doi:10.19920/j.cnki.jmsc.2021.05.002

语调、情绪及市场影响:基于金融情绪词典①

姚加权[1],冯 绪[2*],王赞钧[3],纪荣嵘[4],张 维[2]

(1. 暨南大学管理学院,广州 510632;2. 天津大学管理与经济学部,天津 300072;
3. 厦门大学经济学院,厦门 361005;4. 厦门大学信息科学与技术学院,厦门 361005)

摘要:金融文本的语调与情绪含有上市公司管理层以及个体投资者表达的情感信息,并对股票市场产生影响。通过词典重组和深度学习算法构建了适用于正式文本与非正式文本的金融领域中文情绪词典,并基于词典构建了上市公司的年报语调和社交媒体情绪指标.构建的年报语调指标和社交媒体情绪指标能有效地预测上市公司股票的收益率、成交量、波动率和非预期盈余等市场因素,并优于基于其他广泛使用情绪词典构建的指标.此外,年报语调指标和社交媒体情绪指标对上市公司的股价崩盘风险具有显著的预测作用.为文本大数据在金融市场的应用提供了分析工具,也为大数据时代的金融市场预测和监管等活动提供了决策支持.

关键词:情绪词典;语调;投资者情绪;市场影响

中图分类号:F830.91 **文献标识码**:A **文章编号**:1007-9807(2021)05-0026-21

0 引 言

大数据时代,越来越多的金融领域研究关注了上市公司年报、新闻媒体报道和投资者社交媒体发帖等文本中所包含的语调与情绪[1].语调和情绪是上市公司管理层以及个体投资者情感和心理活动的外在表达[2,3],并且能预测股票的收益、波动以及成交量等重要的市场指标.对金融文本的语调和情绪的研究,有助于为大数据时代背景下金融市场预测和监管等活动提供决策支持,并为习近平总书记强调的"实施国家大数据战略、建设数字中国"的发展目标服务.

构建语调和情绪指标的关键是情绪词典.Loughran 和 McDonald[4] 基于人工筛选提出了适用于上市公司英文年报语调分析的词典,并发现基于此词典构建的年报语调能显著预测年报发布后股票的收益率、成交量和波动率.该词典在英文文本语调研究中得到广泛的应用,如:Feldman 等[5]、Liu 和 McConnell[6]、Rogers 等[7] 等研究均使用了 Loughran 和 McDonald[4] 词典,分别构建了年报、新闻报道、公司公告等文本的语调指标,并发现了这些语调对股价的预测能力.另外有一部分学者通过机器学习方法形成了情绪指标.Antweiler和 Frank[3] 选用 Yahoo! 财经论坛中的 1 000条样本发帖进行人工情绪分类,通过机器学习方法计算训练样本中频繁出现的词汇在情感表达上的先验概率,从而通过朴素贝叶斯学习方法实现对非训练样本发帖的情绪提取,并构建了情绪指标.中文社交媒体方面,Chang 等[8] 和部慧等[9] 选用东方财富股吧中的发帖,同样先人工判断生成训练集,再用机器学习模型对股吧发帖进行情绪分类.

尽管金融文本的语调和情绪相关工作已经开展,但现有的词典还存在一些不足,有可能影响语调和情绪指标的有效构建.首先,现有研究多采用人工判别方法构建基于小样本的情绪词典.这种

① 收稿日期:2019-12-23;修订日期:2021-03-14.
基金项目:国家自然科学基金资助项目(71790594;71871157;71502152);国家社会科学基金重大资助项目(18ZDA092).
通信作者:冯 绪(1984—),男,天津人,副教授,硕士生导师. Email:fengxu@tju.edu.cn

方法构建的语调和情绪指标可能导致样本筛选和判断标准不统一，研究结果无法复制等问题. 其次，在英文情绪词典研究中，Loughran 和 McDonald[4] 的词典具有较高的权威性. 但对于中文情绪词典研究来说，英文年报与中文年报在用词、表达方式等方面有很大的差异，不能简单地将英文年报情绪词典经过翻译后套用在中文年报的分析上，如谢德仁和林乐[10]、汪昌云和武佳薇[11] 等均使用了 Loughran 和 McDonald[4] 的词汇列表，但经过了手工筛选，并进行了适用于中文的用词习惯和语境的翻译工作. 最后，还存在一些针对中文文本分析的通用型词典，如：大连理工大学情感词汇本体库、中国知网词库（HowNet）和清华大学褒贬意词典等，他们的构成基于文学作品、媒体报道等，在金融领域研究的适用性和准确性还存在疑问. 以中国沪深股票市场为例，“庄”字在其他领域并不具备特殊的情绪和语调，但是在股票论坛上则含有强烈的负面情感. 综合来看，构建基于大样本、标准化的金融领域中文情绪词典是目前金融文本语调和情绪相关研究中亟待解决的问题.

基于以上考虑，本文构建了专门针对金融领域的中文情绪词典，并发现基于词典构建的年报语调和社交媒体情绪指标能有效地预测股票收益率、波动率、成交量等因素的变化. 具体而言，在年报文本方面，本文利用词典重组方法，在现有广泛使用词典的基础上提炼和构建了适用于金融领域正式文本研究的情绪词典. 利用 2003 年 ~2015 年间所有中国上市公司年报文本（共计 19 970 份），结合 Engelberg 等[12] 的语调判断方法区分单个年报的正负面情绪. 对 3 份现有通用型中文情绪词典和 Loughran 和 McDonald[4] 情绪词典的中文翻译版进行词语整合，并加入年报语料的分词结果去重得到初始词典，然后运用带惩罚机制词频法提取情绪词生成正式用语情绪词典. 研究发现，根据正式用语情绪词典构建的负面语调指标与上市公司发布年报后的股票交易量、波动率及下季度未预期盈余显著正相关. 除年报文本之外，社交媒体文本以非正式用语为主，正式用语情绪词典可能不再适用，因此也构建了适用于金融领域非正式文本研究的情绪词典. 利用 2011 年 ~2016 年间雪球论坛用户发帖以及 2010 年 ~2017 年间东方财富网股吧发帖（共计 8 130 多万条发帖），以 8 789 条带有情绪识别符号的股票论坛发帖为训练集，结合长短期记忆（long short-term memory，LSTM）模型的深度学习算法，并运用带惩罚机制词频法生成了非正式用语情绪词典. 基于非正式用语情绪词典构建的社交媒体情绪指标则能显著预测未来股票的超额收益、波动率和成交量. 这些结论说明本文提出的两个情绪词典在提取年报语调和社交媒体投资者情绪上是有效的.

此外，也将所构建的语调指标和情绪指标与基于大连理工大学情感词汇本体库、中国知网词库和清华大学褒贬意词典 3 个通用型情绪词典②以及 Loughran 和 McDonald[4] 词典中文翻译版所构建的语调指标和情绪指标进行比对，发现现有广泛使用情绪词典③在构建金融文本语调和情绪时并不能达到理想的效果，所构建的指标仅与少量市场指标存在显著关系. 而基于非正式用语情绪词典构建的年报文本语调指标和基于正式用语情绪词典构建的社交媒体情绪指标也没有很好的预测作用，这说明正式金融文本和非正式金融文本需要用不同的词典进行分析，故构建两个不同的词典是有必要的. 最后本文给出了所构建词典的实际应用场景，用年报语调指标和社交媒体情绪指标分别对个股的股价崩盘风险进行预测，发现两个指标均对股价崩盘风险有较好的预测作用，说明所构建的语调和情绪指标很好地表征了上市公司管理层以及个体投资者的情感和心理活动.

本文的贡献在 3 个方面：1）在研究工具上，基于大样本和机器学习方法构建了两个分别适用于金融领域正式文本和非正式文本研究的情绪词典，为金融领域的中文文本分析提供了简易的语

② 大连理工大学情感词汇本体库：http://ir. dlut. edu. cn/EmotionOntologyDownload；HowNet 情感分析用词语集：http://www. keenage. com/html/c_index. html；清华大学褒贬意词典：http://nlp. csai. tsinghua. edu. cn/site2/index. php/resources/13 - v10

③ 本文将现有的 3 个通用型情绪词典（大连理工大学情感词汇本体库、中国知网词库和清华大学褒贬意词典）以及 Loughran 和 McDonald[4] 词典的中文翻译版本合称为“广泛使用情绪词典”.

调和情绪提取工具,有助于实现未来相关研究结果的可重复性④. 此外本文的词典构建方法避免了人工判断和筛选情绪词所造成的偏差,词典构建方法本身也具有可重复性⑤;2)在经济理论上,本文基于情绪词典构建了中国市场的年报语调指标和社交媒体情绪指标. 一方面,发现语调和情绪指标对股价的崩盘风险有很强的预测作用,这一发现在前人研究中未曾涉及过. 这一结果为股价崩盘风险的相关研究增添了新的影响因素,有助于加深对金融风险形成机理的理解,促进资本市场健康发展. 另一方面,发现语调和情绪指标对股票的超额收益、波动率、成交量等因素有显著的预测作用. 这些结果支持了已有的信息过度反应和投资者情绪的相关理论,为这些理论补充了中国沪深股票市场的证据;3)在构建指标上,对现有广泛使用的情绪词典在金融领域的适用性进行了测度,发现基于这些词典构建的年报语调和社交媒体情绪指标不能准确表征公司管理层以及个体投资者的情感,这些结果为本文基于金融领域专用情绪词典构建语调和情绪指标的必要性提供了支持,并为今后的相关研究提供了借鉴.

1 文献回顾和研究假设

1.1 情绪词典相关研究

英文文本情绪词典在国外已有多部,其中Loughran 和 McDonald[4]发布的词典最有影响力,他们从上市公司年报中提取了高频词,通过人工筛选方式制作了年报语调词典(LM 词典). 在 LM 词典建立之前,常用的还有 Henry[13] 词典、Harvard-IV-4词典和 Diction 词典. Tetlock[14] 采用 Harvard-IV-4 词典提取了新闻报道语调,并和股票市场相关联. Loughran 和 McDonald[15]在文献综述中指出上述 3 部词典存在的缺陷:Henry[13] 词典存在词数少、缺失常用词的问题;Harvard-IV-4词典中 75% 的负面词在金融文档中并不具备负面含义;Diction 词典则是存在词语错误分类的问题. 相较而言,LM 词典更为完备,更适用于金融研究,也因此得到广泛的运用,Kearney 和 Liu[16] 甚至指出 LM 词典在近期的研究中占据了主导地位.

除公司年报之外,很多学者也将 LM 词典运用于分析新闻语调上. Dougal 等[17] 使用 LM 词典研究《华尔街日报》中“与市场同步专栏”语调. Garcia[18] 通过 LM 词典分析了《纽约时报》的财经专栏语调. Liu 和 McConnell[6] 则是将《华尔街日报》、《纽约时报》以及《道琼斯新闻》3 家媒体的报道内容结合起来分析语调. 此外 Solomon 等[19] 还研究了《华尔街日报》、《纽约时报》、《华盛顿邮报》和《今日美国》4 家媒体的新闻语调对共同基金的买入量的影响.

随着互联网的普及,大量研究转向 Yahoo! Finance,SeekingAlpha,Twitter 等社交媒体,讨论社交媒体中的投资者情绪. 由于社交媒体所使用语言的非正式性,此类研究大多构造小的训练集[3, 20],然后通过传统的机器学习方法判断文本情绪. 此外,Chen 等[21] 通过使用 LM 词典分析了 SeekingAlpha 论坛中投资者发帖的语调. 而针对社交媒体文本的权威金融情绪词典尚未出现,这一研究方向还有很大的拓展空间.

目前在金融领域的中文文本分析研究中,权威的中文情绪词典还没有出现. 一部分研究通过传统机器学习的方法来判断情绪[8, 22]. 也有一些研究将 LM 的词典本土化,通过结合语境翻译和人工筛选方式制作中文情绪词典[10-11]. 在缺乏权威情绪词典条件下,还有很多研究通过人工大量阅读方式判断文本情绪[23]. 在已经公布的通用型情绪词典方面,中文文本分析经常使用的情绪词典主要包括大连理工大学信息检索研究室、中国知网和清华大学自然语言处理与社会人文计算实验室所提出的 3 部词典. 其中大连理工大学信息检索研究室的情感词汇本体库以大量情感语料为基础,采用手工情感分类和自动获取强度两种方法,从数据中提取情感信息,该情感词典含有 7 773个负面词与 8 610 个正面词. 知网情感分析用词语集发布于 2007 年,含有 1 254 个中文负面

④ 本文词典对外公开,网址: https://github.com/dictionaries2020/SentimentDictionaries.

⑤ Loughran 和 McDonald[4]词典也是基于人工筛选方法构建而成. 与该词典相比,本文采用词典构建的方法避免了这一问题.

词与 836 个中文正面词. 清华大学褒贬意词典则含有 4 468 个负面词与 5 567 个正面词. 但这 3 部词典是否适用于金融领域的文本分析还尚未有研究给出详尽的测试. 唐国豪等[24]指出，在分析不同领域文本内容时要选择适合的分类词典，对现有情感词典的细分、改进和升级是十分必要的.

1.2　年报语调相关研究

上市公司公布的年报通常被认为含有丰富的信息，年报的语调（特别是负面语调）经常被认为是上市公司管理层对公司预期的表达. 管理层通过更悲观的语调来降低投资者对于未来公司经营表现的预期，管理层语调也影响了公司股票的多项市场指标. Loughran 和 McDonald[4]提取了上市公司年报和其中的管理层讨论与分析（MD&A）部分的正负面语调，发现年报中的负面语调与年报发布后几个交易日的股票超额报酬率、异常交易量、收益波动率以及未预期盈余显著相关. Feldman 等[5]使用 LM 词典来研究年报和季报中管理层讨论与分析板块语调的变化对金融市场造成的影响，发现更为正面的表述将伴随着较高的股票收益率.

综合来看，现有研究普遍认为年报语调与公司的盈余预期相关，可以预测发布日后公司股票交易特征. 以最有代表性的 Loughran 和 McDonald[4]的研究为例，发现年报语调与公司股票的超额收益、交易量、收益波动率显著相关. 这其中的原因在于，投资者可以感受到年报的语调. 当年报越悲观时，将影响投资者对该公司的未来预期，产生过度卖出股票的行为. 且投资者对此信息的反应过度，经常以较低的市场价格卖出股票. 此外本期悲观语调也会导致公司在未来产生更高的未预期盈余，因此悲观语调也与下季度的未预期盈余显著相关. 在金融英文文本分析领域，已有 LM 词典可以便利地进行情感分析. 然而，金融中文文本分析领域缺乏这样行之有效的词典. 本文认为应针对中文年报文本来选择合适的词典对其进行情绪的量化，以达到足够的解释能力，据此提出如下假设：

假设 1　针对年报文本，基于中文正式用语情绪词典构建的年报负面语调指标与年报发布后上市公司股票的交易量、波动率及下季度未预期盈余显著正相关；而基于广泛使用的情绪词典和非正式用语情绪词典构造的年报语调指标无法完全解释股票的交易量、波动率及下季度未预期盈余⑥.

1.3　社交媒体情绪指标

社交媒体的出现为投资者发布信息、观点和投资策略提供了渠道. 社交媒体中的发帖是投资者表达情绪的主要途径，因此也会对市场指标产生影响⑦. 虽然在早期的研究中，Tumarkin 和 Whitelaw[26]和 Das 和 Chen[20]发现股吧论坛信息不会对收益率产生影响，但是 Antweiler 和 Frank[3]使用朴素贝叶斯算法从雅虎财经（Yahoo! finance）上的 150 万条发帖中提取看多和看空情绪之差作为看涨情绪指标，结果发现看涨情绪指标与股票收益率的正相关在统计学意义下显著，而投资者情绪一致性指标与股票的成交量及波动率显著负相关. 随着互联网普及和投资者使用社交媒体发布信息的增加，越来越多的研究发现社交媒体中投资者情绪显著影响了资产价格和其他交易指标. Bollen 等[27]分析了 Twitter 上的发帖并提取了公众情绪，公众情绪与道琼斯指数的日收益率有显著正相关关系. Chen 等[21]分析SeekingAlpha 网站中投资者发帖数据，发现投资者的发帖中含有私有信息，可以显著预测未来股价，而情绪一致性指标也可以显著预测未来股票的波动率和成交量. 段江娇等[28]对中国沪深股票市场东方财富股吧论坛发帖进行了情绪提取，发现东方财富股吧的发帖具有一定的信息含量，股票日收益率与当日论坛情绪显著正相关. 杨晓兰等[22]也分析了新浪财经博客中投资者情绪对对股市政策效应的影响.

综合以上研究可以发现，现有文献在社交媒体投资者情绪相关研究中已经取得了部分共识. 社交媒体发帖是投资者情绪的有效表达，由此生

⑥　由于本文根据 Engelberg 等[12]的方法使用超额收益率作为判断年报正负语调的依据，因此基于正式用语情绪词典构建的年报语调指标天然地满足语调和超额收益率的关系，故在此省略了针对超额收益率的假设检验.

⑦　关于社交媒体和其他类型互联网大数据对资产定价的影响，张学勇和吴雨玲[25]给出了很好的综述.

成的看涨情绪指标与收益率显著正相关,且投资者情绪一致性指标与交易量及波动率显著负相关,即投资者情绪越不一致,交易量及波动率越大. 据此,本文提出如下假设:

假设 2 在股票论坛中,基于中文非正式用语情绪词典构建的投资者看涨情绪指标与股票超额收益显著正相关,构建的情绪一致性指标与交易量及波动率显著负相关. 而基于广泛使用的情绪词典和正式用语情绪词典构造的投资者情绪指标无法满足以上关系.

1.4 股票价格崩盘风险相关研究

为了检验基于词典构建的年报语调指标和社交媒体情绪指标,将这两个指标和股价崩盘风险相关联,考察年报语调指标和社交媒体情绪指标对上市公司股价崩盘风险的预测作用. 从管理层角度来看,股价崩盘的根源是管理层的捂盘行为[29, 30]. 由于管理层和投资者之间存在委托代理冲突. 管理层出于自身薪酬、职业生涯、建立帝国以及晋升等考虑,在信息披露中经常会报喜不报忧,如果好消息和坏消息均随机出现,并且管理者均及时披露两类消息,即消息分布是对称的,则股票回报的分布也对称. 然而,大量研究表明,管理者披露坏消息和好消息分布并不对称. 管理层存在捂盘坏消息的行为倾向,即管理者更倾向于隐瞒或推迟披露坏消息而加速披露好消息,坏消息随时间的推移在公司内部不断积累,由于公司对坏消息的容纳存在上限,一旦累积的负面消息超过了这个上限,坏消息将集中释放出来,进而对公司股价造成极大的负面冲击并最终崩盘. 基于以上原因,公司年报的负面语调指标应该与未来股价崩盘风险负相关,即负面语调越弱(正面语调越强),公司未来股价崩盘风险越大.

从投资者角度来看,投资者在公司负面信息集中释放之前可能通过一些渠道(投资者可能有自身的私有信息渠道[31]. 即便投资者是非知情者,通过对其他投资者公开交易的学习,也可以从中推断出部分知情者知晓的信息[32])提前得知了公司的负面消息. 在无摩擦市场中,投资者可以通过卖空股票,使悲观预期提前进入股票价格,从而降低股价的崩盘风险[33]. 但由于中国市场的卖空机制并不完善,卖空交易的总量很小,主要参与者为机构投资者,个体投资者很少参与到这一过程中. 此外由于制度设计缺陷,本身股价崩盘风险大的股票没有进入融券名单,这导致卖空机制很难发挥作用,甚至加剧了崩盘风险的上升[34]. 因此在中国沪深股票市场,投资者即便提前知晓了公司的负面消息,其悲观预期更多地是通过论坛和社交媒体而非市场交易来表达. 基于以上原因,社交媒体中投资者的负面情绪应该与未来股价崩盘风险正相关,即负面情绪越强(正面情绪越弱),公司未来股价崩盘风险越大. 综合以上两点,本文提出如下假设:

假设 3 构建的负面年报语调指标与股票未来的股价崩盘风险显著负相关,且构建的社交媒体看跌情绪指标与股票未来的股价崩盘风险显著正相关.

2 语调和情绪指标构建

构建金融文本语调和情绪指标的关键是情绪词典,故首先构建用于年报语调分析的正式用语情绪词典和用于社交媒体投资者情绪分析的非正式用语情绪词典. 词典的构建思路如下:在正式用语情绪词典构建方面,参照 Engelberg 等[12]的方法选取年报发布[0, +3]日累积超额收益率作为判断正负面年报的依据(Engelberg 等[12]认为该方法可以有效地避免人工判断文本信息所带来的偏差). 然后采用词典重组法,综合大连理工大学情感词汇本体库、中国知网词库、清华大学褒贬意词典和 Loughran 和 McDonald[4] 词典中文翻译版⑧4 部词典,并加入年报语料的分词结果去重得到初始词典,然后使用带惩罚机制词频法提取情绪词生成正式用语情绪词典,最后基于情绪词典采用词袋模型构建年报语调指标. 在非正式用语情绪词典构建方面,考虑到社交媒体中普遍使

⑧ 本文两位作者协同 3 位会计学专业的研究生共同对该词典进行了中文翻译. 因英文单词可能对应多个中文释义,因此在翻译过程中要求翻译者先列出所有认为正确的译法,再对这 3 份各自独立完成的翻译稿综合考虑后得到 Loughran 和 McDonald[4] 词典的中文翻译. 其中包括 2 337 个负面词和 353 个正面词.

用表情符号(如 Emoji 等)来表达发帖者的情绪，筛选出了使用表情符号并明确表达了情绪的发帖作为训练样本⑨，以此来排除人工分类社交媒体发帖带来的偏差. 然后采用长短期记忆网络模型的深度学习算法分析股票论坛上的帖子情绪⑩，并使用带惩罚机制词频法提取情绪词典. 同样基于词典采用词袋模型构建社交媒体情绪指标. 最终形成的正式用语情绪词典含有 1 633 个负面词和 3 592 个正面词，非正式用语情绪词典则含有 965 个负面词和 912 个正面词. 附录表 1 列出正式用语情绪词典和非正式用语情绪词典正负面各 30 个高频词语. 可以发现，与正式用语情绪词典中的词相比，非正式用语情绪词典较口语化且情绪更加突出. 具体构建细节如下.

2.1　正式用语情绪词典与年报语调指标构建

上市公司年报中所披露的信息其语调相对于社交媒体往往更加隐晦. 下面给出了年报文本样例，其中标黑单词是情绪表达的关键词，而本文的词典构建方法就是将标黑的单词提取出来. 样例如下：

挑战方面：一是茅台酒市场拓展还需进一步加强；二是国内中、低端白酒市场竞争异常激烈，酱香系列酒市场竞争力不强，对公司业绩的贡献度有待提高. 风险方面：一是宏观经济下行压力加大；二是赤水河流域生态环境保护压力增大；三是打假保知任重道远(来源：贵州茅台酒股份有限公司 2015 年年度报告).

本文首先依据年报发布[0，+3]日的累计正负收益率，将年报分为正负面情绪两类，样本数据涵盖 2003 年～2015 年共 19 970 份年报，其中累计收益率为正的有 12 475 份，累计收益率为负的有 7 383 份，累计收益率为零(停牌)的有 112 份，其中收益为零的年报不进入提取情绪词典的数据样本中. 然后综合大连理工大学情感词汇本体库、中国知网词库、清华大学褒贬意词典和 Loughran 和 McDonald[4] 词典中文翻译版，筛选出现在年报中的各词典所包含的词，并加入使用中科院汉语词法分析系统(ICTCLAS)⑪对年报语料的词汇切分结果去重作为初始词典. 最后依照带惩罚机制词频法计算初始词典中待选负面词调整后的词频数值，如下式所示

$$adjusted\ frequency_n = \frac{w_{n,N}}{\sum_n w_{n,N}} \times \frac{1}{1 + \frac{w_{n,P}}{\sum_n w_{n,P}}} \quad (1)$$

式中 $w_{n,N}$ 为待选负面词 n 在收益为负的年报集合 N 中的出现次数；$w_{n,P}$ 为待选负面词在收益为正的年报集合 P 中出现的次数. 待选负面情绪词的频度随着其在负面年报集合出现的比重 $\left(\frac{w_{n,N}}{\sum_n w_{n,N}}\right)$ 增加，同时也随着其在正面年报集合出现的比重 $\left(\frac{w_{n,P}}{\sum_n w_{n,P}}\right)$ 下降. 通过惩罚机制 $\left(\frac{1}{1+\frac{w_{n,P}}{\sum_n w_{n,P}}}\right)$ 的引入，带惩罚机制词频法从总体层面上衡量了待选词的负面程度. 根据带惩罚机制词频法计算得到的数值排序并生成正式用语情绪词典的负面词表. 基于同样的逻辑，可以得到正式用语情绪词典的正面词表⑫.

基于正式用语情绪词典，使用词袋模型(*bag-of-words*)构建年报的负面语调指标，命名为：正式用语词典语调(*Index*)，计算如下式

$$\begin{aligned} Index_i &= Negative_i - Positive_i, \\ Negative_i &= \frac{\sum_n w_{n,i}}{total\ word_i}, \\ Positive_i &= \frac{\sum_p w_{p,i}}{total\ word_i} \end{aligned} \quad (2)$$

式中 $w_{n,i}$ 为负面词 n 在年报 i 中出现的次数；$w_{p,i}$

⑨　标注训练集的过程中存在一个帖子有多个表情符号的情况，而多个表情符号会使得该帖子的情绪较为模糊，本文剔除了这种类型的帖子.

⑩　LSTM 的深度学习需要将文本分词后以向量形式进行编码，因此更适用于具有短文本特性的论坛发帖. 而年报属于长文本，难以有效地将整份年报编码并训练整个神经网络，故本文不使用 LSTM 来构造正式用语词典. 以 LSTM 为代表的深度学习算法在金融实证研究中已经有初步应用，详见苏治等[35].

⑪　中科院汉语词法分析系统(ICTCLAS)：http://ictclas.nlpir.org/. ICTCLAS 系统是当前中文分词领域准确率(高达 98%)表现优秀的系统.

⑫　为了得到正式用语情绪词典的正面词表，只需把式(1)中的符号 n，N，P 分别替换成 p，P，N.

为正面词 p 在年报 i 中出现的次数;$total\ word_i$ 为年报 i 的总词数.

2.2 非正式用语情绪词典与社交媒体情绪指标构建

使用2011 - 04 - 08 ~ 2016 - 04 - 22 所有关于中国上市公司在雪球论坛上的用户发帖,以及东方财富网股吧 2010 - 05 - 01 ~ 2017 - 09 - 30 所有关于上市公司的发帖,共计 8 130 多万条. 在网络股票论坛中,用户发表自己的意见并与其他用户交流,其中附带有明显表情符(Emoji) 的文本将使得帖子情绪显而易见. Derks 等[36] 也指出社交媒体用户一般会以表情符表达自身情感,因此明确的情绪识别符号可以作为用户发帖情绪的判别标准,并且避免由人工判断发帖情绪分类造成的偏差. 例如:

屋漏偏逢连夜雨,踩雷$ 中兴通讯(SZ000063)$,又缝本周大 A 癫痫躺地板,组合下跌 5% 😭(来源:雪球论坛).

本文首先通过用户在发帖或回帖时所加入的表情:[笑](😊)、[大笑](😂)、[哭泣](😭)和[怒了](😡)等明确的情绪识别符号,选出带有明显情绪特征的帖子作为样本⑬,并删除了内容含有广告等噪音的帖子,最终筛选出8 789 个股票论坛发帖作为深度学习算法的训练样本.

本文采用的深度学习算法为 LSTM 网络,是递归神经网络(recurrent neural network,RNN)的特殊变体. 传统 RNN 在处理时序高度相关问题时,因其每次循环更新参数时,都把预测误差前向和反向传播以得到对误差的修正,但同时也存在梯度消失(gradient vanishing) 和梯度爆炸(gradient exploding) 问题,即当梯度前向和反向传播来优化神经网络的参数时,梯度太小或太大(大于1) 导致传播过程中因多次相乘使得对误差的修正最终变为 0 或无限大,从而无法继续优化神经网络的参数. 因此,LSTM 在每一个神经网络中加入了 3 个控制门(gate):输入门(input gate)、输出门(output gate) 及遗忘门(forget gate). 输入门决定当前时序的新数据有多少要进入当前神经网络,输出门决定当前神经元有多少信息要进入下一个神经网络,遗忘门则决定上一个时序的信息有多少要进入当前神经网络,通过这 3 个门的控制,能够有效缓解梯度消失和梯度爆炸问题. 在处理时序高度相关的问题上,例如:机器翻译、对话生成和文本情绪分析等问题,LSTM 有着优异的表现. 此外,中文词汇在文本中所含有的情绪,往往会受到上下文的影响,使得整体句子的情绪表现有所不同,而 LSTM 能够很好地捕捉词与词之间的相依性. 文本中某个单词可能与前后第二至三个单词有关,但离前后第十个单词已没有太大的关系. 因此通过 LSTM 的 3 个控制门,适当地遗忘较远的信息和加强邻近的信息,能够使得模型更准确、更高效地捕捉上下文的信息. 该模型在本文数据分类中的准确率达到 87%⑭. 本文使用训练好的 LSTM 模型对两个论坛上每一个发帖的正负面情绪进行了分类. 将相同情绪的帖子归为一类并分词,计算每个词的带惩罚机制词频法的数值,根据此数值进行排序,最后生成非正式用语情绪词典.

在非正式用语情绪词典的基础上,参照 Antweiler 和 Frank[3] 的方法,构建了看涨情绪指标和情绪一致性指标,分别命名为非正式用语词典情绪(*Bullishness*) 和非正式用语词典情绪一致性(*Agreement*),计算如下式

$$Bullishness_{i,t} = \frac{Positive_{i,t} - Negative_{i,t}}{total\ word_{i,t}}, \quad (3)$$

$$Agreement_{i,t} = 1 - \sqrt{1 - Bullishness_{i,t}^2}$$

式中 $Positive_{i,t}$ 为公司 i 在 t 日的全部帖子中正面词占总词数的比例;$Negative_{i,t}$ 为公司 i 在 t 日的全部帖子中负面词占总词数的比例. $Bullishness_{i,t}$ 反映了股票 i 在 t 日的看涨情绪,而 $Agreement_{i,t}$ 的最大值是 1,即一致看涨或者一致看跌,最小值为 0,即情绪分歧度最大.

⑬ 为了确保帖子的情绪识别符号和文字表达相一致,本文还是对样本进行了人工筛查,删除了识别符表达情感不明确的样本.

⑭ 计算方法为:首先将有标注的帖子分成 10 等份,并用其中 9 份训练 LSTM. 用训练完成的 LSTM 预测剩下一份的情绪,并与标注结果对比,从而计算出准确率. 按以上步骤重复进行 10 次,然后计算 10 次准确率的平均值.

3　实证检验

3.1　年报负面语调指标的描述性统计和回归结果

检验年报负面语调指标所使用的数据均来自于 CSMAR 数据库. 按照假设 1，回归模型中所使用的解释和被解释变量、变量名及其相关含义或计算公式如表 1 所示.

表 1　变量名及其含义和计算公式说明

Table 1 Definition and calculation of variables

变量类型	变量名	含义或计算公式
被解释变量	超额收益率	年报发布后[0，+3] 日累积收益率减去同期市场累积收益率
	异常成交量	年报发布后[0，+3] 日平均交易量除以发布前[－65，－6] 日平均交易量取自然对数
	波动率	年报发布后的[＋6，＋252] 日对数收益率的波动率
	未预期盈余	年报发布后下一份季报的每股盈余减去上一年该季度季报的每股盈余
解释变量	语调／情绪	基于各词典构建的年报语调和投资者情绪指标
	市值	年报发布前一天的流通股市值，取自然对数
	换手率	年报发布前[－252，－6] 天的平均交易量除以发布当天的在外流通股数，取自然对数
控制变量	账面市值比	年报发布前一天的账面市值比，取自然对数
	机构投资者	年报发布前一季度的机构投资者比例
	前期异常收益	年报发布前一个月的 Fama-French 三因子模型 α
	预测分歧	股票分析师对该季度的每股盈余预测的标准差，除以该季度末股价
	预测修正	年报发布前股票分析师对于该季度的每股盈余预测的平均值的月度变动

表 2 对回归模型的变量进行了描述性统计，样本数据涵盖 2003 年～2015 年共 14 339 份年报[15]. 除了基于本文提出的正式用语情绪词典所构建的负面语调指标外，也基于其他词典参照式(2) 构建了负面语调指标，分别是大连理工大学词典语调、知网词典语调、清华大学词典语调、LM 词典语调、非正式用语词典语调，其中本文对 LM 情绪词典进行中文翻译并构建指标 LM 词典语调，而非正式用语词典语调是基于本文的非正式用语情绪词典所构造的年报情绪指标. 从表 2 的结果来看，大连理工大学词典语调总体上较为悲观且语调波动较大. 此外，控制变量中年报发布前一个月股票异常收益(前期异常收益) 的均值为负，表示在本文研究样本中，年报发布之前的公司股价存在负的异常收益.

表 3 为各年报负面语调指标相关系数表，其中本文提出的正式用语情绪词典形成的负面语调指标与其他词典构建的负面语调指标存在显著的正向关系(相关系数 0.171～0.422 之间). 正式用语词典语调和非正式用语词典语调之间的相关系数只有 0.171，表明正式用语词典和非正式用语词典囊括的词语差别较大. 此外，大连理工大学词典语调和清华大学词典语调的相关系数(0.804) 为最大，表明大连理工大学情感词汇本体库与清华大学褒贬意词典在捕捉年报语调的效果方面差异不大.

⑮　本文用于提取正式用语情绪词典的年报有 19 970 份，因被解释变量缺失的缘故，用于实证分析的年报样本只有 14 339 份.

表 2 年报语调回归相关变量的描述性统计

Table 2 Summary statistics of variables that are related to tone measures of annual filings

变量名	样本量	均值	标准差	最小值	最大值
异常成交量	14 244	0.102	0.339	−1.481	3.906
波动率	12 758	0.501	0.153	0.117	3.129
未预期盈余	13 902	−0.033	0.077	−1.130	0.050
正式用语词典语调	14 339	−0.001	0.002	−0.025	0.013
大连理工大学词典语调	14 339	−0.070	0.011	−0.135	−0.015
知网词典语调	14 339	−0.007	0.002	−0.029	0.001
清华大学词典语调	14 339	−0.029	0.009	−0.081	−0.001
LM 词典语调	14 339	−0.025	0.007	−0.036	0.007
非正式用语词典语调	14 339	−0.013	0.005	−0.048	0.019
市值	14 339	22.488	0.979	18.921	27.876
换手率	14 339	−3.792	1.181	−10.126	−0.634
账面市值比	14 339	−1.190	0.294	−1.536	−0.845
机构投资者	14 339	0.040	0.081	0	0.887
前期异常收益	14 339	−0.243	0.130	−0.697	1.195
预测分歧	14 339	0.017	0.025	0	0.448
预测修正	14 339	0.047	0.047	−0.264	0.533

表 3 各年报负面语调指标相关系数表

Table 3 Correlation coefficients of negative tone indexes with different dictionaries

变量名	正式用语词典语调	大连理工大学词典语调	知网词典语调	清华大学词典语调	LM 词典语调	非正式用语词典语调
正式用语词典语调	1.000					
大连理工大学词典语调	0.422**	1.000				
知网词典语调	0.238**	0.417**	1.000			
清华大学词典语调	0.391**	0.804**	0.245**	1.000		
LM 词典语调	0.406**	0.116**	0.132**	0.096**	1.000	
非正式用语词典语调	0.171**	0.108**	0.102**	0.094**	0.032**	1.000

注：** 表示在 5% 的水平下显著.

参照 Loughran 和 McDonald[4] 的实证方法及控制变量的选择，使用 Fama-MacBeth[37] 的两阶段回归来解决最小二乘回归存在的残差相关性，并经过 Newey-West 滞后一阶方法调整异方差和自相关. Fama-Macbeth 两阶段回归第一步是对于每一年的解释变量和被解释变量进行最小二乘回归，得到估计参数，将每个参数视为整体参数的样本值；第二步为对第一步的所有参数求平均值，计算整体数据的估计参数.

表 4 为各情绪词典构建的年报负面语调指标对交易量、波动率及下季度未预期盈余的回归结果⑯. 结果表明，正式用语情绪词典构建的年报负面语调与交易量、波动率及下季度未预期盈余显著正相关. 实证结果与 Loughran 和 McDonald[4] 中所述一致，即年报语调越悲观，未来的交易量与

⑯ 控制变量包括 ln *Size*（市值），ln *Turnover*（换手率），ln(*B/M*)（账面市值比），*Institution Own*（机构投资者），*Pre FF Alpha*（前期异常收益），*Analyst Dispersion*（预测分歧），*Analyst Revisions*（预测修正），各变量定义参见表 1. 回归控制了行业固定效应和时间固定效应. 因版面关系，控制变量的回归结果被省略.

波动率越大，同时，下季度未预期盈余也越大. 年报负面语调和交易量的回归结果表明年报中有更多的负面词，即越悲观，随后的交易量也越大. 年报负面语调与波动率显著正相关，即年报中有更多的负面词，在年报发布后的收益波动率也越大. 而年报负面语调与下季度未预期盈余的显著正相关关系表示管理层在年报中使用更多的负面词来降低外界对该公司的预期，从而在未来有更大的未预期盈余. 此外控制变量前期异常收益，预测分歧和预测修正与未预期盈余有显著关系，即公司过去有更好的表现、分析师预测分歧度越大与分析师预测改变程度越小时，未来的未预期盈余越大.

表 4　交易量、波动率和未预期盈余与各年报语调回归结果

Table 4 The regressions of negative tone indexes with different dictionaries on trading volume, volatility and unexpected earnings

解释变量	被解释变量		
	异常成交量	波动率	未预期盈余
正式用语词典语调	5.948 **	3.878 **	1.132 **
	(2.01)	(2.40)	(2.15)
大连理工大学词典语调	-0.046	0.931	0.246 **
	(-0.37)	(1.51)	(2.28)
知网词典语调	-1.987	-1.876	0.698
	(-1.27)	(-1.48)	(1.80)
清华大学词典语调	0.053	1.198	0.356 ***
	(0.24)	(1.60)	(2.95)
LM 词典语调	3.201 *	1.065	1.001
	(1.88)	(1.51)	(1.62)
非正式用语词典语调	0.559	0.693	-0.247
	(0.80)	(0.75)	(-1.68)

注：括号中是 t 值；* 表示 10% 显著，** 表示 5% 显著，*** 表示 1% 显著.

在表 4 中，其他 5 个词典所衡量的年报负面语调除了大连理工大学词典语调和清华大学词典语调与下季度未预期盈余显著正相关以及 LM 词典语调与交易量显著正相关之外，其余回归结果皆不显著. 非正式用语词典语调对于交易量、波动率及下季度未预期盈余的回归结果皆不显著. 在本文提出的正式用语情绪词典与非正式情绪用语词典中，以非正式用语情绪词典为例：单词“春天”被分类在正面词下，而单词“春天”如果在年报中出现，却不带有任何的情绪. 同样地，在 LM 中比较了 Harvard-IV-4 词典与该研究所提出的情绪词典，例如：单词“tax”在 Harvard-IV-4 的类别为负面词，但如果年报中出现“tax”则不带有任何情绪. 类似情况在本文构建的正式用语和非正式用语情绪词典中并不少见. 如上所述，回归结果也验证了不同类型的文本须使用相应的情绪词典，否则可能无法有效衡量文本语调或投资者情绪. 表 4 的回归结果满足本文的假设 1.

3.2　社交媒体情绪指标的描述性统计和回归结果

附录表 2 为检验非正式用语词典有效性所使用相关变量的描述性统计，本文使用 2011-04-08～2016-04-22 日在雪球论坛以及 2010-05-01～2017-09-30 在东方财富网股吧关于中国上市公司⑰的用户发帖，同样也基于其他词典构建了针对论坛文本的情绪指标与情绪一致性指标. 附录表 3 为通过各词典构建的看涨情绪与情绪一致性的相关系数表. 在看涨情绪指标上，非正式用语词典与大连理工大学情感词汇本体库、中国知网词典、清华大学褒贬意词典及 LM 情绪词典中文翻译版存在显著的负向关系，说明非正式用语情绪词典与其他 4 个词典在股票论坛文本数据的看涨情绪指标构建上有着很大的差异. 在情绪一致性指标上，非正式用语词典与其他词典存在显著的正向关系.

⑰　包括上证 A 股，深证 A 股主板和创业板股票.

表 5 为基于非正式用语词典的社交媒体投资者看涨情绪指标以及情绪一致性指标对股票超额收益、成交量和波动率的最小二乘法回归的实证结果⑱. 可以发现非正式用语词典情绪与超额收益、交易量及波动率显著正相关，这些结论与 Antweiler 和 Frank[3] 的结果一致，也与 Bollen 等[27] 和 Chen 等[21] 等文献的结论保持了一致性. 这一结果也符合段江娇等[28] 在东方财富网股票论坛的样本中发现的投资者看涨情绪与当日股票收益率显著正相关的现象. 此外非正式用语词典情绪一致性与超额收益、交易量及波动率显著负相关，说明投资者情绪越不一致，超额收益、交易量及波动率越高，这与 Antweiler 和 Frank[3]、Chen 等[21] 的结果相一致. 表 5 结果说明本文构建的非正式用语情绪词典在中文社交媒体文本中的应用是有效的.

表 5 超额收益、交易量及波动率与非正式词典构建的社交媒体情绪指标回归结果

Table 5 The regressions of sentiment index with informal dictionary on trading volume, volatility and unexpected earnings

解释变量	被解释变量					
	超额收益率		异常成交量		波动率	
非正式用语词典情绪	0.001 ** (2.17)		0.128 *** (2.44)		0.001 ** (2.14)	
非正式用语词典情绪一致性		−0.001 *** (−3.08)		−0.149 ** (−2.00)		−0.001 ** (−2.13)
市值	−0.001 *** (−6.81)	−0.001 *** (−7.22)	−0.040 ** (−2.08)	−0.042 ** (−2.16)	0.000 *** (6.52)	0.001 *** (6.85)
换手率	−0.001 *** (−5.02)	−0.001 *** (−4.86)	0.067 *** (3.40)	0.744 *** (3.29)	0.000 *** (8.70)	0.001 *** (9.06)
账面市值比	0.000 *** (2.69)	0.000 *** (3.04)	0.055 *** (2.51)	0.054 *** (2.43)	−0.000 *** (−5.64)	−0.001 *** (−5.98)
常数项	0.012 *** (5.08)	0.014 *** (5.64)	1.157 *** (2.68)	1.119 *** (2.58)	−0.004 *** (−3.01)	−0.004 *** (−3.09)
样本量	564 887	564 887	564 822	564 822	564 886	564 886
调整 R^2	0.013	0.013	0.006	0.006	0.102	0.103

注：括号中是 t 值；* 表示 10% 显著，** 表示 5% 显著，*** 表示 1% 显著.

表 6 为基于其他词典构建的投资者看涨情绪指标与投资者情绪一致性指标的回归结果，与表 5 的回归方法一致，因版面关系，本文省略了控制变量，仅列出关键解释变量的回归结果. 其中正式用语词典情绪、大连理工大学词典情绪、清华大学词典情绪、LM 词典情绪与交易量显著正相关，该结果和非正式用语情绪词典的看涨情绪与交易量的关系一致. 然而，正式用语词典情绪、知网词典情绪、大连理工大学词典情绪、清华大学词典情绪、LM 词典情绪与超额收益和收益波动率的关系与已有文献的结论相异，也与非正式用语词典的结果不一致. 此外，在投资者情绪一致性的回归结果中，正式用语词典情绪一致性、知网词典情绪一致性、大连理工大学词典情绪一致性、清华大学词典情绪一致性、LM 词典情绪一致性均与非正式用语情绪词典结果一致. 总体来看，其他词典并不能同时解释超额收益、交易量及波动率. 考虑到股票论坛上的帖子并非全部使用金融领域独特的非正式情绪词汇，因此其他词典使用在非正式用语的文本上，可能具有一定程度的解释能力. 但是这些结果不能全部满足已经发现的社交媒体情绪的相关规律，构建的指标在收益率预测效果上也不及本文所构建的非正式用语情绪词典. 总体来看，表 5 和表 6 的结果与本文假设 2 一致.

⑱ 非正式文本词典相关的回归变量计算，是把每个交易日作为信息发布日 t，计算方法沿用了表 1 的定义. 本文在此部分沿用了表 1 中的部分控制变量，包括 ln *Size*（市值），ln *Turnover*（换手率）和 ln(*B*/*M*)（账面市值比），各变量定义参见表 1，回归控制了公司固定效应和时间固定效应.

表 6　超额收益、交易量及波动率与基于其他词典的社交媒体情绪指标回归结果

Table 6 The regressions of sentiment indexes with different dictionaries on excess return, trading volume and volatility

解释变量	被解释变量		
	超额收益率	异常成交量	波动率
正式用语词典情绪	−0.001**	0.368***	−0.002***
	(−2.41)	(9.34)	(−18.50)
大连理工大学词典情绪	−0.001***	0.391***	−0.004***
	(−4.70)	(9.19)	(−33.38)
知网词典情绪	−0.001***	−0.352***	−0.003***
	(−8.73)	(−9.27)	(−30.72)
清华大学词典情绪	0.000	0.486***	−0.003***
	(0.42)	(11.97)	(−26.43)
LM 词典情绪	−0.001*	0.276***	−0.002***
	(1.82)	(8.24)	(−16.57)
正式用语词典情绪一致性	−0.003***	−0.668***	−0.002***
	(−11.19)	(−13.15)	(−21.30)
大连理工大学词典情绪一致性	−0.003***	−0.720***	−0.004***
	(−10.71)	(−12.05)	(−26.07)
知网词典情绪一致性	−0.004***	−1.066***	−0.004***
	(−15.27)	(−20.77)	(−29.78)
清华大学词典情绪一致性	−0.003***	−0.620***	−0.003***
	(−8.49)	(−11.89)	(−26.16)
LM 词典情绪一致性	−0.002***	−0.572***	−0.002***
	(−8.68)	(−10.66)	(−17.80)

注：括号中是 t 值；* 表示 10% 显著，** 表示 5% 显著，*** 表示 1% 显著.

3.3　稳健性检验

基于本文的两个情绪词典所构建的年报负面语调和社交媒体情绪指标分别与 Loughran 和 McDonald[4] 和 Antweiler 和 Frank[3] 的结论保持一致. 然而这些良好的表现可能来自于：1）词典本身是采用样本内数据构建；2）在正式用语词典的构建过程中，采用累计收益率作为判断标准，而这一判断标准本身有可能导致词典是样本内累计收益率与成交量等指标相关关系的表征，而非年报语调本身的特征. 基于以上两个考虑，采用样本外数据进行了词典的有效性检验. 通过这一检验，可以说明词典对非样本内文本数据进行情绪分类的有效性，同时也可以排除正式用语词典是“样本内累计收益率与成交量等指标相关关系”表征的可能性.

在样本外检验中，另外选取2016年及2017年中国上市公司年报共9 991份和2016－04－23～2018－06－28雪球论坛上的47.3万余条用户发帖. 实证结果如表7. A组结果显示，正式用语情绪词典构建的年报负面语调在样本外与交易量、波动率及下季度未预期盈余显著正相关，因此说明正式用语情绪词典是年报语调的特征，而非样本内累积收益率与成交量等指标相关关系的表征. B组结果显示，非正式用语情绪词典构建的投资者看涨情绪与超额收益、交易量及波动率显著正相关，且非正式用语情绪词典构建的投资者情绪一致性与超额收益、交易量及波动率显著负相关. 综合来看样本外检验结果与样本内检验结果一致，证明了本文构建的情绪词典、年报语调和社交媒体情绪指标的稳健性.

表 7 样本外回归结果

Table 7 The results of out-of-sample regressions

A 组:年报语调回归结果			
解释变量	被解释变量		
	异常成交量	波动率	未预期盈余
正式用语词典语调	6.127 ***	4.128 **	1.376 ***
	(2.62)	(2.15)	(2.47)
市值	0.026 **	−0.012 ***	0.002 **
	(2.02)	(−3.26)	(2.16)
换手率	5.768 ***	0.662 ***	−0.019 ***
	(11.21)	(3.64)	(−5.15)
账面市值比	0.026 ***	−0.036 **	0.007
	(3.89)	(2.02)	(1.10)
机构投资者	0.416	−0.157	0.149 ***
	(0.97)	(−1.14)	(3.86)
前期异常收益	0.126	0.020	0.071 **
	(1.28)	(0.58)	(2.20)
预测分歧			0.602 **
			(2.20)
预测修正			−0.561 ***
			(−3.81)
常数项	−0.106 ***	0.376 ***	−0.066 ***
	(−3.43)	(8.19)	(−5.00)
样本量	9 576	9 481	9 812
调整 R^2	0.726	0.378	0.062

B 组:论坛投资者情绪回归结果						
解释变量	被解释变量					
	超额收益率		异常成交量		波动率	
非正式用语词典情绪	0.001 ***		0.131 ***		0.001 ***	
	(2.42)		(2.69)		(2.58)	
非正式用语词典情绪一致性		−0.001 ***		−0.152 **		−0.001 ***
		(−3.25)		(−2.09)		(−2.37)
市值	−0.001 ***	−0.000 ***	−0.037 **	−0.041 **	0.000 ***	0.001 ***
	(−5.96)	(−6.01)	(−1.98)	(−2.14)	(5.66)	(5.70)
换手率	−0.001 ***	−0.001 ***	0.062 ***	0.701 ***	0.000 ***	0.001 ***
	(−4.34)	(−4.02)	(3.07)	(3.10)	(7.83)	(8.87)
账面市值比	0.000 **	0.000 ***	0.051 ***	0.053 ***	−0.000 ***	−0.001 ***
	(2.10)	(2.69)	(2.49)	(2.39)	(−4.96)	(−5.74)
常数项	0.011 ***	0.012 ***	1.106 **	1.107 ***	−0.004 ***	−0.004 ***
	(4.76)	(4.28)	(2.04)	(2.39)	(−2.97)	(−2.86)
样本量	472 665	472 665	472 605	472 605	472 647	472 647
调整 R^2	0.016	0.014	0.007	0.007	0.104	0.103

注: 括号中是 t 值; * 表示 10% 显著, ** 表示 5% 显著, *** 表示 1% 显著.

同时,本文将正式用语情绪词典与非正式用语情绪词典合并,并且在年报及股票论坛数据上进行检验,结果如表 8. 合并词典语调为基于合并词典所构建的年报负面语调指标,该指标与交易量、波动率及下季度未预期盈余显著正相关. 合并词典情绪是基于合并词典所构建的投资者看涨情绪指标,其与超额收益、交易量及波动率显著正相关. 合并词典情绪一致性则是基于合并词典所构

建的情绪一致性指标，其与超额收益、交易量及波动率显著负相关. 基于合并词典的检验结果与分开词典检验的结果相一致，但检验结果的显著性整体上有所下降. 年报文本及股票论坛文本就内容上来说差异较大，因此建议分别使用正式用语情绪词典和非正式用语情绪词典来分别捕捉文本情绪. 其他如媒体报道等文本，其内容用语介于正式用语及非正式用语之间，对于此类文本建议使用合并词典.

表 8　基于合并词典的年报语调与投资者情绪回归结果

Table 8 The regressions of negative tone and sentiment indexes constructed using combined dictionaries

A 组：年报语调回归结果			
变量	被解释变量		
	异常成交量	波动率	未预期盈余
合并词典语调	5.012 *	3.076 **	1.566 **
	(1.80)	(2.06)	(2.11)
市值	0.033 ***	−0.017 **	0.002 ***
	(3.64)	(−2.07)	(2.46)
换手率	6.407 ***	0.722 ***	−0.021 ***
	(14.86)	(3.62)	(−6.83)
账面市值比	0.036 ***	−0.050 ***	0.007
	(4.78)	(−4.66)	(1.30)
机构投资者	0.398	−0.160	0.153 ***
	(0.95)	(−1.14)	(3.77)
前期异常收益	0.127	0.022	0.072 **
	(1.25)	(0.59)	(2.09)
预测分歧			0.634 **
			(2.17)
预测修正			−0.560 ***
			(−3.81)
常数项	−0.101 ***	0.433 ***	−0.076 ***
	(−2.68)	(8.76)	(−5.54)
样本量	14 244	12 758	13 898
调整 R^2	0.710	0.312	0.052

B 组：论坛投资者情绪回归结果						
变量	被解释变量					
	超额收益率		异常成交量		波动率	
合并词典情绪	0.002 **		0.119 **		0.002 **	
	(2.18)		(1.99)		(2.21)	
合并词典情绪一致性		−0.001 *		−0.102 *		−0.001 *
		(−1.82)		(−1.83)		(−1.84)
市值	−0.001 ***	−0.001 ***	−0.041 **	−0.043 **	0.001 ***	0.001 ***
	(−5.27)	(−8.69)	(−2.11)	(−2.17)	(6.95)	(5.75)
换手率	−0.001 ***	−0.001 ***	0.065 ***	0.759 ***	0.000 ***	0.001 ***
	(−4.64)	(−5.27)	(3.14)	(3.77)	(7.26)	(8.12)
账面市值比	0.000 ***	0.000 ***	0.050 ***	0.059 ***	−0.000 ***	−0.001 ***
	(2.54)	(3.13)	(2.39)	(2.93)	(−4.58)	(−4.62)
常数项	0.013 ***	0.016 ***	1.179 ***	1.179 ***	−0.006 ***	−0.007 ***
	(5.26)	(5.88)	(2.97)	(3.08)	(−4.74)	(−5.12)
样本量	564 887	564 887	564 822	564 822	564 886	564 886
调整 R^2	0.012	0.012	0.006	0.005	0.101	0.101

注：括号中是 t 值；* 表示 10% 显著，** 表示 5% 显著，*** 表示 1% 显著.

3.4 股价崩盘风险预测

为了进一步说明基于词典构建的年报语调指标和社交媒体情绪指标是上市公司管理层以及个体投资者情感和心理活动的表征,将这两个指标和上市公司股价崩盘风险相关联. 根据假设 3,负面年报语调与股票未来的股价崩盘风险显著负相关,且社交媒体看跌情绪与股票未来的股价崩盘风险显著正相关. 由于上文构建的是社交媒体看涨情绪指标,而此处检验更关心投资者看跌情绪的积累是否导致股价崩盘,因此在检验过程中将看涨指标更换为看跌指标(非正式用语词典看跌情绪),计算方法为每支股票日发帖的负面词比例减去正面词比例. 如果假设 3 成立,上文构建的社交媒体看跌情绪指标应该与未来的股价崩盘风险显著正相关.

为检验以上假设,首先构建了股价崩盘风险的代理变量. Bao 等[38] 给出了多种股价崩盘风险的度量指标,参照他本文首先通过式(4) 的残差部分 $\varepsilon_{i,t}$ 估计了公司净日收益率 $W_{i,t}$

$$r_{i,t} = \beta_i + \beta_{1,i} r_{m,t-1} + \beta_{2,i} r_{j,t-1} + \beta_{3,i} r_{m,t} + \beta_{4,i} r_{j,t} + \beta_{5,i} r_{m,t+1} + \beta_{6,i} r_{j,t+1} + \varepsilon_{i,t} \quad (4)$$

式中 $r_{i,t}$ 为 i 公司第 t 日的收益率;$r_{m,t}$ 为市场指数(上证综指或深成指) 第 t 日的收益率;$r_{j,t}$ 为 i 公司所属行业 j(Wind 二级行业指数) 第 t 日的收益率. 而公司净日收益率 $W_{i,t} = \ln(1 + \varepsilon_{i,t})$. 本文使用以下 3 个指标度量季度股价崩盘风险(*CrashRisk*)

1) 崩盘风险 1(*CRASH*) 为公司净日收益率减去公司当季净日收益率均值与 3.09 倍标准差之差的季度累计和.

2) 崩盘风险 2(*NCSKEW*) 为季度内公司日净收益率的负偏态系数,计算式为

$$NCSKEW_{i,t} = \frac{-[n(n-1)^{3/2} \sum W_{i,t}^3]}{(n-1)(n-2)(\sum W_{i,t}^2)^{3/2}} \quad (5)$$

3) 崩盘风险 3(*DUVOL*) 为收益率上下波动率,用来捕捉股价崩盘风险的上下波动. 对于每季度每个上市公司,定义净收益率小于均值的交易日为下跌日,净收益率高于均值的交易日为上涨日. 分别计算出下跌日和上涨日净收益率的标准差,得出下跌波动率和上涨波动率. 最后,以 ln(下跌波动率/上涨波动率) 作为每季度每个上市公司的 *DUVOL* 指标.

在得到上市公司的季度股价崩盘风险后,将本年度年报语调指标和本季度社交媒体情绪指标分别和下季度公司股价崩盘风险相关联. 选取样本期为 2003 年 ~ 2015 年中国上市公司年报和 2011 年第 2 季度至 2018 年第 2 季度雪球论坛上的用户发帖为样本,分别构建基于正式用语情绪词典的年报负面语调指标,以及基于非正式用语情绪词典的季度社交媒体看跌情绪指标. 回归方程如下式

$$CrashRisk_{i,t+1} = \beta_0 + \beta_1 Index_{i,t} + \beta_2 Ctrl_{i,t} + \mu_i + \nu_{t+1} + \varepsilon_{i,t+1} \quad (6)$$

式中 $CrashRisk_{i,t+1}$ 表示公司股价崩盘风险,分别为股价崩盘风险的 3 个代理指标 *CRASH*、*NCSKEW* 和 *DUVOL*;$Index_{i,t}$ 分别为正式用语词典语调和季度非正式用语词典看跌情绪;$Ctrl_{i,t}$ 为控制变量,包括机构投资者占比、账面市值比和本季度的 Fama-French 三因子模型 α;μ_i 为公司固定效应;ν_{t+1} 为季度固定效应;$\varepsilon_{i,t+1}$ 为残差项. 回归系数的标准误在公司层面进行了聚类处理.

实证结果如表 9. A 组给出了基于正式用语情绪词典构建的年报负面语调指标对公司股价崩盘风险的预测结果. 结果显示,年报负面语调指标与下季度的公司股价崩盘风险显著负相关,说明年报表达的悲观预期程度越高,公司管理层越倾向于公布负面消息,因此公司未来发生股价崩盘风险的概率越低. B 组给出了基于非正式用语情绪词典构建的社交媒体看跌情绪指标对公司股价崩盘风险的预测结果. 结果显示,社交媒体看跌情绪指标与下季度的公司股价崩盘风险显著正相关,因此说明社交媒体上投资者情绪越低,投资者对公司前景越悲观,但是由于投资者无法获得卖空途径,悲观情绪难以在交易中释放,导致公司未来发生股价崩盘风险的概率越高. 综合来看表 9 的结果和假设 3 一致,证明了本文构建的年报语调指标和社交媒体情绪指标准确表征了上市公司管理层以及个体投资者的情感和心理活动,具有较好的稳健性.

表 9　年报语调，社交媒体投资者情绪与股价崩盘风险

Table 9 The regressions of negative tone and sentiment indexes on crash risk

A 组：年报语调回归结果						
变量	崩盘风险 1		崩盘风险 2		崩盘风险 3	
	1	2	3	4	5	6
正式用语词典语调	−1.585***	−1.318***	−6.784***	−6.150***	−10.611***	−8.266**
	(−3.57)	(−3.19)	(−3.05)	(−2.85)	(−2.73)	(−2.28)
机构投资者		−0.539***		−0.492***		−0.728***
		(−3.03)		(−3.97)		(−2.70)
账面市值比		0.936		0.277		0.185
		(1.42)		(1.36)		(1.33)
前期异常收益		−6.425***		−16.888***		−29.583***
		(−5.11)		(−8.12)		(−8.74)
常数项	−7.945***	−6.520***	−9.109***	−8.371***	−11.711***	−11.281***
	(−14.10)	(−14.07)	(−13.83)	(−13.25)	(−15.18)	(−15.09)
样本量	24 330	24 330	24 330	24 330	24 330	24 330
调整 R^2	0.719	0.759	0.238	0.320	0.276	0.384
B 组：论坛投资者情绪回归结果						
变量	崩盘风险 1		崩盘风险 2		崩盘风险 3	
	1	2	3	4	5	6
非正式用语词典看跌情绪	0.080**	0.083**	0.257***	0.151**	0.284***	0.171**
	(2.25)	(2.39)	(4.92)	(2.70)	(4.62)	(2.37)
机构投资者		−0.344**		−0.275*		−0.427**
		(−2.67)		(−2.05)		(−2.51)
账面市值比		0.608***		0.049		0.071
		(6.81)		(1.02)		(1.24)
前期异常收益		−4.295***		−15.488***		−16.354***
		(−15.42)		(−16.54)		(−16.39)
常数项	−1.530***	−1.777***	−1.537***	−1.594***	−2.032***	−2.135***
	(−88.74)	(−41.47)	(−62.09)	(−47.76)	(−73.61)	(−54.95)
样本量	19 839	19 839	19 839	19 839	19 839	19 839
调整 R^2	0.609	0.633	0.075	0.173	0.097	0.192

注：括号中是 t 值，* 表示 10%显著，** 表示 5%显著，*** 表示 1%显著

4　结束语

本文构建了中国沪深股票市场的年报语调和社交媒体情绪指标，并发现年报语调和社交媒体情绪指标可以显著预测公司股票未来的收益率、波动率、成交量和未预期盈余，对股价崩盘风险指标也有良好的预测作用. 在语调和情绪指标构建过程中，基于词典重组和深度学习算法（LSTM）构建了适用于金融领域研究的中文情绪词典，包括正式用语情绪词典和非正式用语情绪词典，其中正式用语情绪词典适用于公司年报等正式文本的语调分析，而非正式用语情绪词典则适用于社交媒体等非正式文本的情绪分析. 检验了所构建语调和情绪指标的回归结果与现有的年报语调和投资者情绪相关理论的符合程度，并与根据广泛

使用的情绪词典所构建的语调和情绪指标结果进行了对比,主要结论如下.

1)基于正式用语情绪词典构建的年报负面语调与交易量、波动率及下季度未预期盈余显著正相关,说明了年报语调与年报发布之后的股票市场表现有相关性,证明本文构建的正式用语情绪词典,以及在此基础上形成的年报语调指标是有效的.

2)基于非正式用语情绪词典构建的社交媒体看涨情绪指标与超额收益、成交量、波动率正相关,构建的情绪一致性指标与超额收益、交易量及收益波动率显著负相关,证明本文构建的非正式用语情绪词典,以及在此基础上形成的社交媒体情绪指标是有效的.

3)现有广泛使用的情绪词典在金融领域正式用语文本和非正式用语文本上的语调和情绪提取并不能达到理想的效果,因此在金融领域研究中构建专业情绪词典是有必要的.

4)作为对年报语调与社交媒体情绪指标的应用,研究发现基于本文词典构建的年报负面语调与社交媒体看跌情绪指标可以显著预测上市公司的股价崩盘风险,证明语调和情绪指标是上市公司管理层以及个体投资者的情感和心理活动的准确反映.

本文提出的两个情绪词典,除了应用在年报和股票论坛之外,还有潜力对分析师报告、财经新闻、微博中财经类发帖等文本进行分析,进而构建相关的语调和情绪指标. 此类型的研究在国内目前处于起步阶段,因此两个词典在未来金融领域的中文文本分析研究中有较好的应用空间.

参 考 文 献:

[1]姚加权, 张锟澎, 罗 平. 金融学文本大数据挖掘方法与研究进展[J]. 经济学动态, 2020, (4): 143 - 158.
Yao Jiaquan, Zhang Kunpeng, Luo Ping. Text mining in financial big data and its research progress[J]. Economic Perspectives, 2020, (4): 143 - 158. (in Chinese)

[2]Bodnaruk A, Loughran T, McDonald B. Using 10-K text to gauge financial constraints[J]. Journal of Financial and Quantitative Analysis, 2015, 50(4): 623 - 646.

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

[4]Loughran T, McDonald B. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks[J]. Journal of Finance, 2011, 66(1): 35 - 65.

[5]Feldman R, Govindaraj S, Livnat J, et al. Management's tone change, post earnings announcement drift and accruals[J]. Review of Accounting Studies, 2010, 15(4): 915 - 953.

[6]Liu B, McConnell J J. The role of the media in corporate governance: Do the media influence managers' capital allocation decisions? [J]. Journal of Financial Economics, 2013, 110(1): 1 - 17.

[7]Rogers J L, Buskirk A V, Zechman S L C. Disclosure tone and shareholder litigation[J]. The Accounting Review, 2011, 86(6): 2155 - 2183.

[8]Chang Y C, Hong H G, Tiedents L, et al. Does Diversity Lead to Diverse Opinions? Evidence from Languages and Stock Markets[R]. Stanford: Stanford University Graduate School of Business Research Paper, 2015.

[9]部 慧, 解 峥, 李佳鸿, 等. 基于股评的投资者情绪对股票市场的影响[J]. 管理科学学报, 2018, 21(4): 86 - 101.
Bu Hui, Xie Zheng, Li Jiahong, et al. Investor sentiment extracted from internet stock message boards and its effect on Chinese stock market[J]. Journal of Management Sciences in China, 2018, 21(4): 86 - 101. (in Chinese)

[10]谢德仁, 林 乐. 管理层语调能预示公司未来业绩吗? ——基于我国上市公司年度业绩说明会的文本分析[J]. 会计研究, 2015, (2): 20 - 27.
Xie Deren, Lin Le. Do management tones help to forecast firms' future performance: A textual analysis based on annual earnings communication conferences of listed companies in China[J]. Accounting Research, 2015, (2): 20 - 27. (in Chinese)

[11]汪昌云，武佳薇. 媒体语气，投资者情绪与 IPO 定价[J]. 金融研究，2015，(9)：174 - 189.
Wang Changyun，Wu Jiawei. Media tone，investor sentiment and IPO pricing[J]. Journal of Financial Research，2015，(9)：174 - 189. (in Chinese)

[12]Engelberg J，Reed A V，Ringgenberg M C. How are shorts informed? Short sellers，news，and information processing[J]. Journal of Financial Economics，2012，105(2)：260 - 278.

[13]Henry E. Are investors influenced by how earnings press releases are written? [J]. Journal of Business Communication，2008，45(4)：363 - 407.

[14]Tetlock P C. Giving content to investor sentiment：The role of media in the stock market[J]. Journal of Finance，2007，62(3)：1139 - 1168.

[15]Loughran T，McDonald B. Textual analysis in accounting and finance：A survey[J]. Journal of Accounting Research，2016，54(4)：1187 - 1230.

[16]Kearney C，Liu S. Textual sentiment in finance：A survey of methods and models[J]. International Review of Financial Analysis，2014，33：171 - 185.

[17]Dougal C，Engelberg J，Garcia D，et al. Journalists and the stock market[J]. Review of Financial Studies，2012，25(3)：639 - 679.

[18]Garcia D. Sentiment during recessions[J]. Journal of Finance，2013，68(3)：1267 - 1300.

[19]Solomon D H，Soltes E，Sosyura D. Winners in the spotlight：Media coverage of fund holdings as a driver of flows[J]. Journal of Financial Economics，2014，113(1)：53 - 72.

[20]Das S R，Chen M Y. Yahoo! for Amazon：Sentiment extraction from small talk on the web[J]. Management Science，2007，53(9)：1375 - 1388.

[21]Chen H，De P，Hu Y，et al. Wisdom of crowds：The value of stock opinions transmitted through social media[J]. Review of Financial Studies，2014，27(5)：1367 - 1403.

[22]杨晓兰，王伟超，高　媚. 股市政策对股票市场的影响——基于投资者社会互动的视角[J]. 管理科学学报，2020，23(1)：15 - 32.
Yang Xiaolan，Wang Weichao，Gao Mei. The impact of stock market policies on stock market：From the perspective of investor social interaction[J]. Journal of Management Sciences in China，2020，23(1)：15 - 32. (in Chinese)

[23]李培功，沈艺峰. 媒体的公司治理作用：中国的经验证据[J]. 经济研究，2010，45(4)：14 - 27.
Li Peigong，Shen Yifeng. The corporate governance role of media：Empirical evidence from China[J]. Economic Research Journal，2010，45(4)：14 - 27. (in Chinese)

[24]唐国豪，姜富伟，张定胜. 金融市场文本情绪研究进展[J]. 经济学动态，2016，(11)：137 - 147.
Tang Guohao，Jiang Fuwei，Zhang Dingsheng. Research progress of textual sentiment in financial market[J]. Economic Perspectives，2016，(11)：137 - 147. (in Chinese)

[25]张学勇，吴雨玲. 基于网络大数据挖掘的实证资产定价研究进展[J]. 经济学动态，2018，(6)：129 - 140.
Zhang Xueyong，Wu Yuling. Research progress of empirical asset pricing based on internet big data mining[J]. Economic Perspectives，2018，(6)：129 - 140. (in Chinese)

[26]Tumarkin R，Whitelaw R F. News or noise? Internet postings and stock prices[J]. Financial Analysts Journal，2001，57(3)：41 - 51.

[27]Bollen J，Mao H，Zeng X J. Twitter mood predicts the stock market[J]. Journal of Computational Science，2011，2(1)：1 - 8.

[28]段江娇，刘红忠，曾剑平. 中国股票网络论坛的信息含量分析[J]. 金融研究，2017，(10)：178 - 192.
Duan Jiangjiao，Liu Hongzhong，Zeng Jianping. Analysis on the information content of China's internet stock message boards[J]. Journal of Financial Research，2017，(10)：178 - 192. (in Chinese)

[29]许年行，江轩宇，伊志宏，等. 分析师利益冲突，乐观偏差与股价崩盘风险[J]. 经济研究，2012，47(7)：127 - 140.
Xu Nianhang，Jiang Xuanyu，Yi Zhihong，et al. Conflicts of interest，analyst optimism and stock price crash risk[J]. Economic Research Journal，2012，47(7)：127 - 140. (in Chinese)

[30]彭俞超，倪骁然，沈　吉. 企业“脱实向虚”与金融市场稳定——基于股价崩盘风险的视角[J]. 经济研究，2018，

53(10): 50 – 66.
Peng Yuchao, Ni Xiaoran, Shen Ji. The effect of transforming the economy from substantial to fictitious on financial market stability: An analysis on stock price crash risk[J]. Economic Research Journal, 2018, 53(10): 50 – 66. (in Chinese)
[31] Easley D, O'hara M, Srinivas P S. Option volume and stock prices: Evidence on where informed traders trade[J]. Journal of Finance, 1998, 53(2): 431 – 465.
[32] Grossman S J, Stiglitz J E. On the impossibility of informationally efficient markets[J]. American Economic Review, 1980, 70(3): 393 – 408.
[33] Hong H, Stein J C. Differences of opinion, short-sales constraints, and market crashes[J]. Review of Financial Studies, 2003, 16(2): 487 – 525.
[34] 褚 剑, 方军雄. 中国式融资融券制度安排与股价崩盘风险的恶化[J]. 经济研究, 2016, 51(5): 143 – 158.
Chu Jian, Fang Junxiong. Margin-trading, short-selling and the deterioration of crash risk[J]. Economic Research Journal, 2016, 51(5): 143 – 158. (in Chinese)
[35] 苏 治, 卢 曼, 李德轩. 深度学习的金融实证应用: 动态, 贡献与展望[J]. 金融研究, 2017, (5): 111 – 126.
Su Zhi, Lu Man, Li Dexuan. Deep learning in financial empirical applications: Dynamics, contributions and prospects[J]. Journal of Financial Research, 2017, (5): 111 – 126. (in Chinese)
[36] Derks D, Bos A E R, Grumbkow J V. Emoticons and social interaction on the internet: The importance of social context [J]. Computers in Human Behavior, 2007, 23(1): 842 – 849.
[37] Fama E F, MacBeth J D. Risk, return, and equilibrium: Empirical tests[J]. Journal of Political Economy, 1973, 81(3): 607 – 636.
[38] Bao D, Fung S Y K, Su L. Can shareholders be at rest after adopting clawback provisions? Evidence from stock price crash risk[J]. Contemporary Accounting Research, 2018, 35(3): 1578 – 1615.

Tone, sentiment and market impacts: The construction of Chinese sentiment dictionary in finance

YAO Jia-quan[1], *FENG Xu*[2*], *WANG Zan-jun*[3], *JI Rong-rong*[4], *ZHANG Wei*[2]
1. School of Management, Jinan University, Guangzhou 510632, China;
2. College of Management and Economics, Tianjin University, Tianjin 300072, China;
3. The School of Economics, Xiamen University, Xiamen 361005, China;
4. School of Information Science and Engineering, Xiamen University, Xiamen 361005, China

Abstract: Tone and sentiment in financial contexts, containing emotional information expressed by managers in public listed firms and individual investors, affect stock market. By restructuring dictionaries and using deep learning model LSTM, the paper constructs two Chinese sentiment dictionaries for formal and informal texts in Finance respectively. Based on the constructed dictionaries, tone measures of annual filings and sentiment proxies of social media for Chinese public firms are proposed. Our tone measures of annual filings and sentiment proxies of social media can effectively predict stock return, trading volume, return volatility, unexpected earnings and other market factors and perform better than indices made by other commonly used sentiment lexicons. Additionally, our tone measures of annual filings and sentiment proxies of social media have predictive abilities for crash risk of public firms. This research provides an analytical tool for big data application in financial market and offers decision-making supports in financial market forecasting, monitoring, and other activities in the big data era.

Key words: sentiment dictionary; tone; investor sentiment; market impacts

附录

附表 1　部分正式用语情绪词典及非正式用语情绪词典

Appendix Table 1 Examples of words in formal and informal sentiment dictionaries

正式用语情绪词典：									
负面									
风险	亏损	违反	损害	舞弊	严重	约束	手段	坏帐	负担
越权	不道德	毁损	异常	谴责	严峻	委靡	困顿	失利	守旧
不健全	仿造	倒闭	侮辱	压制	冒进	刁难	危害	压迫	低迷
正面									
平稳	崛起	精神	和谐	突出	合格	力争	透明	成熟	迅速
倾心	保密	清晰	积极性	严正	丰硕	乐观	从优	信誉	充实
不屈	威信	完备	创新	勇气	飙升	富余	干劲	庆祝	强悍
非正式用语情绪词典：									
负面									
垃圾	下跌	回调	割肉	套牢	风险	减持	抛售	可悲	低迷
向下	跌破	无耻	狗屎	利空	困顿	可笑	跳空	倒霉	赔钱
烂股	小人	绝望	卑鄙	压制	不值	草包	担心	丢脸	烦心
正面									
涨停	崛起	胜利	献花	发财	暴涨	战斗机	稳赚	过瘾	幸运
黑马	赚翻天	爽歪歪	止跌	恭喜	开心	舒服	漂亮	牛股	完美
赚大	期待	好样	创新	勇气	神奇	明智	成功	飙升	支持

附表 2　投资者情绪回归相关变量的描述性统计

Appendix Table 2 Summary statistics of regression variables related to investor sentiment indexes

变量	样本量	均值	标准差	最小值	最大值
超额收益率	569 198	-0.000 3	0.077	-0.815	0.556
异常成交量	569 133	1.378	13.235	-45.144	2 202.645
波动率	569 197	0.064	0.040	0	0.271
非正式用语词典情绪	564 892	-0.403	0.340	-1	1
正式用语词典情绪	461 425	0.416	0.522	-1	1
大连理工大学词典情绪	545 779	0.267	0.426	-1	1
知网词典情绪	462 392	0.323	0.551	-1	1
清华大学词典情绪	496 216	0.493	0.477	-1	1
LM 词典情绪	552 137	0.012	0.146	-1	1
非正式用语词典情绪一致性	557 261	0.455	0.315	0	1
正式用语词典情绪一致性	461 425	0.381	0.414	0	1
大连理工大学词典情绪一致性	545 779	0.192	0.308	0	1
知网词典情绪一致性	462 392	0.356	0.419	0	1
清华大学词典情绪一致性	492 216	0.379	0.378	0	1
LM 词典情绪一致性	552 137	0.323	0.406	0	1
市值	569 201	22.613	0.968	18.959	28.144
换手率	569 199	-3.550	0.989	-11.512	-0.243
账面市值比	569 201	-1.469	0.829	-9.873	1.693

附表 3 投资者看涨情绪与情绪一致性相关系数表

Appendix Table 3 correlation coefficients of Investor bullish sentiment and sentiment consistency

变量	非正式用语词典情绪	正式用语词典情绪	大连理工大学词典情绪	知网词典情绪	清华大学词典情绪	LM 词典情绪
非正式用语词典情绪	1.000					
正式用语词典情绪	0.107**	1.000				
大连理工大学词典情绪	-0.142**	0.534**	1.000			
知网词典情绪	-0.035**	0.225**	0.512**	1.000		
清华大学词典情绪	-0.031**	0.663**	0.640**	0.177**	1.000	
LM 词典情绪	-0.021**	0.477**	0.504**	0.519**	-0.037**	1.000

变量	非正式用语词典情绪一致性	正式用语词典情绪一致性	大连理工大学词典情绪一致性	知网词典情绪一致性	清华大学词典情绪一致性	LM 词典情绪一致性
非正式用语词典情绪一致性	1.000					
正式用语词典情绪一致性	0.178**	1.000				
大连理工大学词典情绪一致性	0.315**	0.409**	1.000			
知网词典情绪一致性	0.152**	0.428**	0.416**	1.000		
清华大学词典情绪一致性	0.133**	0.572**	0.483**	0.359**	1.000	
LM 词典情绪一致性	0.441**	0.468**	0.472**	0.436**	0.506**	1.000

注：** 表示在 5%的水平下显著.