投资者情绪的识别与统计
——基于非结构数据的分析

耿晓媛

摘　要：投资者的情绪特征主要来自于投资者的心理认知，基于行为金融学的前景理论、期望理论与后悔理论，深入细致地分析了影响投资者情绪的内涵，进而编制投资者情绪特征指数进行测度，以此研究投资者情绪特征指数在行为金融学中的影响机制。同时应用爬虫技术爬取相关投资者情绪的数据，经过合理清洗、筛选与朴素贝叶斯处理等过程，对投资者情绪特征进行统计实践。研究发现投资者情绪特征具有可甄别性、可预测性；基于对投资者情绪特征变量的统计测度方法的分析，论证这些统计测度方法不仅可以实现对投资者风险偏好特征的提取与测度，而且可行；积极情绪投资者对投资风险性较高的新兴行业板块更加青睐，且与消极情绪投资者呈现出显著的分布差异，但是积极情绪投资者与消极情绪投资者在投资收益率方面并无显著差异。

关键词：非结构数据；投资者情绪；机器学习；朴素贝叶斯

DOI：10.3773/j.issn.1006-4885.2023.11.156

中图分类号：F830.91　**文献标识码：**A　**文章编码：**1002-9753（2023）11-0156-14

1　引　言

行为金融理论是一个非常宽泛理论体系，它包含了多种不同的理论，比如前景理论等，该理论主张，多倾向于以投资人的风险偏好表示其行为选择，以此为切入点深入细致地分析经济现象后所存在的理论机制，尤其是研究经济现象发生过程中相关方的行为及其对经济现象产生的各种影响；不仅如此，还对未知环境下行为人做出的决策进行深入全面地分析，使得经济学的解释能力得到进一步提升。不过，受各种因素的制约，个人投资者风险偏好的精准化、实时化测度始终都未得到有效攻克。在大数据发展环境下，基于结构特征的不同，可将数据大体划分为三类，分别是结构性、半结构性、非结构性数据。在传统统计学分析中，其分析对象皆是比较简单且便于理解的结构性数据，也就是我们通常所讲的狭义数据。在大数据发展环境下，统计研究中涉及到数据结构多样，涵盖了上述三类不同结构形式的数据。其中，结构性数据能够以二维表的形式进行清晰直观地排列与描述，往往以统一的格式安全规范地存储于数据库内；非结构性数据简单来讲指的是无法按照一致的结构形式进行排列和描述的数据信息，往往指的是利用网络社交等获取的信息，譬如图像、

基金项目：国家社会科学基金一般项目（项目编号：19BTJ057）。

作者简介：耿晓媛（1981 年 –），女，黑龙江大庆人，黑龙江八一农垦大学经济管理学院教授，研究方向：金融风险技术分析，应用统计、技术经济及管理、会计理论与实务等。

视频等：半结构性数据是介于上述两种结构形式之间的信息，其结构与内容并未明确划分，而是相互融合，譬如 E-Mail 或者网页新闻等。大数据通过大规模、多元化的数据结构为人们分析个体决策提供了重要依据，借助数据挖掘等一系列先进成熟的前沿技术，将能够反映个体投资者风险偏好的大规模非结构数据信息科学合理地转变为可以体现其心理波动变化的结构性信息，增强样本数据的有效性。因此，本文以非结构性数据为研究的切入点进行研究。

现有关于投资者情绪识别与统计的方法可以分为以下三类：一是使用直接调查的主观指标。该方法是通过问卷调查方式，直接调查投资者对股票市场未来行情走势的判断或对未来经济和投资前景的信心状况来直接反映投资者情绪，如投资者智能指数、央视看盘指数、巨潮投资者信心指数（黄燕芬等，2019[1]；王德青等，2021[2]）。二是使用市场公开交易数据构建的单一客观指标。这种方法是采集金融市场中与情绪有关的公开交易数据并运用的单一客观指标来间接反映投资者的心态和情绪。如首日公开发行（IPO）发行量及首日收益、市场换手率、封闭式基金折价率、共同基金净赎回、非主力资金在股票或股票组合的买卖不平衡程度等（吴海燕和杨朝军，2012[3]；贺刚等，2018[4]；何诚颖等，2021[5]）。三是基于多项市场指标构建的复合指标。这部分研究主要是指采集多项市场中与情绪有关的公开交易数据并构建的复合指标来间接反映投资者的情绪状况，此方面较为典型的是 Baker 和 Wurgler（2006）[6] 创建的 B-W 方法和易志高和茅宁（2006）[7] 通过主成分分析构建的中国股票市场投资者情绪的综合指数 CICSI。但是第一类度量指标由于问卷设计和调查群体等因素影响具有较强主观性，而第二类度量指标虽然能够客观反应投资者的情绪心理，却忽视外在的宏观经济因素影响，因而前两种指标不能全面准确地测度情绪变化，学者们大多采用第三类复合指标构建情绪指标。此外，随着互联网大数据的发展，基于网络搜索引擎（张永杰等，2018[8]）和社交媒体平台数据（Piñ 等，2021[9]）构建投资者情绪指数的研究日益丰富。如谷歌搜索引擎指数、百度搜索引擎指数、新闻情感指数以及财经门户网站的股票论发帖数量（Gao 等，2020[10]；卢米雪，2022[11]；高扬等，2023[12]）。其中雪球论坛是我国最典型的财经论坛的代表之一，其以用户群体素质更高、号称“聪明的投资者都在这里”而闻名（陈浪南和苏湃，2017[13]；熊艳，2022[14]）。

目前国内学术界大多数的研究还是基于传统指标进行的，基于互联网社交媒体指标探究其对股市影响的研究还较少。

目前为止，在研究投资者情绪识别方面的文献中，主要是以传统的结构数据进行分析的，鲜少有以非结构数据为切入点研究投资者的风险偏好，基于此，本文关于投资者情绪方面研究的切入点为非结构数据。

通过科学合理的方法将大量非结构性数据灵活合理地转变为用于传统统计分析的结构性数据，促进统计学方法在更多领域和行业得到大力推广和积极应用。本文的分析可知，选择非结构性数据进行关于行为金融的分析更具有经济性，因此本文将深入细致的分析影响投资者风险偏好的主要因素中非结构性数据进行识别与合成，比如注意力、情绪等，同时对上述因素间的关系展开全面深入地研究，准确合理地明确投资人风险偏好的作用机理，为后续研究做好基础。

2 投资者情绪特征变量的识别

2.1 投资者情绪的内涵界定

情绪是一个非常抽象的概念，可将其视为各种主观认知经验与体会的结合体，是随着人们成长和发展而形成的一种独特心理现象。Izard（1977）[15] 在进行深入细致的分析之后表示，情绪是一个非常复杂的概念，不仅涉及到了认知元素，同时也涵盖了对外界事物主观评判。在人格系统中，情绪是不可或缺的重要元素，呈现出明显的个体差异性特征。无论是外界因素，还是个人主观因素，它们均会对情绪产生非常大的影响，可以说，情绪不仅是个体对相关事物的一种主观体会，同时也是对外界事物的一种客观生理反应，是一种较为独特的社会表达方法。另外，该学者还明确指出，情绪涉及到三大因素，一是主观体验，二是生理基础，三是表情行为。情绪不仅表现出较强的生物学价值，并且社会性色彩非常浓郁。Levin 等（2003）[16] 在对情绪展开深入细致的分析后表示，人们对某种事物产生的情绪较为复杂，会因外界事物的性质不同而呈现出较大差异。80 年代后，随着神经成像技术的日益发展及广泛应用，人们通过此技术了解到，神经元回路能

够有效调节情绪，它们在对接收到的数据进行处理后会输出相应的情绪。

从行为金融学的层面来讲，情绪可以说是个体对外界事物进行感知判断之后而形成的一种产物。情绪既能够是个体生理特征的表现，也可以是行为影响结果。投资者的情绪是不固定的，会受到诸多因素的影响与干扰，比如人格特质，它可以被视为最具代表性的投资者风险偏好，即便是在相同的环境下，面对同样的事与物，投资者的情绪也会因人格特质的不同而呈现出明显差别。

2.2 投资者情绪的特征变量提取

在行为金融学当中，情绪属于相当关键的指标之一，它不仅会对个体的投资者风险偏好产生较为深刻的影响，在行为金融学当中同样有着非常关键的地位。Long 等（1990）[17]在经过深层次的探究以后，创建了经典的噪声交易模型（DSSW），它着重介绍了外生有偏信息的投资者交易行为，对噪声交易者在金融领域的立足能力展开了深入细致地研究。该模型能够依托于对噪声交易者特征的全面揭示及系统研究，科学合理地描述其对金融资产定价所产生的重要影响，并且针对其创造可观回报的原因进行了科学合理地阐释。一般来讲，在外生信息方面处于不利位置的交易者即为噪声交易者；噪声主要是指市场运作过程中私自捏造或者被曲解的数据，它是造成股价和其自身价值之间存在不同的重要因素。在市场领域，交易者往往会依托于其采集到的数据对市场变化趋势、股价走向进行科学合理地自主判定，尽管内在价值不会随着人的意志转变而发生变化，其形成时间明显在于价格，不过在现实交易中，噪声交易者对于价格的曲解会形成“共识”。这种共识源于大量投资者心理的波动，内在价值和价格之间所存在的差值即为噪声。

在行为金融学研究过程中，投资者情绪可被视为一个重要的参考指标，利用情绪指数能够全面系统地分析情绪和行为人决策间的关系。以噪声交易模型为例进行说明，该模型主张，噪声会造成产品价格与其实际价值之间产生一定差别，可将其用于对风险资产后期回报的合理评估，大量研究证实，情绪的波动势必会引起资产价格变化。对于损失的未知性，投资人一般会因情绪的随机波动而呈现出较大的认知偏差，对其选择行为产生进一步影响。Arkes 等（1988）[18]在进行深入细致地分析后表示，当保持乐观积极的情绪时，投资者往往会夸大市场行情。Kaufmann 等（1997）[19]在对相关研究成果进行系统全面地梳理后，通过广泛深入地研究发现，无论是投资者的情绪，还是其心理变动，均会对投资选择产生较为显著的影响。诸多心理实验研究证实，情绪是一个非常宽泛且抽象的概念，不仅使得认知风险水平发生变化，也会使得预期收益率出现相应的改变。Hastie（2001）[20]在进行深入细致地分析后表示，情绪的波动势必会对投资者决策产生较为显著的影响。目前，在行为金融学研究领域，投资者情绪如何影响经济行为成为备受业内人士高度关注的重要研究课题。

2.3 投资者情绪的统计测度——情绪指数

作为一种能够反映情绪变化的重要参数，情绪指数不仅能够对投资者的情绪变化情况进行科学合理地度量，也能够对社会情绪指数的变化情况进行科学合理地度量。作为一种能够有效量化且应用比较广泛的统计数据，情绪指数可被用于评估人体情绪变化情况。在实际应用中，可依托于创建情绪指标体系的方式深入细致地分析各个变量间的关系，基于此构建严谨合理的分析模型，在进行严格规范地归一化处理后，便捷高效地求解出相应的情绪指数。其求解步骤如下：

根据前文所知情绪是一个比较抽象且复杂的概念，基于不同的方法或者按照不同的标准，可将其划分为不同的类型，根据各种子情绪，科学合理地构建指标体系。

情绪从长期的角度来讲具有一定的稳定性。所以，当前情绪和稳定状况下的情绪之间所产生的偏差即为我们通常所讲的情绪波动，基于此进行科学合理地计算与分析，有效剔除由于各种情绪占比不同而产生的量纲偏差。若 x_i 为第 i 种情绪的测度值，e_i 为行为人投资者第 i 种情绪比重，即式（1）、（2）与（3）：

$$e_i = \frac{X_i}{\sum_{i=1}^{\infty} X_i} \tag{1}$$

$\overline{e_i}$ 表示情绪的一般水平值，第 i 种情绪的波动值是：

$$\tilde{e}_i = \frac{e_i - \overline{e}_i}{\overline{e}_i} \quad (2)$$

通过方差测定情绪，以此描述情绪的变化水平即：

$$D(X) = \sum_{i=1}^{\infty} (e_i - \overline{e}_i)^2 \quad (3)$$

综上，均值 - 方差组合可以对情绪波动程度进行合理测定。

第一，确定权重。应针对各种不同的情绪科学合理地确定权重。

第二，求解情绪指数，见式（4）：

$$F_1 = \text{情绪指数} = \frac{\text{期望实现值}}{\text{内心期望值}} \quad (4)$$

通常来讲，个体的期望值愈高，其情绪指数愈小，此时积极情绪较为严重；反之，情绪消极情绪较多。期望值愈高，情绪冲突愈强，通过成功而获得的满足感愈小，但是基于失败而造成的消极情绪愈严重。整体来讲，情绪指数能够对情绪变化水平进行合理测定。

3 投资者情绪指数的构造

本文在分析时所采用的数据主要源于雪球网，其收集整理了大量社交数据、交易数据，也就是前文提到的“社交投资平台”。结合“社交投资平台”的有关含义可知，雪球网除了具有“信号发送者”以及“信号接收者”之前，还涉及到“复制交易”的有关功能。本文在分析时使用的数据全部来源于雪球网，同时采用 Python 爬虫技术中的开源爬虫框架 Scrapy 对该网站上的用户好友关系、雪球组合、雪球实盘等相关数据与记录进行抓取。获取的基本思路是在遍历雪球网的 120 万雪球数据时，剔除掉其中已经关闭的组合，总计抓取 794151 个雪球组合和雪球实盘，并将这些信息进行保存，保存的信息包括组合的 ID、有效性、当前状态下的组合收益、创建日期、关注度、投资行为、投资版块、情绪、挂单数据与实际交易数据等。本文在分析时采用的数据全部为雪球组合与雪球实盘。在这样的社交网络投资平台中，往往把信号发送者界定为主动交易者的身份，而信号接收者往往进行的是“复制交易”。

3.1 数据清洗与筛选

雪球组合创立于 2014 年 10 月 20 日，雪球实盘则创立于 2016 年 6 月 27 日。因此本文数据爬取的时间起点为 2016 年 7 月 1 日，截至到 2020 年 6 月 30 日。本研究总计抓取了 794151 个雪球组合与雪球实盘（不涉及已经关闭的组合），包含所有雪球组合 782603 个，所有雪球实盘的共 11547 个。雪球网曾经公开表示，其平台中大概拥有 70 万个处在存续状态下的雪球组合与大约 1.02 万条处在存续状态下的雪球实盘，倘若把本文抓取的量与雪球公布的官方数据进行对比，由此可知本文抓取的数量与现实状况相符，同时大体上可以评估且不会出现大量遗漏的问题。

（1）存续期筛选

本文将其存续期都界定为首次买卖与最后一次交易跨越的时段。不过，由于本文选取的一部分雪球组合与实盘在形成和完成的时间上相隔只有几天，因此通常这些功能均属于用户持有尝试的心理，开启了组合或者实盘的功能并及时关掉的，这些组合已经没有办法反映出投资人的投资理念，应予以全部删除。其次，有些组合刚刚建立，并未有足够长期的历史业绩能够用于对其做出评估。在数据清洗的过程中，本文发现了样本存在一些非正常值与错误，其中多数原因是由于网站后台的数据分析出现有误，比如有一些组合在不到二个月时间里净值增长了一千倍，少部分则是在抓取的过程中形成的。综上，我们把存续期少于六十天的雪球组合和雪球实盘从统计中删除。

（2）交易市场筛选

雪球组合准许客户在 A 股、港股以及美股中任选一个市场展开投资，不过同一个组合仅可以在一个市场投放，不能同时在两个市场投放。历史数据证明，A 股组合中所占的比重最高，接近 95%。据此，本文

研究将着重关注 A 股市场，在雪球实盘上目前基本还是只支持 A 股，从而剔除了在美股和香港市场的投资组合。

（3）收益组合筛选

在数据抓取过程中发现，一些用户的雪球实盘公司尽管尚未关闭，但实际上在一至二日之间就对相关的财产进行了处理与清空，此种等同于退出雪球实盘的情形会造成净值的急剧下降，从而造成异常负收益率的产生，为防止此类极端情形出现，本文采用首末两段 1% 的缩尾方法对收益组合进行筛选。

（4）交易行为筛选

在雪球组合中，除了对用户的主动调仓状况展开记录之外，同时还对各种非用户因素（诸如分红除权等）产生的持仓变动进行记录。考虑到非用户因素引发的调仓行为无法反映用户在资金方面的真实意愿，所以这种交易也将被删除。

除此之外，也有部分用户做出的调仓行为是因为市价不满足等原因而无法完成交易，所以雪球网的后台会将这类行为标记成异常的调仓记录。鉴于此，本文在分析时也删除了这类交易。

另外，本文中还规定，进入样本的雪球组合或者雪球实盘都应当具有全部的基础信息、调仓历史记录、业绩历史记录三部分数据，并且进入样本的用户也应当已经持有了某个雪球组合或雪球实盘，以及该组合和实盘的基础信息都齐全。

通过上述的数据处理后，得到的组合共计有 63.5 万个，其中大都是雪球组合，数量达到 62.6 万个，雪球实盘有 0.9 万个。

3.2 描述统计分析

（1）调仓统计

本文认为每变动一个证券都算一个调仓行为。假设在一个组合中对多个证券进行操作，则每只证券的买卖行为均算一个调仓。如果在同一个组合中，买进 A 股，抛出 B 股，则其属于两次调仓行为。此种调仓记录方法不仅可以很好的体现调仓的频率，同时还能够展现出调仓中涉及的股票数。

如果调仓后，该股票在组合中的比重发生变化，则前后变化的比重为调仓幅度，即调仓后该股票的比重 - 调仓前该股票的比重。这样，以相对数形式衡量的调仓幅度不但有利于投资组合间相互对比，而且同时取得预测值之后可避免买进和卖出的调仓相互对抗。

通过抓取后的数据发现，雪球实盘用户会在间隔 15 天内展开一次调仓，而雪球组合用户会在间隔 149 天内展开一次调仓。从调整幅度上分析，雪球组合用户大概日均调仓幅度为 15.4%，而雪球实盘用户更高，达到 20.1%。

根据调仓行为数据，本文在捕捉与刻画用户的行为时需要用到月数据，甚至是周数据。如果按月数据口径下，本文会得到 48 个月维度下的数据，；如果按周数据口径下，本文会得到 200 个周维度下的数据。本文后续的分析中主要采用周维度下的数据。

（2）用户行为统计

根据本文清洗后的数据可知，这些用户中男人占据了大多数，差不多是女生数量的十倍，特别是在雪球实盘，其男用户数量差不多是女生用户的二十倍。用户数最多的五大地区，分别是 : 广东、北京、上海、浙江以及江苏。

雪球网当中设置了 5 项可以反映出用户行为的重要变量，分别为 : 一、关注数，其是指此用户关注别的用户的数量；二、粉丝数，其是指关注此用户的其它用户数量；三、自选股数，用户关注的股票数量，不但涉及个股，且包括组合；四、掌握的组合数；五、发帖数。在雪球平台上用户的发帖量，不但涉及股票方面的评论，还涉及组合调仓方面的阐述。前二项体现出不同用户间的互动情况，剩余三项反映出用户在股票间的互动情况。

纵观这五个反映用户活动程度的变量中，可总结出以下二个特征：首先实盘用户通常要比组合用户更为活跃。通过爬取的历史数据显示，实盘用户平均关注 86 个其它用户，而组合用户的平均关注数却只有

45；又比如，实盘用户平均有 112 个自选股，而组合用户平均有 53 个自选股，同样是组合用户的二倍。

用户活动的统计结果表现出相当显著的左偏分布特点，也就是活跃用户有很多粉丝、关注很多股票，同时主动参加评论，但实际上这些用户均是“沉默的”，即“马太效应”。基于此，为保证本文后续研究地顺利进行，本文采取随机抽样的方式随机抽取 100 位雪球用户为样本，其描述统计结果，具体见表 1。

表 1　Python 爬虫技术爬取的数据进行描述性统计

	雪球组合	雪球实盘	总计/平均
总人数（人）	57	43	100
关注数（平均）（个/人）	33	62	51
粉丝数（平均）（个人/）	190	358	203
自选股数（平均）（个/人）	56	107	56
拥有组合个数（平均）（个/人）	8	10	9
发帖数（平均）（个/人）	190	710	75

注：表中的总计为雪球组合与雪球实盘两组数据加总；平均为以雪球组合与雪球实盘两组各指标数据为基础数据、各组人数为权数并做加权平均处理。

对比表 1 与上述总体数据的结果，表 1 中的数据基本符合总体数据的趋势，因此，为本文后续分析结果的可靠性提供了保障。

（3）收益统计

本文针对雪球组合与雪球实盘的整体收益展开分析，并主要采用了如下二种指标来表现收益：总体收益率和年化收益率。

其中：总收益计算见式（5）。

$$\text{总收益} = (\text{当前的净值} - 1) \times 100\% \quad (5)$$

年化收益是借助复利法展开计算的。对建立还不到 1 年的组合，其年化收益与收益总额相等。为避免异常值的影响，本文将针对首尾进行 1% 的缩尾处理，也就是将周收益当中最大的 1% 和最小的 1% 进行剔除。因为尽管在 IPO 的过程中签会增加投资收益，但是中签结果却无法反映投资人的实际投入水平，所以本文将在 IPO 中签的实盘用户从样本中去掉。

本文计算周数据口径下的收益结果。为了直接体现收益率和雪球用户对投资组合中的风险偏好之间的关联，即越风险偏好型投资人所关注的组合就越多。为此，本文还给出了有五人及以上关注的组合及有一人及以上关注的组合的实际收益。结合下表 2、3 与 4 可知，雪球组合与实盘的收益率都具有着以下特征：

首先，雪球实盘的收益率高于雪球组合，且方差也更大。其次，关注的数量也和收益率成正有关。很显然，收益越多，便会吸引越多的人关注。诸如对于不少于 5 人关注的雪球组合，其年化收益的中位数为 74.8%，而对于所有样本总体而言，其收益的中位数是 25.1%，仅为 1/3。

表 2　雪球组合和雪球实盘的收益（年化收益，%）

	最小值	中位数	平均数	最大值	方差	标准差
雪球组合（ZH）	−61.9	1.6	1.3	56.48	25	5
雪球实盘（SP）	−61.7	8.05	25.1	51.92	157	12.53
总计/平均	−61.6	1.8	1.6	56.48	26.2	5.12

表 3　雪球组合和雪球实盘的收益
（5 人及以上人数关注的组合的总收益率，%）

	最小值	中位数	平均数	最大值	方差	标准差
雪球组合（ZH）	–61.9	24.6	38.4	111.15	94	9.70
雪球实盘（SP）	–53.9	29.4	74.8	36.34	230	15.17
总计/平均	–61.9	24.6	38.4	111.15	94	9.70

表 4　雪球组合和雪球实盘的收益
（1 人及以上人数关注的组合的年化收益率，%）

	最小值	中位数	平均数	最大值	方差	标准差
雪球组合（ZH）	–61.9	15.2	19	33.88	40	6.32
雪球实盘（SP）	–53.9	29.4	74.8	36.34	230	15.17
总计/平均	–61.8	15.6	19.8	33.88	42	6.48

本文针对半月收益同样实施了缩尾处理，经过处理以后总计拥有 1.5 万条数据用于观测，结果见表 5。

表 5　雪球组合与实盘的月数据口径下的收益结果

	最小值	中位数	平均数	最大值	方差
雪球组合（ZH）	–7.9	1.33	1.36	6.66	1.69
雪球实盘（SP）	–6.7	1.22	1.15	6.37	2.09
总计/平均	–7.89	1.33	1.36	6.66	1.7

3.3　指数编制数据的预处理

依托网络爬虫技术取得的数据属于文本数据，同时其中含有大量不完整、重复以及没有价值的数据，所以必须对这些文本数据实施预处理。具体来说，是针对庞大的数据实施整理、筛选等处理方式，主要是将重复和不完整的数据进行剔除，将大小写字母、繁简字进行统一，剔除和用户情绪之间没有关联的帖子，剔除没有价值的字段，剔除停用词等。同时，还做了文本分词处理，也就是将文本数据中的句子拆分成对应的词语，以确保机器学习模型可以正确的判断各个句子的意思。

通过以上这些预处理，让数据噪音大大减少，实现对数据的初步筛选，对后续进一步挖掘评论当中的情绪有所帮助。将本文最后的数据通过以上预处理流程，剩下 3.89 万条文本数据。

3.4　基于机器学习模型的投资者情感分类

在展开分析前，需要对文本数据展开预处理，剔除数据中的无用信息，进而从整体上提升数据的有效性。随后开始构建情感分类模型，其能够对每条评论数据中包含的情绪展开判断，进而给投资者的情绪指数提供必要的支持。

（1）人工标记与训练集的实践

机器学习是模型结合已掌握的数据集来找寻最佳参数的过程。为了让模型分类效果实现更加准确，挑选训练集变得非常关键。因为本文取得的数据是用户的原始评论，而不同评论和特定情绪之间并未呈现出一一映射，也不存在现成的训练样本，对于帖子中情感极性缺少评判标准，所以本文利用人工标记的形式来取得训练集，用于机器学习。

本文运用随机抽样法由通过预处理之后的数据当中，按照随机原则选取了 1 万条。依据二分类原则对这

些数据加以标记：其中，标记值为 1 代表投资者对当前的风险偏好情绪比较积极，其中根据对投资倾向的幅度不同加以区分，程度由弱至强分别为“大盘无忧”、“看涨、涨”以及“大胆加仓”，在标记为“1”类别中分别赋值为 1、2 和 3；标记值为 -1 代表投资者对当前的风险偏好情绪比较消极，其中根据对投资倾向的幅度不同加以区分，程度由弱至强分别为“下跌、跌”、“逃”以及“割肉”，在标记为“-1”类别中分别赋值为 -1、-2 和 -3。投资者的情感分类词典是结合现有的文献研究成果经整理而成，见表 6，据此标记从预处理后数据中的 1 万条评论。

表 6　投资者投资情感分类词典词汇示例

情感分类	类别	情感词典词汇示例	赋值
积极情感	1	大盘无忧	1
		看涨、涨	2
		大胆加仓	3
消极情感	–1	下跌、跌	–1
		逃	–2
		割肉	–3

（2）基于朴素贝叶斯的评论文本投资者投资情感分类的实现

应用留出法把以上人工标记评论文本当中选取 8000 条当作训练集，剩余 2000 条当作测试集，针对投资者的情感分类展开精度检验，同时符合：人工标记样本训练集 $\cup$ 测试集 = 全集，训练集 $\cap$ 测试集 = Φ。剩余的 2.89 万条评论划分成文本集 15 个，再分别使用训练完的模型展开测评。

基于多项式朴素贝叶斯模型展开对应的预测分类。针对分类任务而言，在已知全部有关概率的情形下，该模型考虑怎样结合上述概率与误判损失来筛选最佳的类别标记。朴素贝叶斯分类器遵循属性条件独立性假设原则，也就是说针对已知事物的类别，倘若全部属性彼此独立，同时各个属性可以对分类结果产生影响，具体见式（6）：

$$P(c|x) = \frac{P(c)P(x|c)}{P(x)} = \frac{P(c)}{P(x)}\prod_{i=1}^{d} P(x_i|c) \quad (6)$$

其中，d 代表属性类型，x_i 是在第 i 个属性上的取值。因为针对全部类型而言，$P(x)$ 一样，所以在贝叶斯基础上的评判标准为式（7）：

$$h_{nb}(x) = \underset{c \in y}{\arg\max}\ p(c)\prod_{i=1}^{d} p(x_i\ c) \quad (7)$$

基于 Python 语言建立多项式朴素贝叶斯模型，从标注的 10000 条数据中选择 8000 条展开分类学习，剩余的 2000 条当作精度验证。表 7 给出了模型分类的精度：

表 7　基于多项式朴素贝叶斯算法的精度表

文本集	准确率
训练集	0.8879
测试集	0.8513

通过表 7 可知，基于多项式朴素贝叶斯算法的训练集与测试集的准确率均大于 85%，预测的投资者情感分类效果比较理想。

然后，本文在随机选取 1 万条数据之后剩余的 28900 条评论当中，划分成 15 个文本集，各个文本集大概包括两千条评论，分别使用训练好的模型展开投资者情感分类预测。当中八千条训练集（人工标记）、两千条测试集（人工标记），28900 条预测集（机器自动标记），总共 38900 条文本数据，最后把所有评论划分成“满意”与“不满意”2 个文本集合，具体统计情况见表 8。

表 8　总体评论文本投资者情感分类分布情况

评论文本情感	赋值	数量	比重（%）
积极（占63.08%）	1	9813	39.99%
	2	8785	35.80%
	3	5940	24.21%
	小计	24538	100.00%
消极（36.92）	–1	5231	36.42%
	–2	4568	31.81%
	–3	4563	31.77%
	小计	14362	100.00%
合计		38900	100.00%

根据前述式（1）、（2）、（3）可以分别计算不同类别情绪投资者的方差，分别为投资者积极情感的方差 σ_1^2 为 0.6170，投资者消极情感的方差 σ_2^2 为 0.6798。

根据公式（7），投资者的内心期望值可用投资者交易投资时的挂单价格 P^e 进行衡量，本文将挂单价格定义为投资者在交易时委托系统提交的依据委托价格计算的交易金额，交易者提交的委托价格即为投资者的内心期望值，投资者的期望实现值可用投资者交易时的真实价格 P 进行衡量，因此将式（7）进行改进，见式（8）：

$$F_1 = \text{情绪指数} = \frac{\text{期望实现值}}{\text{内心期望值}} = \frac{\sum P\sigma_i^2 C_i}{\sum P^e \sigma_i^2 C_i} \quad (8)$$

其中，σ_i^2（$i = 1,-1$）为投资者情感为积极与消极时的方差，C_i（$i = 1,-1$）为投资者情感为积极与消极的不同类别。

根据对 100 位投资者 38900 条评论中的情感提取，依据式（8）计算 100 位投资者分数不同类别中投资者的情绪指数，并将 100 位投资者的情绪指数在时间维度上进行平均数处理，投资者个体投资者情绪指数参见表 9，投资者投资情绪指数在个体维度和时间维度的综合均值如表 10 所示。

表 9　投资者个体投资情绪指数

序号	类型	情绪指数	序号	类型	情绪指数	序号	类型	情绪指数	序号	类型	情绪指数
1	积极	79%	26	积极	57%	51	积极	86%	76	消极	102%
2	积极	91%	27	积极	77%	52	积极	89%	77	消极	109%
3	积极	82%	28	积极	57%	53	积极	79%	78	消极	134%
4	积极	96%	29	积极	69%	54	积极	88%	79	消极	117%
5	积极	98%	30	积极	92%	55	积极	85%	80	消极	108%
6	积极	81%	31	积极	97%	56	积极	91%	81	消极	129%

续表

序号	类型	情绪指数	序号	类型	情绪指数	序号	类型	情绪指数	序号	类型	情绪指数
7	积极	79%	32	积极	93%	57	积极	92%	82	消极	102%
8	积极	89%	33	积极	87%	58	积极	82%	83	消极	104%
9	积极	85%	34	积极	85%	59	积极	79%	84	消极	150%
10	积极	87%	35	积极	82%	60	积极	67%	85	消极	129%
11	积极	67%	36	积极	96%	61	消极	135%	86	消极	139%
12	积极	92%	37	积极	87%	62	消极	115%	87	消极	116%
13	积极	71%	38	积极	99%	63	消极	136%	88	消极	169%
14	积极	94%	39	积极	87%	64	消极	108%	89	消极	121%
15	积极	86%	40	积极	96%	65	消极	115%	90	消极	124%
16	积极	74%	41	积极	58%	66	消极	167%	91	消极	101%
17	积极	95%	42	积极	96%	67	消极	117%	92	消极	102%
18	积极	74%	43	积极	87%	68	消极	128%	93	消极	101%
19	积极	69%	44	积极	81%	69	消极	129%	94	消极	109%
20	积极	68%	45	积极	97%	70	消极	118%	95	消极	126%
21	积极	82%	46	积极	95%	71	消极	136%	96	消极	136%
22	积极	91%	47	积极	79%	72	消极	147%	97	消极	125%
23	积极	70%	48	积极	89%	73	消极	119%	98	消极	117%
24	积极	69%	49	积极	97%	74	消极	167%	99	消极	128%
25	积极	59%	50	积极	83%	75	消极	136%	100	消极	146%

表 10　投资者投资情绪指数均值表

情感分类	积极	消极	总体
人数	60	40	100
情绪指数	0.83	1.25	0.99

根据表 9 中的结果可知，积极投资情绪投资者的投资情绪指数小于 100%，消极投资情绪投资者的投资情绪指数基本上均大于 100%。根据行为金融理论，投资者的消极心理的作用相较于积极心理的作用更为明显，表 10 的结果也证明了这一点，投资者的消极情绪更具有渲染色彩，表现力更强。

为确认投资者情绪数据的稳定性，本文针对表 9 中的数据，借助 ADF 统计量展开稳定性测试，具体结果见表 11。

表 11　投资者投资情绪指数稳定性检验

	ADF统计量	5%临界值	P值	结论
情绪指数	–7.126	–2.867	0.0001	平稳

基于稳定性检验角度，投资者情绪指数的 ADF 统计量达到 -7.126，在 5% 的置信水平下，

其阈值是 -2.867，可见其已经超过了阈值，表明投资者情绪指数是稳定的。

4　投资者的情绪指数与投资板块、收益率的关系

4.1　投资者的情绪指数与投资板块的关系

证券市场由不同板块股票构成，但在不同板块间股票的投资风险却并不相同。因此，为验证不同情绪组在风险不同的板块之间投资行为是否存在差异，本文将爬取的 100 位投资者在不同板块交易的交易数据进行整理，发现这 100 位投资者投资行为主要集中于银行、保险、消费、医药、食品及新兴行业等板块；结合之前的分析，本文分别将不同投资情绪投资者的投资板快分布进行整理，见图 1。

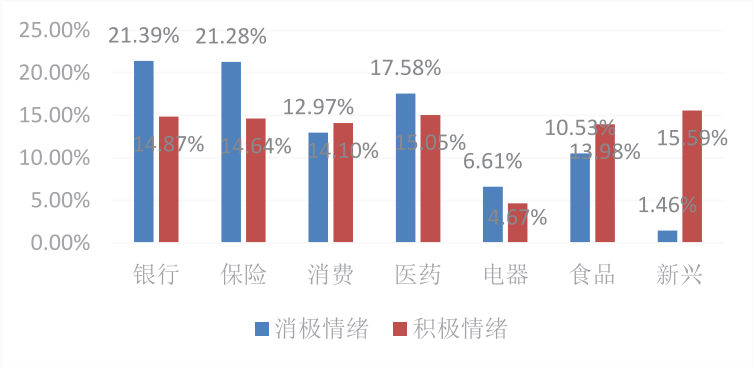

图 1　不同投资情绪投资者投资板快情况

由图 1 所示，消极情绪投资者在银行、保险、医药、电器等板块投资所占比重超过积极情绪投资者，其中在银行（消极情绪投资者 21.39%）、保险（消极情绪投资者 21.28%）等板块所占比重最高；而积极情绪投资者在新兴（积极情绪投资者 15.59%）、食品（积极情绪投资者 13.98%）、消费（积极情绪投资者 14.10%）等板块所占比重超过消极情绪投资者，尤其是在新兴行业板块表现最为明显。

通过对比分析可以看出，积极情绪投资者对更具风险性的新兴行业板块投资的可能性增加，且在持有多种板块股票时，积极情绪投资者对此类股票的交易也会增加，使得交易版块种类越多，积极情绪投资者所占比重越大。

4.2　投资者的情绪指数与投资收益率的关系

根据经典金融学理论认为，积极情绪投资者相对于消极情绪投资者要承担更多的投资风险，因此在做出投资行为时积极情绪投资者需要用更高的投资收益来对承担的投资风险进行补偿。在证券市场中，积极情绪投资者为了追逐更高的投资收益，其在投资方式和投资标的选择上相对于消极情绪投资者要更为激进，而消极情绪投资者则相对更加保守，这种在投资方式和投资标的选择上的差异通常会使得积极情绪投资者与消极情绪投资者获得不同的收益率。因此，为了更深入地研究不同投资者情绪是否在各自投资收益率上存在显著差异，本文将 100 位投资者的情绪指数与各自投资收益率进行独立样本 T 检验。

表 12 显示了 100 位投资者的情绪指数以及投资收益率情况。其中积极情绪投资者的标准差相对于消极情绪投资者标准差较大，这可能是积极情绪投资者为追逐更高的投资收益率，在投资市场中表现得更为活跃，涉及的投资范围更高等现象所致。同时，根据表 12，可见投资者无论是具有消极的投资情绪还是积极的投资情绪，两组间的投资收益率并无显著差异。

表 12　不同投资情绪与收益率的独立样本 T 检验

情绪类型	人数	收益率均值（%）	标准差	t值	p–value（双尾）
消极	60	23.68	4.68	–2.01**	0.03
积极	40	28.93	16.03		

注：** 表示通过显著性水平为 5% 的显著性检验。

5 结论与建议

5.1 结论

第一，本文在研究投资者情绪特征问题时，加入衡量投资者在投资时的心理特征的数据信息，而且这种数据信息特征不能来源于传统的数据获取方式，比如调查问卷等，而是来源于更具客观、真实性的非结构性数据。

第二，本研究以非结构性数据为切入点，以机器学习的方法进行运用，不仅区别了以往关于投资者情绪特征研究的切入点，并且较以往研究更具客观性、合理性以及有效性，在一定程度上具有创新性。文章还介绍了数据清洗与筛选的过程，同时对其展开描述性统计分析，为投资者情绪分析提供依据。

第三，在不同的投资板块中，积极情绪投资者对投资风险性较高的新兴行业板块更加青睐，且与消极情绪投资者呈现出显著的分布差异，但是积极情绪投资者与消极情绪投资者在投资收益率方面并无显著差异。

5.2 建议

第一，根据本文的研究结果，通过构建的投资者情绪指数可以有效地识别不同情绪的投资者。因此，在证券市场中，可以将本文构建的投资者情绪指数作为创新投资标的与投资决策的参考指标，进而进一步提高证券市场的资源配置效率。

第二，研究结果表明，积极情绪投资者与消极情绪投资者在投资收益率方面并无显著差异，可以根据这一研究结果对投资者的投资行为带来一定的指导作用，即在投资行为选择上，积极情绪投资者要控制住为追逐更高的投资收益率而发生的频繁交易行为，因为这并不会带来更高的收益率，而消极情绪投资者则需稳定住对投资行为的情绪。

参考文献：

References：

［1］黄燕芬 , 洪文斌 , 余华义 . 市场情绪如何影响城市房价［J］. 经济理论与经济管理 ,2019,7:75-88.

Huang Y F, Hong W B, Yu H Y. How Market Sentiment Affects Urban Housing Prices ［J］. Economic Theory and Management, 2019,7: 75-88.

［2］王德青 , 田思华 , 朱建平等 . 中国股市投资者情绪指数的函数型构建方法研究［J］. 数理统计与管理 ,2021, 1:162-174.

Wang D Q, Tian S H, Zhu J P et al. A Study on the Functional Construction Method of Investor Sentiment Index in Chinese Stock Market ［J］. Mathematical Statistics and Management, 2021, 1: 162-174.

［3］吴海燕 , 杨朝军 . 金融市场投资者情绪研究述评［J］. 现代管理科学 ,2012,2:12-14+17.

Wu H Y, Yang C J. Review of Research on Investor Sentiment in Financial Markets ［J］. Modern Management Science, 2012,2: 12-14+17.

［4］贺刚 , 朱淑珍 , 顾海峰 . 投资者情绪综合测度指数的构建［J］. 统计与决策 ,2018, 17:149-153.

He G, Zhu S Z, Gu H F. Construction of a Comprehensive Investor Sentiment Measurement Index ［J］. Statistics and Decision, 2018,17: 149-153.

［5］何诚颖 , 陈锐 , 薛冰等 . 投资者情绪、有限套利与股价异象［J］. 经济研究 ,2021, 1:58-73.

He C Y, Chen R, Xue B et al. Investor Sentiment, Limited Arbitrage, and Stock Price Anomalies ［J］. Economic Research, 2021, 1: 58-73.

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

［7］易志高 , 茅宁 . 中国股市投资者情绪测量研究 :CICSI 的构建［J］. 金融研究 ,2009,11:174-184.

Yi Z G, Mao N. Research on Measuring Investor Sentiment in China’s Stock Market: The Construction of CICSI［J］.

Financial Research, 2009,11: 174-184.

［8］张永杰 , 张昱昭 , 金曦等 . 媒体关注与成交量 : 基于百度媒体指数的研究［J］. 系统工程理论与实践 ,2018, 3:576-584.
Zhang Y J, Zhang Y Z, Jin X et al. Media Attention and Trading Volume: A Study Based on Baidu Media Index ［J］. Systems Engineering Theory and Practice, 2018, 3: 576-584.

［9］Piñ, J., Eiro-Chousa, M. Á., et al. The Influence of Investor Sentiment on the Green Bond Market［J］. Technological Forecasting & Social Change,2021,162.

［10］Gao Z, Ren H, Zhang B. Googling Investor Sentiment Around the World［J］. Journal of Financial and Quantitative Analysis,2020,55(2):549-580.

［11］卢米雪 . 投资者情绪的测量及其对股市波动率的影响效应研究［J］. 宏观经济研究 ,2022,9:106-119.
Lu M X. Measurement of Investor Sentiment and Its Impact on Stock Market Volatility ［J］. Macroeconomic Research, 2022,9: 106-119.

［12］高扬 , 赵昆 , 王耀君 . 投资者情绪、股票流动性与股价泡沫——基于 GASDF 检验法的分析［J］. 运筹与管理 ,2023, 7:156-161.
Gao Y, Zhao K, Wang Y J. Investor Sentiment, Stock Liquidity and Stock Price Foam: An Analysis Based on the GASDF Test ［J］. Operations Research and Management, 2023,7: 156-161.

［13］陈浪南 , 苏湃 . 社交媒体对股票市场影响的实证研究［J］. 投资研究 ,2017, 11:17-35.
Chen L N, Su P. Empirical Study on the Impact of Social Media on the Stock Market ［J］. Investment Research, 2017, 11: 17-35.

［14］熊艳 . 论坛发帖与股价行为 : 情绪宣泄还是信息传递 ?［J］. 中央财经大学学报 ,2022,5:29-45.
Xiong Y. Forum Posting and Stock Price Behavior: Emotional Release or Information Transmission? ［J］Journal of Central University of Finance and Economics, 2022,5: 29-45.

［15］Izard, C. E. Human Emotions［M］. New York: Plenum Press, 1977.

［16］Levin, M. A., Levin, P. I., Heath, E. C. Product Category Dependent Consumer Preferences for Online and Offline Shopping Features and Their Influence on Multichannel Retail Alliances ［J］. Journal of Electronic Commerce Research, 2003, 4(3):85-93.

［17］Long, D. E., Sleifer, A., Summers, L. H., Waldmann, R. J. Positive Feedback Investment Strategies and Destabilizing Rational Speculation ［J］. Journal of Finance, 1990, 45(2): 379-395.

［18］Arkes, H., Herren, L., Isen, A. The Role of Potential Loss in the Influence of Affect on Risk Taking Behavior ［J］. Organizational Behavior and Human Decision Processes, 1988, 42(2):181-193.

［19］Kaufmann, G., Vbsburg, S. Paradoxical Mood Effects on Creative Problem-solving［J］. Cognition and Emotion, 1997, 11(2): 151-170.

［20］Hastie, R. Problems for Judgment and Decision Making［J］. Annual Review of Psychology, 2001,52(1): 653-683.

（本文责编：林　琳）

Identification and Statistics of Investor Emotions: Analysis Based on Unstructured Data

GENG Xiao-yuan

Abstract: *The emotional characteristics of investors mainly come from their psychological cognition. Based on the prospect theory, expectation theory, and regret theory of behavioral finance, this paper deeply and meticulously analyzes the connotations that affect investor emotions, and then prepares an investor emotional characteristic index for measurement, in order to study the impact mechanism of investor emotional characteristic index in behavioral finance. At the same time, crawler technology is applied to crawl relevant investor sentiment data, and through reasonable cleaning, screening, and naive Bayesian processing, statistical practice is carried out on investor sentiment characteristics. Research has found that the emotional characteristics of investors are distinguishable and predictable; Based on the analysis of statistical measurement methods for investor sentiment characteristic variables, it is demonstrated that these statistical measurement methods not only can extract and measure investor risk preference features, but also are feasible; Positive sentiment investors prefer emerging industry sectors with higher investment risks, and exhibit significant distribution differences compared to negative sentiment investors. However, there is no significant difference in investment returns between positive sentiment investors and negative sentiment investors.*

Key words: unstructured data; investor sentiment; machine learning; Naive Bayes