上 海 财 经 大 学

本 科 生 毕 业 论 文 开 题 报 告 表

姓　　名：付舟行

学　　号：2021111662

专　　业：工商管理（商务分析）

学　　院：商学院

指导教师：田中俊

2024 年 12 月 26 日填写

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
论文选题意义	资本市场作为现代经济的核心组成部分，其价格发现机制和资源配置效率受到多种因素的影响。经典金融理论，如有效市场假说，强调理性决策和信息的完全反映。然而，现实市场中普遍存在的过度波动、资产泡沫及市场异象，促使行为金融学理论应运而生，该理论认为投资者的心理因素，尤其是情绪，在资产定价中扮演着不可忽视的角色。投资者情绪，作为一种可能偏离基本面信息的集体性市场信念或预期，能够系统性地影响投资决策，进而驱动资产价格偏离其内在价值，对市场短期波动、长期趋势乃至稳定性产生深远影响。特别是在中国 A 股市场，个人投资者占比较高，市场情绪的影响可能更为显著。 与此同时，信息技术的飞速发展催生了海量的非结构化文本数据，如社交媒体讨论、在线股票论坛评论等。这些数据蕴含着丰富、实时、直接反映市场参与者观点的宝贵信息，尤其是投资者情绪。传统依赖间接市场指标（如封闭式基金折价率、IPO 数量、交易量等）或调查问卷的情绪度量方法，存在时效性差、覆盖面窄、易混入基本面信息等局限。自然语言处理（NLP）技术的崛起，特别是基于深度学习的预训练语言模型（如 BERT 及其变种），为直接、大规模、精细化地从文本大数据中量化投资者情绪提供了强大的技术武器，开启了金融研究的新范式。 （1）理论意义 本研究旨在深化对投资者情绪影响资产定价机制的理解。首先，通过运用先进的自然语言处理技术（具体采用 StructBERT 模型）直接从高频、海量的在线股吧评论中量化投资者情绪，本研究克服了传统情绪代理变量的局限性，为行为金融学提供了更直接、动态且微观的证据，尤其是在中国这一重要新兴市场背景下的实证支持。其次，本研究系统性地检验了基于 NLP 量化的投资者情绪对股票短期收益率的预测能力，并探究了其影响的行业异质性和时效性特征，有助于揭示情绪信息在中国 A 股市场的传导机制和价格发现过程，丰富和发展了资产定价理论。最后，在方法论层面，本研究验证了深度学习模型在处理复杂金融文本、从另类数据中萃取有效市场信号方面的有效性，为计算金融和金融科技领域的研究提供了方法论参考和技术路径验证。 （2）现实意义 本研究的成果具有显著的实践价值。对于投资者而言，本研究构建的情绪指标体系可作为投资决策的有效参考。通过实时监测和分析在线社区的投资者情绪，投资者能够更敏锐地捕捉市场短期预期变化，识别潜在的交易机会或规避由情绪驱动的非理性市场波动，尤其是在情绪影响显著的特定行业（如研究发现的计算机、电子、食品饮料、医药生物等），从而优化投资组合管理和交易策略。对于金融机构和分析师，本研究提供了一种新型的市场分析工具，将情绪维度纳入考量，有助于构建更全面的市场预测模型，提升市场分析和预测的准确性。对于监管机构，本研究的方法和发现有助于实时监测市场情绪的异常波动，识别潜在的过度投机、市场操纵风险或系统性恐慌情绪积聚，为维护市场稳定、保护投资者利益提供重要的早期预警信号和决策支持，促进资本市场的健康发展。

研究内容范围	**核心研究问题** 本研究的核心目标是探讨基于自然语言处理技术从在线股票论坛（东方财富网股吧）评论中量化的投资者情绪，是否以及如何影响中国 A 股市场股票的短期收益率，并考察这种影响是否存在行业差异和时间衰减效应。 **具体研究内容** 数据获取与处理： 利用 Python 网络爬虫技术，定向采集东方财富网股吧中特定股票（覆盖 A 股 8 个主要行业的 111 只代表性股票）在特定时期（2025 年 2 月）的投资者评论文本数据。同时，通过 Akshare 接口获取相应的股票日度交易数据（价格、成交量、换手率等）和行业指数数据。对原始文本数据进行严格的清洗和预处理，构建高质量的语料库。 投资者情绪量化： 应用先进的自然语言处理技术，具体采用基于 Transformer 架构的 StructBERT 深度学习预训练语言模型，对预处理后的每条股评文本进行情感倾向分析（正面/负面），并输出情感得分，实现对非结构化文本中蕴含的投资者情绪的自动化、精细化量化。 情绪指标体系构建： 基于单条评论的情感分析结果，在个股层面按日度频率构建多维度的情绪指标体系，不仅包括反映整体情绪水平的核心指标（如平均情绪得分、正面情绪比例），也探索情绪波动性、一致性及评论热度等辅助指标。 情绪影响的实证检验： 构建面板数据集，整合情绪指标与股票市场交易数据。运用计量经济学方法，重点采用面板数据双向固定效应模型，实证检验投资者情绪指标对未来不同期限（1 日、3 日、5 日）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。 异质性与稳健性分析： 进行分行业的回归分析，探究投资者情绪影响在不同行业板块间的差异。讨论并尝试处理潜在的内生性问题（如进行格兰杰因果检验）。通过替换核心情绪指标、考察不同预测期限的收益率等方式进行稳健性检验，确保研究结论的可靠性。 **研究方法** 文献研究法： 系统梳理国内外关于投资者情绪理论、情绪度量方法（特别是基于文本分析的方法）、情绪对资本市场影响（收益率、波动性、交易量等）以及自然语言处理技术在金融领域应用的相关文献，为本研究提供理论基础、研究视角和方法借鉴。 自然语言处理（NLP）与机器学习： 采用基于深度学习的 StructBERT 模型对大规模中文股评文本进行情感分析和量化。对比分析了 StructBERT 与 RoBERTa 模型的效果，最终选择 StructBERT 作为主要分析工具。涉及文本预处理、模型应用、情感评分生成等技术环节。 数理统计与计量经济学分析： （1）描述性统计： 对情绪指标、市场交易数据进行统计描述，了解数据分布特征。 （2）面板数据分析： 构建股票-日度面板数据，运用固定效应模型估计投资者情绪对未来股票收益率的影响，并解读系数的经济意义和统计显著性。 （3）检验与处理： 进行单位根检验（ADF-Fisher, IPS）、协整检验（Kao）、相关性分析、多重共线性检验（VIF）确保模型设定的合理性。进行格兰杰因果检验探讨变量间的动态关系。 （4）异质性与稳健性检验： 通过分行业回归、替换关键变量、改变被解释变量等方法，检验核心结论的稳定性和适用边界。

查阅主要文献	[1] Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. Journal of Finance, 59(3), 1259-1294. [2] Zhang, Y., Song, W., Shen, D., & Zhang, W. (2016). Market reaction to internet news: Information diffusion and price pressure. Economic Modelling, 56, 43-49. [3] Li, T., Van Dalen, J., & Van Rees, P. J. (2018). More than just noise? Examining the information content of stock microblogs on financial markets. Journal of Information Technology, 33(1), 50-69. [4] Chen, C. P., Tseng, T. H., & Yang, T. H. (2018, June). Sentiment Analysis on Social Network: Using Emoticon Characteristics for Twitter Polarity Classification. In International Journal of Computational Linguistics & Chinese Language Processing, Volume 23, Number 1, June 2018. [5] Chen, S., Zhang, W., Feng, X., & Xiong, X. (2020). Asymmetry of retail investors' attention and asymmetric volatility: Evidence from China. Finance Research Letters, 36, 101334. [6] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 26. [7] Lai, H. H., Chang, T. P., Hu, C. H., & Chou, P. C. (2022). Can Google search volume index predict the returns and trading volumes of stocks in a retail investor dominant market? Cogent Economics & Finance, 10(1), 2014640. [8] Booker, A., Curtis, A., & Richardson, V. J. (2018). Investor disagreement, disclosure processing costs, and trading volume: Evidence from investors who interact on social media. Social Science Electronic Publishing. [9] Chen, J., Tang, G., Yao, J., & Zhou, G. (2022). Investor attention and stock returns. Journal of Financial and Quantitative Analysis, 57(2), 455-484. [10] Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10‐Ks. The Journal of Finance, 66(1), 35-65. [11] 李岩, 金德环. 投资者情绪与股票收益关系的实证检验[J]. 统计与决策, 2018, 34:166-169. [12] 马勇, 杨雯葳, 姜伊晴. 投资者情绪如何影响公司股价[J]. 金融论坛, 2020, 25:57-67. [13] 宋泽芳, 李元. 投资者情绪与股票特征关系[J]. 系统工程理论与实践, 2012, 32:27-33. [14] 王春. 投资者情绪对股票市场收益和波动的影响——基于开放式股票型基金资金净流入的实证研究[J]. 中国管理科学, 2014, 22:49-56. [15] 王美今, 孙建军. 中国股市收益、收益波动与投资者情绪[J]. 经济研究, 2004, 10:75-83. [18] 闫伟, 杨春鹏. 金融市场中投资者情绪研究进展[J]. 华南理工大学学报（社会科学版）, 2011, 13:33-43. [19] 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, 7:13-17. [20] 赵汝为, 熊熊, 沈德华. 投资者情绪与股价崩盘风险：来自中国市场的经验证据[J]. 管理评论, 2019, 31:50-60.

<table>
<tr><td rowspan="9">进度安排</td><td>序号</td><td>论文各阶段名称</td><td>日期</td><td>备注</td></tr>
<tr><td>1</td><td>问题定义</td><td>2024.12.26-
2024.12.31</td><td>明确研究问题和目标。</td></tr>
<tr><td>2</td><td>文献阅读</td><td>2025.01.01-
2025.01.05</td><td>了解已有的研究和理论基础，完成文献综述。</td></tr>
<tr><td>3</td><td>研究设计</td><td>2025.01.06-
2025.01.10</td><td>确定研究方法和框架。</td></tr>
<tr><td>4</td><td>数据收集&预处理</td><td>2025.01.11-
2025.01.15</td><td>爬虫收集数据，完成数据预处理，以准备数据分析。</td></tr>
<tr><td>5</td><td>数据分析</td><td>2025.01.16-
2025.01.31</td><td>使用机器学习算法、回归等对数据进行分析。</td></tr>
<tr><td>6</td><td>论文初稿撰写</td><td>2025.02.01-
2025.02.07</td><td>根据分析结果撰写论文初稿。</td></tr>
<tr><td>7</td><td>论文修改与完善</td><td>2025.02.08-
2025.02.17</td><td>根据导师的指导进行论文修改，论文查重与格式审查。</td></tr>
<tr><td>8</td><td>论文定稿与答辩准备</td><td>2025.02.18-</td><td>完成论文的最终修改，准备答辩演讲稿。</td></tr>
<tr><td>目前进展情况</td><td colspan="4">明确研究问题，进入文献阅读阶段

学生签名：
2024 年 12 月 30 日</td></tr>
<tr><td>是否同意该生进入论文工作阶段</td><td colspan="4">同意

导师：
2024 年 12 月 31 日</td></tr>
</table>

注：本表由学生填写，经指导教师和学院相关负责人审阅后，由各学院备案。