研究提案

中国股市的行业关注度指数：基于深度学习方法的估算

韩冰

2024 年 1 月 4 日

摘要

对于中国股市，我们定义了 SAI（行业关注度指数），根据某一特定行业最具代表性的股票的在线帖子数量来量化散户投资者对该行业的关注程度。我们定义了异常 SAI 以区分额外的关注度，并定义了情绪 SAI 用于情绪分析。帖子的文本数据来自中国大陆最活跃的在线股票论坛东方财富网（又称股吧）。我们重建了一个新的情绪词典，并设计了一个深度学习模型来对股票帖子的情绪倾向进行分类。我们进行了一系列回归分析，以检验 SAIs 的预测能力以及它们与股票收益和交易量的相关性。

关键词：行业投资者关注度、情绪分析、深度学习、中国股市、自然语言处理（NLP）

介绍

股市并非铁板一块。它被方便地划分为不同的板块，将从事不同业务类型的公司归为一类。像标准普尔 500 指数（SPX）这样的主要股票指数，能为整个市场提供一个宏观视角，但追踪股市板块——比如能源、医疗保健和科技板块——能帮助投资者清晰地了解特定行业随时间推移的市场表现。

在中国，投资者在股票论坛上发帖的频率很好地反映了对各行业的关注程度。基于从股票论坛获取的文本分析结果，我们可以构建一系列指标来衡量投资者对不同行业板块的关注度，并进一步研究这些指标与市场的关系。目前尚无相关研究对行业关注度进行衡量。

已发表的文献表明，投资者关注度对股票价格具有预测作用，并且与股票收益和交易量相关。我们将尝试从行业视角来检验这些结论在中国股市是否同样成立。

另一方面，人工智能技术和替代数据在金融分析中发挥着越来越重要的作用。一段时间以来，研究人员专注于如何利用大数据研究投资者的关注度和情绪，这些信息包含在像推特这样的社交平台或金融网站上的机器可读新闻中。然而，基于股票论坛数据的研究相对有限，尤其是在中国股市。我们的研究将探索这一领域，并使用更先进的自然语言处理和人工智能方法来处理非结构化文本数据。

让我们关注一则最新消息。几天前，2023 年神经信息处理系统大会（NeuroIPS）公布了获奖论文，其中十年奖颁给了十年前的 NeuroIPS 论文《词和短语的分布式表示及其组合》。该论文介绍了开创性的词嵌入技术 word2vec，展示了从大量无结构文本中学习的能力，并推动了自然语言处理（NLP）新时代的到来。

实际上，诸如 word2vec 这类极具潜力的先进自然语言处理技术在金融领域尚未得到广泛应用。在我的研究中，我将介绍如何利用深度神经网络模型优化文本处理结果。我们采用了一些先进技术，比如基于似然比检验的短语检测和基于词嵌入的神经网络。这样做的目的是获取更丰富的词典，并获得更准确的情感分类结果。

文献综述

安特韦勒和弗兰克（2004 年）研究了雅虎财经在线论坛上的帖子，发现以帖子数量衡量的关注度指标能够有效预测股票收益和市场波动。同期帖子情绪的分歧与股票交易量呈正相关。

Zhi 等人（2011 年）获取了个股的每周谷歌搜索指数，并直接通过股票的搜索频率来衡量投资者的关注度。研究表明，使用搜索指数能够更及时地衡量投资者的关注度，而且搜索指数的上升能够预测未来两周内股价的上涨以及一年内的股价反转。

正如张等人（2016 年）所发现的那样，推特上的情绪对道琼斯工业平均指数具有一定的预测作用。当推特上的情绪表达强烈，比如出现大量诸如希望和担忧之类的情绪因素时，道琼斯工业平均指数次日将会下跌。

斯普伦格等人（2014 年）研究了推特上专门讨论股市的论坛，提取关键词对个股和大型公司问题进行深入研究。结果表明，推特文章的情绪与股票收益和交易量之间存在关联。

李等人（2018 年）研究了股票微博消息（即推文）与金融市场指标的相关程度以及信息有效聚合的机制。他们收集了超过 120 万条与标准普尔 100 指数成分股公司相关的消息，并对这些数据进行了日度和 15 分钟的分析。他们发现，消息的情绪与日度异常股票收益呈正相关，且消息量能够预测 15 分钟后的收益、成交量和波动率。

陈 C P 等人（2018 年）首次使用卷积神经网络（CNN）来计算中国散户投资者的情绪，并比较了 CNN 和支持向量机（SVM）模型的预测性能。他们的研究发现，CNN 的预测准确率大致与 SVM 相当，但 CNN 模型在分类上更具决定性。

陈 Z 等人（2023 年）利用罗宾汉投资者数据和谷歌搜索量指数来衡量活跃的散户投资者的关注度，发现活跃的散户投资者的关注度会受到近期股票收益的影响，并且对大盘股的影响更为显著。

陈 S 等人（2020 年）提出了一种新的代理变量，用于衡量中国 A 股市场在线股票论坛上带有正负中性态度的帖子所体现的零售商的不对称关注。结果表明，该不对称关注代理变量与波动率不对称性显著正相关。此外，我们发现负面信息的出现会引发更高的波动率。

不对称性以及作为中介的不对称关注将更多的负面信息流引入市场，从而引发高不对称波动率。此外，该代理变量与特定的财务杠杆是相互独立的，但其对不对称波动率的影响会随着市场系统性风险的增加而增强。

迪克森（2022 年）所著的这本书涵盖了利用另类数据进行投资的基础知识，并阐述了对另类数据进行预处理和建模的最佳实践。

克拉斯（2019 年）所著的这本书探讨了机器学习领域的进展以及如何将其应用于金融行业。书中对机器学习的工作原理进行了清晰的解释和专业的讨论，重点在于金融应用，并涵盖了包括神经网络、生成对抗网络（GANs）和强化学习在内的先进机器学习方法。

Mikolov 等人（2013 年）提出了连续词袋模型（Continuous Skip-gram 模型），这是一种高效的方法，能够学习到高质量的分布式向量表示，从而捕捉大量精确的句法和语义词关系。

赖等人（2022 年）研究了谷歌搜索量指数（GSVI）——投资者关注度的一个代理指标——能否预测台湾证券交易所 50 指数成分股的超额收益和异常交易量。该研究还基于股价的正向或负向冲击，探讨了 GSVI 的潜在动机。

布克尔等人（2018 年）利用面向投资者的 StockTwits 社交媒体网络上的帖子，对收益公告前后投资者意见分歧、信息披露处理成本以及交易量之间的关系进行了新的研究。他们发现，收益公告前的意见分歧以及收益公告期间意见分歧的增加都与交易量呈正相关。

陈 J 等人（2022 年）基于文献中的代理变量提出了一项投资者关注度指数，发现该指数能显著预测股票市场的风险溢价，而每个单独的代理变量预测能力较弱。这种预测能力主要源于暂时性价格压力的反转以及对高波动性股票更强的预测能力。

洛赫兰和麦克唐纳（2011 年）开发了一种替代的负面词汇表，以及另外五种词汇表，这些词汇表能更好地反映财务文本的语气。他们将这些词汇表与 10-K 报告、交易量、收益波动性、欺诈、重大缺陷和意外收益联系起来。

法玛和麦克贝斯（1973 年）对纽约证券交易所普通股的平均收益与风险之间的关系进行了检验。这些检验的理论基础是“双参数”投资组合模型以及从该模型推导出的市场均衡模型。

问题

我们如何衡量中国股市各板块（尤其是散户投资者）的关注度?

对于中国股市，我们定义了 SAI（行业关注度指数），根据某一特定行业最具代表性的股票的网络帖子数量来量化投资者对该行业的关注程度。我们关注如何获取异常的 SAI 值，以区分哪些行业突然在市场上获得了额外的关注。我们还关注 SAI 中的情绪特征。

活跃的股票市场论坛中有哪些可用的文本数据可以支持我们的研究?

☒ 零售投资者在在线股票论坛上的发帖很好地反映了对特定板块的关注度。我们从东方财富网（又称股吧）收集文本数据，这是中国大陆最大、最活跃的股票交流平台。

在文本数据分析中，我们可以运用哪些先进的人工智能方法?

我们采用词性标注方法重建了一个新的情感词典，并设计了一个深度学习模型来对股票帖子的情感倾向进行分类。

我们所定义的交易数据（例如股票收益和交易量）与投资者关注度指标之间存在怎样的关联?

我们将进行一系列回归分析，以检验行业关注度指数的预测能力及其与股票收益和交易量的相关性。

方法

关键变量定义

1. 行业关注度指数（SAI）

我们将定义一个指标来衡量投资者对不同行业指数的关注度，并将其称为行业关注度指数（SAI）。对于行业 *i*

$$SAI_i = \sum_{k=1}^{N} V_{i,k}$$

其中 N 为能最大程度代表该板块的所选股票数量，而 $V_{i,k}$ 为股票 k 在单个时间周期（例如最近 30 天）内的成交量。

2. 异常 SAI

不同板块之间的投资者关注度无法横向比较，因为板块内帖子数量与板块规模之间存在一定的正相关关系。一些成分股较少、总市值较低的板块，其帖子数量肯定不会像规模较大的板块那么多，所以我们不能用绝对值来进行比较。一个可行的方法是使用异常关注度指数（SAI）来区分哪些板块突然在市场上受到了额外的关注。

我们将异常的 SAI 定义为

$$abnormalSAI_t \boxtimes \log(SAI_t) - \log[\text{Med}(SAI_{t\boxtimes 1}, ..., SAI_{t\boxtimes r})]$$

其中，log（SAIt） 是第 *t 周 SAI 的对数*，而 log[Med(SAIt−1,...,*SAIt−r*)] *是*前 r 周 SAI 中位数的对数。

Zhi 等人（2011 年）在其论文中证实了这种定义方式是有效的。直观来看，较长时间窗口内的中位数能够以一种不受近期波动影响的方式捕捉到“正常”的关注度水平。这种方法还有一个优点，即消除了时间趋势和其他低频季节性因素。较大的正异常 SAI 明显代表了投资者关注度的激增，并且可以在横截面上进行不同股票之间的比较。

3. 情绪 SAI

我们将创建一个分类任务，将投资者的帖子分为两类。每篇帖子被标记为积极情绪（标记为 1）或消极情绪（标记为 0）。根据分类结果，我们可以按如下方式计算情绪 SAI：

$$sentimentSAI_i = \frac{1}{N}\sum_{k=1}^{N}\frac{V_{i,k}^{positive}}{V_{i,k}}$$

其中 $V_{i,k\ positive}$ 表示股票 *k*,行业 *i* *的积极情绪帖子数量*。

数据

1. 文本数据

该帖子的文本数据来自在线股票论坛东方财富网（又称股吧）。这是中国大陆最大、最活跃的股票交流平台，散户投资者在此发帖和评论。这些帖子和评论通常都很简短，平均不到 20 个字。

需要收集的帖子包括主要行业板块中***最***具***代表性***的***股票***。基于这些股票所收集的文本将作为该板块的语料库。之所以这样做，是因为在行业指数主题下投资者的讨论并不活跃，而在代表性股票下则非常活跃。同时，与指数论坛相比，个股论坛上的帖子更能表达出强烈的买卖偏好。因此，这非常适合用于投资者关注度分析。我们将收集一定时期的数据（通常不少于一年）。

2. 市场交易数据

我们收集的市场数据包括每日开盘价和收盘价等主要指标，以及成交量和成交额。数据收集对象包括行业指数（SI）以及每个 SI 的代表性股票。我们将使用来自中国金融数据服务提供商 RoyalFlush 网络技术的 SI 类别和每日数据（表 1）。

表 1.由皇家顺网络技术公司发布的行业指数数据演示

部门	日期	开放的	高的	低	关闭	体积	金额
881101	20201013	3299.52	3305.41	3261.98	3283.49	382250000	3850680000
881102	20201013	2796.46	2850.65	2775.58	2846.61	301790000	6608240000
881103	20201013	3965.01	3996.09	3935.66	3989.45	651560000	10429000000
881104	20201013	10602.2	10906.8	10588.9	10835.1	140700000	2852140000

最具代表性的股票

尽管单只股票的帖子数量并不总是很多，但一组成分股能够有效地弥补这一不足，并且有资格代表整个行业。截至目前，皇家顺子已发布超过 80 个行业及板块指数。大多数板块的成分股数量超过 30 只，且这些股票通常权重较高，投资者集中。

我们不会从所有板块收集数据，而是从中选取一部分。每个板块所选股票的数量 N 也将根据数据进行平衡。

时间区间。更重要的是，我们将选取与该行业指数历史走势关联度最高的成分股（例如相关系数值最高），但不能依据其权重来选择，以避免个别股票的异常波动产生影响。

网站提交数据

我们将开发一个网络爬虫程序，用于收集东方财富股吧中的散户投资者帖子，并将其插入数据库。对于行业和成分股信息，我们还需要执行一个单独的程序来收集行业列表和成分股列表。将这两组信息合并后，我们得到的原始数据集如表 2 所示。第一列“代码”代表股票代码。

表 2.东方财富网投资者帖子的原始数据集演示

代码	部门	时间	内容	作者
600999	安全	2020 年 1 月 13 日 13:25	招商银行都撑不住了，大盘真的不行。	股友 2Ev***
002230	计算机应用	2020 年 2 月 13 日 13 点 42 分：**	三天内会创新高的	股友 pnb***

文本数据预处理

我们需要删除不符合要求的帖子，包括那些疑似广告的帖子以及过短的帖子。我们采用 Jieba 分词器进行分词。Jieba 分词器是一款高效的中文分词工具，支持多种分词模式。

我们将尝试使用一种基于似然比检验（一种统计方法***)的文本短语检测工具（附录 1***），该工具无需手动设置语法规则即可检测出散户投资者评论中的常见短语，并提高分词的准确性。这样做的目的是扩大我们的分词结果，使文本分析更具稳健性。

为了适应 AI 模型的嵌入层，我们将所有序列通过截断或在末尾添加零的方式调整为等长。我们将所有帖子的长度固定为 40 个字符。这一步可以通过 Python 编写的开源人工神经网络库 Keras 提供的方法来实现。

情感 SAI 分类任务

1. 一个新的、有针对性的情感词典

大多数学者选择现有的通用情感词典和词库作为参考来进行研究，这导致其对特定语境缺乏针对性（例如，股市评论中专业术语较多，社交媒体中则充斥着大量俚语和表情符号）。应当专门构建情感词典。

对于股票市场的帖子或评论，我们可以去除一般性的情感词汇，并使用词性标注（POS）方法尽可能多地保留人名、动作和形容词。通过统计词频，我们可以找出新的替代情感词汇，并通过亚马逊机械土耳其人（Mturk）对其进行评分。当然，这项工作也可以手动完成。

2. 深度学习模型

本研究构建了一个 BiLSTM-CNN-Attention 情感分析模型，用于对股票帖子的情感倾向进行分类。该模型的结构可能包括词嵌入层、BiLSTM 层、CNN 层、注意力层和输出层。最终版本将取决于研究过程中的具体情况。有关模型的更多详细信息，请参阅***附录 2***。此外，还需要适合中文自然语言处理的预训练资源。

回归方法

法玛和麦克贝斯（1973 年）采用了一种两步回归法，这是一种衡量这些风险因素对资产或投资组合收益解释准确性的实用方法。第一步是对每个阶段的解释变量和因变量进行最小二乘回归，以获得估计参数，并将每个参数视为总体参数的一个样本值；第二步是计算第一步中所有参数的平均值，从而得出总体数据的估计参数。我们将使用这种方法进行实证检验。

我们将采用 Loughran 和 McDonald（2011）的实证方法来选择解释变量，包括超额收益、异常交易量、收益波动率等。需要对回归模型中的变量进行描述性统计，包括样本量、均值、标准差、最小值和最大值。***SAIs 对交易量***、波动率***和意外收益***的***回归结果应显示出显著***的***相关性***。

使用样本外数据进行稳健性测试。通过此项测试，我们能够证明我们的方法在对非样本文本数据进行预处理和情感分类方面的有效性。所呈现的显著相关性也应与样本数据保持一致。

进一步讨论

1. 遗憾指数

我们发现，与个股相关的帖子和评论更多地反映了行为金融学中所描述的投资者心理偏差。例如，有很多样本包含与“后悔”相关的词汇，无论是后悔没买还是后悔没卖。我们还希望测试后悔情绪与股票收益之间是否存在显著的相关性，以及投资者关注度的增加在多大程度上转化为实际交易（买入或卖出）。

表 3.与“后悔”相关的投资者帖子示例

代码	部门	时间	内容	作者	对...的遗憾
600999	安全	2020 年 3 月 25 日 11 时 31 分：**	16.88 这个价格卖了真后悔啊 我后悔没在 16.88 元时卖掉。	股 友 JV z***	不出售
600999	安全	2020 年 7 月 6 日 10 点 25 分：**	周五在 3 点卖出，现在后悔了。	股 友 1r 2***	售罄（过早）

参考文献

[1] Antweiler, W. , & Frank, M. Z. . (2004). Is all that talk just noise? the information content of internet stock message boards. Journal of Finance, 59 (3), 1259-1294.

[2] Zhi, D. A. , Engelberg, J. , & Gao, P. . (2011). In search of attention. Journal of Finance, 66(5), 1461-1499.

[3] Zhang, Y. , Song, W. , Shen, D. , & Zhang, W. . (2016). Market reaction to internet news: information diffusion and price pressure. Economic Modelling, 56, 43-49.

[4] Sprenger, T. O. , Tumasjan, A. , Sandner, P. G. , & Welpe, I. M. . (2014). Tweets and trades: the information content of stock microblogs. European Financial Management, 20(5), 926–957.

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

[6] Chen, C. P., Tseng, T. H., & Yang, T. H. (2018, June). Sentiment Analysis on Social Network: Using Emoticon Characteristics for Twitter Polarity Classification. In International Journal of Computational Linguistics & Chinese Language Processing, Volume 23, Number 1, June 2018.

[7] Chen, Z., & Craig, K. A. (2023). Active attention, retail investor base, and stock returns. Journal of Behavioral and Experimental Finance, 100820.

[8] Chen, S., Zhang, W., Feng, X., & Xiong, X. (2020). Asymmetry of retail investors' attention and asymmetric volatility: Evidence from China. Finance Research Letters, 36, 101334.

[9] Matthew Dixon (2022) The Book of Alternative Data: A Guide for Investors, Traders and Risk Managers, Quantitative

Finance, 22:8, 1427-1428, DOI: 10.1080/14697688.2022.2078736

[10] Klaas, J. (2019). Machine learning for finance: principles and practice for financial insiders. Packt Publishing Ltd.

[11] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their

compositionality. Advances in neural information processing systems, 26.

[12] Lai, H. H., Chang, T. P., Hu, C. H., & Chou, P. C. (2022). Can google search volume index predict the returns and trading volumes of stocks in a retail investor dominant market. Cogent Economics & Finance, 10(1), 2014640.

[13] Booker, A. , Curtis, A. , & Richardson, V. J. . (2018). Investor disagreement, disclosure processing costs, and trading volume: evidence from investors who interact on social media. Social Science Electronic Publishing.

[14] Chen, J., Tang, G., Yao, J., & Zhou, G. (2022). Investor attention and stock returns. Journal of Financial and Quantitative Analysis, 57(2), 455-484.

[15] Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10‐Ks. The Journal of finance, 66(1), 35-65.

[16] Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. Journal of political economy, 81(3), 607-636.

附录 1

一种基于似然比检验的文本短语检测方法及装置 ***[未应用专利]***

介绍

本发明属于自然语言处理中的文本短语检测领域。它采用文本分词技术和似然比检验，尽可能避免检测过程中的手动干预，提高了短语检测的适用性和可靠性。

现有技术方案（通过查找最相关的专利）试图通过遍历所有可能的词语组合来寻找匹配项，从而预设适用于大多数短语的词性依赖规则。如果语料库仅限于特定领域且包含大量专业术语，那么这可能是最佳方法。然而，这种默认规则需要大量的人工干预，并且随着语料库的发展需要频繁更新，在实际应用中很难完全实现，而且它也不适用于推文、博客和新闻文章。此外，在计算词语组合的构词概率时，通常需要手动预设模型阈值，其大小也会影响最终短语检测的质量，因此可靠性不高。

方法论

分词的功能是根据规则提取文本中包含的所有词语，不同的分词工具具有不同的分词规则。目前，jieba 分词器是一种高效的中文分词工具，支持多种分词模式。本文采用基于 Python 实现的 jieba 分词器的精确模式。

似然比被定义为在受约束条件下似然函数的最大值与在无约束条件下似然函数的最大值之比。似然比检验的步骤是：对于给定的一对词，该方法在观测数据集上检验两个假设。第一个假设（零假设）表明词 1 的出现与词 2 无关。第二个假设（备择假设）表明看到词 1 会改变看到词 2 的可能性。如果我们接受备择假设，这意味着这两个词可以构成一个常见短语。哪个假设成立是通过计算语料库中实际观察到的词频来确定的。

总体处理流程是通过结巴分词器将输入文本分词，过滤停用词，组合 n 元语法，并计算似然比。根据似然比的计算结果对 n 元语法进行排序，选择似然比最小的 n 元语法作为最终的检测结果。

附录 2

双向长短期记忆网络 - 卷积神经网络 - 注意力模型

本研究构建了一个 BiLSTM-CNN-Attention 情感分析模型，用于提取股票帖子中的情感倾向。该模型的结构包括

1. 词嵌入层：文本的词向量由预训练的 word2vec 模型构建，分词后的句子被映射为低维稠密向量，每个词对应一个向量。生成的词向量包含语义信息，有利于在下一层进行进一步的特征提取。

2. 双向长短期记忆（BiLSTM）层：双向长短期记忆（BiLSTM）模型通过堆叠两层长短期记忆（LSTM）网络，突破了模型只能基于前一时刻的时间信息来预测下一时刻输出的局限性，能够更好地结合上下文进行输出。

3. 卷积神经网络层：卷积神经网络（CNN）在双向长短期记忆网络（BiLSTM）在每个时间步的输出上执行卷积操作，进一步提取文本特征，并在卷积后添加最大池化，以防止过拟合，减少参数和计算复杂度。

4. 注意力层：通过为注意力机制分配不同的概率权重，并将其与卷积神经网络（CNN）的输出向量相乘，然后通过 softmax 函数进行归一化计算，从而得到注意力机制的权重矩阵的值。

5. 输出层：输出层通过 S 型函数映射结果，以获得情感分类的结果。

可用的预训练资源

一些人工智能巨头（如腾讯和百度）公开披露的预训练模型或算法将有助于我们提高中文文本自然语言处理任务的效率。我们将选择其中一些，包括：1）能够完成诸如句子级情感分类和方面级情感分类等典型情感分析任务的预训练模型。2）为中文单词和短语提供有限维度向量表示（即嵌入）的嵌入语料库。我们将引入这些预训练模型来构建我们的词嵌入层和双向长短期记忆层，并对其进行微调以适应我们的股票分析场景。