基于影响力加权的在线投资者情绪对股票收益的影响*

王高山　王越　董宜麟　张新
（山东财经大学管理科学与工程学院，山东 济南 250014）

摘　要　为更好地捕捉市场总体情绪，本文在利用机器学习算法对在线投资者评论进行情感分类的基础上，考虑评论的阅读量、点赞量和评论量等信息，构建基于影响力加权的在线投资者情绪指数，并建立加权和未加权在线投资者情绪与股票收益的回归模型，在控制股票市值、账面市值比、Beta 等变量基础上，发现在线投资者情绪对股票收益具有显著正向影响，且基于影响力加权的情绪指数相比未加权的情绪指数更能反映股票收益变化。这意味着在线评论影响力蕴含着对投资决策和市场监管有价值的信息。

关键词　在线评论，影响力，机器学习，投资者情绪，股票收益

中图分类号　C931.6，F830.91

1　引言

随着在线社区和社交媒体的发展，越来越多的投资者涌入此类平台，他们在平台上搜索股票信息，并发表观点或看法（通常称为在线投资者评论）[1]，而投资者通过在线评论的文本方式所表达出来的情绪，称为在线投资者情绪。行为金融学理论认为，投资者并非完全理性，容易受情绪的影响而改变投资决策[2, 3]。此外，随着投资者交流互动，情绪也在不断传递、传染、扩散，从而加剧对股市的干扰。可见投资者情绪对股市具有重要的影响，有学者认为其与专业投资信息和公司信息并列成为影响股票运动的三大信息来源[4]。然而，多数文献在构建在线投资者情绪时仅利用了正面评论和负面评论的数量等信息，很少考虑发帖人或帖子内容的影响力。网络上存在“沉默的大多数”效应[5-7]，很多用户并不发表在线评论，仅仅是浏览和查看[8]。发表在线评论的用户难以代表用户总体，从而造成测量上的偏差。

由于在线评论的阅读量、点赞量、评论量和转发量等信息包含了许多未发表评论用户的观点或态度[9, 10]，因此，如果能够充分利用这些信息就能更好地捕捉市场总体情绪。根据心理学和社会学理论，帖子的点赞量、评论量和转发量等信息反映了该帖子的影响力[11, 12]，此影响力指帖子能够引起阅读者的共鸣或认同从而具有改变别人态度和行为的能力[13]，帖子的影响力越大说明其反映了越多人的观点或态度，在计算投资者情绪时应该给予比较高的权重。Li 等[14]利用推特转发量、提及量、粉丝数等计算影响力权重，但主要针对的是用户影响力，且忽视了推文的阅读量；Wang 和 Zhu[15]以及 Shen 等[16]利用帖子的阅读量对投资者情绪进行加权，但没有考虑帖子的点赞量和评论量。本文在上述基础上综合利用帖子

* 基金项目：教育部人文社会科学研究规划基金项目（22YJA630086）。
通信作者：王高山，山东财经大学管理科学与工程学院，教授，E-mail：gaoshanwang@126.com。

的阅读量、点赞量和评论量等信息计算帖子影响力而非用户影响力，并用于修正投资者情绪，能够更好地反映市场总体情绪。

投资者情绪对股票收益具有重要影响，然而由于研究所选取的股票市场、样本、数据周期、数据时间跨度以及投资者情绪指标等不同，对于这种影响的大小及其方向仍存在不同结论。早期文献采用换手率等间接指标测量投资者情绪，存在不足之处：间接指标本身从市场数据中提取得出，却又用于预测和分析市场数据，有可能会存在双向因果关系[17]。本文从在线评论中提取投资者情绪，是对投资者情绪的直接测量，且相比传统问卷方式获得的直接测量指标在成本、时间和频率上更有优势。具体而言，本文利用 Python 编写爬虫程序每天定时获取 24 小时内股票评论及其阅读量、评论量和点赞量等数据。已有研究往往是一次性采集，获取的是发帖日到采集日之间的累积数据[18]，因此无法计算每一天的影响力。本文是连续每天采集，获取的是每日观测值，可以利用这些数据构建每天的基于影响力加权的在线投资者情绪指数，从而更能反映市场总体情绪。同时，本文设定了投资者评论情感极性人工标注的准则（附录 1），构建了投资者评论语料库，从而使得机器学习算法能够对充满俚语、反讽等特征的投资者评论进行准确度较高的分类[19]。

2 文献综述

当前金融学不再满足于以价格反映已有信息的有效市场假说，而是深入研究投资者心理、行为及认知局限等对股票收益的影响，从而对投资者情绪测量以及投资者情绪和股票收益的关系进行多视角研究，相关文献述评如下。

2.1 投资者情绪的测量

对于投资者情绪的测量，基本上可以分为传统方法、基于情感词典的方法和基于机器学习的方法。

传统方法按所采用指标的不同，又可分为直接测量法和间接测量法。直接测量法通常采用调查问卷的方式去获取投资者的态度从而观测投资者情绪，主要有投资者智慧情绪指数、个体投资者协会指数、分析师情绪指数、CBSI（Consensus Bullish Sentiment Index，一致看涨情绪指数）、“央视看盘”指数、好淡指数、投资者信心指数和消费者信心指数等。优点是直接反映了投资者的情绪，缺点则是耗时费力且难以获得短周期、高频率的投资者情绪指数[20]。间接测量法主要利用能反映（部分）情绪的股票交易数据指标来度量投资者情绪[21]，如封闭式基金折价、IPO（initial public offering，首次公开发行）发行量及首日收益、交易量、共同基金净赎回、零股买卖比例等。还有一些学者采用主成分分析方法、偏最小二乘法和 k 步偏最小二乘法对相关间接指标进行降维处理来构造综合指标[22-24]。虽然间接测量法所采用的指标容易获取，但始终不是投资者情绪的直接表现，其既有包含情绪的成分，也有不包含情绪的杂质。

基于情感词典的方法则将人工制定规则和现有情感词相结合，用现有情感词典来分析文本中正负情感词的数目，统计后所获得的正向情感词数若大于负向情感词数，则定文本情感为正向，反之为负向。其优点在于实现方法简单透明、计算词语极性成本相对较低。其中，最具代表性的国外公开情感词典主要有 GI（Harvard General Inquirer Dictionary，哈佛大学通用文本分析查询系统词典）、Harvard Ⅳ[25]和 SentiWordNet[26]等，而国内公开情感词典主要有知网 HowNet[27]、清华大学情感词典等。这些词典通常用词正式、词汇量规模小[28, 29]，不适合存在大量口语、反讽的金融领域[30]。所以，一些学者尝试建立金融领域情感词典，以 Loughran 和 McDonald[31]的 LM 词典最具代表性。国内学者也为构建中文金融领域

情感词典做出了贡献，如姚加权等[32]结合了深度学习和情感词典来构建适用于中文的金融领域情感词典，祝清麟等[33]标注了 5000 多个金融领域中文情感词来构建情感词典。然而这些词典尚未公开，即尚无权威的中文金融领域情感词典可用[34]，也无法对其进行比较。再者，由于人类语言的语义性、复杂性和多样性，仅采用情感词典的方法很难大幅提高准确性[19, 35]，再加上构造情感词典会耗费大量人力，且随着新词语的涌现，需要对情感词典进行不断的更新，因此许多研究人员选择采用基于机器学习的方法。

基于机器学习的方法又可分为无监督、半监督和有监督三类，其中有监督的机器学习方法，在给定足够大的训练集以及合适的特征信息时，能够获得比情感词典方法和无监督机器学习方法更高的准确率[36-38]。该方法基本思路为：首先，使用训练好的分类器对在线评论进行情感分类，如正面、负面和中性，或看空、看多和看平；其次，利用不同种类评论的数量构造统计量来表示投资者情绪，最常用的构造方法为下述三种，分别用式（1）～式（3）来表示[39]。

$$S=\frac{\mathrm{Neg}}{T} \tag{1}$$

$$S=\frac{\mathrm{Pos}-\mathrm{Neg}}{\mathrm{Pos}+\mathrm{Neg}} \tag{2}$$

$$S=\ln\left(\frac{1+\mathrm{Pos}}{1+\mathrm{Neg}}\right) \tag{3}$$

其中，S 为投资者情绪；Neg 为特定时间内负面评论数；Pos 为相同时间内正面评论数；T 为同样时间内全部评论数。式（1）实际上是负面评论比率；式（2）反映了总体正面或负面情绪的标准化比率；式（3）是正面评论相对于负面评论的比例取对数，加 1 是为了防止分母为 0。Wu 等[40]认为式（3）是最为稳健的。

也有学者同时使用正面情绪和负面情绪来开展研究[41, 42]，如式（4）和式（5）所示，其中 PS 和 NS 分别为正面在线投资者情绪和负面在线投资者情绪，Total 为所有评论数量。

$$\mathrm{PS}=\frac{\mathrm{Pos}}{\mathrm{Total}} \tag{4}$$

$$\mathrm{NS}=\frac{\mathrm{Neg}}{\mathrm{Total}} \tag{5}$$

2.2 影响力的测量

早期研究者对影响力的研究多基于用户关系网络拓扑结构、个体属性特征、随机游走的个体影响力排序等方法，其中最常见的方法之一就是基于 PageRank 算法来进行影响力研究，该算法常根据用户间关注-粉丝情况或回复情况来构建网络拓扑结构[43]，但在现实生活中，粉丝数目中包括“僵尸粉”和出于礼貌而互加关注的粉丝，同时只关注粉丝数会忽视用户基于内容的互动情况，所以单纯以“粉丝数”为评价指标的用户影响力评价模型可信度大大降低[44]。转发量反映了用户内容的价值，提及量反映了用户吸引他人参与会话的能力[45, 46]，因此学者不断结合转发量、提及量对 PageRank 算法加以改进，使其更为全面地概括用户影响力，如欧阳纯萍等[47]提出一种 FDRank 算法，Cheng 等[48]则将用户多维社交行为活动与用户兴趣相结合，对用户行为有效性进行量化并作为权重，最后合理地融入 PageRank 算法。

也有学者摒弃了 PageRank 的思想，而直接采用转发量、回复量和点赞量等指标计算用户影响力[49, 50]，并依此思路进一步对单个帖子的影响力进行测度，其背后蕴含的问题为“有多少人认同该帖子的观点和意图，并保持跟此帖一致的思想和行为”。这一思路也存在一定缺陷，认同可能来自两个方面，一是用户一开始并没有这种看法，看到这条帖子后，看法发生了改变，对帖子非常赞同；二是用

户本来就持同样的观点。严格来讲，前者属于影响力，后者属于同质性。Aral 等[51]试图区分影响力和同质性，然而目前仍难以测度网络上的点赞究竟是观点发生改变后的赞同还是本来大家就是同质性的，所以，学者们仍主要采用点赞量、转发量等来计算影响力。

根据不同社交媒体的特征，研究者提出了不同的影响力计算方法。针对微博平台，李华和朱荔[52]采用了一种更为新颖的影响力计算公式，先通过发帖人的粉丝数计算一个因子得分，再通过该帖子的转发数和评论数计算另一个因子得分，对两个因子赋予权重，最后计算综合影响力。针对微信平台，朱丹[11]利用微信的单篇阅读数、单篇在看数和单篇点赞数来计算微信单篇推文的影响力。针对 Twitter 平台，Berger 等[12]收集了 2012～2021 年 ESTRO（European Society of Therapeutic Radiology and Oncology，欧洲放射肿瘤学学会）会议前后的 Twitter 帖子，并将帖子中“点赞”“转发推文”和粉丝数视为领导力/影响力的指标。

在行为金融领域，测量投资者情绪时，主要目的是准确估计具有同类情绪的投资者数量，即避免因忽略“沉默的大多数”而产生的偏误。根据行为金融学的观点，情绪影响投资决策行为，因此人们主要关注情绪和股票收益之间的关系，但情绪是如何产生的，即是受别人影响还是本来自身就是如此，则不影响上述关系的分析。因此，虽然影响力和同质性不同，点赞等行为既可能来自影响力也可能来自同质性，但在分析某一时刻投资者情绪对未来股票收益影响时，重要的是该时刻持有同样态度的投资者数量，而非为什么他们持有这种态度。所以，本文仍采用目前主流的方法，利用点赞量、阅读量、评论量等测算影响力，并用于对情绪进行赋权。

2.3 投资者情绪与股票收益的关系

行为金融学认为个体投资者并非完全理性，他们缺乏专业投资经验，更容易因受到心理偏见、市场情绪和吸引注意力的事件的影响而产生过度自信、后悔等心理情绪[53]。同时，噪声交易模型认为，如果将投资者情绪视为资产定价的系统性风险因素，那股票价格将受情绪影响而显著偏离基本价值[54-56]。

学者对投资者情绪与股票收益间的关系进行了大量的实证研究，但采取的情绪指标不尽相同。例如，Wang 等[57]使用消费者信心指数作为投资者情绪的代理指标，评估了投资者情绪对未来股票回报的影响，发现二者呈负相关关系。王美今和孙建军[58]使用“央视看盘”指数来度量投资者情绪，发现投资者情绪可以影响股票收益。上述文献采用的是直接指标，但直接指标获取难度大且频率低，所以研究人员也经常使用间接指标来研究情绪与股票收益关系。Lee 等[59]使用封闭式基金折价作为投资者情绪的代理指标，发现情绪波动与中小市值股票的收益波动高度相关。林枫娇[60]将平均隔夜收益率作为投资者情绪的度量指标，发现投资者情绪与股票收益显著正相关。李鸿翔[61]使用上证 50ETF（Exchange Traded Funds，交易所交易基金）波动率指数（iVIX）度量投资者情绪，发现 iVIX 与上证 50 指数显著负相关。

随着文本分析技术的不断成熟，越来越多的学者倾向于从股评中提取投资者情绪。鲁万波等[62]使用中文情感词典建立投资者情绪指数，发现该指数对上证指数下行风险具有一定的预测能力。易洪波等[63]同样利用情感词典对投资者情绪指数进行构建，发现投资者多方情绪比空方情绪对收益率的影响更加明显。然而，由于基于词典的方法不能很好地处理语义，基于机器学习方法从文本中提取投资者情绪逐渐成为主流。李岩和金德环[64]使用朴素贝叶斯算法构建投资者情绪，结果证明投资者情绪与股票收益间存在显著正相关关系。许天阳[65]使用支持向量机算法获取投资者情绪指数，发现其对上证指数收益率具有短期正向影响。虽然从在线评论中提取投资者情绪的技术已经比较成熟，但往往使用的是发帖人的数据，而忽视了“沉默的大多数”的意见，因此有学者用帖子的转发量、提及量及用户粉丝数等数据对投资者情绪进行修正，发现修正后的情绪指数对股票收益具有较好的预测效果[14-16]。

总体而言，投资者情绪对股票收益的显著影响已具有广泛共识，但由于研究者采用的度量指标不同、数据时间周期不同以及股票样本的不同等原因，对于影响的大小、方向和持续时间并无一致结论。对于在线投资者情绪，虽然已有学者用影响力指标来进行修正，但使用的影响力模型及数据仍有进一步改进的空间，如可使用帖子的日阅读量、点赞量和评论量等信息来构建综合指标，从而获取更好的预测效果。

3 理论分析与研究假设

不论是有效市场假说还是行为金融学理论，均认为新的信息会对股票市场产生影响，不同之处在于有效市场假说认为新信息的发生是随机的且立即被市场吸收，从而股票价格遵循随机游走模式，而行为金融学理论认为，投资者处理信息的能力是有限的，在这个过程中，投资者情绪对股票价格具有一定的预测作用，背后的机制为：投资者具有有限理性和认知局限，他们在投资决策过程中往往受到不确定信息、波动性环境的影响从而产生认知偏差，进而引起一致的投资决策偏误，对股票价格的形成产生明显的系统性干扰[66]。更进一步，根据认知心理学理论，投资行为可被视为一个系统的信息处理过程，包括对投资感觉的输入、变换、精炼、加工、存储直至形成具体投资行为的全过程，每个阶段都有可能因存在认知偏差而导致股价异常[67]。

按照情绪一致性理论，积极情绪对人的决策行为的影响都应该是积极的，处于积极心态下的投资者通常会做出较为乐观的决定，容易高估企业预期收益，低估相应风险，投资股市意愿提高[68]。如果整体市场情绪表现为积极，场外投资者也会跃跃欲试想要购买股票[69]。如果投资者情绪趋向消极，则投资者在消极情绪状态下对投资决策更为慎重甚至回避，更为担心股票的风险，容易低估股票价值。此外，情绪的传递需要时间，即投资者情绪具有滞后效应[70, 71]，投资者情绪不会立即反映到股票收益中。所以大量研究人员对滞后一期情绪与股票收益关系进行了研究。例如，Seok 等[72]研究了韩国股市中投资者情绪与股票回报之间的关系，发现情绪（第 t–1 天）与股票回报（第 t 天）呈正相关关系。Bouteska 等[73]研究第 t–1 天投资者情绪指数对第 t 天韩国 KOSPI（Korea Composite Stock Price Index，韩国综合股票价格指数）收益率的影响，结果表明两者呈正相关关系。

因此，本文提出如下假设。

H1：在沪深股市，第 t 天股票收益受第 $t-1$ 天投资者情绪的正向影响。

多数研究在构建在线投资者情绪指数时，往往只采用正面评论和负面评论数量，如式（1）～式（5）所示，未考虑在线评论的影响力，这就导致所构建的在线投资者情绪指数仅包含发帖人的情绪，没有包含未发帖人的情绪，而后者的数量非常庞大。因此，在构建在线投资者情绪指数时，考虑在线评论的影响力，以其作为权重构建在线投资者情绪指数，就能够包含许多未发帖投资者的情绪，从而更为准确地反映市场总体情绪。

根据群体动力学理论和传播心理学理论，人的内在需要和周围环境的相互作用决定了人的心理和行为，一条评论被点赞、评论和转发是因为它引起了阅读者的情感共鸣、理解和认同。点赞、评论和转发的次数越多，说明认同的人越多，也表明围绕此主题形成了一个认知一致的群体，而具有相同认知或信念的群体成员数量越多，越容易对股票产生影响[11, 12]。根据心理学和社会学理论，帖子的点赞量和评论量等信息反映了该帖子的影响力[11, 12, 74]，因此，在构建在线投资者情绪指数时，考虑帖子的影响力能够更好地预测股市变化。

Li 等[14]采用朴素贝叶斯算法对 Twitter 文本进行情感分类，并分别使用转发量、提及量、粉丝数和 Kred（https://www.score.kred/）影响力指标进行加权，构建不同类型的加权在线投资者情绪指数，针对标准普尔 100 指数成分股每日和每 15 分钟数据进行实证分析，发现在日线数据上，不同类型的加权在线投资者情

绪均对同期股票异常收益具有显著影响；并且他们发现，当用转发量作为权重时，加权在线投资者情绪比未进行任何加权的投资者情绪更能预测下一日股票异常收益。Altuner 和 Kilimci[75]同样利用 Twitter 转发数据进行投资者情绪指数修正，并用于股价预测。Wang 和 Zhu[15]利用阅读量进行加权，利用修正后的在线投资者情绪指数预测沪深 300 指数，取得较好的效果。从东方财富网可以获取在线投资者评论的阅读量、点赞量和评论量等信息，因此，综合利用这些信息构建加权情绪指数，预计可以取得更好的收益预测效果。

据此，本文提出如下假设。

H2：利用在线投资者评论及其阅读量、点赞量和评论量等信息构建的第 $t-1$ 天基于影响力加权的在线投资者情绪，能比未加权的在线投资者情绪（以下称为简单在线投资者情绪）更好地预测第 t 天的股票收益。

4 研究方法

4.1 基于影响力加权的在线投资者情绪指数构建

深度学习模型在处理长文本和大规模数据集时比传统机器学习模型更具有优势，但在一些小型化的应用中，如多分类、小数据集和短文本的分类任务上，传统机器学习模型仍能够取得不错的性能表现，且训练成本比深度学习要低[76-78]。本文的训练集有 22 000 条标注的评论，且是短文本，进行三分类任务，因此暂时选择传统机器学习模型进行情感分类并构建在线投资者情绪指数，未来在取得更大规模的训练集和更丰富的文本时，可采用深度学习模型。本文的情感分析过程主要包括以下几个阶段：数据采集、数据清洗、人工标注、中文分词、特征选择、算法训练、情感分类、在线评论的影响力计算和在线投资者情绪指数计算，除人工标注外，其余步骤均可自动实现，如图 1 所示。

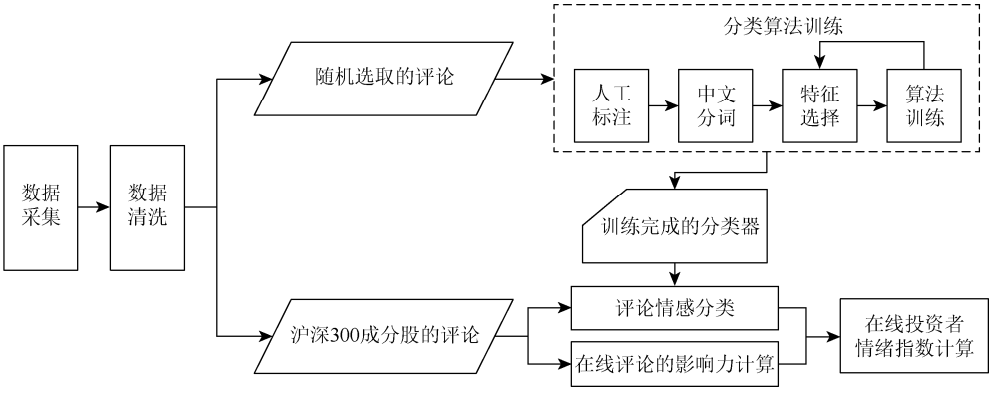

图 1 在线投资者情绪指数构建流程

在数据采集部分，利用 Python 构建网络爬虫，对个人投资者发表的评论内容、阅读量、点赞量、评论量、评论时间和评论人基本信息进行采集。由于本文主要关注个人投资者情绪，然而评论中存在一些公司或机构发布的资讯，因此对其做了进一步的清洗。

接下来采用有监督机器学习方法训练情感分类器，影响分类器性能的因素主要有训练数据的选取、训练数据的标注以及特征集的选择。为保证参与训练样本的代表性，本文对沪深股市股票按照总资产进行排序，按照随机数表，采用系统抽样法选取 25 只股票，然后获取它们的评论数据，共计 22 000 条。为了保证数据标注的公正性、客观性和科学性，本文组建了由 2 位教师、3 位研究生和 3 位本科生组成

的评论标注团队，按照附录 1 的准则对上述评论进行标注。如果一条帖子反映的是积极情绪，标注为 1；消极情绪，标注为–1；中性则标注为 0。标注人分为 3 组，每组包括 1 位研究生和 1 位本科生，前两组各标注 7300 条，最后一组标注 7400 条。每位标注人各自独立标注，然后各组汇总标注结果，如果 2 位标注人标注结果一致，采用标注的结果；如果结果不一致，由标注人和 2 位教师讨论决定结果；最后再由 2 位教师对所有标注的评论进行审核。

文本特征的选择也对机器学习算法的准确性具有重要影响，构建不同的特征集会产生不同的结果。特征词的提取有 DF（document frequency，文档频度）、TF-IDF（term frequency-inverse document frequency，词频–逆文档频率）、IG（information gain，信息增益）、MI（mutual information，互信息）、CHI（chi-square statistic，卡方统计量）等方法，其中 IG 和 CHI 相对较好，然而“没有唯一解”是特征词提取领域一个普遍存在的现象，需要针对不同研究领域、问题和对象，结合人工提取和自动提取两者的优势来进行特征选择[79]。本文采用 CHI 方法进行特征词提取。首先利用 Jieba（https://pypi.org/project/jieba/）分词工具对标注文本进行分词，利用 NLTK（Natural Language Toolkit，自然语言工具包，http://www.nltk.org/）进行词条频率统计，计算每个词条的卡方统计量，按照卡方统计量从大到小选择特征词条，通常数量在 1000～1500 较好，本文通过设定不同的特征词数量进行算法训练，然后利用训练后的算法进行情感分类，再根据分类的准确率来选择特征词数量，经过反复训练，发现特征词数量为 1200 效果最好，其中前 10 位特征词如表 1 所示。

表 1　卡方统计量前 10 位的特征词

序号	特征词	卡方统计量	序号	特征词	卡方统计量
1	垃圾	243.502	6	持股	61.779
2	垃圾股	132.564	7	鼓掌	56.677
3	加仓	94.779	8	坚定	56.499
4	涨停	87.488	9	拉升	56.144
5	买入	85.917	10	满仓	55.224

根据标注文本及其特征，本文继续构造训练集和测试集，分别占文本总量的 80%和 20%，然后选取机器学习算法进行训练。采用 Python 机器学习库 scikit-learn 中的六种机器学习算法（BernoulliNB，MultinomialNB，LogisticRegression，SVC，LinearSVC 和 NuSVC）[①]进行训练，并获得六种算法模型的混淆矩阵和矩阵中元素代表真正例（true positive，TP）、伪正例（false positive，FP）、真反例（true negative，TN）和伪反例（false negative，FN）等统计量，利用这些统计量可以计算分类模型的准确率（Accuracy）、精确率（Precision）、召回率（Recall）和 F_1 值，用于评估模型的性能，如式（6）～式（9）所示。

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{6}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

① BernoulliNB，Bernoulli Naïve Bayes，伯努利朴素贝叶斯；MultinomialNB，Multinomial Naïve Bayes，多项式朴素贝叶斯；LogisticRegression，logistic regression，逻辑回归；SVC，support vector classification，支持向量分类；LinearSVC，linear support vector classification，线性支持向量分类；NuSVC，nu-support vector classification，核支持向量分类，Nu 是一个参数。

$$F_1=\frac{2\times \text{Precision}\times \text{Recall}}{\text{Precision}+\text{Recall}} \tag{9}$$

上述公式是两分类问题的性能评估指标，本文是三分类问题（积极评论、消极评论和中性评论），可以计算每一类的指标，由于本文的重点是积极类和消极类，因此，表 2 展示了这两类的性能指标。此外，也计算了每种模型的总体准确率，如式（10）所示。

表 2 六种机器学习算法的性能指标

算法	总体准确率/%	积极类				消极类			
		准确率/%	精确率/%	召回率/%	F_1/%	准确率/%	精确率/%	召回率/%	F_1/%
BernoulliNB	75.16	86.47	81.26	77.23	79.19	84.54	74.76	80.95	77.73
MultinomialNB	75.64	86.54	80.13	79.30	79.71	84.89	75.19	81.57	78.25
LogisticRegression	72.95	85.16	78.63	76.19	77.39	83.78	76.16	74.74	75.44
SVC	43.82	70.32	85.33	13.25	22.94	72.46	82.31	22.15	34.91
LinearSVC	73.84	85.51	79.10	76.81	77.94	84.47	77.10	75.98	76.54
NuSVC	71.84	84.68	80.85	70.81	75.50	83.57	77.40	71.64	74.41

$$\text{Accuracy}_{\text{overall}}=\frac{\text{TP}_{\text{Pos}}+\text{TP}_{\text{Neu}}+\text{TP}_{\text{Neg}}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}} \tag{10}$$

其中，TP_{Pos} 为积极类的 TP 值；TP_{Neg} 为消极类的 TP 值；TP_{Neu} 为中性类的 TP 值；$\text{Accuracy}_{\text{overall}}$ 为总体准确率。

在机器学习分类任务中，同样的数据和模型，三分类的总体准确率一般低于两分类。本文是三分类问题，从表 2 可以看出 MultinomialNB 算法的总体准确率达到 75.64%，积极类和消极类的准确率分别达到 86.54%和 84.89%。因此，本文选择 MultinomialNB 算法以及相应的特征集对股票评论进行情感分类，其基本公式如式（11）所示：

$$P(X_j=x_{jl}\mid Y=C_k)=\frac{m_{jl}+\lambda}{m_k+n\lambda} \tag{11}$$

其中，$P(X_j=x_{jl}\mid Y=C_k)$ 为第 k 个类别的第 j 维特征的第 l 个取值的条件概率；m_{jl} 为第 k 类第 j 维特征的观测值为第 l 个取值的样本数量；m_k 为训练集中属于第 k 类的样本数量；n 为第 j 维特征值不重复的样本数量；λ 为拉普拉斯平滑，为大于 0 的常数，通常取 1。

得到训练好的分类器后，就可对沪深 300 成分股的评论数据进行情感分类，然后根据式（3）构建在线投资者情绪指数。为更全面地反映市场总体情绪，本文在式（3）的基础上，加入帖子影响力作为权重，如式（12）所示。

$$S_{t,k}=\ln\left(\frac{1+\sum w_{i,t,k}P_{i,t,k}}{1+\sum w_{i,t,k}\left|N_{i,t,k}\right|}\right) \tag{12}$$

其中，$S_{t,k}$ 为 t 时间区间内股票 k 的投资者情绪；$w_{i,t,k}$ 为 t 时间区间内股票 k 第 i 条帖子的影响力；$P_{i,t,k}$ 为 t 时间区间内股票 k 第 i 条正面情感帖子的情感值；$N_{i,t,k}$ 为 t 时间区间内股票 k 第 i 条负面情感帖子的情感值。

对 $w_{i,t,k}$ 的计算参照李华和朱荔[52]的做法。他们使用发帖人粉丝数、帖子的转发量和评论量来计算一条微博的影响力。由于东方财富网并不像微博一样在用户发帖时同时向该用户的粉丝进行广播或通知，因此发帖人粉丝量对帖子的影响力不具有贡献，但东方财富网提供帖子的阅读量信息，因此用阅读量代替发帖人粉丝量更为合适，本文定义股评的阅读量评分 RScore_i 的计算公式如式（13）所示：

$$\mathrm{RScore}_i = \frac{\ln(\mathrm{Reading}_i + 1)}{\ln(\mathrm{Reading}^{\max} + 1)} \tag{13}$$

其中，$\mathrm{Reading}_i$ 为帖子 i 的阅读量；$\mathrm{Reading}^{\max}$ 为阅读量最多的帖子的阅读量。

然而，阅读量自身不能完全说明该条股评的影响力，阅读量高只能说明标题对读者的吸引力或流行度（popularity）[80]，读了之后能不能产生共鸣或认同，则是另一个问题，而只有产生了共鸣或认同，才能产生影响力。Berger 等[12]认为，点赞、评论或转发他人的推文会给该信息和信息发起者带来更多的受众和更高的可信度，同时也是信息接收者表示支持、认可，甚至是尊重的表现，因此很多学者使用“点赞”“转发”“评论”来代表社交媒体帖子影响力[12, 81, 82]。东方财富网提供帖子的阅读量、评论量和点赞量数据，并不提供转发功能，所以本文进一步考虑评论量和点赞量的影响，如式（14）所示：

$$\mathrm{FLScore}_i = \min\left(\frac{\ln(\mathrm{Comment}_i + \mathrm{Like}_i + 1)}{r}, 1\right) \tag{14}$$

其中，$\mathrm{FLScore}_i$ 为帖子 i 因为被评论和点赞而得到的分值；$\mathrm{Comment}_i$ 为帖子 i 的评论量；Like_i 为帖子 i 的点赞量；r 为评论和点赞数之和取对数后的一个阈值，当股评的评论和点赞数之和取对数后的值大于该阈值时认为 $\mathrm{FLScore}_i$ 为 1。

综合利用阅读量得分以及评论和点赞量得分计算股评的影响力得分，计算公式如式（15）所示：

$$w_i = \alpha \times \mathrm{RScore}_i + (1-\alpha) \times \mathrm{FLScore}_i \tag{15}$$

其中，w_i 为帖子 i 的影响力得分；$\alpha \in [0,1]$，为阅读量得分的权重。因为阅读量虽然在一定程度上能够反映影响力，但其反映的更多是帖子的受关注程度而非影响力[83]，而评论量和点赞量才更能体现帖子的影响力，因此可以给予后者较高权重[12, 80, 84]，冯锐和李闻[81]以及姚婷等[82]在确定微博影响力时也认为评论量、转发量和点赞量应该赋予较大权重，所以本文认为赋予 α 较小值比较合理，为进一步进行选择，本文计算了 α 取不同值时的在线投资者情绪指数，使用 OLS（ordinary least squares，普通最小二乘法）对投资者情绪和股票收益等变量进行回归方程检验，发现总体上 R^2 随 α 增加呈下降趋势，当 $\alpha = 0.1$ 时，R^2 最高，如图 2 所示。因此，本文选择 $\alpha = 0.1$。

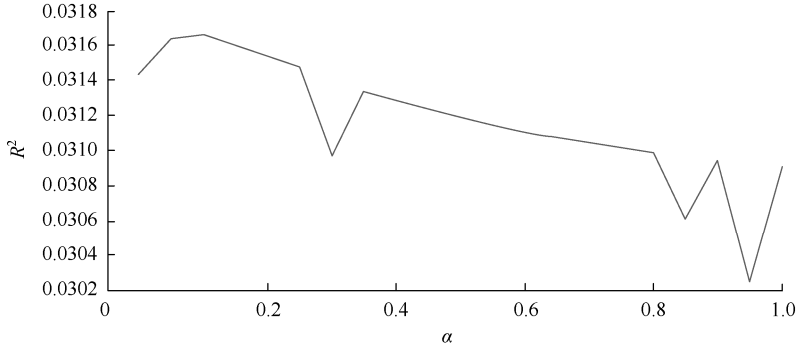

图 2　α 不同取值对回归方程 R^2 的影响

4.2　研究模型

现有文献大多以市场指数级别的投资者情绪和股票收益为对象进行研究[85-87]，即使用时间序列数据开展的研究相对较多，而利用面板数据开展的研究相对较少。与横截面数据模型和时间序列数据模型相比，面板数据模型具有突出的优点，能够对更复杂的行为模型进行研究。本文选用沪深 300 成分股每天的投资者情绪和股票收益，建立面板数据模型进行实证研究。

Schmeling[88]认为第 $t-1$ 期的投资者情绪对第 t 期的收益具有显著影响，其基本模型如式（16）所示：

$$R_{i,t}=\alpha+\beta\times S_{i,t-1}+\varepsilon_t \tag{16}$$

其中，$R_{i,t}=\ln C_{i,t}-\ln C_{i,t-1}$；$R_{i,t}$ 为股票 i 第 t 期收益率；$C_{i,t}$ 为股票 i 第 t 期收盘价；$C_{i,t-1}$ 为股票 i 第 $t-1$ 期收盘价；$S_{i,t-1}$ 为第 t–1 期的投资者情绪；ε_t 为误差项。

根据之前的研究，股市中可能存在周末效应[89]，因此，本文在式（16）中增加 Monday_t（星期一）和 Friday_t（星期五）两个虚拟变量作为控制变量。此外，股票收益还可能受股票市值、账面市值比、Beta 和过去收益等因素的影响[90]，因此，本文也把它们作为控制变量。总体模型如式（17）所示：

$$R_{i,t}=\beta_0+\beta_1\times S_{i,t-1}+\beta_2\times \text{MV}_{i,t-1}+\beta_3\times \text{Bmr}_{i,t-1}+\beta_4\times \text{Beta}_{i,t-1}+\beta_5\times R_{i,t-1}+\beta_6\text{Monday}_t+\beta_7\text{Friday}_t+\varepsilon_t \tag{17}$$

其中，$\text{MV}_{i,t-1}$ 为股票 i 在 t–1 期的股票市值；$\text{Bmr}_{i,t-1}$ 为股票 i 在 t–1 期的账面市值比；$\text{Beta}_{i,t-1}$ 为股票 i 在 t–1 期的 Beta 值；Monday_t 和 Friday_t 为虚拟变量。

4.3 数据来源

股票评论数据来自东方财富网股吧论坛（http://guba.eastmoney.com/），东亚前海证券有限责任公司 2022 年 4 月 14 日发布的东方财富网研究报告中引述艾瑞咨询和 Alexa 公司[①]数据显示，东方财富网覆盖人数 5990 万人，仅次于新浪，日均页面浏览量达 5911 万次，流量优势显著。本文选取沪深 300 成分股作为研究对象，东方财富网股吧论坛上存有这些股票的历史评论数据，但其阅读量、评论量和点赞量是累积数据，因此，本文利用 Python 语言编写爬虫程序每天定时获取 24 小时内股票评论及其阅读量、评论量和点赞量等数据（这也是本文在数据获取方面的特色之处，之前的研究往往是一次性采集，获取的是发帖日到采集日之间的累积阅读量、评论量和点赞量，无法计算每一天的影响力[18]；本文是连续每日采集，获取的是每日观测值），共获取评论 1 045 311 条。根据式（12）计算每只股票每天的加权在线投资者情绪（S_{weight}）；股票交易数据从国泰安 CSMAR（China Stock Market and Accounting Research，中国经济金融研究）数据库获得；同时本文也根据式（3）计算了未加权在线投资者情绪（简单在线投资者情绪，S），并在式（4）和式（5）基础上考虑帖子影响力分别计算正面在线投资者情绪和负面在线投资者情绪，最后形成面板数据，总样本量为 17 946 个，描述性统计量如表 3 所示。

表 3 沪深 300 成分股在线投资者情绪和股票收益率（日线数据）描述性统计

Variable（变量）	N（样本容量）	Mean（均值）	Std.（标准差）	Min.（最小值）	Max.（最大值）
R	17 946	–0.0013	0.0255	–0.1061	0.0965
MV	17 946	0.4693	0.2383	0	1
Beta	17 946	1.0380	0.3299	–0.3217	2.2837
Bmr	17 946	0.7079	0.4590	0.0375	2.9470
S_{weight}	17 946	–0.0932	0.3024	–1.7401	1.1535
S	17 946	–0.2034	0.6454	–2.8526	2.4849
PS	17 946	0.2170	0.1243	0	1
NS	17 946	–0.2709	0.1332	–1	0

注：时间范围为 2018 年 7 月 27 日至 2018 年 11 月 14 日

在线投资者情绪和股票收益均为平稳时间序列，通过了 ADF（augmented Dickey-Fuller，增广迪基–富勒）单位根检验，图 3 为沪深 300 成分股中的贵州茅台（600519.SH）的在线投资者情绪和股票收益时间序列（归一化后结果）。

① Alexa 是一家专门发布网站世界排名的网站，是 Amazon 的子公司。

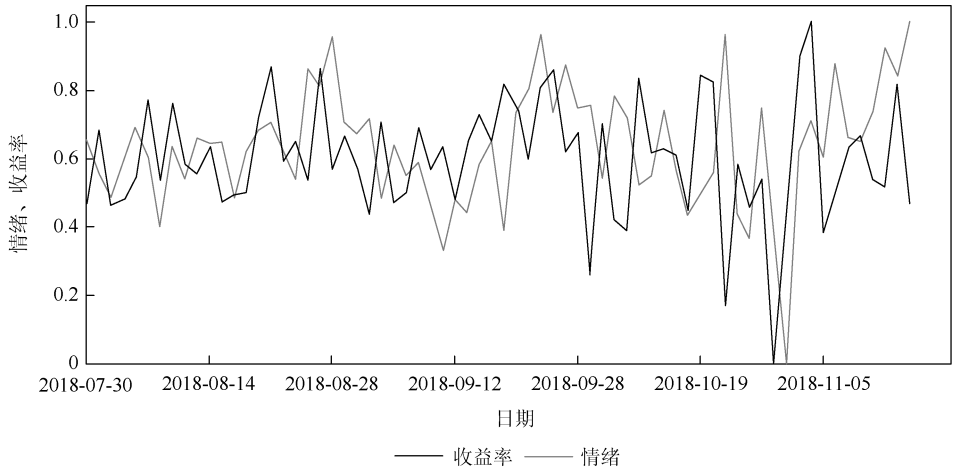

图 3　贵州茅台的在线投资者情绪和股票收益（归一化后结果）

5　实证分析

5.1　在线投资者情绪和股票收益的回归结果

经 Hausman（豪斯曼）检验，本文选择个体固定效应模型而非随机效应模型；经 White（怀特）检验，回归方程满足异方差性，因此本文对比了稳健标准误修正的 OLS 和 FE（fixed effects，固定效应）模型来对式（17）进行回归估计。在动态面板数据模型中，因变量的滞后项可能与模型随机扰动项存在相关性，从而使得 OLS 估计是有偏的，即使 FE 也是不一致的[91, 92]，所以本文同时对比了差分 GMM（generalized method of moments，广义矩方法）估计结果。另外，本文也对比了基于影响力加权的在线投资者情绪（S_{weight}）和简单在线投资者情绪与股票收益的关系，结果如表 4 所示。可以看出，不论是采用 OLS、FE，还是差分 GMM，S_{weight} 均对股票收益产生显著影响，H1 成立，而 S 的影响仅在 OLS 中显著，在 FE 和差分 GMM 中均不显著，表明基于影响力加权的在线投资者情绪能够更好地反映股票收益。

表 4　在线投资者情绪的回归结果

变量	（1）OLS	（2）FE	（3）差分 GMM	（4）OLS	（5）FE	（6）差分 GMM
l.S_{weight}	0.0031*** （4.46）	0.0024*** （3.44）	0.0089*** （3.98）			
l.S				0.0008** （2.76）	0.0006 （1.94）	–0.0001 （–0.01）
l.R	0.0028 （0.27）	–0.0032 （–0.31）	–0.0188* （–2.13）	0.008 （0.78）	0.0008 （0.07）	0.0001 （0.01）
l.Bmr	0.0033*** （7.43）	0.0035 （0.38）	0.1180* （2.06）	0.0033*** （7.39）	0.0035 （0.38）	0.0635 （0.82）
l.Beta	0.0003 （0.39）	0.0048** （3.06）	0.0227** （2.94）	0.0002 （0.34）	0.0048** （3.12）	0.0241** （2.89）
l.MV	–0.0153*** （–17.89）	–0.0197*** （–8.68）	–0.169*** （–12.18）	–0.0153*** （–17.80）	–0.0196*** （–8.69）	–0.1710*** （–8.78）

续表

变量	（1）OLS	（2）FE	（3）差分 GMM	（4）OLS	（5）FE	（6）差分 GMM
Monday	–0.0009 （–1.47）	–0.0010* （–2.16）	–0.0013** （–3.05）	–0.0009 （–1.50）	–0.0010* （–2.19）	–0.0015*** （–3.36）
Friday	0.0039*** （8.70）	0.0039*** （9.87）	0.0026*** （7.05）	0.0040*** （8.75）	0.0039*** （9.91）	0.0030*** （8.16）
Constant	0.00219* （2.19）	–0.0006 （–0.08）	–0.0308 （–0.63）	0.0021* （2.09）	–0.0008 （–0.11）	0.0071 （0.11）
R^2	0.0308	0.0330		0.0300	0.0325	

***表示显著性水平为 0.001，**表示显著性水平为 0.01，*表示显著性水平为 0.05

5.2 基于影响力加权的在线投资者情绪和简单在线投资者情绪的进一步比较

1. 配对样本 *t* 检验

为进一步分析基于影响力加权的在线投资者情绪和简单在线投资者情绪的不同，本文首先对它们进行配对样本 *t* 检验，结果如表 5 所示，从中可以看出，两者的差异是显著的。

表 5 配对样本 *t* 检验结果

差值	均值	标准误	*t* 值	*p* 值
S_{weight}-S	0.1102	0.0034	32.6347	0.000

2. 相关系数比较

本文采用皮尔逊相关系数检验和斯皮尔曼相关系数检验，分别对基于影响力加权的在线投资者情绪和简单在线投资者情绪与股票收益的相关系数进行比较。如表 6 所示。可以看出，从皮尔逊相关系数检验结果来看，基于影响力加权的在线投资者情绪与股票收益的相关系数为 0.0362，简单在线投资者情绪与股票收益的相关系数为 0.0194，前者明显大于后者。从斯皮尔曼相关系数检验结果来看，基于影响力加权的在线投资者情绪与股票收益的相关系数为 0.0290，而简单在线投资者情绪与股票收益的相关系数为 0.0200，同样前者明显大于后者。

表 6 在线投资者情绪与股票收益的相关系数

检验方法	基于影响力加权的在线投资者情绪	简单在线投资者情绪
皮尔逊相关系数	0.0362***	0.0194***
斯皮尔曼相关系数	0.0290***	0.0200***

***表示显著性水平为 0.001

进一步使用 suest 进行系数的差异性检验，结果如表 7 所示，说明两个系数具有显著差异。综上分析结果表明，基于影响力加权的在线投资者情绪指数比简单在线投资者情绪指数对股票收益更具有预测力，可以得出 H2 成立。

表 7 系数差异性检验结果

假设检验	卡方值	*p* 值
系数差异	20.29	0.000

5.3 积极情绪和消极情绪的作用分析

为进一步探讨在线投资者情绪中的积极成分和消极成分各自的作用，本文继续采用稳健标准误修正的 OLS 和 FE 模型来分别对正面在线投资者情绪和负面在线投资者情绪与股票收益（R）的关系进行分析，结果如表 8 所示，总体来看，积极成分对股票收益的影响相对显著，用稳健标准误修正的 OLS 和 FE 估计时，正面在线投资者情绪对股票收益具有显著正向影响，而负面在线投资者情绪的影响不显著。表明在线投资者情绪对股票收益的影响主要是通过其中的积极成分产生，且随着积极情绪增强，股票收益会增高。

表 8 正面在线投资者情绪、负面在线投资者情绪和股票收益的回归结果

变量	（1）OLS	（2）FE	（3）OLS	（4）FE
PS	0.0040* （2.25）	0.0047* （2.35）		
NS			0.0024 （1.47）	0.0002 （0.09）
l.R	0.0100 （0.98）	0.0011 （0.11）	0.0103 （1.02）	0.0038 （0.37）
l.Bmr	0.0033*** （7.45）	0.0034 （0.39）	0.0033*** （7.29）	0.0036 （0.39）
l.Beta	0.0002 （0.24）	0.0049** （3.14）	0.0002 （0.34）	0.0049** （3.15）
l.MV	–0.0152*** （–17.76）	–0.0196*** （–8.67）	–0.0152*** （–17.75）	–0.0196*** （–8.66）
Monday	–0.0009 （–1.49）	–0.0010* （–2.15）	–0.0009 （–1.51）	–0.0010* （–2.21）
Friday	0.0040*** （8.74）	0.0039*** （9.86）	0.0040*** （8.81）	0.0039*** （9.97）
Constant	0.0011 （1.01）	–0.0021 （–0.29）	0.0026* （2.41）	–0.0011 （–0.14）
R^2	0.0299	0.0327	0.0297	0.0323

***表示显著性水平为 0.001，**表示显著性水平为 0.01，*表示显著性水平为 0.05

5.4 稳健性检验

国泰安 CSMAR 数据库提供两种看涨情绪指数 BullishSentIndexA 和 BullishSentIndexB，分别按式（2）和式（3）计算，其股评来源为东方财富网股吧和新浪股吧（https://guba.sina.com.cn/）。本文利用 BullishSentIndexA 和 BullishSentIndexB 两种指数数据对式（17）进行重新运算，发现这两个情绪指数均跟股票收益显著正相关，说明 H1 成立。本文也进一步对基于影响力加权的在线投资者情绪、简单投资者情绪与 BullishSentIndexA 和 BullishSentIndexB 两种情绪指数进行了相关性分析，发现它们之间具有显著正相关性，且基于影响力加权的在线投资者情绪与 BullishSentIndexA 和 BullishSentIndexB 的相关系数比简单在线投资者情绪与后者的相关系数要高，说明 H2 成立。

5.5 结果讨论

本文的实证结果表明在线投资者情绪对股票收益具有显著的正向影响，这说明社交媒体或股票论坛中的评论内容蕴含着预测股票变化的信息，尤其是社交网络中的点赞和评论等用户行为中蕴含着大量有价值信息，点赞和评论行为背后是“沉默的大多数”投资者对帖子的情感共鸣、认同或支持，能够更好地体现市场总体意见，从而能够更好地预测短期股票收益。

其背后的逻辑是：社交媒体日益成为互联网的主要应用，越来越多的投资者通过论坛、股吧等社交媒体发表意见、观点和获取信息，不同的观点和信息在网络上交织碰撞，根据复杂适应系统理论，投资者的相互作用生成宏观的股市现象，投资者在社交媒体上的互动是一种信息交流和相互学习的过程，他们从中可以坚定、修正或抛弃自己已有的信念和期望，从而对短期股票收益产生影响。

本文的研究结果表明，在线投资者论坛或股吧可以看作整个股票系统中的承载投资者态度和观点的信息子系统，根据态度和行为的关系理论，这一信息子系统应该具有预测股票运动的能力，而本文的方法能够对其中的信息进行提取。这也意味着，随着在线投资者论坛用户数量日益增多，市场管理者应利用大数据分析手段加强对论坛的监管和引导，跟踪其中的舆情动态，防范潜在的股市风险。

最后，本文计算得到的是个股的在线投资者情绪，采用的是面板数据模型，相比基于指数级在线投资者情绪的时间序列模型能够进一步揭示出个股层面上在线投资者情绪和股票收益的关系。因此，利用本文所构建的在线投资者情绪指数，可以进一步在传统金融计量模型基础上增加在线投资者情绪变量，建立更有效的个股预测模型，为投资者提供决策参考，同时，市场监管者可以了解个股的舆情动向，从而进一步深化市场监管。

6 结束语

行为金融学理论和认知心理学理论认为投资者情绪会对股票收益产生影响，大量研究者构建了投资者情绪与股票收益关系模型来分析两者关系或预测股票收益。然而已有的模型未充分考虑网络上“沉默的大多数”效应，本文利用在线评论当日阅读量、点赞量、评论量来构造基于影响力加权的在线投资者情绪指数，能够更好地反映市场总体情绪，进一步完善了现有的基于投资者情绪的股票收益预测模型，具有一定的理论贡献。

在实践方面，本文的研究结果对投资者和监管者均具有决策参考作用。对于投资者而言，在线评论及其阅读量、点赞量、评论量等信息可以帮助他们做出决策，而管理者也可据此进行舆情管理，防范金融风险，构建良好股市环境。另外，本文建立了包括数据采集、数据清洗、人工标注、特征选择、机器学习算法训练、情感分类、影响力计算和情绪指数计算等功能的分析处理平台，提取了股评文本特征，构建了特征集合，训练了具有较高准确度的情感分类器，能够克服在线投资者评论文法自由不利于提取情绪的困难，可为研究者进行金融大数据分析提供参考。

在本文研究过程中，自然语言处理技术发展迅速，尤其是以 BERT（Bidirectional Encoder Representations from Transformers，基于变换器的双向编码器表示技术）和 GPT（Generative Pre-trained Transformer，生成式预训练变换模型）为代表的大语言模型表现出强大的能力，GPT 擅长生成式任务，相比 GPT，BERT 更适合文本分类等任务。本文在 BERT 预训练模型基础上对所建在线投资者评论语料

库进行再训练，三分类总体准确率达到 77.17%，相比 MultinomialNB 算法提高约 1.5 个百分点，说明 BERT 更有优势，但目前优势还不明显，主要是因为 BERT 预训练模型所用语料并非金融领域，未来可进一步利用金融、股评领域大量文本重新进行预训练，建立金融领域 BERT 预训练模型，再开展下游文本分类等任务，有望取得更好的效果。

参 考 文 献

[1] Gan B Q，Alexeev V，Bird R，et al. Sensitivity to sentiment：news vs social media[J]. International Review of Financial Analysis，2020，67：101390.

[2] 宫汝凯. 信息不对称、过度自信与股价变动[J]. 金融研究，2021，（6）：152-169.

[3] Tan X Y，Zhang Z L，Zhao X J，et al. Investor sentiment and limits of arbitrage：evidence from Chinese stock market[J]. International Review of Economics & Finance，2021，75：577-595.

[4] Li Q，Wang J，Wang F，et al. The role of social sentiment in stock markets：a view from joint effects of multiple information sources[J]. Multimedia Tools and Applications，2017，76（10）：12315-12345.

[5] Gao G D，Greenwood B N，Agarwal R，et al. Vocal minority and silent majority：how do online ratings reflect population perceptions of quality[J]. MIS Quarterly，2015，39（3）：565-590.

[6] Long W，Zhong Y Q. The neglected cohort：the impact of silent majority in social media on stock returns[J]. Finance Research Letters，2023，52：103363.

[7] Chang K F，Huang Y H，Li W C，et al. Promotion of Internet users' aggressive participation via the mediators of flow experience and identification[J]. Frontiers in Psychology，2022，13：836303.

[8] Hong Y，Hu J T，Zhao Y X. Would You go invisible on social media? An empirical study on the antecedents of users' lurking behavior[J]. Technological Forecasting and Social Change，2023，187：122237.

[9] Lu J H，Zhang M S，Zheng Y，et al. Communication of uncertainty about preliminary evidence and the spread of its inferred misinformation during the COVID-19 pandemic-a weibo case study[J]. International Journal of Environmental Research and Public Health，2021，18（22）：11933.

[10] Sui M X，Hawkins I，Wang R. When falsehood wins? Varied effects of sensational elements on users' engagement with real and fake posts[J]. Computers in Human Behavior，2023，142：107654.

[11] 朱丹. 基于 WCI 的高校图书馆微信推文影响力及运营策略探究：以武汉大学图书馆为例[J]. 图书馆工作与研究，2022，（9）：104-112.

[12] Berger T，Payan N，Fleury E，et al. Gender-related and geographic trends in interactions between radiotherapy professionals on Twitter[J]. Physics and Imaging in Radiation Oncology，2022，24：129-135.

[13] 钱萌，王子鸣，程树林. 社交网络中的用户影响力研究综述[J]. 电脑知识与技术，2023，19（2）：72-74.

[14] Li T，van Dalen J，van Rees P J. More than just noise? Examining the information content of stock microblogs on financial markets[J]. Journal of Information Technology，2018，33（1）：50-69.

[15] Wang J J，Zhu S Z. A novel stock index direction prediction based on dual classifier coupling and investor sentiment analysis[J]. Cognitive Computation，2023，15（3）：1023-1041.

[16] Shen Y R，Liu C，Sun X L，et al. Investor sentiment and the Chinese new energy stock market：a risk-return perspective[J]. International Review of Economics & Finance，2023，84：395-408.

[17] Anand A，Basu S，Pathak J，et al. The impact of sentiment on emerging stock markets[J]. International Review of Economics & Finance，2021，75：161-177.

[18] 李志恒，李红，崔昭文，等. 基于 PCA 的地震官方微博影响力评价研究[J]. 中国地震，2021，37（3）：649-658.

[19] 刘兵. 情感分析：挖掘观点、情感和情绪[M]. 刘康，赵军，译. 北京：机械工业出版社，2017.

[20] Zhang W G，Gong X E，Wang C，et al. Predicting stock market volatility based on textual sentiment：a nonlinear analysis[J]. Journal of Forecasting，2021，40（8）：1479-1500.

[21] Kim K，Ryu D，Yu J. Is a sentiment-based trading strategy profitable?[J]. Investment Analysts Journal，2022，51（2）：94-107.
[22] Song Z Y，Yu C R. Investor sentiment indices based on k-step PLS algorithm：a group of powerful predictors of stock market returns[J]. International Review of Financial Analysis，2022，83：102321.
[23] Bouteska A，Mefteh-Wali S，Dang T. Predictive power of investor sentiment for Bitcoin returns：evidence from COVID-19 pandemic[J]. Technological Forecasting and Social Change，2022，184：121999.
[24] Song Z Y，Gong X M，Zhang C，et al. Investor sentiment based on scaled PCA method：a powerful predictor of realized volatility in the Chinese stock market[J]. International Review of Economics & Finance，2023，83：528-545.
[25] Sul H K，Dennis A R，Yuan L I. Trading on twitter：using social media sentiment to predict stock returns[J]. Decision Sciences，2017，48（3）：454-488.
[26] Punetha N，Jain G. Bayesian game model based unsupervised sentiment analysis of product reviews[J]. Expert Systems With Applications，2023，214：119128.
[27] 蒋翠清，郭铁博，刘尧. 基于中文社交媒体文本的领域情感词典构建方法研究[J]. 数据分析与知识发现，2019，3（2）：98-107.
[28] 赵妍妍，秦兵，石秋慧，等. 大规模情感词典的构建及其在情感分类中的应用[J]. 中文信息学报，2017，31（2）：187-193.
[29] Wang Q L，Xu W，Zheng H. Combining the wisdom of crowds and technical analysis for financial market prediction using deep random subspace ensembles[J]. Neurocomputing，2018，299：51-61.
[30] 顾文涛，王儒，郑肃豪，等. 金融市场收益率方向预测模型研究：基于文本大数据方法[J]. 统计研究，2020，37（11）：68-79.
[31] 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.
[32] 姚加权，冯绪，王赞钧，等. 语调、情绪及市场影响：基于金融情绪词典[J]. 管理科学学报，2021，24（5）：26-46.
[33] 祝清麟，梁斌，徐睿峰，等. 结合金融领域情感词典和注意力机制的细粒度情感分析[J]. 中文信息学报，2022，36（8）：109-117.
[34] Sun Y C，Zeng X P，Zhou S Y，et al. What investors say is what the market says：measuring China’s real investor sentiment[J]. Personal and Ubiquitous Computing，2021，25（3）：587-599.
[35] 李合龙，任昌松，柳欣茹，等. 金融市场文本情绪研究综述[J]. 数据分析与知识发现，2023：1-25.
[36] Mardjo A，Choksuchat C. HyVADRF：hybrid VADER-random forest and GWO for bitcoin tweet sentiment analysis[J]. IEEE Access，2022，10：101889-101897.
[37] Kanavos A，Nodarakis N，Sioutas S，et al. Large scale implementations for twitter sentiment classification[J]. Algorithms，2017，10（1）：33.
[38] van Atteveldt W，van der Velden M A C G，Boukes M. The validity of sentiment analysis：comparing manual annotation，crowd-coding，dictionary approaches，and machine learning algorithms[J]. Communication Methods and Measures，2021，15（2）：121-140.
[39] Herrera G P，Constantino M，Su J J，et al. Renewable energy stocks forecast using Twitter investor sentiment and deep learning[J]. Energy Economics，2022，114：106285.
[40] Wu D D，Zheng L J，Olson D L. A decision support approach for online stock forum sentiment analysis[J]. IEEE Transactions on Systems，Man，and Cybernetics：Systems，2014，44（8）：1077-1087.
[41] Oliveira N，Cortez P，Areal N. The impact of microblogging data for stock market prediction：using Twitter to predict returns，volatility，trading volume and survey sentiment indices[J]. Expert Systems With Applications，2017，73：125-144.
[42] Obaid K，Pukthuanthong K. A picture is worth a thousand words：measuring investor sentiment by combining machine learning and photos from news[J]. Journal of Financial Economics，2022，144（1）：273-297.
[43] 李威，廖健，曾剑平. 网络社交媒体用户影响力的动态量化方法[J]. 计算机应用研究，2020，37（S2）：50-53.

[44] 文馨，陈能成，肖长江. 基于 Spark GraphX 和社交网络大数据的用户影响力分析[J]. 计算机应用研究，2018，35（3）：830-834.
[45] Yang B，Liu C，Cheng X S，et al. Understanding users' group behavioral decisions about sharing articles in social media：an elaboration likelihood model perspective[J]. Group Decision and Negotiation，2022，31（4）：819-842.
[46] Hong L，Yin J E，Xia L L，et al. Improved short-video user impact assessment method based on PageRank algorithm[J]. Intelligent Automation & Soft Computing，2021，29（2）：437-449.
[47] 欧阳纯萍，陈湘龙，刘永彬. 基于网络新闻评论的四度用户影响力分析模型[J]. 计算机工程与设计，2021，42（9）：2671-2678.
[48] Cheng S L，Wang Z M，Qian M，et al. Calculating influence based on the fusion of interest similarity and information dissemination ability[J]. Turkish Journal of Electrical Engineering and Computer Sciences，2022，30（3）：592-608.
[49] 王利，于磊，吴渝. 基于 Swarm 模型的微博用户影响力评价方法[J]. 计算机工程与应用，2021，57（2）：267-272.
[50] 王林，潘陈益，朱文静. 基于 h 指数、g 指数和 p 指数的微博影响力评价对比研究[J]. 现代情报，2018，38（6）：11-18，61.
[51] Aral S，Muchnik L，Sundararajan A. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks[J]. Proceedings of the National Academy of Sciences of the United States of America，2009，106（51）：21544-21549.
[52] 李华，朱荔. 基于影响力的微博新兴热点事件检测[J]. 计算机应用与软件，2016，33（5）：98-101，165.
[53] Kaniel R，Saar G，Titman S. Individual investor trading and stock returns[J]. The Journal of Finance，2008，63（1）：273-310.
[54] Zhang W G，Gong X E，Wang C，et al. Predicting stock market volatility based on textual sentiment：a nonlinear analysis[J]. Journal of Forecasting，2021，40（8）：1479-1500.
[55] Gao B，Hao H H，Xie J. Does retail investors beat institutional investors? Explanation of game stop's stock price anomalies[J]. PLoS One，2022，17（10）：e0268387.
[56] 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.
[57] Wang W Z，Su C，Duxbury D. Investor sentiment and stock returns：global evidence[J]. Journal of Empirical Finance，2021，63：365-391.
[58] 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.
[59] 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.
[60] 林枫娇. 投资者情绪与股票收益关系研究：基于隔夜收益率[J]. 投资研究，2022，41（11）：137-159.
[61] 李鸿翔. 投资者情绪与股票收益率的关系研究：基于上证 50ETF 波动率指数的实证分析[J]. 中国物价，2020，（8）：74-77.
[62] 鲁万波，张萌，郑天照. 股票评论信息能够预测股票市场的下行风险吗?[J]. 统计与信息论坛，2022，37（8）：53-66.
[63] 易洪波，赖娟娟，董大勇. 网络论坛不同投资者情绪对交易市场的影响：基于 VAR 模型的实证分析[J]. 财经论丛，2015，（1）：46-54.
[64] 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（15）：166-169.
[65] 许天阳. 网络社交媒体中投资者情绪对股票市场的影响研究[J]. 上海管理科学，2018，40（3）：67-74.
[66] 曹国华，任成林，林川. 投资者情绪、管理者过度乐观与“IPO 之谜”[J]. 重庆大学学报（社会科学版），2019，25（1）：29-48.
[67] Chang Y C，Shao R，Wang N. Can stock message board sentiment predict future returns? Local versus nonlocal posts[J]. Journal of Behavioral and Experimental Finance，2022，34：100625.
[68] 石善冲，朱颖楠，赵志刚，等. 基于微信文本挖掘的投资者情绪与股票市场表现[J]. 系统工程理论与实践，2018，38（6）：1404-1412.
[69] Mahmoudi N，Docherty P，Melia A. Firm-level investor sentiment and corporate announcement returns[J]. Journal of

Banking & Finance，2022，144：106586.
[70] Baker M，Wurgler J. Investor sentiment and the cross-section of stock returns[J]. The Journal of Finance，2006，61（4）：1645-1680.
[71] Aboody D，Even-Tov O，Lehavy R，et al. Overnight returns and firm-specific investor sentiment[J]. Journal of Financial and Quantitative Analysis，2018，53（2）：485-505.
[72] Seok S I，Cho H，Ryu D. Firm-specific investor sentiment and daily stock returns[J]. The North American Journal of Economics and Finance，2019，50：100857.
[73] Bouteska A，Sharif T，Abedin M Z. Does investor sentiment create value for asset pricing? An empirical investigation of the KOSPI-listed firms[J]. International Journal of Finance & Economics，2023，1-23.
[74] Bovet A，Makse H A. Influence of fake news in Twitter during the 2016 US presidential election[J]. Nature Communications，2019，10（1）：7.
[75] Altuner A B，Kilimci Z H. A novel deep reinforcement learning based stock price prediction using knowledge graph and community aware sentiments[J]. Turkish Journal of Electrical Engineering and Computer Sciences，2022，30（4）：1506-1524.
[76] Qaiser A，Hina S，Kazi A K，et al. Fake news encoder classifier（FNEC）for online published news related to COVID-19 vaccines[J]. Intelligent Automation & Soft Computing，2023，37（1）：73-90.
[77] Roy S S，Roy A，Samui P，et al. Hateful sentiment detection in real-time tweets：an LSTM-based comparative approach[J]. IEEE Transactions on Computational Social Systems，2023，（99）：1-10.
[78] Wang Z F，Yao L S，Shao X Y，et al. A combination of TEXTCNN model and Bayesian classifier for microblog sentiment analysis[J]. Journal of Combinatorial Optimization，2023，45（4）：109.
[79] 许鑫. 基于文本特征计算的信息分析方法[M]. 上海：上海科学技术文献出版社，2015.
[80] Lee E，Kim Y G，Seo Y D，et al. An evaluation method for content analysis based on twitter content influence[J]. International Journal of Software Engineering and Knowledge Engineering，2017，27（5）：841-867.
[81] 冯锐，李闻. 社交媒体影响力评价指标体系的构建[J]. 现代传播（中国传媒大学学报），2017，39（3）：63-69.
[82] 姚婷，赵锦栋，杨莉. 突发环境事件中微博影响力的预测研究[J]. 智能计算机与应用，2022，12（10）：36-42.
[83] Yu D G，Chen N，Ran X. Computational modeling of Weibo user influence based on information interactive network[J]. Online Information Review，2016，40（7）：867-881.
[84] 叶佳鑫，熊回香，易明，等. 融合影响力传播的社交网络群推荐方法[J]. 情报学报，2022，41（4）：364-374.
[85] 张继海，刘雅玫. 我国股票市场投资者情绪与市场收益研究：基于个人与机构的比较分析[J]. 山东社会科学，2020，（2）：81-86.
[86] 尹海员，吴兴颖. 投资者日度情绪、超额收益率与市场流动性：基于 DCC-GARCH 模型的时变相关性研究[J]. 北京理工大学学报（社会科学版），2019，21（5）：76-87，114.
[87] 程萧潇. 场景效应还是内容效应? 财经新闻、网络舆情对股市行情的实证检验[J]. 统计与信息论坛，2019，34（7）：69-75.
[88] Schmeling M. Investor sentiment and stock returns：some international evidence[J]. Journal of Empirical Finance，2009，16（3）：394-408.
[89] Batrinca B，Hesse C W，Treleaven P C. Examining drivers of trading volume in European markets[J]. International Journal of Finance & Economics，2018，23（2）：134-154.
[90] Wang G S，Yu G J，Shen X H. The effect of online environmental news on green industry stocks：the mediating role of investor sentiment[J]. Physica A：Statistical Mechanics and Its Applications，2021，573：125979.
[91] Angrist J D，Pischke J S. Mostly Harmless Econometrics：An Empiricist’s Companion[M]. Princeton：Princeton University Press，2009.
[92] Roodman D. How to do Xtabond2：an introduction to difference and system GMM in Stata[J]. The Stata Journal：Promoting Communications on Statistics and Stata，2009，9（1）：86-136.

Effect of Influence-based Online Investor Sentiment on Stock Returns

WANG Gaoshan，WANG Yue，DONG Yilin，ZHANG Xin
（School of Management Science and Engineering，Shandong University of Finance and Economics，Jinan 250014，China）

Abstract In order to better capture the overall market sentiment，this paper used machine learning algorithms to classify online investor comments and developed an influence-based investor sentiment index，taking into account information such as reading volume，number of likes，and number of comments. Furthermore，the paper developed regression models of weighted sentiment，unweighted sentiment，and stock returns. After controlling for variables such as stock market value，book-to-market ratio，and Beta，we found that online investor sentiment has a significant and positive impact on stock returns. Moreover，the influence-based sentiment index can better reflect changes in stock returns compared to the unweighted sentiment index，implying that the influence of online comments contains valuable information for investment decision-making and market regulation.

Key words Online comment，Influence，Machine learning，Investor sentiment，Stock returns

作者简介

王高山（1977—），男，山东财经大学管理科学与工程学院教授、博士生导师，研究方向为大数据与商务分析。E-mail：gaoshanwang@126.com。

王越（1997—），女，山东财经大学管理科学与工程学院 2021 级硕士研究生，研究方向为金融大数据。E-mail：wy971008@126.com。

董宜麟（1999—），女，山东财经大学管理科学与工程学院 2021 级硕士研究生，研究方向为金融大数据。E-mail：dong1999yx@163.com。

张新（1967—），男，山东财经大学管理科学与工程学院教授、博士生导师，研究方向为信息管理与信息系统。E-mail：zhangxin@sdufe.edu.cn。

附录 人工标注准则

1. 积极评论

（1）发布与股票相应的公司利好的消息。
（2）建仓、加仓、持仓等买进类评论。
（3）主力或庄家增持等利好消息。
（4）通过对指标的分析得出金叉等利好消息。
（5）认定此股会反弹、大涨、补涨、涨停等看涨类评论。
举例：
（1）加仓。
（2）抓紧时间上车。

2. 消极评论

（1）发布与股票相应的公司利空的消息。
（2）清仓、减仓、减持、拉高卖出等卖出类评论。
（3）主力或庄家撤资、出货等利空消息。
（4）通过对指标的分析得出死叉等利空消息。
（5）坚决认定此股为垃圾股的评论。
（6）认定此股会跳水、大跌、补跌、跌停等看跌类评论。
举例：
（1）此票垃圾。
（2）都清仓逃命吧，越快越好。

3. 中性评论

（1）概述股票现状的评论。
（2）与当前股票信息完全无关的评论。
（3）用疑问语气询问该股票发展走向的评论。
（4）只含有语气词或表情的评论。
（5）评论信息是对自己以往交易情况的评价。
（6）对两只股票或指数进行无意义对比的评论。
（7）无法单纯靠评论信息对股票未来发展做出判断的评论。
举例：
（1）今天有公告没?
（2）万物互联是分享经济。