第 21 卷第 4 期　　经 济 学（季 刊）　　Vol. 21, No. 4
2021 年 7 月　　China Economic Quarterly　　July, 2021

媒体文本情绪与股票回报预测

姜富伟　孟令超　唐国豪*

摘　要　本文在 Loughran and MacDonald（2011）词典的基础上通过人工筛选和 word2vec 算法扩充，构建了一个更新更全面的中文金融情感词典。我们使用该情感词典计算我国财经媒体文本情绪指标，发现媒体文本情绪可以更准确地衡量我国股市投资者情绪的变化，对我国股票回报有显著的样本内和样本外预测能力。媒体文本情绪对一些重要的宏观经济指标也有显著的预测能力，具有重要的学术和实践应用价值。

关键词　媒体文本情绪，情感词典，收益预测
DOI：10.13821/j.cnki.ceq.2021.04.10

一、引　　言

投资者情绪是行为金融和资产定价研究领域的重点研究内容，大量研究表明非理性投资者情绪可以使公司股价偏离它们的基本面价值（De Long *et al.*，1990）。投资者情绪研究的难点与重点在于如何测量投资者情绪（Baker and Wurgler，2007）。以 Baker and Wurgler（2006，2007）和 Huang *et al.*（2015）为例，过去的研究大多通过选取一些与股票市场交易相关的指标来计算投资者情绪。近年来，随着文本分析技术的应用，大量学者开始通过提取金融文本中的情绪信息展开更深入的研究（Li，2006；Jiang *et al.*，2019）。

本文基于人工智能文本分析方法，从财经新闻报道中提取投资者情绪信息。文本情绪分析技术为投资者情绪领域的相关研究开创了广阔的新空间，并具有四个突出的优点：（1）媒体文本与金融市场交易数据信息互补性强，在以往的研究中往往被忽视；（2）文本数据规模大，从这一海量数据中提取情绪信息有利于减少测量误差；（3）属于直接情绪测度；（4）考虑到文本数据更新的高频性，我们有机会构建日频甚至分钟频等更高频的情绪指数。

* 姜富伟，中央财经大学金融学院；孟令超，北京大学经济学院；唐国豪，湖南大学金融与统计学院。通信作者及地址：孟令超，北京市海淀区颐和园路 5 号北京大学经济学院，100871；电话：18810712516；E-mail：lingchao_meng@pku.edu.cn。本文感谢国家自然科学基金（72072193、72003062、71872195）和中央财经大学创新群体项目资助；感谢北京大学、中国人民大学、上海交通大学、中央财经大学以及邹恒甫、周国富、沈艳、黄卓、王一鸣、许志伟、《经济学》（季刊）编辑部和三位匿名审稿人对本文的宝贵意见，当然文责自负。

近几年来，国内也有不少文献开始探究中文文本中蕴含的情绪信息，例如杨晓兰等（2016）、游家兴和吴静（2012）以及汪昌云和武佳薇（2015）等。同时，Li *et al*.（2019）试图从论坛数据中构建文本情绪指数。现有研究仍有如下拓展空间：第一，缺乏一个广泛接受的、专业的、开源的中文金融情感词典。金融领域的文本分析大多使用情感词典法，但是当前的中文情感词典均为通用情感词典。这些词典中含有在金融语境下不再适用的词语，还会遗漏许多金融专有情感词汇，在金融语境下适用性不强。因此，一个专门的金融情感词典是必不可少的。第二，当前大部分研究中使用的文本来自股票论坛或上市公司公告，对经济金融新闻报道文本的研究仍然不足。然而，作为市场广大投资者接触信息的最主要渠道，新闻报道文本中蕴含的情绪对市场存在着无法忽视的影响。第三，由于文本数据量级较大，难以收集与处理，因此现有文献中构建的文本情绪时间跨度均较短，这对研究金融市场与文本情绪之间的关系产生较大制约。

本文基于我国中文财经新闻文本构建了一个全新的中文金融情感词典。[1]在国外研究中，Loughran and MacDonald（2011）专门构建了一个英文的金融情感词典（以下称作LM词典），这一词典一经推出便得到了广泛使用[2]，极大促进了英文金融文本分析的研究。然而当前国内缺乏一个类似的、被广泛接受且公开的中文金融情感词典，本文的研究尝试填补这一空白，为未来金融文本分析的研究奠定良好基础。在构建中文金融情感词典时，本文应用了“洋为中用”“古为今用”和word2vec算法扩充三种方法。“洋为中用”指的是将英文的LM词典转化为中文版本，“古为今用”指的是从现有中文通用情感词典中筛选出适合金融语境的词语。为保证词典完备性，本文还利用word2vec算法（一种深度学习算法），从语料中挖掘与前两部分词语高度相关并具有合适情感倾向的词语。通过将三种方法得到的词语合并，本文得到了最终的中文金融情感词典。

在金融情感词典构建方面，本文有以下三点贡献：第一，将LM词典进行了中文化。LM情感词典是一款广为应用的英文情感词典，但是目前尚无一款可以直接应用的中文版LM情感词典存在，为此，本文通过爬虫程序对LM词典进行了全面客观的翻译。第二，对比了LM词典与中文通用词典的优劣性。尽管在英文语境下金融情感词典的必要性得到了证明与认可，但在中文语境下这一点尚缺乏直接检验。为此，我们对比了几款应用较为广泛的通用情感词典与中文LM词典的表现。从实证结果来看，中文LM词典的表现的确优于通用情感词典。第三，我们针对中文语境对LM词典进行了扩展

[1] 该词典可以下载自https://github.com/MengLingchao/Chinese_financial_sentiment_dictionary，访问时间：2021年7月8日。

[2] 截至2021年7月8日，Loughran and MacDonald（2011）的最新引用次数为3 079。

与完善，从而最终构建了更全面的中文金融情感词典。LM 词典本身是一个英文金融情感词典，更加适用于英文语境。因此，将 LM 词典简单翻译为中文后仍然有很大的改进空间。本文在此基础上通过从通用情感词典中筛选及 word2vec 扩充两种方式，对中文 LM 词典进行了扩充和改进，从而形成了最终的中文金融情感词典。

接着我们应用金融情感词典计算了我国股票市场媒体文本情绪指数。本文语料数据来源于 infobank 数据库中的经济新闻库，这一数据库收集了自 1992 年以来权威纸质媒体和主流互联网网站上的经济金融新闻报道，时间跨度极长且覆盖范围极广。借助这一语料数据，本文得以计算了 1992 年以来的文本情绪指数，涵盖了中国股票市场的大部分历史。为了对比本文构建的中文金融情感词典与现有通用情感词典的优劣性，本文也同时利用通用情感词典计算了文本情绪序列，以作为对照。

然后我们探究了媒体文本情绪对我国整体股票市场的样本内与样本外预测能力。在样本内预测检验部分，本文发现文本情绪指标可以显著地正向预测整体股票市场回报，说明投资者们会受到金融新闻中文本情绪的影响，并根据情绪来调整自己的投资决策。他们在情绪积极时，会将更多资金投入股票市场中，或采取一些更为激进的投资方式，从而使得市场回报增加。在样本外预测检验方面，本文应用 R^2_{OS} 统计量来评估文本情绪指标对市场回报的样本外预测能力。检验结果显示，文本情绪指标的样本外预测能力显著强于历史均值预测基准。上述结果只在以金融情感词典计算的媒体文本情绪中才显著。使用通用情感词典计算的文本情绪的实证表现明显弱于金融词典文本情绪，没有显著的预测能力。这是因为通用情感词典含有大量在金融语境下不适用的词汇，还会遗漏许多金融专业词汇，使用这种词典不能捕捉到真实准确的金融文本情绪信息。总之，本文构建的金融情感词典在金融经济语境下的适用程度要远强于通用情感词典。

我们还进行了资产配置检验，发现投资者使用文本情绪指标进行资产配置时，可以获得正的确定性等价收益，说明文本情绪指标具有较强的实际投资价值。本文还对比了文本情绪与常见经济指标对整体股票市场的预测能力。从实证结果来看，一方面，文本情绪的单变量预测能力不弱于常见经济指标，另一方面，将文本情绪指标纳入经济指标预测方程时，可以显著增强经济指标的预测能力，说明文本情绪指标含有增量信息。这些发现都揭示出了文本情绪指标极强的应用价值。

本文最后探究了文本情绪预测能力的来源与作用机制。首先，本文发现媒体文本情绪可以显著地影响投资者对宏观经济的预期，而投资者会根据预期调整金融市场参与程度，从而让市场回报产生相应的反应。其次，本文发现基于风险补偿的理论并不能很好地解释文本情绪的预测能力，说明这一指标对股票产生影响的方式更加符合 De Long *et al*.（1990）噪音交易者模型中

的非理性传播渠道。

后文行文结构如下：第二部分是文献综述，第三部分介绍数据来源，第四和第五部分分别介绍文章构建词典与计算文本情绪的方法，第六部分为实证研究部分，最后为文章结论。

二、文献回顾

文本情感分析技术近年来在经济金融领域得到了广泛的应用。Tetlock（2007）首先研究了媒体新闻报道对市场的影响，发现媒体报道情绪对整体经济走势存在影响。随后的研究显示各类文本中蕴含的情绪均对市场存在显著影响。例如，Loughran and MacDonald（2011）发现公司年报文件中负面词汇比例越高，未来的股票收益降幅越大。Jiang *et al*.（2019）则从上市公司年报和电话会议提取了管理层经理人情绪指标，并发现这一指标对于市场回报有着很强的预测能力。唐国豪等（2016）、沈艳等（2019）和Zhou（2018）等对文本分析技术进行了细致的文献总结。

在国内，也有不少学者利用文本分析技术对相关问题进行了研究。Li *et al*.（2019）在构建中国市场文本情绪方面做出了探索。游家兴和吴静（2012）发现新闻报道中的情绪会使得股票价格偏离基本价值，并且这种影响在公司信息不透明时更加明显。杨晓兰等（2016）利用东方财富网股吧评论数据研究了本地关注和帖子文本情绪对股票收益率的影响。汪昌云和武佳薇（2015）则发现公司上市前新闻媒体报道的语气会对IPO抑价率产生影响。

现有文献大多基于文本情感词典的方法计算文本情绪，这种方法认为文章的情感由情感词语决定。因此，情感词典的质量决定了文本分析的质量。由于金融语境的特殊性，使用常见的通用情感词典在处理金融相关文本时会产生偏差。Li（2010）提出通用情感词典无法区分部分在金融领域中被认为是积极或消极的词语。而Loughran and McDonald（2011）则对通用情感词典在金融语境下适用性的问题进行了更加详细的分析。他们发现，通用情感词典中73.8%的负面词汇在金融语境下都不再具有负面含义，说明使用通用情感词典分析金融文本可能产生极大的误差。为了克服上述问题，Loughran and MacDonald（2011）结合词语在金融文本语料中的具体使用情况，筛选出了合适的词语组成金融情感词典（LM词典）。LM词典在金融会计领域的文本分析研究中得到了广泛的运用，对国外金融文本分析的研究起到了极大的促进作用。

在中文语境下，常见的中文通用情感词典同样存在与金融文本契合度不高的问题。[3]但是目前尚没有被广泛认可的中文金融情感词典，这极大地制约

[3] 在附录中，我们列举了中文通用情感词典中与金融语境不契合的示例词语。限于篇幅，附录从略，留存备索。

了中文金融情感分析的发展。一些现有文献尝试构建这样一个词典，但本文所构建的词典在方法论与适用范围上与这些词典具有较大区别。例如汪昌云和武佳薇（2015）从现有的多个词典中筛选词语组成情感词典，未考虑从语料中发现新词，而 Yan *et al*.（2019）中的情感词典则专门适用于招股说明书文本。与之对比，本文金融情感词典的构建方法更加全面，覆盖范围更加广泛，对于改进中文金融文本分析效果具有十分重大的意义。

三、数据来源

本文对财经新闻媒体报道的文本情绪进行研究，这类文本覆盖内容广，受众人群多，对市场存在着较大的潜在影响。文本素材来源于 infobank 数据库中的经济新闻库。该数据库收集自 1992 年至今的新闻数据，全部新闻数据均为每日更新。图 1 展示了数据库中各年份的新闻条数。其中，1992 年新闻条数最少，但也超过了两万条，其余各年份新闻条数基本都在十万条以上，大量年份新闻条数达到了二三十万，2001 年新闻条数最多，高达四十万。这一数据量可以满足研究需求。infobank 数据库中的新闻来源主要包括三类。第一类为综合性报纸，代表有《人民日报》《光明日报》等，在这一类报纸上，通常会报道与整体经济形势以及国家大政方针相关的新闻，会对整个金融市场产生全面性的影响。第二类为专业性的财经类报纸，典型代表包括《中国证券报》《经济日报》等，这一类报纸上通常报道与经济金融直接相关的新闻。第三类为财经类网站，例如东方财富、新浪财经等。财经类网站的新闻报道越来越成为投资者们接触金融市场相关信息的主要渠道，因而搜集网站上的新闻报道也是非常有必要的。这三类素材基本覆盖了我国市场上经济金融新闻报道的主要来源。宏观数据以及股票市场数据分别来自中经网数据库以及国泰安数据库。

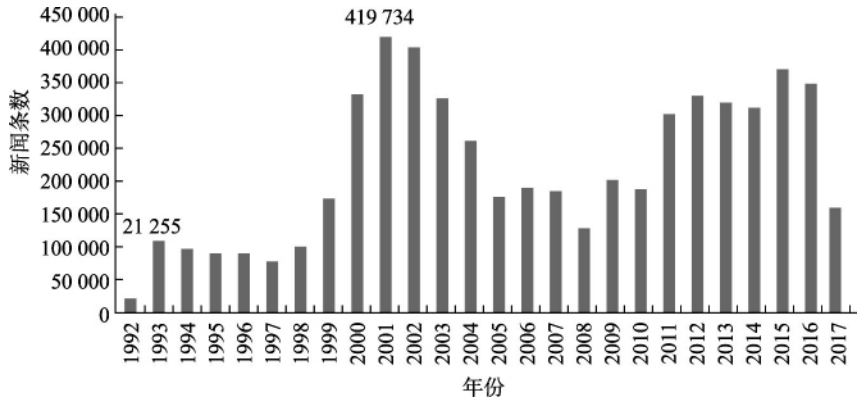

图 1　各年份新闻数据条数

四、词典构建

（一）整体方法

本文按照图 2 思路构建中文金融情感词典。我们将英文 LM 金融词典转化为中文版本[4]（洋为中用），并从中文通用情感词典[5]中筛选出在金融语境下适用的情感词汇（古为今用）。为了避免金融情感词语的遗漏，我们利用 word2vec 算法（一种深度学习算法）从语料中找到与前两部分词语高度相关并且具有合适情感倾向的词语，从而扩充词典。最后，将上述三种方法得到的词语合并去重后得到最终的中文金融情感词典。

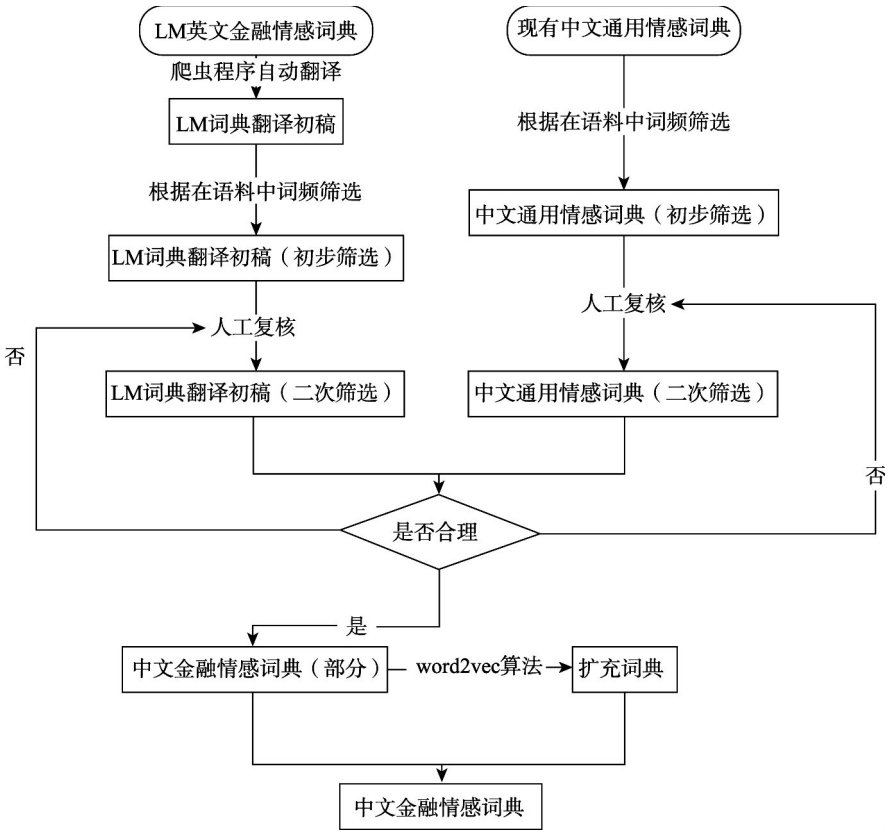

图 2 词典构建思路

[4] 利用爬虫程序从有道词典网站爬取词语释义，自动实现对 LM 金融情感词典的翻译。

[5] 在古为今用部分，为了避免不同通用情感词典之间差异的影响，同时也为了保证词语的完备性，我们将三个应用程度较为广泛的词典［知网 Hownet 情感词典（简称知网词典）、清华大军李军中文褒贬义词典（简称清华词典）以及台湾大学 NTUSD 简体中文情感词典（简称 NTUSD 词典）］合并去重，以此作为所使用的通用情感词典。如无特殊说明，后文所提到的通用情感词典均指此三个词典的合集。

（二）情感词语的筛选

我们对 LM 词典中文翻译与通用词典中不符合金融语境的词语进行筛选。第一种筛选是根据词频筛选。我们对语料进行分词处理[6]，并去除停用词（的、了、呢等无意义词以及标点符号），随后统计词频数。如果某词语的词频数为 0，说明该词语在金融语境下基本不会出现，属于无意义词语或不相关词语，可以直接删除。其他词语继续通过人工方式审核。

人工筛选主要剔除了如下三类词语。第一类为不具有情感倾向的词语，例如财务报告、董事会等。第二类为具有情感倾向但与金融语境不相关的词语，例如“寡廉鲜耻”等。如果一个词语与金融相关度较低且词频数不高，那么也予以剔除处理。第三类为在金融语境下意义发生变化的词语。例如，“负债”在日常语境下具有负面情感，它用以描述个人承担着债务。但在金融语境下，“负债”是企业日常经营活动中所必然产生的特征，它只是企业状态的一种客观描述，并不具备明显的情感倾向，如资产负债表中的“负债”不具备负面倾向。

（三）word2vec 算法与词典扩展

我们进一步利用 word2vec 算法从语料库中训练词语向量并计算词语相似度，从而提取与前文筛选后情感词语高度相关的词语，并挑选出具有合适情感倾向的词语，以此实现发现新词并扩充词典的目的。

word2vec 算法由 Mikolov *et al*.（2013）提出，它是一种深度学习算法，可以将词语转化为相对低维的向量，且向量元素含有词语信息。从而可以根据 word2vec 向量计算词语相似度，这是其相对于传统词语向量表示法的最大优点。word2vec 模型包含 CBoW 模型和 Skip-gram 模型两种形式。CBoW 模型根据文本上下文对目标词进行预测，而 Skip-gram 根据目标词对文本上下临近词语进行预测，如图 3 所示。根据 Mikolov *et al*.（2013），Skip-gram 模型估计准确率更高，且在低频词上表现更加明显。考虑到本文利用 word2vec 模型的目的是发现新情感词汇，这些词汇中难免会存在一些低频词汇，本文应用 skip-gram 模型进行估计，并将词语向量维度设置为 200 维。

对于两个给定的词语向量 A 和 B，余弦相似度的计算方法为：

$$similarity=\cos(\theta)=\frac{A\times B}{|A||B|}=\frac{\sum_{i=1}^{n}A_i\times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\times\sqrt{\sum_{i=1}^{n}(B_i)^2}}. \quad (1)$$

[6] 使用 Python 中的 Jieba 分词库完成。

余弦相似度越大，代表两个向量越接近，即两个词语越相关。我们根据余弦相似度对每个词语选取了五个相似词语，并在此基础上进行后续筛选，以得到 word2vec 情感词语扩充集。

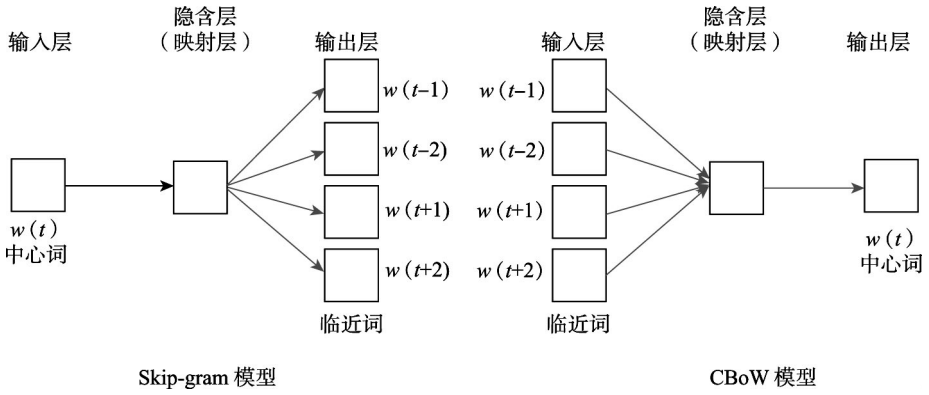

图 3 word2vec 算法中 Skip-gram 和 CBoW 模型示意图

（四）词典展示

表 1 展示了我们构建的中文金融情感词典的相关信息。[7] 占原词典比例指的是该部分词语占其来源词典词语数量的比例，即词语数量/来源通用情感词典词语数。我们发现现有通用情感词典中能够入选金融情感词典的比例最大只能达到 40%左右，大量词语在金融语境下不再适用，这突出了构建金融情感词典的必要性。

表 1 中文金融情感词典的构成信息

消极词语部分（5 890）				
来源		词语数量	原词典总词数	占原词典比例（%）
LM 词典中文翻译		1 562		
通用词典筛选	清华词典	1 945	4 469	43.5
	知网词典	534	1 254	42.5
	NTUSD 词典	1 243	8 274	15.0
word2vec 词典扩充		606		
积极词语部分（3 338）				
来源		词语数量	原词典总词数	占原词典比例
LM 词典中文翻译		458		

[7] 与绝大多数情感词典一致，本文词典将情感词划分为积极词与消极词两大类，未再根据情感程度继续细分。这是因为语料中重度情感词较少，我们没有足够的数据再对情感词的情感程度进行细分，并赋予不同权重，而武断的分类可能降低词典的适用程度。

（续表）

积极词语部分（3 338）				
来源		词语数量	原词典总词数	占原词典比例
通用词典筛选	清华词典	1 928	5 567	34.6
	知网词典	304	837	36.3
	NTUSD 词典	255	2 811	9.07
word2vec 词典扩充		393		

表 2 展示了中文金融情感词典三个来源中的代表性词语。LM 词典中文翻译中的词语多为金融领域的专有词汇，与金融领域关系极为密切，这部分词语在通用情感词典中是极为少见的。通用情感词典筛选得到的词语则多为日常语境中常见的情感词汇，这些词汇在金融语境下仍然出现概率较大，而且情感意义保持一致，因此也被纳入金融情感词典中。word2vec 词典扩充得到的词语与金融语境也有很强的相关性，但是词语的口语化与习语化特征比 LM 词典的翻译词语更为明显。可以看出，LM 词典中文翻译、通用情感词典筛选与 word2vec 词典扩充三部分词语特征差别较大，它们互为补充，共同构成了一个完善的中文金融情感词典。

表 2　中文金融情感词典展示

词语倾向	来源	词语	词语倾向	来源	词语
负向词语	LM 词典中文翻译	跌 被降级的 管理不善 旷工 漏税	正向词语	LM 词典中文翻译	涨 晋升 先发优势 独家经营 超额完成
	通用情感词典筛选	诽谤 担心 艰苦 薄弱 惩罚		通用情感词典筛选	一帆风顺 井然有序 可靠的 合法的 完美
	word2vec 词典扩充	败下阵来 变相涨价 操作失误 炒鱿鱼 大跌眼镜		word2vec 词典扩充	爱岗敬业 大好时机 高回报 绝对优势 可喜成绩

五、媒体文本情绪计算

（一）文本情绪计算方法

考虑到否定词可以改变情感倾向，本文将情感词及否定词组合为情感单元，并采用如下步骤计算文本情绪。第一步我们对新闻文本进行分词处理，并去除停用词语。然后，我们利用情感词典筛选出所有情感词语。第二步为构建情感单元。我们假设情感词只受到在其之前的词汇的影响，将从前一个情感词之后开始到该情感词为止作为一个情感单元。图4为情感单元的图形表示。

图4 情感单元示意图

第三步为赋予情感得分。我们将积极词权重设为1，消极词权重设为－1。考虑到双重否定表肯定的情况，否定词的权重为$(-1)^n$，其中n为否定词出现的次数。从而，文章情绪的计算公式如下所示：

$$SENTIMENT=\frac{\sum_{i=1}^{N}(-1)^n\times W_i}{N}, \tag{2}$$

$$W_i=\begin{cases}1, & positive\\ -1, & negative\end{cases}, \tag{3}$$

其中，W_i为情感词权重，N为一篇文章中的情感单元总数，$SENTIMENT$即为一篇文章的情绪值。基于单篇文章的情绪指数，可以通过求当日或当月内所有新闻文本情绪平均值的方法获得日度与月度文本情绪。

（二）文本情绪指数对比

本文利用所构建的中文金融情感词典与通用情感词典计算了两条月度文本情绪序列，并将其标准化为均值为0、标准差为1的序列，时间区间为1992年10月至2017年7月，并将其绘制如图5所示。

由图5，金融情感词典文本情绪更加准确。以2008年金融危机为例，市场情绪偏负面，但是通用词典文本情绪没有明显下降。在金融危机之后，我国政府推出四万亿计划，提振市场信心，但通用词典文本情绪却反而表现出了大幅下降趋势。上述现象都是通用情感词典对特定股票事件捕捉不够准确的表现，说明通用情感词典对金融文本情绪的计算不够准确。与之相反，金

融词典文本情绪准确地刻画出了各主要事件期间的情绪变化，明显优于通用情感词典的表现。

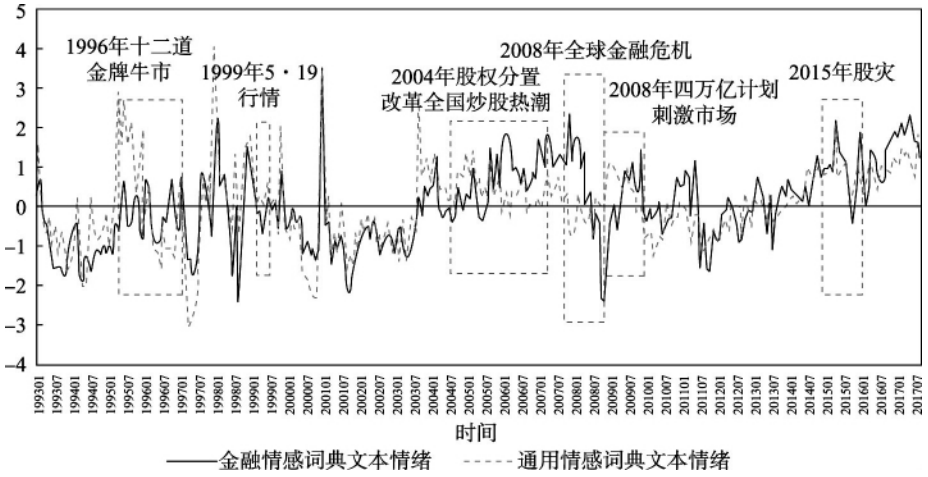

图 5　文本情绪走势与事件识别

六、实证分析

借助媒体文本情绪分析，我们实证研究投资者情绪对我国股票市场回报的预测能力。本文分别利用完整的金融情感词典（简称全金融词典）、金融情感词典中的 LM 词典翻译部分（简称 LM 词典）和通用情感词典（简称通用词典）计算了文本情绪，并检验三种文本情绪对股票市场的预测能力。这既可以挖掘文本情绪对股票市场的预测能力，也可以对比三种情感词典的表现。考虑到 2000 年之前中国股票市场成熟度较低，市场机制不健全，数据质量较差，为此在实证分析部分，我们将样本区间统一设定为 2000 年 1 月到 2016 年 12 月。

（一）文本情绪预测能力检验

1. 非参数检验

我们首先参照 Baker and Wurgler（2007）中的方法进行非参数检验，即根据投资者情绪值的高低将月份区分为情绪高涨期与情绪低落期，并统计两种情绪时期的下一月份的平均回报情况。为了避免情绪时期划分方式对检验结果的影响，我们分别根据当期情绪值是否大于 0 与是否大于全样本中位数划分情绪高涨期与情绪低落期。

如图 6 所示，在不同设置下情绪低落期的下月平均回报都要低于情绪高涨期的下月平均回报。这一结果与 De Long *et al*.（1990）提出的 DSSW 噪音交易者模型结论一致，市场上噪音交易者对股票未来价格的错误估计（过度

乐观或悲观的情绪）将推动风险资产价格高于或低于其基本价值。当金融新闻报道负面倾向明显时，投资者对未来的判断偏向悲观，即“情绪低落”，从而他们会采取与自己情绪相“匹配”的投资决策，例如将资金撤离股票市场转而投资无风险资产，而这种投资决策将导致下一月的整体股票市场回报偏低。同理，在情绪高涨期，投资者们对股票市场未来预期会更加乐观，他们所选择的投资行为将最终导致下一月股票市场整体回报走高。

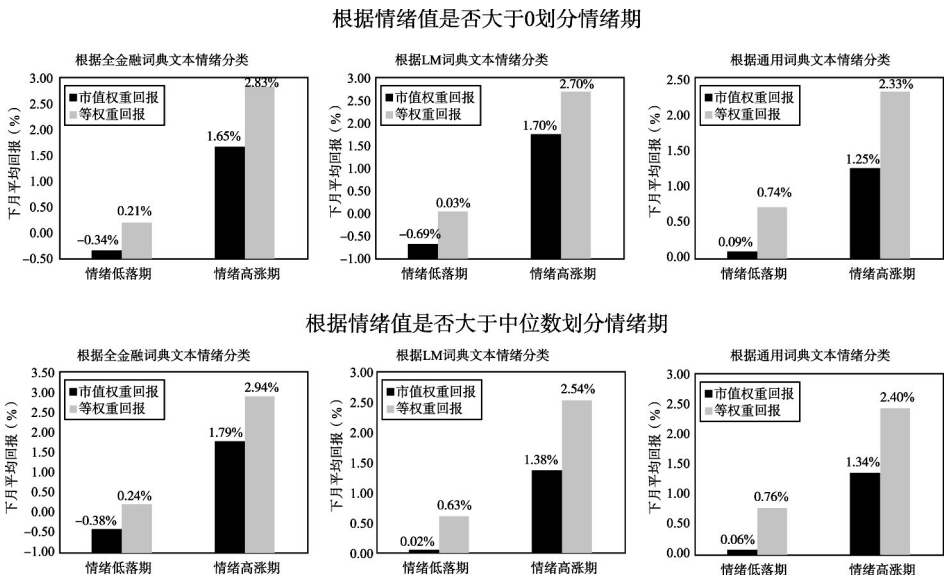

图 6 非参数检验结果

2．样本内预测能力的检验

在本小节，本文将验证文本情绪对股票回报的样本内预测能力。参照 Jiang *et al*.（2019），设置如下预测性回归方程：

$$R_t^M = \alpha + \beta SENTIMENT_{t-1} + \varepsilon_t, \tag{4}$$

其中，R_t^M 是 t 期 A 股等权与市值加权市场回报率，$SENTIMENT_{t-1}$ 是 $t-1$ 期的情绪变量，分别是全金融词典情绪、LM 词典情绪以及通用词典情绪，ε_t 是随机误差项。

如表 3 所示，各情感词典文本情绪与市场回报之间都表现出了显著的正相关性，说明前一期情绪越高涨，接下来的市场回报越高，与前文非参数检验的发现一致。从R^2统计量来看，三种文本情绪的R^2都大于 0.5%。根据 Campbell and Thompson（2008），在月度频度上，只要预测性回归方程的R^2统计量大于 0.5%时就具有显著的经济意义，而以本文构建的全金融词典计算的文本情绪R^2能达到 3%左右。使用完整的金融情感词典计算的文本情绪具有最大的R^2，说明这种词典最适合于金融语境。根据通用词典计算的文本情绪具有相对较小的R^2，小于两种金融词典文本情绪，说明其对股票市场的预

测能力是不完善的。总之，上述结果对比说明，只有采用金融情感词典计算文本情绪才能尽可能强地预测股票市场回报，而通用词典的表现则不尽如人意。

表 3　情绪与整体股票市场关系

文本情绪与等权重市场回报					
	α（%）	t 值	β（%）	t 值	R^2（%）
全金融词典文本情绪	1.30*	1.87	1.79**	2.48	2.98
LM 词典文本情绪	1.45**	2.09	1.44*	2.00	1.95
通用词典文本情绪	1.52**	2.17	1.29	1.50	1.11
文本情绪与市值加权市场回报					
	α（%）	t 值	β（%）	t 值	R^2（%）
全金融词典文本情绪	0.47	0.82	1.29**	2.17	2.30
LM 词典文本情绪	0.57	1.00	1.26**	2.12	2.18
通用词典文本情绪	0.63	1.09	1.10	1.55	1.19

注：***、**和*分别代表参数估计在 1%、5%和 10%水平下显著。

我们在附录中检验了日度文本情绪对股票市场回报的预测能力。在日度检验中，三种文本情绪均能够正向预测未来股票回报，且全金融词典文本情绪具有最好的表现。同时由于在日度频度下，新闻媒体报道的文本情绪与市场回报之间更容易出现相互影响，我们利用 VAR 模型检验了两者之间的关系。结果显示，在 VAR 模型下，文本情绪可以显著预测股票市场回报，说明媒体报道与文本情绪并不是对过去市场信息的简单反映，也包含与未来回报相关的信息。相应细节可以参阅附录。

3．样本外预测能力的检验

在本小节，本文将探讨文本情绪的样本外预测能力。参照 Welch and Goyal（2008），文本情绪对股票市场月度回报的样本外预测方程如式（5）：

$$\hat{R}_t^M = \hat{\alpha}_{t-1} + \hat{\beta}_{t-1} SENTIMENT_{t-1} + \varepsilon_t, \tag{5}$$

其中，$\hat{\alpha}_{t-1}$ 和 $\hat{\beta}_{t-1}$ 是把 $\{R_s^M\}_{s=2}^{t-1}$ 对 $\{SENTIMENT_{s-1}\}_{s=2}^{t-1}$ 回归得到的最小二乘（OLS）估计，此处市场回报选取 A 股市场等权重回报，其余变量与前文相同。设 P 为设定的初始训练区间大小，则我们需要对时刻 $t = P+1$，$P+2$，…，T 的市场月度回报进行样本外预测，从而共有 $q = T-P$ 个样本外预测值，也即 $\{\hat{R}_t^M\}_{t=P+1}^{T}$。初始训练区间内数据用于估计第一个预测性回归方程参数，并得到第一个样本外预测值。此后，随时间推移，我们不断增添新的训练样本，以滚动回归的方式得到新的回归方程参数，并以此计算对应的下一期样本外预测值。为了增强检验的稳健性，我们对区间做了两种划分。第

一种划分（区间一）下初始训练区间为2000M01至2006M12，样本外预测区间为2007M01至2016M12；第二种划分（区间二）下初始训练区间为2000M01至2003M12，样本外预测区间为2004M01至2016M12。

本文应用Campbell and Thompson（2008）中的R^2_{OS}统计量来评估不同文本情绪指标对市场回报的样本外预测能力。R^2_{OS}统计量的表达式如下：

$$R^2_{OS}=1-\frac{\sum_{P+1}^{T}(R_t^M-\hat{R}_t^M)^2}{\sum_{P+1}^{T}(R_t^M-\overline{R}_t^M)^2}, \tag{6}$$

其中，$\hat{R}_t^M$代表预测性回归方程得到的未来回报预测，公式如上文（5）所示。$\overline{R}_t^M$代表历史均值预测基准，即将股票回报的历史均值当作未来回报预测值，计算方法如（7）所示。

$$\overline{R}_t^M=\frac{1}{t-1}\sum_{s=1}^{t-1}R_s^M. \tag{7}$$

表4汇报了样本外预测结果。全金融词典文本情绪的样本外预测R^2_{OS}统计量均大于0，尤其，在第二种区间设置下可达到1.51%。这说明使用该词典计算的文本情绪对整体股票市场回报的样本预测能力要明显强于历史均值预测基准，具有实际应用的价值。其次，通过对比还可以发现，完整金融情感词典所计算的文本情绪在三种文本情绪具有最大的R^2_{OS}统计量，样本外预测表现最佳。这体现出了本文所构建的金融情感词典在金融语境下的适用性。

表4 样本外预测结果 单位：%

	区间一	区间二
全金融词典文本情绪	1.18	1.51
LM词典文本情绪	−2.18	−1.18
通用词典文本情绪	0.66	0.20

注：区间一代表初始训练区间为2000M01至2006M12，样本外预测区间为2007M01至2016M12；区间二代表初始训练区间为2000M01至2003M12，样本外预测区间为2004M01至2016M12。

4．资产配置检验

我们进一步从资产配置的角度探究文本情绪的应用价值。参照Campbell and Thompson（2008）等，我们计算了当均值方差投资者借助文本情绪来实现对股票和无风险资产间最优化配置时，他们可获得的确定性等价收益和夏普比率。根据均值方差投资理论，在$t-1$期末，投资者们在t期的投资组合中分配的股票资产的最优化权重为ω_{t-1}，计算方式如下：

$$\omega_{t-1}=\frac{1}{\gamma}\frac{\hat{R}_t^M}{\hat{\sigma}_t^2}, \tag{8}$$

其中，γ 为风险厌恶系数，我们取值为 3；$\hat{R}_t^M$ 是市场回报的样本外估计，定义与计算方法与上文相同；$\hat{\sigma}_t^2$ 为样本外方差估计，假设投资者使用 36 个月的滚动窗口来估计未来方差。该投资组合在 t 期获得的实现回报为：

$$R_t^P = \omega_{t-1} R_t^M + R_t^f, \tag{9}$$

其中，R_t^f 为无风险回报，取定期-整存整取-一年利率。限制 ω_{t-1} 取值范围在 0 到 1.5 之间，从而排除卖空并允许最大 1.5 倍杠杆。该投资组合的确定性等价收益如下所示：

$$CER_P = \hat{\mu}_P - 0.5\gamma\hat{\sigma}_P^2. \tag{10}$$

我们采用了与前文相同的初始训练区间和样本外预测区间设置。$\hat{\mu}_P$ 和 $\hat{\sigma}_P^2$ 为样本外预测区间上的样本均值与方差。我们假设投资者分别应用文本情绪与历史均值来生成样本外回报预测并构建投资组合，两者之间的差值即为应用文本情绪获得的确定性等价收益，我们将该差值乘以 12 以转化为年化回报。此外，我们还计算了投资组合的夏普比率。同时，我们还考虑了存在交易费用的情形，并将交易费用设为 50bp。

根据表 5，三种文本情绪均可获得正的确定性等价收益和夏普比率，且交易费用影响不大，证明了文本情绪指标的应用价值。从具体数值上来看，两种金融词典文本情绪得到的确定性等价收益基本在 1.5%以上。然而，通用词典文本情绪的确定性等价收益只在 0.2%左右，而夏普比率也呈现出同样的特征，说明通用词典文本情绪指标的实际应用价值相对有限。

表 5　资产配置检验结果

	无交易费用		交易费用（50bp）	
情绪指标	确定性等价收益	夏普比率	确定性等价收益	夏普比率
		区间一		
全金融词典文本情绪	1.76%	0.21	1.60%	0.20
LM 词典文本情绪	1.67%	0.21	1.50%	0.20
通用词典文本情绪	0.25%	0.18	0.20%	0.17
		区间二		
全金融词典文本情绪	1.48%	0.19	1.33%	0.18
LM 词典文本情绪	1.71%	0.21	1.55%	0.20
通用词典文本情绪	0.18%	0.16	0.14%	0.15

注：区间一代表初始训练区间为 2000M01 至 2006M12，样本外预测区间为 2007M01 至 2016M12；区间二代表初始训练区间为 2000M01 至 2003M12，样本外预测区间为 2004M01 至 2016M12。

5．与宏观经济指标预测能力对比

许多宏观经济变量对股票回报具有预测能力。例如，姜富伟等（2011）

发现几个经济指标在中国对股票市场回报有着较强的预测能力。本文选取了几个常见的宏观经济指标，包括工业增加值、国家统计局编制的宏观经济景气指数、制造业采购经理指数（PMI）、居民消费价格指数（CPI）以及流通中货币（M0）。我们对这些指标进行了平稳性检验以及平稳化处理，并标准化均值为 0、方差为 1 的序列。

我们将对比文本情绪与经济指标对股票市场回报的预测能力。首先，我们会对比文本情绪与经济指标的单变量预测能力，如式（11）所示。其中，Z_{t-1}^{k} 是经济指标之一。接下来我们检验文本情绪与经济指标的双变量预测能力，如式（12）所示。如果情绪指标可以带来预测能力的增值，那么双变量回归的预测表现应该改善，这可以说明情绪指标含有增量信息。股票市场回报选择等权重回报。

$$R_t^M = \alpha + \Psi Z_{t-1}^{k} + \varepsilon_t，\ k = 1，\cdots，5. \tag{11}$$

$$R_t^M = \alpha + \Psi Z_{t-1}^{k} + \varepsilon_t + \beta SENTIMENT_{t-1}，\ k = 1，\cdots，5. \tag{12}$$

根据表 6，只有一个经济指标的单变量预测回归结果在统计意义上显著。大部分经济指标的预测R^2都在 1%以下，有一些指标的R^2甚至在 0.5%以下，而文本情绪的R^2基本在 2%左右。因此，文本情绪指标可以作为宏观经济指标的有效补充，为股票市场回报的预测提供一个新角度。从双变量回归结果来看，大部分情况下文本情绪在统计意义上仍然显著，而且系数方向与单变量回归一致。这说明文本情绪对股票市场的预测能力是稳定的，不易受到其他因素的影响，这减轻了数据挖掘的担忧。同时，双变量回归方程的R^2出现显著增加，并以全金融词典文本情绪最为明显，它可以使得 R^2 出现 2%以上的改善。这说明文本情绪不是完全由宏观基本面驱动的，情绪信息可以和经济指标中的宏观信息形成有效互补。最后，对比三类文本词典情绪带来的预测能力改进可以发现，全金融词典文本情绪改进最强，通用词典文本情绪改进最差，这再次体现了金融情感词典的应用价值。

表 6 情绪指标与经济指标预测能力对比

	单变量回归		双变量回归			
				全金融词典	LM 词典	通用词典
居民消费价格指数	Ψ	−0.0169**	Ψ	−0.0204***	−0.0176**	−0.0173**
			β	0.0215***	0.0152**	0.0137
	R^2（%）	2.9017	R^2（%）	7.0850	5.0714	4.1527
制造业采购经理指数	Ψ	0.0020	Ψ	−0.0080	−0.0092	0.0018
			β	0.0204*	0.0146	0.0283*
	R^2（%）	0.0317	R^2（%）	2.3544	1.5147	2.4507

（续表）

	单变量回归		双变量回归			
				全金融词典	LM 词典	通用词典
流通中货币	Ψ	0.0090	Ψ	0.0091	0.0098	0.0090
			β	0.0183**	0.0148*	0.0136
	R^2（%）	0.8265	R^2（%）	3.9732	2.8849	2.0643
工业增加值	Ψ	0.0038	Ψ	0.0031	0.0030	0.0035
			β	0.0181**	0.0141**	0.0135
	R^2（%）	0.1474	R^2（%）	3.2191	2.0034	1.3671
宏观经济景气指数	Ψ	0.0027	Ψ	−0.0002	−0.0020	0.0035
			β	0.0183**	0.0149*	0.0140
	R^2（%）	0.0737	R^2（%）	3.1217	1.9444	1.3702

注：***、**和*分别代表参数估计在 1%、5%和 10%水平下显著。由于数据可获得性，带有采购经理人指数的回归方程样本区间为 2005M01—2016M12，其他方程均为 2000M01—2016M12。

（二）机制解释

1. 文本情绪与投资者预期

媒体新闻报道可以影响投资者预期的形成。根据 Sims（2003）等，投资者没有能力来关注所有事件信息，因此往往借助新闻媒体公布的信息来形成预期与做出对应决策。文本情绪作为媒体新闻报道中的重要信息，它对投资者预期的形成也具有潜在的重要影响，这可以进一步影响市场回报。当新闻报道乐观时，投资者会形成乐观预期，从而提高股票市场参与度，使得流入股市的资金增多，并推高当期股票回报。为了检验这一假设，本文将分别检验文本情绪对几大反映市场预期与信心的指数和基金净流量的预测能力。

本文选取的信心指数包括宏观经济景气指数、制造业采购经理人指数、非制造业采购经理人指数（商务活动）、消费者信心指数和企业家信心指数。其中企业家信心指数为季度指数，其余指数均为月度数据。这些指数可以反映市场对宏观经济的预期。回归方程如下：

$$Macro_t^k = \alpha + \beta SENTIMENT_{t-1} + \varphi Macro_{t-1}^k + \varepsilon_t,\ k = 1, \cdots, 5, \quad (13)$$

其中，$Macro_t^k$ 指的是五种信心指数，$SENTIMENT_{t-1}$ 指的是 $t-1$ 期的情绪变量，ε_t 是随机误差项。由于企业家信心指数为季度数据，我们通过求平均的方式将月度文本情绪季度化，并进行预测回归检验。其余指数的预测性均在月度频率上进行，与本文主体部分一致。

由表 7，文本情绪可以显著地正向预测下一期信心指数。以全金融词典文本情绪为例，各指数均至少在 10%的水平上可以被显著预测。这一结果可支

持本文假设，即媒体新闻报道中的文本情绪可以影响投资者对宏观市场预期的形成。

表7 情绪指标与宏观经济信心指数

	全金融词典文本情绪	LM词典文本情绪	通用词典文本情绪
宏观经济景气指数	0.1380***	0.1886***	0.0470
	(3.58)	(4.86)	(0.96)
制造业采购经理人指数	0.5562***	0.6260***	0.1399
	(3.75)	(4.63)	(0.65)
非制造业采购经理人指数	0.1942**	0.2535***	−0.0405
	(1.99)	(2.81)	(−0.25)
消费者信心指数	0.2550*	0.4102***	−0.0178
	(1.94)	(3.04)	(−0.11)
企业家信心指数	2.1498**	3.8153***	1.2363
	(2.17)	(3.88)	(0.92)

注：表中每个指数对应的第一行报告方程的估计系数，第二行为对应的 t 值。***、**和*分别代表参数估计在1%、5%和10%水平下显著。

接下来我们检验文本情绪对投资者的股票市场参与度的预测能力。我们将基金净流量作为投资者股票市场参与度的代理变量，参照 Sirri and Tufano (1998)、Cha and Lee (2001) 等，基金净流量定义如下：

$$FLOW_{i,t}=TNA_{i,t}-TNA_{i,t-1}\times(1+R_{i,t}), \tag{14}$$

其中，$TNA_{i,t}$ 是基金 i 在 t 期末的基金资产净值，$R_{i,t}$ 是基金 i 在 t 期的回报率，从而 $FLOW_{i,t}$ 表示了基金 i 在除去因投资产生的基金增长后的新资金流入。对单只基金的净流量加总可得市场基金总净流量。我们的基金样本包含纯股票型基金和偏股型混合基金。回归方程设置如下：

$$FLOW_t=\alpha+\beta SENTIMENT_{t-1}+\varphi FLOW_{t-1}+\varepsilon_t, \tag{15}$$

其中，$FLOW_t$ 是市场层面总基金净流量数据，$SENTIMENT_{t-1}$ 指的是 $t-1$ 期的情绪变量，ε_t 是随机误差项。由于基金净流量数据为季度，故将月度情绪数据转为季度，转化方式同上。

从表8可以看出，在控制了净流量滞后数据后，文本情绪可以显著地正向预测基金净流量，说明当媒体文本情绪高涨时，投资者们的确会显著增加自己对股票的投资与参与程度。当基金净流量增大，流入股市资金增多时，股票市场回报会在多种途径影响下显著提高（Cha and Lee，2001；Warther，1995）。

表 8　文本情绪与基金净流量

	α	β	t 值	φ	t 值
全金融词典文本情绪	202.53	522.76*	(1.96)	−0.36***	(−3.00)
LM 词典文本情绪	268.72	563.37**	(2.05)	−0.37***	(−3.09)
通用词典文本情绪	271.18	411.28	(1.24)	−0.31***	(−2.67)

注：***、**和*分别代表参数估计在 1%、5%和 10%水平下显著。

综合上述内容，我们可以得到文本情绪的确影响投资者对宏观市场预期的形成，而投资者们往往会根据预期进行投资。当媒体文本情绪乐观时，投资者们受到影响形成乐观预期，从而增大股市投资，使得流入股市的资金增多，从而推动股市回报增加。这是文本情绪对股票市场预测能力来源的一种解释。

2．文本情绪与市场风险

在本小节，我们将检验市场波动风险是否可以解释文本情绪的预测能力。Merton（1980）与 French *et al*.（1987）表明当市场波动较低时，市场风险也较低，导致下一期的风险溢价更低。因此，文本情绪预测能力的潜在来源之一是其可以预测市场波动预期的时间变动。

为了检验这一假设，参照 Huang *et al*.（2015），我们估计了如下预测性回归方程：

$$LVOL_t = \alpha + \beta SENTIMENT_{t-1} + \varphi LVOL_{t-1} + \varepsilon_t, \tag{16}$$

其中，$LVOL_t = \log(\sqrt{SVAR_t})$，是 t 期月度股票市场波动率，月度股票市场回报方差 $SVAR_t$ 定义如下：

$$SVAR_t = \sum_{i=1}^{N_t} \widetilde{R}_{i,t}^2, \tag{17}$$

其中，$\widetilde{R}_{i,t}$ 是 t 月第 i 个交易日的去均值后市场回报，去均值方法为当日回报减去 60 日市场回报移动平均。N_t 为 t 月交易日数量。

从表 9 可以看到，三种文本情绪对等权重和市值权重市场波动都不具有显著的预测能力，这与波动风险假设是不符的。基于上述结果，尽管我们不能完全排除基于风险的解释，但至少市场风险并不是文本情绪预测能力的主要来源。这表明本文的发现更加符合 DSSW 噪音交易者模型的结论，即文本情绪对市场情绪的影响主要通过非理性渠道传递，与 Huang *et al*.（2015）中的发现一致。

表 9　文本情绪与市场风险

	等权重市场波动			
	β	t 值	φ	t 值
全金融词典文本情绪	0.0018	(0.13)	0.8922***	(27.57)
LM 词典文本情绪	0.0014	(0.10)	0.8930***	(27.55)
通用词典文本情绪	0.0126	(0.79)	0.8893***	(27.41)

（续表）

	市值权重市场波动			
	β	t 值	φ	t 值
全金融词典文本情绪	−0.0000	（−0.00）	0.8917***	（27.10）
LM词典文本情绪	0.0026	（0.19）	0.8923***	（27.10）
通用词典文本情绪	0.0075	（0.47）	0.8901***	（27.00）

注：***、**和*分别代表参数估计在1%、5%和10%水平下显著。

七、结 论

本文构建了一个全新的中文金融情感词典，并探讨了媒体文本情绪与金融市场回报间的预测关系。为了更好地测度财经新闻文本情绪，本文首先运用“洋为中用”“古为今用”和word2vec三种方法构建了一个全新的中文金融情感词典，该词典共9 228个词语，其中消极词5 890个，积极词3 338个。利用我们的中文金融文本情感词典计算的文本情绪，可以更准确地捕捉投资者情绪变动对股票市场的影响，其准确度远远优于现有的通用情感词典，体现出了本文词典的价值，可以为后续金融文本分析研究提供有力工具。

实证分析部分，我们发现文本情绪对市场回报有着显著的预测能力。值得注意的是，使用通用情感词典计算的文本情绪的实证结果一般不显著，表现远差于我们的金融情感词典。因此，在金融经济应用中，本文所构建的金融情感词典的表现要远优于通用情感词典。

本文还对文本情绪预测能力的经济理论来源进行了探索，发现媒体文本情绪可以显著地影响投资者对宏观经济的预期，而投资者会根据预期调整金融市场参与程度，从而让市场回报产生相应的反应。同时还发现基于风险补偿的理论并不能很好地解释文本情绪的预测能力，说明媒体文本情绪对股票价格产生影响的方式更加符合DSSW噪音交易者模型的传播渠道。本文还有许多潜在的研究方向。举例来说，由于经济金融新闻频度极高，因此我们可以通过调节新闻频度来计算日度的文本情绪，这是其他投资者情绪衡量方法不能实现的。本文对日度文本情绪与股票市场关联进行了简单的探索，但未来还可以做更深的挖掘。

参 考 文 献

[1] Baker, M., and J. Wurgler, “Investor Sentiment and the Cross-Section of Stock Returns”, *Journal of Finance*, 2006, 61 (4), 1645-1680.

[2] Baker, M., and J. Wurgler, “Investor Sentiment in the Stock Market”, *Journal of Economic Perspectives*, 2007, 21 (2), 129-151.

[3] Campbell, J. Y., and S. B. Thompson, "Predicting Excess Stock Returns out of Sample: Can Anything Beat the Historical Average?", *Review of Financial Studies*, 2008, 21 (4), 1509-1531.

[4] Cha, H., and B. Lee, "The Market Demand Curve for Common Stocks: Evidence from Equity Mutual Fund Flows", *Journal of Financial and Quantitative Analysis*, 2001, 36 (2), 195-220.

[5] De Long, J. B., A. Shleifer, L. H. Summers, and R. J. Waldmann, "Noise Trader Risk in Financial-Markets", *Journal of Political Economy*, 1990, 98 (4), 703-738.

[6] French, K. R., G. W. Schwert, and R. F. Stambaugh, "Expected Stock Returns and Volatility", *Journal of Financial Economics*, 1987, 19 (1), 3-29.

[7] Huang, D. S., F. W. Jiang, J. Tu, and G. F. Zhou, "Investor Sentiment Aligned: A Powerful Predictor of Stock Returns", *Review of Financial Studies*, 2015, 28 (3), 791-837.

[8] Jiang, F. W., J. Lee, X. M. Martin, and G. F. Zhou, "Manager Sentiment and Stock Returns", *Journal of Financial Economics*, 2019, 132 (1), 126-149.

[9] 姜富伟、涂俊、D. E. Rapach、J. K. Strauss、周国富，"中国股票市场可预测性的实证研究"，《金融研究》，2011 年第 9 期，第 107—121 页。

[10] Li, F., "Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports?", *Working Paper*, 2006.

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

[12] Li, J., Y. Chen, Y. Shen, J. Y. Wang, and Z. Huang, "Measuring China's Stock Market Sentiment", *Working Paper*, 2019.

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

[14] Merton, R. C., "On Estimating the Expected Return on the Market: An Exploratory Investigation", *Journal of Financial Economics*, 1980, 8 (4), 323-361.

[15] Mikolov, T., K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space", *Working Paper*, 2013.

[16] 沈艳、陈赟、黄卓，"文本大数据分析在经济学和金融学中的应用：一个文献综述"，《经济学》(季刊)，2019 年第 18 卷第 4 期，第 1153—1186 页。

[17] Sims, C. A., "Implications of Rational Inattention", *Journal of Monetary Economics*, 2003, 50 (3), 665-690.

[18] Sirri, E. R., and P. Tufano, "Costly Search and Mutual Fund Flows", *Journal of Finance*, 1998, 53 (5), 1589-1622.

[19] 唐国豪、姜富伟、张定胜，"金融市场文本情绪研究进展"，《经济学动态》，2016 年第 11 期，第 137—147 页。

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

[21] Warther, V. A., "Aggregate Mutual Fund Flows and Security Returns", *Journal of Financial Economics*, 1995, 39 (2), 209-235.

[22] 汪昌云、武佳薇，"媒体语气、投资者情绪与 IPO 定价"，《金融研究》，2015 年第 9 期，第 174—189 页。

[23] Welch, I., and A. Goyal, "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction", *Review of Financial Studies*, 2008, 21 (4), 1455-1508.

[24] Yan, Y., X. Xiong, J. G. Meng, and G. Zou, "Uncertainty and IPO Initial Returns: Evidence

from the Tone Analysis of China's IPO Prospectuses", *Pacific-Basin Finance Journal*, 2019, 57 (10), 101-135.

[25] 杨晓兰、沈翰彬、祝宇，"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》，2016年第12期，第143—158页。

[26] 游家兴、吴静，"沉默的螺旋：媒体情绪与资产误定价"，《经济研究》，2012年第7期，第141—152页。

[27] Zhou, G. F., "Measuring Investor Sentiment", *Annual Review of Financial Economics*, 2018, 10 (1), 239-259.

Media Textual Sentiment and Chinese Stock Return Predictability

FUWEI JIANG
(*Central University of Finance and Economics*)
LINGCHAO MENG*
(*Peking University*)
GUOHAO TANG
(*Hunan University*)

Abstract We construct a novel Chinese financial sentiment word dictionary based on the Loughran and MacDonald (2011) dictionary, word2vec algorithm, and hand collection. The media textual sentiment calculated by this dictionary captures changes in investor sentiment well, and has significant predictive ability for Chinese stock market returns both in and out of sample, which is greater than the predictive power of commonly used macroeconomic indicators. In addition, textual sentiment is also a significant predictor for macroeconomic condition.

Keywords media textual sentiment, sentiment dictionary, return predictability

JEL Classification C53, G11, G12

* Corresponding Author: Lingchao Meng, School of Economics, Peking University, No. 5 Yiheyuan Road, Haidian District, Beijing, 100871, China; Tel: 86-18810712516; E-mail: lingchao_meng@pku.edu.cn.