金融学文本大数据挖掘方法与研究进展*

姚加权　张锟澎　罗平

摘要:在金融学领域的传统实证研究中,所用数据多局限于财务报表和股票市场数据等结构化数据。而在大数据时代,计算机技术的进步使得数据类型不断丰富,研究者开始将非结构化的文本大数据引入到金融学领域的研究中,其主要包括上市公司披露文本、财经媒体报道、社交网络文本、网络搜索指数以及 P2P 网络借贷文本等,并对文本的可读性、语气语调、相似性以及语义特征展开研究。本文首先介绍了金融学领域文本大数据挖掘步骤和方法,描述了语料获取、预处理过程、文档表示以及文档的特征抽取;然后根据不同的文本信息来源,梳理了金融学文本大数据的研究进展;最后对未来金融学文本大数据的研究方法和研究内容进行了展望。

关键词:文本大数据　文本分析　机器学习　深度学习　数据挖掘

一、引言

在金融学领域的传统实证研究文献中,研究数据多局限于财务报表数据、股票市场数据等结构化数据(structured data)。而在大数据时代,计算机技术的不断提高使得数据类型更加丰富,文本大数据已经成为计算机可以解读和分析的数据,并能够对非传统领域的经济现象展开研究(Loughran & McDonald,2016;Teoh,2018)。这种非结构化数据(unstructured data)在公司对外披露以及股票市场中所占的比重较大,传递形式和表达方式更为多样化,尤其是在中国这种"听话听音,听锣听声"的高语境传播环境中(林乐和谢德仁,2016),文本大数据在金融学领域中拥有较高的研究价值。文本分析(textual analysis)是指,运用特定的方法挖掘文本信息内容,从而对文本的可读性、情绪语调、语义特征以及相似性等文本特征进行分析。通过对上市公司披露文本、财经媒体报道、社交网络文本、网络搜索指数以及 P2P 网络借贷文本等文本大数据进行挖掘和分析,研究者能够从文本的披露行为(Loughran et al,2009;曾庆生等,2018)、文本的情绪和语调(Li,2010b;Loughran & McDonald,2011;Garcia,2013;Davis et al,2015;汪昌云和武佳薇,2015)以及文本信息的市场反应(Antweiler & Frank,2004;游家兴和吴静,2012)等方面展开研究,从而为金融学领域提供更丰富的研究内容和研究视角。

文本分析型研究有较长的历史,Jones & Shoemaker(1994)以及 Cole & Jones(2005)分别对会计文本内容以及管理层讨论与分析(management discussion and analysis,MD&A)的相关文献进行了综述。随后,Li(2010a)着重于计算机语言学、自然语言处理以及统计学的大样本文本分析,按照不同主题调查了企业披露文本的相关研究。再之后,Loughran & McDonald(2016)对国外会计和金融领域中的文本分析文献以及相关方法进行了调查和描述。Guo et al(2016)总结了机器学习方法

* 姚加权、张锟澎,暨南大学管理学院,邮政编码:510632,电子邮箱:jiaquanyao@gmail.com,zkpworking0729@163.com;罗平,中国科学院智能信息处理重点实验室、中国科学院计算所和中国科学院大学,邮政编码:100190,电子邮箱:luop@ict.ac.cn。本文受国家自然科学基金项目(71502152, U1811461),国家社科基金重大项目(18ZDA092)和国家重点研发计划课题(2017YFB1002104)的资助。感谢冯绪、梁平汉、唐国豪、王靖一、杨海生和赵轶星的宝贵意见。本文曾在 2019 海峡两岸暨港澳金融科技青年学者论坛、2019 中国数字金融研究联盟学术年会、第十八期香樟经济学 Seminar(广州)和 2019 中国"金融科技学"年会报告,感谢与会专家的评论,感谢审稿人的修改意见,文责自负。

在财务文本大数据分析中的应用。Gentzkow et al(2019)则描述了文本大数据的分析方法以及在经济学中的应用。Cong et al(2019)描述了金融市场中典型的英文文本来源,并讨论了神经网络模型与生成统计模型在文本分析领域的应用。在文本分析研究综述方面,国内研究者如唐国豪等(2016)整理了国内外基于文本情绪分析的行为金融研究进展,并总结了主要的文本分析方法。沈艳等(2018)综述了英文文本大数据分析在经济学和金融学领域中的应用,辅助以中文文本文献。张学勇和吴雨玲(2018)以国外文献为主,从网络新闻数据、搜索引擎数据、社交网络数据以及网络论坛数据四个方面,梳理了资产定价领域中运用网络大数据挖掘技术分析投资者心理和行为的研究内容。

综合来看,以往文献着重于介绍文本分析的主要方法,但缺乏对文本大数据挖掘步骤和方法的详细介绍。本文主要在以下方面拓展了以往的研究:首先,详细介绍了文本大数据挖掘步骤和方法,描述了文本的语料获取、预处理过程、文档表示以及文档的特征抽取。其次,介绍了国内外金融学文本大数据挖掘的主要文本信息来源,并根据不同的文本信息来源,梳理了金融学文本大数据的国内外研究进展,以便把握文本大数据目前在金融学领域中的研究方向和重点领域。最后,提出了未来的研究展望,期望有助于国内研究者进一步拓展文本大数据在金融学和经济学领域的应用。

二、金融学文本大数据挖掘步骤和方法

在金融学文本大数据研究中,早期的部分学者采用人工阅读的方式识别文本信息。但随着文本数量的增大,该方法不仅耗时耗力,且提取信息的精度也由于阅读者理解能力的差别而受到制约。因此,多数学者开始将计算机处理技术引入文本大数据的分析中,该过程主要包括:语料获取、文本的预处理、文档表示以及文档的特征抽取。然后,研究者再根据需要将抽取的文档特征应用到具体的相关性分析或因果分析中,具体如图1所示。

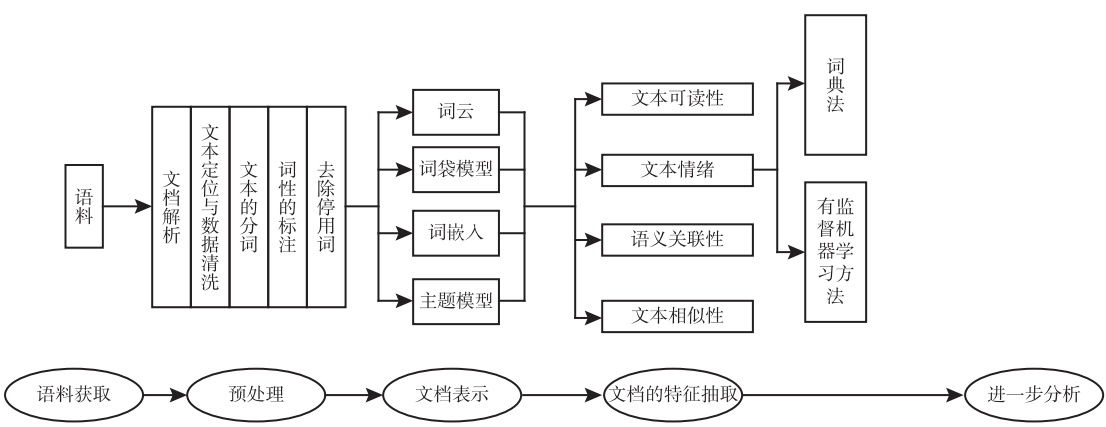

图1 文本大数据分析流程图

(一)语料获取

语料获取的方法主要包括:(1)手工收集。然而,该过程需要消耗大量的时间和人力成本。(2)网络抓取。由于文本量的提高以及文本大数据获取困难,大多数学者选择运用编程语言直接从网络中爬取文本大数据(Loughran & McDonald,2014;Blankespoor et al,2014)。该方法一方面能够及时地获取文本信息,另一方面还可以通过编程语言对文本格式和内容等进行整理,以便进行下一步分析。

(二)文本的预处理

在语料获取后,研究者需要对文本进行预处理,该过程主要包括文档解析、文本定位与数据清洗、文本的分词标注、词性的标注(part-of-speech tagging)以及停用词去除(stop words)五个步骤。

1. 文档解析。在信息披露监管制度下,企业需要以电子文档的形式定期或非定期地公开发布相关信息的文档。然而,这些文档仅仅实现了电子化存储,方便读者在电子设备上阅读,但这并不意

味着机器可以自动处理,实现“机器可阅读”。在计算机领域,信息披露的电子化文档被统称为富格式文档(richly formatted documents)。这些文档包含文本段落、表格、图表等多种内容模态,通常会组织为层次化的目录结构,并经过美化的排版和格式处理以呈现给读者阅读。从文档的格式来看,绝大多数金融市场要求的信息披露文档是 PDF 格式。因此,解析富格式文档经常是进行文本预处理的第一步,即获取里面的信息内容。

在进行 PDF 文档解析的过程中需要注意两个方面:一方面,PDF 文档的生成不是一个可逆的过程,当我们使用 Word 或 Excel 的编辑器将文档导出为 PDF 文档后,虽然文档的排版格式等视觉呈现可得以保证,但文档内部的结构信息被部分或完全地丢失了。另一方面,解析后的文档是文本分析的基础,不精确的 PDF 解析可能会给后续的文本分析带来严重的影响。因此,针对金融学文本大数据,需要慎重选择文档结构的解析工具。①

2. *文本定位与数据清洗*。一方面,研究者需要运用计算机程序对文本信息进行定位。例如,MD&A 部分是较多学者的研究对象,研究人员可以运用正则表达式来定位财务报告正文中 MD&A 部分的开头和结尾,进而将该部分内容提取出来。另一方面,研究者还需要对文本中被视为噪音的内容进行清洗和删除(Jiang et al,2019),主要包括广告、超文本标记语言(HTML)、直译式脚本语言(JavaScript)等代码以及图片等。

3. *文本的分词*。在英文文本中,单词被空格分开就自动完成了分词。另外,还可以通过词形还原(lemmatization)和词干提取(stemming)对单词进一步地划分。但中文文字之间没有空格切分,而且词语才是能够独立运用的最小语言单位。因此,研究者需要对中文文本进行专门的分词处理。目前,大多数学者采用了 Python 开源“jieba”中文分词模块来对企业财务报告、年度业绩说明会以及股票论坛帖子进行分词。中文文本分词存在三个难点,即切分颗粒度、歧义词的识别和新词的识别。切分颗粒度太小,容易破坏词语的意思。例如,容易将“机器学习”切分成“机器”和“学习”。针对歧义词,则应该选择合适的分词模式。例如,在使用“jieba”分词时,为了提高分词的精确度,应当选择精准分词模式。针对新词(如公司名称、产品名以及关键人物姓名),用户可以自定义词典以便帮助分词软件对新词进行识别。

4. *词性的标注*。词性是识别语义信息的重要语法特征,例如名词、动词、连接词等,词性标注就是对切分后词语的词性做标记。通过词性标注,计算机能够识别词语的种类、消除词语歧义,进而能识别语法结构,降低计算机语义分析的难度。中英文在词性标注方面具有较大差异,英文单词在词性划分方面较为严谨,能通过词尾变换来揭示词性的变化,例如“-ing”、“-ness”和“-ment”等均对确认词性给予了具体的提示。但是,中文词语未对词性做出明确规范,主要靠语法和语义来识别词性,即“英语重形合、汉语重意合”。

5. *停用词去除*。为了提高文本挖掘信息的精度,还需要对文本中的停用词进行剔除。停用词是指对句子语法结构很重要但本身传达意义较少的词语,其增加了文本数据的维度,提高了文本分析的成本。在英文文本中,停用词主要包括冠词(the,a)、连词(and,or)以及动词“to be”等(Gentzkow et al,2019)。但在中文文本中,应当根据中文的语言习惯确定停用词,除了标点符号和特殊符号以外,还包括表示逻辑关系的连接词(和、然而等)以及俚语等。另外,停用词还需要根据研究的内容来决定。例如,当研究文本情感时,保留语气词以及特定的标点符号,均有利于衡量文本的情感程度。

(三)文档表示

文本数据属于稀疏的高维度数据,计算机处理存在困难,因此对文本数据进行预处理后,还需要将文档中的信息以特定的方式表示出来,从而便于研究人员或者计算机进行下一步的分析。文档表示方法主要包括:词云(word cloud)、词袋模型(bag of words,BOW)、词嵌入(word embedding)和主题模型(topic model)。

①我们详细举例说明了选择精准 PDF 解析工具的重要性,详见本文的在线附录,有兴趣的读者可以联系作者获取。

1. 词云。词云是一种文本大数据的可视化技术。文本可视化是指将文本中比较复杂的内容和规律转化为视觉符号表达出来，进而能够使人们利用与生俱来的视觉感知快速获取文本中所蕴含的关键信息。词云技术能够描述词语在文本中出现的频率，当词语出现频率较高时，会以较大且醒目的形式呈现。

2. 词袋模型。词袋模型是一种建立在文字词组语序不重要的假设之上，将文本看作是若干个词语的集合，只计算每个词语出现次数的一种文本向量化的表示方法。该模型主要包括：独热表示法(one-hot representation)以及词频－逆文档频率法(term frequency-inverse document frequency，TF-IDF)。独热表示法操作简单。假设有两个文档“文本大数据在经济学中的应用”和“文本大数据在金融学中的应用”，基于这两个文本文档可以构建如下词表：[“文本”，“大数据”，“在”，“经济学”，“金融学”，“中”，“的”，“应用”]，按照该顺序进行词袋化后得到两个文档的词袋向量分别为：[1，1，1，1，0，1，1，1]和[1，1，1，0，1，1，1，1]，其中，“1”和“0”分别表示文档中有无出现这个词。然而，文档中并非每个词语均有相同的机会出现，大多数文本中只有极少数词语被经常使用，绝大多数词语很少被使用。因此，需要对每一个词语赋予其权重来更好地表示每个词语在文档中的作用。Loughran & McDonald(2011)运用 TF-IDF 方法计算了特定词语在文档中的权重。其基本公式如下：

$$idf_i = log\frac{N}{df_i} \tag{1}$$

$$tf-idf_{i,j} = \begin{cases} \frac{(1+log(tf_{i,j}))}{(1+log(a_j))}log\frac{N}{df_i} & 若\ tf_{i,j} \geqslant 1 \\ 0 & 其他 \end{cases} \tag{2}$$

式(1)中 df_i 定义为包含词语 i 的文档数量，N 表示文档集合中的文档总数，idf_i 为逆文档频率。式(2)中 $tf_{i,j}$ 为第 j 个文档中词语 i 出现的总次数，a_j 为第 j 个文档中包含的词语数，$tf-idf_{i,j}$ 则为第 j 个文档中词语 i 的权重。但是，词袋模型存在以下问题：第一，忽略了文档中词语顺序和词语之间的语义关系，可能会产生歧义；第二，向量的维度取决于文档中词语的数量，当文档中词语数量过多时，很可能产生维度灾难。

3. 词嵌入。词嵌入是将维数为所有词的数量的高维空间嵌入到低维连续向量空间中的技术。通过词嵌入技术，可以将词语映射成低维连续向量空间中的向量，通过向量之间的距离和位置来表示文档中词语的上下文、语法和语义的相似性以及与其他词语的关系。在金融学文本分析中，Word2Vec 技术是常用的词嵌入技术，包括 CBOW(continuous bag of words)和 Skip-Gram 神经网络模型，可以通过训练使神经网络捕捉到更多词语之间的上下文信息，从而将每个词语映射成更低维度、稠密且包含更多语义信息的向量(Mikolov et al，2013)。在 Word2Vec 技术中，词嵌入向量能得到不同词之间的类比关系，最经典的例子就是“*king-queen*＝*man-woman*”，如图 2 所示。

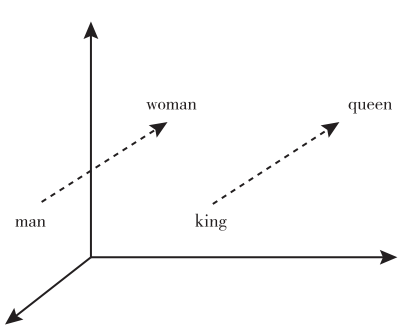

图 2　词嵌入模型实例图

4. 主题模型。最常用的主题模型是 LDA(latent dirichlet allocation)模型(Blei et al，2003)。LDA 模型是在大规模语料集中提取主题信息的无监督机器学习方法，它假设文档生成包括两个步

骤:第一步,假定每个文档均有对应的主题分布,在文档的主题分布中抽取一个主题;第二步,假定每个主题都有对应的词语分布,从上一步抽取的主题所对应的词语分布中抽取一个词语。通过将这两步迭代拟合到文档中的每个词语,即可得出每个文档的主题分布和每个主题的词语分布。

Huang et al(2018)指出 LDA 模型具有以下优势:首先,该模型克服了手动编码的局限性,能够对大量文本文档进行分类;其次,LDA 模型能够提供可靠且具有可复制性的文本主题分类,排除了人工文本分类的主观性;最后,LDA 模型不需要研究者为划分类别预先指定相应的规则和关键词。然而,该模型的局限性在于预设主题个数的方式中加入了人的主观因素,这会影响主题个数的选择,进而影响主题的生成和文本的主题归类①。

(四)文档的特征抽取

目前,在金融学领域中针对文档的特征抽取方面主要包括四个方面:文本可读性(textual readability)、文本情绪(textual sentiment)、语义关联性(textual relatedness)以及文本相似性(textual similarity)。

1. 文本可读性。文本的可读性反映了读者理解文本信息的难易程度,文本可读性较低时,投资者会难以理解文本编辑者所传达的信息,进而会影响到投资者的投资行为。Li(2008)将迷雾指数(fog)应用到了文本分析中,指出迷雾指数越小时,年报的可读性越强。另外,学者们还利用了年报中的字数(You & Zhang,2009)和年报电子文档的大小(Loughran & McDonald,2014)来衡量年报的可读性。以往多数研究采用迷雾指数来衡量文本可读性(Li,2010b;Lehavy et al,2011),但这种方式仍具有一些问题。例如,如果将文本中每个句子的词语随机排序,那么文章将完全无法理解,但是原来的文本与随机排序后的文本所计算的迷雾指数完全相同(Jones & Shoemaker,1994)。此外,Loughran & McDonald(2014)研究指出,迷雾指数在测量商业文本可读性时具有局限性。正如Loughran & McDonald(2016)提出,当衡量企业信息披露可读性时,将公司复杂性和年报可读性分开是困难的。当公司具有多种业务时,内部业务结构复杂的公司很可能会因为业务复杂性而披露难以阅读和理解的年报。因此,在衡量公司年报文本可读性时,应当考虑将企业的业务复杂性因素剔除。

2. 文本情绪。目前,文本情绪的提取方法主要包括词典法(dictionary-based approach)和有监督机器学习方法(supervised machine learning)。

词典法是指运用情绪词典来研究文本情绪或者语气语调的方法。一旦确定了词典,就可以使相关研究具有可复制性。另外,在构建词典时,金融领域的知识也尤为重要,只有将金融学知识应用于词典的构建过程中才能使词典更适用于金融文本的分析。针对英文文本大数据,国外已经形成了多部具有影响力的词典,例如 Henry 词典(Henry Word Lists,2008)、LM 词典(Loughran and McDonald Word Lists,2011)、哈佛大学通用调查词典(Harvard General Inquirer Word Lists, GI)、文辞乐观与悲观词典(Diction Optimism and Pessimism Word Lists)等。结合以上词典,已有文献能够对媒体报道情绪(Tetlock,2007;Tetlock et al,2008;Solomon et al,2014)、电话会议文本语气语调(Price et al,2012)以及财务报告的语气语调(Feldman et al,2010;Merkley,2014)进行分析。针对中文文本大数据,大多数学者在参考英文词典及其他词库的基础上构建自己的词典展开研究(汪昌云和武佳薇,2015;曾庆生等,2018)。另外,姚加权等(2019)通过词典重组和深度学习算法构建了针对金融领域正式文本和非正式文本的中文情绪词典。该文指出基于该词典构建的上市公司年报语调指标和社交媒体情绪指标能够有效地预测上市公司股票的收益率、成交量等市场因素以及股价崩盘风险。

Jegadeesh & Wu(2013)指出,在词典法中选择合适的加权方法至少跟选择准确的词典一样重

①为了避免该局限性,学者们研究了一系列方法,例如复杂度得分(perplexity score)(Blei et al,2003)和层次狄利克雷过程(hierarchical Dirichlet process)(Teh et al,2006)。

要。在金融学领域中，多数学者采用了简单比例加总权重法衡量文本情绪，具体公式为：

$$Pos = \text{积极、正面词的个数} / \text{文本总词数}, \tag{3}$$

$$Neg = \text{消极、负面词的个数} / \text{文本总词数} \tag{4}$$

$$Tone = \frac{(Pos - Neg)}{(Pos + Neg)} \tag{5}$$

公式(5)中 $Tone$ 为管理层净正面语调指标，$-1 \leqslant Tone \leqslant 1$，当 Pos 大于 Neg 时，$Tone$ 越大，从而说明管理层语调更加正面。此外，林乐和谢德仁(2016)在稳健性检验中进一步采用了 TF-IDF 方法来衡量词语权重。然而，词典法的局限性在于，一方面，构建针对特定文本的词典时，需要相关领域的专业知识，这样就导致构建出来的特定词典无法简单地应用于其他文本。另一方面，词典法仅关注特定的关键词，从而会忽略文档的上下文关系。

除了词典法，学者们还运用有监督机器学习方法对文本情绪进行了分类。其中，有监督机器学习方法是指将有标签的数据集分为训练集和测试集，利用训练集来训练模型，然后将训练的模型应用到测试集中，使用测试集的预测结果来评估模型。在金融学文本情绪分类中，学者常用的有监督机器学习方法为朴素贝叶斯和支持向量机。

朴素贝叶斯(naïve Bayesian)是一种基于贝叶斯理论的有监督机器学习算法。首先，输入训练集学习文档词语归类关系，得到文档归类的先验概率以及条件概率分布。其次，根据贝叶斯条件概率公式计算已知文档属于不同文档类别的条件概率。最后，基于最大后验假设把该文档归为具有最大后验概率的一类。Loughran & McDonald(2016)指出了朴素贝叶斯方法的三个优点：首先，该方法是文本分析中最古老且成熟的方法之一；其次，该方法基于机器学习来阅读文本，使得研究大量文本信息成为可能；最后，文本度量规则的确定排除了研究人员的主观性。然而，该方法建立在词语集合中词语属性相互独立的假设之上，但实际上该假设很难成立。在金融学领域，朴素贝叶斯方法已经广泛地运用到了文本分析中(Antweiler & Frank，2004；Das & Chen，2007；Li，2010b；Jegadeesh & Wu，2013)。

支持向量机(support vector machine，SVM)是一种基于统计学习理论和结构风险最小原理的有监督机器学习算法。其基本思想为，将语料库中的文档通过核函数映射为高维度特征空间中的一个样本点，然后根据训练集，在特征空间中找到最优分类的超平面，使得它能尽可能多地将两类数据点正确分开，同时使分开的两类数据点距离超平面最远。Guo et al(2016)指出，与朴素贝叶斯方法相比，支持向量机能够实现更好的样本外预测精度。朴素贝叶斯方法在学习中涵盖了所有文档，引入了较多噪音，而该方法只关注不同类型惩罚函数选择的支持向量，从而能够避免过度拟合的问题。而且，该方法可利用核函数解决线性不可分问题。该方法的弊端在于：分类结果对核函数的选择比较敏感，难以针对具体问题选择出最佳的核函数。Antweiler & Frank(2004)运用该方法对互联网股票留言板中的文本情绪进行了分类。

需要指出的是，在运用有监督机器学习方法时，检验和评价模型的分类效果至关重要。其主要方法为交叉验证法，该方法的基本思想为将原始数据切分为较小子集，并将其随机重新组合为训练集和测试集，在此基础上对模型反复进行训练、测试，并根据结果对模型进行选择。此外，学者还可以通过一系列指标对模型的分类水平进行评估，例如正确率、精度、召回率、F1 值等指标。

Henry & Leone(2016)指出，与词典法相比，有监督机器学习方法的局限性在于该方法需要人工编码提供训练集，这比用词典法进行词频测量所需的工作量更大。此外，在对训练集进行人工编码的过程中，编码结果容易受到研究人员的主观行为影响，进而影响机器学习的分类效果，并且研究的可复制性较差。尽管存在以上缺陷，但对没有预定词典的文本来说，有监督机器学习方法仍是一种合理且有效的分类工具，特别是在分类精确度比可复制性更重要的情况下。

3. 语义关联性。语义关联性就是根据某一类词语去识别文本语义特征的过程。具体而言，首

先依照某一类关键词构建词表,然后计算词表中词语在文档中的词频,进而识别出文本中与关键词语义相关的语义特征。例如,Loughran et al(2009)根据"corporate responsibility""social responsibility""socially responsible"等关键词识别了公司对道德相关术语的使用情况。此外,学者还可以运用词嵌入技术,根据空间中词向量之间的距离(即语法和语义的相似性)来处理词语语义关联性问题。例如,Li et al(2020)运用 Word2Vec 技术扩展了不同类型企业文化的关键词。

4. 文本相似性。目前,较多学者运用了余弦相似度指标来衡量财务报告的相似性(Hoberg & Phillips,2010,2016;王雄元等,2018)和专利文本的相似程度(Kelly et al,2018)。假设文本 d_1 和 d_2 对应的文本向量分别为 $a=(w_{a1},w_{a2},\cdots,w_{an})$,$b=(w_{b1},w_{b2},\cdots,w_{bn})$,则文本 d_1 和 d_2 的余弦相似度计算公式如下:

$$cosine\ similarity(d_1,d_2)=\frac{a\times b}{\|a\|\times\|b\|}=\frac{\sum_1^n w_{ai}\times w_{bi}}{\sqrt{\sum_1^n w_{ai}^2}\times\sqrt{\sum_1^n w_{bi}^2}} \tag{6}$$

其中,n 为特征个数,w_{ai},w_{bi} 为特征 i 在两个文本中的权重。该公式取值在 0 和 1 之间,数值越大表明文档相似度越大。

通过总结金融学文本大数据挖掘步骤和方法可以发现,中英文文本大数据分析的主要区别在于文本的预处理过程。相对于英文文本大数据,中文文本大数据的预处理过程更加复杂。在对中文文本大数据进行文本特征抽取前,研究者需要进行分词处理。而且,中文词语未对词性做出明确规范,主要依靠文本的语法和语义来识别词性。在文本可读性方面,尽管多数学者参考迷雾指数构建中文文本的可读性指标,但是他们运用的常用词或复杂词的词典不同,从而导致构建的可读性指标也不尽相同。

三、金融学文本大数据分析

本文接下来将根据不同文本信息来源介绍金融学文本大数据的研究内容,国外研究的文本类型主要包括:上市公司披露的文本信息(如财务报告、电话会议、招股说明书等),财经媒体报道,社交网络文本以及搜索指数等。针对以上文本来源,国外文献运用文本挖掘方法抽取文本特征(例如可读性、语气语调、相似性和语义特征等),并探讨了这些文本特征与企业表现、股票市场行为之间的关系。

(一)对上市公司披露文本信息的研究

上市公司披露的文本信息能够反映企业披露行为,也可用于衡量企业财务和经营状况以及向股票市场传达出公司管理层对未来发展的信心程度。其中,针对财务报告、电话会议文本以及招股说明书的研究比较广泛,研究重点在于文本的可读性、语气语调以及相似性。

1. 上市公司披露文本的可读性研究。可读性较高的公司披露文本能够更好地向投资者传达公司信息。Li(2008)运用迷雾指数衡量了财务报告的可读性,发现财务报告可读性较高的企业利润更加持久。Lehavy et al(2011)指出提高企业财务报告的可读性能够降低分析师盈余预测的离散度,提高分析师盈余预测的准确度。Guay et al(2016)发现,财务报告可读性较差的公司会通过自愿披露的方式来减轻可读性的负面影响。Lo et al(2017)发现管理者会策略性地操控财务报告可读性,以误导或影响投资者对企业业绩的评价。Bushee et al(2018)将电话会议文本语言的复杂性区分为信息成分和混淆成分,发现信息成分能够降低信息不对称程度,而混淆成分会加剧信息不对称。

2. 上市公司披露文本的语气语调研究。在语气语调的研究方面,学者们认为公司披露文本的语气语调能够用于预测企业表现和股票市场的变化。Mayew et al(2015)研究发现财务报告中 MD&A 部分的语气具有增量信息,能够预测企业的破产概率和持续经营能力。Davis et al(2015)指出,管理者的乐观情绪会影响电话会议中管理者的语气。Allee & Deangelis(2015)考察了电话会议

中语气分散程度，发现管理者语气分散程度与企业未来业绩和管理者的决策相关，而且语气分散程度还会影响分析师和投资者对信息的感知。Bochkay et al(2020)开发了一个极端语气词典，研究发现，管理者在电话会议中使用极端词汇后，企业的股票交易量会显著增加，股价反应会更加强烈。Jiang et al(2019)基于公司财务报告以及电话会议文本构建了经理情绪指标，指出该经理情绪指标能够有效预测股票收益，且该指标的预测能力超过常用的宏观经济变量及投资者情绪。

3. 上市公司披露文本的相似性研究。相似性也是公司披露文本信息的重要特征，一方面，基于企业之间财务报告内容的相似性可以研究不同企业之间的关系。例如，Hoberg & Phillip(2010)发现两家公司财务报告中产品描述部分越相似，那么两者之间发生并购的可能性越高，且并购绩效越好。Hoberg & Phillips(2016)根据不同公司财务报告产品描述部分的相似性构建了公司特有的产品市场竞争对手集合，并基于此形成了时变的行业划分标准。另一方面，企业之间和企业不同时期财务报告的相似性也为研究企业披露行为模板化提供了契机(Brown & Tucker，2011；Lang & Stice-Lawrence，2015)。

4. 上市公司披露文本的语义特征研究。还有一系列文献针对公司披露文本的语义特征展开研究。例如，Buehlmaier & Whited(2018)基于MD&A部分运用朴素贝叶斯方法构建了融资约束指标，发现受融资约束的公司的股票收益较高。Bochkay et al(2019)根据电话会议研究了企业CEO信息披露的风格，指出CEO的前瞻性信息披露行为及乐观态度会在其任期内下降，外部雇用和经验不足的CEO更倾向于披露前瞻性信息，而年轻的CEO在信息披露方面表现出更大程度的乐观态度。Hanley & Hoberg(2019)将LDA模型和Word2Vec技术相结合，从银行年报中提取与风险相关的语义主题，并结合投资者的交易模式研究发现，金融行业新显露的风险信号有助于监管金融市场的稳定性。

(二)对财经媒体报道的研究

财经媒体报道为研究股票市场提供了丰富的文本大数据，常见的英文媒体报道有《华尔街日报》、《纽约时报》以及《金融时报》等。

1. 媒体关注和媒体情绪研究。部分文献根据财经媒体发布新闻的数量及正负面新闻研究媒体报道对股票市场的影响。Hillert et al(2014)基于美国45家报纸在1989－2010年间发表的约220万篇文章进行研究发现，媒体报道会加剧投资者的偏见，收益的可预测性在媒体关注度高的公司中更强。Baloria & Heese(2018)指出受到FNC(Fox News Channel)媒体倾向性报道威胁的民主党关联企业在大选前会隐瞒负面消息，而在大选后释放负面消息，证明企业会为了名誉资本而避免负面媒体报道。Frank & Sanati(2018)研究发现，股票市场存在对好消息反应过度、对坏消息反应不足的现象，正面新闻冲击后股价会出现反转，负面新闻冲击会引起股价漂移。此外，还有文献研究了媒体情绪对未来房价的预测作用(Soo，2018)。

2. 经济政策不确定性研究。媒体报道中还包含着经济政策的信息，Baker et al(2016)根据多个主要经济体中具有代表性的媒体报道，运用文本挖掘技术构建了经济政策不确定性指数(economic policy uncertainty，EPU)，该指数能够连续且定量地描述经济政策的不确定性。随后，Gulen & Ion(2016)运用该指数研究了经济政策不确定性对企业投资的影响，发现宏观经济政策不确定性会影响微观企业的财务决策，抑制企业投资。Bonaime et al(2018)基于该指数研究了经济政策不确定性对企业并购的影响，发现经济政策不确定性的上升会减少并购交易的价值和数量。

3. 媒体偏向、谣言和假新闻研究。财经媒体报道还存在本地偏向现象。Gurun & Butler(2012)指出媒体在报道本地公司新闻时使用更少的否定词，出现该现象的原因在于本地公司投入了更多广告支出。他们还发现，异常的本地媒体偏向与企业股票估值密切相关。另外，市场上还存在公司谣言和虚假新闻。Ahern & Sosyura(2015)通过对公司并购的谣言进行研究发现，媒体更偏向于发布具有新闻价值公司的谣言，媒体发布的谣言会争夺投资者有限的的注意力，从而导致股票价格的过度反应和随后逆转。Kogan et al(2019)发现虚假新闻会提高所涉及股票的交易量和价格波动性，当虚假新闻曝光后，假新闻所在平台上所有新闻对股票交易量和价格波动性的影

响会降低。

(三)对社交网络文本的研究

随着社交网络的兴起,学者开始将微博和股票论坛等社交网络纳入金融学领域的研究范围内,社交网络中的文本信息与股票市场的关系是该领域的研究重点。

1. 社交网络文本情绪研究。早期,Antweiler & Frank(2004)以雅虎财经网络论坛中的帖子为研究对象,发现以帖子数量衡量的关注度指标能够有效地预测股票收益率和市场波动情况,帖子情绪分歧与同期股票交易量正相关。Chen et al(2014)以股票论坛 Seeking Alpha 上的文章和评论为对象进行研究发现,文章和评论的语气能够预测企业未来的股票收益。Cookson & Niessner(2020)通过对美国股票论坛 StockTwits 上的帖子和用户信息进行研究发现,大约一半的投资者分歧是由不同投资理念造成的,而且投资者分歧可以有效预测股票市场中的异常交易量。

2. 策略性信息披露研究。公司也可以通过社交网络平台策略性地发布信息。例如,Blankespoor et al(2014)发现公司在 Twitter 上发布新闻的链接能够降低公司股票的买卖差价,提高交易深度。研究也表明,社交网络作为信息传播的途径,能够优化投资者获取信息的能力,降低投资者信息搜寻成本。Jung et al(2018)分析了标普 1500 家公司在 Twitter 上的信息发布情况,研究发现,当存在坏消息时公司会减少 Twitter 的发文量,对于投资者成熟程度较低的公司和社交媒体受众面较广的公司而言,这种策略性信息披露行为更明显。

(四)对搜索指数的研究

在互联网技术不断发展的背景下,网络搜索指数是衡量投资者对股票关注程度的有效指标。Da et al(2011)获取了单个股票每周的谷歌搜索指数,利用股票的搜索频率直接测量了投资者关注度。研究指出,运用搜索指数能够更及时地度量投资者关注度,搜索指数的增加能够预测未来两周股价的上涨以及一年内的股价反转。Chi & Shanthikumar(2017)通过搜索位置检验了投资者对不同地区股票的关注情况,发现投资者存在“本地偏见”,即投资者更倾向于关注本地的股票,这种“本地偏见”还会影响市场对盈余公告的反应。此外,已有文献还研究了特定关键词的搜索指数与股票市场的关系。例如,Da et al(2015)通过确定 118 个词的搜索量与同期市场收益之间的历史关系,选择 30 个最负面的词语作为特定搜索关键词,并使用其构建 FEARS 指数,发现该指数预测了股票市场短期收益逆转、波动性的暂时性增加以及共同基金从股票基金流入债券基金。

(五)对其他英文文本的研究

还有学者针对其他文本进行了研究。例如,De Franco et al(2015)运用 2002—2009 年间分析师报告研究发现,可读性较高的分析师报告能够提高股票交易量。Hwang & Kim(2017)分析了封闭式投资公司对外披露的年度股东报告的可读性对公司价值的影响,研究发现,在相对不透明的信息环境中,当公司发布可读性较差的年度股东报告时,投资者会产生怀疑等负面情绪,导致公司发生折价交易,从而降低了公司价值。Green et al(2019)运用 Glassdoor 上员工对雇主的评级信息进行研究发现,员工对雇主的评价与企业的销售增长和盈利能力相关,并能够预测一个季度后的未预期盈余。Huang(2018)发现亚马逊(Amazon. com)上的消费者产品评论可以预测公司股票收益。Chen et al(2019)基于美国专利文本数据研究了金融科技创新对金融行业的价值影响。Ryans(2019)运用朴素贝叶斯文本分类方法,结合公司未来发生财务重述和资产减值的情况,将问询函划分为重述问询函和非重述问询函、减值问询函和非减值问询函,研究了问询函与财务报告质量的相关性。Bandiera et al(2020)以 CEO 日志为研究文本,通过机器学习算法将 CEO 划分为“领导者”和“管理者”,研究了 CEO 行为与企业绩效的关系。

四、金融学文本大数据的国内研究进展

国内研究的文本信息来源主要包括:上市公司披露的文本信息(如财务报告、业绩说明会文本、招股说明书)、财经媒体报道、社交网络文本、网络搜索指数以及 P2P 网络借贷文本。国内文献通过

提取各种信息的文本特征对企业表现、股票市场以及网络借贷市场展开研究。与国外相比，随着国内 P2P 网络借贷的兴起，该类文本成为国内文本分析的研究重点之一。

1. 对上市公司披露文本信息的研究。针对上市公司年报，丘心颖等(2016)研究了上市公司年报可读性与分析师信息解读之间的关系，发现年报可读性与分析师预测质量不存在显著关系，说明中国分析师未能有效发挥专业解读信息的作用。曾庆生等(2018)研究了年报语调与年报披露后内部人交易行为之间的关系，发现企业管理者在编制年报时存在“口是心非”的现象，积极的年报语调却伴随着管理者较高的股票卖出规模。王雄元等(2018)基于年报 MD&A 中与风险相关的信息进行研究发现，前后两年年报中风险段落内容的相似性能够降低审计费用。针对公司披露的其他文本，林乐和谢德仁(2016)指出投资者能够识别上市公司年度业绩说明会中的管理层语调。Yan et al(2019)发现 IPO 招股说明书中不确定性或负面语调与 IPO 初期股票收益及后续股票收益的波动显著相关，而且降低了股票的长期回报率。

2. 对财经媒体报道的研究。证监会规定上市公司必须在《上海证券报》《中国证券报》《证券时报》《金融时报》《经济日报》《中国改革报》《中国日报》《证券市场周刊》“七报一刊”中公布企业重大信息。另外，我国还拥有百度新闻、新浪财经、和讯网等网络新闻媒体，这均为研究中国股票市场提供了丰富的媒体报道信息。游家兴和吴静(2012)以财经报纸为研究文本进行研究发现，媒体情绪越高涨或越低落时，资产定价偏误的现象越严重。汪昌云和武佳薇(2015)以财经媒体报道为研究对象进行研究发现，媒体负面语气的下降会提高 IPO 抑价率、IPO 超募资金比例和承销商费用占比。王靖一和黄益平(2018)还研究了网络媒体情绪对网贷市场的影响。此外，媒体报道可以划分为市场导向媒体报道和政策导向媒体报道。Piotroski et al(2017)指出中国媒体的集团化改革使得政策导向的媒体报道更加集中于政治目标，市场导向的媒体报道更加集中于商业目标。You et al(2018)从信息监督的角度研究发现，与政策导向媒体报道相比，市场导向的媒体报道能够提供更多关于企业方面的信息，且只有市场导向的媒体报道能够对公司治理产生重大影响。此外，在衡量宏观经济政策不确定性方面，Huang & Luk(2020)利用中国大陆多家报纸构建了新的并且频率更高的中国 EPU 指数，研究发现，新的中国 EPU 指数能够预测中国的股票价格、就业和产出情况。

3. 对社交网络文本的研究。何贤杰等(2016)基于新浪微博博文研究指出，公司治理水平越高，公司越倾向于开通微博并且发布更多的公司信息。东方财富股吧及雪球网等股票论坛为研究我国股票市场中投资者关注和投资者情绪提供了机会。Huang et al(2016)基于东方财富股吧发帖信息发现，中国投资者也存在“本地偏见”现象，这种偏见在欠发达地区、大型公司、非沪深 300 指数、低成交量且名称表明公司所在地的股票中尤其明显。孙书娜和孙谦(2018)研究发现，根据“雪球网”用户自选股信息构建的投资者关注在短期内会提高股票价格和股票交易量，但该影响会随着时间的推移而逐渐衰减。另外，Jiang, Liu & Yang(2019)表明股票论坛中投资者之间的交流也会对股票收益产生影响。

4. 对搜索指数的研究。在对搜索指数的研究方面，部分学者基于网络搜索指数构建了投资者个股关注指标，研究了投资者关注与资产定价的关系。例如，俞庆进和张兵(2012)以百度指数作为投资者关注度指标进行研究发现，投资者关注对股票当期收益有正向价格压力，但这种压力会在短期内实现反转。还有学者运用特定关键词的搜索指数展开研究。例如，曾建光(2015)根据“余额宝被盗”的百度搜索指数构建了投资者网络安全风险感知指标，研究发现，投资者对互联网安全风险感知越强烈，要求的风险补偿越高，并且移动互联网投资者风险感知要强于电脑端投资者。

5. 对 P2P 网络借贷文本的研究。在我国金融改革和金融创新的背景下，P2P 网络借贷(peer to peer lending)掀起了新的热潮，一些学者对 P2P 网络借贷成功率的影响因素展开研究。陈霄等(2018)发现可读性较强的借款描述能够向投资者传递积极信息，提高借款的成功率。彭红枫和林川(2018)分析了借款描述中特定词汇比重对借贷成功率的影响，研究发现，积极语气词汇和金融词汇

比重与借款成功率正相关,消极语气词汇比重、强语气词汇比重和弱语气词汇比重与借款成功率负相关。此外,还有不少学者基于 P2P 网络借贷文本对借款利率、筹资效率等方面展开研究。

除了对以上五个文本数据来源进行分析以外,还有学者针对分析师报告、年报问询函、私人会议总结报告等其他中文文本进行了广泛研究。①

五、金融学文本大数据研究展望

本文总结并介绍了金融学文本大数据挖掘步骤和方法,描述了文本的语料获取、预处理过程,文档表示以及文档特征的抽取。另外,本文根据不同的文本信息来源,梳理了国内外金融学文本大数据的研究内容。从已有文献来看,对文本大数据的挖掘与分析正处于蓬勃发展阶段。本文认为未来针对金融学文本大数据挖掘与分析还可以从以下几个方面进一步深入探讨。

1. *丰富研究内容,开拓更多文本数据来源*。文本大数据在金融学领域中的研究内容和信息来源还可以进一步细化和丰富。例如在财经媒体报道方面,研究者可以不仅对媒体报道的数量和情绪进行分析,还可以对报道事件类型进行判断,从而识别出有关企业并购、IPO、财务舞弊、高管个人新闻等方面的媒体报道,研究不同事件类型的报道对企业和股票市场的影响。在股票论坛方面,研究者可以根据投资者对企业的关注度来构建企业网络,研究不同企业之间的竞争关系。从数据来源方面,还可以开拓更多的文本大数据。例如微信公众号、政府工作报告、国务院政策文件、法院裁判文书、招聘网站、企业发布的业绩修正公告、社会责任鉴证意见、内部控制评价报告等。

2. *运用新的文本信息提取方法*。目前,在金融学领域的文本分析研究中,应用较为广泛的仍是无法反映上下文含义的“词袋”方法。然而,在自然语言处理领域(natural language processing,NLP)仍有许多新的分析方法和工具,它们在金融学领域的文本分析中尚未得到足够的重视,但具有较大的潜力。例如:(1)命名实体识别(named entity recognition,NER)。NER 属于 NLP 领域的重要基础工具,能够识别待处理文本中的命名实体,从而提取出时间、地点、人名、机构、货币、百分比和日期,常用的 NER 工具有 Stanford NER。(2)关系提取(relation extraction)。通常运用有监督的机器学习方法从含有实体对的句子中提取出实体对之间的对应关系,并对它们的共现性进行分析。(3)文本摘要(summarization),即使用计算机算法压缩文本内容的过程,摘要的长度取决于压缩率。Cardinaels et al(2018)研究指出,基于算法形成的摘要比管理层披露摘要的积极程度更低,且算法摘要能够使投资者对企业股价做出较为保守的估计。

3. *将深度学习引入文本领域的学术研究*。深度学习方法在 NLP 领域得到迅猛发展。深度学习模型主要包括:卷积神经网络(convolutional neural networks,CNN)模型、循环神经网络(recurrent neural networks,RNN)模型以及其变体长短期记忆网络(long short term memory networks,LSTM)模型,生成对抗网络(GAN),强化学习,以及目前在 NLP 领域流行的 BERT、XLNet 等模型。将深度学习引入文本领域将会展开更丰富的研究内容,同时还会提高文本信息提取的准确性。例如,Cao et al(2018)利用 LSTM 模型的变种在企业披露文本中检查出了不一致错误。

深度学习与传统机器学习方法在特征表示及模型参数数量等方面具有较大差异,如表 1 所示。Heaton et al(2016)指出深度学习方法在金融领域中的研究具有以下优势:第一,模型考虑了与预测问题相关的尽可能多的数据信息;第二,能够捕获输入数据间的非线性关系,提高样本内的拟合程度;第三,能够有效避免浅层结构的过度拟合问题。此外,当训练集数据量提高到一定程度时,深度学习信息提取准确度要明显高于传统机器学习方法。

---

①我们详细梳理了近年来关于金融学文本大数据分析的大约 170 篇国内期刊论文,详见本文的在线附录,有兴趣的读者可以联系作者获取。

表 1 传统机器学习与深度学习方法的比较

指标	传统机器学习模型（例如 logistic、SVM 等）	深度学习模型（例如 CNN、RNN 等）
特征表示	需要手动提取特征，输入被表示为向量	无须手动提取特征，特征被表示自动学习
模型参数数量	与特征个数线性相关	可达到千万、亿个参数规模
所需训练样本（以保证高准确率）	即使增大训练样本，性能也可能无法提升	巨大（百万、千万个观测值级别）
可处理任务	输出为单个随机变量	输出为随机变量的结构体
计算代价	小	大（往往需要 GPU 进行运算）

4. 构建具有针对性的中文情绪词典。词典法以预先设定的词典为基础来计算文本中不同类型词语的词频，并结合合适的加权方法来提取文本信息。然而，在中文文本情绪分类方面，词典法的应用目前仍处于探索阶段。多数学者选择已有的英文情绪词典以及词库作为参照来构建中文文本情绪词典，这就导致了构建的词典缺乏针对中文语境的问题。另外，不同来源及类别的文本信息在语言使用方面具有不同特点。例如，企业年报中专业术语较多，社交网络媒体中俚语和表情符号使用较多（姚加权等，2019）。因此，应当针对不同的中文文本内容构建具有针对性的中文文本情绪词典，并在未来的研究中持续验证且更新词典内容。

5. 改进文本可读性指标。目前在文本可读性的衡量方面，多数学者参考迷雾指数展开分析。但是，语言的语序和逻辑关系是影响可读性的重要因素，如果仅考虑句子的长度和复杂字词的比例而忽略语序和逻辑，仍无法准确衡量读者对文本的理解程度。目前，已有新的指标对文本的可读性进行衡量。例如，StyleWriter 软件包中的 Bog 指数（Bonsall IV et al，2017），捕获了语言学家强调的且又简单的英文文本特征。任宏达和王琨（2018）指出，运用机器学习方法衡量的可读性指标具有全面性和综合性，还可以克服自然语言的障碍。因此可以预见的是，在未来的研究中，更多学者会考虑采用新的指标展开研究，并使用机器学习甚至是深度学习模型，构建更加综合和准确的可读性指标。

另外，基于文本中表格内容的统计信息可能会构造出更有效的可读性指标。在以往的研究中，一般会将文档中的表格删除，只分析文本段落中的内容。然而，表格内容包含的数字化信息往往比文字信息更加客观和容易理解。如前所述，利用更精准的文档结构识别技术识别出文档中的所有表格后，可以计算每页平均的表格数量，文档中数字和文字的相对比例以及数字在表格和文本段落中出现的比例等也可以视为可读性指标。

6. 提高研究的可复制性。文本大数据的非结构化特征使得将其转化为结构化数据的过程比较复杂，其转变方法会影响研究的可复制性，即他人能否按照文章描述的研究思路和方法得出一致的结论。已经有学者在论文中以附录的形式详细介绍研究的文本分析方法与步骤（Huang et al，2018；姚加权等，2019）。因此，在未来的文本分析中，为了提高研究的可复制性，作者应当详细记录文档的预处理过程、文档表示以及特征的抽取方式。无论是使用词典法还是使用较为复杂的机器学习和深度学习方法，研究者都应当详细地揭示影响研究结果的关键词、词典以及具体思路和算法等。

参考文献：

陈霄 叶德珠 邓洁，2018：《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》第 3 期。

何贤杰 等，2016：《上市公司网络新媒体信息披露研究：基于微博的实证分析》，《财经研究》第 3 期。

林乐 谢德仁，2016：《投资者会听话听音吗？——基于管理层语调视角的实证研究》，《财经研究》第 7 期。

彭红枫 林川，2018：《言之有物：网络借贷中语言有用吗？——来自人人贷借款描述的经验证据》，《金融研究》第 11 期。

丘心颖 郑小翠 邓可斌，2016：《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）》第 4 期。

任宏达 王琨，2018：《社会关系与企业信息披露质量——基于中国上市公司年报的文本分析》，《南开管理评论》第 5 期。

沈艳 陈赟 黄卓,2018:《文本大数据分析在经济学和金融学中的应用》,北京大学工作论文。
孙书娜 孙谦,2018:《投资者关注和股市表现——基于雪球关注度的研究》,《管理科学学报》第 6 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。
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。
王靖一 黄益平,2018:《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)》第 4 期。
王雄元 高曦 何捷,2018:《年报风险信息披露与审计费用——基于文本余弦相似度视角》,《审计研究》第 5 期。
姚加权 等,2019:《语调、情绪及其市场影响——基于金融领域中文情绪词典的研究》,暨南大学工作论文。
游家兴 吴静,2012:《沉默的螺旋:媒体情绪与资产误定价》,《经济研究》第 7 期。
俞庆进 张兵,2012:《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》第 8 期。
曾建光,2015:《网络安全风险感知与互联网金融的资产定价》,《经济研究》第 7 期。
曾庆生 等,2018:《年报语调与内部人交易:"表里如一"还是"口是心非"?》,《管理世界》第 9 期。
张学勇 吴雨玲,2018:《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》第 6 期。
Allee, K. D. & M. D. Deangelis(2015), "The structure of voluntary disclosure narratives: Evidence from tone dispersion", *Journal of Accounting Research* 53(2):241－274.
Ahern, K. R. & D. Sosyura(2015), "Rumor has it: Sensationalism in financial media", *Review of Financial Studies* 28(7):2050－2093.
Antweiler, W. & M. Frank(2004), "Is all that talk just noise? The information content of internet stock message boards", *Journal of Finance* 59(3):1259－1294.
Baker, S. R. et al(2016), "Measuring economic policy uncertainty", *Quarterly Journal of Economics* 131(4):1593－1636.
Baloria, V. P. & J. Heese(2018), "The effects of media slant on firm behavior", *Journal of Financial Economics* 129(1):184－202.
Bandiera, O. et al(2020), "CEO behavior and firm performance", *Journal of Political Economy*, forthcoming.
Blankespoor, E. et al(2014), "The role of dissemination in market liquidity: Evidence from firms' use of twitter", *Accounting Review* 89(1):79－112.
Blei, D. M. et al(2003), "Latent Dirichlet allocation", *Journal of Machine Learning Research* 3: 993－1022.
Brown, S. V. & J. W. Tucker(2011), "Large-sample evidence on firm's year-over-year MD&A modifications", *Journal of Accounting Research* 49(2):309－346.
Bochkay, K. et al(2020), "Hyperbole or reality? Investor responses to extreme language in earnings conference calls", *Accounting Review* 95(2):31－60.
Bochkay, K. et al(2019), "Dynamics of CEO disclosure style", *Accounting Review* 94(4): 103－140.
Bonsall IV, S. et al(2017), "A plain English measure of financial reporting readability", *Journal of Accounting and Economics* 63(2－3):329－357.
Bonaime, A. et al(2018), "Does policy uncertainty affect mergers and acquisitions?", *Journal of Financial Economics* 129(3):531－558.
Buehlmaier, M. M. & T. M. Whited(2018), "Are financial constraints priced? Evidence from textual analysis", *Review of Financial Studies* 31(7):2693－2728.
Bushee, B. J., I. D. Gow & D. J. Taylor(2018), "Linguistic complexity in firm disclosures: Obfuscation or information?", *Journal of Accounting Research* 56(1):85－121.
Cao, Y. et al(2018), "Towards automatic numerical cross-checking: Extracting formulas from text", In: *Proceedings of the* 2018 *World Wide Web Conference*, pp. 1795－1804.
Cardinaels, E. et al(2018), "Automatic summaries of earnings releases: Attributes and effects on investors' judgments", Tilburg University Working Paper, https://ssrn.com/abstract=2904384.
Chen, H. et al(2014), "Wisdom of crowds: The value of stock opinions transmitted through social media", *Review of Financial Studies* 27(5):1367－1403.
Chen, M. A., Q. Wu & B. Yang(2019), "How valuable is FinTech innovation?", *Review of Financial Studies* 35(5):2062－2106.
Chi, S. & D. M. Shanthikumar(2017), "Local bias in google search and the market response around earnings announcements", *Accounting Review* 92(4):115－143.

Cole, C. J. & C. L. Jones(2005), "Management discussion and analysis: A review and implication for future research", *Journal of Accounting Literature* 24:135—174.

Cookson, J. A. & M. Niessner(2020), "Why don't we agree? Evidence from a social network of investors", *Journal of Finance* 75(1):173—228.

Cong, L. W., T. Liang & X. Zhang(2019), "Analyzing textual information at scale", Cornell University Working Paper, https://ssrn.com/abstract=3449822.

Da, Z., J. Engelberg & P. Gao(2011), "In search of attention", *Journal of Finance* 66(5): 1461—1499.

Da, Z., J. Engelberg & P. Gao(2015), "The sum of all fears investor sentiment and asset prices", *Review of Financial Studies* 28(1):1—32.

Das, S. R. & M. Y. Chen(2007), "Yahoo! for amazon: Sentiment extraction from small talk on the web", *Management Science* 53(9):1375—1388.

Davis, A. K. et al(2015), "The effect of manager-specific optimism on the tone of earnings conference calls", *Review of Accounting Studies* 20(2):639—673.

De Franco, G. et al(2015), "Analyst report readability", *Contemporary Accounting Research* 32(1): 76—104.

Feldman, R. et al(2010), "Management's tone change, post earnings announcement drift and accruals", *Review of Accounting Studies* 15(4):915—953.

Frank, M. Z. & A. Sanati(2018), "How does the stock market absorb shocks?", *Journal of Financial Economics* 129(1):136—153.

Garcia, D.(2013), "Sentiment during recessions", *Journal of Finance* 68(3):1267—1300.

Gentzkow, M. et al(2019), "Text as data", *Journal of Economic Literature* 57(3):535—574.

Green, T. C. et al(2019), "Crowdsourced employer reviews and stock returns", *Journal of Financial Economics* 134(1):236—251.

Guay, W. et al(2016), "Guiding through the fog: Financial statement complexity and voluntary disclosure", *Journal of Accounting and Economics* 62(2—3):234—269.

Gulen, H. & M. Ion(2016), "Policy uncertainty and corporate investment", *Review of Financial Studies* 29(3):523—564.

Guo, L., F. Shi & J. Tu(2016), "Textual analysis and machine leaning: Crack unstructured data in finance and accounting", *Journal of Finance and Data Science* 2(3):153—170.

Gurun, U. G. & A. W. Butler(2012), "Don't believe the hype: Local media slant, local advertising, and firm value", *Journal of Finance* 67(2):561—598.

Hanley, K. W. & G. Hoberg(2019), "Dynamic interpretation of emerging risks in the financial sector", *Review of Financial Studies* 32(12):4543—4603.

Heaton, J. B. et al(2016), "Deep learning in finance: Deep portfolios", *Applied Stochastic Models in Business and Industry* 33(1):3—12.

Henry, E. & A. J. Leone(2016), "Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone", *Accounting Review* 91(1): 153—178.

Hillert, A. et al(2014), "Media makes momentum", *Review of Financial Studies* 27(12): 3467—3501.

Hoberg, G. & G. Phillips(2010), "Product market synergies and competition in mergers and acquisitions: A text-based analysis", *Review of Financial Studies* 23(10):3773—3811.

Hoberg, G. & G. Phillips(2016), "Text-based network industries and endogenous product differentiation", *Journal of Political Economy* 124(5):1423—1465.

Huang, Y., H. Qiu & Z. Wu(2016), "Local bias in investor attention: Evidence from China's internet stock message boards", *Journal of Empirical Finance* 38:338—354.

Huang, J.(2018), "The customer knows best: The investment value of consumer opinions", *Journal of Financial Economics* 128(1):164—182.

Huang, A. H. et al(2018), "Analyst information discovery and interpretation roles: A topic modeling approach", *Management Science* 64(4):2833—2855.

Huang, Y. & P. Luk(2020), "Measuring economic policy uncertainty in China", *China Economic Review* 59, No. 101367.

Hwang, B. H. & H. H. Kim(2017), "It pays to write well", *Journal of Financial Economics* 124(2): 373—394.

Jegadeesh, N. & D. Wu(2013), "Word power: A new approach for content analysis", *Journal of Financial Economics* 110(3):712—729.

Jiang, F. et al(2019), "Manager sentiment and stock returns", *Journal of Financial Economics* 132(1):126—149.

Jiang, L., J. Liu & B. Yang(2019), "Communication and comovement: Evidence from online stock forums", *Financial Management* 48(3):805—847.

Jones, M. J. & P. A. Shoemaker(1994), "Accounting narratives: A review of empirical studies of content and readability", *Journal of Accounting Literature* 13:142—184.

Jung, M. J. et al(2018), "Do firms strategically disseminate? Evidence from corporate use of social media", *Accounting Review* 93(4):225—252.

Kelly, B. et al(2018), "Measuring technological innovation over the long run", NBER Working Paper, No. w25266.

Kogan, S. et al(2019), "Fake news: Evidence from financial markets", MIT Sloan School of Management Working Paper, https://ssrn.com/abstract=3237763.

Lang, W. & L. Stice-Lawrence(2015), "Textual analysis and international financial reporting: Large sample evidence", *Journal of Accounting and Economics* 60(2—3):110—135.

Lehavy, R. et al(2011), "The effect of annual report readability on analyst following and the properties of their earnings forecasts", *Accounting Review* 86(3):1087—1115.

Li, F.(2008), "Annual report readability, current earnings, and earnings persistence", *Journal of Accounting and Economics* 45(2—3): 221—247.

Li, F.(2010a), "Textual analysis of corporate disclosure: A survey of the literature", *Journal of Accounting Literature* 29:143—165.

Li, F.(2010b), "The information content of forward-looking statement in corporate filings: A naïve Bayesian machines learning approach", *Journal of Accounting Research* 48(5):1049—1102.

Li, K. et al(2020), "Measuring corporate culture using machine learning", University of British Columbia Working Paper, https://ssrn.com/abstract=3256608.

Lo, K. et al(2017), "Earnings management and annual report readability", *Journal of Accounting and Economics* 63(1):1—25.

Loughran, T. & 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. & B. McDonald(2014), "Measuring readability in financial disclosures", *Journal of Finance* 69(4): 1643—1671.

Loughran, T. & B. McDonald(2016), "Textual analysis in accounting and finance: A survey", *Journal of Accounting Research* 54(4):1187—1230.

Loughran, T., B. McDonald & H. Yun(2009), "A wolf in sheep's clothing: The use of ethics-related terms in 10—K reports", *Journal of Business Ethics* 89(1):39—49.

Mayew, W. J. et al(2015), "MD&A disclosure and the firm's ability to continue as a going concern", *Accounting Review* 90(4):1621—1651.

Merkley, K. J.(2014), "Narrative disclosure and earnings performance: Evidence from R&D disclosures", *Accounting Review* 89(2):725—757.

Mikolov, T. et al(2013), "Distributed representations of words and phrases and their compositionality", In: *Advances in Neural Information Processing Systems* 26 (*NIPS* 2013), pp. 3111—3119.

Piotroski, J. D., T. J. Wong & T. Zhang(2017), "Political bias in corporate news: The role of conglomeration reform in China", *Journal of Law and Economics* 60(1):173—207.

Price, S. M. et al(2012), "Earnings conference calls and stock returns: The incremental informativeness of textual tone", *Journal of Banking & Finance* 36(4):992—1011.

Ryans, J. P.(2019), "Textual classification of SEC comment letters", *Review of Accounting Studies*, forthcoming.

Solomon, D. H. et al(2014), "Winners in the spotlight: Media coverage of fund holdings as a driver of flows", *Journal of Financial Economics* 113(1):53—72.

Soo, C. K(2018), "Quantifying sentiment with news media across local housing markets", *Review of Financial Studies* 31(10):3689－3719.

Teh, Y. W. et al(2006), "Sharing clusters among related groups: Hierarchical Dirichlet processes", *Journal of the American Statistical Association* 101(476):1566－1581.

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

Tetlock, P. C. et al(2008), "More than words: Quantifying language to measure firms' fundamentals", *Journal of Finance* 63(3):1437－1467.

Teoh, S. H. (2018), "The promise and challenges of news datasets for accounting research", *Accounting, Organizations and Society* 68－69:109－117.

Yan, Y. et al(2019), "Uncertainty and IPO initial returns: Evidence from the tone analysis of China's IPO prospectuses", *Pacific-Basin Finance Journal* 57, No. 101075.

You, H. & X. Zhang(2009), "Financial reporting complexity and investor underreaction to 10－K information", *Review of Accounting Studies* 14(4):559－586.

You, J., B. Zhang & L. Zhang(2018), "Who captures the power of the pen?", *Review of Financial Studies* 31(1): 43－96.

Text Mining in Financial Big Data and Its Research Progress

YAO Jiaquan[1] ZHANG Kunpeng[1] LUO Ping[2,3]

(1. Jinan University, Guangzhou, China; 2. Chinese Academy of Sciences, Beijing, China;
3. University of Chinese Academy of Sciences, Beijing, China)

Abstract: Traditional empirical studies in the field of finance usually rely on structured data such as financial statements and stock market trading data. In the era of big data, data types have enriched with the improvement of computer technology and researchers have begun to introduce textual big data into the field of finance, mainly including the disclosure documents of listed companies, financial media reports, social network texts, internet search index, P2P online lending texts, and have examined the readability, tone, similarity and semantic characteristics of the text. This paper first introduces the steps and methods of textual big data mining in the field of finance, describing the corpus acquisition, preprocessing, document representation and the extraction process of document features. In addition, according to different sources of textual information, this paper introduces the research progress in financial textual big data. Finally, this paper provides a comprehensive research prospect on the research methods and topics of financial textual big data.

Keywords: Textual Big Data; Textual Analysis; Machine Learning; Deep Learning; Data Mining

（责任编辑：刘洪愧）
（校对：李仁贵）