中国上证指数的情绪指标的构建及社群媒体文本分析论证与预测性研究

【摘要】: 情绪指数的构建在中国金融市场环境下一直没有一个规范化的范式，并且所构建的指数往往没有好的预测效果。正是基于此，本文决定首先根据 BW 模型，运用逐步回归的方法选择符合中国市场的情绪代理变量,再基于文本分析对所构建的情绪指标失去趋势性的区间进行解释，最后拟用机器学习的集成算法来寻找更优的预测模型。在筛选代理变量中，本文发现了新的符合中国市场环境的代理变量，并用主成分分析合成了短期情绪指数和长期情绪指数；而在文本分析过程中，本文根据形式上的最大词频指标和内容上的情绪打分发现在情绪指数没有明显趋势的区间，社群媒体的文本偏于中性。最后本文检验并分析了传统 VAR 方法，采用集成学习的方法集成预测，进一步解释与佐证了情绪指数预测效果不佳的结论。

【关键词】: 情绪指数；文本分析；委员会机器

[Abstract]: For a long time, there has been no norms regarding sentiment measurement in China’s financial market and that there has been proved repeatedly that the sentiment indices lacks predictability. This thesis firstly builds the long-term sentiment index and short-term sentiment index based on BW model, using stepwise regression to pick out agency variables. And secondly uses natural language processing to explain the interval where the sentiment indices differ thus losing tendency. And finally it builds integrate classifier to find optimal model. In selection of agency variables, it finds new variables that suit China’s financial market. In the process of texts, it finds explanation for the lack of tendency from both the frequency of words and context of sentiment. And in the search of predict model, it reviews the VAR models and uses the ensemble learning to prove again that there is little predictability concerning sentiment index.

[Key Words]: Investor Sentiment; Natural Language Processing; Ensemble Learning

目录

一、引言

传统的金融理论的基本假设是投资者是理性的，但是现实情况并非如此，特别是在有效市场理论（EMH）被证明弱有效或基本无效的市场环境中，非理性的投资行为十分明显。行为金融学理论认为，投资者会对接收到的金融信息产生不同的认知，而这些有偏差的认知经过聚集效应与放大效应，便能对整个金融市场产生影响，最终导致市场波动性变大。因此对投资者及市场情绪的量化研究，具有前瞻的理论与使用意义。

目前对于市场情绪的量化评测方法，国内外学者展开了多角度而深入的研究，试图采用多种情绪代理指标来揣测市场情绪的变动来形成相关的情绪指数，最终达到根据情绪指数预测收益率波动性的目的。这些指标各有优劣，统计方法也多种多样，最为成熟与应用最广的指标是 Baker 和 Wurgler 在 2006 年通过主成分分析合成的情绪指数。这种方法来构建情绪指数，所选取的代理指标都是间接指标，即通过市场交易数据，得到反映市场情绪的统计量。这种指标避免了像问卷调查形式的主观指标所带来的主观性偏差，是现在主流研究较为常用的方法。然而后来有许多论文指出 BW 的情绪指数在中国市场的预测性上表现不佳。这启示本文：第一需要寻找更适合中国市场的变量来构造情绪指数，第二需要考察与解释信息失去预测性的原因，第三需要研究能够通过其他的统计技术获得更好的预测效果。

这三个思考层次是递进的关系。首先对变量的选取，应当分析这些变量背后所代理的金融意义、这些金融意义在中国市场环境下是否能通过 BW 模型的变量来代理、若不能，该选择哪些变量进行代理、又如何筛选这些变量。

第二个层次是解释情绪信息。在第一个层次中所构造的情绪指标，更多的是从统计学的角度进行考察，具有不可避免的局限性。信息解释的维度应更加多元。近年来研究最多的是社群媒体对金融市场的影响，特别是对金融市场、社群媒体、文本信息的处理三者相结合的交叉研究主题，一直是近年来国外学术热点。但利用文本分析技术对中文社交媒体做深层次挖掘有效信息的研究，在中国学术领域基本为空白。因此本文希望对情绪指数失去趋势性的区间进行文本分析，从社群媒体的角度考察情绪指数是否反映了市场情绪。

最后一个层次是希望能用数据挖掘的技术方法获得更好的预测效果，如果可以挖掘出更好的预测效果则弥补了情绪指数预测性不强的问题。如果不能挖掘出更好的结果，则进一步佐证情绪指数的形成滞后于股价的波动，与前人研究结果保持一致。这一步更多的是从纯粹的统计方法上进行挖掘，金融的解释意义未见得明显。

本文接下来组织结构如下：第二部分文献综述，发现前人研究的成果与不足，寻找研究的空白与创新之处；第三部分进行情绪指数构建的研究；第四部分对所构建的情绪指数结合社群媒体进行文本分析；第五部分寻找更优的预测模型；第六部分是结论与展望。

二、文献综述

2.1 国外学者研究

对情绪指数的研究最早也最为认可的是 Baker 和 Wurgler 在 2006 年的论文。他们运用主成分分析的方法，从 6 个变量中提取两个主成分作为情绪指数。这六个变量是①到期债券的折价率；②NYSE 股票的自然对数周转率；③IPO 总数；④IPO 首日平均收益率；⑤债股比；⑥溢出红利。

其后的研究，分别在纳入变量，统计方法，与地区修正性上做了改进与延伸。在纳入变量上，Jun Sik Kim 等（2017）将情绪指数与期权对股票率（O/S Ratio）的可预测性做了检验，发现情绪指数能影响到 O/S Ratio；Steven E.Sibley 等（2015）认为情绪指数更主要受到短期利率和流动性风险影响，在主成分提取时换取标准风险（Standard Risk）商业周期等变量，发现在横截面数据上，股票收益确实当且仅当收到商业周期变量的影响；考察情绪指数是否构成动量因子，并发现显著结果(Xing Han 等 2015)；纳入新闻因素，构建“累计新闻情绪指数（ANSI）”，考察其对中国台湾股票市场的影响（Yu-Chen Wei 等，2016）；基于会计报表考察情绪指数对会计信息的影响（Feng Niu 等，2015）等；在统计方法上的创新，主要有 Jyoti Kumari 等运用 GARCH（1，1）、TGARCH、和 GARCH-VAR 模型检验情绪指数对印度股票的可预测性；基于 PLS 的情绪指数构建（Dashan Hunag 等，2014）；基于信息博弈均衡所提出的信息资产定价模型指数（Stambaugh 等,2012）等；基于 VAR 和 Granger 因果检验考察情绪指数对不同部门收益率波动情况（Aleksandra 等，2017）等；

对地区修正性的模型必要性是有效市场理论（以下简称 EMH）不成立。而通过实证检验，也确实发现在 EMH 不成立的市场，情绪指数的构建应有所区别。Waseda 等（2017）发现 BW 模型中的六个因素不能很好地反应中国市场的情绪指数；Huang 等(2015)指出 BW 模型对中国市场不具有预测性，并基于 PLS 提出了线性修正模型，Stelios Bekiros 等（2016）在他的基础上又提出了一种非线性的模型；Joseph 等（2016）针对泰国 SET 和 Mai 两个市场的联动性提取一个情绪变量作为两个市场的预测等等。

2.2 国内学者研究

目前国内学者对情绪指数的研究分为两类。

第一类是选择各种模型的框架下研究股票市场情绪指数的度量问题，如将市场情绪指数作为状态变量，股票指数作为观测变量，建立状态空间方程组，实现市场情绪指数的度量（蒋文江等，2016）；构建向量自回归 VAR 模型，检验了投资者情绪指数、分析师推荐指数和上证指数收益率之间的动态关系，并分析之间的格兰杰因果关系（段江娇等，2014）；基于卡尔曼滤波技术构建复合指数，并检验其有效性（蔡志刚等，2016；靖荣华等，2016）。

第二类是选择各种变量作为情绪变量的代理变量来实现市场情绪指数的代理问题。这些代理变量，可能是统计意义上的代理变量，如主成分分析的方法，基于封闭式基金折价、成交量、消费者信心指数、新开户数 4 个指标构造投资者情绪指数（闫芬娟等，2013）；基于封闭式基金折价、交易量、波动率、行业 beta 值和每股流通市值这 5 个单项情绪指标提取主成分，构建了 23 个申万一级行业的投资者综合情绪指数(IICSI)（卢米雪等，2014）；而对主成分分析法采取单个指标来构建指数可能造成精度偏差的问题，修正研究又有基于偏最小二乘法所构建的投资者情绪综合指数（王镇等，2014）。这些变量的选取，还有直接使用具体的市场因素作为代理。4 个情绪代理变量分别与宏观经济变量进行回归，将各次回归得到的残差作为新的代理变量（马若微等，2015）；在 Fama 和 French 三因子模型的基础上，引入市场情绪因子，分别对股票市场收益率进行回归分析（万方明等，2013；都星汉等，2014）；余额宝情绪指数作为刻画散户投资者入市意愿情绪的代理变量，检验余额宝情绪指数和上证综指、深圳成指、创业板指、中小板指收益率及成交量之间的动态关系（蒋文璐等，2016）。

从技术方法上看，主要有以下两种：

第一、主成分分析提取情绪成分（或直接选取代理变量），然后使用向量自回归和格兰杰因果检验来检验该情绪成分的有效性；

第二、基于文本分析技术来构建情绪指数，董大勇和肖作平(2011)、Wu (2012)、施荣盛和陈工孟(2012)选用东方财金融市场富股票论坛中的发帖量作为代理；段江娇等（2014）进一步在帖子数量上分析了帖子内容。类似研究还有《网络媒体对股票市场的影响》（金雪军等，2013），采用文本挖掘工具提取每条发帖所蕴含的情绪并用数字编码，然后将加总这些编码，得到所需要的指数。程琬芸等（2013）文章中表示：“利用文本分析技术对中文社交媒体做深层次挖掘有效信息的研究基本空白”。

2.3 总结与创新

综合大量国内外文献，我发现：在对情绪指数的度量与预测性问题上，国内外学者研究已经很深入，且统计方法较为一致，最为常用的 VAR 和 Granger 因果检验。但许多文献也指出，对于 EMH 不能成立的市场，情绪指数对收益率很难具有预测效应。

而在情绪指数的构建和变量选取问题上，国内外研究暂时没有统一的结论，特别是对文本信息的挖掘上，对中文社交媒体做深层次挖掘的研究基本为空白。

基于以上文献综述，本文决定从情绪指数构建、文本信息挖掘、和预测性三个方面给予更为深入的研究和探讨。

在情绪指数构建方面，本文创新之处在于根据 BW 模型来选择更加贴近中国市场的变量，筛选方法是对上证 50 的成分股，每一只进行逐步回归，根据入选率来筛选。之后再利用主成分分析的方法对所筛选的变量进行降维。

在文本信息挖掘方面，本文的创新之处在于从形式到内容对情绪指数无趋势性的区间进行分析，弥补中文社群媒体做深层次挖掘有效信息的研究空白。

在预测性方面，本文的创新之处在于综合了传统 VAR 模型与委员会机器，对上证 50 的成分股逐一分析，得到的结果更为全面可靠。

三、情绪指数构建

3.1 代理变量的金融意义的分析

根据 Baker 和 Wurgler 在 2006 年的论文，本文对情绪指数的构建仍采取代理变量和主成分分析的方法。原文从 6 个变量中提取两个主成分作为情绪指数。这六个变量是①到期债券的折价率；②NYSE 股票的自然对数周转率；③IPO 总数；④IPO 首日平均收益率；⑤债股比；⑥溢出红利。首先本文将探讨这六个变量背后的金融意义，及其与中国金融市场的适应性分析：

1）折价率的选取：

根据 Baker（2005，2006）年的研究，对折价率的选取是基于股市、债市、和利率市场的联动性。在我国，这种联动性常用的代理变量分为两种：一是如封闭式基金的折价率、当期到期债券的折价率这种直接指标，另一是如 50ETF 能量强弱指数的技术指标。本文将全面考察这两类指标，并从中选取一个指标代理联动性情绪；

2）股票的自然对数周转率选取：

根据 Baker（2005，2006）年的研究，对股票自然对数周转率的变量选取是对交易因素的代理。在我国，这类指标以各种换手率作为代表。本文将选取证券在指定区间内转手买卖的频率作为换手率指标。

3）溢出红利：

在 Baker（2005，2006）年的研究中，溢出红利（The Dividend Premium）定义成：”Log difference of the average market-to-book ratios of dividend payers and non-payers（平均红利支付公司与不支付红利公司的市帐比的对数差）”。而这种定义在我国的市场环境下很难有找到代理变量的数据，原因在于我国大多数上市公司股利的发放在周期内并无规律与规范。

因此本文需要寻找变量重新代理这一因素。这一变量所想表达的实际金融意义及其背后假设是迎合理论（catering theory）：即上市公司会迎合投资者的投资者需求，当市场上的红利水平已经处于高位时。根据 Laura Yue Liu(2010)的研究，当红利水平处于高位的时候，代理问题（agency problems）亦处于高位，并实证跟踪了美国 1962 年到 2004 年的数据证明了这种相关性。

这启示本文能否用代理问题的常用变量来协同代理溢出红利这个因素。常见的研究代理问题的变量是市值与面值的对数差 Nissim, D., Ziv, A（2001）等。虽然还有其他变量可以代理，但仍存在红利数据的选取问题，因此本文首先考虑对数差作为主要代理变量。

4）债股比、IPO 总数与首日平均收益率

债股比衡量的是公司的资本结构。IPO 总数与首日收益率这两个变量衡量的市场热情与向好期望。因这三类数据易于获得，因此本文将直接使用这三类数据，而不再选取新的代理变量。

下表是经过上述分析后决定进行筛选的代理变量。

表 3-1 代理变量待筛选表

变量名	代表含义
F1	封闭式基金的折价率
F2	当期到期债券的折价率
F3	50ETF 能量强弱指数
F4	市场平均的换手率
F5	当月 IPO 总数
F6	IPO 首日平均收益率
F7	股票市值与债券市值的比率
F8	股票市值与面值的对数差

3.2 变量筛选

本文筛选变量的方法是根据 Fama and French（1993）的 5 因素模型对上证 50 的每一只成分股进行回归，然后进行逐步回归的方法筛选变量，根据入选次数选择与筛除这八个变量。

回归模型可以写作：

$$R_i = \beta_0 + \beta_1 \mathrm{Rm} + \beta_2 smb + \beta_3 umd + \beta_4 rmw + \beta_5 cma + \sum \alpha_i F_i \quad (3.1)$$

其中 Rm 是市场风险因子，smb 是规模风险因子，umd 是惯性因子，rmw 是盈利因子，cma 是投资模式因子。

SMB 构建方法是：t 年 6 月使用流通市值进行排序，计算 t 年 7 到 12 月及 t+1 年 1 到 6 月份，小盘股组合和大盘股组合的（流通市值加权及等权重）收益率之差。

UMD 构建方法是：每个月按前 2-12 个月的累积收益率排序，计算高收益股票组合和低收益股票组合的（流通市值加权及等权重）收益率之差。具体构建方法见 Carhart（1997）。

RMW 构建方法是：t 年 6 月使用盈利能力（盈利能力计算方法：t-1 年 12 月的营业收人减去营业成本、利息支出、销售费用、管理费用后与 t-1 年 12 月的账面价值之比）进行排序，计算 t 年 7 到 12 月及 t+1 年 1 到 6 月份，高盈利股票组合和低盈利组合的（流通市值加权及等权重）收益率之差。

CMA 构建方法是：t 年 6 月使用投资水平（投资水平的计算：t-1 财年的新增总资产除以 t-2 财年末的总资产）进行排序，计算 t 年 7 到 12 月及 t+1 年 1 到 6 月份，低投资比例股票组合和高投资比例股票组合的（流通市值加权及等权重）收益率之差。具体构建方法见 fama and french (2015）。

因变量是上证 50 的成分股月收益率，数据来源是 Wind 数据库。

数据长度是从 2003-2 月份到 2017-3 月份，因素模型数据来源是中国资产管理研究中心（http://sf.cufe.edu.cn/kxyj/kyjg/zgzcglyjzx/index.htm），F1，F2，F5，F6，F8 来自于深圳国泰安教育技术股份有限公司，F3，F4，F7 来自于 Wind 数据库。

逐步回归模型的算法设计思路是：

② 回归基模型；

②规定入选变量所应具备的显著性水平，本文定 BIC 指标（sle）；

③规定筛出变量所应具备的显著性水平，本文定为 BIC 指标（sls），并确保 sls>sle;

④

循环直到收敛：

对纳入 Fi 后的模型进行回归；

if(Fi 的显著性(信息量指标)<=sle)：

Fi 入选；

再次检查 Fj（j!=i）的显著性；

if(Fj>sls):

Fj 落选；

详细的程序实现参考附录 1。

在进行回归前，首先将所选取的 8 个变量与市场收益率的分布图进行可视化，可视化结果见下图。

可以发现，所入选的 8 个指标里面，大部分与 Rm 的分布具有趋同性，从某种意义上说选择这些变量是合理的。

因为量纲的不一致，在回归前还须对变量进行标化处理。此外还需对缺失值进行删除。下表是 50 次逐步回归后的入选结果：

表 3-2 逐步回归入选次数表

指标	F1	F2	F3	F4	F5	F6	F7	F8
入选次数	50	37	41	49	50	50	50	50
入选率	100.0%	74.0%	82.0%	98.0%	100.0%	100.0%	100.0%	100.0%

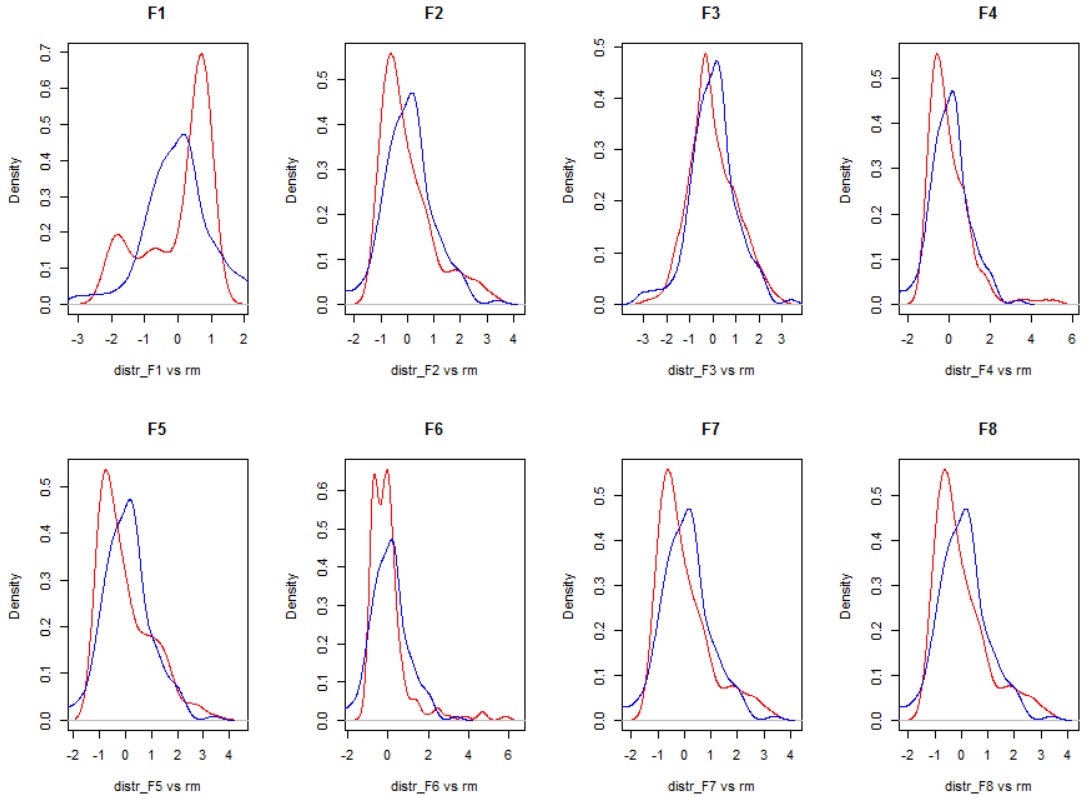

图 3-1 待筛选变量密度分布图与市场收益密度分布图对照图

从表中可以看出，F1，F5，F6，F7，F8 全部入选，这与原 BW 模型的③-⑥指标基本一致。在到期债券收益率上，实证结果显示封闭基金的到期收益率的折价率比一般债券市场的收益率更能反应两个市场的联动性。在反应市场交易情绪的指标上，技术指标与换手率的差异不是特别大。进一步分析，换手率的分母是总股本，而 50ETF 能量指数的分母是上证 50ETF 的参考基金单位净值，分子部分都是流通交易量。两者在定义上并不存在本质上的区别。考虑到平均换手率纳入了所有流通中股票数，因此可能更能体现市场的交易意愿，因此本文选取 F4 作为交易行为的代理变量。

至此，代理变量的选取和筛选工作完成，所选取的情绪因子为 F1、F4、F5、F6、F7、F8。

3.3 构建情绪指数

针对所选取的这六个代理变量，本文进行了主成分分析。通过 100 次随机数据模拟的平行分析，得到碎石图如下图。

通过碎石图可以发现在选择 2 个主成分时即可保留大部分信息。下表进一步显示出了各个变量在主成分上的得分。在第一主成分中，F6 与 F8 得分较少；在第二主成分中，主要由 F4 与 F6 因子组成。由各变量的金融意义，大致可以认为，第一主成分反映了投资者的中长期投资情绪，第二主成分反映了投资者的较短期的投资情绪。式 3.2、3.2 表示出来了这两个情绪指数。

至此，情绪指数的构建部分完成了。接下来将从文本挖掘的角度对社群媒体文本信息和所构建的情绪指数进行比较分析。

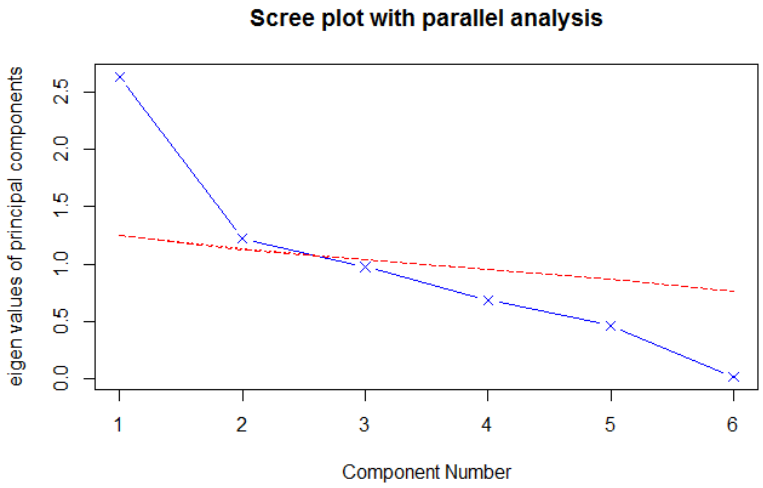

图 3-2 情绪指数构建-主成分分析碎石图

表 3-3 情绪指数得分表

变量	RC1	RC2
F1	0.29	0.09
F4	0.27	-0.52
F5	0.33	-0.09
F6	0.03	0.64
F7	0.35	0.17
F8	0.01	-0.29

$$L\_term = 0.29F1 + 0.27F4 + 0.33F5 + 0.03F6 + 0.35F7 + 0.01F8 \quad (3.2)$$

$$S\_term = 0.09F1 - 0.52F4 - 0.09F5 + 0.64F6 + 0.17F7 - 0.29F8 \quad (3.3)$$

四、文本分析对情绪指数的解释

4.1 情绪指数的趋势性

根据上一节得到的情绪指数可以知道，如果长期和短期的情绪指数运动趋势一致，那么通过方差贡献率加权而得到的综合情绪指数也具有该种运动趋势。若两种情绪指数运动趋势相反，那么经过方差贡献率加权而得到的综合情绪指数的趋势性将被抵消，这也是情绪指数不具有明显趋势的原因。

因此，本文首先做出这两个情绪主成分在 2003-2 月份到 2017-3 月份的时间序列图。

红线是按方差贡献率加权得到的情绪指数线，蓝线是中长期的投资者情绪，黑线是短期的投资者情绪。由图可以看出，两个指数在大多数时间保持了趋同性，在 2011 年 11 月份到 2012 年 5 月份近半年的时间两者运动有明显的不一致情况。总体上看短期情绪指数波动的振幅更为剧烈。

情绪指数趋同运动情况说明了所选取的代理变量能够充分反应投资者的情绪，此时对社群网络的文本数据挖掘和分析意义将不明显。在两者运动不趋同的

情况下，说明市场情绪在统计意义上存在着分歧，进而导致加权后的情绪指数失去明显的趋势性。因此在对文本挖掘与情绪指数的关联性上的研究，本文选取阶段是 2011 年 11 月份到 2012 年 5 月份半年的文本挖掘。

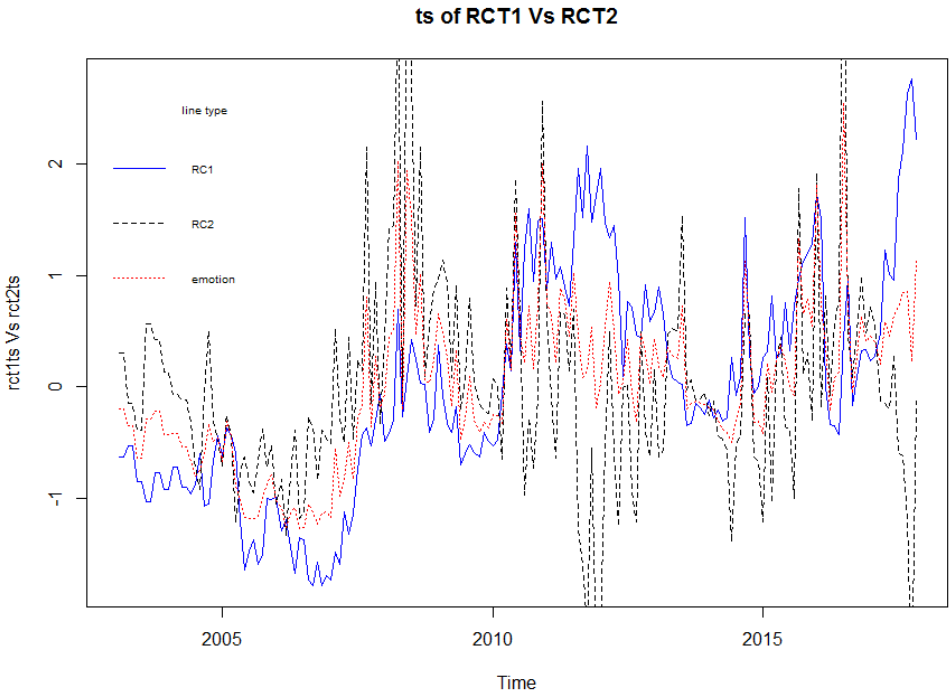

图 4-1 情绪指数时间序列图

对文本数据的挖掘，分为形式上的挖掘与内容上的挖掘。本文接下来将分别对形式和内容进行挖掘。

4.2 形式挖掘

所谓形式挖掘主要是指社群媒体信息的外显性特征进行挖掘，如发帖数量，最大词频等。董大勇和肖作平(2011)、Wu (2012)、施荣盛和陈工孟(2012)选用东方财金融市场富股票论坛中的发帖量作为代理，用每个帖子的最大词频的情绪作为该帖子的情绪。情绪度量被表示为：

$$\sum \frac{\text{积极帖数}}{\text{总帖数}} \quad (4.1)$$

这种方法的最大优势是简明易操作，但这种方式的最大缺点是没有考虑情绪的强度，即最大词频很可能并不是能够传达确切文本信息的因素。

但是这种方法在对模式化文本,如社群媒体报道，进行处理时往往是有效率的。因此本文也将采用这种方法对模式化的文本进行分析。

在对形式情绪分析上，本文选取的数据库是中国经济新闻库（http://www.bjinfobank.com/DataList.do?method=toSpecSearch&dbCode=HK），选取的时间范围是 2011 年 11 月到 2012 年 5 月，所选取的关键词是“市场情绪”、“投资者情绪”、和“上证综合指数”，选取的匹配模式是模糊查询。一共命中 89 篇新闻，来源有经济参考报、中国证券报、上海证券报、腾讯网、财经网等，基本涵盖各级别社群媒体以及网络新媒体。

对最大词频的统计算法，最常用的是开源的程序包 jieba 分词。该分词器的原理是：基于 Trie 树结构实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（DAG)；采用了动态规划查找最大概率路径，找出基于词频的最大切分组合；对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。因算法解释起来比较繁琐，且与主题无关，在此处略去。

下表是最大词频表，下图是根据词频生成的词云图。

表 4-1 前 20 最大词频

词汇	出现数	词频
经济	75	10.90%
市场	68	9.88%
政策	63	9.16%
投资	57	8.28%
板块	45	6.54%
出现	40	5.81%
增长	39	5.67%
反弹	38	5.52%
行业	38	5.52%
上涨	37	5.38%
预期	37	5.38%
投资者	36	5.23%
认为	35	5.09%
数据	33	4.80%
月份	30	4.36%
可能	27	3.92%
大盘	25	3.63%
利好	25	3.63%
下跌	25	3.63%
CPI	24	3.49%

图 4-2 根据词频生成的词云图

可以看出，最大词频的情绪都是一些中性词汇。而根据情绪度量公式，算出这半年的情绪度量为 0.63，也说明了通过最大词频得到的帖数情绪为中性稍积极。

但这个结果也间接说明了，情绪指数失去明显倾向性与形式情绪分析结果不明显是有一定相关性的。

4.3 内容挖掘

所谓内容挖掘，是指对每篇文本数据的每一个分句进行情绪打分，以情绪的平均得分作为该文本的得分。常用的对情感打分的方法分为绝对打分法与相对比较法。本文先尝试使用绝对打分法来实验。

4.3.1 绝对打分法

绝对打分是分词后对词性与程度设置权重，然后对每一分句的情感值加权平均得到文本的情感值。这个值是绝对的，具有可加性与可比性，易于直接比较与计算。这个方法比较成熟地算法可以如下表示：

Step1：进行分词与分句；

Step2：

For 每一分句 in 文本：

① 对每一分句，查找情感词，并根据对照词典，记录情感词的符号与位置；

② 在情感词前面寻找程度副词，以程度副词为权重，对情感词加权得到情感值；

③ 往情感词后面寻找否定词，否定词数记为 i，再对②中情感值乘以$(-1)^i$ ；

④ 查看每一分句最后的符号，对叹号与问号设置对应的情感数，加在③中的情感值中；

⑤ 得到每一分句的情感值，以数组形式储存；

加总后平均得到文本的情感值。

本文所使用的情感词、情绪词以及否定词参考的是 GitHub 上现成的词库（ https://github.com/dongxiexidian/Chinese/tree/master/%E4%B8%AD%E6%96%87%E5%88%86%E8%AF%8D%E8%AF%8D%E5%BA%93%E6%95%B4%E7%90%86/%E4%B8%AD%E6%96%87%E5%88%86%E8%AF%8D%E8%AF%8D%E5%BA%93%E6%95%B4%E7%90%86）。

这个算法的完整实现可参考附录 2。

下表是根据绝对打分法得到的每一篇文章的情感值，完整表格可参考附录 3。

表 4-2 情绪打分表

Piece	Positive	Negative
1	73.18	222.62
2	58.00	226.58
3	70.03	228.83
4	69.88	209.12
……	……	……
86	41.21	266.27
87	225.02	97.34
88	66.89	231.47
89	232.45	82.25
mean	138.07	169.44

可以看出，在 89 篇文本中，积极情绪得分实际上是弱于消极情绪得分的。事实上，通过对每一分句打分的方法得到的情绪值比通过最大词频得到的情绪值更为准确。因为在实词前面的程度副词或否定词，每一篇可能都不一样，导致最大词频只搜集了实词，而没能搜集到实词的上下文环境。

4.3.2 相对比较法

相似比较法的思想是选取一篇基文本，这篇文本所代表的情绪最为客观。然后再将其他文本用向量化的方式表示，与基文本计算距离，距离的偏离程度代表该文本的情绪值。

这个模型的难点在于基文本的选取和文本的向量化。由于中国目前尚未建立权威的投资者情绪报告或市场情绪指数报告，若采用多篇文本成分分析与合成的方法，实质上与绝对比较法相似。对文本的向量化，较常用的方法是 LSI。LSI 是概率主题模型的一种，核心思想是：每篇文本中有多个概率分布不同的主题；每个主题中都包含所有已知词，但是这些词在不同主题中的概率分布不同。LSI 通过奇异值分解的方法计算出文本中各个主题的概率分布。从而根据每个主题上的得分得到主题向量。

鉴于基文本难以选取的问题，相对比较法本文将跳过此方法。

4.4 小结

本节针对上节所构造的情绪指数不明显的时间段（2011 年 11 月到 2012 年 5 月份），选取了 89 篇来自新闻媒体和网络媒体的报道，分别进行了形式上的分析与内容上的分析，结果发现，在形式上，最大词频体现为中性词，整体的情感倾向为中性偏积极。在内容上进行分析，本文分别分析了绝对情绪值法和相对比较法两种方法。在绝对情绪值法上，本文对 89 篇报道的每一篇的情绪值进行了计算，发现积极情绪的整体得分事实上略低于消极情绪的整体得分。而相对比较法，由于中国股票市场尚不存在比较权威的基文本，若采用多篇文本合成的方法生成基向量，将导致该方法与绝对情绪值法无显著区别。

本节的结论是，对于情绪指数趋势不明显的时间段，进行适当的文本分析可以部分解释趋势不明显的原因。接下来本文将讨论情绪指数的预测性的问题。

五、预测性

传统的因子模型在前人研究中也被证明预测效力不足（Waseda 等（2017）），究其原因，本质上是因子模型适用于解释风险的分散因素，而不适用于预先风险因素再进行预测。许多实证结果也（Aleksandra 等，2017；Huang 等，2015；段江娇等，2014；万方明等，2013；都星汉等，2014；等）都显示 VAR 模型并不能合理对指数收益率进行预测。正是因为传统的时间序列或者回归模型解释能力强，但预测效果不强，本文试图采用机器学习的方法，拟获得更好的预测效果。不过在使用机器学习的方法以前，本文也将首先考察 VAR 模型。在考察 VAR 模型过程中，本文拟获得最优滞后阶，并提取该变量作为训练模型的变量。

5.1 VAR

向量自回归模型简称 VAR 模型，是一种常用的计量经济模型，1980 年由克里斯托弗·西姆斯（Christopher Sims）提出。VAR 模型是用模型中所有当期变量对所有变量的若干滞后变量进行回归。VAR 模型用来估计联合内生变量的动态

关系，而不带有任何事先约束条件。它是 AR 模型的推广，此模型目前已得到广泛应用。

向量自回归(VAR)是基于数据的统计性质建立模型，VAR 模型把系统中每一个内生变量作为系统中所有内生变量的滞后值的函数来构造模型，从而将单变量自回归模型推广到由多元时间序列变量组成的“向量”自回归模型。

形式如下：

$$Y_t = c + \sum_{i=1}^{p} A_i Y_{t-i} + e_t \ (5.1)$$

其中Y_t，c是n×2维向量，p是滞后阶数，e_t是误差向量，满足:

$\mathrm{E}(e_t) = 0$（误差均值为0）

$\mathrm{E}(e_t e_t^T) = \omega$（误差的协方差矩阵是一个正定矩阵）

$\mathrm{E}(e_t e_{t-k}^T) = 0$（不存在自相关性）

本文考察的是情绪指数和股价收益率之间的动态关系。因此所选取的变量为3.3节所构建的情绪指数，根据方差贡献率加权而得到的综合情绪指数（记为emo）与上证综合指数的收益率（记为rtn）。时间区间还是与之前一致，从2013-2月到2017-3月年。将指数收益率去量纲化并不影响模型结果，因此本文先将指数收益率标准化，做出了两者无量纲的时间序列图如下：

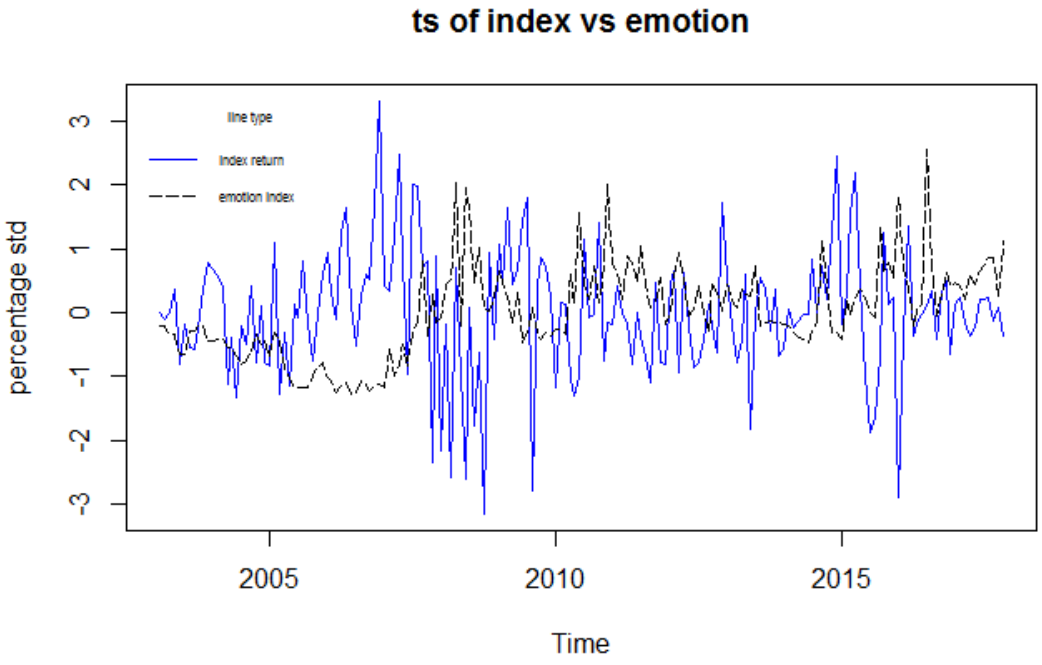

图 5-1 指数收益率与情绪指数的时间序列图

可以看出情绪指数在很多情况下是滞后与大盘指数收益率的，这也间接支持了很多论文发现用情绪指数的时间序列模型很难有预测效应的结论。因此本文将

不使用 VAR 模型进行预测，而是使用 VAR 模型确定滞后阶数，再用机器学习的方法拟合与预测。

下表是分别对情绪指数与收益率进行平稳性检验得到的结果：

表 5-1 平稳性检验

T 值 rtn	-7.598	19.244	28.866
T 值 emo	-4.717	7.457	11.158
	1%	5%	10%
tau3	-3.99	-3.43	-3.13
phi2	6.22	4.75	4.07
phi3	8.43	6.49	5.47

可以看出指数收益率与情绪指数的统计量值都小于 1%，5%，10%的临界值，因此两个序列都是平稳的。因此不用进行协整检验。接下来进行滞后阶数的确定。下表是 10 阶内各准则下的最优阶数：

表 5-2 各准则下的最优阶数

AIC	**HQ**	**SC**	**FPE**
9	2	2	9

根据结果，不同的信息准则有不同的滞后阶数，选择 2 阶或者 49 阶都是可以的，一般来说选择在相同条件下更加简洁的模型，因此选择 2 阶滞后。

接下来选取两阶滞后进行格兰杰因果检验。格兰杰因果检验的结果如下表：

表 5-3 Granger 检验表

rtn~emo				
	Res.Df	Df	F	Pr(>F)
1	171			
2	173	-2	2.580	0.079
emo~rtn				
	Res.Df	Df	F	Pr(>F)
1	171			
2	173	-2	3.215	0.043

检验结果表示，当原假设为“emo 不是引起 rtn 变化的 Granger 原因”，P 值为 0.07>0.05，我们无法拒绝原假设；而当原假设为“rtn 不是引起 emo 变化的 Granger 原因”时，P 值为 0.042585<0.05，我们可以拒绝原假设。因此，可以证明：rtn 是 emo 的 Granger 原因。这也进一步证明了前人研究的结果：即用一般时间序列模型，情绪指数不能很好地预测收益率。

经过 VAR 模型拟合后的参数表如下。

因此拟合后模型可以写作：

$$\begin{pmatrix} rtn_{i,t} \\ emo_{i,t} \end{pmatrix} = \begin{pmatrix} 0.612 \\ 0.008 \end{pmatrix} + \begin{pmatrix} 0.112 & 0.124 \\ -0.009 & 0.445 \end{pmatrix} \begin{pmatrix} rtn_{i,t-1} \\ emo_{i,t-1} \end{pmatrix} + \begin{pmatrix} 0.083 & -2.069 \\ 0.009 & 0.353 \end{pmatrix} \begin{pmatrix} rtn_{i,t-2} \\ emo_{i,t-2} \end{pmatrix} + \begin{pmatrix} e_{1,t} \\ e_{2,t} \end{pmatrix} \tag{5.2}$$

但是查看参数表可以知道，大多数参数并没有良好的显著性，这也间接说明 VAR 模型是失拟的。

表 5-4 VAR 参数表

	rtn.Estimate	rtn.Std..Error	rtn.t.value	Pr(>|t|)
rtn.l1	0.112	0.076	1.463	0.145
emo.l1	0.124	1.179	0.105	0.917
rtn.l2	0.083	0.077	1.082	0.281
emo.l2	-2.069	1.161	-1.782	0.077
const	0.612	0.605	1.012	0.313
	emo.Estimate	emo.Std..Error	emo.t.value	Pr(>|t|)
rtn.l1	-0.009	0.005	-1.915	0.057
emo.l1	0.445	0.073	6.117	0.000***
rtn.l2	0.009	0.005	1.848	0.066
emo.l2	0.353	0.072	4.928	0.000***
const	0.008	0.037	0.202	0.840

接下来用这个模型做脉冲响应分析，得到的脉冲响应图如下：

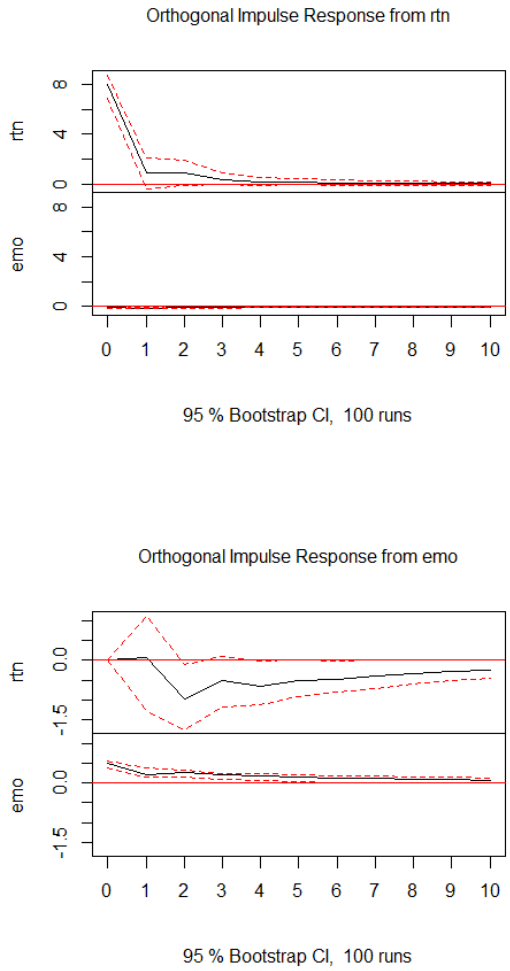

图 5-2 脉冲响应图

根据图形 rtn 自身以及 emo 的波动对 rtn 有反向的冲击。rtn 对自身的影响没有滞后期，并且自身波动的影响随着时期的增加会越来越小。emo 波动对 rtn 的影响在第二期达到最大，此后随时间增加波动减小。

rtn 波动对 emo 几乎没有产生影响，这显然是不符合常理的，也间接说明 VAR 模型没有预测性。emo 波动对自身有一个反向的冲击，这个冲击从第二期后又逐渐减少。

至此，本文已经证明了前人研究的结论：根据 BW 模型得出的情绪指数不能通过 VAR 等传统时间序列方法得到良好的预测性。

5.2 委员会机器

正是由于 VAR 等传统时间序列模型拟合效果不佳，本文考虑使用机器学习的统计工作进行预测探索。首先先要考察输入变量和输出变量的选择问题，本文决定纳入 VAR 模型的滞后项和因子模型的一阶滞后项。接着本文将探讨委员会机器及其算法，最后本文将输出结果。

5.2.1 变量选择

因子模型更偏向于解释情绪指数是否能对收益率产生影响，回归结果显示所构建的情绪指数是显著的。说明情绪指数可以解释部分系统性风险。但是通过传统的时间序列模型，单纯考察情绪指数与收益率的关系，众多实证结果都显示情绪指数的预测效力不足。因此本文考虑，预测模型的变量选择纳入因子模型的因素，以及时间序列模型的滞后期数，兼顾两个模型的变量选择。

此外，数值模型稳定性较差，加上情绪指数有一定滞后性，因此考察运动趋势更为合理，因此本模型将回归问题改造成分类问题。

表 5-5 委员会机器参数选择表

变量类型	变量名称	备注
输出变量（factor 类）	上阵 50 成分股收益率的运动趋势	当月收益率为正记为 1，为负记为 0，记为 rtn
数值变量（数值类，下同）	指数收益率一阶滞后项	rtn_lag1
	指数收益率二阶滞后项	rtn_lag2
	情绪指数一阶滞后项	emo_lag1
	情绪指数二阶滞后项	emo_lag2
	规模风险因子一阶滞后项	smb_lag1
	惯性因子一阶滞后项	umd_lag1
	盈利因子一阶滞后项	rmw_lag1
	投资模式因子一阶滞后项	cma_lag1

上表是最终选择的变量列表及其名称。

5.2.2 集成学习

委员会机器是一种分类器的思想，将各种统计方法的预测结果进行投票，投票的权重根据错误率来判定。本文将选取 QDA、Bagging、SVM 和 ANN 四个模型来训练模型，然后根据（1-错误率）为权重进行投票。

二次判别分析（QDA）是一种线性判别分析算法，具有与线性判别法（LDA）类似的算法特征，区别仅在于：当不同分类样本的协方差矩阵相同时，使用线性判别分析；当不同分类样本的协方差矩阵不同时，则应该使用二次判别。根据数据性质，本文选取 QDA 显然比 LDA 更加合适。

Bagging 算法是树方法的延伸，它的特点在于随机采样(bootsrap)，也即放回抽样，这样之前采集到的样本在放回后有可能继续被采集到。对于 Bagging 算法，一般会随机采集和训练集样本数 m 一样个数的样本。这样得到的采样集和训练集样本的个数相同，但是样本内容不同。Bagging 算法本身是一种不带权重的集成学习算法，符合弱分类器的选取。

支持向量机（SVM）的算法思想是寻找最优分隔平面，使得样本空间落在支持向量两边的距离尽可能的远。寻找支持向量与最优分隔平面的过程最终可以证明为一个凸优化问题，由 SMO，PSO 等算法可以找到最优解，进而求得最优分隔。

人工神经网络（ANN）是由大量处理单元互联组成的非线性、自适应信息处理系统。它是在现代神经科学研究成果的基础上提出的，试图通过模拟大脑神经网络处理、记忆信息的方式进行信息处理。算法由输入向量开始，进入多层感知器，根据初始设定的权重和阈值，经过反复地训练模型，找到权重和阈值的最优估计，最终得到输出。

可以看出，这四个模型的分类思想不同，并且基模型没有进行过度拟合，属于弱分类器，符合集成学习算法中对基分类器的要求。因此本文选择用这四种方法的模型作为投票人。

委员会机器的算法设计如下：

For each stock in list：

1) 清洗数据，去掉缺失行，留出第一行数据；
2) 划分样本空间，以1/3的数据作为预测集，2/3的数据作为训练集；
3) 分别进行QDA、Bagging、SVM、ANN四个模型的训练；
4) 将训练模型运用在测试集上，并记录四维错误向量（错误向量定义成 mean(prdict!= test$rtn)）；
5) 用训练模型预测第一行，记预测结果为r $= \frac{\sum_{i=1}^{4} flag_i(1-err[i])}{4(4-sum(err))}$

   If r>0.5:

   预期最近一期结果为上涨；

   Else：

   预期最后一期结果为下跌；
6) 将预测结果与第一行因变量实际数据进行比较，若符合记i为1，不符合记为0。

整个算法的预测效果为$\frac{\sum i}{50}$。

各个模型参数设置如下表。

各股票的错误向量部分如下表，详见附录4。

用错误向量为权重，来预测最近一期的运动趋势，50只股票有25只预测成功，预测成功率为50%，预测效果不甚理想。

表5-6 关键参数设置表

Bagging	nbagg=200
	minsplit=2
	out of bag=True
SVM	cross=80
	kernel=”poly”
	gamma=1
	degree=2
ANN	size=5

表5-7 错误向量表（部分）

stock#	QDA-ERR	Bag-ERR	SVM-ERR	ANN-ERR
1	46.91%	43.21%	62.96%	56.17%
2	54.32%	48.15%	62.96%	48.77%
3	42.13%	56.18%	48.88%	43.26%
4	43.82%	62.36%	51.12%	48.88%
5	51.23%	58.64%	61.11%	51.85%
6	46.63%	66.85%	41.01%	64.04%
7	56.79%	48.15%	55.56%	54.32%
8	41.98%	44.44%	60.49%	55.56%
9	62.92%	46.63%	60.67%	41.57%
10	41.36%	59.88%	49.38%	54.94%
……	……	……	……	……
46	48.88%	52.81%	50.56%	45.51%
47	60.49%	43.21%	51.85%	54.94%
48	42.70%	44.38%	52.25%	64.61%
49	41.57%	41.01%	48.88%	59.55%
50	46.07%	41.01%	45.51%	43.82%
Mean	50.17%	51.22%	53.98%	54.04%

这个结果再一次说明了情绪指数预测效果不佳的结论。由于所选取的四个模型都有自适应的参数效果，模型本身的金融解释效果不强，对这一预测结果也不能得出有意义的金融意义，但是再次佐证了情绪指数不具有预测效力的结论。

六、结论与展望

本文通过逐步回归，得到了更符合中国金融市场的情绪代理变量，并经过主成分分析得到了长期情绪指数与短期情绪指数。在长期情绪指数与短期情绪指数运动趋势相一致的区间，根据方差贡献率加权而得到的综合情绪指数将有较好的趋势性。在运动趋势不一致的区间，综合情绪指数的运动趋势将不再明显。对此，本文选取其中的一个区间，对 89 篇社群媒体的文本信息进行了文本挖掘。在对形式的文本挖掘过程中，本文发现最大词频多为中性词，根据帖数计算的情绪因子为 0.63，也呈现出中性偏积极的情绪。在对文本内容挖掘的过程中，本文计算

出这些文本的积极情绪的分数略低于消极情绪的指数，但仍呈现出偏中性的情绪。这里得到的结论是在情绪指数运动趋势不明显的区间，文本分析也呈现出不明显的情绪成分。最后，本文对情绪指数的预测性进行了探索。首先本文考察了传统时间序列模型 VAR，分析了 VAR 模型预测效果不佳的原因。然后本文通过集成学习算法构造出了集成分类器，结果发现情绪指数的预测效果仍显不佳。这一结论与前人的研究基本一致。

本文最大的局限性是在数据挖掘过程中仍没能找到更具有预测性的模型，这启示以后的文章对情绪指数的预测性应当继续探索。此外，对文本的挖掘，本文是作为情绪指数运动趋势的补充与解释，有很多文章已经开始试图将情绪分数与收益率进行回归与相关分析。这启示今后的文章可以在选择好合理的文本数据后，直接对情绪指数进行挖掘，而不需要通过代理变量间接挖掘。最后本文的算法设计思路还可以更加复杂与精细化，如对情绪的打分算法设计还有提升的空间，这启示在算法与金融的交叉学科，还有很多值得挖掘的地方。

参考文献：

[1]靖荣华,贺晓玲.我国投资者情绪指数的构建[J].当代经济,2016(34):43-47.

[2]蒋文江,李彩雯,刘鹏懿.度量股票市场情绪指数的新方法——基于状态空间模型[J].海南师范大学学报(自然科学版),2016,29(03):242-248.

[3]蔡志刚,赖明明.中国股市投资者情绪指数构建与有效性检验[J].金融发展研究,2016(07):24-30.

[4]朱振,蒋文璐.余额宝情绪指数与中国股票市场间相互影响的实证分析[J].中国物价,2016(05):36-39.

[5]都星汉,李玥.我国股票投资者情绪指数构建的研究[J].经贸实践,2016(02):1-2.

[6]郭晓菲. 基于投资者情绪指数的上证综指预测研究[A]. 中国统计教育学会.2015 年（第四届）全国大学生统计建模大赛论文[C].中国统计教育学会:,2015:24.

[7]马若微,张娜.我国股票市场投资者情绪 SENT 指数的构建——基于上证 A 股公司的面板数据[J].中央财经大学学报,2015(07):42-49.

[8]段江娇,刘红忠,曾剑平.投资者情绪指数、分析师推荐指数与股指收益率的影响研究——基于我国东方财富网股吧论坛、新浪网分析师个股评级数据[J].上海金融,2014(11):60-64.

[9]林建雄. 投资者情绪与沪深 300 指数收益率研究[D].厦门大学,2014.

[10]程琬芸,林杰.社交媒体的投资者涨跌情绪与证券市场指数[J].管理科学,2013,26(05):111-119.

[11]鲁训法,黎建强.中国股市指数与投资者情绪指数的相互关系[J].系统工程理论与实践,2012,32(03):621-629.

[12]杨春鹏,淳于松涛,杨德平,姜伟.投资者情绪指数研究综述[J].青岛大学学报(自然科学版),2007(01):86-92.

[13]丁志国,苏治.投资者情绪、内在价值估计与证券价格波动——市场情绪指数假说[J].管理世界,2005(02):143-145.

[14]Aleksandra. The influence of investor sentiment on sector indices[R]. Poznan University:JEL Classification, 2017. 14-40

[15]Jun, Sik, Kim, Da-Hea, Kim. Investor sentiment and return of the option to stock volume ratio[J]. Financial Management, 2017, 3(3): 767-796

[16]Yuxi, Yang. Construction of investor sentiment index in the Chinese stock market[J]. Advanced Applied Informatics, 2017, 6(2): 23-28

[17]Steven, E, Sibley. The information content of the sentiment index[J]. Journal of Banking and Finance, 2015, 2(3): 164-179

[18]Patrick, Roger. The 99% market sentiment index[J]. Finance, 2014, 3(35): 54-96

[19]Baker, Wurgler. Investment sentiment and the cross-section of stock returns[J]. The Journal of Finance, 2006, 61(4): 1645-1680

[20]Baker, Wurgler. Investment sentiment and the stock market[J]. The Journal of Economic Perspectives, 2012, 2(104): 272-287

[21]Baker, Wurgler. Global,local and contagious investor sentiment[J]. Review of Financial Studies, 2008, 21(2): 785-818

[22]Stelios, Bekiros. A non-linear approach for prediction stock returns and volatility with the use of investor sentiment indices[J]. Applied Economics, 2016, 48(31): 2896-2899

[23]Laura, Yue, Liu. What is dividend premium[J]. Journal of Finance, 2001, 3(56): 2111-2133

致谢

在论文完成过程中，我真挚感谢我的导师刘良副教授的帮助。他督促了我们导师组所有成员完成论文，并在必要的时候提供了知识和技术上的指导。通过这次论文写作，我坚定了我研究生、博士生的研究方向与职业规划。我喜欢处理数据，并用数学模型去解决现实生活的问题。这次论文写作将是对我很好的激励。回首四年，白驹过隙。我感谢我在管理学院的每一位老师和在数学学院的每一位老师，是他们带我成熟与成长。最后我希望感谢我的父母，我希望我能通过学术上的成就来回报他们的养育之恩。

附录

附录 1：逐步回归算法关键代码

```
###mydata is the rtn of 50 stocks###
library(Hmisc)
for(i in 1:50){
  impute(y[,i],mean)
}##delete N/A
for(i in 1:50){
  mydata$s=y[,i]
  null=lm(s~rm+smb+umd+rmw+cma,data=mydata)
  full=lm(s~.,data=mydata)
  step(null,scope=list(upper=full,lower=null),
       direction="both",k=log(size(mydata)[2]))
}
```

附录 2：情绪打分关键代码

```
def sentiment_score_list(dataset):#dataset 即输入文本
    seg_sentence = dataset.split('。')
    count1 = []
    count2 = []
    for sen in seg_sentence: #循环遍历每一条语句
        segtmp = jieba.lcut(sen, cut_all=False) #把句子进行分词
        i=0#记录扫描词的位置
        a = 0 #记录情感词的位置
        poscount = 0 #积极词的第一次分值
        poscount2 = 0 #积极词反转后的分值
        poscount3 = 0 #积极词的最后分值
        negcount = 0
        negcount2 = 0
        negcount3 = 0
        for word in segtmp:
            if word in posdict:  # 判断词语是否是情感词
                poscount += 1
                c = 0
                for w in segtmp[a:i]:  # 扫描情感词前的程度词
                    if w in mostdict:
                        poscount *= 4.0
                    elif w in verydict:
                        poscount *= 3.0
                    elif w in moredict:
                        poscount *= 2.0
                    elif w in ishdict:
                        poscount *= 0.5
                    elif w in deny_word:
```

```
                    c += 1
            if judgeodd(c) == 'odd':   # 扫描情感词前的否定词数
                poscount *= -1.0
                poscount2 += poscount
                poscount = 0
                poscount3 = poscount + poscount2 + poscount3
                poscount2 = 0
            else:
                poscount3 = poscount + poscount2 + poscount3
                poscount = 0
            a = i + 1   # 情感词的位置变化

        elif word in negdict:   # 消极情感的分析，与上面一致
            negcount += 1
            d = 0
            for w in segtmp[a:i]:
                if w in mostdict:
                    negcount *= 4.0
                elif w in verydict:
                    negcount *= 3.0
                elif w in moredict:
                    negcount *= 2.0
                elif w in ishdict:
                    negcount *= 0.5
                elif w in degree_word:
                    d += 1
            if judgeodd(d) == 'odd':
                negcount *= -1.0
                negcount2 += negcount
                negcount = 0
                negcount3 = negcount + negcount2 + negcount3
                negcount2 = 0
            else:
                negcount3 = negcount + negcount2 + negcount3
                negcount = 0
            a = i + 1
        elif word == '！' or word == '!':   ##判断句子是否有感叹号
            for w2 in segtmp[::-1]:   # 扫描感叹号前的情感词
                if w2 in posdict or negdict:
                    poscount3 += 2
                    negcount3 += 2
                    break
        i += 1 # 扫描词位置前移
return count2
```

```
def sentiment_score(senti_score_list):
    score = []
    for review in senti_score_list:
        score_array = np.array(review)
        Pos = np.sum(score_array[:, 0])
        Neg = np.sum(score_array[:, 1])
        AvgPos = np.mean(score_array[:, 0])
        AvgPos = float('%.1f'%AvgPos)
        AvgNeg = np.mean(score_array[:, 1])
        AvgNeg = float('%.1f'%AvgNeg)
        StdPos = np.std(score_array[:, 0])
        StdPos = float('%.1f'%StdPos)
        StdNeg = np.std(score_array[:, 1])
        StdNeg = float('%.1f'%StdNeg)
        score.append([Pos, Neg, AvgPos, AvgNeg, StdPos, StdNeg])
    return score
```

附录 3：情绪得分完整表

Piece	Positive	Negative
1	73.18	222.62
2	58	226.58
3	70.03	228.83
4	69.88	209.12
5	186.42	116.08
6	228.41	97.15
7	63.33	231.62
8	65.92	290.31
9	74.89	193.7
10	60.11	266.86
11	81.68	238.33
12	222.98	67.38
13	65.96	279.02
14	267.05	113.45
15	55.63	225.8
16	188.43	104.87
17	226.16	105.4
18	238.02	115.19
19	156.26	134.07
20	155.01	104.33
21	75.91	252.09
22	46.07	214.85
23	187.62	114.13

24	181.29	99.38
25	216.87	100.43
26	246.61	111.14
27	67.28	241.4
28	211.66	108.11
29	183.59	102.08
30	199.37	107.62
31	176.63	75.97
32	66.41	227.39
33	67.64	228.77
34	200.42	95.86
35	173.56	80.2
36	179.67	79
37	192.91	91.76
38	64.74	289.85
39	219.4	110.64
40	125.56	110.75
41	209.37	105.27
42	152.36	112.09
43	67.22	220.02
44	60.21	195.74
45	162.25	104.88
46	61.68	328.06
47	135.63	82.98
48	51.55	274.12
49	229.73	112.39
50	107.62	122.57
51	217.54	110.78
52	64.63	299.7
53	69.43	257.58
54	62.56	246.07
55	56.64	255.51
56	216.24	114.83
57	50.81	246.65
58	43.43	266.24
59	185.08	107.95
60	58.61	305.78
61	56.04	242.37
62	50.62	297.73
63	186.43	93.19
64	190.11	77.02
65	241.71	128.17
66	186.76	66.93

67	46.71	288.27
68	184.84	80.44
69	253.81	104.81
70	192.05	85.06
71	182.59	131.23
72	334.08	99.8
73	44.84	277.68
74	39.33	284.31
75	189.6	102.87
76	43.28	238.52
77	144.24	110.12
78	59.7	286.52
79	257.05	90.51
80	174.77	59.52
81	190.38	89.62
82	52.17	371.61
83	63.98	310.35
84	238.92	130.06
85	165.72	75.13
86	41.21	266.27
87	225.02	97.34
88	66.89	231.47
89	232.45	82.25
mean	138.07	169.44

附录 4：错误向量完整表

stock#	QDA-ERR	Bag-ERR	SVM-ERR	ANN-ERR
1	46.91%	43.21%	62.96%	56.17%
2	54.32%	48.15%	62.96%	48.77%
3	42.13%	56.18%	48.88%	43.26%
4	43.82%	62.36%	51.12%	48.88%
5	51.23%	58.64%	61.11%	51.85%
6	46.63%	66.85%	41.01%	64.04%
7	56.79%	48.15%	55.56%	54.32%
8	41.98%	44.44%	60.49%	55.56%
9	62.92%	46.63%	60.67%	41.57%
10	41.36%	59.88%	49.38%	54.94%
11	54.32%	48.15%	47.53%	53.70%
12	58.43%	46.63%	51.12%	46.07%
13	50.00%	56.18%	65.17%	63.48%
14	42.70%	55.06%	56.74%	52.81%
15	62.36%	44.94%	59.55%	65.73%
16	47.19%	41.01%	50.00%	60.11%

17	41.36%	61.73%	60.49%	58.64%
18	62.35%	49.38%	50.00%	49.38%
19	47.53%	58.64%	59.26%	61.73%
20	48.77%	57.41%	50.00%	58.02%
21	45.68%	46.91%	57.41%	54.32%
22	57.30%	44.94%	49.44%	50.00%
23	46.91%	49.38%	51.85%	53.09%
24	51.69%	61.24%	54.49%	58.99%
25	43.21%	59.26%	59.88%	48.77%
26	50.00%	46.30%	57.41%	58.02%
27	57.87%	42.13%	53.93%	50.00%
28	58.64%	42.59%	54.32%	63.58%
29	42.13%	63.48%	60.11%	65.73%
30	55.62%	47.19%	65.73%	43.26%
31	61.11%	50.62%	47.53%	54.32%
32	48.31%	45.51%	48.31%	44.94%
33	56.17%	53.09%	54.94%	58.64%
34	48.77%	43.83%	48.15%	50.00%
35	43.83%	44.44%	50.62%	50.00%
36	59.55%	55.62%	57.30%	44.38%
37	50.62%	46.91%	51.23%	51.85%
38	41.01%	53.93%	44.38%	67.42%
39	61.11%	58.64%	56.79%	50.00%
40	57.30%	43.82%	60.11%	51.12%
41	42.59%	46.30%	52.47%	61.11%
42	43.21%	57.41%	59.26%	61.73%
43	57.87%	62.92%	48.31%	53.93%
44	42.59%	58.02%	50.00%	51.85%
45	42.59%	60.49%	51.85%	47.53%
46	48.88%	52.81%	50.56%	45.51%
47	60.49%	43.21%	51.85%	54.94%
48	42.70%	44.38%	52.25%	64.61%
49	41.57%	41.01%	48.88%	59.55%
50	46.07%	41.01%	45.51%	43.82%
mean	50.17%	51.22%	53.98%	54.04%