语调、情绪与文本分析

-构建金融领域中文情绪词典

姚加权　冯绪　王赞钧　纪荣嵘　张维

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

作者简介: 姚加权，暨南大学管理学院教授，Email: jiaquanyao@gmail.com.
冯绪（通讯作者），天津大学管理与经济学部副教授，Email: fengxu@tju.edu.cn.
王赞钧，厦门大学经济学院硕士研究生，Email: qaz1992726@gmail.com.
纪荣嵘，厦门大学信息科学与技术学院教授，Email: jirongrong@gmail.com.
张维，天津大学管理与经济学部教授，Email: weiz@tju.edu.cn.

*本文得到国家自然科学基金重大项目(71790594)、面上项目(71871157)及青年项目(71502152)的资助。本文所著中文情绪词典将由天津大学社会计算研究中心、暨南大学管理学院和厦门大学媒体分析与计算组联合发布。感谢北京大学国家发展研究院沈艳，加拿大麦吉尔大学谭红平，西南财经大学罗荣华，中科院计算所罗平等老师的建设性意见，感谢厦门大学媒体分析与计算组陈福海和周奕毅的协助。文责自负。

语调、情绪与文本分析 - 构建金融领域中文情绪词典

Tone, Sentiment and Textual Analysis: the Construction of Chinese Sentiment Dictionary in Finance

YAO Jiaquan　FENG Xu　WANG Zanjun　JI Rongrong　ZHANG Wei

(School of Management, Jinan University;
College of Management and Economics, Tianjin University;
School of Economics, Xiamen University;
School of Information Science and Engineering, Xiamen University)

摘　要:本文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。分别通过词典重组和长短期记忆模型深度学习算法对上市公司年报和社交媒体发帖进行文本分析，构建了正式用语情绪词典和非正式用语情绪词典。我们对词典的可用性进行了检验，并证明了对正式和非正式文本分别构建情绪词典的必要性。根据本文构建的情绪词典计算所得到的情绪指标能有效地预测上市公司股票的收益率、成交量、波动率等市场因素，并优于其他广泛使用的情绪词典。

关键词: 情绪词典；语调；投资者情绪

Abstract: This paper proposes two Chinese sentiment dictionaries used for formal and informal texts in Finance respectively. We analyze the textual content in annual filings and social media via restructuring dictionaries and deep learning model LSTM respectively in order to construct sentiment dictionary for formal text and sentiment dictionary for informal text. We check the validity of the two dictionaries and show the necessity to construct sentiment dictionaries specifically for formal and informal texts. This study shows the sentiment indices made by the dictionaries can predict stock return, trading volume, return volatility and other market factors for public firms and the two dictionaries perform better than other commonly used lexicons.

Key words: Sentiment Dictionary, Tone, Investor Sentiment
JEL Classification: G10, G30

一、引言

大数据时代，越来越多金融领域研究通过文本分析方法提取上市公司年报、新闻报道和社交媒体发帖中所包含的情绪与语调。而情绪词典是进行准确文本分析的关键所在。现有的研究(尤其是中文文本研究)大多通过人工方法构建了基于小样本的情绪词典，但是存在情绪提取准确率低，研究结果无法重复等问题。另外有一部分学者通过机器学习的方法来判断文本的情绪(Antweiler and Frank(2004), Li(2010))，但是运用机器学习算法，特别是有监督学习涉及的训练集，对大部分研究者执行起来存在很大困难[1]。因此构建一个基于大样本、标准化的中文情绪词典是目前金融市场文本情绪相关研究中亟待解决的问题。

在国外的金融领域文本情绪研究中，Loughran and McDonald(2011)提出了一个适用于上市公司英文年报语调分析的情绪词典[2]。然而，对于中文研究来说，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，不能将英文年报情绪词典经过翻译后简单地套用在中文年报的分析上[3]。此外，针对中文文本的一些通用型词典，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，它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。因此构建专门针对金融领域的中文情绪词典能有效地增加金融领域文本情绪与语调提取的准确度。

基于以上考虑，在正式用语文本方面，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研究的中文情绪词典。本文利用 2003-2015 年间所有中国上市公司年报文本(共计 19970 份)，结合 Engelberg et al.(2012)的语调判断方法对三份现有通用型中文情绪词典和 Loughran and McDonald (2011)情绪词典的中文翻译版进行正负面词语提取，并通过带惩罚机制词频法构建了适用于中文年报文本分析的正式用语情绪词典。此外，考虑到社交媒体中的文本以非正式用语为主，基于正式用语的情绪词典可能不再适用，因此本文利用 2011-2016 年间社交媒体平台的文本(雪球网和东方财富网共计 8130 多万条发帖)，以 8789 条带有情绪识别符号的股票论坛发帖为训练集，结合长短期记忆模型(Long Short-Term Memory, LSTM)的深度学习算法，构建了适用于中文社交媒体文本分析的非正式用语情绪词典。最终形成的正式用语情绪词典含有 1633 个负面词和 3592 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。

为了检验情绪词典的有效性，本文根据两个情绪词典分别构建了情绪指标,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，而根据非正式用语情绪词典构建的投资者情绪指标则能显著预测未来股票超额收益、波动率、成交量等指标。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均是有效的。此外，本文也将所构建的语调指标与大连理工情感词汇本体库、中国

[1]英文文本方面,Antweiler and Frank (2004)选用 Yahoo 财经论坛中的 1000 条发帖人工进行情绪分类,并把分类结果作为机器学习模型的训练集。中文文本方面,杨晓兰等（2016）选用东方财富股吧中的 2000 条发帖作为机器学习模型训练集进行人工情绪分类。

[2]Loughran and McDonald(2011)词典在英文文本中得到广泛的研究应用，如：Feldman et al.(2010)、Liu and McConnell(2013)、Kearney and Liu(2014)等。

[3]谢德仁和林乐(2015)、汪昌云和武佳薇(2015)等均使用了 Loughran and McDonald(2011)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。

知网词库和清华大学褒贬意词典三个通用型情绪词典[4]以及 Loughran and McDonald (2011)词典中文翻译版所构建的语调指标进行比对，发现现有广泛使用的情绪词典[5]在金融领域研究中的语调和情绪提取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。

本文在以下方面拓展了现有的研究:(1)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一个简易的情绪提取工具，且有助于实现未来相关领域研究结果的可重复性。此外本文使用的词典构建避免了人工判断和筛选情绪词所造成的偏差和遗漏，词典构建方法本身也具有可重复性。为了促进金融领域中文文本分析相关研究的开展，本文构建的两部情绪词典将对外公开；(2)将现有的广泛使用的情绪词典在金融领域研究的适用性进行了测度，发现这些词典并不适用于年报和社交媒体的语调和情绪提取，这些结果为金融领域情绪词典构建的必要性提供了支持，并且为今后的相关研究提供了借鉴；(3)基于本文词典所构建的情绪指标与股票的超额收益、波动率、成交量等因素显著相关，这些结果支持了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市场的证据。

本文的结构安排:第二部分为文献回顾和研究假设；第三部分为情绪词典构建；第四部分为实证结果；第五部分为本文结论。

二、相关文献回顾和研究假设

(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 and McDonald(2011)发布的词典最有影响力。Loughran and McDonald(2011)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(2008)词典、Harvard-IV-4 词典和 Diction 词典。Tetlock(2007)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 and McDonald(2016)在文献综述中指出上述三个词典存在的缺陷：Henry(2008)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 and Liu(2014)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。如：Dougal et al.(2012)使用 LM 词典研究《华尔街日报》中“与市场同步专栏”语调。Garcia(2013)通过 LM 词典分析了《纽约时报》的财经专栏语调。Liu and McConnell(2013)则是将《华尔街日报》、《纽约时报》以及《道琼斯新闻》三家媒体的报导内容结合起来分析语调。此外 Solomon et al. (2014)还研究了《华尔街日报》、《纽约时报》、《华盛顿邮报》和《今日美国》四家媒体的新闻语调对共同基金的买入量等指标的影响。

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

[4]大连理工情感词汇本体库 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

[5]本文将现有的三个通用型情绪词典（大连理工情感词汇本体库、中国知网词库和清华大学褒贬意词典）以及 Loughran and McDonald (2011)词典的中文翻译版本合称为“广泛使用的情绪词典”。

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，此类研究大多构造一个小的训练集（如：Antweiler and Frank(2004)、Das and Chen (2007)、Kim and Kim (2014)等），然后通过传统的机器学习方法判断文本情绪。此外，Chen et al.(2014)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（如：杨晓兰等(2016)、Chang et al.(2015)等）。也有一些研究将Loughran and McDonald(2011)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（如：谢德仁和林乐(2015)、汪昌云和武佳薇(2015)等）。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（如：李培功和沈艺峰(2010)、游家兴等(2018)，以及杨道广等(2017)）。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三部词典。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7773个负面词与8610个正面词。知网情感分析用词语集发布于2007年，含有1254个中文负面词与836个中文正面词。清华大学褒贬意词典则含有4468个负面词与5567个正面词。但这三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(2016)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。

(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，随着文本分析方法在金融市场的应用，越来越多的研究开始分析年报语调与公司股票交易特征的关系。其中Loughran and McDonald(2011)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，并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。而MD&A部分的负面语调和股票超额报酬率没有显著关系，该研究认为MD&A部分相对于整份年报来说，并没有包含更多的情绪。Feldman et al.(2010)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。Price et al.(2012)使用Henry(2008)词典证实了公司电话会议问答环节的正面语调与未来股票收益率存在显著的正相关关系。Hanley and Hoberg(2010)使用Harvard-IV-4词典对公司招股说明书风险评述部分进行分析，发现该部分的语气显著影响IPO首日收益率。Rogers et al. (2011)利用Henry(2008)词典和Loughran and McDonald(2011)词典发现公司过于乐观地披露信息可能导致更多的法律诉讼。谢德仁和林乐(2015)使用中国市场数据发现年报和管理层语调与公司未来业绩关联。汪昌云和武佳薇(2015)则分析了公司上市前不同时段内的主流财经媒体报导的语调，发现媒体语调能够更好地解释IPO抑价率、首日换手率及超募比例。

综合来看，现有研究普遍认为年报语调可以预测发布日后公司股票交易特征。以最有代表性的Loughran and McDonald(2011)的研究为例，发现年报语调与公司股票的超额收益、交易量、收益波动率及下季度未预期盈余显著相关。其中悲观的年报语调与年报发布后的超额收益显著负相关，并与交易量、波动率及下季度未预期盈余显著正相关。从投资者层面来看，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，经常以较大的偏离市场价格卖出股票。从公司管理层面来看，管理层通过更悲观的语调来降低投资者对于下一季公司经营表现的预期，进而在未来可能产生更高的未预期盈余。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一套行之有效的词典。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，据此提出假设 1:

假设 1:针对年报文本，通过本文构建的中文正式用语情绪词典得到的负面语调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]。

(3)社交媒体投资者情绪指标

社交媒体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否含有有效的信息？以及投资者在社交媒体中的情绪是否对市场产生影响？这些问题成为此方向研究的主要驱动力。现有的研究大多利用文本分析提取社交媒体中的投资者情绪，并检验投资者情绪对资产价格的影响。在早期的研究中，Tumarkin and Whitelaw(2001)以及 Das and Chen(2007)均发现股吧论坛信息不会对收益率产生影响，因此认为社交媒体信息内容以噪音为主。但是 Antweiler and Frank(2004)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。此外，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 et al. (2011)分析了 Twitter 上的发贴并提取了公众情绪，发现在日频度下，公众情绪与道琼斯指数的收益率有显著正相关关系。Chen et al.(2014)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，此外投资者情绪的一致性也可以显著预测未来股票的波动率和成交量。段江娇等(2017)对中国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(2016)也分析了东方财富股吧中投资者发帖的情绪与本地股票关注度的关系。

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。社交媒体中的投资者看涨情绪与收益率显著正相关，且投资者情绪一致性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，本文提出假设 2:

假设 2:在股票论坛中，通过本文构建的非正式用语情绪词典计算出投资者看涨情绪指标和投资者情绪一致性指标。其中看涨情绪与超额收益显著正相关，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情绪指标无法满足这些关系。

三、情绪词典构建

本文构建正式用语情绪词典和非正式用语情绪词典的思路如下：在正式用语情绪词典构建方面，本文参照 Engelberg et al.(2012)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 et al.(2012)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。

[6]由于本文根据 Engelberg et al.(2012)的方法使用超额收益率作为判断年报正负语调的依据，因此正式用语情绪词典构建的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。

然后采用词典重组法，综合大连理工情感词汇本体库、中国知网词库、清华大学褒贬意词典和Loughran and McDonald(2011)词典中文翻译版并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，进而构建得到年报语调指标；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用表情符号(如 Emoji 表情等)来表达发帖者的情绪，本文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[7]，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(Long Short-Term Memory, LSTM)模型的深度学习算法分析股票论坛上的帖子情绪[8]，同样使用带惩罚机制词频法提取情绪词典。最终形成的正式用语情绪词典含有 1633 个负面词和 3592 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 个词语。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。

(一)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，以图 1 为例，该图为贵州茅台酒股份有限公司 2015 年部分年报。图中所圈选的单词带有情绪，而本文的词典构建方法就是将圈选的单词提取出来。

[此处插入图 1]

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 2003-2015 年共 19970 份年报，其中累计收益率为正的有 12475 份，累计收益率为负的有 7383 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本文提取情绪词典的数据样本中。然后综合大连理工情感词汇本体库、中国知网词库、清华大学褒贬意词典和 Loughran and McDonald(2011)词典中文翻译版，筛选出各词典正负面词出现在年报中的词作为初始情绪词典。并使用中科院汉语词法分析系统 ICTCLAS[9]后对年报进行词汇切分。最后依照带惩罚机制词频法计算初始情绪词典中负面词调整后的词频数值，如计算公式(1)所示:

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

其中，$w_{n,N}$为待选负面词 n 在收益为负的年报集合N中的出现次数，$w_{n,P}$为初始负面词在收益为正的年报集合P中出现的次数。待选负面情绪词的负面程度($adjusted\ weight_n$)随着其在负面年报集合出现的比重($\frac{w_{n,N}}{\sum_n w_{n,N}}$)增加,同时也随着其在正面年报集合出现的比重($\frac{w_{n,P}}{\sum_n w_{n,P}}$)下降。通过惩罚机制($\frac{1}{1+\frac{w_{n,P}}{\sum_n w_{n,P}}}$)的引入，带惩罚机制词频法从总体层面上衡量了待选负面词的负面程度。本文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

[7] 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。

[8] LSTM 的结构特殊,通过三个门的加入,适当忘记较远的信息并加强较近的信息,符合论坛帖子短文本的特性。而年报属于长文本, 使用 LSTM 无法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。

[9] 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最高的系统(高达 98%)。

的逻辑，可以得到正式用语情绪词典的正面词表[10]。

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,计算公式如(2):

$$Index_i = Negative_i - Positive_i \quad (2)$$

$$Negative_i = \frac{\sum_n w_{n,i}}{total\ word_i}, Positive_i = \frac{\sum_p w_{p,i}}{total\ word_i}$$

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

(二)非正式用语情绪词典与投资者情绪指标构建

使用2011年4月8日至2016年4月22日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年5月1日至2017年9月30日所有关于上市公司的发帖，共计8130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。附带有明显表情符（emoji）的文本将使得帖子情绪显而易见，以图2为例，该图为雪球论坛上的帖子。首先通过用户在发帖或回帖时所加入的表情:[笑](😊)、[大笑](😁)、[哭泣](😭)、[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，并经过人工检查确保帖子的情绪识别符号和文字表达的情绪相一致，以及确保帖子内容不含广告和其他噪音，最终筛选出8789个股票论坛发帖作为深度学习算法的训练样本。

[此处插入图 2]

本文采用的深度学习算法为长短期记忆网络(LSTM)，是递归神经网络(Recurrent Neural Network, RNN)的一种特殊变体。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三个控制门(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，通过这三个门的控制，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，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 的三个控制门，适当的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。具体 LSTM 模型的运作方式请参考本文附录。该模型在本文数据分类中的准确率达到 87%[11]。本文使用训练好的长短期记忆模型对两个论坛的每一个发帖属于正面或负

[10]为了得到正式用语情绪词典的正面词表，只需把公式（1）种的符号 n,N,P 分别替换成 p,P,N。

[11]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

面情绪进行分类。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，本文参照 Antweiler and Frank(2004)的方法，构建了看涨情绪指标$Bullishness$ 和情绪一致性指标 $Agreement$ 作为检验变量，计算公式如(3):

$$Bullishness_{i,t} = \frac{Positive_{i,t} - Negative_{i,t}}{total\ word_{i,t}}, Agreement_{i,t} = 1 - \sqrt{1 - Bullishness_{i,t}^2} \quad (3)$$

其中$Positive_{i,t}$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，而情绪一致性指标($Agreement_{i,t}$)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。

四、实证检验

(一)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来均自于 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。

[此处插入表 1]

表 2 对回归模型的变量进行了描述性统计，样本数据涵盖 2003-2015 年共 14339 份年报[12]。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公式(2)构建了负面语调指标，分别是 *Index Dalian*、*Index HowNet*、*Index Tsinghua*、*Index LM* 和 *Index Informal*，其中本文对 Loughran and McDonald(2011) 情绪词典进行中文翻译[13]并构建指标 *Index LM*，而 *Index Informal* 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，*Index Dalian* 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 (*Pre FF Alpha*) 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。

[此处插入表 2]

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)。*Index Formal* 和 *Index Informal* 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，*Index Dalian* 和 *Index Tsinghua* 的相关系数(0.804)为最大，表明大连理工

下 1 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。

[12] 本文用于提取正式用语情绪词典的年报有 19970 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 14339 份。

[13] 本文两位作者协同三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，再对这三份各自独立完成的翻译稿综合考虑后得到 Loughran and McDonald(2011)词典的中文翻译。

情感词汇本体库与清华大学褒贬意词典在年报语调的捕捉效果上的差异不大。

[此处插入表 3]

参照 Loughran and McDonald(2011)的实证方法及控制变量的选择，本文使用 Fama-MacBeth (1973)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。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 and McDonald(2011)中所述一致，即年报语调越悲观，未来的交易量与波动率越大，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 *Pre FF Alpha*，*Analyst Dispersion* 和 *Analyst Revisions* 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。

在表 4 中，其他五个词典所衡量的年报负面语调除了 *Index Dalian* 和 *Index Tsinghua* 与下季度未预期盈余显著正相关以及 *Index LM* 与交易量显著正相关之外，其余回归结果皆不显著。*Index Informal* 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 Loughran and McDonald(2011)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。如上所述及回归结果，验证了不同类型的文本应使用相对应的情绪词典，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。

[此处插入表 4]

(二)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 年 4 月 8 日至 2016 年 4 月 22 日中国上市公司[14]在雪球论坛和东方财富论坛上的用户发帖，同样也基于其他词典构建了针对论坛文本的情绪指标，其中 *Bullishness Informal* 和 *Agreement Informal* 是基于非正式用语情绪词典所构建的投资者看涨情绪指标与投资者情绪一致性指标。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

[14] 包括上证 A 股，深证 A 股主板和创业板股票。

连理工情感词汇本体库、中国知网词典、清华大学褒贬意词典及 Loughran and McDonald (2011)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四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。

表 5 为社交媒体情绪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，并控制了每家公司的固定效应。可以发现投资者看涨情绪 *Bullishness Informal* 与超额收益、交易量及波动率显著正相关，这些结论与 Antweiler and Frank(2004)的结果一致，也与 Bollen et al.(2011)和 Chen et al.(2014)等文献的结论保持了一致性。这一结果也符合段江娇等(2017)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。此外投资者情绪一致性 *Agreement Informal* 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 and Frank(2004)、Chen et al.(2014)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。

[此处插入表 5]

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，仅列出关键被解释变量的回归结果。其中 *Bullishness Formal*、*Bullishness Dalian*、*Bullishness Tsinghua* 和 *Bullishness LM* 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，*Bullishness Formal*、*Bullishness Dalian*、*Bullishness HowNet*、*Bullishness Tsinghua* 及 *Bullishness LM* 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，*Agreement Formal*、*Agreement Dalian*、*Agreement HowNet*、*Agreement Tsinghua* 和 *Agreement LM* 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。

[此处插入表 6]

(三)样本外检验

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 and McDonald(2011)和 Antweiler and Frank(2004)的结论保持了一致。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表征。基于以上两个考虑，本文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，本文选取 2016 及 2017 年中国上市公司年报共 9991 份和 2016 年 4 月 23 日至 2018 年 6 月 28 日雪球论坛上的 47.3 万余条用户发帖。实证结果如表 7。Panel A 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表语调的表征，而非样本内累积收益率与成交量等指标相关关系的表征。Panel B 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，证明了本文构建的正式用语情绪词典和非正式用语词典的稳健性。

[此处插入表 7]

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。*Index Combined* 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。*Bullishness Combined* 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。*Agreement Combined* 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。其他如报章新闻等文本，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。

[此处插入表 8]

五、结论

本文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。本文构建的词典包括两个：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财会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语调分析。为检验所构建词典的有效性，本文根据两个情绪词典分别计算了年报语调指标和社交媒体情绪指标，考察这些指标是否能预测上市公司的超额收益、成交量、波动率、下季度未预期盈余等因素，检验情绪指标回归的结果是否和现有的年报语调与投资者情绪相关理论符合，并和广泛使用的情绪词典构建的情绪指标结果做了对比。发现的主要结论如下:

(1)通过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波动有相关性，证明本文提出的正式用语情绪词典在提取年报语调上是有效的。

(2)通过非正式用语情绪词典从股票论坛文本构建了投资者看涨情绪指标和投资者情绪的一致性指标,发现投资者看涨情绪与超额收益正相关且投资者情绪的一致性与交易量及收益波动率显著负相关，说明本文提出的非正式用语情绪词典在量化社交媒体上投资者情绪亦是有效的。

(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。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，还有潜力使用在如分析师报告、财经新闻、微博中财经类发帖等文本分析，关于此类型的研究在国内正处于起步阶段，因此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，并计划对外公开以方便相关研究者使用。

参考文献

段江娇、刘红忠、曾剑平，2017：《中国股票网络论坛的信息含量分析》，《金融研究》，第10期。

李培功、沈艺峰，2010：《媒体的公司治理作用:中国的经验证据》，《经济研究》，第4期。

唐国豪、姜富伟、张定胜，2016：《金融市场文本情绪研究进展》，《经济学动态》，第11期。

谢德仁、林乐，2015：《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》，《会计研究》，第2期。

汪昌云、武佳薇，2015 ：《媒体语气、投资者情绪与IPO定价》，《金融研究》，第9期。

杨道广、陈汉文、刘启亮，2017：《媒体压力与企业创新》，《经济研究》，第8期。

杨晓兰、沈翰彬、祝宇，2016：《本地偏好、投资者情绪与股票收益率: 来自网络论坛的经验证据》，《金融研究》，第12期。

游家兴、陈志锋、肖曾昱、薛小琳，2018：《财经媒体地域偏见实证研究》，《经济研究》，第4期。

Antweiler, W., and M. Z. Frank, 2004, "Is All that Talk Just Noise? The Information Content of Internet Stock Message Boards", *Journal of Finance* ,59(3), 1259—1294.

Bollen, J. , H. Mao, and X. Zeng, 2011,"Twitter Mood Predicts the Stock Market",*Journal of Computational Science* ,2(1),1—8.

Chang, Y., H. G. Harrison, T. Larissa, N. Wang, and B. Zhao, 2015, "Does Diversity Lead to Diverse Opinions? Evidence from Languages and Stock Markets", Stanford University Graduate School of Business research paper.

Chen, H., P. De, Y. J. Hu, and B. H. Hwang, 2014,"Wisdom of Crowds: The Value of Stock Opinions Transmitted through Social Media",*Review of Financial Studies* ,27(5),1367—1403.

Das, S. R., and M. Y. Chen, 2007, "Yahoo! For Amazon: Sentiment Extraction from Small Talk on the Web" ,*Management Science* ,53(9),1375—1388.

Dougal, C., J. Engelberg, D. Garcia, and C. A. Parsons, 2012,"Journalists and the Stock Market", *Review of Financial Studies* ,*25*(3),639—679.

Engelberg, J. E., A. V. Reed, and M. C. Ringgenberg, 2012, "How Are Shorts Informed? Short Sellers, News, and Information Processing", *Journal of Financial Economics* ,105(2) ,260—278.

Fama, E. F., and J. D. MacBeth, 1973, "Risk, Return, and Equilibrium: Empirical Tests", *Journal of Political Economy* ,81(3), 607—636.

Fama, E. F., and K. R. French, 1993, "Common Risk Factors in the Returns on Stocks and Bonds" ,*Journal of Financial Economics* ,33(1),3—56.

Feldman, R., S. Govindaraj, J. Livnat, and B. Segal, 2010, "Management's Tone Change, Post Earnings Announcement Drift and Accruals" , *Review of Accounting Studies* ,15(4),915—953.

Garcia, D., 2013, "Sentiment During Recessions", *Journal of Finance* ,68(3),1267—1300.

Hanley, K. W., and G. Hoberg, 2010, "The Information Content of IPO Prospectuses", *Review of Financial Studies* ,23(7),2821—2864.

Henry, E., 2008, "Are Investors Influenced by How Earnings Press Releases Are Written?",

Journal of Business Communication (1973) ,45(4), 363—407.

Kearney, C., and S. Liu, 2014, "Textual Sentiment in Finance: A Survey of Methods and Models" , *International Review of Financial Analysis* ,33,171—185.

Kim, S., and D. Kim, 2014, "Investor Sentiment from Internet Message Postings and the Predictability of Stock Returns" , *Journal of Economic Behavior & Organization* ,107,708—729.

Li, F., 2010, "The Information Content of Forward-looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach", *Journal of Accounting Research* ,48(5),1049—1102.

Liu, B., and J. J. McConnell, 2013, "The Role of the Media in Corporate Governance: Do the Media Influence Managers' Capital Allocation Decisions?", *Journal of Financial Economics* ,110(1), 1—17.

Loughran, T., and B. McDonald, 2011, "When Is A Liability Not A Liability? Textual Analysis, Dictionaries, and 10-Ks", *Journal of Finance* ,66(1), 35—65.

Loughran, T., and B. McDonald, 2016, "Textual Analysis in Accounting and Finance: A Survey", *Journal of Accounting Research* ,54(4), 1187—1230.

Newey, W. K., and K. D. West, 1986, "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix" , *Econometrica* ,55,703—708.

Price, S. M., J. S. Doran, D. R. Peterson, and B. A. Bliss, 2012, "Earnings Conference Calls and Stock Returns: The Incremental Informativeness of Textual Tone", *Journal of Banking & Finance* ,36(4), 992—1011.

Rogers, J. L., A. V. Buskirk, and S. LC. Zechman, 2011, "Disclosure Tone and Shareholder Litigation" , *The Accounting Review* ,86(6),2155—2183.

Solomon, D. H., E. Soltes, and D. Sosyura, 2014, "Winners in the Spotlight: Media Coverage of Fund Holdings as A Driver of Flows" , *Journal of Financial Economics* ,113(1),53—72.

Tetlock, P. C., 2007, "Giving Content to Investor Sentiment: The Role of Media in the Stock Market" , *Journal of Finance* ,62(3), 1139—1168.

Tumarkin, R., and R. F. Whitelaw, 2001,"News or Noise? Internet Postings and Stock Prices", *Financial Analysts Journal* ,57(3), 41—51.

(四) 可能面对的风险

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。

数据来源:贵州茅台酒股份有限公司 2015 年年度报告

图 1 年报中所隐含的情绪

数据来源:雪球

图 2 股票论坛帖子带有的情绪

表 1 变量名及其含义或计算公式说明

	变量名	含义或计算公式
被解释变量	*Excess Return*	年报发布后[0,+3]日累积收益率减去同期市场累积收益率
	Abnormal Volume	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
	Volatility	年报发布后的[+6，+252]日对数收益率的波动率
	Unexpected Earnings	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	*Index*	年报负面语调，计算方式参考公式(2)
控制变量	*log(Size)*	年报发布前一天的流通股市值取自然对数
	log(Turnover)	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
	log(B/M)	年报发布前一天的账面市值比，取自然对数
	Institution Own	年报发布前一季度的机构投资者比例
	Pre FF Alpha	年报发布前一个月的 Fama-French 三因子模型 Alpha
	Analyst Dispersion	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
	Analyst Revisions	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 年报语调回归相关变量的描述性统计

变量名	样本量	均值	标准差	最小值	最大值
Abnormal Volume	14,244	0.102	0.339	-1.481	3.906
Volatility	12,758	0.501	0.153	0.117	3.129
Unexpected Earnings	13,902	-0.033	0.077	-1.130	0.050
Index Formal	14,339	-0.001	0.002	-0.025	0.013
Index Dalian	14,339	-0.070	0.011	-0.135	-0.015
Index HowNet	14,339	-0.007	0.002	-0.029	0.001
Index Tsinghua	14,339	-0.029	0.009	-0.081	-0.001
Index LM	14,339	-0.025	0.007	-0.036	0.007
Index Informal	14,339	-0.013	0.005	-0.048	0.019
log(Size)	14,339	1.215	0.483	0.648	1.782
log(Turnover)	14,339	-3.792	1.181	-10.126	-0.634
log(B/M)	14,339	-1.190	0.294	-1.536	-0.845
Institution Own	14,339	0.040	0.081	0	0.887
Pre FF Alpha	14,339	-0.243	0.130	-0.697	1.195
Analyst Dispersion	14,339	0.017	0.025	0	0.448
Analyst Revisions	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

	Index Formal	*Index Dalian*	*Index HowNet*	*Index Tsinghua*	*Index LM*	*Index Informal*
Index Formal	1.000					
Index Dalian	0.422**	1.000				
Index HowNet	0.238**	0.417**	1.000			
Index Tsinghua	0.391**	0.804**	0.245**	1.000		
Index LM	0.406**	0.116**	0.132**	0.096**	1.000	
Index Informal	0.171**	0.108**	0.102**	0.094**	0.032**	1.000

注: **表示在 5%的水平下显著

表 4 交易量、波动率和未预期盈余与各年报语调回归结果

解释变量	被解释变量																	
	Abnormal Volume						*Volatility*						*Unexpected Earnings*					
Index Formal	5.948**						3.878**						1.132**					
	(2.01)						(2.40)						(2.15)					
Index Dalian		-0.046						0.931						0.246**				
		(-0.37)						(1.51)						(2.28)				
Index HowNet			-1.987						-1.876						0.698			
			(-1.27)						(-1.48)						(1.80)			
Index Tsinghua				0.053						1.198						0.356***		
				(0.24)						(1.60)						(2.95)		
Index LM					3.201*						1.065						1.001	
					(1.88)						(1.51)						(1.62)	
Index Informal						0.559						0.693						-0.247
						(0.80)						(0.75)						(-1.68)
log(Size)	0.038***	0.008**	0.011***	0.008**	0.033**	0.039***	-0.018***	-0.014**	-0.016**	-0.016**	-0.014*	-0.018***	0.002***	0.002***	0.002***	0.002***	0.002**	0.002**
	(4.92)	(2.28)	(5.86)	(2.02)	(2.14)	(5.96)	(-3.49)	(-2.17)	(-2.14)	(-2.08)	(-1.88)	(-3.60)	(2.51)	(2.97)	(2.96)	(2.83)	(2.03)	(2.23)
log(Turnover)	6.025***	0.041***	0.039***	0.041***	6.407***	6.476***	0.713***	0.032***	0.029***	0.032***	0.705***	0.702***	-0.023***	-0.023***	-0.023***	-0.022***	-0.22***	-0.023***
	(15.61)	(17.04)	(18.58)	(17.35)	(15.07)	(15.54)	(3.81)	(5.19)	(5.04)	(5.17)	(2.56)	(3.79)	(-6.55)	(-7.24)	(-7.14)	(-7.15)	(-7.01)	(-6.67)
log(B/M)	0.034***	0.011***	0.009***	0.012***	0.034***	0.035***	-0.051***	-0.046***	-0.047***	-0.043***	-0.049***	-0.055***	0.008	0.007	0.008	0.007	0.007	0.009
	(4.63)	(4.07)	(3.27)	(3.27)	(4.46)	(5.20)	(-4.85)	(-4.16)	(-4.64)	(-4.05)	(-4.12)	(-4.45)	(1.27)	(1.51)	(1.57)	(1.45)	(1.34)	(1.29)
Institution Own	0.427	0.167	0.128	0.176	0.429	0.471	-0.164	-0.163	-0.166	-0.145	-0.166	-0.083	0.158***	0.157***	0.157***	0.161***	0.157***	0.155***
	(1.01)	(0.98)	(0.99)	(0.98)	(1.02)	(1.03)	(-1.16)	(-1.28)	(-1.26)	(-1.29)	(-1.18)	(-1.24)	(4.12)	(4.61)	(4.77)	(4.72)	(3.92)	(3.95)
Pre FF Alpha	0.132	0.022	0.040	0.020	0.125	0.136	0.023	0.002	0.030	0.002	0.024	0.015	0.074**	0.072**	0.073**	0.073**	0.072**	0.073**
	(1.31)	(0.52)	(1.45)	(0.43)	(1.27)	(1.32)	(0.61)	(0.07)	(0.79)	(0.08)	(0.63)	(0.37)	(2.36)	(2.36)	(2.36)	(2.38)	(2.27)	(2.35)
Analyst Dispersion													0.646***	0.640***	0.646***	0.649***	0.644**	0.651***
													(2.57)	(2.95)	(2.85)	(3.08)	(2.10)	(2.61)
Analyst Revisions													-0.594***	-0.591***	-0.596***	-0.589***	-0.582***	-0.602***
													(-4.17)	(-4.12)	(-4.18)	(-4.09)	(-4.02)	(-4.15)
常数	-0.117***	0.203***	0.187***	0.207***	-0.108***	-0.106***	0.440***	0.641***	0.561***	0.613***	0.426***	0.445***	-0.078***	-0.063***	-0.075***	-0.069***	-0.077***	-0.079***
	(-3.66)	(19.20)	(9.15)	(24.67)	(-3.20)	(-2.67)	(9.95)	(9.53)	(9.02)	(9.71)	(9.25)	(8.36)	(-5.66)	(-4.06)	(-6.37)	(-5.04)	(-5.51)	(-5.77)
样本量	14,244	14,244	14,244	14,244	14,244	14,244	12,758	12,758	12,759	12,758	12,758	12,758	13,898	13,898	13,898	13,898	13,898	13,898
Avg.R^2	0.712	0.437	0.454	0.439	0.710	0.705	0.315	0.359	0.349	0.364	0.312	0.321	0.053	0.054	0.053	0.055	0.052	0.053

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

表 5 超额收益、交易量及波动率与非正式词典构建的投资者情绪指标回归结果

解释变量	被解释变量					
	Excess Return		*Abnormal Volume*		*Volatility*	
Bullishness Informal	0.001**		0.128***		0.001**	
	(2.17)		(2.44)		(2.14)	
Agreement Informal		-0.001***		-0.149**		-0.001**
		(-3.08)		(-2.00)		(-2.13)
log(Size)	-0.001***	-0.001***	-0.040**	-0.042**	0.000***	0.001***
	(-6.81)	(-7.22)	(-2.08)	(-2.16)	(6.52)	(6.85)
log(Turnover)	-0.001***	-0.001***	0.067***	0.744***	0.000***	0.001***
	(-5.02)	(-4.86)	(3.40)	(3.29)	(8.70)	(9.06)
log(B/M)	0.000***	0.000***	0.055***	0.054***	-0.000***	-0.001***
	(2.69)	(3.04)	(2.51)	(2.43)	(-5.64)	(-5.98)
常数	0.012***	0.014***	1.157***	1.119***	-0.004***	-0.004***
	(5.08)	(5.64)	(2.68)	(2.58)	(-3.01)	(-3.09)
样本量	564,887	564,887	564,822	564,822	564,886	564,886
Adj.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 超额收益、交易量与波动率与基于其他词典的投资者情绪指标回归结果

解释变量	被解释变量		
	Excess Return	*Abnormal Volume*	*Volatility*
Bullishness Formal	-0.001**	0.368***	-0.002***
	(-2.41)	(9.34)	(-18.50)
Bullishness Dalian	-0.001***	0.391***	-0.004***
	(-4.70)	(9.19)	(-33.38)
Bullishness HowNet	-0.001***	-0.352***	-0.003***
	(-8.73)	(-9.27)	(-30.72)
Bullishness Tsinghua	0.000	0.486***	-0.003***
	(0.42)	(11.97)	(-26.43)
Bullishness LM	-0.001*	0.276***	-0.002***
	(1.82)	(8.24)	(-16.57)
Agreement Formal	-0.003***	-0.668***	-0.002***
	(-11.19)	(-13.15)	(-21.30)
Agreement Dalian	-0.003***	-0.720***	-0.004***
	(-10.71)	(-12.05)	(-26.07)
Agreement HowNet	-0.004***	-1.066***	-0.004***
	(-15.27)	(-20.77)	(-29.78)
Agreement Tsinghua	-0.003***	-0.620***	-0.003***
	(-8.49)	(-11.89)	(-26.16)
Agreement LM	-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

表 7 样本外回归结果

Panel A:年报语调回归结果

解释变量	被解释变量		
	Abnormal Volume	*Volatility*	*Unexpected Earnings*
Index Formal	6.127***	4.128**	1.376***
	(2.62)	(2.15)	(2.47)
log(Size)	0.026**	-0.012***	0.002**
	(2.02)	(-3.26)	(2.16)
log(Turnover)	5.768***	0.662***	-0.019***
	(11.21)	(3.64)	(-5.15)
log(B/M)	0.026***	-0.036**	0.007
	(3.89)	(2.02)	(1.10)
Institution Own	0.416	-0.157	0.149***
	(0.97)	(-1.14)	(3.86)
Per FF Alpha	0.126	0.020	0.071**
	(1.28)	(0.58)	(2.20)
Analyst Dispersion			0.602**
			(2.20)
Analyst Revision			-0.561***
			(-3.81)
常数	-0.106***	0.376***	-0.066***
	(-3.43)	(8.19)	(-5.00)
样本量	9,576	9,481	9,812
Avg.R^2	0.726	0.378	0.062

Panel B:论坛投资者情绪回归结果

解释变量	被解释变量					
	Excess Return		*Abnormal Volume*		*Volatility*	
Bullishness Informal	0.001***		0.131***		0.001***	
	(2.42)		(2.69)		(2.58)	
Agreement Informal		-0.001***		-0.152**		-0.001***
		(-3.25)		(-2.09)		(-2.37)
log(Size)	-0.001***	-0.000***	-0.037**	-0.041**	0.000***	0.001***
	(-5.96)	(-6.01)	(-1.98)	(-2.14)	(5.66)	(5.70)
log(Turnover)	-0.001***	-0.001***	0.062***	0.701***	0.000***	0.001***
	(-4.34)	(-4.02)	(3.07)	(3.10)	(7.83)	(8.87)
log(B/M)	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	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
Adj.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

表 8 基于合并词典的年报语调与投资者情绪回归结果

Panel A:年报语调回归结果

解释变量	被解释变量		
	Abnormal Volume	*Volatility*	*Unexpected Earnings*
Index Combined	5.012*	3.076**	1.566**
	(1.80)	(2.06)	(2.11)
log(Size)	0.033***	-0.017**	0.002***
	(3.64)	(-2.07)	(2.46)
log(Turnover)	6.407***	0.722***	-0.021***
	(14.86)	(3.62)	(-6.83)
log(B/M)	0.036***	-0.050***	0.007
	(4.78)	(-4.66)	(1.30)
Institution Own	0.398	-0.160	0.153***
	(0.95)	(-1.14)	(3.77)
Pre FF Alpha	0.127	0.022	0.072**
	(1.25)	(0.59)	(2.09)
Analyst Dispersions			0.634**
			(2.17)
Analyst Revisions			-0.560***
			(-3.81)
常数	-0.101***	0.433***	-0.076***
	(-2.68)	(8.76)	(-5.54)
样本量	14,244	12,758	13,898
Avg.R^2	0.710	0.312	0.052

Panel B:论坛投资者情绪回归结果

解释变量	被解释变量					
	Excess Return		*Abnormal Volume*		*Volatility*	
Bullishness Combined	0.002**		0.119**		0.002**	
	(2.18)		(1.99)		(2.21)	
Agreement Combined		-0.001*		-0.102*		-0.001*
		(-1.82)		(-1.83)		(-1.84)
log(Size)	-0.001***	-0.001***	-0.041**	-0.043**	0.001***	0.001***
	(-5.27)	(-8.69)	(-2.11)	(-2.17)	(6.95)	(5.75)
log(Turnover)	-0.001***	-0.001***	0.065***	0.759***	0.000***	0.001***
	(-4.64)	(-5.27)	(3.14)	(3.77)	(7.26)	(8.12)
log(B/M)	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	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
Adj.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

附录

长短期记忆网络模型(LSTM)

在文本分析的任务中，需要将一段文本转换为计算机可处理的形式，转换的过程有两种方法:独热编码(One-Hot Encoding)及词嵌入(Word Embedding)，因独热编码方法无法很好地将词与词间的关系联系起来，本文使用词嵌入作为转换文本的方法。

1.词嵌入方法

在语言学中，有意义的文本之最小单位是词，在处理情绪分析问题时，同样要以词出发。以一段文本为例子:"大盘跌成这样，我亏多了"，首先对这段文本进行分割并删除无意义的词，即分割成:"大盘"、"跌"、"这样"、"我"、"亏"、"多"。而词与词之间的关系在整段文本所表达的含义占有重要地位，本文使用 Skip-gram 模型量化词与词之间的关系。

Skip-gram 模型为无监督式机器学习方法。其中心思想为给定一个词，通过训练好的模型，预测出该词前后词出现的概率，即给定"跌"，预测"大盘"和"这样"出现的概率；给定"这样"，预测"跌"和"我"出现概率；给定"我"，预测"这样"和"亏"的出现概率；给定"亏"，预测"我"和"多"的出现概率。通过如此滑动窗口的数据采样过程，便能够很好地将词与词间的关系考虑进来。以下叙述模型的运作过程：

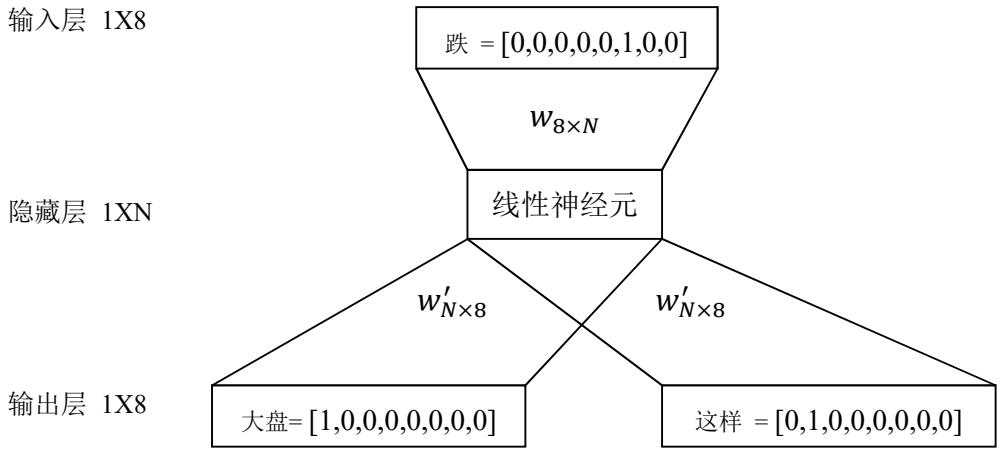
假设训练数据为两段文本:"大盘跌成这样，我亏多了"和"大盘涨成这样，我赚多了"，分割成单词序列，即"大盘"、"跌"、"涨"、"这样"、"我"、"亏"、"赚"、"多"，并依照出现次数进行排序。接下来对单词进行编码。

语料库	出现次数
大盘	2
这样	2
我	2
多	2
跌	1
涨	1
赚	1
亏	1

编码 →

大盘 = [1,0,0,0,0,0,0,0]
这样 = [0,1,0,0,0,0,0,0]
我 = [0,0,1,0,0,0,0,0]
多 = [0,0,0,1,0,0,0,0]
跌 = [0,0,0,0,1,0,0,0]
涨 = [0,0,0,0,0,1,0,0]
赚 = [0,0,0,0,0,0,1,0]
亏 = [0,0,0,0,0,0,0,1]

编码完成之后，并依照单词的前后文关系成为 Skip-gram 模型的输入与输出，如附录图 1:

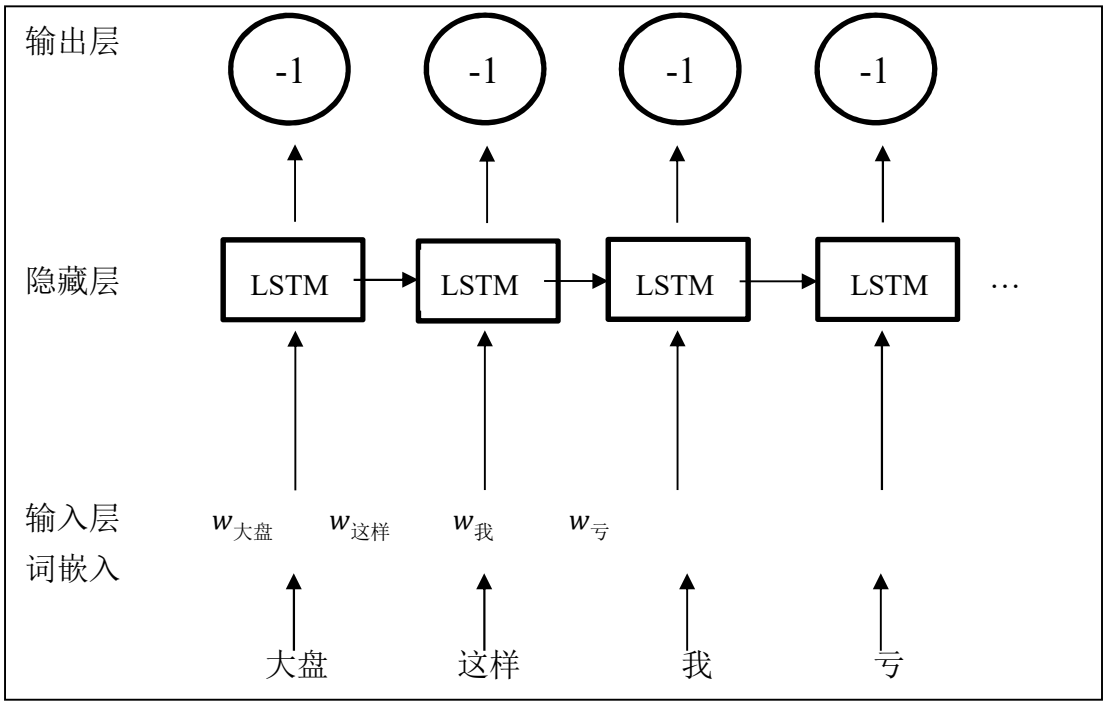


附录图 1

借由整个语料库及其上下词，Skip-gram 模型使得每个输入单词依次经过权重矩阵 W 和 W' 处理后能够接近其上下词所代表的向量。其中，隐藏层向量 w 能够从整体语料库的层面上代表该单词，也完成了"跌"这个单词向量化的过程。

2. 长短期记忆网络模型

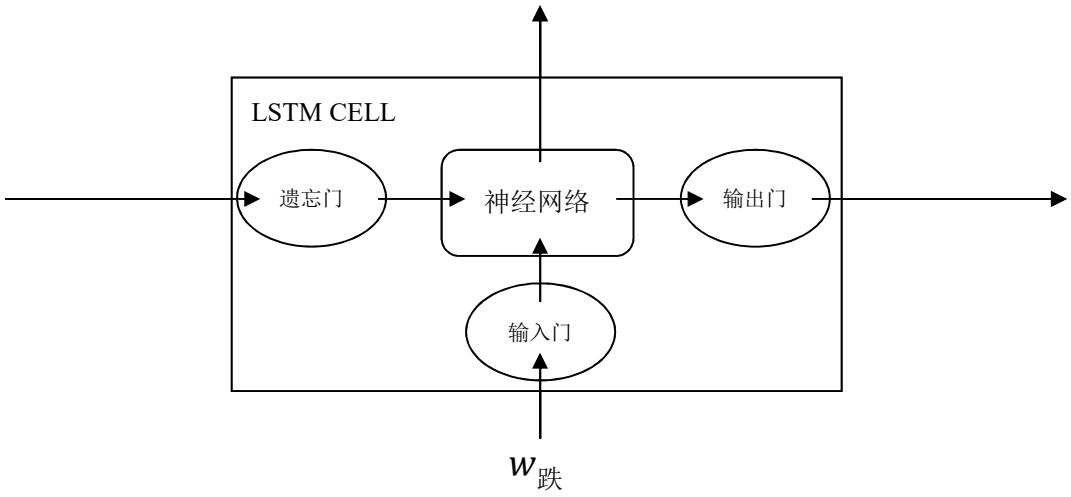
本文使用已判定为正负面情绪的文本，训练长短期记忆模型，并对其他未判定的文本进行情绪预测。其中，模型的输入为向量化后的单词，输出则是正负面情绪的标签，+1 为正面，-1 为负面。具体模型运作方式附录图 2。同样以"大盘跌成这样，我亏多了"为例子，这是一个负面情绪的句子，因此附录图 2 的输出层的标签为-1。



附录图 2

由输入层开始，模型将一段句子的单词出现顺序视为时间序列。从第一个词开始，词向量"大盘"首先进入隐藏层，并拟合出权重矩阵逼近输出层的值，即-1，下一步，此权重矩阵乘上词向量"跌"， 接着同样通过隐藏层拟合出新的权重矩阵进入下一个 LSTM 神经网络，如此进行到最后一个词向量"多"，最终得到在文字出现序列上不断迭代更新的权重矩阵。LSTM 模型能够学习长期的依赖关系，LSTM 中有一个四层的网络，它们以一种特殊的方式交互着，能够识别需要抛弃的信息和需要保留的信息，对于每一个 LSTM 神经元，它的输入值是前一个 LSTM 神经元和当前输入的词向量， LSTM 学习到的权重向量使得输入值的重要部分被存储下来并作为下一个 LSTM 神经元的输入，因此它可以识别时间序列中的长期信息，一段句子中单词有先后顺序，可视为时间序列，在词向量序列长期记忆表现为上下文信息， 应用 LSTM 可以捕获词与词之间的关系。

附录图 2 里面的 LSTM CELL 具体如附录图 3 所示。每一个 LSTM CELL 皆拥有神经网络，且权值共享，即训练过程中不仅迭代更新当前 LSTM CELL 的神经网络权值，且同时更新所有 LSTM CELL，借此达到考虑了上下文关系的功能。当更新所有 LSTM CELL 的权值时，通过遗忘门和输出门的控制来防止模型失效(梯度消失或梯度爆炸)。



附录图 3

附录表 1 部分正式用语情绪词典及非正式用语情绪词典

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

附录表 2 投资者情绪回归相关变量的描述性统计

变量名	样本量	均值	标准差	最小值	最大值
Excess Return	569,198	-0.0003	0.077	-0.815	0.556
Abnormal Volume	569,133	1.378	13.235	-45.144	2202.645
Volatility	569,197	0.064	0.040	0	0.271
Bullishness Informal	564,892	-0.403	0.340	-1	1
Bullishness Formal	461,425	0.416	0.522	-1	1
Bullishness Dalian	545,779	0.267	0.426	-1	1
Bullishness HowNet	462,392	0.323	0.551	-1	1
Bullishness Tsinghua	496,216	0.493	0.477	-1	1
Bullishness LM	552,137	0.012	0.146	-1	1
Agreement Informal	557,261	0.455	0.315	0	1
Agreement Formal	461,425	0.381	0.414	0	1
Agreement Dalian	545,779	0.192	0.308	0	1
Agreement HowNet	462,392	0.356	0.419	0	1
Agreement Tsinghua	492,216	0.379	0.378	0	1
Agreement LM	552,137	0.323	0.406	0	1
log(Size)	569,201	22.613	0.968	18.959	28.144
log(Turnover)	569,199	-3.550	0.989	-11.512	-0.243
log(B/M)	569,201	-1.469	0.829	-9.873	1.693

附录表 3 投资者看涨情绪与情绪一致性相关系数表

	Bullishness Informal	*Bullishness Formal*	*Bullishness Dalian*	*Bullishness HowNet*	*Bullishness Tsinghua*	*Bullishness LM*
Bullishness Informal	1.000					
Bullishness Formal	0.107**	1.000				
Bullishness Dalian	-0.142**	0.534**	1.000			
Bullishness HowNet	-0.035**	0.225**	0.512**	1.000		
Bullishness Tsinghua	-0.031**	0.663**	0.640**	0.177**	1.000	
Bullishness LM	-0.021**	0.477**	0.504**	0.519**	-0.037**	1.000
	Agreement Informal	*Agreement Formal*	*Agreement Dalian*	*Agreement HowNet*	*Agreement Tsinghua*	*Agreement LM*
Agreement Informal	1.000					
Agreement Formal	0.178**	1.000				
Agreement Dalian	0.315**	0.409**	1.000			
Agreement HowNet	0.152**	0.428**	0.416**	1.000		
Agreement Tsinghua	0.133**	0.572**	0.483**	0.359**	1.000	
Agreement LM	0.441**	0.468**	0.472**	0.436**	0.506**	1.000

注:**表示在 5%的水平下显著