ATLANTIS PRESS

Advances in Social Science, Education and Humanities Research (ASSEHR), volume 314

4th International Conference on Social Sciences and Economic Development (ICSSED 2019)

An Empirical Study of Stock Return and Investor Sentiment Based on Text Mining and LSTM

Tianyu Ren

School of Management and Engineering, Nanjing University, Nanjing, Jiangsu, China

704226242@qq.com

Keywords: Investor sentiment; natural language processing; text sentiment analysis; deep learning; LSTM; stock return; expanded asset pricing model.

Abstract. Based on the development of social network and big data, we adopt the unstructured text-based investors' comment data mining from the stock bar forum, and use long short-term memory neural network for text sentiment analysis to build a more accurate investor sentiment indicator. Based on this indicator, an empirical study on the component stocks of the GEM Composite Index is conducted to explore the impact of investor sentiment on stock return. Through a full sample stock selection test, we find that the performance of the portfolio based on investor sentiment indicator performs significantly better than the benchmark. Further more, compared with the basic Fama-French three factor model, the goodness of fit and significance of the asset pricing model with investor sentiment factor added are both improved, indicating that the investor sentiment index we constructed can capture the investors' sentiment in the market well, and has a good explanatory power for stock return.

基于文本挖掘和LSTM的投资者情绪与股票收益率的实证研究

任天语

南京大学工程管理学院，南京，江苏，中国

704226242@qq.com

关键词: 投资者情绪; 自然语言处理; 文本情感分析; 深度学习; LSTM; 股票收益率; 拓展的资产定价模型

中文摘要. 本文基于互联网社交平台和大数据技术的发展，采用从股吧论坛挖掘的非结构化的文本型投资者评论数据，使用长短期记忆神经网络进行文本情感分析，从而构建更精确的投资者情绪指标。并基于该指标对创业板综指的成分股进行实证研究来探讨投资者情绪对股票收益率的影响，经过全样本选股测试发现基于投资者情绪指标构建的投资组合的表现明显优于基准。进一步，引入投资者情绪因子的资产定价模型的拟合优度和显著性较基础的Fama-French三因子模型均有所改善，说明本文构建的投资者情绪指标比较好的捕捉了市场投资者情绪，并且对股票收益率具有较好的解释力。

1. 引言

行为金融学认为投资者的行为决策在一定的投资环境下并非独立，由于投资者之间的模仿学习、情绪传染等，极易形成羊群效应，导致股价异常波动。

有关投资者情绪度量的方法一般有调查问卷方式，如美国个体投资者协会指数(Fisher and Statman, 2009),央视看盘指数(Sun Jianjun, 2004)等；基于市场数据或互联网数值型数据的方式，如IPO首日溢价及数量(Brown and Cliff, 2005)，股票型基金重仓持股现金流(Zhuang Xintian, 2011)，百度搜索指数、财经论坛发帖量、浏览量、回复量(Antweiler and Frank, 2004; Tetlock, 2007)以及使用主成分分析(Brown and Cliff, 2004),因子分析(Zhang Zhongyuan, 2016)方法构建的综合性指标。鲜少有文献使用非结构化文本数据进行度量，同时在自然语言处理与文本情感分析领域，以往通常使用情感词典直接计算情感关键词得分并加权处理得到投资者的情感倾向性，无法考虑文本语法结构、上下文关联及非线性因素，而长短期记忆神经网络可以对本文建模，其特殊的门控结构能优化记忆并很好地捕捉文本的长期时间关联，解决梯度爆炸或消失的问题。

本文基于大数据和社交网络的背景，挖掘投资者评论等非结构化数据，使用深度学习分析文本信息中的投资者情绪，并通过实证分析探讨投资者情绪对股票收益率的影响。由此有助于投资者作出理性的行为决策，政府完善相关舆论监督机制，促进证券市场更稳健地运作。

2．研究方法

2.1 文本情感分类及投资者情绪指标构建

文本情感分类整体流程如图1。其中中文分词采用jieba分词，词向量训练采用Word2Vec模型，并通过剔除词频小于10的词语进行降维。深度神经网络分类器为LSTM。

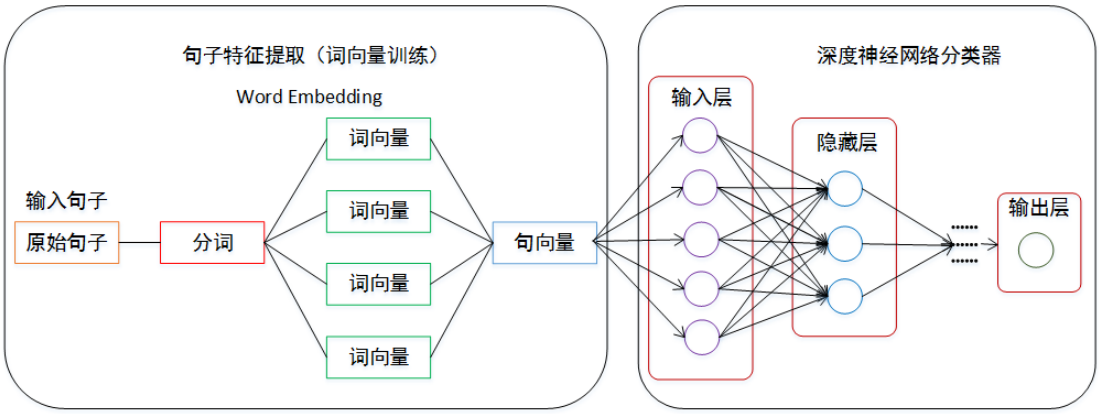

图 1 文本情感分类流程框架

通过训练好的LSTM模型预测每条评论的情感得分为：1(积极)、-1(消极)、0(中性)，构建投资者情绪指标$SENT_{it}$为每个交易日个股所有评论的情感得分均值，对于非交易日，将其情感信息纳入最近的一个交易日取均值。

2.2 单因子测试

为检测因子的单调性，按因子大小将股票池分为5组构成投资组合，每个投资组合内的股票等权配置。做多因子值最大的一组做空因子值最小的一组称作long-short对冲组合。

基于情绪因子的单调性，根据因子值选择股票构建投资组合，以创业板指数为基准，在2011年1月至2017年12月整个时间区间进行全样本选股回测。

2.3 引入情绪因子的资产定价模型

以Fama-French三因素模型作为基准模型并在中国创业板市场进行实证检验，进一步将本文构建的投资者情绪因子引入三因素模型中，拓展后的资产定价模型如下：

$$R_{it} - R_{ft} = \alpha_i + \beta_{i1}(R_{mt} - R_{ft}) + \beta_{i2}SMB_t + \beta_{i3}HML_t + \beta_{i4}SENT_{it} + \varepsilon_{it} \quad (1)$$

$$R_{it} - R_{ft} = \alpha_i + \beta_{i1}(R_{mt} - R_{ft}) + \beta_{i2}SMB_t + \beta_{i3}HML_t + \beta_{i4}SENT_{it-1} + \varepsilon_{it} \quad (2)$$

其中：$R_{it}-R_{ft}$表示股票投资组合的超额收益率，$R_{mt}-R_{ft}$表示市场投资组合的超额收益率，SMB_t表示规模因子，HML_t表示账面市值比因子，$SENT_{it}$表示投资者情绪因子。

根据规模因子将样本划分为大规模(B)与小规模(S)两组，根据账面市值比因子将样本划分为高(H)、低(L)、中(M)三组，交叉可形成6种组合：S/H、S/M、S/L、B/H、B/M、B/L，对各组合的平均收益率分别建模。

3．实证分析

本文选取创业板综指的成分股作为投资者情绪对股票收益影响的研究标的，一共挖掘了719只成分股自2011年1月至2017年12月共40730019条评论数据。

3.1 使用深度学习模型进行文本情感分类

使用基于TensorFlow的Keras来搭建LSTM模型，通过1000多条语料(其中80%为训练集，20%为验证集)的学习，模型最终在验证集上达到了90.18%的准确率，训练过程如图2。其中train和validation分别表示在训练集和验证集上的准确率。可以看出：训练集和验证集上的准确率差距总体变小，验证集上准确率逐渐提高，说明模型没有过拟合。

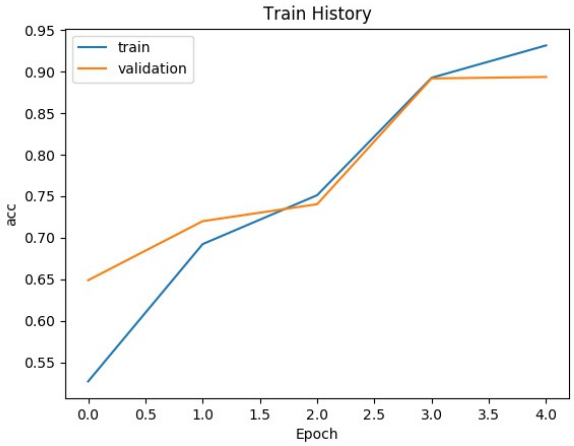

图 2 LSTM 训练过程图

3.2 单因子测试

图3左为分组测试的年化收益率图，可以看出投资者情绪指标对股票具备很好的区分效果，与股票收益率呈正相关。进一步，选择因子值最大的前100只股票构建投资组合进行选股回测，图3右为回测净值图，回测指标见表1。从净值曲线及各项指标均可以看出选股组合表现明显优于基准，说明投资者情绪指标能够在贯穿牛熊市期间表现出良好的预测及选股能力。

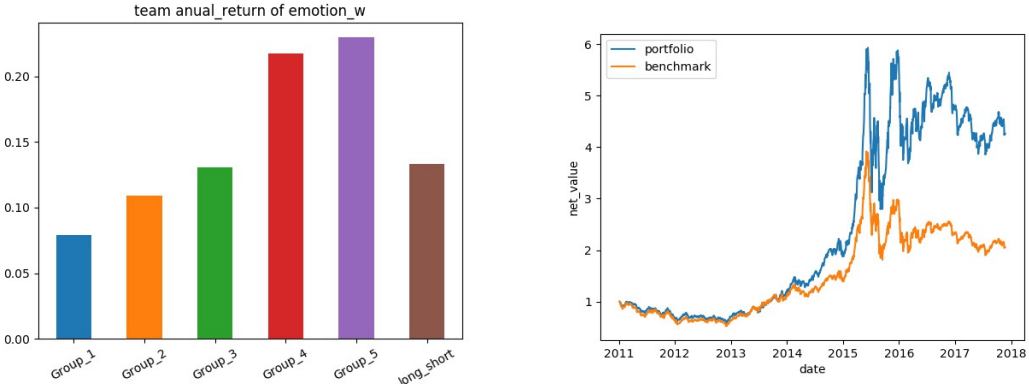

图3 单因子测试结果(其中左图为因子分层测试图，右图为选股回测净值图)

表1　选股回测指标结果

	年化收益率	累计收益率	年化波动率	夏普比率	最大回撤
投资组合	0.2441	3.2629	0.3570	0.7920	0.5292
基准	0.1146	1.0544	0.3217	0.4992	0.5371

3.3　基于投资者情绪因子的多因子模型分析

首先对Fama-French三因子模型在创业板上进行实证检验，以月度(20个交易日)为基准，结果如表2。可以看出，三因子模型在中国创业板股票上具备较好的解释力，可以作为本文研究的基准模型。但并非所有组合的截距项都具备显著性，模型可能存在一定的改进空间。

表2　Fama-French资产定价模型回归结果

模型	α_i	β_{i1}	β_{i2}	β_{i3}	R-squared
S/H	-0.0049(**)	1.0558(***)	0.4461(*)	0.4107(***)	0.984
S/M	0.0009	1.0837(**)	0.3517(***)	0.2227(*)	0.980
S/L	0.0022(*)	0.7962(***)	0.7788(***)	-0.7536(**)	0.968
B/H	0.0033(**)	0.8336(***)	-0.4699(***)	0.3031(***)	0.827
B/M	-0.0013	1.0089(***)	-0.1509(**)	0.1094(***)	0.963
B/L	-0.0038(**)	1.0931(***)	-0.8026(***)	-0.5326(**)	0.986

检验引入情绪因子后的资产定价模型，模型(1)结果如表3。整体来看，各组合的拟合优度和显著性均有所提升，说明引入本文构建的情绪因子后，拓展的资产定价模型具备更强的解释力。模型(2)结果如表4，滞后一期的情绪因子对未来一期的收益率均具备显著影响，投资者情绪对股票收益率具备一定的预测能力。

表3　引入投资者情绪的资产定价模型(1)回归结果

模型	α_i	β_{i1}	β_{i2}	β_{i3}	β_{i4}	R-squared
S/H	-0.0038(***)	1.0559(***)	0.4482(***)	0.4073(***)	0.0097(***)	0.984
S/M	0.0022(***)	1.0852(***)	0.3531(***)	0.2186(***)	0.0146(***)	0.981
S/L	0.0008(**)	0.7896(***)	0.7871(***)	-0.7498(***)	-0.0229(***)	0.969
B/H	0.0015(*)	0.8327(***)	-0.4785(***)	0.3172(***)	-0.0322(***)	0.836
B/M	0.0005(**)	1.0094(***)	-0.1464(***)	0.0991(***)	0.0257(***)	0.965
B/L	-0.0036(***)	1.0935(***)	-0.8028(***)	-0.5337(***)	0.0035	0.986

表4　引入投资者情绪的资产定价模型(2)回归结果

模型	α_i	β_{i1}	β_{i2}	β_{i3}	β_{i4}	R-squared
S/H	-0.0038(***)	1.0558(***)	0.4480(***)	0.4075(***)	0.0097(***)	0.984
S/M	0.0023(***)	1.0851(***)	0.3529(***)	0.2188(***)	0.0147(***)	0.981
S/L	0.0007(*)	0.7894(***)	0.7889(***)	-0.7500(***)	-0.0233(***)	0.969
B/H	0.0014(*)	0.8328(***)	-0.4770(***)	0.3170(***)	-0.0323(***)	0.837
B/M	0.0005(*)	1.0093(***)	-0.1465(***)	0.0994(***)	0.0253(***)	0.965
B/L	-0.0034(***)	1.0941(***)	-0.8058(***)	-0.5336(***)	0.0042(*)	0.986

4．结束语

投资者情绪是行为金融学备受关注的研究领域之一，如何更加充分利用有效信息、准确度量投资者情绪对研究结果的准确性至关重要。以往度量投资者情绪多以客观指标为主，一些主观指标也往往缺乏代表性和时效性，本文基于互联网大数据的优势，使用文本挖掘和深度学习方法进行文本情感分析，从而构建更直观准确的投资者情绪指标。

通过对创业板成分股进行实证分析发现，投资者情绪与股票收益率具备正相关关系，基于此构建的选股策略的表现明显优于基准。同时，以Fama-French三因子模型为基准，引入情绪因子后拓展的资产定价模型具备更强的解释力，说明本文构建的投资者情绪指标较好的捕捉了市场投资者情绪，对股票收益率具有显著影响和一定的预测能力。

基于此，本文建议可加强网络使用规范并完善监督机制，防止庄托利用网络社交平台有目的性地散播消息或因舆论传播引发羊群效应，避免资产价格受到舆情冲击而发生异常波动。

References

[1] E. F. Fama and K. R. French, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics*, vol.33, pp. 3-56, 1993.

[2] W. Antweiler and M. Z. Frank, Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards, *Journal of Finance*, vol.59, pp. 1259-1294, 2004.

[3] G. W. Brown and M. T. Cliff, Investor sentiment and the near-term stock market, *Journal of Empirical Finance*, vol.11, pp. 1-27, 2004.

[4] G. W. Brown and M. T. Cliff, Investor Sentiment and Asset Valuation, *Journal of Business*, vol.78, pp. 405-440, 2005.

[5] P. C. Tetlock, Giving Content to Investor Sentiment: The Role of Media in the Stock Market, *The Journal of Finance*, vol.62, pp. 30, 2007.

[6] K. L. Fisher and M. Statman, Consumer Confidence and Stock Returns, *Ssrn Electronic Journal*, vol.30, pp. 225-236, 2009.

[7] Sun Jianjun and Wang Meijin, China's stock market returns, income fluctuations and investor sentiment, *Economic Research*, vol.10, pp. 75-83, 2004.

[8] Zhuang Xintian and Chi Lixu, The Impact of Investor Emotions on Stock Returns in China——Based on Panel Data Research, *Management Review*, vol.23, pp. 41-48, 2011.

[9] Zhang Zhongyuan and Ma Qiang, Bank transfer, stock market income and investor sentiment, *Journal of Yunnan University of Finance and Economics*, vol.32, 2016.