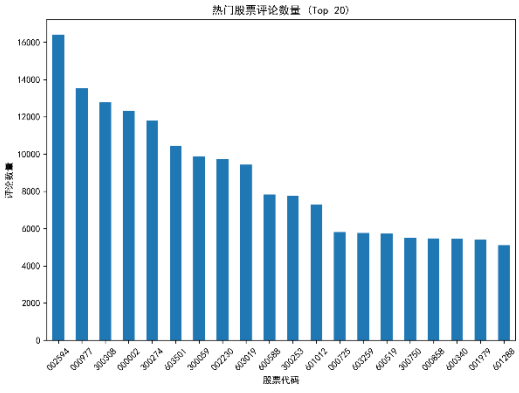


图 3-7 热门股票评论数（top20）

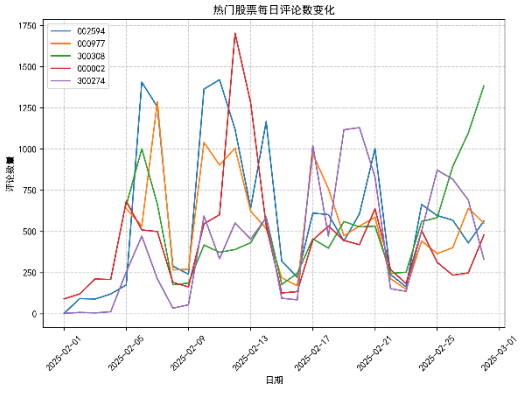


图 3-8 top5 股票每日评论数变化

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(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。

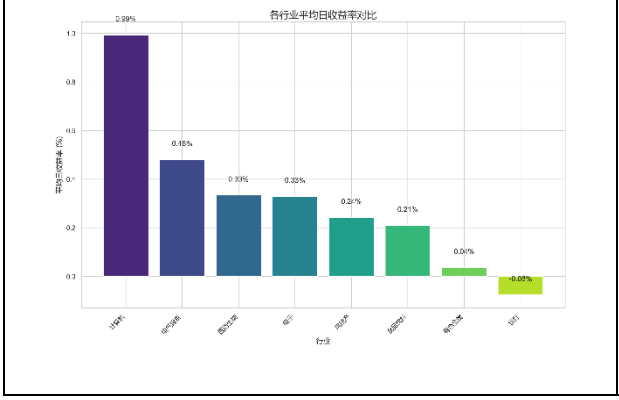


图 3-9 各行业平均日收益率

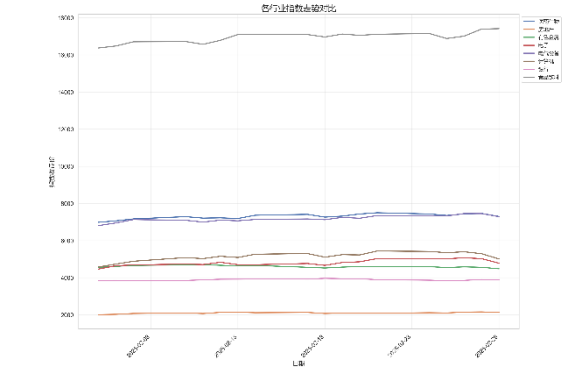


图 3-10 行业指数走势

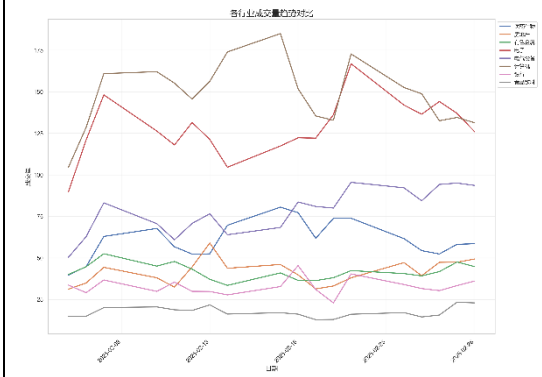


图 3-11 行业成交量趋势

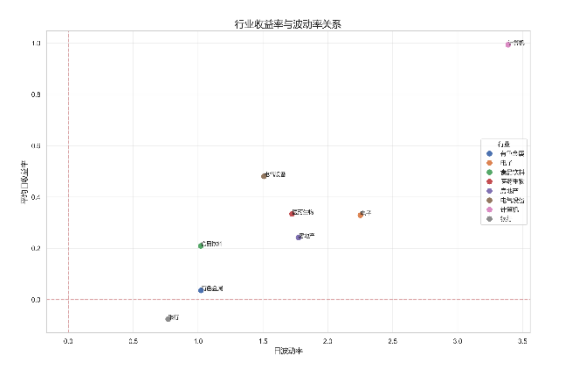


图 3-12 行业收益率与波动率散点图

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。综合评估后选取 StructBERT 分类结果作为情绪变量。

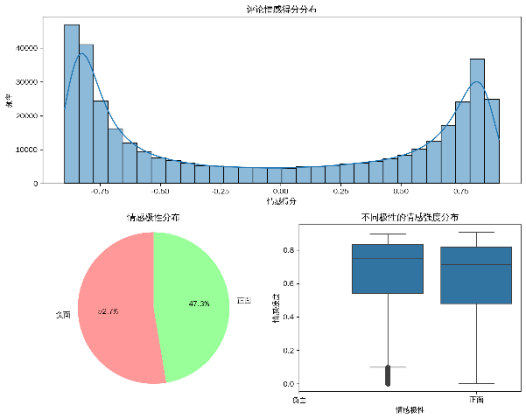


图 3-13 情感分布（StructBERT）

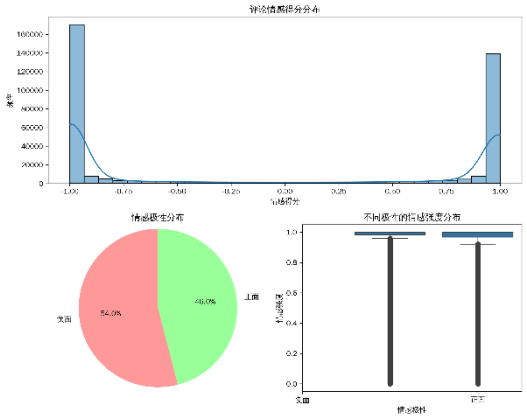


图 3-14 情感分布（RoBERTa）

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，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)、评论的平均情感强度(avg_intensity)、情绪净值(sentiment_net)和一致性(sentiment_consensus)，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。同时计算了情绪净值(sentiment_dispersion)指标来衡量行业内部情绪的分歧程度,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02.01-02.28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 avg_sentiment 等情绪指标。具体而言，本研究构建了一系列情绪变量，包括 avg_sentiment(平均情绪得分)、sentiment_std(情绪标准差)、avg_intensity(平均情绪强度)、comment_count(评论数量)以及 sentiment_consensus(情绪一致性)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 positive_ratio(正面情绪比例)和 ma_3d/5d/10d(移动平均)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。

从情绪指标来看，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。

就市场交易指标而言，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

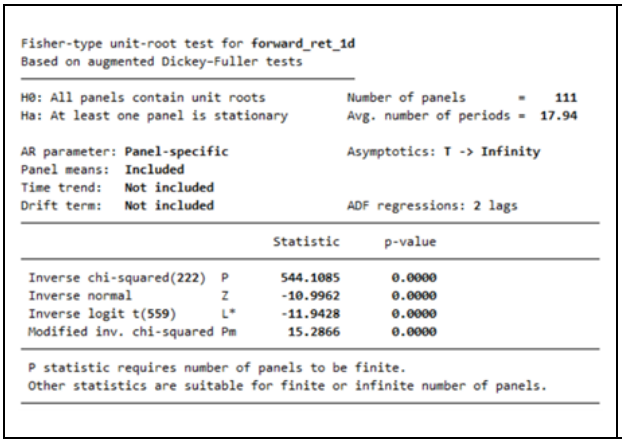
Variable	Obs	Mean	Std. Dev.	Min	Max
stock code	1991	383270.28	261941.11	2	688111
time	1991	9.522	5.184	1	18
avg sentiment	1991	-.077	.15	-.783	.512
sentiment std	1989	.664	.04	.121	.866
positive ratio	1991	.456	.103	0	1
avg intensity	1991	.638	.041	.131	.791
comment count	1991	171.956	189.799	1	1732
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

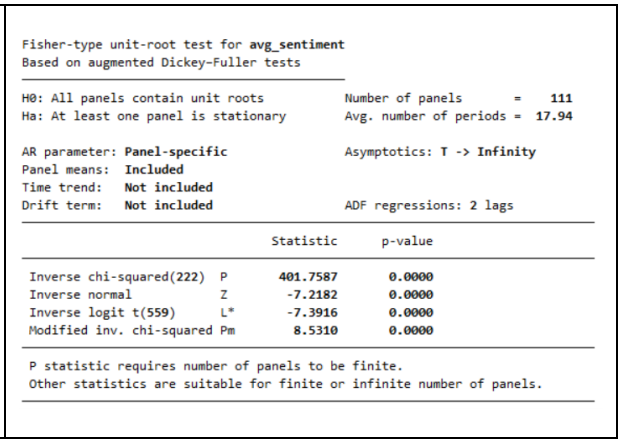
H0: All panels contain unit roots — Number of panels = 111
Ha: At least one panel is stationary — Avg. number of periods = 17.94

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

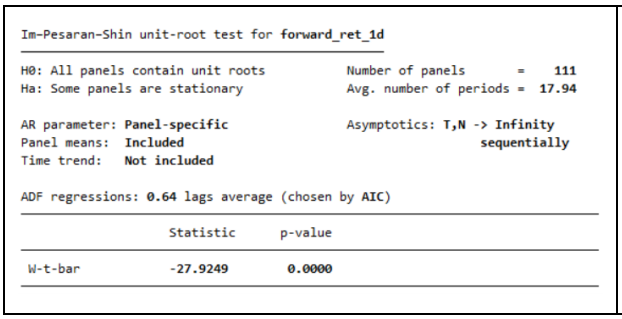
H0: All panels contain unit roots — Number of panels = 111
Ha: At least one panel is stationary — Avg. number of periods = 17.94

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

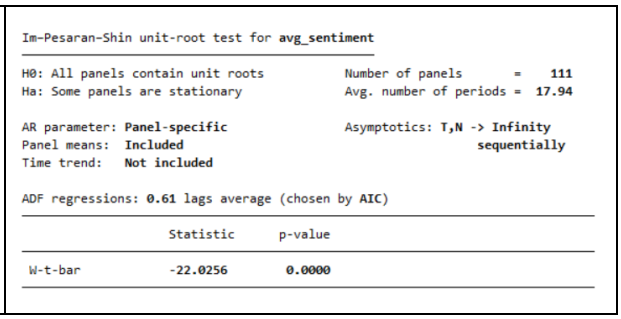
H0: All panels contain unit roots — Number of panels = 111
Ha: Some panels are stationary — Avg. number of periods = 17.94

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，为后续分析情绪对收益率的预测提供了理论基础。

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：

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

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（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。

首先，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。值得注意的是，在加入时间固定效应后（模型 2），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；银行、医药生物和房地产行业的系数分别为 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、银行、医药生物和房地产等行业，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

一、工具变量讨论

在研究投资者情绪与股票收益率的关系时，潜在的内生性问题不容忽视。内生性可能来源于以下几个方面：（1）遗漏变量：未纳入模型的变量同时影响情绪和收益率；（2）反向因果：当日股价表现可能影响投资者情绪；（3）测量误差：情绪变量的测量可能存在不准确性。

为缓解内生性问题，本研究采用工具变量法进行估计。理想的工具变量应满足两个条件：与内生解释变量（avg_sentiment）高度相关，且除通过 avg_sentiment 间接影响外，不直接影响被解释变量（forward_ret_1d）。

基于现有文献和变量特性，我们选择

情绪指标的滞后项？

天气？地理？

新闻媒体相关？

重大节假日？文娱事件

周一效应？月初/月末效应？季节性因素？

螺纹钢期货主力合约的成交量 (Vo) 和持仓量(0pen)作为商品期货市场投机氛围的替代指标？

二、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的因果关系方向，我们进行了格兰杰因果关系检验。格兰杰因果性的概念基于预测能力：如果变量 X 的历史信息有助于预测变量 Y 的未来值，则称 X 格兰杰导致 Y。

我们构建了包含 forward_ret_1d 和 avg_sentiment 两个变量的面板向量自回归模型（PVAR），滞后阶数选择为 2，基于信息准则确定。检验结果表明，在 5%的显著性水平上，avg_sentiment 格兰杰导致 forward_ret_1d，而 forward_ret_1d 不格兰杰导致 avg_sentiment。这表明情绪变量的历史信息有助于预测未来收益率，而过去的收益率对预测未来情绪不具显著贡献。

这一发现进一步支持了情绪对收益率具有预测作用的假设，且排除了收益率对情绪的反向因果可能性，增强了我们主要结论的可靠性。格兰杰因果关系检验的结果与前述工具变量估计结果相互补充，共同表明在控制潜在内生性后，投资者情绪对股票收益率的预测作用依然存在。

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

首先，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。

其次，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。这一发现对构建有效的情绪指标体系具有重要启示：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，而非情绪的变化趋势。

表 4-7 不同情绪指标的比较

	(1) Model 2	(2) Model 5	(3) Model 6	(4) 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

总体而言，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。这些发现对理解情绪在资本市场中的传导机制和实践应用具有重要价值。

第五章 研究结论

第一节 主要研究结论

一、情绪解释效果

二、行业差异特征

第二节 政策建议与启示

网络水军管理/新闻、论坛等舆论传媒工具的管理

第三节 研究局限

一、数据限制

二、方法局限

三、外部有效性

第四节 未来研究展望

一、方法改进方向

二、研究扩展方向

三、应用推广建议

参考文献

列出作者直接阅读过或在正文中引用过的文献资料。撰写论文时，需注意引用权威和最新的文献。

参考文献需在引文右上角用方括号“[]”标明序号，如“基本机构[1]”，并在参考文献中列出。每一条参考文献著录均以“.”结束。参考文献要另起一页，一律放在正文之后，不得放在各章节之后。

参考文献采用顺序编码制，需符合《信息与文献 参考文献著录规则》（GB/T 7714-2015）规范要求，文献类型和标识代码为：普通图书[M]、会议录[C]、汇编[G]、报纸[N]、期刊[J]、学位论文[D]、报告[R]、标准[S]、专利[P]、数据库[DB]、计算机程序[CP]、电子公告[EB]、档案[A]、舆图[CM]、数据集[DS]、其他[Z]。

参考文献中主要责任者的个人作者采用姓在前名在后的著录形式，当作者不超过 3 个时，全部照录。超过 3 个，著录的前 3 个作者其后加“,等”（,et al）。欧美著者的名可用缩写字母，缩写名后省略缩写点，姓和缩写名全大写。用汉语拼音书写的人名，姓全大写，名可缩写，取每个汉字拼音的首字母。

参考文献为五号宋体，英文及数字为五号 Times New Roman 字体，两端对齐。参考文献中的标点符号均为英文标点，常用的参考文献著录项目和格式示例如下：

附录 A 附录名称

对于一些不宜放入正文中、但作为毕业设计（论文）又不可残缺的组成部分或具有重要参考价值的内容，可编入毕业设计（论文）的附录中，例如，正文内过于冗长的公式推导、方便他人阅读所需的辅助性数学工具或表格、重复性数据和图表、非常必要的程序说明和程序全文、关键调查问卷或方案等。

附录的格式与正文相同，如有多个附录需依顺序用大写字母 A，B，C，……编序号，如附录 A，附录 B，附录 C，……。只有一个附录时也要编序号，即附录 A。每个附录应有标题，如：“附录 A 参考文献著录规则及注意事项”。

附录一般与论文全文装订在一起，与正文一起编页码。

致　　谢

学位论文正文和附录之后，一般应放置致谢（后记或说明），主要感谢指导老师和对论文工作有直接贡献和帮助的人士和单位。致谢言语应谦虚诚恳，实事求是。字数一般不超过 1000 个汉字。

“致谢”用三号黑体加粗居中，两字之间空 4 个半角空格。致谢内容为小四号宋体，1.5 倍行距。