研究计划

中国股票市场行业关注指数:基于深度学习方法的估计

宾汉

2024年1月4日

摘要

对于中国股市，我们定义了行业关注指数(Sector Attention Index, SAI)，根据该行业最具代表性的股票的网上发帖量来量化散户对该行业的关注程度。我们定义了异常SAI来区分额外注意力，定义了情感SAI来进行情感分析。帖子的文本数据取自中国大陆最活跃的在线股票论坛Eastmoney(又名Guba)。我们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中的情感倾向进行分类。我们进行了一系列回归分析，以测试SAIs的预测能力及其与股票收益和交易量的相关性。

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

介绍

股票市场并不是铁板一块。它被方便地分为不同的板块，根据不同的公司经营的业务类型对不同的公司进行分组。主要的股票指数，如标准普尔500指数(SPX)，提供了整个市场的大视图，但跟踪股票市场部门-如能源，医疗保健和技术-可以帮助投资者清楚地跟踪特定行业的市场表现。

在中国，投资者在股票论坛上发帖的频率很好地反映了该行业的关注度。基于从股票论坛获得的文本分析结果，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。目前尚无衡量行业关注度的相关研究。

已发表的文献表明，投资者关注对股价具有预测能力，并与股票收益和交易量相关。我们将尝试检验这些结论在行业视角方面是否也有效，特别是在中国股市中。

另一方面，人工智能技术和替代数据在财务分析中发挥着越来越重要的作用。一段时间以来，研究人员关注的是如何利用大数据来研究投资者的关注和情绪，这些关注和情绪包含在推特等社交平台或金融网站上的机读新闻中。然而，基于股票论坛数据的研究相对有限，尤其是针对中国股市的研究。我们的研究将探索这一领域，并使用更先进的基于NLP和人工智能的方法来处理非结构化文本数据。

让我们关注一条最新的新闻。几天前，NeuroIPS(神经信息处理系统会议和研讨会)2023发布了获奖论文，时间测试奖授予了十年前的NeuroIPS论文“单词和短语及其组成的分布式表示”。这项工作引入了突破性的词嵌入技术word2vec，展示了从大量非结构化文本中学习的能力，并推动了自然语言处理(NLP)新时代的到来。

现实中，word2vec等极具潜力的先进NLP技术尚未在金融领域得到广泛应用。在我的研究中，我将介绍如何使用深度神经网络模型来优化文本处理结果。我们已经使用了一些先进的技术，比如基于似然比测试的短语检测和基于词嵌入的神经网络。这样做的目的是为了捕获更丰富的词典，并获得更准确的情感分类结果。

文献综述

Antweiler和Frank(2004)研究了雅虎财经在线论坛上的帖子，发现通过帖子数量衡量的关注指标可以有效预测股票收益和市场波动。帖子情绪分化与同期股票交易量呈正相关。

Zhi et al.(2011)获得个股每周谷歌搜索指数，利用股票搜索频率直接衡量投资者关注度。研究表明，使用搜索指数可以更及时地衡量投资者关注度，搜索指数的增加可以预测未来两周的股价上涨和一年内的股价反转。

Zhang et al.(2016)发现，Twitter上的情绪对道琼斯工业平均指数有一定的预测作用。当Twitter上的情绪表达强烈，比如表现出大量的希望、担忧等情绪因素时，道琼斯工业平均指数就会在第二天下跌。

Sprenger et al.(2014)研究了Twitter上专门讨论股票市场的论坛，提取关键词对个股和公司重大问题进行深入研究。结果显示，Twitter文章的情绪与股票收益和交易量之间存在相关性。

Li et al.(2018)考察了股票微博消息(即推文)与金融市场指标的关联程度，以及导致信息高效聚合的机制。他们收集了120多万条与标普100指数成分股公司相关的微博信息，并以每日和15分钟为基础分析了这些数据。他们发现，消息的情绪与每日异常股票收益呈正相关，消息量可以预测15分钟的后续收益、交易量和波动率。

Chen C P et al.(2018)首先使用CNN计算中国散户投资者的情绪，并比较了CNN和SVM模型的预测性能。他们的研究发现，CNN的预测精度与SVM大致相当，但CNN模型在分类上更具决定性。

Chen Z et al.(2023)利用罗宾汉投资者数据和谷歌搜索量指数(Google Search Volume Index)来衡量活跃的散户投资者关注度，发现活跃的散户投资者关注度受到近期股票收益的影响，并且对于较大的股票更为显著。

Chen S et al.(2020)提出了一种新的代理方法来衡量中国a股市场在线股票留言板上发表的带有签名(积极、消极和中性态度)帖子的零售商的不对称关注。结果表明，不对称注意力的代理显著且与波动不对称正相关。此外，我们发现负面信息的到来会导致更高的波动率

不对称和不对称注意力作为中介，将更多的负面信息流入市场，从而引发高不对称波动。此外，这种代理是来自特殊金融杠杆的自变量，但其对不对称波动的影响随着市场系统风险的增加而增加。

Dixon(2022)的书涵盖了使用替代数据进行投资的基础知识，并列出了使用替代数据进行预处理和建模的最佳实践。

Klaas(2019)的书探讨了机器学习的进展，以及如何将它们应用于金融行业。它对机器学习的工作原理进行了清晰的解释和专家讨论，重点是金融应用，并涵盖了包括神经网络、gan和强化学习在内的高级机器学习方法。

Mikolov等人(2013)引入了一个连续的Skip-gram模型，这是一种学习高质量分布式向量表示的有效方法，可以捕获大量精确的句法和语义词关系。

Lai et al.(2022)检验了谷歌搜索量指数(Google search volume index, GSVI)能否预测TPEx 50指数成分股的超额收益和异常交易量。本文还探讨了基于股价正面或负面冲击的GSVI背后的动机。

Booker等人(2018)利用以投资者为中心的StockTwits社交媒体网络上的帖子，就投资者分歧、披露处理成本和收益公告周围的交易量产生了新的见解。他们发现，公告前的分歧和围绕盈利公告分歧的增加都与交易量呈正相关。

Chen J et al.(2022)在文献中提出了一个基于代理的投资者关注指数，并发现它可以显著地预测股市风险溢价，而每个代理单独的预测能力都很小。预测能力主要来源于对暂时性价格压力的逆转，以及对高方差股票较强的预测能力。

Loughran和McDonald(2011)开发了另一种负面词表，以及其他五个词表，可以更好地反映金融文本的语气。他们将单词列表与10-K申报报表、交易量、回报波动性、欺诈、实质性弱点和意外收益联系起来。

Fama和MacBeth(1973)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。检验的理论基础是“双参数”投资组合模型和由双参数投资组合模型衍生出来的市场均衡模型。

问题

我们如何衡量投资者(尤其是散户)对中国股市板块的关注程度?

对于中国股票市场，我们定义了SAI(行业关注指数)，根据该行业最具代表性的股票的在线帖子量来量化投资者对特定行业的关注。我们关心的是如何得到一个异常的SAI来区分哪些行业在市场上突然受到了额外的关注。我们还关注SAI中的情绪特征。

在活跃的股市论坛中有哪些文本数据可以支持我们的研究?

<s:1>网上股票论坛的散户帖子很好地反映了对特定行业的关注程度。我们收集的文字数据来自中国大陆最大、最活跃的股票交流平台东钱(又名古吧)。

在分析文字数据时，我们可以使用哪些先进的AI方法?

<s:1>我们使用词性标注方法重建一个新的情感词典，并设计一个深度学习模型，对股票帖子中的情感倾向进行分类。

交易数据(例如股票收益和交易金额)与我们定义的投资者关注指标之间的相关性是什么?

<s:1>我们将进行一系列回归分析，检验行业关注指数的预测能力及其与股票收益、交易金额的相关性。

方法

关键变量定义

1. 行业关注指数(SAI)

我们将定义一个衡量投资者对不同行业指数关注程度的指标，我们称之为行业关注指数(sector attention Index, SAI)。对于行业*i*

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

其中*N*是可以最大程度代表该行业的选定股票的数量，$V_{i,k}$是股票*k*在单个时间周期内(例如最近30天)的post volume。

2. SAI异常

不同行业之间的投资者关注度不能横向比较，因为岗位数量与行业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，肯定不会有较大板块那么多的帖子，所以我们不能用绝对值来比较。一种可行的方法是使用异常SAI来区分哪些行业在市场中突然受到额外的关注。

我们将异常SAI定义为

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

其中log (SAIt)为第*t周SAI的对数*，log [Med(SAIt−1，*⋯]*， *SAI−r)]*为前*r*周SAI中位数的对数。

Zhi et al.(2011)在他们的论文中证实了这种定义方式是有效的。直观地说，较长时间窗口内的中位数以一种对最近的跳跃具有鲁棒性的方式捕获了“正常”的注意力水平。它还具有去除时间趋势和其他低频季节性因素的优势。一个较大的正异常SAI清楚地表示投资者关注的激增，并且可以在横截面上对股票进行比较。

3. 情绪SAI

我们将创建一个分类任务，将投资者帖子分为两类。每个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，我们可以计算出情绪SAI如下

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

其中$V_{i,k}^{positive}$为股票*k*,行业*i*的*正面情绪帖子量*。

数据

1. 文本数据

帖子的文字数据取自在线股票论坛Eastmoney(又名Guba)。这是中国大陆最大、最活跃的股票交流平台，散户在这里发帖、评论。这些帖子和评论往往很短，平均不到20个字。

要收集的帖子包括主要行业板块中***最具代表性***的***个股***。基于这些股票收集的文本被用作该行业的语料库。这样做是因为投资者讨论在行业指数主题下并不活跃，而在代表性股票下非常活跃。同时，与指数论坛相比，个股论坛的帖子表达了强烈的买入和卖出偏好。因此，非常适合做投资者关注分析。我们会收集一定时期(通常不少于一年)的数据。

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，以及交易量和金额。收集目标包括行业指数(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 Network Technology的SI类别和日常数据(表1)。

表1。行业指数数据演示，由RoyalFlush Network Technology发布

Sector	Date	Open	High	Low	Close	Volume	Amount
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

最具代表性的个股

虽然单一股票的职位数量并不总是相当多，但一组成分股可以有效地弥补这一劣势，并有资格代表整个部门。截至目前，RoyalFlush已经发布了80多个行业&板块指数。大多数行业都有超过30只成分股，它们往往是权重高、投资者集中的股票。

我们不会收集所有板块的数据，而是从中选择一部分。每个板块的选股数量N也会根据数据进行平衡

时期。更重要的是，我们将衡量与行业指数历史最相关的成分股(例如具有最高相关系数值)，但不能根据权重选择它们，以避免个股异常波动的影响。

发布数据来自网站

我们将开发一个网络爬虫程序，收集散户投资者的帖子在东方货币的股票论坛，并将其插入数据库。对于板块和成分股信息，我们还需要执行一个单独的程序来收集板块列表和成分股列表。将这两组信息合并后，我们得到了如表2所示的原始数据集。第一列Code表示一个股票代码。

表2。原始数据集演示，东方货币投资者帖子

Code	Sector	Time	Content	Author
600999	Security	2020/1/13 13:25:**	招行都撑不住，大盘真不行 China Merchants Bank cannot hold on, the overall market is really not good.	股友2Ev***
002230	Computer application	2020/2/13 13:42:**	三天内会创新高的 It will reach a new high within three days	股友pnb***

文本数据预处理

我们需要删除不符合要求的帖子，包括有广告嫌疑的帖子和太短的帖子。我们采用Jieba分词器进行分词。Jieba分词器是一种高效的汉语分词工具，支持多种分词模式。

我们将尝试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)***，它可以在不手动设置语法规则的情况下检测散户评论中的常见短语，并增强分词的准确性。这样做的目的是为了扩展我们的分词结果，使文本分析更加鲁棒。

为了适应AI模型的嵌入层，我们通过裁剪或在末尾加零的方式使所有序列的长度相等。我们将所有帖子的长度固定为40个字符。这一步可以通过Keras提供的方法来实现，Keras是一个用Python编写的开源人工神经网络库。

面向情感SAI的分类任务

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绪词典和词库作为参考来进行研究，这导致缺乏对具体语境的针对性(例如，股市评论中技术术语含量高，社交媒体中俚语和表情符号含量高)。情感词典应该有针对性地构建

用于股票市场的帖子或评论。我们可以去掉一般的情绪词，使用词性标注(POS)的方法，尽可能保留人名、行为和形容词。通过对单词出现的频率进行排序，我们可以挑选出可供选择的新情感词，并用Amazon Mechanical Turk (Mturk)给它们打分。当然，这项工作也可以手工完成。

2. 深度学习模型

本研究构建了BiLSTM-CNN-Attention情绪分析模型，对股票帖子中的情绪倾向进行分类。这个模型的结构可能包括词嵌入层、BiLSTM层、CNN层、关注层和输出层。最终的版本将取决于研究过程中的情况。关于模型的更多细节请参考***附录2***。对于中文NLP来说，适当的预训练资源也是必要的。

回归方法

Fama和MacBeth(1973)提出了一种两步回归法，这是一种衡量这些风险因素如何正确解释资产或投资组合回报的实用方法。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，将每个参数视为整体参数的样本值;第二步是对第一步中所有参数取平均值，计算整体数据的估计参数。我们将使用这种方法进行实证检验。

我们将参考Loughran和McDonald(2011)的实证方法来选择被解释的变量，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***回归结果应该显示出显著***的***相关性***。

使用样本外数据进行稳健性测试。通过这个测试，我们可以证明我们的方法在非样本文本数据的预处理和情感分类方面的有效性。呈现的显著相关性也应该与样本数据一致。

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描述的投资者的心理偏差。比如，有大量的样本带有与“后悔”相关的词汇，无论是后悔不买还是后悔不卖。我们还想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

Code	Sector	Time	Content	Author	Regret for
600999	Security	2020/3/25 11:31:**	16.88 没卖后悔啊 I regret not selling at 16.88 yuan	股友 JVz***	Not selling
600999	Security	2020/7/6 10:25:**	周五 3 个点跑的，后悔 Sold at 3 points on Friday, regretting	股友 1r2***	Sold out early

参考

[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

(未实施的专利)

介绍

本发明属于NPL中的文本短语检测领域。它采用了文本分割技术和似然比检验，尽可能避免了检测过程中的人工干预，提高了短语检测的适用性和可靠性。

现有技术解决方案(通过搜索最相关的专利)试图通过遍历所有可选的单词组合来找到匹配，从而预先假设一个适用于大多数短语的词性依赖规则。如果语料库被限制在一个特定的领域，并且有许多专门的词，那么这应该是最好的方法。但是，这种默认规则需要大量的人为干预，并且需要随着语料库的发展而频繁更新，这在实际应用中很难完全实现，也不适合推文、博客、新闻文章。另外，在计算词组合的构词概率时，往往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

分词的功能是基于规则提取文本中包含的所有词，不同的分词设备有不同的分词规则。Jieba分词器是目前一款高效的中文分词设备，支持多种分词模式。本礼物采用了jieba分词器的精确模式，它是基于Python实现的。

似然比定义为约束条件下似然函数的最大值与无约束条件下似然函数的最大值之比。似然比检验的过程是: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)声明单词1的出现与单词2无关。第二个假设(备选假设)指出，看到单词1会改变看到单词2的可能性。如果我们接受备选假设，就意味着这两个词可以形成一个共同的短语。哪种假设成立是通过计算语料库中实际观察到的单词的频率来确定的。

一般的处理流程是通过jieba tokenizer将输入文本分割成单词，过滤停止词，组装n-gram，计算似然比。根据似然比的计算结果对n个图进行排序，选择似然比最小的n个图作为最终的检测结果。

附录2

bilstm - cnn -注意力模型

本研究建立了一个BiLSTM-CNN-Attention情感分析模型来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:通过预训练的word2vec模型构建文本的词向量，将分词后的句子映射成低维密集向量，每个词对应一个向量。生成的词向量包含语义信息，有利于下一层进一步的特征提取。

2. BiLSTM层:双向长短期记忆(Bidirectional Long - Short - Term Memory, BiLSTM)模型利用两层LSTM的叠加，打破了模型只能根据前一次的时间信息预测下一次输出的限制，可以更好地结合上下文进行输出。

3. CNN层:CNN(递归神经网络)在每个时间步对BiLSTM的输出进行卷积运算，进一步提取文本特征，并在卷积后加入max pooling，防止过拟合，减少参数和计算复杂度。

4. 注意层:通过分配不同的概率权值与CNN输出向量相乘得到注意机制的权矩阵，然后通过softmax函数进行计数得到权矩阵的值。

5. 输出层:输出层通过sigmoid函数对结果进行映射，得到情感分类的结果。

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帮助我们提高中文文本NLP任务的效率。我们将选择其中的一些，包括1)预训练的模型，它可以完成典型的情感分析任务，如句子级情感分类和方面级情感分类。2)中文单词和短语的嵌入语料库，它为中文单词提供有限维向量表示，即嵌入。我们将引入这些预训练的模型来构建我们的词嵌入层和BiLSTM层，然后对它们进行微调以适应我们的股票分析场景。