



投资者情绪与证券投资收益: 机器学习对网络社交媒体的文本分析

万漪萍 赵岩 赵留彦

摘要 本文研究投资者有关个股的情绪对股票收益的预测价值。我们通过网络爬虫收集了线上投资者论坛中有关上证 50 指数成分股在 2019—2023 年的 700 多万条评论信息, 通过自然语言处理进行情感分类, 从而构建了关于每只股票的周度投资者情绪指数。研究发现, 投资者看涨情绪指数和关注度指数对于短期股价变化具有显著的预测能力, 并且基于这两个因子所构建的量化投资策略能够获得明显超额收益。因而, 在考察投资者情绪的效应时, 研究者不仅要考虑到整体市场情绪, 还应该关注投资者有关个股情绪的异质性。这些发现提供了有效市场理论的反面证据。

关键词 投资者情绪 投资者关注度 社交媒体 文本分析

一、引言

投资者情绪如何影响股票价格? 这是经济学家和金融业界长期争论的问题。至少自凯恩斯的“动物精神”概念提出开始, 不少研究者便认识到, 市场中由于大量投资者的交易行为仅受情绪驱动, 股价会在较长时间内偏离其基本面价值 (例如, De Long et al., 1990 ; Shiller, 1990)。股价的大幅波动一定程度上可以归因于投资者情绪因素而非基本面的变化。不过, 也有很多研究者并不认可情绪效应, 其中有代表性的观点是, 理性交易者会积极捕捉情绪交易者错误定价所造成的套利机会, 从而消除情绪效应。只有理性交易者无法充分利用这些机会时 (例如存在卖空限制), 才有可能产生情绪效应 (Miller, 1977 ; Shleifer & Vishny, 1997 ; Shen et al., 2017)。

投资者情绪并没有精确定义, 一般认为是指交易者的一种带有投机性倾向的信念, 它反映了交易者对资产价格的估值偏差, 即交易者的估值离了该资产未来现金流的现值 (Baker & Wurgler, 2007)。大量文献讨论了投资者情绪对股票收益的预测意义。例如, Baker & Wurgler (2006) 构建了投资者情绪指数, 认为投资者情绪高涨时股市收益会倾向于下降, 这一效应在市值较小、波动率较高的成长型公司上体现更为明显; Firth et al. (2015) 发现透明度较低的公司更容易受到投资者情绪的影响; Schmeling (2009) 和 Baker et al. (2012) 考察了多个发达国家股市中投资者情绪的预测能力。在中国 A 股市场, Xu & Zhou (2018) 也发现市场的投资者情绪对未来收益有预测能力。

研究中的一个关键问题是如何度量投

万漪萍 (通讯作者), 北京大学软件与微电子学院金融信息工程系, E-mail: 1901210501@pku.edu.cn; 赵岩, 北京大学软件与微电子学院金融信息工程系副教授; 赵留彦, 北京大学经济学院教授。



投资者情绪。投资者情绪指标的传统构建方法可分为直接法和间接法。直接法是通过问卷调查获得受访者的观点。我国一些研究机构发布的情绪指数就是基于直接法编制的。^①也有研究者使用消费者信心指数衡量情绪 (Lemmon & Portniaguina, 2006)。直接法比较直观, 不过也存在如下局限。第一, 可靠性较差, 因为受访者几乎没有诚实回复的动机 (Da et al., 2015)。第二, 频率较低, 对情绪的反应有滞后性。调查通常每月开展一次, 而当市场有效程度相对较高时, 错误定价只会在短期内存在, 所以使用低频数据难以发现投资者情绪的效应。第三, 考虑到样本量和成本, 调查一般只针对整体市场情绪, 而无法获得投资者关于个股的情绪。然而, 人们对股票的投机行为可能仅受有关个股的情绪驱使, 而与整体市场情绪无关 (Avery & Chevalier, 1999)。

间接法是通过可观察的资产价格或投资者交易行为间接推测投资者情绪, 例如, 通过封闭式基金折价率 (Lee, 1991; 伍燕然和韩立岩, 2007)、IPO 数量或 IPO 抑价率 (Ljungqvist & Wilhelm, 2003)、股票换手率 (Baker & Stein, 2004) 等。也有投资者基于多个间接指标和主成分分析法来构建综合性的投资者情绪指数。例如, Baker & Wurgler (2006)、Huang et al. (2015)、易志高和茅宁 (2009)。间接法相对客观, 但不够准确, 因为这些被视

作代表投资者情绪的指标很大程度上还受到情绪之外的因素影响。例如, Pástor & Veronesi (2005、2006) 表明, 不少被视为情绪因素所导致股价波动的证据, 本质上可以通过基本面价值变化来解释。^②除了投资者看涨/看跌情绪之外, 以往文献还表明, 投资者对股票的关注度和分歧度也可能影响到股票收益。例如, Kumar & Lee (2006) 发现, 市场关注度高的股票, 投资者情绪和股票收益之间的关系更强。^③

近年来, 研究者将文本分析和机器学习技术应用于网络信息的分析, 为构建投资者情绪指标提供了一条新的方法, 一定程度上克服了上述两种传统方法的不足 (Loughran & McDonald, 2016; Jiang et al., 2019)。这些方法依靠迅速发展的自然语言处理技术, 从海量投资者评论信息中有效提取情感信息。这是对投资者情绪的直接度量, 但相比传统直接法, 具有更广泛的市场覆盖面和更高频率, 不仅可用于度量市场层面情绪, 还可用于度量个股层面情绪。

本文通过对互联网社交平台中投资者的评论信息进行数据挖掘, 构建有关个股的投资者情绪指数, 并研究投资者情绪对股价变化的预测意义。我们通过网络爬虫爬取了 2019—2023 年东方财富网股吧中有关上证 50 指数成分股的约 700 万条评论信息, 进而使用大量主流的机器学习和深度学习方法对这些评论信息进行情感分

① 例如, 巨潮资讯网的投资者信心指数、北京大学的中国投资者情绪指数、上海财经大学的上海投资者信心指数。

② Sibley et al. (2016) 认为, Baker & Wurgler (2006) 的情绪指数对股票收益有预测能力, 主要是因为该指数所反映的商业周期和风险成分的变化, 而不是投资者情绪本身。

③ Barber & Odean (2008) 和 Lou (2014) 分别使用股票交易量和公司广告支出来衡量投资者关注度, Da et al. (2011) 使用谷歌搜索量来衡量关注度, Hillert et al. (2018) 用分析师预测离差衡量投资者意见分歧程度。中国 A 股相关的投资者关注度文献见董大勇和肖作平 (2011)、赵龙凯等 (2013) 及 Huang et al. (2016)。



类, 比较了各种方法执行情感分类任务的效果。最后选取效果最优的双向长短期记忆神经网络 (Bi-LSTM) 方法, 构建了关于每只股票的三维度投资者情绪指数: 看涨指数、意见趋同指数和关注度指数。基于这些指标, 我们构建了股票投资组合策略并进行回测, 以检验投资者情绪对于股票价格的预测意义。

研究从三个方面拓展了关于我国投资者情绪的研究。第一, 改进了针对个股的投资者情绪指标构建方法, 并检验了该情绪指标对于未来股价变化的预测能力。本文使用自己标注的股吧评论数据作为训练集, 选择适用于金融领域评论的“Chinese-Word-Vectors”和 Bi-LSTM 模型进行情感分类, 分类的准确率得到了提升。在此基础上, 检验了投资者情绪指标对于未来股价变化的预测能力。此前, 有研究从股民的网络搜索或评论信息中测量投资者情绪 (杨晓兰 等, 2016; 段江娇 等, 2017), 不过关于投资者情绪的预测价值仍缺乏研究。第二, 通过分析互联网社交媒体信息研究市场参与者的行为。随着互联网发展, 特别是新冠疫情以来, 投资者之间的沟通主要转为线上形式, 网络平台信息已成为投资者搜集并交流信息的主要渠道。互联网的信息传播速度远远快于传统媒体, 这可能放大投资者情绪的影响。不少文献关注了互联网对于投资者的影响 (Drake et al., 2012; Gao et al., 2020), 这些研究多是基于互联网搜索量而忽略了具体内容。与这些研究不同, 我们的方法通过对投资者评论信息的数据挖掘提取具体的文本信息, 使用海量的股吧评论数据, 能够更加

有效且全面地衡量投资者情绪。

本文以下的结构安排为: 第二节介绍数据集, 即投资者网络交流平台中有关上证 50 指数每只成分股的评论信息; 第三节通过自然语言处理对这些评论信息进行情感分类, 从而构建关于每只股票的投资者情绪指数; 第四节考察投资者情绪指数对于股票收益的预测能力, 并构建基于个股情绪的量化投资策略; 最后一节是结论。

二、数据

本文的研究样本为东方财富网股吧^①之中上证 50 指数每只成分股论坛的评论信息, 时间区间为 2019 年 1 月 1 日至 2023 年 12 月 31 日。股吧论坛是我国开放程度较高、较为活跃的投资者网络交流平台, 投资者的情感可以通过股吧交流得以明确呈现并迅速传播扩散。股吧的参与群体主要是投资 A 股的散户 (个人投资者), 我们以股吧作为研究对象, 比较贴合 A 股市场上散户占主导地位的特征。本文使用东方财富网作为研究对象, 是出于以下考虑。第一, 该股吧是目前中国最受投资者欢迎的股票网络交流平台之一, 日活跃用户数量和浏览量在所有股吧中居于首位, 目前其日浏览量超过一亿。第二, 该网站帖子的留存时间长, 有个股、主题、行业、概念等分类, 同时保留了发帖人、发帖时间、标题、点击量和评论量等信息, 为研究提供了比较全面的数据。上证 50 指数由上海证券交易所中规模大、流动性好的 50 只股票组成, 综合反映了最具市场影响力的大盘企业整体状况。这些成分股的股吧中评论者人数众多, 帖子更新速度快, 可以反

① <https://guba.eastmoney.com>.



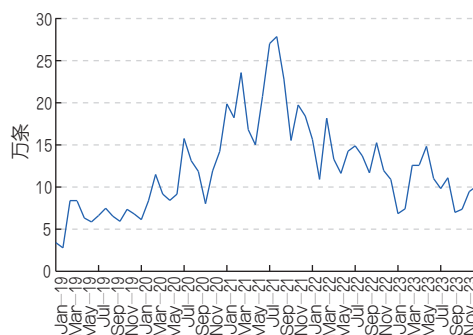
映投资者情绪的高频变化状况。相比而言，其他股票的评论信息要少得多，特别是那些市值较小、交易不活跃的股票，评论帖更新慢（甚至连续多日不更新），因而信息时效性差。

表 1 上证 50 指数成分股股吧的帖子信息统计

统计维度	数值
样本期	2019.1.1—2023.12.31
个股数	44
发帖总数	731 万
个股最大帖数	61.5 万（贵州茅台）
个股最小帖数	3.9 万（陕西煤业）
日均帖数	4 007
日最高帖数	18 069
日最低帖数	72

数据来源：东方财富网股吧。

首先利用爬虫技术收集每只样本股论坛中的所有评论信息。^① 上证 50 指数的成分股每半年会有所调整，每次调整不超过 5 只。以 2023 年底的 50 只成分股为准，剔除了其中 2019 年 1 月 1 日之后新上市的 6 只股票^②，针对剩余的 44 只股票股吧帖子信息进行爬取，共获得 758 万条帖子。^③ 我们关于每帖抓取的信息包括其标题、URL、作者、时间、浏览数及评论数。爬取文本信息之后，我们再进行数据清洗，这包括以下几步：第一，删掉重复的文本信息，也就是说对连续且内容



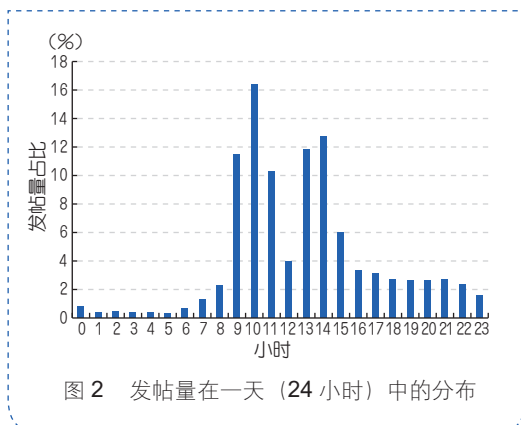
数据来源：东方财富网股吧。

图 1 上证 50 指数股吧月度评论量
(2019.1—2023.12)

完全相同的帖子去重；第二，剔除官方发布的新闻、公告以及机构路演信息，例如内含“要闻”“资讯”“播报”“公告”等词语的帖子，以及投资者和上市公司的互动信息；第三，删除转发的内容，即内部文字中包含“转”“转帖”“转发”等词语的帖子；第四，删除广告帖，即含有“荐股”“推荐”等词语的帖子；第五，去除匿名用户的发帖，因为注册用户更可能是持有或关注股票的投资者，而受雇佣的“水军”一般以匿名身份出现。按上述流程进行数据预处理后，最终剩下 731 万条帖子记录作为分析样本。帖子的基本信息见表 1。

图 1 显示，2019 年间，上证 50 指数成分股吧每月的发帖量一般在 3 万至 8 万之间波动。特别是样本期初，每月发帖量

- ① 利用 Python3.7 来编写网络爬虫程序，利用 requests 库与 BeautifulSoup 库发送 HTTP 网页请求，获取网页源代码后，解析代码可得到原始评论数据，并以 csv 文件形式存储。
- ② 被剔除的 6 只股票为三峡能源（600905）、中国电信（601728）、海光信息（688041）、金山办公（688111）、天合光能（688599）及中芯国际（688981）。
- ③ 在股吧中，绝大多数网民习惯将观点写在主贴标题中，正文部分只有“如题”二字或者重复标题内容。回帖则大多为“顶”“支持楼主”等缺乏信息含量的“灌水”，难以判断回复者的情感倾向。所以，我们在抓取评论内容时，为减少工作量，只针对主贴标题内容而不包括回复信息。



仅有3万多条,每只个股相关发帖量月均约770条。自2020年1月开始发帖量有明显上升,一般维持在每月10万条以上,有些月份甚至超过25万条。这表明样本期内东方财富网股吧热度不断上升,特别是在疫情之下,人们的交流渠道更多通过线上形式开展。图2是股吧发帖量在一天24小时中的分布状况。可以看出,发帖主要集中在股票市场交易时段9:30—11:30以及13:00—15:00,20:00—22:00的发帖量也相对较多。此外,交易日的股吧发帖量明显高于非交易日。平均而言,一周之内周一至周五每天发帖量占比均为18%~19%,而周六和周日每天的发帖量占比不足4%。

三、构建投资者情绪指标

(一) 情感分析方法

获得每只股票的评论帖子之后,对每个帖子进行情感分类。目前,文本的情感分析方法主要有两类。一类是按照金融专业情感词典进行分类,然后通过句法分析对文本进行情感分类(Chen et al., 2014;

Jiang et al., 2019; Lv et al., 2021)。^①这种情感分类方式比较直观,不过准确度严重依赖于情感词典的构建质量。由于网络用语的变化很快,很难及时构建完备的情感词典。此外,该方法中每个词汇被打上唯一分数,但同一词汇或语句在不同语境下的情感差异会非常大。词汇汇总得到的情感分数忽略了这种差异性,从而会影响分类的效果。另一类是基于机器学习算法分类,选取一部分人工情感分类的语句作为训练集,进行机器学习。面对海量文本数据,机器学习算法借助计算机力量进行情感分类,是近年来文本数据和自然语言处理中的重要进步。研究表明,在情感分析任务中机器学习算法的表现优于情感词典法(Behrendt & Schmidt, 2018)。

机器学习不是从词语而是从句子角度进行情感分析。它关注的是句子间的相似性,因而不需要情感词典,只是需要一些标注好标签的数据集训练。该方法可分为三步:第一步,从文本(语料)中抽样并人工标注,例如在情感分析中文本常被分为“看涨”“中立”“看跌”三类;第二步,通过one-hot、n-gram、TF-IDF、TextRank、Word2Vec等方法,将文本信息用向量形式表示;第三步,将转换后的文本向量输入机器学习模型,用以训练模型、调整参数,得到情感分类器,由此对所有语料进行分类。针对每一条文本,机器学习的结果可以为确切分类类别(“看涨”“中立”或“看跌”),也可以是其属于某一类别的概率。在投资者情绪分析中常用的机器学习

① 基础情感词典一般仅包括没有专业领域差异的日常词汇(例如《How Net情感词典》《BosonNLP情感词典》《DUTIR情感词典》《NTUSD情感词典》)。由于金融投资领域还有大量含有明确情绪色彩的专业用语(例如“多”“空”),因而研究中有必要将专业情感词汇加入基础情感词典,形成金融专业情感词典,例如《CSFD中文金融领域情感词典》。



习算法有支持向量机 (SVM) 分类、朴素贝叶斯分类、长短期记忆神经网络 (LSTM) 和双向长短期记忆神经网络 (Bi-LSTM)。下文中, 我们分别使用以上几种机器学习算法, 对股吧评论帖进行情感分类, 并比较分类器效果, 选择最优算法进行后续研究。

在使用机器学习方法进行分类之前, 还有必要先对评论信息进行文本预处理工作。预处理的程序包括三步: 文本分词、去除停用词和文本向量化。

第一步, 文本分词。与英文不同, 中文语句中各词语之间无空格符切分, 因此, 对文本数据挖掘之前, 要先对句子分词, 这里我们使用 Python 中的 jieba 库进行分词。

第二步, 对分词后的文本数据去除停用词。中文文本中存在较多停用词 (例如, “了” “的” “也” 之类), 它们对情感分析任务帮助不大, 反而会大幅增加后续训练的运算成本。因此, 我们整合了常见的停用词表 (包括哈工大停用词表、百度停用词表、四川大学机器智能实验室停用词库), 去除了相关停用词。^①

第三步, 文本向量化, 也就是将自然语言文本转化为计算机可识别的信息。我们使用 TF-IDF (词频-逆文本频率) 和 Word2Vec 方法将文本信息转化为文本向量, 并输入各类机器学习模型中 (包括 SVM、朴素贝叶斯、LSTM 和 Bi-LSTM), 比较分类效果。TF-IDF 方法是反映某个词条 (记为) 在整个语料库中重要性的指标,

其思路是, 一个区分度高的词条, 应该表现为在某一文档中出现频率高, 而在其他文档中出现频率低。公式表示为:

$$TF-IDF \quad (1)$$

其中, TF_w 是词条 w 在某一文档中出现的频率 (即词频): $TF_w = \text{某一文档中 } w \text{ 出现次数} / \text{该文档中所有词条数量}$; IDF_w 是包含词条 w 的文档逆频率 (逆文本频率): $IDF_w = \log[\text{语料库的文档总数} / (\text{包含 } w \text{ 的文档数} + 1)]$ 。针对已完成分词和去停用词处理之后的语料, 我们计算词条在每一个帖子中的 TF-IDF 值, 得到矩阵 X (元素 x_{ij} 为第 j 个词语在第 i 个帖子中的 TF-IDF 值)。将该矩阵作为训练集和测试集, 输入到朴素贝叶斯分类或支持向量机 (SVM) 分类器模型中, 对帖子进行情感分类。

Word2Vec 方法是通过某个词的前后文信息推算其词向量, 推算方法为通过背景词计算中心词 (CROW) 或通过中心词计算背景词 (Skip-Gram)。我们利用 Chinese-word-vectors (由 DBIIR 实验室开源) 中利用金融新闻训练好的 Skip-Gram 词向量, 将已完成前两步处理之后的语料转化为词向量。然后, 将词向量输入到两类不同的机器学习模型中——长短期记忆神经网络 (LSTM) 或双向长短期记忆神经网络 (Bi-LSTM) ——对每条帖子进行情感分类。

使用机器学习方法执行本文情感分类任务时, 训练样本的分类质量十分关键。我们从所有样本 (585 万条帖子) 中随机抽取了 13 286 条帖子, 通过作者人工阅读

① 例如, 一个帖子内容为 “很好, 一波又开始了”。对其分词结果为: [‘很’, ‘好’, ‘一波’, ‘又’, ‘开始’, ‘了’], 去停用词之后的结果为: [‘很’, ‘好’, ‘一波’, ‘又’, ‘开始’]。停用词表来源: <https://github.com/goto456/stopwords>。



表 2 机器学习模型对训练样本进行情感分类的效果 (%)

	准确率	精确率	召回率	F1 值
Bi-LSTM	82.27	82.33	82.27	82.27
LSTM	81.93	81.95	81.93	81.93
逻辑回归	76.09	76.02	76.09	75.60
SVM	76.08	75.88	76.08	75.76
朴素贝叶斯	73.50	75.74	73.50	71.72
决策树	67.76	67.38	67.76	67.37
KNN	65.72	65.83	65.72	65.52

注: 情感分类是依据每个模型将每条帖子的情感态度分为“看跌”“中性”“看涨”三类。准确率 (accuracy) 是判断正确 (把真实为“看涨”的帖子归类为“看涨”, 把真实为“中性”的帖子归类为“中性”, 把真实为“看跌”的帖子归类为“看跌”) 的帖子数与总帖子数之比, 精确率 (Precision) 是指在所有被分类为“看涨”的帖子中, 真实为“看涨”的帖子所占的比例。召回率是所有真实为“看涨”的帖子中, 被正确预测的比例。F1 值是精确率和召回率的调和平均值。

将这些帖子分为“看跌”“中立”“看涨”三类, 分别用 -1, 0 和 1 表示发帖者的情感态度。^① 为避免人工分类方式的主观性对情感分类效果产生误差, 我们还使用了已开源的标注情感分类的金融语料 9 214 条。这样, 我们所使用的训练样本共有 2.25 万条已标注情感标签的语料, 每种情感语料分别占 1/3 比例。表 2 是使用各种机器学习模型对训练样本进行情感分类的效果比较, 从中可以看出, Bi-LSTM 和 LSTM 的效果明显优于其他模型, 特别是 Bi-LSTM 各方面的效果都更占优势。因此, 我们选择用 Bi-LSTM 对每个帖子进行情感三分类, 构建投资者情绪指标。

表 3 是我们应用 Bi-LSTM 模型对训练样本及所有 (585 万条) 评论帖执行情感三分类任务的效果。其中, 从 2.25 万条训

练样本来看, Bi-LSTM 的分类效果与我们的人工分类效果十分接近。例如, 人工分类结果中被归入“看跌”类的帖子数目占比 33.33%, 而使用 Bi-LSTM 进行分类时, 其中 29.24% 被归入“看跌”类, 剩余 2.41% 和 1.68% 被错误归入“中立”和“看涨”类, 分类准确率达到 88%。^② 对于人工分类结果中被归入“中立”和“看涨”类的帖子, Bi-LSTM 分类的准确率也都接近 90%。可以看出, Bi-LSTM 倾向于把本来归为“看涨”“看跌”的帖子判为“中立”, 这意味着分类器会一定程度上弱化投资者意见, 不过这种偏向的影响不太大, 总体分类表现仍优秀。一般来说, 准确率 70% 即可认为模型的分类结果有效。可见 Bi-LSTM 模型对样本的分类结果可信度高。使用该模型对所有样本进行情感三分类, 结果为“看

① 看涨贴的例子: “国之重器, 可以涨了!” “[……] 大家赶紧补仓上车吧!” “明天涨停不敢说, 5 个点是板上钉钉了: 8 寸晶圆涨价供不应求”。看跌贴的例子: “我的乖乖明天又要跌下来” “一泻千里了, 跌停吧” “这要收割散户吗”。中性贴的例子: “反复震荡, 绝对了” “是蓄势还是出货?”。

② 准确率即是以 29.24% 除以 33.33%。



表 3 Bi-LSTM 模型在训练样本和全部样本中的情感分类

类别	人工分类结果	Bi-LSTM 分类结果 (%)		
		看跌	中立	看涨
训练样本：看跌	33.33%	29.24	2.41	1.68
训练样本：中立	33.33%	1.48	30.37	1.48
训练样本：看涨	33.33%	1.52	2.42	29.39
训练样本	2.25 万条	32.25	35.20	32.55
全样本	731 万条	29.98	47.36	22.67

跌”“中立”和“看涨”的帖子占比分别为 29.7%、46.5% 和 23.8%。

（二）投资者情绪指标

通过 Bi-LSTM 模型对全部文本信息进行情感分类之后，我们获得了每只股票相关评论帖的看跌情绪、中立情绪和看涨情绪帖子的数量。在此基础上，我们使用 Antweiler & Frank (2004) 的方法，构建了三个维度的投资者情绪指标：情感指标、关注度指标和意见趋同指标^①。情感指标反映投资者看跌、看涨或者中立。

四、投资者情绪与股票收益

（一）投资者情绪对当期股票收益的解释能力

表 4 是变量的描述性统计。使用周度面板数据，包括 44 只股票，共 252 周，除去个别缺失观测后，共 11 086 个观测。研究期内，所有样本股的周度收益率均值为 0.084%（折算为年化收益率 4.5%），表明样本期内股票市场温和上涨。股票市场指数的超额收益率为 0.107%，与上证 50 指数的收益情况接近。研究期内，投资者看

涨指数 (BI_{it}) 的均值和中位数都为负值，意味着就所有样本股平均而言，看跌的帖子数略高于看涨的帖子数。也就是说，投资者总体的情感表现相比市场的实际状况偏悲观，这可能是受到 2018 年和 2020 年上半年熊市的影响。投资者意见趋同指数的均值和中位数接近于零，说明投资者的观点分歧非常明显。有趣的是， HML_t 的均值为负数（年化值为 -1.2%），也就是说，在研究期间内价值股收益率平均略低于成长股收益率。这与文献中常见的价值股溢价并不一致。考虑到样本期只有五年，这一结果是否稳健还值得进一步讨论，或许并不构成对成熟市场中经典的“价值股溢价之谜”的挑战，不过至少可以表明“价值股溢价之谜”在中国市场中并不是十分稳健的现象。表 4 还列出了股票换手率信息。全部股票周度换手率（一周中被交易的股数除以全部流通股数）均值为 7.3%，折合日度换手率 1.5%。考虑到这些股票多为大型国有企业，一半以上流通股数并不会真正上市交易，所以真实的日度换手率均值应高于 3%，可见市场交易十分活跃。

① 三个维度投资者情绪指标的模型构建、表 4~7，见增强出版，中国知网—《金融市场研究》。



接下来使用周度面板数据估计方程(5), 结果见表5。^① 根据 Hausman 检验结果, 选择双向固定效应模型。由于面板数据存在截面相关与异方差问题, 采用 Prais-Winsten 方法计算面板矫正的稳健标准误。解释变量仅包括三个投资者情绪指标时, 模型拟合优度为 35%。考虑到被解释变量是个股收益率, 其中包含大量随机成分, 这表明我们构造的个股情绪指标具有显著解释能力。

以第 3 列结果和第 1、2 列相比较, 可见加入 Fama-French 三因子前后, 每个情绪因子的系数没有明显的改变。故仅以第 3 列结果为例来进行说明。投资者情感指数 ($BI_{i,t}$) 的系数估计值为 1.43, 高度显著, 意味着给定其他变量, 投资者情感指数每增加一个标准差 (0.45), 股票的周收益率平均会提高 0.64 个百分点。从绝对水平来看, 表明投资者看涨情绪与同期股票收益有很强的正相关性。其次, 投资者意见趋同指数 ($AI_{i,t}$) 的系数负值, 意味着投资者观点高度分歧时股票收益率往往较高。具体而言, 意见趋同指数每降低一个标准差 (0.037), 股票的周收益率平均会提高 0.015 个百分点。这可能是因为股市中卖空手段缺乏, 当投资者意见分歧较大时, 看多的投资者可以选择买入或持有股票, 而看空的投资者却不能将其观点转化为具体投资行为, 因此市场价格偏向反映看多者的观点。不过, 从经济意义上看, 特别是相比投资者情感的效应而言, 意见趋同指数的影响并不大。此外, 投资者关注度 ($ATTI_{i,t}$) 的系数几乎为零, 表明关注度与当期收益

率没有显著相关性。

表 5 中第 4 列结果是使用 Fama-MacBeth 回归方法的估计结果, 相比第 1 列结果, Fama-MacBeth 方法所估计的 $AI_{i,t}$ 和 $ATTI_{i,t}$ 的系数变得显著, 不过经济意义差别并不大。具体而言, 根据 Fama-MacBeth 方法的结果, 意见趋同指数每提高一个标准差, 股票的周收益率平均会降低 0.029 个百分点。投资者关注度 ($ATTI_{i,t}$) 的系数为 0.046, 意味着关注度每提高一个标准差, 当期收益率平均会提高 0.044 个百分点。总体而言, 二者的影响比较小, 特别是相比投资者情感的效应而言更是如此。

(二) 投资者情绪对股票收益的预测能力

通过面板数据的固定效应回归, 我们发现投资者有关个股的情绪因子对于同期个股收益率变化具有解释能力, 不过, 这未必意味着是情绪因子导致了收益率变化。因为, 情绪因子和股票收益率可能同时受其他因素的影响——比如宏观经济指标、行业信息或公司个体信息, 此外, 也有可能是股价的变化引起了投资者情绪的变化。为了考察情绪指数是否含有关于股票收益率的预测信息, 我们进一步估计包含个股收益率以及三个情绪因子的面板向量自回归模型 (Panel Vector Autoregression, PVAR)。根据贝叶斯信息准则 (BIC), 设定 PVAR 模型的滞后阶数为 3。基于 PVAR 模型的格兰杰因果关系检验结果见表 6。每个投资者情绪因子不是收益率格兰杰原因的原假设均在 1% 显著水平上被拒绝, 说明投资者对个股的看涨情绪、关注度和意见趋同性对未来股价

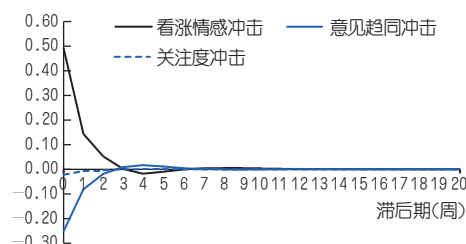
① 为了防止出现伪回归情况, 采用 Fisher-ADF 和 Fisher-PP 方法对解释变量与被解释变量开展平稳性检验, 结果表明这些变量均不存在单位根。



变化都有一定预测能力。不过，模型的脉冲响应函数（图3）表明，投资者情感指标（ $BI_{i,t}$ ）的冲击对股票未来收益率的影响最大，投资者关注度（ $ATTI_{i,t}$ ）的冲击对股票未来收益率的影响较小，而投资者意见趋同因子（ $AI_{i,t}$ ）的冲击几乎没有影响。

（三）基于投资者情绪的量化投资策略

我们发现投资者情绪对股票收益具有统计意义上显著的预测能力，那么，这种预测是否具有经济价值？换言之，考虑到交易成本后，基于投资者情绪进行量化投资是否可以获得超额收益？为回答这些问题，我们基于前文的投资者情绪指标建立股票投资组合，并进行回测。回测使用的样本仍为前述上证50指数的成分股（剔除2019年后上市的新股后，共44只股票）。在 t 期的最后一个交易日，根据当期的投资者情绪因子对所有样本股进行排序，构建 $t+1$ 期的两个投资组合。具体地，选取 t 期投资者看涨情感因子值（ $BI_{i,t}$ ）排序最高的10只股票构成一个组合（ $BI_{H,t}$ ）、排序最低的10只股票构成另一个组合（ $BI_{L,t}$ ），每个组合使用流通市值加权法构建，即每只股票配置比例由其在10只股票中的流通市值占比决定。此外，我们使用所有成分股，按流通市值比例加权构建基准组合。关于每期时长（调仓频率），考虑月度频率和周度频率，分别在每月或每周的最后一个交易日考虑改变投资组合。前后两期相比，如果排序最高（或最低）的10只股票有变化，则进行调仓，即按照排序变化的股票相应改变组合 $BI_{H,t}$ 或 $BI_{L,t}$ 。我们设定



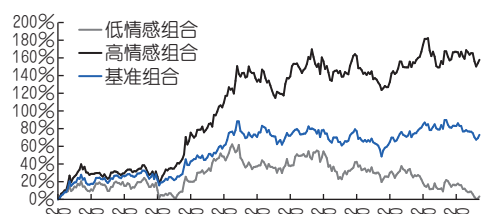
注：脉冲响应函数来自滞后阶数为8的面板向量自回归模型，图中三条线分别表示投资者看涨指数（ $BI_{i,t}$ ）、意见趋同指数（ $ATTI_{i,t}$ ）和关注度指数（ $AI_{i,t}$ ）一个标准差冲击对周度股票收益的长期累积效应。

图3 股票收益率对投资者情绪冲击的脉冲反应函数

交易费用包括：印花税千分之1（仅在卖出时收费），佣金率万分之1.5（买、卖双向收费）。^①按照同样思路，我们还可以根据另外两个因子——投资者关注度（ $ATTI_{i,t}$ ）和投资者意见趋同度（ $AI_{i,t}$ ）——分别构建投资组合。

表7列出了周度调仓频率下选股策略的回测结果。按照投资者情感因子构建组合时，高情感投资组合表现明显优于低情感组合和基准组合。高情感组合年化收益率为21%，远高于低情感组合的0.6%。此外，两类组合的波动率几乎相同。因此，高情感组合取得了更高的夏普率。相反，以投资者关注度因子构建组合时，低关注度组合表现更优——收益率较高且波动率较小。不过，以投资者意见趋同指标所构建的投资策略则表现不佳，低趋同组合和高趋同组合的收益率都低于基准组

① 印花税率由政策统一规定，佣金率则是A股市场中散户普遍面临的数值。选股策略及量化回测评价指标计算均由我们编写的Python3.7程序完成。



注: 回测使用的样本为上证 50 指数的成分股。在 t 期的最后一个交易日, 分别根据当期的三个投资者情绪指数对个股进行排序, 构建 $t+1$ 期的两个投资组合。高情感组合由 t 期投资者看涨情感因子值 ($BI_{i,t}$) 最高的 10 只股票组成, 低情感组合由情感因子值最低的 10 只股票组成, 基准组合由所有样本股按流通市值比例加权构成。

数据来源: 金融交易数据通过 Tushare 平台的 python 接口提取。

图 4 基于投资者情感因子的投资策略业绩回测

合。表 7 还列出了月度调仓频率的结果, 这与周度频率结果没有本质差异。^① 图 4 是基于投资者情感因子构建的投资策略业绩表现。样本期内, 高情感组合相比基准组合的累积超额收益十分明显, 这种超额收益主要体现在 2020 年新冠疫情突袭而至之时。^②

五、结论

我们使用网络爬虫爬取了 2019—2023 年间东方财富网股吧中有关上证 50 指数成分股的评论信息, 通过机器学习方法——特别是 Bi-LSTM 模型——对这些评论信

息进行情感分类, 从而对每只股票构建了三个维度投资者情绪指数: 投资者情感指标、投资者意见趋同指标和投资者关注度指标 3。我们发现, 投资者情感指标和投资者关注度指标对于短期股价变化具有预测能力, 基于这两个因子所构建的量化投资策略, 考虑交易成本后仍然可以获得超额收益, 但基于投资者意见趋同指标的选股策略无效。

本文结果表明, 单只股票收益很大程度上受投资者关于个股的情绪影响, 因而, 考察投资者情绪的效应时, 研究者不能仅考虑到整体市场情绪, 还应该重点关注投资者对个股情绪的异质性。这凸显了机器学习和深度学习技术在投资者情绪研究中的优势。应用近年来迅速发展的自然语言处理技术, 研究者可以从海量的投资者评论信息中迅速有效提取发帖人的情感信息。相比传统方法, 这样所构建出的投资者情绪指标具有更广的覆盖面、更高的频率。不仅如此, 自然语言处理技术的另一个重要优势在于它可以方便地用于度量个股层面的投资者情绪, 而传统方法一般仅用于度量市场层面的情绪。^[N]

学术编辑: 卢超群

- ① 月度和周度基准组合的年化收益率略有差异。这是因为周度组合中, 每周会按照成分股流通市值调整持仓比例, 而月度组合是每月调整持仓比例, 所以两个基准组合的个股配置比例略有别。
- ② 基于投资者关注度因子构建的投资策略超额收益较小, 基于投资者意见趋同因子构建的投资策略没有超额收益, 考虑到篇幅限制, 这里未列出二者的业绩表现走势图。
- ③ 我们爬取了东方财富网股吧 2019—2023 年所有上证指数成分股的帖子信息, 并以此构建了日度投资者情绪指标, 但发现该指标对未来股票收益缺乏预测价值。这可能是由于大量股票评论不活跃 (不少股票甚至连续多个交易日无评论信息), 因而所构建的指标不能及时反映投资者情绪变化。近年来, 该股吧活跃用户量和发帖数量迅猛增长。为了检验本文结果的稳健性, 我们可尝试用同类方法构建关于更多个股、更高频率的投资者情绪指数。



参考文献

- [1] 董大勇,肖作平.证券信息交流家乡偏误及其对股票价格的影响:来自股票论坛的证据[J].管理世界,2011(01): 52-61+188.
- [2] 段江娇,刘红忠,曾剑平.中国股票网络论坛的信息含量分析[J].金融研究,2017(10):178-192.
- [3] 伍燕然,韩立岩.不完全理性、投资者情绪与封闭式基金之谜 [J].经济研究,2007(03):117-129.
- [4] 杨晓兰,沈翰彬,祝宇.本地偏好、投资者情绪与股票收益率:来自网络论坛的经验证据[J].金融研究,2016(12):143-158.
- [5] 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.
- [6] 赵龙凯,陆子昱,王致远.众里寻“股”千百度——股票收益率与百度搜索量关系的实证探究[J].金融研究,2013(04):183-195.
- [7] Antweiler W, Frank M Z. Is all that talk just noise? The information content of internet stock message boards[J]. Journal of Finance, 2004, 59(3): 1259-1294.
- [8] Avery C, Chevalier J. Identifying investor sentiment from price paths: the case of football betting[J]. Journal of Business, 1999, 72(4): 493-521.
- [9] Baker M, Stein J C. Market liquidity as a sentiment indicator[J]. Journal of Financial Markets, 2004, 7(3): 271-299.
- [10] Baker M, Wurgler J. Investor sentiment and the cross-section of stock returns[J]. Journal of Finance, 2006, 61(4): 1645-1680.
- [11] Baker M, Wurgler J. Investor sentiment in the stock market[J]. Journal of Economic Perspectives, 2007, 21(2): 129-152.
- [12] Baker M, Wurgler J, Yuan Y. Global, local, and contagious investor sentiment[J]. Journal of Financial Economics, 2012, 104(2): 272-287.
- [13] Barber B M, Odean T. All that glitters: the effect of attention and news on the buying behavior of individual and institutional investors[J]. Review of Financial Studies, 2008, 21(2): 785-818.
- [14] Behrendt S, Schmidt A. The Twitter myth revisited: Intraday investor sentiment, Twitter activity and individual-level stock return volatility[J]. Journal of Banking and Finance, 2018, 96: 355-367.
- [15] Chen H, De P, Hu Y J, et al. Wisdom of crowds: the value of stock opinions transmitted through social media[J]. Review of Financial Studies, 2014, 27(5): 1367-1403.
- [16] Da Z, Engelberg J, Gao P. In search of attention[J]. Journal of Finance, 2011, 66(5): 1461-1499.
- [17] 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.
- [18] Drake M S, Roulstone D T, Thornock J R. Investor information demand: evidence from Google searches around earnings announcements[J]. Journal of Accounting Research, 2012, 50(4): 1001-1040.
- [19] Fama E F, French K R. Common risk factors in the returns on stocks and bonds[J]. Journal of Financial Economics, 1993, 33: 3-56.
- [20] Firth M, Wang K, Wong S M. Corporate transparency and the impact of investor sentiment on stock prices[J]. Management Science, 2015, 61(7): 1630-1647.
- [21] Gao Z, Ren H, Zhang B. Googling investor sentiment around the world[J]. Journal of Financial and Quantitative Analysis, 2020, 55(2): 549-580.
- [22] Hillert A, Jacobs H, Müller S. Journalist disagreement[J]. Journal of Financial Markets, 2018, 41: 57-76.
- [23] Huang D, Jiang F, Tu J, Zhou G. Investor sentiment aligned: a powerful predictor of stock returns[J]. Review of Financial Studies, 2015, 28(3): 791-837.
- [24] Huang Y, Qiu H, Wu Z. Local bias in investor attention: evidence from China's internet stock message boards[J]. Journal of Empirical Finance, 2016, 38: 338-354.
- [25] Jiang F, Lee J, Martin X, Zhou G. Manager sentiment and stock returns[J]. Journal of Financial Economics, 2019, 132(1): 126-149.
- [26] Klemola A, Nikkinen J, Peltomäki J. Changes in investors' market attention and near-term stock market returns[J]. Journal of Behavioral Finance, 2016, 17(1): 18-30.



- [27] Kumar A, Lee C M. Retail investor sentiment and return comovements[J]. *Journal of Finance*, 2006, 61(5):2451-2486.
- [28] Lee C M, Ma P, Wang C C. Search-based peer firms: aggregating investor perceptions through internet co-searches[J]. *Journal of Financial Economics*, 2015, 116(2):410-431.
- [29] Lee C M, Shleifer A, Thaler R H. Investor sentiment and the closed-end fund puzzle[J]. *Journal of Finance*, 1991, 46(1):75-109.
- [30] Lv Y, Piao J, Li B, et al. Does online investor sentiment impact stock returns? Evidence from the Chinese stock market[J]. *Applied Economics Letters*, 2015, 29(15):1434-1438.
- [31] Lemmon M, Portniaguina E. Consumer confidence and asset prices: some empirical evidence[J]. *Review of Financial Studies*, 2006, 19(4):1499-1529.
- [32] Ljungqvist A, Nanda V, Singh R. Hot markets, investor sentiment, and IPO pricing[J]. *Journal of Business*, 2006, 79(4):1667-1702.
- [33] Lou D. Attracting investor attention through advertising[J]. *Review of Financial Studies*, 2014, 27(6):1797-1829.
- [34] Loughran T, McDonald B. Textual analysis in accounting and finance: a survey[J]. *Journal of Accounting Research*, 2016, 54(4):1187-1230.
- [35] Miller E M. Risk, uncertainty, and divergence of opinion[J]. *Journal of Finance*, 1977, 32(4):1151-1168.
- [36] Pástor L, Veronesi P. Rational IPO waves[J]. *Journal of Finance*, 2005, 60(4):1713-1757.
- [37] Pástor L, Veronesi P. Was there a Nasdaq bubble in the late 1990s[J]. *Journal of Financial Economics*, 2006, 81(1):61-100.
- [38] Schmeling M. Investor sentiment and stock returns: some international evidence[J]. *Journal of Empirical Finance*, 2009, 16(3):394-408.
- [39] Shen J, Yu J, Zhao S. Investor sentiment and economic forces[J]. *Journal of Monetary Economics*, 2017, 86:1-21.
- [40] Shiller R J. Speculative prices and popular models[J]. *Journal of Economic Perspectives*, 1990, 4(2):55-65.
- [41] Shleifer A, Vishny R W. The limits of arbitrage[J]. *Journal of Finance*, 1997, 52(1):35-55.
- [42] Sibley S E, Wang Y, Xing Y, et al. The information content of the sentiment index[J]. *Journal of Banking and Finance*, 2016, 62:164-179.
- [43] Stambaugh R F, Yu J, Yuan Y. The short of it: investor sentiment and anomalies[J]. *Journal of Financial Economics*, 2012, 104(2):288-302.
- [44] Xu H C, Zhou W X. A weekly sentiment index and the cross-section of stock returns[J]. *Finance Research Letters*, 2018, 27:135-139.

Stock Returns and Investor Sentiment: Machine Learning for Textual Analysis of Online Social Media

WAN Yiping ZHAO Yan ZHAN Liuyan
(Peking University)

Abstract This paper assesses the link between investor sentiment and stock investment returns. Researchers made use of a web crawler to capture 7 million posts on the constituent stocks of the Shanghai Stock Exchange 50 Index by means of the online Eastmoney Investor Forum from 2019 to 2023. Relying on textual analysis of the captured social media posts, a weekly measure of investor sentiment was created for each stock. The study found that the weekly measures of bullish investor sentiment and investor attention had significant predictive ability for short-term stock returns, and a quantitative investment strategy based on these measures could obtain substantial excess returns. These findings were consistent across a number of different models and specifications, providing further evidence against the efficient market hypothesis. Moreover, while market sentiment plays a role in explaining the overall mispricing of assets, stock-specific sentiment is likely to be informative.

Keywords Investor Sentiment, Investor Attention, Social Media, Textual Analysis

JEL Classification G10 G12 G14