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。

1. 数据来源与处理
2. 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 |

1. 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。

1. 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。

1. 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。

1. 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 |

1. 情绪分析模型构建
2. 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。

1. 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。

1. 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。

|  |  |
| --- | --- |
|  |  |
| 图3-13 情感分布（StructBERT） | 图3-14 情感分布（RoBERTa） |

1. 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。

1. 模型设定及变量说明
2. 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：

(1)

其中，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；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。

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。