Operations Research and Fuzziology 运筹与模糊学, 2023, 13(5), 5231-5237
Published Online October 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.135524

Hans汉斯

基于朴素贝叶斯算法的投资者情绪对于股票收益率的影响

徐　可

上海理工大学管理学院，上海

收稿日期：2023年8月17日；录用日期：2023年10月11日；发布日期：2023年10月20日

摘　要

马科维茨的投资组合理论假设投资者是理性的，然而，最近几年股票市场出现了剧烈的波动，这种波动很难用宏观经济因素或公司自身情况来解释，行为金融学应运而生，投资者情绪在股市中的作用也备受关注。随着中国股市的发展，投资者的情绪很容易对股票市场产生影响，从而导致股票收益的波动，鉴于投资者情绪状态无法直接观测到，对其如何进行度量一直是一个难题。本文在传统金融学的基础上引入了行为金融理论，考虑了市场因素和公司基本面因素，并加入了投资者情绪因子，用以解释股票收益率的波动。在实证方面，研究使用了沪深300指数成分股中300个企业的数据，并使用朴素贝叶斯分类算法构建了投资者情绪指标。通过个股技术指标来衡量投资者情绪，填补了沪深300成分股中个股情绪面板数据的空白。同时，还构建了多元线性模型进行回归，分析了投资者情绪对股票收益率的影响。结果表明，随着投资者情绪得分的提高，股票的回报率也会随之上升，投资者情绪的变化在一定的程度上会对股票市场的投资收益产生影响，更清晰体现出股市中投资者情绪对股票风险的影响机理，这对个人和机构投资者在选股时具有重要意义。

关键词

投资者情绪，朴素贝叶斯算法，回归分析，股票收益率

The Impact of Investor Sentiment on Stock Returns Based on the Naive Bayes Algorithm

Ke Xu

Business School, University of Shanghai for Science and Technology, Shanghai

Received: Aug. 17th, 2023; accepted: Oct. 11th, 2023; published: Oct. 20th, 2023

文章引用: 徐可. 基于朴素贝叶斯算法的投资者情绪对于股票收益率的影响[J]. 运筹与模糊学, 2023, 13(5): 5231-5237.
DOI: 10.12677/orf.2023.135524

Abstract

Harry Markowitz's portfolio theory assumes that investors are rational; however, the stock market has experienced dramatic volatility in recent years, which is difficult to explain in terms of macroeconomic factors or a company's own circumstances. Behavioral finance has emerged, and the role of investor sentiment in the stock market has received much attention. With the development of China's stock market, investors' emotions can easily have an impact on the stock market, which leads to the fluctuation of stock returns, and given that investors' emotional state cannot be directly observed, it has been a difficult problem to measure how it is measured. This paper introduces behavioral finance theory based on traditional finance, considers market factors and firm fundamentals, and adds an investor sentiment factor to explain the volatility of stock returns. In terms of empirical evidence, the study uses data from 300 companies in the constituent stocks of the CSI 300 index and constructs an investor sentiment indicator using a simple Bayesian classification algorithm. By measuring investor sentiment through technical indicators of individual stocks, it fills the gap of panel data of individual stock sentiment in CSI 300 constituent stocks. A multivariate linear model is also constructed for regression to analyze the impact of investor sentiment on stock returns. The results show that as the score of investor sentiment increases, the return of stocks will also increase, and the change of investor sentiment will have an impact on the investment return of the stock market to a certain extent, and these conclusions reflect more clearly the mechanism of the impact of investor sentiment on stock risk in the stock market, which is of great significance to individual and institutional investors in stock selection.

Keywords

Investor Sentiment, Naive Bayes Algorithm, Regression Analysis, Stock Returns

1. 引言

传统金融学理论假设市场是有效的，投资者是理性人，但市场上存在很多传统金融学理论无法解释的金融异象，比如严重偏离基本面的股市暴涨暴跌、投资组合分散化不足等，因此学者们开始寻求理论上的突破，在这种背景下产生了行为金融理论。行为金融理论对传统金融理论所作的假设进行反思并试图从“人”的角度来解释市场行为，充分考虑了投资者的心理等因素。现有研究结果表明股票的价格会受到投资主体行为的影响，情绪作为影响投资者心理活动进而影响其行为的一种因素逐渐成为该领域重要的研究问题。

学者们放松了理性人的假设，从投资者的认知偏差、投资者情绪和投资动机等方面展开研究，经典的研究成果有 Kahneman 和 Tversky (1979) [1]提出的“展望理论”，Bell (1982) [2]、Loomes 和 Sugden (1982) [3]分别独立提出的“后悔理论”等，这些理论用投资者的非理性行为解释金融异象的来源。此外，De Long 等(1990) [4]提出了噪音交易模型(DSSW)，Barberis 和 Shleifer 等(1998) [5]提出了投资者心态模型(BSV)等，都充分证明了情绪影响股价。中国 A 股市场经过三十多年的发展，总市值已位居世界第二，但我国投资者仍以中小投资者为主，持股市值在 50 万以下的中小投资者占比高达 97% [6]。中小投资者相比机构投资者专业知识更少，其投资决策极易受到自身情绪的影响。

2. 基于朴素贝叶斯算法的投资者情绪量化

2.1. 样本选取与数据来源

本文选择沪深 300 成分股中 300 家上市公司作为样本，考虑到股吧发帖量、数据可获得性，样本区间选取为 2022 年 1 月 1 日至 2022 年 6 月 30 日，观测值频率为日度数据。为保证数据完整性和连续性，剔除以下样本：1) 个股讨论帖子数目少于 10 个，无法满足挖掘需要的个股；2) 剔除金融类公司；3) 剔除 ST 和*ST 的公司。最终获得约 24 万个日度观测值。

经过对多个股票论坛的比较，选择东方财富网的股吧论坛发帖作为文本挖掘的来源，可以保证足够的样本容量，且相较其他论坛有更长的历史追溯性。

2.2. 投资者情绪提取与度量

朴素贝叶斯是基于贝叶斯定理的分类算法，它假设“特征”之间是相互独立的，这也是“朴素”说法的由来，是由训练数据学习得到联合概率分布 $P(X,Y)$，最后求出后验概率分布 $P(Y|X)$的一种预测模型，基本原理如下：

1) 假设待分类文本 $d=\{f_1,f_2,f_3,\cdots,f_n\}$，其中 n 是特征维数，也即待分类文本可以划分成 n 个特征，最终用特征向量来表示待分类文本；

2) 样本数据中共有 m 个不同的标签，假设标签集合可表示为 $L=\{l_1,l_2,l_3,\cdots,l_m\}$；

3) 根据训练数据学习计算先验概率 $P(l_i)=N_i/N$，也即计算每个标签的概率，其中 N_i 是训练数据集中属于标签 l_i 的样本数量，N 是训练数据集中的样本总数；

4) 计算条件概率 $P(d\,|\,l_i)$。由于朴素贝叶斯假定各特征之间相互独立，有：

$$P(d\,|\,l_i)=P(f_1,f_2,f_3,\cdots,f_n)=P(f_1|l_i)P(f_2|l_i)P(f_3|l_i)\cdots P(f_n|l_i)=\prod_1^n P(f_i|l_i)$$

朴素贝叶斯分类算法就是通过训练数据学习得到先验概率和条件概率，从而计算出联合概率分布，最终求出待分类文本 d 属于各个标签 l_i 的概率值 $P(l_i|d)$，比较这些概率值，最大概率对应的标签就是待分类文本 d 的标签，公式如下：

$$P(l_i|d)=\frac{P(d|l_i)P(l_i)}{P(d)}=\frac{\prod_1^n P(f_i|l_i)}{P(d)}$$

5) 综上，朴素贝叶斯分类模型可以表示为：

$$P(l_i|d)=\max_{l_i\in L}\prod_1^n P(f_i|l_i)P(l_i)$$

2.3. 投资者情绪度量

本文选取东方财富网中个股股吧的实时发帖内容。

1) 首先，借助 Python 爬取了约 24 万条文本内容，每条文本内容都包含股票代码、发帖人、发帖时间、评论内容、点击数和阅读数等标签。

2) 其次，运用 Excel 和 Python 对文本进行数据预处理，剔除无效发帖、重复发帖、空缺行、外部链接、杂乱符号等，初步获得待处理文本。再次，运用 Python 中的 Jieba 分词对待处理文本进行分词，同时去除停用词(如“的”，“了”，“是”等)，见表 1。

3) 最后，由于分词之后的文本还是非结构化文本信息，使用 TF-IDF 和卡方统计(chi-square)筛选出

Table 1. Processed data
表 1. 处理后数据

股票代码	日期	时间	发帖人 ID	点击数	阅读数	处理后评论内容
600000	6.30	22:44	风马牛不相*****	540	1	(剔除该数据)
600000	6.30	21:54	身高一米九*****	497	1	清仓了买入了
600000	6.30	20:50	股友 6217X*****	402	1	好股票
600000	6.30	18:46	小**	692	7	今天表现不错上涨就好
600000	6.30	18:39	小楼*****	302	0	明天高开低走大幅下跌
...	...	...	...	...	...	...
300999	4.13	9:21	东财网友 73385136397*****	125	0	金龙鱼满仓进城
300999	4.13	9:30	梦醒*****	120	0	金龙鱼你真行
300999	4.13	9:22	踏浪*****	115	0	金龙鱼搞什么妖
300999	4.13	9:20	夜衍**	245	2	底部大宗交易吓散户的以下是历史大底当年我买伊利翻倍一模一样

排名靠前的前 K 个特征来表示待分类文本，以 1000 个数据做例子，前 k 个特征值，设置 k 为 10，将每一条待分类文本转化为向量形式供计算机学习。

4) 为分析每条文本包含的情绪倾向，利用 Python 编写基于朴素贝叶斯的情感分类模型。首先需要准备训练数据集，人工将帖子分为“乐观”、“悲观”和“中性”三类，最终得到各类样本数量如下：乐观 2313 条，悲观 3069 条，中性 2618 条。接着用训练数据集基于 Python 中的 Sklearn 库中自带的 Naïve.bayes 来训练模型。

为了测试模型准确率，随机挑选 100 帖子，用训练好的模型预测其所属标签，然后将模型预测的结果和人工标注结果进行比对，测得模型准确率为 82%。最后，用训练好的模型预测所有文本的情绪倾向，基于当日得到的乐观/悲观帖子数构建日度投资者情绪：$Score_{i,t} = \ln\left[\left(1+pos_t\right)/\left(1+neg_t\right)\right]$

上式中，$post$ 代表第 t 日的乐观帖子数；neg_t 代表第 t 日的悲观帖子数。

3. 投资者情绪影响股票收益率的回归分析

3.1. 变量选取

本文选取沪深 300 指数成分股内 300 家企业 2022 年半年报披露的数据为样本，同时筛选处理：

Table 2. Variable selection and their explanations
表 2. 变量选取及其解释

变量符号		指标解释
被解释变量	Return	平均收益率的期间年化值。
解释变量	Score	通过计算得出的个股投资者情绪分数。
控制变量	Size	流通市值的对数。
	Lev	资产负债率，总负债/总资产。
	ROA	资产收益率，净利润/总资产。
	CFPS	每股现金流量净额。
	Turnover	换手率，成交量/流通总股数。

1) 剔除样本期间 ST、*ST 和 PT 类股票；

2) 删除了缺失数据；

3) 剔除离群值，对模型中的变量进行了 1%~99%的缩尾处理。

最终形成了 298 家企业在 2022 年的截面数据，共计 2086 条观测值。数据来自同花顺 iFinD 数据库，使用的计量软件为 STATA16。

表 2 列示了回归变量的名称、符号以及处理方法。

3.2. 模型构建

为了保证实证结果的真实准确性，根据以往学者们的研究结论，本文将公司规模、企业资本结构、企业盈利能力、换手率、每股现金流量净额等可能影响股票收益率的各种因素作为控制变量，此回归模型为：

$$Return_i = \alpha_0 + sScore_i + mSize_i + bLev_i + oROA_i + nTurnover_i + rCFPS_i + u_i$$

为了统一变量量纲，对公司流通市值取对数得到 *Size*。*Return* 是股票 *i* 在第 2022 年年中的年化收益率，α_0 为模型截距项，*Score* 为股票 *i* 在第 2022 年年中的投资者情绪，控制变量为股票 *i* 在第 2022 年的公司市值的对数 *mSize*、资产负债率 *bLev*、资产收益率 *oROA*、股票换手率 *nTurnover*、每股现金流量净额 *rCFPS*，*s*、*m*、*b*、*o*、*n*、*r* 分别是控制变量相应的系数。

3.3. 实证研究及结果分析

在上文模型构建与数据选取的基础上，运用多元线性回归模型进行实证研究，投资者情绪对于股票收益率的影响是否显著。

3.3.1. 相关性分析

在系统性地对各项披露的数据进行整合分析后，对所得结果进行回归性的相关性检验，对各个存在变量之间的相关性进行整体性的初步判断，另外，也需要严谨的检验在各个相关变量之间是否有多重共线关系的显性存在。

Table 3. Correlation analysis of variables
表 3. 各变量的相关性分析

	Return	Score	Size	Lev	ROA	CFPS	Turnover
Return	1						
Score	0.106*	1					
Size	0.053	-0.003	1				
Lev	0.006	-0.282***	0.206***	1			
ROA	0.097*	0.207***	-0.020	-0.603***	1		
CFPS	-0.004	0.229***	0.083	-0.097*	0.169***	1	
Turnover	0.031	0.523***	-0.410***	-0.155***	0.109*	0.150***	1

***p < 0.01，**p < 0.05，*p < 0.1。
(注：*、**、***分别表示数据的双尾检验在 10%、5%、1%的显著性水平上显著。)

由表 3 可以看出，本文所构建实证模型的解释变量我国投资者情绪得分(Score)与被解释变量(Return)的相关系数为 0.106，可以看出我国投资者情绪得分与股票收益率之间存在正相关关系。从相关性分析结

果可以看出，本文所使用的变量的相关系数的绝对值均未超过 0.75，证明变量直接没有较为突出显性的多重共线关系，模型可靠。

3.3.2. 回归分析

本文对模型进行回归分析，结果见表 4。

Table 4. Regression analysis of the impact of investor sentiment on stock returns
表 4. 投资者情绪影响股票收益率的回归分析

	(1)
	Return
Score	134.31*
	(1.696)
Size	83.02
	(0.442)
Lev	13.38
	(1.588)
ROA	46.41**
	(2.101)
CFPS	−70.37
	(−0.783)
Turnover	−0.20
	(−0.165)
_cons	−1761.57
	(−0.808)
N	298
r2_a	0.011

注：*、**、***分别表示在 10%、5%、1%的显著性水平上显著(双尾检验)；括号内为在企业层面经过聚类(cluster)调整的 t 检验值。

由此可知，本文的关键解释变量 Score 在 10%的水平上显著为正，即在控制企业规模(Size)、资产负债率(Lev)、资产收益率(ROA)、股票换手率(Turnover)、每股现金流量净额(CFPS)等变量情况下，回归结果显示在 10%的置信水平下我国投资者情绪得分(Score)与股票收益率(Return)正向相关。换言之，随着投资者情绪得分的提高，股票的回报率也会随之上升。

4. 主要结论

本文选取了全部沪深 300 成分股内上市公司 2022 年 1 月至 2022 年 6 月共 6 个月内披露的各个初始数据及其他可计算数据，通过朴素贝叶斯分类算法构建了投资者情绪指标(Score)。同时选取沪深 300 指数成分股内 300 家企业 2022 年半年报披露的数据为样本，通过线性回归模型检验投资者情绪对股票收益率的影响。根据回归结果可以看出：本文的关键解释变量 Score 在 10%的水平上显著为正，即在控制企

业规模(Size)、资产负债率(Lev)等变量情况下，回归结果显示在 10%的置信水平下我国投资者情绪得分(Score)与股票收益率(Return)正向相关。换言之，随着投资者情绪得分的提高，股票的回报率也会随之上升，投资者情绪的变化在一定的程度上会对股票市场的投资收益产生影响。

5. 相关建议

中国金融证券市场主要由个人投资者推动，专业机构投资者参与有限。然而，由于缺乏专业经验，对市场了解不足，散户投资者往往容易跟风，做出冲动的决定，尤其是在市场大幅下跌时。投资者情绪激化，导致极度恐惧和负面情绪。这种极度恐惧会进一步加剧负面情绪。因此，作为个人投资者，关键是要不断加强专业知识学习，了解投资价值，摒弃投机观念，建立健全投资风险机制，提高风险意识，更好地控制自己的情绪，以多元化的投资心态，规避风险，从而减少市场波动，提高投资收益。

中国资本市场中的专业机构投资者只占金融市场的约 3%。然而，与散户相比，机构投资者具有更大的优势，这是由于在资产配置方面，他们大多运用大数据、量化投资等技术，往往优先考虑长期投资，注重对企业的基本面分析，并利用金融衍生品对冲极端风险。这种谨慎理性的投资策略将对中国股市产生积极影响。因此，政府应营造更加良好的投资环境，降低金融市场的准入门槛，鼓励机构投资者的创新与发展，提高其市场占有率。机构投资者应规范投资标准，主动增强金融市场责任感，积极提高市场参与度。

参考文献

[1] Daniel, K. and Amos, T. (1979) On the Interpretation of Intuitive Probability: A Reply to Jonathan Cohen. *Cognition*, 7, 409-411. https://doi.org/10.1016/0010-0277(79)90024-6

[2] Bell, D.E. (1982) Regret in Decision Making under Uncertainty. *Operations Research*, **30**, 803-1022. https://doi.org/10.1287/opre.30.5.961

[3] Loomes, G. and Sugden, R. (1982) Regret Theory: An Alternative Theory of Rational Choice under Uncertainty. *The Economic Journal*, **92**, 805-824. https://doi.org/10.2307/2232669

[4] De Long, J.B., Shleifer, A., Summers, L.H. and Waldmann, R.J. (1990) Noise Trader Risk in Financial Markets. *Journal of Political Economy*, **98**, No. 4. https://doi.org/10.1086/261703

[5] Barberis, N., Shleifer, A. and Vishny, R. (1998) A Model of Investor Sentiment. *Journal of Financial Economics*, **49**, 307-343. https://doi.org/10.1016/S0304-405X(98)00027-0

[6] 黄灵. 投资者情绪、知情交易与资本市场反应[J]. 时代金融, 2022(11): 28-31.