第 18 卷第 4 期　　经 济 学（季 刊）　　Vol. 18，No. 4
2019 年 7 月　　China Economic Quarterly　　July，2019

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

沈 艳　陈 赟　黄 卓*

摘 要　本文对文本大数据分析在经济学和金融学中的应用进行综述。文本大数据具有来源多样化、数据量增长快和高频等特征，为经济学和金融学研究提供了新的分析视角。本文梳理了文本大数据的信息提取步骤，总结了词典法、机器学习方法和深度学习方法的实现原理和技术特点，并对文本大数据在经济学和金融学中的应用研究的数据来源、处理方法和实证结果进行了全面梳理。本文还讨论了基于文本大数据的实证分析的新特征和未来研究趋势。

关键词　文本大数据，机器学习，投资者情绪
DOI：10.13821/j.cnki.ceq.2019.03.01

一、引　　言

得益于互联网的快速发展和计算机技术的进步，文本大数据在经济学和金融学领域的应用方兴未艾。在经济学领域，文本大数据被用于刻画经济政策不确定性（Baker *et al.*，2016）、对行业进行动态分类（Hoberg and Phillips，2016）、度量和预测经济周期（Shapiro *et al.*，2018；Thorsrud，2019）和度量媒体报道偏差及新闻需求（Gentzkow and Shapiro，2010）等问题。在金融学领域，文本数据被用于刻画关注度（Fang and Peress，2009；Da *et al.*，2011；Hillert *et al.*，2014；Ben-Rephael *et al.*，2017）、情绪或语调（Antweiler and Frank，2004；Tetlock，2007；Li，2010；Loughran and McDonald，2011；Garcia，2013；Jegadeesh and Wu，2013；Kim and Kim，2014；Tsukioka *et al.*，2018）、可读性（Li，2008；Loughran and McDonald，2014）、新闻隐含波动率（Manela and Moreira，2017）和意见分歧（Antweiler and Frank，2004；Hillert *et al.*，2018）等方面。非结构化文本大数据的运用在拓宽经济和金融实证研究领域的同时，也带来了新挑战。

* 北京大学国家发展研究院，北京大学数字金融研究中心。通信作者及地址：陈赟，北京市海淀区颐和园路 5 号北京大学国家发展研究院，100871；电话：13264714992；E-mail：yunchen@pku.edu.cn。
本研究受国家社会科学基金重大项目（18ZDA091）资助。

作为新数据源，文本大数据至少有三个特征。一是数据来源多样化。相对于主要由政府和机构主导收集的传统数据，文本大数据的发布主体有个人（如投资者、消费者）、企业、媒体、机构和政府相关职能部门等；其具体形式丰富多样，如推特（Twitter），微博，论坛帖子，消费者对产品的评价，微信公众号，上市公司年报，电话录音文稿，招聘广告，公司年报、季报、公告，IPO 招股说明书，分析师研究报告，会议纪要，有影响力的政治、经济、金融领域人物的演讲，央行等政府机构定期和不定期发布的各类信息，等等。二是数据体量呈几何级增长。囿于数据收集成本，传统数据收集往往需要借助纸质媒介，体量较小。随着文本信息从纸质媒介向以互联网为媒介的方式转移，文本数据收集和传输成本大幅度降低，为计算机领域的自然语言处理方法（Natural Language Processing，NLP）提供了应用场景。三是时频高。传统数据需要经过系统性的组织和安排来收集，常用的经济和金融领域数据多为年度、季度、月度、周度数据，频率更高的数据可得性不足，不足以满足对经济和金融领域高频数据分析的应用需要。而文本大数据的频率可以高达秒级（如网民在网络平台上发布的消息和观点的时间颗粒度），这为高频研究提供了数据基础。

文本大数据为经典研究问题提供了新视角（Gentzkow *et al.*，2019），也可用于研究新的问题。例如，投资者情绪如何影响资产定价是经典问题，而投资者情绪的度量是实证研究的关键。传统度量方法包括选择市场变量作为投资者情绪代理变量的市场变量法和采用调查问卷收集到的答案来度量情绪的调查法。Baker and Wurgler（2006）对六个市场变量采用主成分分析法构建的情绪指数采用了市场变量法，而密歇根大学消费者信心指数则是调查法的主要代表。由于市场变量法获得的指数更可能是关于情绪的均衡结果（Qiu and Welch，2006），因此不只包含情绪（Sibley *et al.*，2016），而调查法频率低、成本高、受访者答案未必是其真实意图的表述，现有投资者情绪度量手段均有缺陷。通过收集反映投资者情绪的言论形成的文本数据（如论坛帖子、微博）提供了直接度量情绪的新渠道（Antweiler and Frank，2004；Tetlock，2007）。又如，经济不确定性会影响经济周期（Bachmann *et al.*，2013；Baker *et al.*，2016）和金融市场（Bali *et al.*，2017）。在度量经济不确定性的方法上，除了市场变量法外（Jurado *et al.*，2015），文本大数据方法最近广受关注。采用新闻文本数据构造的经济政策不确定性指数可以实现高时频，还可以实现统一标准下来度量各国、各地区的经济不确定性（Baker *et al.*，2016）。再如，Manela and Moreira（2017）利用 1890—2009 年间《华尔街日报》（*The Wall Street Journal*）的新闻构造的新闻隐含波动率指数，不仅为理解波动率提供了新渠道，还近似出尚不存在 VIX 指标的历史金融市场的风险状况。

将文本大数据应用于经济学和金融学研究的核心挑战在于如何准确、有

效率地从文本中提取出需要的信息，并考察其对相应问题的解释或预测能力。如图 1 所示，令 Ψ 代表采用的原始文本库，Y 代表要解释或者预测的经济或金融现象，要考察 Ψ 对 Y 的解释能力，需要经过三个步骤。第一，将文本库 Ψ 内所有文本转化数据矩阵 A；第二，通过计量或者统计方法 F，将 A 转换成目标信息序列 V，如关注度、情绪、不确定性等指数；第三，用提取出的 V 来解释或预测 Y。

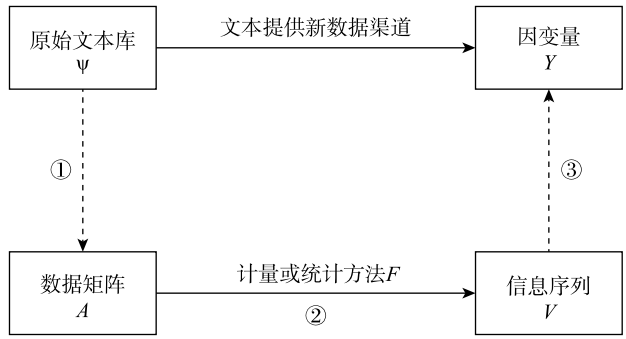

图 1　文本信息提取步骤

目前对实现上述三步转换需要解决的问题和相应的实际应用，现有中文文献缺乏较为系统的梳理。本文有两个主要目标，一是介绍从原始文本库 Ψ 到解释或预测 Y 的过程中，不同步骤面临的主要挑战、解决方法及其特点。由于第三步（利用结构化数据来完成解释或预测 Y 的工作）是计量经济学和统计学的研究重点，本文侧重介绍前面两步所涉及的方法。二是简要梳理国内外运用文本数据在经济和金融领域的应用，并探讨未来的研究方向。我们在第二部分介绍文本信息提取方法，即如何从原始文本库出发，提取出需要的信息序列；在第三部分梳理经济和金融各领域文本数据的应用；第四部分为结论和展望。

二、文本大数据信息提取方法

从中文原始文本中提取出研究者需要的信息，就需要让计算机学习用类似人的思维模式来分析和处理语言。本部分着重介绍将非结构化的帖子、文章、演讲词、会议记录等原始文本库 Ψ 作为输入，通过一定的转换方法得到结构化数据矩阵 A；再以数据矩阵 A 为输入，利用统计或者计量模型，输出目标信息序列 V 所涉及的方法。

（一）原始文本库到数据矩阵的结构化转换

从形式上看，一个中文文本是由汉字（包括标点符号等）组成的一个字符串。如果将文本从大到小分解，可能得到篇、章、节、段、句子、词组、

词和字。自然语言理解中的主要困难和障碍，是同一个字（词）的含义在不同的场景或语境下有变化；同时由于文字的丰富多样性，在转换为数据矩阵后往往需要处理高维稀疏矩阵相关的问题。在本小节着重介绍确定文本数据基础单位的分词技术和词嵌入（Word Embedding）技术，即将词转换为向量的方法。经过上述转换，非结构化文本可用矩阵形式表示，其中每一行记录同一个体的不同属性信息，而同一列数据记录不同个体的同一属性相关资料。

1. 分词技术

在英文环境下，单词被空格分隔开，因此单词就实现了分词。实证运用中也会将单个词语扩展成长度为 n 的词组，即 n 元词组（n-gram）。例如 Gentzkow and Shapiro（2010）发现，在分析不同党派的演讲内容时，词组比单词包含更多的信息，也更能反映政党的语言色彩。由于 n 的值越大，总词组的数量就越多，表示文本的矩阵或向量的维数将呈几何级增长，因此常用的 n-gram 模型中 n 取值一般为 1、2 和 3。

由于中文中汉字为连续序列，分析文本就需要按照一定的规范将汉字序列切分成词或词组，即中文分词。根据分割原理，可将现有的分词方法归纳为基于字符串匹配、基于理解和基于统计这三类。字符串匹配法将待分析的汉字串与前定的词典词条匹配，若某个字符串可在词典中找到，则记为识别出一个词。该方法的好处是简便快速，但忽略歧义问题。基于理解的分词方法则在分词的同时进行句法、语义分析，以改进对歧义词的处理。基于统计的分词方法则先用机器学习模型学习已经切分好的词语的规律，进而实现对未知文本的切分，常用方法包括最大概率分词法和最大熵分词法等。

目前经济和金融文献中用到的中文分词方法往往能结合上述三种方法，如自然语言处理与信息检索共享平台 NLPIR[1]（汪昌云和武佳薇，2015），中科院汉语词法分析系统（段江娇等，2017），Python 软件包“jieba”（王靖一和黄益平，2018；Li *et al.*，2019）等。需要注意的是，由于一些特定领域的文本包含一些对信息提取比较重要的专有词语（如上市公司名称、金融术语等），因此常常需要根据研究问题拓展现有词典，以提高软件识别和分割词语的准确度。

2. 词转换为向量的技术

完成分词之后需要完成的是如何将文本进一步转化为数字化矩阵。如果将一篇文本视作从所有词语库中挑选若干词形成的组合，这一转换的主要挑战往往是如何对由词语构成的高维矩阵实现降维的问题。要理解这一点首先需要介绍独热表示法（One-Hot Representation）。

（1）独热表示法

独热表示法最早的应用是在自然语言处理和信息检索领域。在金融领域，

[1] http：//www.nlpir.org/。

Manela and Moreira（2017）使用该方法，根据《华尔街日报》新闻的标题和摘要中全部词语出现的频率来提取新闻数据的特征。

独热表示法的特点是忽略语法和语序等要素，将文本数据看作是若干独立词汇的集合。首先，根据文本中出现的全部词语构建一个词表[2]，并将每个词按顺序编号 1，2，3…，N。然后，将词语 j 用一个 N 维向量w_j来表示，该向量的第 j 个位置的元素为 1，其余均为 0。在每一个词都转换为一个向量后，通过加总所有词的向量，文本 t 就可以转化为 $1\times N$ 的向量W_t，其中w_{tj}（$j=1$，…，N）是第 j 个词语在文本 t 中出现的频率。若一共有 $t=1$，…，T 个文本，采用独热表示法之后，原始文本库 Ψ 就可以转化为 $T\times N$ 的数字矩阵。

例如，原始文本库 Ψ 由两条帖子组成。第一条的内容是“明天涨停。后天涨停没戏。”第二条是“玛丽有个小绵羊”。分词后得“明天、涨停、后天、没戏、玛丽、有、个、小、绵羊”九个不同词语，即 $N=9$。用独热表示法则“明天”用向量［1，0，0，0，0，0，0，0，0］表示，“涨停”为［0，1，0，0，0，0，0，0，0］，以此类推。于是第一个帖子可用向量［1，2，1，1，0，0，0，0，0］表示，第二个帖子即［0，0，0，0，1，1，1，1，1］。

上述步骤显示，独热表示法操作简单；但数据量大时转换后的矩阵往往是高维稀疏数据矩阵。这是由于词向量维数由词语数量决定，并且大部分词语出现频率低，因此文本对应的向量中绝大部分元素的值为零。另外，独热表示法可能会因忽略上下文结构而产生歧义。例如上例中第一个帖子转换成的向量也可以是“明天涨停没戏。后天涨停。”的转换结果，这和原文的含义产生了偏差。

要解决文本数据是高维稀疏矩阵的问题有两种策略，一是采取多种措施对数字化文本矩阵实现降维，Gentzkow *et al*.（2019）对相应的降维方法已经做了系统总结。另一个思路则是采用词语嵌入技术，直接在词语转换成数字化矩阵时就将词语转化为低维向量。

（2）词嵌入技术

词嵌入技术是指把一个维数为所有词的数量的高维空间“嵌入”一个维数低得多的连续向量空间中涉及的模型和技术，即$e_j=E\times W_j$，其中e_j表示第 j 个词通过嵌入矩阵 E（embedding matrix）映射到实数域上的词向量[3]，W_j为第 j 个词的独热向量表示。由于该向量的每个元素值可以是连续值而不只是 0 或者 1，e_j的维度N_e可以远低于 N。

独热表示法可以看作是最简单的词嵌入方法，即$e_j=E\times W_j=W_j$。常用

[2] 通常是去掉“的、地、得、和”等停用词和标点符号后得到的全部词语。

[3] 吴恩达，“序列模型”课程，https：//mooc. study. 163. com/university/deeplearning _ ai#/c，访问时间：2019 年 7 月 15 日。

词嵌入算法包括 Mikolov *et al*.（2013）提出的 Word2Vec 技术和 Pennington *et al*.（2014）开发的 GloVe（Global Vectors for Word Representation）技术，其中 Word2Vec 的应用更为广泛。Word2Vec 的主要思想是先用向量代表各个词，然后通过神经网络模型，在大量的文本语料数据上来学习这些向量的参数。训练后的模型不仅可以将每个词语映射到一个低维的空间上（通常为 100—1 000 维），每个维数上的取值为连续值；并且根据不同词语的向量距离可以度量词语间的相似程度，也解决了独热表示法下不同词语相互独立的问题。

Word2Vec 技术在计算语言学等领域得到了广泛的应用，并且在和其他的统计模型结合进行文本分析时具有很好的表现，但在经济金融领域的应用相对较少（Gentzkow *et al*.，2019）。近年来，该方法也逐渐得到重视。例如，王靖一和黄益平（2018）利用该技术拓展了金融科技情绪词典。Li *et al*.（2019）对比了独热表示法和 Word2Vec 两种方法，发现相比于独热表示法，使用 Word2vec 来表示文本特征能够显著提高文本情绪的分类准确性。

（二）数据矩阵的信息提取

根据事先是否存在有标签的训练数据，经济和金融领域文本相关的问题可以采用有监督学习或无监督学习这两类方法来分析。其中，无监督学习的主要方法包括词典法和主题分类模型等，而支持向量机等机器学习经典方法和深度学习方法近年在经济和金融领域的运用更多属于有监督学习。

1. 无监督学习方法

（1）词典法

词典法是一种传统的文本大数据分析方法。该方法从预先设定的词典出发，通过统计文本数据中不同类别词语出现的次数，结合不同的加权方法来提取文本信息。在经济金融领域中，词典法得到广泛运用（如 Tetlock，2007；Tetlock *et al*.，2008；Loughran and McDonald，2011；Garcia，2013；Da *et al*.，2014；Zhang *et al*.，2016；Renault，2017；Li *et al*.，2019；等等）。

使用词典法的一个关键环节是选择或构建合适的词典，这里词典包括了特定词典，也包括作者构造的特定词语或词组的集合。文献中常用的英文特定词典包括 Harvard IV-4 词典[4]、Henry 词典、Diction 词典[5]和 Loughran and McDonald 词典[6]。早期文本情绪构造多使用 Harvard IV-4 词典（Tetlock，2007；Tetlock *et al*.，2008；Jegadeesh and Wu，2013），它包含心理和社会学常涉及的 1 045 个正面词语和 1 160 个负面词语，但并非为金融领域文本专

[4] http：//www. wjh. harvard. edu/～inquirer/homecat. htm，访问时间：2019 年 7 月 10 日。
[5] https：//www. dictionsoftware. com/，访问时间：2019 年 7 月 10 日。
[6] https：//sraf. nd. edu/textual-analysis/resources/，访问时间：2019 年 7 月 10 日。

门创建。Henry 词典是专门为金融文本构建的词典（Henry，2008），Price *et al*.（2012）发现 Henry 词典比 Harvard IV-4 词典更准确地度量了上市公司盈利披露电话会议文字稿中的语调，但 Henry 词典包含的负面词汇较少。Diction 词典包含了 686 个正面词汇和 920 个负面词汇，主要应用于会计领域（Rogers *et al*.，2011；Davis *et al*.，2012）。Loughran and McDonald（简称 LM）词典由 Loughran and McDonald（2011）从上市公司的 10-K 文件中人工收集并整理构造出来，他们的实证结果表明 LM 词典在度量文本情绪方面比 Harvard IV-4 词典和 Diction 词典的效果更好，因此目前用词典法分析金融、会计领域文本情绪时多采用 LM 词典（Garcia，2013；Huang *et al*.，2014；Loughran and McDonald，2014；Solomon *et al*.，2014；Zhang *et al*.，2016；Jiang *et al*.，2019；等等）。

在特定词典外，使用作者自行构造的词或者词组的代表性研究包括 Da *et al*.（2014），Baker *et al*.（2016），Hoberg and Phillips（2016）和 Box（2018）。Da *et al*.（2014）选取了 118 个与经济相关的词语（词组），利用这些词语在谷歌搜索中的搜索频次，构建了用来度量投资者情绪的 FEARS（Financial and Economic Attitude Revealed by Search）指数。Hoberg and Phillips（2016）基于 1996—2008 年间的 10-K 文件，以不同的产品描述词语集作为基础，将不同上市公司作了基于文本的行业分类。Baker *et al*.（2016）则选取和经济、政策、不确定三个类别相关的一些词语，通过统计同时包含这些词语的新闻的比例，构建了经济政策不确定性（Economic Policy Uncertainty，EPU）指数。

在中文语境下使用词典法，需要注意的是直接翻译的英文词典可能并不适用。Li *et al*.（2019）随机抽取 2008—2018 年间某股票论坛四万条帖子，人工挑取其中正、负面词语，构建了适用于中国股吧论坛的金融情绪词典。他们发现，与直接使用翻译的 LM 词典相比，该词典能将情绪分类准确率提高 30%。在中文相关文献有作者就具体问题构建的中文词典。例如，汪昌云和武佳薇（2015）手动整理新闻报道，结合《现代汉语词典》《最新汉英经济金融常用术语》、LM 词典中文版以及知网—中文信息结构库等词库，构建了中国财经媒体领域的正负面词库。王靖一和黄益平（2018）根据和讯网上的新闻，构建了适用于金融科技领域的情感词词典等。

在确定词典后，另一个要处理的问题是如何确定词语权重。Jegadeesh and Wu（2013）指出，选择合适的加权方法有时比构建完备且精确的词典更重要。常用的加权方法有等权重、词频-逆文档（Term Frequency-Inverse Document Frequency，TF-IDF）加权和对应变量加权这三种。顾名思义，等权重法假定文本中每个词语的重要程度相同。TF-IDF 加权方法则同时考虑词语在文本中出现的次数（频率）和多少文档包含该词语这两个维度，对在文本中频繁出现但并没有实际含义的词语赋予较少的权重，而给予有重要含义

但出现次数较少的词语较大权重。对应变量加权是指借用文本中词语与对应变量（市场收益率、波动率指数等）的关系来确定词语的权重。

不同权重法各有千秋。等权重法因简便易行而广为使用（如 Hoberg and Phillips，2016），不过 Loughran and McDonald（2011）发现在 10-K 文本下，TF-IDF 加权法比等权重法更能降低词语分类错误，可以实现更为有效的信息提取。Li *et al*.（2019）的研究结果也表明，在中文语境下使用 TF-IDF 能够比等权重法得到更准确的情绪分类。对应变量加权法的优点是能在一定程度上避免权重选择的主观性，其效果也不依赖于词典是否完整，因此文本分析结果可能比人为主观设定权重更为准确。例如，Jegadeesh and Wu（2013）根据市场收益率与 10-K 文件中词语的关系，将词语与市场变量对应起来，从而确定词语的权重。利用这种方法还能自动判断出 10-K 文件中的某个词语属于正面还是负面。Renault（2017）根据网络论坛 StockTwits 上带有标签（看涨和看跌）的帖子，统计词语在这两类文档出现的频率，为不同词语正负情感的强弱程度作加权。

总体而言，只要运用得当，词典法从文本中提取信息的能力较强，这种优势对于短文本和词语间逻辑关系较弱的应用更为明显。例如，Renault（2017）发现词典法和机器学习方法在识别论坛情绪准确率方面不相上下。Li *et al*.（2019）对中文论坛帖子数据的分析也有类似的结论。因此实际应用中，词典法常常可以作为文本大数据分析的一种基准方法。

（2）主题分类模型

在经济和金融领域的一个应用需求是在没有事先标注集的情况下，对文本按主题做分类。由于一篇文本的主题可能有多个，这类分类问题不同于按照事先标注集、将一篇文本仅归入一类的应用。主题分类问题的代表模型是由 Blei *et al*.（2003）提出的隐含狄利克雷分配（Latent Dirichlet Allocation，LDA）模型，它是一种概率主题模型。

LDA 模型假定全部文档 M 中存在 K 个主题，每个文档 m 包含 N_m 个词语，并且每个词都是由其中的一个主题生成。主题服从一个多项式分布 θ_m，而每个主题 k 与词汇表中的 V 个单词的一个多项式分布 φ_k 相对应，并且假定分布 θ_m 和分布 φ_k 具有共轭的狄利克雷分布，该共轭的狄利克雷分布的超参数为 α 和 β。通过预设文档中的主题个数，LDA 模型可以将每篇文档的主题以概率分布的形式给出，其中每个主题对应一类词语分布，根据词语分布可以挑选出一些关键词对该主题进行描述。

在 LDA 模型下，文档的生成过程如图 2 所示：①从狄利克雷分布 α 中抽样得到文档 m 的主题多项式分布 θ_m，从狄利克雷分布 β 抽样得到主题 k 的词语多项式分布 φ_k，$k=1, \cdots, K$；②从主题多项式分布 θ_m 中抽样得到文档 m 的第 n 个词的主题 $Z_{m,n}$；③从主题 $Z_{m,n}$ 对应的词语分布 $\varphi_{Z_{m,n}}$ 抽取词语 $W_{m,n}$；④重复上述步骤 N_m 次。因此，所有已知的和隐藏的变量的联合分布可以表示为：

$$P\ (W_m, Z_m, \theta_m, \Phi; \alpha, \beta) = \prod_{n=1}^{N_m} P(\theta_m; \alpha) P(Z_{m,n} | \theta_m) P(\Phi; \beta) P(W_{m,n} | \varphi_{Z_{m,n}}),$$

其中 $\Phi = \{\varphi_k\}_{k=1}^{K}$，模型中唯一可观测的变量是词语 $W_{m,n}$。实际应用中，可以通过 Gibss 抽样方法来估计 LDA 模型的参数，从而得到每篇文档的主题分布 θ_m 和每个主题对应的词语分布 φ_k。

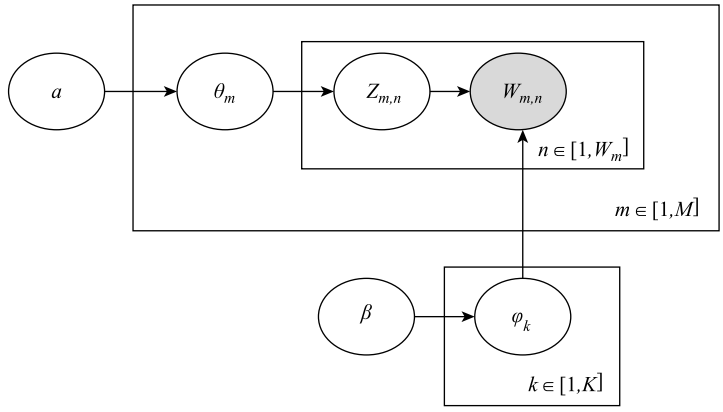

图 2　LDA 模型图示法

注：图片来自 Thorsrud (2019)。

LDA 的一个局限性是需要人为地给出一个主题数量，而主题数量的选择会影响主题的生成和文档的归类。选择文档主题个数 K 的方法通常包括：根据主题个数计算得到复杂度得分（perplexity score）（Blei *et al.*，2003），交叉验证法（Airoldi *et al.*，2010），预设一些初始值、再根据主题的解释能力来调整主题个数（Gentzkow *et al.*，2019）等。LDA 模型的一个拓展是 Teh *et al.*（2006）提出的层次狄利克雷过程（Hierarchical Dirichlet Processes）。该方法不需要事先设定 K，而是将主题个数作为未知的模型参数并结合贝叶斯非参技术来估计。LDA 的另一个局限性是忽略了主题分布随时间可能存在的演进变化，相对应的拓展是 Blei and Laffery（2006）的动态主题模型（Dynamic Topic Models）。对这些拓展的细节本文不复赘述。

近年来主题分类模型在经济和金融领域逐渐得到运用。例如，Thorsrud (2019) 使用 LDA 模型从新闻数据中提取出了 80 个主题并估计日度频率的新闻即时经济周期指数；Wang *et al.*（2018）使用 LDA 和 HDP 从近两千万新闻中分离出金融科技主题，并且构建了金融科技情绪指数等。

2. 有监督学习方法

(1) 经典的有监督机器学习方法

经典机器学习方法包括朴素贝叶斯、支持向量机、决策树、K 近邻算法、AdaBoost、最大熵法等。在金融领域的文本分析中，较为常用的传统机器学习方法包括朴素贝叶斯（Naïve Bayes）和支持向量机（Support Vector Ma-

chine，SVM）。

朴素贝叶斯算法（Murphy，2012）是一种基于贝叶斯理论的有监督学习算法。在处理文本分类问题时常见步骤如下。首先根据训练集学习文本中词语与所属类别的关系，得到朴素贝叶斯分类器的先验分布（即文本属于不同类别的先验概率），以及条件概率分布（即给定分类类别下某个词语出现的概率）。其次，使用这些概率，根据文本中的词语特征，结合贝叶斯条件概率公式，计算该文档属于不同类别的条件概率。最后，按照最大后验假设将文本分类为具有最大后验概率的一类。

Antweiler and Frank（2004）较早采用朴素贝叶斯方法研究文本情绪。他们挑选了 1 000 条雅虎财经上的帖子，人工将其分类为买入、卖出、持有。接着利用这些人工标注帖训练朴素贝叶斯算法，并将其应用到剩余的未分类帖子上，最后根据买入、卖出帖子的数量构建了 Bullishness 指数，用于度量文本情绪。此后，在经济金融领域不少文献将朴素贝叶斯算法应用到不同类型的文本上，如雅虎财经上的股票讨论帖（Das and Chen，2007；Kim and Kim，2014）、公司年报文件（Li，2010；Jegadeesh and Wu，2013）、分析师报告（Huang *et al*.，2014）和网络论坛发帖（段江娇等，2017）。

支持向量机（Vapnik，1996）是一种有监督学习算法，既可以用于分类也可以用于回归分析。其基本原理是，首先将每个文本投射为高维空间的一个点，通过寻找到一个超平面，将这些点按照其对应的标签（如正、负情绪等）进行分割，使得每个类别的点到这个超平面的最近距离最大化。使用支持向量机进行分类和回归分析前的步骤包括，首先采用独热表示法或者 Word2vec 等方法将文本转换为向量，然后根据训练集学习文本向量与所属类别的关系，再对将根据训练集得到的模型做交叉验证（cross-validation），最后将训练出的最优模型用于预测所有文本的分类。

SVM 相关应用主要出现在金融领域。如 Manela and Moreira（2017）使用独热表示法，将 1890—2009 年间《华尔街日报》头版新闻向量化；再使用支持向量回归法提取新闻隐含波动率指数。Tsukioka *et al*.（2018）使用 SVM 方法度量日文环境下的股票论坛投资者情绪。Li *et al*.（2019）使用 SVM 对中国网络论坛帖子进行情感分类。

此外，K 近邻算法（杨晓兰等，2016），最大熵法（Renault，2017）等也有运用，其表现和朴素贝叶斯、支持向量机方法相近。

（2）深度学习法

文本分析中，SVM 等分类器虽然可以处理一定的非线性，但作为线性分类器，这类方法往往只能将输入数据切分为非常简单的区域，也容易导致过拟合等问题（Gentzkow *et al*. 2019）。随着大数据可得性的增加、人工智能软硬件技术的发展，深度学习方法在自然语言处理领域的强大功能逐渐显现。作为机器学习的分支，深度学习试图通过模仿人脑的神经网络，使用多重非

线性变换构成的多个处理层对数据进行高层抽象，以实现分类等目标。这类方法可用于有监督和无监督学习，但目前尚未在经济领域有广泛运用，在金融领域的运用主要是有监督学习（Li *et al*.，2019）。

神经网络（Neural Network）（Bishop，1995）是基于模仿人脑的神经网络来实现人工智能的机器学习模型，包含输入层、隐藏层、输出层等结构，可用于处理文本分类问题，其原理是输入层的特征向量通过隐含层的变换到达输出层，在输出层得到分类结果，通常使用反向传播算法对神经网络模型进行训练。

深度学习常用模型包括深度神经网络（Deep Neural Network，DNN）、卷积神经网络（Convolutional Neural Network，CNN）和循环神经网络（Recurrent Neural Network，RNN）等。作为神经网络模型的拓展，DNN（Hinton and Salakhutdinov，2006；LeCun *et al*.，2015）可以通过增加网络层数、减少每层网络节点数，以及使用不同的传输函数克服训练过程中的梯度消失现象等方法，处理文本分类、翻译、语义分析等复杂的自然语言处理任务。针对 DNN 参数数量巨大，没有考虑数据的固有局部特征等缺陷，Kim（2014）提出了 CNN 方法。CNN 模型进行文本分类时，不仅限制了参数个数，还通过考虑词语在文本中的上下结构来挖掘文本内的局部结构。Kim（2014）发现，将 CNN 和 Word2vec 嵌套在一起使用可以在对文本进行分类时达到非常高的准确率。第三种深度学习方法 RNN（Elman，1990；Mikolov *et al*.，2010）处理文本分类问题的思想是借用 RNN 模型的递归结构来捕捉上下文信息。深度学习的常用模型还在不断拓展中，如 CNN 和 RNN 模型组合在一起的 RCNN 模型（Lai *et al*.，2015）等。这些深度神经网络模型的好处是可以提供非线性分类，但代价是模型待训练参数很多，通常一层结构的 CNN 模型就需要训练数万的参数，因此训练样本足够大是模型效果的基本保障。

目前在经济金融领域使用深度神经网络模型提取文本信息的文献较少。Li *et al*.（2019）首次采用 CNN 来计算中国散户投资者情绪，并比较了 CNN 与 SVM 等模型的预测效果。他们的研究发现，在采用四万条训练数据集的情况下，训练出的 CNN 模型的预测准确性与 SVM 大致相当，但是在分类中 CNN 模型的分类更为果断；随着训练数据集的增大，CNN 的优势可能会进一步显现。

无论是采用经典机器学习方法还是新兴的深度学习法，有监督训练都需要两个要素：高质量的标注数据作为训练集和明确的模型选择标准。由于训练集质量会直接影响最终信息提取效果，做相关研究应事先评估构建标注数据需要耗费的成本。在模型选择标准方面，理想模型不仅要能避免样本内过拟合，也要有较好的样本外表现。通常需要采用交叉验证的方法来评估模型：首先将标注集按照一定的比例随机分为训练集、验证集和测试集；再在训练集上训练模型，根据其在验证集上的表现来调整模型参数；最后将模型应用

到测试集上计算准确率，作为评估模型样本外表现的标准。

综上所述，选择文本数据信息提取方法需综合考虑文本数据的来源、语言环境、内容长短以及需提取信息的特征等因素，同时评估各类方法的成本和收益。在条件允许的情况下，可同时考虑简单方法和复杂方法，通过分析比较两类方法的差异来提高信息提取的准确性。当然，使用复杂方法时需要保证这些方法的透明性和可复制性（Loughran and McDonald，2016）。最后还要注意的是，数据的结构化转换和文本数据信息提取这两步的执行顺序需要依靠具体问题来决定，有时需要反复尝试才能找到最佳方案。

三、文本大数据分析在经济学和金融学中的应用

目前，文本大数据分析在经济学和金融学中的应用日渐广泛。经济学应用主要体现在刻画经济政策不确定性、对行业进行动态分类、预测经济周期、度量媒体报道偏差及新闻需求和量化央行政策沟通内容等问题中。在金融学中的应用主要包括在刻画关注度、情绪、可读性、隐含波动率和意见不一致性等五个方面的指标。

（一）在经济学中的应用

1. 经济政策不确定性指数

经济政策不确定性是经济中的个体对未来政策的变动和当前政府政策影响的不确定性程度（Gentzkow *et al.*，2019）。传统方法从市场变量出发来度量经济不确定性[7]，这些度量方法存在市场变量时间跨度短、频率低，不同国家指标不具可比性等弱点。而各国主流新闻媒体的新闻文本历史跨度长、频率高，因此 Baker *et al.*（2016）另辟蹊径，采用新闻文本数据来度量经济政策不确定性的 EPU 指数。

从构建方法看，EPU 指数属于作者自行选择词或词组的词典法。首先作者收集美国 10 家主流新闻媒体从 1985 年以来的新闻数据，使用机器自动统计各媒体新闻中同时包含经济（economic/economics）、不确定（uncertain/uncertainty）和政策[8]三类词语的月度文章数量。为控制新闻数量的时间趋势，作者对上述文章数量作标准化处理，再对这十个标准化的序列按月作平均，最后将序列转换为均值 100 的指数。

---

[7] 如：市场波动率指数 VIX 或 VXO（反映对权益资产未来收益的不确定性）、衡量投资者风险厌恶程度的方差风险溢价（Bali and Zhou，2016）、横截面分散程度（上市公司股票收益率、企业利润增长率、分析师 GDP 预测、不同行业的全要素生产率等变量的标准差）、基于大量经济变量中不可预测部分条件波动率的不确定性指数（Jurado *et al.*，2015）等。

[8] 政策类词语为 Congress，deficit，Federal Reserve，legislation，regulation，White House，从 1985—2012 年新闻中出现频率最高的 15 个词语中选出。

在从三个角度[9]验证了 EPU 指数能较好度量经济政策不确定性后，Baker *et al*.（2016）对 EPU 指数还做了一系列拓展：①构建 11 个主要经济体月度频率的 EPU 指数；②构建货币政策、财政政策、国防等 11 个政策分类的 EPU 子指数；③构造英国和美国 1900—2011 年的日度 EPU 指数。

研究 EPU 与其他经济变量间的关系是近期文献的热点之一[10]。如 EPU 对企业层面的股价波动率、投资率、就业增长率的影响，对市场加总层面的投资、产出、就业的影响（Baker *et al*.，2016），对公司投资的影响（Gulen and Ion，2016），对股市波动率的影响（Pástor and Veronesi，2013）；对市场超额收益率的预测（Brogaard and Detzel，2015）等。国内也有不少文献使用 Baker *et al*.（2016）提供的中国市场 EPU 指数来研究经济政策不确定性对微观企业经营活动和决策行为的影响，如经济政策不确定性如何影响公司现金持有水平（王红建等，2014）、企业投资行为（李凤羽和杨墨竹，2015；饶品贵等，2017）、企业创新（顾夏铭等，2018）、国有企业和非国有企业的杠杆率（纪洋等，2018）、分析师盈余预测修正（陈胜蓝和李占婷，2017）、企业金融化趋势（彭俞超等，2018）、企业资本结构动态调整（顾研和周强龙，2018）和公司提供的商业信用（陈胜蓝和刘晓玲，2018）等。

当然，基于英文《南华早报》（*South China Morning Post*）度量的中国市场 EPU 指数也存在一定的局限性（饶品贵等，2017）。目前度量中国经济不确定性的指标除了 EPU 外，还有 Huang *et al*.（2018）和黄卓等（2018）参照 Jurado *et al*.（2015）的方法分别构建的中国经济不确定性指数和金融不确定指数，但从中文文本数据出发度量中国经济不确定性的研究尚属空白。

除了市场层面的不确定性指数，Handley and Li（2018）还构建了公司层面的不确定性指数（Company Reported Uncertainty Index，CRUX）。他们使用 1994—2016 年上市公司的年报和季报等文件（约 100 万份），根据文件中“不确定”词语[11]的占比来度量公司层面的不确定性，并在公司和市场层面研究了不确定性程度对经济活动的影响。他们发现该不确定性指数显著负向影响投资增长率、GDP 增长率以及就业增长率等经济变量。

2. 行业分类

产业组织中的一个核心问题是定义行业边界和行业竞争力，传统的行业分类通常较为固定（如：标准行业分类 SIC 和北美产业分类体系 NAICS 等）。Hoberg and Phillips（2016）根据上市公司年报中对企业产品描述的内容提出

[9] 这三个角度是①EPU 指数和重大经济政治事件相吻合；②采用人工阅读 12 000 多份报纸构建的指数和使用机器构建的两种 EPU 指数相关系数高达 0.86；③EPU 指数与其他常用的不确定性度量指标有很高的相关系数，特征相似。

[10] EPU 除了对经济变量有重要影响，对金融市场也有影响，如：金融市场的溢出效应和联动性等。黄卓等（2017）对研究 EPU 与金融市场关系的相关文献进行了总结。

[11] 包括：uncertain、uncertainty、uncertainties 和 uncertainly。

了一种新的行业分类方法，即文本网络行业分类法（Text-based Network Industry Classification）。首先统计全部文档中包含用来描述产品的不同词语的个数 W，并构建对应的词表；接着采用独热表示法将每个企业的产品描述文档转换为长度为 W 的向量，即如果文档中包含某词语，则对应位置的元素取值为 1，否则为 0。然后，计算不同文档向量的余弦相似性作为不同公司产品相似度的度量。最后根据这些相似度得分，结合聚类算法，可以将不同公司分组到不同的行业，最终得到 300 个行业分类（与 SIC 和 NAICS 中的行业数量一致）。基于这种时变的行业分类标准，他们检验了公司如何对产品市场的外部和内部变化做出反应，并评估在军事和软件行业等大的外生行业冲击下，企业对产品生产调整的反应程度。他们发现外生行业冲击会对相似企业数量、产品差异性、企业生产的产品种类等产生重大影响。除了用于研究外生冲击对行业内竞争和企业产品生产决策的影响外，他们还认为上述分类可用于解释行业内盈利能力、销售增长率、市场风险等不同特征的差异性。

Hoberg and Phillips（2018）根据 Hoberg and Phillips（2016）提出的文本网络行业分类方法，重新检验了行业收益率的动量效应。他们发现相比于使用传统的行业分类，根据文本网络行业分类产生的行业收益率动量效应更加稳健并且具有强度大、持续时间长等特点。这是由于与按照传统 SIC 分类的行业内公司关联度相比，按照文本网络分类的行业内的公司关联程度可见性更低。因此，后者容易产生更严重的市场反应不足现象，从而导致长期并且显著的动量效应。他们进一步验证了这种动量效应可以被关注理论来解释。

3. 度量和预测经济周期

如何追踪和实时预测经济周期是经济学中的一个重要问题。由于衡量经济活动的主要变量 GDP 增长率无法实时观测，传统做法是使用市场上存在的一些即时指示变量，如金融市场、劳动力市场的数据等来作为反映经济活动的一致性指标。但这些方法存在的问题是，一方面这些指示变量和 GDP 增长率之间的关系不稳定，另一方面使用高频金融数据只能反映经济层面的一部分信息，很难判断是何种信息因素在影响或反映经济变动状况。与传统数据相比，新闻数据覆盖领域广泛，信息可以被很多经济个体所获取，并且新闻内容可能与经济当前和未来状态密切相关。基于这一思想，Thorsrud（2019）基于挪威日度频率的商业新闻数据，结合季度 GDP 增长率数据，构建了日度经济周期指数。他首先使用 LDA 模型从新闻数据中提取出了 80 个主题（财政政策、税收、货币政策等）。然后他根据各个主题的语调（正面或负面），在混频时变动态因子模型的框架下，估计出了日度频率的新闻即时经济周期指数。他发现，相比于使用现有的经济变量和一些复杂的经济模型，该指数能够更准确地预测和划分经济周期，并且样本外预测也有很高的准确性。另外，不同时期文章所包含的主题不同，因此对该经济周期指数分解出影响其波动的新闻类别，能够进一步推测出驱动或反映经济波动的因素。

除了构建经济周期之外，文献也有研究致力于考察媒体情绪和经济周期之间的关系。例如，Shapiro *et al*.（2018）使用 1980—2015 年美国 16 家主流新闻媒体的经济和金融相关的新闻数据，结合词典法和机器学习方法（商业公司提供的软件包）构建了反映经济状况的月度频率的情绪指数，该情绪指数包括负面、忧虑、满意等多个度量维度。他们研究了这些新闻情绪与当前经济状态的相关性，以及这些情绪的变动对当前和未来经济状态的影响程度。他们发现新闻情绪中包含了能够预测未来经济状况的信息：媒体情绪与联邦基金利率、非农就业率、行业产出、实际个人消费支出等重要的经济周期指示变量存在很强的同期关系，并且对通货膨胀率和联邦基金利率等经济变量有预测能力。他们还发现新闻冲击和总需求冲击作用相似，即当媒体情绪变差的时候，就业率、通货膨胀率、联邦基金利率均会显著下降。

4. 媒体报道偏差

文本大数据有助于度量一些对经济政治生活比较重要、但过去无法量化的指标，如媒体报道偏差（media slant）。Gentzkow and Shapiro（2010）使用美国新闻报纸数据，根据媒体新闻语言与国会共和党和民主党语言的相似性，构建了媒体报道偏差指数。他们首先利用 2005 年国会议员发言记录数据，提取与民主党和共和党国会议员意识形态高度相关的 1 000 个短语。然后将这些短语和政党意识形态对应起来，根据这 1 000 个短语在议员报告中出现的相对频率和议员的政治形态，回归找出最能预测党派的特征性短语和相应回归系数。最后使用 2000—2005 年英文新闻头条数据，通过统计新闻报纸中这些短语出现的频率并结合回归系数对报纸的报道偏差进行分类。该方法在样本内的分类估计结果和真实分类结果的相关系数为 0.61；而将分类出的媒体报道偏差和用户对这些报纸的政治形态评级数据进行比对，发现两者的相关系数高达 0.4。作者认为分类效果较好，并进一步研究媒体报道偏差与读者需求之间的关系。

5. 量化央行政策沟通

央行使用定期发布政策报告、新闻发布会和讲话等政策沟通手段向市场传递货币政策信息。从这些政策性文本数据中提取的信息，常被用于研究央行政策沟通内容对金融市场的影响。Lucca and Trebbi（2009）根据联邦公开市场委员会（The Federal Open Market Committee，FOMC）的会议内容，结合词典法量化了 FOMC 会议内容对于未来政策利率的预测方向和强度，并发现会议内容的变化是预测国库券收益率变化的主要因素。Hansen and McMahon（2016）则结合 LDA 主题模型，从 FOMC 会议内容中提取了与当前经济状态相关的五个主题，并使用词典法度量了 FOMC 声明中的语调。根据这些信息，他们研究了央行沟通内容对于市场的影响是否持续以及对实际经济变量是否有影响。Hansen *et al*.（2018）在此基础上，采用 LDA 主题模型的方法提取信息，进一步研究了央行政策沟通透明程度对货币政策制定者商

议过程的影响。

对于中国人民银行政策沟通效果的研究主要从事件分析出发，即是否有政策沟通，研究的角度包括：政策沟通的类别、参与政策沟通的人员信息等（McMahon *et al.*，2018）。McMahon *et al.*（2018）指出中国人民银行常用的沟通渠道包括：《货币政策执行报告》、货币政策委员会会议记录、行长与副行长的新闻发布会和讲话、公开市场操作报告等。鉴于中文文本的复杂性，并且人民银行政策沟通内容中的语言往往经过仔细斟酌，简单的机器学习方法不容易直接识别，因此少有研究直接从沟通内容中提取信息。随着全球对中国货币政策信息的兴趣日益高涨以及中国的全球影响力不断增强，会有更多研究致力于从政策沟通内容中获取对人民银行对未来货币政策的预期方向及强弱程度的变化等问题更为丰富的解读。因此，使用恰当的机器学习方法来量化人民银行政策沟通内容可能是未来值得关注的研究领域。

（二）在金融学中的应用

文本大数据在金融学的应用主要涉及度量关注度、情绪、可读性、隐含波动率、意见分歧和行业关联性等六方面指标的构建，以及这些指标与市场表现之间的关系。基于文本的行业分类相关研究，我们已在经济学应用部分作了简述，这里侧重对前五个方面文献的总结。

1. 关注度指数

金融理论指出，关注是一种稀缺资源（Kahneman，1973），信息需要先被投资者关注到，才能通过投资者交易行为传递到资产价格中（Ben-Rephael *et al.*，2017），因此关注是信息反应的前提。从度量个体角度，关注度可分为投资者关注度（散户投资者和机构投资者）、媒体关注度和分析师关注度。由于现有文献对分析师关注的研究较少，本文主要梳理投资者关注和媒体关注相关应用。

（1）投资者关注度

散户和机构投资者是金融市场的直接参与者，研究他们的关注行为有助于理解资产价格的变动。Barber and Odean（2008）认为，由于购买股票时散户从他们关注的股票列表中做选择，但卖出股票时只能从持仓中选择，散户投资者关注度增加会导致暂时的价格上升。机构投资者持有更多股票、信息加工能力更强，因此通常不存在有限关注度约束。要检验 Barber and Odean（2008）理论，关键是如何度量两类投资者的关注。

传统的关注度度量方法选择市场变量等作为关注度的代理变量，如交易量（Barber and Odean，2008；Hou *et al.*，2009）、超额收益率（Barber and Odean，2008）、广告费用（Lou，2014）等。但 Da *et al.*（2011）指出，与

投资者关注无关的因素也可以引发这些变量的变动。近年来的研究开始直接用文本大数据构建散户关注度指标。

用文本数据度量散户投资者关注度的方法主要有两类，一类是利用网络搜索引擎统计对上市公司的搜索次数，另一类是利用网络论坛上股民对特定股票的发帖数量。相较于传统交易数据或财务报表数据，网络大数据时刻记录着投资者的行为，能够直接揭示投资者心理状态（张学勇和吴雨玲，2018）。Da *et al*.（2011）最早提出并使用搜索次数度量投资者关注，他们根据谷歌趋势提供的搜索指数，使用 Russell 3000 成分股的代码作为关键字，构建了特定股票的投资者关注度。他们的发现与 Barber and Odean（2008）的关注理论一致，高散户关注度预测了短期更高的收益率，但长期存在收益率反转。Antweiler and Frank（2004）则使用雅虎财经网络论坛的帖子数量来近似关注度，发现关注度对收益率和市场波动率均有预测能力，但对收益率的预测并不具有经济上的显著性。Tsukioka *et al*.（2018）使用雅虎财经日本板块上 654 家公司的帖子数据度量投资者关注度，并发现投资者关注可以解释日本上市公司的 IPO 抑价现象。

国内采用文本数据度量散户投资者关注的研究与国外做法类似，也是或者采用搜索指数或者使用论坛发帖量。在使用搜索指数方面，宋双杰等（2011）使用中国 A 股 825 家上市公司的名称作为关键词，从谷歌趋势上获取这些公司的每周搜索量数据，并参照 Da *et al*.（2011）用周度异常搜索量来度量投资者关注，他们发现投资者关注可解释中国市场的 IPO 异象。俞庆进和张兵（2012）则采用百度搜索构建创业板 196 家公司个体投资者关注度，并发现中国创业板市场也存在投资者有限关注现象。张谊浩等（2014）使用百度搜索指数研究了投资者网络搜索行为与资产定价间的关系，在关注度和短长期收益率以及交易量的关系方面，得到的结论和 Da *et al*.（2011）在美国市场的发现基本一致。在使用网络论坛发帖量来度量散户投资者关注度的文献中，Huang *et al*.（2016）的发现是中国市场上投资者的关注具有本地偏好特征；杨晓兰等（2016）发现本地关注度与交易量存在正相关；段江娇等（2017）则发现帖子数与当日及未来的股票收益率显著负相关，但与当日及未来的股票波动率显著正相关。

由于缺乏直接反映机构投资者关注的文本数据，直接使用文本数据度量机构投资者关注的研究较少，Ben-Rephael *et al*.（2017）是首篇这类文献。通过分析 Bloomberg 的用户特征，他们发现 Bloomberg 使用者主要是机构投资者，并采用 Bloomberg 终端记录的用户对股票新闻的搜索和阅读频率数据来度量美国市场机构投资者关注度。他们发现与散户投资者关注相比，机构投资者关注对重大消息和事件反映更迅速，并且机构投资者关注领先于散户

投资者关注。

（2）媒体关注度

媒体关注度（media coverage）反映的是媒体对于特定上市公司、行业或市场的关注程度，通常通过统计特定新闻媒体所发布的与金融市场、上市公司相关的新闻数量来构建。作为金融市场的信息制造和传播者，媒体的关注一方面可以影响市场参与者的关注，另一方面也影响市场信息的传播效率和模式。媒体关注对市场影响的研究，主要从它对资产价格、对管理层行为和分析师行为影响等角度展开。

从对资产价格影响的角度看，Fang and Peress（2009）选取了 1993—2002 年《纽约时报》（*The New York Times*）、《今日美国》（*USA Today*）、《华尔街日报》和《华盛顿邮报》（*The Washington Post*）上关于 NYSE 和 NASDAQ 上市公司的新闻报道数据，从横截面研究了媒体关注与资产收益率的关系，并发现媒体关注低的公司的股票未来收益比媒体报道程度高的公司更高。Zou *et al*.（2019）发现中国股票市场上媒体关注和公司未来股票收益率之间也存在类似关系。Hillert *et al*.（2014）使用 1989—2010 年 45 家美国报纸约 220 万条新闻数据研究了媒体关注与股票市场动量效应的关系，他们的发现可总结为受关注更高的公司的收益率可预测性更强，因此他们认为媒体关注会导致更严重的投资者偏差。

从对管理层行为影响的角度看，Dyck *et al*.（2008）使用俄罗斯 1999—2002 年的公司治理的违规数据，研究了媒体报道与公司违规行为的关系。他们发现《金融时报》（*Financial Times*）和《华尔街日报》等国际媒体对违规事件的关注度越高，公司纠正违规行为的概率越高。周开国等（2016）在中国市场上研究了媒体监督与上市公司违规频率之间的关系，也发现媒体关注度的提高会延长公司的违规间隔、降低违规频率。

从对分析师行为影响的角度看，周开国等（2014）发现媒体关注度可以影响分析师关注度，从而提高其盈余预测的准确度。谭松涛等（2015）则发现媒体关注度能够降低分析师的预测乐观度和预测偏差。

2. 文本情绪

因为情绪的变化可能会导致资产价格偏离正常水平（De Long *et al*.，1990），度量情绪（sentiment）是文本大数据在金融领域的一大应用。文献中情绪常有正面和负面、乐观和悲观、积极和消极、牛市和熊市、看涨和看跌等不同表述，也常用“语调”（tone）来表示“情绪”。根据情绪主题的不同，文本情绪研究对象主要包括媒体语调（媒体新闻）、管理层语调（上市公司年报的管理层讨论与分析、盈利电话会议和其他公开信息披露文件）、投资者情绪（网络论坛发帖）等。与情绪度量有关的文献主要从媒体情绪、管理层情

绪和投资者情绪三个方面展开。

(1) 媒体情绪（语调）

媒体情绪度量媒体报道内容中包含的乐观与悲观情绪。国外文献使用《华尔街日报》《纽约时报》《华盛顿邮报》等文本数据来度量媒体情绪，并研究媒体情绪与股票市场的关系。这些研究主要从媒体情绪对大盘和上市公司的影响、情绪影响在繁荣期和衰退期的非对称性，以及正负词语影响非对称性等角度展开。例如，Tetlock（2007）研究了《华尔街日报》专栏文章与随后股市收益和交易量之间的关系，发现消极词语频率上升时股市收益率会下降。Tetlock *et al*.（2008）使用 1980—2004 年《华尔街日报》和道琼斯新闻社上与标普 500 公司相关的约 35 万条新闻数据，发现新闻中负面词语比例越高的公司，在下个交易日股票收益率和下个季度公司盈利都更低。Garcia（2013）使用 1905—2005 年《纽约时报》上的金融新闻来研究经济繁荣期和衰退期媒体情绪对资产价格影响的不对称性，并发现新闻在日度频率上对收益率的预测能力主要存在于经济衰退期，这是对 Tetlock（2007）的进一步拓展。除了综合考虑正负词语的总体情绪指数外，Garcia（2013）和 Zhang *et al*.（2016）还研究了正面和负面语调作用的非对称性。其中，Garcia（2013）发现正负语调均能预测大盘收益率，而 Zhang *et al*.（2016）则发现公司相关新闻中，正面和负面词语比例对股票市场变量（收益率、波动率、交易量）间的影响存在非对称性，因此建议实证研究应同时考虑文本中的正负语调。

和国外文献类似，国内文献也用国内主流财经媒体报刊数据来度量媒体情绪，但无论是度量媒体情绪还是考察媒体情绪与市场变量之间的关系的研究均处于起步阶段。就方法来看，游家兴和吴静（2012）选取了 2004—2010 年国内 8 家主流财经报纸上的新闻，通过人工阅读新闻报道态度倾向的方法来衡量媒体情绪；汪昌云和武佳薇（2015）使用 6 家主流财经媒体的新闻数据，结合自定义的财经媒体情绪词典统计了新闻中的正负面词语数量并构建媒体正、负面语气指数。在运用媒体情绪研究的问题上，游家兴和吴静（2012）关注情绪对沪深 A 股上市公司资产错误定价的影响，而汪昌云和武佳薇（2015）则研究了 IPO 抑价率的变化。在媒体情绪与其他市场变量，以及上市企业层面媒体情绪与市场变量之间的关系方面尚有很大研究空间。

随着我国互联网金融行业的快速发展，国内文献还研究了媒体情绪与网络借贷之间的关系。王靖一和黄益平（2018）使用和讯网 1 702 万余条新闻数据，构建了 2013 年 1 月至 2017 年 9 月间的金融科技情绪指数，用于反映媒体对金融科技的正负情感态度。他们发现媒体情绪对于个体网络借贷具有显著影响，媒体情绪转向乐观时会提高网络借贷平台交易量的增长率，并且这种

影响在问题平台上更强。张皓星和黄益平（2018）使用该指数进一步研究了互联网金融情绪与互金平台贷款违约率的关系。他们发现当金融科技情绪变差时，网络借贷违约概率会显著增加。除了平台层面的实证研究，文献还探讨了网贷借款人行为、投资人行为和网贷市场监管等（黄益平和黄卓，2018），但对媒体情绪如何影响微观个体行为的研究较少。

（2）管理层语调

除了使用财务报表直接报告公司经营状况外，上市公司还要定期发布公司的财务报告（季报、年报）、季度盈余公告、管理层盈余公告、招股说明书等文件。这些文件包含上市公司管理层对当前经营状况的分析和未来发展方向的讨论，因此往往能反映管理层的决策和意图。

从这类文本数据中提取文本情绪并研究其对上市公司市场表现的影响就属于对管理层语调的研究。例如，Li（2010）从美国上市公司年报和季报文件的管理层讨论与分析（Management Discussion and Analysis，MD&A）前瞻性说明部分构建管理层语调，发现这些语调与公司未来盈利正相关。Loughran and McDonald（2011）利用 1994—2008 年美国上市公司的年报文件，发现年报语调与收益率、交易量、波动率、未预期盈利等市场变量相关。Jegadeesh and Wu（2013）研究了 1995—2010 年上市公司的年报语调与年报发布期的超额收益率之间的关系，发现市场对年报的内容反应不足。他们还采用同样的方法度量了 IPO 招股说明书的语调，并发现 IPO 招股说明书的语调与 IPO 抑价显著负相关。Jiang *et al*.（2019）使用上市公司财报和电话会议记录等文本数据衡量了市场层面的经理人情绪，并发现经理人情绪能显著反向预测未来市场收益率，并且这种预测能力强于现有的宏观变量和投资者情绪度量指标。

国内文献从公司业绩、投资者交易行为等角度研究了管理层语调的影响。谢德仁和林乐（2015）使用 2005—2012 年中国上市公司年度业绩说明会的文本数据，发现业绩说明会中的管理层语调与未来公司的业绩显著正相关。曾庆生等（2018）研究了 2007—2014 年 A 股非金融公司年报语调与公司高管的交易行为，发现积极的年报语调预示公司高管随后的卖出股票规模大、净买入股票规模小。

（3）投资者情绪

实证检验投资者情绪与资产价格之间的关系首先需要测度投资者情绪。De Long *et al*.（1990）将投资者情绪定义为噪声交易者（Noise Trader）关于股票未来股价偏离理性套利者信念的程度。传统的投资者情绪度量方法分

市场变量法和调查法。Baker and Wurgler（2006）的投资者情绪指数[12]是目前文献中应用最广泛的基于市场变量法的投资者情绪指数。他们选取封闭式基金折价率、NYSE 股票换手率、IPO 数量及上市首日收益率、新发行权益份额和股利溢价作为情绪的代理变量来构建情绪指数。调查法则通过问卷调查（电话、邮件等）来收集个体对当前或未来经济状况、金融市场走向的看法和态度，并将这些问卷结果汇总成指数。密歇根大学的消费者信心指数[13]是调查法的经典代表。传统方法的弱点在于，第一，作为投资者情绪代理变量的市场变量可能不只反映投资者情绪，还是情绪与其他经济因素相互作用后的均衡结果（Qiu and Welch，2006；Da *et al.*，2014）；第二，调查法虽然直接度量投资者情绪，但其实施成本高、构建情绪指数频率较低、时间跨度也比较短。

文本大数据为度量投资者情绪提供了新的数据源。一方面，由于投资者倾向于选择在网络论坛上发布与股票相关的评论帖子或者做出相关搜索，这些文本数据能直接反映他们对公司未来的看法，对市场当前状态的解读，以及与自身投资决策相关的信息。另一方面，这些数据具有易获得、时间跨度长、覆盖公司数量多等特点，满足了从不同频率、不同层面研究情绪与资产价格关系的需求。

国外度量投资者情绪的文献较为丰富，表 1 总结了其中的代表文献。其中度量投资者情绪的文本数据主要来源于雅虎财经帖子（Antweiler and Frank，2004；Kim and Kim，2014；Tsukioka *et al.*，2018）、微博平台（Renault，2017）、推特（Behrendt and Schmidt，2018）、专业数据库（Sun *et al.*，2016）和谷歌搜索（Da *et al.*，2014；Gao *et al.*，2019）。相应的，数据类型主要是帖子、推文或者搜索的关键词等。数据频率则涵盖了周频、日频，甚至更高频率（如：日内半小时频率的情绪指数）。利用构造出的文本情绪指数，文献研究了文本情绪和同期收益率、未来收益率、波动率、交易量、IPO 折价之间的关系。这些研究发现，在日度频率上，投资者情绪与同期收益率正相关，但基本对未来收益率、波动率、交易量没有很强的预测能力（Antweiler and Frank，2004）。在更高的半小时频率上，一些研究发现了投资者情绪对收益率存在一定日内预测能力（如 Sun *et al.*，2016；Renault，2017）；对于波动率的日内预测能力也有一定证据，但其经济意义不显著（Behrendt and Schmidt，2018）。目前为止，国外文献基本指向基于文本的投资者情绪的预测能力主要体现在对日内收益率的可预测性，而这一预测能力主要由于噪声交易者的交易行为导致（Sun *et al.*，2016）。美国股票市场比

[12] http：//people. stern. nyu. edu/jwurgler/，访问时间：2019 年 7 月 15 日。
[13] http：//www. sca. isr. umich. edu/，访问时间：2019 年 7 月 15 日。

较成熟，并且以机构投资者为主，散户投资者占比远小于中国散户投资者占比。因此，散户投资者情绪对于市场表现的影响主要在日内体现。

表 1 投资者情绪相关研究

文献	数据来源	数据类型	情绪指数类型	主要发现
Antweiler and Frank（2004）	2000 年雅虎财经和 Raging Bull	道琼斯工业平均指数和道琼斯互联网指数的 45 家公司约 150 万条帖子数据	个股层面投资者情绪指数	情绪与同期收益率显著正相关，与未来收益率不相关
Kim and Kim（2014）	2005—2010 年雅虎财经	91 家公司约 3 200 万条帖子	个股层面投资者情绪指数	情绪不能预测收益率、波动率、交易量，但受收益率影响
Tsukioka *et al*.（2018）	2001—2010 年雅虎财经日本股票版块	654 家公司相关的帖子数据	个股层面投资者情绪指数	投资者情绪可用于解释 IPO 抑价现象
Sun *et al*.（2016）	1998—2011 年汤普森路透数据库	标普 500 指数对应的 1 分钟频率情绪数据	市场层面投资者情绪指数	日内半小时情绪变化可预测日内收益率
Renault（2017）	2012—2016 年美国社交媒体平台 Stock Twits	约 6 000 万条帖子	市场层面投资者情绪指数	第一半小时投资者情绪变化能预测标普 500 指数 ETF 最后半小时收益率，但下个交易日反转
Behrendt and Schmidt（2018）	2015—2017 年推特	道琼斯指数成分股情绪数据（1 分钟频率）	个股 Twitter 情绪	情绪与日内波动率存在反馈效应，但经济意义不显著
Da *et al*.（2011）	2004—2011 年谷歌搜索	118 词日搜索频率	FEARS 指数近似市场情绪	FEARS 能预测股票市场波动，与收益率存在同期正相关，与随后收益率负相关
Gao *et al*.（2019）	2004—2014 年谷歌搜索	词搜索频率	周度国别投资者情绪指数	投资者情绪指数负向预测未来一周股票市场的收益率；并检验定价困难和有限套利等可预测性渠道

国内文献关于从文本数据度量投资者情绪的研究与国外研究类似，也是从网络平台数据出发，根据股民发布的帖子来构建中国股票市场的投资者情绪指数，并检验投资者情绪与市场变量的关系。杨晓兰等（2016）使用东方财富股吧上与创业板股票相关的约一年时间的 90 多万条帖子构建了投资者情绪指数，并研究情绪指数与收益率、交易量等市场变量间的同期关系。段江

娇等（2017）使用东方财富网 2011—2012 年上证 A 股约 466 万条帖子构建了日度频率的投资者情绪指数，他们也发现论坛情绪与公司股票收益率存在同期相关性，但并不能预测未来股票收益率。

简而言之，现有国内外文献都观察到投资者情绪与市场变量间的同期关系或者投资者情绪受市场变量影响，但投资者情绪的预测能力有限。这一现象一方面可能由于在有效的市场中，情绪对于市场变化的作用有限，另一方面也可能是由于情绪与市场变量同时变化的内生性导致低估了情绪的作用。Li *et al*.（2019）则考虑利用中国股市开盘收盘时间段特征来建立因果关系。具体而言，他们使用 2008—2018 年中国股吧论坛的数据，结合词典法和机器学习方法构建了投资者隔夜情绪，即每个交易日收盘后到第二个交易日上午 9 点 15 分集合竞价前这一时段的投资者情绪。由于这段时间内只有投资者在论坛发表言论而没有价格信息，因此隔夜情绪与第二个交易日的开盘价之间只存在单向关系。他们发现隔夜投资者情绪能够显著预测隔夜收益率、第二天的波动率以及交易量。根据他们的研究，网络论坛帖子反映的投资者情绪不只是噪音，还包含了关于未来市场的重要信息。

3. 文本可读性

除了文本的数量和语调，文本数据中另一个维度的信息是文本的可读性（复杂性）。它从文本的内在逻辑和结构出发，根据文本的大小、文本中的词语数量、句子的长度、复杂词语的比例以及包含的图表数量等文本特征，来衡量获取文本中信息的难易程度。由于文本可读性会影响读者获取文本信息的难易程度，文本信息发布者（上市公司、借款人）可通过发布可读性差的信息来增加读者信息获取成本的方式来隐藏坏消息；或采用可读性高的方式，以提高信息传播速度和影响力的方法来发布好消息。因此研究文本可读性，对于理解信息发布者、接受者的行为以及市场的反应至关重要。

近年来金融文献中常用的文本可读性度量指标是迷雾指数（Fog Index），它最早由 Gunning（1952）提出。迷雾指数使用两个标准来判断文本可读性：文本的平均词语长度和文本中复杂词语的比例，其中复杂词语的定义是至少有两个音节的词语（英文环境）。迷雾指数越高则文本的可读性越低。Li（2008）是最早使用迷雾指数来研究年报可读性与公司业绩表现的文献。他将迷雾指数和年报中的词语数量结合起来，度量了 1994—2004 年美国上市公司年报的可读性，实证发现年报可读性差（迷雾指数高或词语数量多）的公司会表现出更低的盈利水平。此后不少文献参照 Li（2008）的做法来度量年报可读性，包括研究年报可读性与投资者的投资行为之间的关系（如：Miller，2010；Lawrence，2013），年报可读性与分析师关注和预测行为的关系（如：Lehavy *et al*.，2011）。

当然，使用迷雾指数度量文本可读性也存在一些局限性。Loughran and

MacDonald（2014）指出，如果按照迷雾指数的标准定义复杂词语，那么金融文本中复杂词语比重非常高，但这些词语很容易被投资者和分析师理解，所以迷雾指数并不适用于评估金融文档的可读性。他们提出的替代方式是直接使用公司年报文件的大小。这种方法简单、易复制，不需要对文档进行解析，可以减少可读性的度量误差。他们使用 1994—2011 年美国上市公司 66 707 份年报文档数据比较了两种可读性度量指标，发现相比于迷雾指数，年报文件的大小更能准确地度量文本的可读性：年报文件越大的公司在发布年报文件后的收益率、波动率、盈余预测误差、盈余预测分歧都会增大。

中文文献关于年报可读性的度量通常是在国外研究的基础上，结合中文的特点和语言环境进行调整和改进。丘心颖等（2016）根据年报中完整句子（含有主谓结构的句子）的占比、基础词汇（汉语水平考试 1—3 级的词汇）占比以及汉字的笔画数等多个指标构建了中文年报可读性指数，研究了年报可读性对分析师关注、预测信息含量以及预测准确性的影响。孟庆斌等（2017）对 A 股上市公司年报的 MD&A 内容进行分析，根据文本中常用汉字词语所占比重来度量文本可读性。他们发现可读性越高时，MD&A 中的信息含量对股价崩盘风险的影响越强。王克敏等（2018）则根据文本逻辑和字词的复杂性来共同刻画年报文本的可读性，研究了管理者操作年报复杂性的动机。除了年报的可读性，国内文献还对其他文本数据的可读性进行了研究。陈霄等（2018）使用中国“人人贷”平台上的借款订单数据，研究了借款描述可读性与网络借款成功率之间的关系。他们采用借款描述中，除去标点符号外字词的数量来衡量借款描述的可读性。他们发现可读性较高的借款描述向投资者传递了积极信号，可以提高借款的成功率。

4. 新闻隐含波动率指数

除了用来度量媒体关注度和媒体语调外，新闻文本还被用来度量金融市场的不确定性。Manela and Moreira（2017）使用《华尔街日报》1890—2009 年头版数据，采用支持向量回归法将新闻文本数据中出现的词语和市场上的波动率指数（VIX）相对应，并构建了新闻隐含波动率指数（News Implied Volatility，NVIX）。

该指数的具体构建方法如下：①从每个月的所有新闻文章中提取全部词语出现的频率，采用独热表示法构建向量X_t，即X_t的长度为所有新闻中单词的个数，每个位置的值表示该词语在该月文章中出现的频率；②将X_t与 VIX 指数v_t构建映射：$v_t=w_0+w\times X_t+\upsilon_t$；③ 将 VIX 数据样本拆分为训练集和测试集，在训练集上使用支持向量回归方法拟合上述方程，得到系数的估计值；④根据每个月构建的新闻向量X_t，向前估计 NVIX 指数。

基于新闻数据构建的 NVIX 指数跨度从 1890 年到 2009 年，该指数与历史上的重要事件（第一次世界大战、第二次世界大战、大萧条等）非常吻合，

间接验证了该指数很好地刻画了市场的不确定性。他们发现 NVIX 可以正向预测市场 6 到 24 个月的收益率。进一步将新闻中的词语分为四类：股票市场、战争、政府、金融中介，他们发现 NVIX 预测收益率的能力主要受新闻中与战争、政府相关的词语的比例的影响。

NVIX 指数的构建思想还可以应用到其他文本数据上，通过选择不同的文本特征X_t，结合不同的市场变量v_t，包括超额收益率、交易量、波动率等，寻找这些文本特征跟市场变量之间的对应关系，提取更丰富的文本隐含信息。

5. 投资者分歧

投资者分歧衡量了投资者的异质信念。传统金融理论指出，投资者分歧会产生交易（Harris and Raviv，1993），因此文献关心分歧与交易量、价格之间的关系。常用的度量投资者分歧的指标包括分析师预测分散程度（Yu，2011）、经济政策不确定性指数（Bollerslev *et al*.，2018）和对经济变量预测的分散程度等。近年来，文献中开始从文本数据出发，构建直接度量投资者分歧的指标。

Antweiler and Frank（2004）使用网络留言板的帖子数据，计算出帖子的情绪得分，然后根据帖子情绪的标准差构建了投资者分歧指数。他们发现投资者分歧与同期的交易量显著正相关，验证了投资者分歧产生交易的理论。段江娇等（2017）在中国市场也有类似的发现。他们使用东方财富股吧上的帖子数据构建了日度频率的投资者分歧，发现投资者分歧越大，未来两天的交易量也越大。

投资者分歧对价格也会产生影响。Miller（1977）指出当市场上的投资者观点不同时，乐观的交易者会推动价格上升，而悲观的交易者由于存在卖空约束的限制，并不能完全消除由乐观交易者导致的错误定价，导致资产价格被高估。因此当投资者的分歧很大时，资产价格会被高估，未来收益率会更低。Hillert *et al*.（2018）使用 1989—2010 年美国主流媒体的新闻数据，先用词典法计算出每篇新闻的语调，再计算公司层面的媒体分歧程度（公司 i 在第 t 天的分歧程度为当天与该公司相关的全部新闻语调的标准差），最后将公司层面的媒体分歧程度加总并平均得到日度市场层面的媒体分歧程度指数。他们发现，媒体分歧程度与第二天的市场收益率显著负相关，并且这种关系在经济处于衰退时期更强。在公司层面，他们还发现媒体分歧对于 Beta、分析师预测分散程度、换手率、特质性波动率更高的公司的影响更大。

在本部分我们对文本大数据在经济和金融领域的运用作了简单的梳理，自然不能穷尽目前日渐涌现的各个子领域的新文献，而一些文献也不能简单分类到一个子领域。例如，Soo（2018）使用 2000—2013 年美国 34 个城市的地方房地产行业新闻数据，采用词典法构建了房地产情绪指数，他们发现媒体新闻所反映的房地产行业情绪能够预测未来的房价变化，这是媒体情绪相

关文献在经济学领域的运用。又如 Boudoukh *et al*.（2019）使用媒体新闻数据，结合 SVM 方法来识别新闻是否与公司特定事件相关联。他们根据新闻关联程度将样本分类为：没有新闻、不相关新闻、相关新闻等信息时期，研究比较了股票收益率的波动率在不同信息时期以及交易和非交易时间的差别。这是从媒体文本数据中提取其他维度信息的研究。另外，手机、摄像机等产生的数据也以文本形式被用于研究中，如 Athey *et al*.（2018）采用手机产生的数据研究了消费者对餐厅的选择。当然，文本大数据在经济和金融外的其他领域，如社会学、政治学等领域也有精彩运用，可参考 Gentzkow *et al*.（2019）去参阅相关经典文献。

四、结论和展望

和经典的数据分析方法相比，文本大数据给经济学和金融学的实证研究至少带来了四个变化。第一，经典实证分析采用的数据往往已经是结构化数据，而文本大数据往往是非结构化数据。非结构化数据向结构化数据转换的实现方式并不是一个简单的问题，而不同转换方式会直接影响后续分析结果，因此可以预见的是，高质量的文本大数据分析，需要对这一转换过程做更为详细的介绍。第二，经典实证分析数据中的变量定义往往比较清晰，这种清晰的边界往往是通过在收集数据时设计的问卷已经准确定义了变量的含义来实现的，又或者是根据实际经济和金融活动运行的需要来事先界定好的。而文本大数据的数据来源是新闻媒体、网络论坛、公司财报等文本文件，本身并不包含清晰的变量，如何提取信息，并且如何论证作者提取的就是目标信息，也将是文本大数据分析的重要步骤。第三，经典实证方法往往采用获取的全样本数据做回归分析，然后通过不同的设定和变化做稳健性检验。将数据分为训练集、验证集和测试集，展开交叉验证的方法虽然早就是经典方法，在过去的经济和金融实证分析中的应用还不够充分。由于这些方法在文本大数据做预测的分析范式中较为常见，这些做法对于经济和金融实证研究范式也会产生一定影响。第四，使用文本大数据需要有跨学科领域的人才。比如卷积神经网络、循环神经网络等模型在经济和金融领域的运用，需要研究人员不仅对经济和金融领域有较为深入的掌握，同时对不同算法的特点和优劣都要有较为丰富的知识。

未来几年，在经济和金融领域运用文本大数据研究方面，可能会有如下趋势。一是研究将开拓更为丰富的数据源。目前文本数据库的主要数据源包括新闻、网络论坛帖子、公司财报、消费者评价、重要人物的发言等。但还有大量的文本数据尚未被研究者所使用，如政府工作报告和规划、网络自媒体公众号文章、微博大 V 观点、书籍、档案、专利网站、法院判决、医生处

方等。运用有监督或者无监督的机器学习方法、深度学习方法来分析这些数据在未来几年将是研究热点。二是采用文本数据研究的问题会更为深入和广泛，如如何在中文语境以及复杂的长文本下构建情绪指数并用于预测；如何从文本数据的文件大小、数字和汉字的占比、图表数量、句子长短等角度加强对文本可读性的研究，又如如何从国内主流新闻数据提取隐含信息（波动率、不确定性等），并考察文本信息与资本市场、宏观经济层面、微观个体表现等之间的关系。三是采用文本大数据展开的研究将不仅满足于基于相关关系的预测问题，因果关系相关的研究也将逐渐进入研究人员的视野。例如，Athey（2017）就考虑了如何将机器学习用到基于文本、摄像头等产生的数据而展开的政策效果的评估中。四是对文本数据常用的机器学习、深度学习方法和经典计量经济学方法之间的联系和差异的评估将得到重视。经典计量经济学框架下，观察到的因变量值是一个数据生成过程的实现值，而采用机器学习、深度学习方法中的有监督学习需要使用颇为主观的标注数据，无监督学习的提炼过程也并非简单明了，由此生成的数据是否能够准确反映待提炼的潜在变量，需要更多研究。最后，由于文本大数据分析往往需要同时运用经济、金融、计算机、心理学等多个领域的知识和技术，对高素质的跨学科人才将产生较大需求，因此研究机构和高校可按照自身的学科优势来培养跨学科、复合型的研究人才。

参 考 文 献

[1] Airoldi, E. M., E. A. Erosheva, S. E. Fienberg, C. Joutard, T. Love, and S. Shringarpure, "Reconceptualizing the Classification of PNAS Articles", *Proceedings of the National Academy of Sciences*, 2010, 107 (49), 20899-20904.

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

[3] Athey, S., "Beyond Prediction: Using Big Data for Policy Problems", *Science*, 2017, 355 (6324), 483-485.

[4] Athey, S., D. Blei, R. Donnelly, F. Ruiz, and T. Schmidt, "Estimating Heterogenous Consumer Preferences for Restaurants and Travel Time Using Mobile Location Data", Working Paper, 2018.

[5] Bachmann, R., S. Elstner, and E. R. Sims, "Uncertainty and Economic Activity: Evidence from Business Survey Data", *American Economic Journal: Macroeconomics*, 2013, 5 (2), 217-249.

[6] Baker, S. R., N. Bloom, and S. J. Davis, "Measuring Economic Policy Uncertainty", *The Quarterly Journal of Economics*, 2016, 131 (4), 1593-1636.

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

[8] Bali, T. G., S. J. Brown, and Y. Tang, "Is Economic Uncertainty Priced in the Cross-Section of Stock Returns?", *Journal of Financial Economics*, 2017, 126 (3), 471-489.

[9] Bali, T. G., and H. Zhou, "Risk, Uncertainty, and Expected Returns", *Journal of Financial and*

Quantitative Analysis, 2016, 51 (3), 707-735.

[10] Barber, B. M., and T. Odean, "All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors", *The Review of Financial Studies*, 2008, 21 (2), 785-818.

[11] Behrendt, S., and A. Schmidt, "The Twitter Myth Revisited: Intraday Investor Sentiment, Twitter Activity and Individual-Level Stock Return Volatility", *Journal of Banking & Finance*, 2018, 96, 355-367.

[12] Ben-Rephael, A., Z. Da, and R. D. Israelsen, "It Depends on Where You Search: Institutional Investor Attention and Underreaction to News", *The Review of Financial Studies*, 2017, 30 (9), 3009-3047.

[13] Bishop, C. M., *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[14] Blei, D. M., and J. D. Lafferty, "Dynamic Topic Models", *Proceedings of the 23rd International Conference on Machine Learning*. ACM, 2006, 113-120.

[15] Blei, D. M., A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, 2003, 3, 993-1022.

[16] Bollerslev, T., J. Li, and Y. Xue, "Volume, Volatility, and Public News Announcements", *The Review of Economic Studies*, 2018, 85 (4), 2005-2041.

[17] Boudoukh, J., R. Feldman, S. Kogan, and M. Richardson, "Information, Trading, and Volatility: Evidence from Firm-Specific News", *The Review of Financial Studies*, 2019, 32 (3), 992-1033.

[18] Box, T., "Qualitative Similarity and Stock Price Comovement", *Journal of Banking & Finance*, 2018, 91, 49-69.

[19] Brogaard, J., and A. Detzel, "The Asset-Pricing Implications of Government Economic Policy Uncertainty", *Management Science*, 2015, 61 (1), 3-18.

[20] 陈胜蓝、李占婷，"经济政策不确定性与分析师盈余预测修正"，《世界经济》，2017 年第 7 期，第 169—192 页。

[21] 陈胜蓝、刘晓玲，"经济政策不确定性与公司商业信用供给"，《金融研究》，2018 年第 5 期，第 172—190 页。

[22] 陈霄、叶德珠、邓洁，"借款描述的可读性能够提高网络借款成功率吗"，《中国工业经济》，2018 年第 3 期，第 174—192 页。

[23] Da, Z., J. Engelberg, and P. Gao, "In Search of Attention", *The Journal of Finance*, 2011, 66 (5), 1461-1499.

[24] Da, Z., J. Engelberg, and P. Gao, "The Sum of All FEARS Investor Sentiment and Asset Prices", *The Review of Financial Studies*, 2014, 28 (1), 1-32.

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

[26] Davis, A. K., J. M. Piger, and L. M. Sedor, "Beyond the Numbers: Measuring the Information Content of Earnings Press Release Language", *Contemporary Accounting Research*, 2012, 29 (3), 845-868.

[27] 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.

[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"，《金融研究》，2017 年第 10 期，第 178—192 页。

[29] Dyck, A., N. Volchkova, and L. Zingales, "The Corporate Governance Role of the Media: Evidence from Russia", *The Journal of Finance*, 2008, 63 (3), 1093-1135.

[30] Elman, J. L, "Finding Structure in Time", *Cognitive Science*, 1990, 14 (2), 179-211.

[31] Fang, L., and J. Peress, "Media Coverage and the Cross-Section of Stock Returns", *The Journal of Finance*, 2009, 64 (5), 2023-2052.

[32] Gao, Z., H. Ren, and B. Zhang, "Googling Investor Sentiment around the World", *Journal of Financial and Quantitative Analysis*, 2019, forthcoming.

[33] Garcia, D., "Sentiment during Recessions", *The Journal of Finance*, 2013, 68 (3), 1267-1300.

[34] Gentzkow, M., B. T. Kelly, and M. Taddy, "Text as Data", *Journal of Economic Literature*, 2019, 57 (3), 535-574.

[35] Gentzkow, M., and J. M. Shapiro, "What Drives Media Slant? Evidence from U. S. Daily Newspapers", *Econometrica*, 2010, 78 (1), 35-71.

[36] 顾夏铭、陈勇民、潘士远，"经济政策不确定性与创新——基于我国上市公司的实证分析"，《经济研究》，2018 年第 2 期，第 109—123 页。

[37] 顾研、周强龙，"政策不确定性、财务柔性价值与资本结构动态调整"，《世界经济》，2018 年第 6 期，第 102—126 页。

[38] Gulen, H., and M. Ion, "Policy Uncertainty and Corporate Investment", *The Review of Financial Studies*, 2016, 29 (3), 523-564.

[39] Gunning, R., *The Technique of Clear Writing*. New York: McGraw-Hill, 1952.

[40] Handley, K., and J. F. Li, "Measuring the Effects of Firm Uncertainty on Economic Activity: New Evidence from One Million Documents", Working Paper, 2018.

[41] Hansen, S., and M. McMahon, "Shocking Language: Understanding the Macroeconomic Effects of Central Bank Communication", *Journal of International Economics*, 2016, 99, S114-S133.

[42] Hansen, S., M. McMahon, and A. Prat, "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach", *The Quarterly Journal of Economics*, 2018, 133 (2), 801-870.

[43] Harris, M., and A. Raviv, "Differences of Opinion Make a Horse Race", *The Review of Financial Studies*, 1993, 6 (3), 473-506.

[44] Henry, E., "Are Investors Influenced by How Earnings Press Releases are Written?", *Journal of Business Communication*, 2008, 45 (4), 363-407.

[45] Hillert, A., H. Jacobs, and S. Müller, "Media Makes Momentum", *The Review of Financial Studies*, 2014, 27 (12), 3467-3501.

[46] Hillert, A., H. Jacobs, and S. Müller, "Journalist Disagreement", *Journal of Financial Markets*, 2018, 41, 57-76.

[47] Hinton, G. E., and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks", *Science*, 2006, 313 (5786), 504-507.

[48] Hoberg, G., and G. Phillips, "Text-Based Network Industries and Endogenous Product Differentiation", *Journal of Political Economy*, 2016, 124 (5), 1423-1465.

[49] Hoberg, G., and G. Phillips, "Text-Based Industry Momentum", *Journal of Financial and Quantitative Analysis*, 2018, 53 (6), 2355-2388.

[50] Hou, K., W. Xiong, and L. Peng, "A Tale of Two Anomalies: The Implications of Investor Attention for Price and Earnings Momentum", Working Paper, 2009.

[51] 黄益平、黄卓，“中国的数字金融发展：现在与未来”，《经济学》（季刊），2018 年第 17 卷第 4 期，第 1489—1502 页。

[52] Huang, Y., H. Qiu, and Z. Wu, “Local Bias in Investor Attention: Evidence from China's Internet Stock Message Boards”, *Journal of Empirical Finance*, 2016, 38, 338-354.

[53] 黄卓、邱晗、沈艳、童晨，“测量中国的金融不确定性：基于大数据的方法”，《金融研究》，2018 年第 11 期，第 30—46 页。

[54] 黄卓、童晨、梁方，“经济不确定性对金融市场的影响：一个文献综述”，《金融科学》，2017 年第 2 期，第 20—35 页。

[55] Huang, Z., C. Tong, H. Qiu, and Y. Shen, “The Spillover of Macroeconomic Uncertainty between the U. S. and China”, *Economics Letters*, 2018, 171, 123-127.

[56] Huang, A. H., A. Y. Zang, and R. Zheng, “Evidence on the Information Content of Text in Analyst Reports”, *The Accounting Review*, 2014, 89 (6), 2151-2180.

[57] Jegadeesh, N., and D. Wu, “Word Power: A New Approach for Content Analysis”, *Journal of Financial Economics*, 2013, 110 (3), 712-729.

[58] 纪洋、王旭、谭语嫣、黄益平，“经济政策不确定性、政府隐性担保与企业杠杆率分化”，《经济学》（季刊），2018 年第 17 卷第 2 期，第 449—470 页。

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

[60] Jurado, K., S. C. Ludvigson, and S. Ng, “Measuring Uncertainty”, *American Economic Review*, 2015, 105 (3), 1177-1216.

[61] Kahneman, D., *Attention and Effort*. Englewood Cliffs, New Jersey: Prentice-Hall, 1973.

[62] Kim, S., and D. Kim, “Investor Sentiment from Internet Message Postings and the Predictability of Stock Returns”, *Journal of Economic Behavior & Organization*, 2014, 107, 708-729.

[63] Kim, Y., “Convolutional Neural Networks for Sentence Classification”, in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[64] Lai, S., L. Xu, K. Liu, and J. Zhao, “Recurrent Convolutional Neural Networks for Text Classification”, in *Proceedings of AAAI*, 2015.

[65] Lawrence, A., “Individual Investors and Financial Disclosure”, *Journal of Accounting and Economics*, 2013, 56 (1), 130-147.

[66] LeCun, Y., Y. Bengio, and G. Hinton, “Deep Learning”, *Nature*, 2015, 521, 436-444.

[67] Lehavy, R., F. Li, and K. Merkley, “The Effect of Annual Report Readability on Analyst Following and the Properties of Their Earnings Forecasts”, *The Accounting Review*, 2011, 86 (3), 1087-1115.

[68] Li, F., “Annual Report Readability, Current Earnings, and Earnings Persistence”, *Journal of Accounting and Economics*, 2008, 45 (2-3), 221-247.

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

[70] 李凤羽、杨墨竹，“经济政策不确定性会抑制企业投资吗？——基于中国经济政策不确定指数的实证研究”，《金融研究》，2015 年第 4 期，第 115—129 页。

[71] Li, J., Y. Chen, Y. Shen, J. Wang, Z. Huang, “Measuring China's Stock Market Sentiment”, Working Paper, 2019.

[72] Lou, D., "Attracting Investor Attention through Advertising", *The Review of Financial Studies*, 2014, 27 (6), 1797-1829.

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

[74] Loughran, T., and B. McDonald, "Measuring Readability in Financial Disclosures", *The Journal of Finance*, 2014, 69 (4), 1643-1671.

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

[76] Lucca, D. O., and F. Trebbi, "Measuring Central Bank Communication: An Automated Approach with Application to FOMC Statements", National Bureau of Economic Research Working Paper, 2009.

[77] Manela, A., and A. Moreira, "News Implied Volatility and Disaster Concerns", *Journal of Financial Economics*, 2017, 123 (1), 137-162.

[78] McMahon, M., A. Schipke, and X. Li, "中国的货币政策沟通：框架、影响和建议", IMF 工作论文, 2018。

[79] 孟庆斌、杨俊华、鲁冰，"管理层讨论与分析披露的信息含量与股价崩盘风险——基于文本向量化方法的研究"，《中国工业经济》，2017 年第 12 期，第 132—150 页。

[80] Mikolov, T., M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, "Recurrent Neural Network Based Language Model", in *Eleventh Annual Conference of the International Speech Communication Association*, 2010, 1045-1048.

[81] Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality", in *Advances in Neural Information Processing Systems*, 2013, 3111-3119.

[82] Miller, B. P., "The Effects of Reporting Complexity on Small and Large Investor Trading", *The Accounting Review*, 2010, 85 (6), 2107-2143.

[83] Miller, E. M., "Risk, Uncertainty, and Divergence of Opinion", *The Journal of Finance*, 1977, 32 (4), 1151-1168.

[84] Murphy, K. P., *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[85] Pástor, L'., and P. Veronesi, "Political Uncertainty and Risk Premia", *Journal of Financial Economics*, 2013, 110 (3), 520-545.

[86] 彭俞超、韩珣、李建军，"经济政策不确定性与企业金融化"，《中国工业经济》，2018 年第 1 期，第 137—155 页。

[87] Pennington, J., R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation", in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, 1532-1543.

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

[89] Qiu, L., and I. Welch, "Investor Sentiment Measures", *National* Bureau of Economic Research Working Paper, 2006.

[90] 丘心颖、郑小翠、邓可斌，"分析师能有效发挥专业解读信息的作用吗？——基于汉字年报复杂性指标的研究"，《经济学》(季刊)，2016 年第 15 卷第 4 期，1483—1506 页。

[91] 饶品贵、岳衡、姜国华，"经济政策不确定性与企业投资行为研究"，《世界经济》，2017 年第 2

期，第 27—51 页。

[92] Renault, T., "Intraday Online Investor Sentiment and Return Patterns in the U. S. Stock Market", *Journal of Banking & Finance*, 2017, 84, 25-40.

[93] Rogers, J. L., A. V. Buskirk, and S. L. C. Zechman, "Disclosure Tone and Shareholder Litigation", *The Accounting Review*, 2011, 86 (6), 2155-2183.

[94] Shapiro, A. H., M. Sudhof, and D. Wilson, "Measuring News Sentiment", Federal Reserve Bank of San Francisco Working Paper, 2018.

[95] Sibley, S. E., Y. Wang, Y. Xing, and X. Zhang, "The Information Content of the Sentiment Index", *Journal of Banking & Finance*, 2016, 62, 164-179.

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

[97] 宋双杰、曹晖、杨坤，"投资者关注与 IPO 异象——来自网络搜索量的经验证据"，《经济研究》，2011 年增 1 期，第 145—155 页。

[98] Soo, C. K., "Quantifying Sentiment with News Media across Local Housing Markets", *The Review of Financial Studies*, 2018, 31 (10), 3689-3719.

[99] Sun, L., M. Najand, and J. Shen, "Stock Return Predictability and Investor Sentiment: A High-Frequency Perspective", *Journal of Banking & Finance*, 2016, 73, 147-164.

[100] 谭松涛、甘顺利、阚铄，"媒体报道能够降低分析师预测偏差吗?"，《金融研究》，2015 年第 5 期，第 192—206 页。

[101] 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。

[102] Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet Processes", *Journal of the American Statistical Association*, 2006, 101 (476), 1566-1581.

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

[104] Tetlock, P. C., M. Saar-Tsechansky, and S. Macskassy, "More Than Words: Quantifying Language to Measure Firms' Fundamentals", *The Journal of Finance*, 2008, 63 (3), 1437-1467.

[105] Thorsrud, L. A., "Words Are the New Numbers: A Newsy Coincident Index of the Business Cycle", *Journal of Business & Economic Statistics*, 2019, forthcoming.

[106] Tsukioka, Y., J. Yanagi, and T. Takada, "Investor Sentiment Extracted from Internet Stock Message Boards and IPO Puzzles", *International Review of Economics & Finance*, 2018, 56, 205-217.

[107] Vapnik, V., *The Nature of Statistical Learning Theory*. New York: Springer, 1996.

[108] 王靖一、黄益平，"金融科技媒体情绪的刻画与对网贷市场的影响"，《经济学》（季刊），2018 年第 17 卷第 4 期，第 1623—1650 页。

[109] 王红建、李青原、刑斐，"经济政策不确定性、现金持有水平及其市场价值"，《金融研究》，2014 年第 9 期，第 53—68 页。

[110] 王克敏、王华杰、李栋栋、戴杏云，"年报文本信息复杂性与管理者自利——来自中国上市公司的证据"，《管理世界》，2018 年第 12 期，第 120—132 页。

[111] 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。

[112] Wang, J., Y. Shen, and Y. Huang, "How Does News Sentiment Affect the Peer-to-Peer Lending

Market in China?", Working Paper, 2018.

[113] 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”，《会计研究》，2015 年第 2 期，第 20—27 页。

[114] 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。

[115] 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。

[116] Yu, J., "Disagreement and Return Predictability of Stock Portfolios", *Journal of Financial Economics*, 2011, 99 (1), 162-183.

[117] 俞庆进、张兵，“投资者有限关注与股票收益——以百度指数作为关注度的一项实证研究”，《金融研究》，2012 年第 8 期，第 152—165 页。

[118] 曾庆生、周波、张程、陈信元，“年报语调与内部人交易：‘表里如一’还是‘口是心非’?”，《管理世界》，2018 年第 34 卷第 9 期，第 143—160 页。

[119] 张皓星、黄益平，“情绪、违约率与反向挤兑——来自某互金企业的证据”，《经济学》（季刊），2018 年第 17 卷第 4 期，第 1503—1524 页。

[120] Zhang, J. L., W. K. Härdle, C. Y. Chen, and E. Bommes, "Distillation of News Flow Into Analysis of Stock Reactions", *Journal of Business & Economic Statistics*, 2016, 34 (4), 547-563.

[121] 张谊浩、李元、苏中锋、张泽林，“网络搜索能预测股票市场吗?”，《金融研究》，2014 年第 2 期，第 193—206 页。

[122] 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”，《经济学动态》，2018 年第 6 期，第 129—140 页。

[123] 周开国、应千伟、陈晓娴，“媒体关注度、分析师关注度与盈余预测准确度”，《金融研究》，2014 年第 2 期，第 139—152 页。

[124] 周开国、应千伟、钟畅，“媒体监督能够起到外部治理的作用吗？——来自中国上市公司违规的证据”，《金融研究》，2016 年第 6 期，第 193—206 页。

[125] Zou, L., K. D. Cao, and Y. Wang, "Media Coverage and the Cross-Section of Stock Returns: The Chinese Evidence", *International Review of Finance*, 2019, forthcoming.

A Literature Review of Textual Analysis in Economic and Financial Research

YAN SHEN　YUN CHEN　ZHUO HUANG*

(*Peking University*)

Abstract　In this paper we conduct a literature review of textual analysis in economic and financial studies. Textual data exhibit the characteristics of diverse data source, rapidly grow-

* Corresponding Author: Yun Chen, National School of Development, Peking University, Haidian District, Beijing, 100871, China; Tel: 86-13264714992; E-mail: yunchen@pku.edu.cn.

ing data volume and high frequency. We summarize the procedures of extracting information from textual data and discuss the features of dictionary-based approach, machine learning and deep learning approaches. We then review the data sources, methods and empirical results of economic and financial studies using textual analysis. Finally, we summarize the new features of using textual analysis in empirical studies and point out future research directions in this field.

Key Words textual analysis, machine learning, investor sentiment

JEL Classification C43, G12, G14