# 基于 LDA 主题模型的上市公司违规识别
## ——以中国 A 股上市银行为例

张　熠,徐　阳,李维萍

(南京审计大学 信息工程学院,江苏 南京 211815)

[摘　要]以我国 2010—2019 年的 A 股上市银行年报为样本,利用 LDA 主题模型深度挖掘年报语义信息并构建银行年报的主题指标,在多种机器学习模型上对比主题指标与常用的财务指标、文本特征指标及其与主题指标的合并指标在检测上市银行违规时的性能。研究发现:年报文本主题内容对上市银行的违规行为有一定的预测作用,且与单一传统指标相比,主题指标可以提升传统指标的违规识别性能。研究结果为使用年报文本主题信息和机器学习方法识别上市银行违规的有效性提供了直接的证据,为市场构建了一种有效的违规识别指标体系,为审计师找到了一种较为高效的违规识别方法,有助于进一步规避与防范审计风险。

[关键词]上市公司违规识别;年度报告;LDA 主题模型;机器学习;违规预测;财务报表

[中图分类号]F275　　[文献标志码]A　　[文章编号]1004 - 4833(2022)05 - 0107 - 10

## 一、引言

近些年来,我国上市公司违规丑闻频发,对投资者的决策以及证券市场的秩序都造成了巨大影响,甚至影响国家的经济运行。因此目前识别上市公司违规的研究层出不穷。传统违规识别是基于年报中的结构化数据构建财务指标[1-3]。随着文本分析技术的不断发展,学者们开始重视年报中的非结构化数据即文本数据,利用文本分析技术提取文本信息,构建语义、语调等相关指标用于违规预测[4-6]。《公开发行证券的公司信息披露内容与格式准则第 2 号——年度报告的内容与格式(2021 年修订)》中指出,公司年报中应该包括经营情况讨论与分析、董事会报告、监事会报告等内容。这些非结构化文本主观性强,其编写者更有可能对其进行粉饰夸大甚至编制虚假财务报告,从而造成重大错报风险。在传统审计工作中,随着企业规模扩大,企业的经济活动愈加复杂,审计工作量也随之变大,多数情况下在审计过程中发现的都是财务报表中存在的技术性错误,审计风险依然存在。此外,在有限的时间内,注册会计师在审计时会依据经验将注意力更多地放置在高风险领域。而随着时间的推移,企业的违规手段愈加复杂和隐蔽,仅仅依靠审计师的经验和传统的财务报告分析手段并不能识别出更多的违规行为。因此根据传统的审计模式与方法,审计人员仍有较大造成审计失败的风险。为了进一步规避审计风险、减少审计失败,亟需提高注册会计师识别企业违规的能力。本文基于 Brown 的思想,利用 LDA(Latent Dirichlet Allocation)主题模型提取上市公司年报的潜在主题内容,构建主题指标[4],捕获年报中管理者是否存在违规意图并获得审计线索。与传统的审计模式相比,利用主题挖掘技术可以更快速地锁定更多的审计疑点,降低审计风险,提高审计效率,辅助注册会计师更精准、更快速地判断上市公司是否存在违规行为并出具正确的审计意见。

由于我国的审查制度以及相关的法律法规等都在不断完善,对于常用的违规手段都能够监察到位,但仍有上市公司为了谋取巨额利润而铤而走险。为了避免被审计人员发现其违规行为,上市公司的违规手段也在随着时间的推移不断地发生变化,采用更新颖、更隐蔽的违规方法与手段。为了探究基于年报文本所构建的主题指标是否可以识别出不断变化的违规手法,本文采用滚动窗口的方法来研究年报主题与上市公司是否存在违规行为之间的相关性,并观察各时间窗口内与违规相关的主题的演化情况。同时,为了检验主题指标是否可有效地对年报中的违规进行预测,本文在每个时间窗口上运行 LDA 主题模型构建主题指标,基于主题指标运用多种机

[收稿日期]2022 - 03 - 18

[基金项目]江苏省社会科学基金项目(21GLD009)

[作者简介]张熠(1980— ),男,安徽桐城人,南京审计大学信息工程学院讲师,从事审计大数据、金融大数据研究;徐阳(1999— ),女,山东菏泽人,南京审计大学信息工程学院硕士研究生,从事审计大数据研究;李维萍(1963— ),男,四川德阳人;南京审计大学信息工程学院教授,博士生导师,从事审计大数据、金融大数据研究,E-mail:320319@ nau. edu. cn,通讯作者。

器学习算法构建上市公司违规预测模型,并对每个时间窗口后一年的违规情况进行预测。此外,由于违规样本占比较低,样本数据存在不平衡现象,本文使采用多种指标评估了模型的预测能力并对主题指标、财务指标、文本特征指标以及合并指标的预测效果进行了对比分析。

本文的主要贡献主要体现在:首先,不同于之前研究中用到的传统财务指标和文本特征指标,本文通过挖掘银行年报的潜在主题信息来构建主题指标,并用于对上市公司违规识别,且提升了传统指标违规识别的准确性,进一步降低了审计风险并提升了审计效率。其次,本文分析了主题指标与违规的相关性以及与违规显著相关的主题随时间推移呈现的变化情况,得到了尽管违规手法愈加隐蔽且复杂,主题指标仍可以有效识别的结论。最后,在相同的输入样本的基础上,检验了不同指标体系、不同机器学习算法在上市银行违规识别上的优劣,找到了更为高效的智能化违规识别方法。

## 二、文献回顾

关于上市公司违规识别的研究可以划分为两个方面,分别是基于财务信息的违规识别研究以及基于非财务信息的违规识别研究。在传统的基于结构化财务数据预测上市公司违规的研究中,为了识别多种类型的违规,所选变量应尽量涵盖公司业绩的各个方面,因此一般所使用的初始财务变量呈现出数量多且复杂的特点。Dechow 等为了发现美国上市公司存在的财务舞弊现象,从应计质量、财务绩效、非财务绩效、表外活动和资本市场等五个方面选择 28 个结构化变量,建立舞弊识别模型[1]。针对中国上市公司的舞弊现象,有研究也从财务杠杆、营运能力、盈利能力等方面选择财务变量构建舞弊识别模型[7]。尽管常规的财务指标在违规预测中可以表现出很好的预测性能,但在财务变量的选取过程中存在较强的主观性,对模型的分类效果也存在相应的干扰[8]。同时随着上市公司违规手法的愈加高明与隐蔽,仅利用财务指标也无法识别出更多的违规现象。有研究发现,与财务信息相比,非财务信息在反映公司经营活动和未来发展前景上表现更加突出[9]。随着文本分析技术和自然语言处理技术的不断发展,学者们便将注意力转移到年报中的文本上,并利用年报文本构建相关指标从而挖掘年报文本与违规之间的内在联系。有研究发现舞弊公司年报中的管理层与讨论(MD&A)部分中会增加美化公司绩效的内容[10]。此外,与非舞弊年报相比,舞弊年报中的句子会较多使用被动语态和不确定的词汇[11]。国内在这方面的研究主要基于情感分析,研究发现年报中的情感特征有助于财务舞弊的识别[12]。通过梳理国内外的研究发现,用于识别违规的文本特征指标主要涉及文本语调、可读性、相似性等,但以上指标只能反映文本披露形式,并不能揭示文本披露信息的内在含义[4]。本文通过对年报文本进行主题挖掘,提取年报潜在主题信息并构建适合中国市场的违规识别模型,可以辅助注册会计师发现更多的审计疑点,从而进一步规避了审计风险。

本文利用 LDA 主题模型提取年报潜在主题。LDA 主题模型由 Blei 等人在 2003 年提出,该主题模型是一个概率主题模型,通过建模后可获得文本所对应的主题概率分布[13]。目前在国内 LDA 主题模型主要用于社交网络、情报分析等领域[14-19]。如关鹏等将 LDA 主题模型应用在科技情报分析中,对基于不同科学文献文本语料库而建立的模型的主题发现效果进行对比评价[14]。目前将 LDA 主题模型应用到经济和金融领域的研究较少,有研究将 LDA 主题模型应用到财经新闻文本上并基于此分析主题强度与孟买股票交易所敏感指数的每日收盘价等指标之间的相关关系,并将主题热度用于预测指数的涨跌[19]。近几年在国内也有学者将 LDA 主题模型应用到财经文本上,在此基础上研究主题的强度、热度或其他特征,并将结果进行可视化[20-21]。如傅魁等人对 LDA 主题模型进行扩展,提出 SGC-LDA 财经文本主题研究模型[20]。

综上,国内外的学者们在构建违规预测模型时关注到了财务信息与文本披露的形式,并基于此来构建财务指标和文本特征指标,但较少研究年报中所披露的主题内容并将其应用于上市公司违规识别中。本文采用 LDA 主题模型对年报文本建模,构建反映文本语义的主题指标,并用于识别上市银行是否存在违规行为。与财务指标和文本特征指标相比,主题指标蕴含了更为丰富的上下文信息和语义信息,从理论上来说可以更为准确地识别出公司管理者是否有违规意图。因此与单一传统指标相比,基于主题指标的违规识别模型打破了原有的审计模式,充分利用中文年报的非财务信息,将主题挖掘技术运用到审计中,帮助注册会计师更快速获得丰富的审计线索,从而减少审计失败的可能性,更大程度地规避审计风险。目前在国内还没有将年报主题信息用于违规识别方面的研究,因此研究文本主题指标与违规之间的联系对于中国市场来说具有重要意义。

## 三、理论分析和研究假设

### (一)年报文本主题与违规

我国年报中披露的管理层讨论与分析、董事会报告等文本信息不仅总结了公司上一年的经营成果、财务状况,也对公司即将要发生的重大事项和未来可能发生的变化进行了讨论与分析,因此年度报告中的文本数据可以反映公司将来的发展风险和趋势。此外,在上市公司年报中,文本数据所占篇幅明显高于财务数据,尤其在近几年,年报篇幅逐渐增加,除去三大报表,非结构化文本占比明显上升。因此通过对上市公司年报中的文本信息进行挖掘和分析,了解公司整体的业绩和发展趋势,可以获取更多的有效信息,无论是对于投资者的选择,还是注册会计师的决策,都是不可缺少的。

即使上市公司年报文本中蕴含着丰富的信息,但由于文本篇幅过长,完全依靠人工阅读、理解并直接提取文本中的有效信息难度非常大,且效率非常低。因此在本文中使用 LDA 主题模型对年报文本潜在的主题进行挖掘。主题挖掘是利用主题模型挖掘语料中的隐藏信息,发现一系列非结构化文本中的主题,也就是找出表达文本中心思想的主题词。同时,本文选用 LDA 主题模型这一最为通用的主题模型,提取具有语义信息的主题。因此,采用 LDA 主题模型所发掘出的年报文本的主题信息可以很好地反映出年报的潜在语义。由于年报中的非结构化文本在编写时自由度较大,主观性较强,可以传达公司许多内部信息,管理者为了牟利或掩盖本身经营问题可能会对年报文字部分进行美化,在用词遣句上避重就轻或进行选择性披露[22],导致其年报文本内容发生变化,继而导致使用 LDA 主题模型对年度报告挖掘后得到的主题信息也会发生改变。因此,年报文本主题是与上市公司违规显著相关的,通过年报主题可以反映出由于要掩盖违规行为所导致的年报内容的变化。此外,随着时间的推移,我国的会计准则、审计准则和监管手段也在不断进行修正与完善,在一定程度上阻止了某些违规行为的出现,但总会产生一些新的违规手段与方法[23],为了避免被审计人员发现,发生违规行为的年报中讨论的重点会随着时间的推移而变化,即可用于识别违规行为的主题并不是一成不变的,会随着违规手段的变化而变化。因此,可利用年报文本的主题信息去识别新出现的且更隐蔽的违规行为。基于以上的理论分析,本文提出假设 H1。

H1:年报文本主题会随着上市公司违规手段的变化而发生变化。

### (二)主题指标、财务指标与文本特征指标

目前识别上市公司违规的方法大多是基于结构化的财务数据或股票市场数据,但利用此类数据存在的一个缺点是违规公司会故意操纵当期的绩效指标和会计交易数据以便与本公司之前的业绩数据或同行的业绩数据保持一致,使得违规行为不易被发现[10],而且有研究指出结构化的财务报表数据可提供给投资者的信息是有限的[9]。因此,仅仅利用财务指标构建的违规识别模型的效果存在一定的不足。为了弥补财务数据的缺陷,研究人员利用财务报告中的非结构化文本数据去发现上市公司是否违规。有研究利用年报文本的语言结构特征来构建与违规相关的指标,如文本可读性、语调等。尽管研究发现利用文本特征指标来识别舞弊有一定的效果[24],但在一些研究中对于文本特征指标仍然存在质疑,即文本特征是否可以真正捕获到管理者违规的意图[25]。有语言学研究表明,很难从披露文本的文本特征中辨别出其中是否存在欺骗或混淆视听的内容[26]。此外,Loughran 和 McDonald 指出常用的文本语言特征指标并不能反映出文本的上下文和语义信息,从而导致利用文本语言特征指标建立的违规识别模型效果就会有所限制[24]。为了进一步提升违规识别模型的效果,本文基于 Brown 的思想,运用 LDA 主题模型构造年报的主题指标[4],提取年报文本中潜在的语义信息,反映出年报文本表达的真正含义并捕获公司管理者的违规意图。综上,基于年报非结构化文本的主题指标不仅蕴含丰富信息,还具有语义内涵,可以反映出文本披露的具体内容,可以弥补财务指标和文本特征指标在预测违规时的不足。因此,相对而言,在传统单一指标的基础上,主题指标可以提升识别上市公司违规的性能。基于以上分析,本文提出假设 2。

H2:在财务指标和文本特征指标的基础上,年报文本主题指标可以进一步提升违规识别模型的性能。

## 四、研究设计

### (一)样本选择与数据来源

由于我国相关的法律法规等都在不断推进,若选择时间过早的样本,样本对应的上市银行的治理结构、经营环

境间存在着较大差异,导致样本数据不可比。另外考虑到近期产生违规的公司还未被证监会认定,同时为了研究用于预测违规主题的变化过程,本文选取我国 36 家 A 股上市银行在 2010—2019 年间发布的年报作为研究样本。其中上市银行的年度报告均从巨潮资讯网中下载得到;文本语言特征数据是对银行年报进行文本分析以及人工计算整理的方式取得;财务数据来自于国泰安数据库。最终本文得到 215 个公司 – 年度层面的上市公司数据。

(二)变量定义

1. 被解释变量

本文的被解释变量为是否违规(*Fraud*),违规数据来源于国泰安经济金融研究数据库(*CSMAR*)和色诺芬数据库(*CCER*),并经过人工合并得到。若上市公司在上期年报和当期年报发布之间发生了违规行为,则将上市公司当期样本的 *Fraud* 变量赋值为 1,若上市公司未发生违规,则赋值为 0。最终本文得到 68 个违规样本,147 个非违规样本。

2. 解释变量

本文的解释变量为文本的主题指标变量。同时为了对比主题指标的预测效果,本文将财务变量、文本特征变量也作为解释变量。

(1)主题指标变量(*Topic*)

本文使用 LDA 主题模型进行年报主题指标的构建。LDA 主题模型可以得到相应数据集的两个概率分布,分别是“文档 – 主题”概率分布以及“主题 – 词”概率分布,其中“文档 – 主题”概率分布就是我们所构建的主题指标。此外,为了研究随着时间推移,与违规相关的主题指标的变化情况,本文采取滚动窗口的方式,将 2010—2019 年的样本区间划分为五个时间窗口,在五个时间窗口上分别运行 LDA 主题模型并构建相应的主题指标。表 1 呈现了每个时间窗口所构建的主题指标变量的定义。LDA 主题模型是无监督机器学习模型,只需提供文本集合和要生成的主题数。其中主题数对于 LDA 主题模型的聚类效果有很大影响。但目前如何得到主题模型的最优主题数这一问题尚未有最佳的方法。在以往的研究中选择最优主题数常用的方法是最小困惑度法,困惑度是指所构建的主题模型对一篇文档属于某一主题的不确定程度。困惑度越小,表示模型对于文本的主题选择越不“困惑”。但研究发现基于最小困惑度法得到的最优主题数数量过多,主题间相似度高,存在冗余情况[14]。

**表 1　主题指标变量定义**

| 变量名称 | 变量符号 | 变量定义 |
|---|---|---|
| 主题指标变量 | $X_{k1}$ | 基于 2010—2014 年间全部年报文本所得到的“文档 – 主题”概率分布 |
| | $X_{k2}$ | 基于 2011—2015 年间全部年报文本所得到的“文档 – 主题”概率分布 |
| | $X_{k3}$ | 基于 2012—2016 年间全部年报文本所得到的“文档 – 主题”概率分布 |
| | $X_{k4}$ | 基于 2013—2017 年间全部年报文本所得到的“文档 – 主题”概率分布 |
| | $X_{k5}$ | 基于 2014—2018 年间全部年报文本所得到的“文档 – 主题”概率分布 |

本文首先尝试使用最小困惑度方法来确定最终要产生的主题数量,将主题数分别设置为 1 至 51,分别训练 LDA 模型并计算模型困惑度。结果如图 1 所示。结果显示根据最小困惑度方法得到的最优主题数为 36 个。对 36 个主题进行可视化,观察主题的分布情况。结果如图 2 所示(图 2 中列示了 11 个主题圈,其他由于占比较小未列示,有需要可联系作者)。图中每一个圆圈代表一个主题,从可视化图中可以看出圆圈之间存在很多重叠部分,表示 36 个主题间存在很高的重复性即存在冗余主题数,验证了通过最小困惑度方法得到的最优主题数量过多。

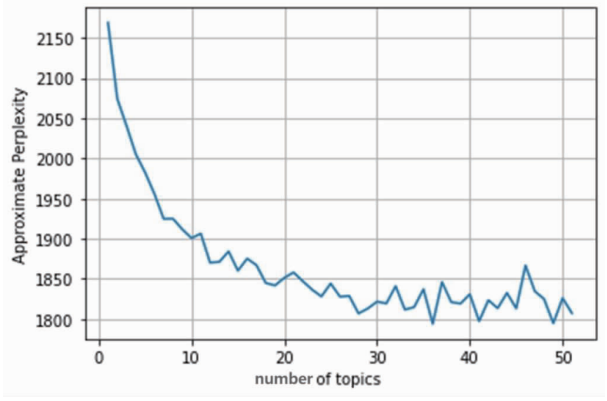


**图 1　困惑度与主题数的关系**

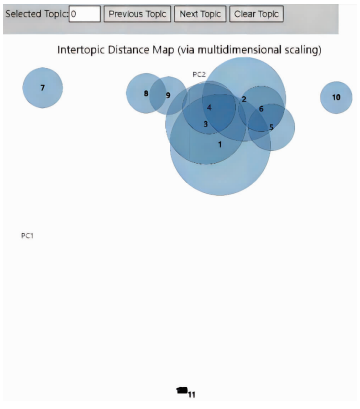


**图 2　36 个主题可视化结果**

为了避免主题相似度过高,本文借鉴王泽贤的主题数选择方法即最小冗余主题数法[27],即在初设最大主题数的基础上,采用二分法逐步减小主题数量,使得最终得到的主题两两之间 JS 散度为 0。最终得到的主题数结果如表 2 所示。

**表 2　最优主题数**

| 年度 | 文件数 | 主题数 |
|---|---|---|
| 2010—2014 | 80 | 9 |
| 2011—2015 | 80 | 9 |
| 2012—2016 | 89 | 9 |
| 2013—2017 | 99 | 12 |
| 2014—2018 | 115 | 10 |

(2)财务变量(*F-score*)

本文基于美国学者 Dechow 研究的 F-score 舞弊识别模型中的变量体系[1],并对其变量进行筛选,最终选择 12 个变量作为本文的财务变量,分别为公司资产、应收账款变动、总应计、软资产占比、现金销售变动、资产回报率变动、净值市价比、前期持有期收益率、并购情况、会计师事务所是否为四大、融资现金流量、重组情况。财务变量的具体定义如表 3 所示。

**表 3　财务变量定义**

| 变量名称 | 变量符号 | 变量定义 |
|---|---|---|
| 公司资产 | *Size* | 公司总资产取对数 |
| 应收账款变动 | Δ*Receivables* | Δ 应收账款/平均总资产 |
| 总应计 | *RSST accruals* | Δ(总资产 - 总负债 - 现金)/平均总资产 |
| 软资产占比 | *% Soft assets* | (总资产 - *PP&E* - 现金及现金等价物)/总资产 |
| 现金销售变动 | Δ*Cash sales* | [(当期营业收入 - Δ 应收账款) - (前期营业收入 - Δ 应收账款)]/(前期营业收入 - Δ 应收账款) |
| 资产回报率变动 | Δ*Return on assets* | 当期净利润/当期平均总资产 - 前期净利润/前期平均总资产 |
| 净值市价比 | *Book-to-market* | (总资产 - 总负债)/股票市值 |
| 前期持有期收益率 | *Lag*(*Mkt-Adj Return*) | 前期个股回报率 - 加权综合市场回报率 |
| 并购情况 | *Merger* | 当期完成并购时为 1,否则为 0 |
| 是否为四大 | *Tot Financing* | 聘请的会计师事务所为四大时为 1,否则为 0 |
| 融资现金流量 | *Big*4 | 净融资现金流量/平均总资产 |
| 重组情况 | *Restructuring* | 当期完成重组时为 1,否则为 0 |

注:表 3 中变量符号中 Δ 表示变化量,指该变量当期值与上期值之差;% 表示某变量占总资产的比例

(3)文本特征变量(*Style*)

本文基于 Brown 等使用的文本语言特征变量及其构建方法[4],对 36 家上市银行在 2010—2019 年间披露的年报全文,通过文本分析等方式构建文本特征指标。由于美国与中国的年报在披露标准与内容上有所不同,因此本文对其变量进行筛选。最终选择的文本特征变量分别为着重号数量、换行数、标签数、平均句长、词长标准差、段长标准差、平均重复句数、句长标准差、唯一词比例、Coleman-Liau 指数、Fog 指数、主动句比例、被动句比例、消极词比例、积极词比例。变量的具体定义如表 4 所示。

**表 4　文本特征变量定义**

| 变量名称 | 变量符号 | 变量定义 |
|---|---|---|
| 着重号数量 | *Log*(*Bullets*) | 年报中着重号数量取对数 |
| 换行数 | *Newlines* | 年报中多余的换行数 |
| 标签数 | *Tags* | 解析后年报中 *HTML* 标签数 |
| 平均句长 | *Sentence Length* | 年报中句子长度的平均数 |
| 词长标准差 | *Word Stddev* | 年报中词语长度的标准差 |
| 段长标准差 | *Paragraph Stddev* | 年报中段落长度的标准差 |
| 平均重复句数 | *Repetitions* | 年报中重复句数量/句子总数 |
| 句长标准差 | *Sentence Stddev* | 年报中句子长度的标准差 |
| 唯一词比例 | *Type Token Ratio* | 年报中只出现 1 次的词语占词语总数的比例 |
| *Coleman-Liau* 指数 | *Coleman-Liau Index* | $5.88 \times C/W - 29.6 \times S/W - 15.8$,其中 $C$ 是指年报中的总字数,$W$ 是指总词数,$S$ 是指句子总数 |
| *Fog* 指数 | *Fog* | $0.4 \times (W/S + 100 \times W'/W)$,$W'$ 是指复杂词语数,此处指字数大于五个字的词语数 |
| 主动句比例 | *% Active Voice* | 年报中主动句占句子总数的比例 |
| 被动句比例 | *% Passive Voice* | 年报中被动句占句子总数的比例 |
| 消极词比例 | *% Negative* | 根据清华大学李军中文褒贬义词典,年报中消极词占总词数的比例 |
| 积极词比例 | *% Positive* | 根据清华大学李军中文褒贬义词典,年报中积极词占总词数的比例 |

(三)模型构建

随着计算机技术的发展,机器学习已经逐渐成为研究的热点,若能使用机器学习识别上市公司的违规行为,对于审计师、投资者以及各类监管机构都能起到重要的辅助作用。识别上市公司的违规行为适用于机器学习模型的分类算法。本文选取目前常用的机器学习分类算法,分别为逻辑回归模型(Logistic Regression)、K - 近邻模型(K-nearest neighbor,简称 KNN)、支持向量机(Suppprt Vector Machine,简称 SVM)、随机森林(Random Forest,简称 RF)、AdaBoost(Adaptive Boosting)、多层感知器(Multilayer Perceptron,简称 MLP)。本文在五个时间窗口上,将主题指标、财务指标、文本特征指标、主题指标 + 财务指标、主题指标 + 文本特征指标分别作为以上机器学

习模型的输入指标,并对每个时间窗口后一年的违规情况进行预测,比较不同指标体系、不同机器学习模型在识别上市公司违规时的优劣,探究加入主题指标是否能够提高财务指标或文本特征指标的预测效果。

$$LOGIT(Fraud) = \alpha_0 + \sum_{j=1}^{12}\alpha_j F\text{-}score_{j,i,t} + \varepsilon_{i,t}, t \in [T-5, T-1], i \in Firms \quad (1)$$

$$LOGIT(Fraud) = \beta_0 + \sum_{j=1}^{15}\beta_j Style_{j,i,t} + \xi_{i,t}, t \in [T-5, T-1], i \in Firms \quad (2)$$

$$LOGIT(Fraud) = \gamma_0 + \sum_{j=1}^{k}\gamma_j Topic_{j,i,t} + \eta_{i,t}, t \in [T-5, T-1], i \in Firms \quad (3)$$

$$LOGIT(Fraud) = \delta_0 + \sum_{j=1}^{12}\delta_j F\text{-}score_{j,i,t} + \sum_{j=1}^{k}\delta_{j+12} Topic_{j,i,t} + \theta_{i,t}, t \in [T-5, T-1], i \in Firms \quad (4)$$

$$LOGIT(Fraud) = \phi_0 + \sum_{j=1}^{15}\phi_j Style_{j,i,t} + \sum_{j=1}^{k}\phi_{j+15} Topic_{j,i,t} + \vartheta_{i,t}, t \in [T-5, T-1], i \in Firms \quad (5)$$

本文构建以上逻辑回归模型验证主题指标的有效性,并与财务指标、文本特征指标及合并指标进行对比分析。模型中 *Topic* 为上文中构建的主题指标,*F-score* 为财务指标,*Style* 为文本特征指标。同样地,本文还将构建 KNN、SVM、RF、AdaBoost 和 MLP 模型,并对不同模型的违规识别效果进行对比。

## 五、实证结果与分析

### (一)描述性统计

#### 1. 财务变量描述性统计

本文财务变量的描述性统计如表 5 所示。表中的财务变量均来自 CSMAR 数据库,涵盖了我国 36 家上市银行在 2010—2019 年间的财务数据,各个财务变量的具体定义如上文表 3 所示。本文将样本划分为两类样本,分别是正常样本与违规样本,并标记为 0、1,为了对比两类样本在同一财务变量上的差别,在表 5 中对每一类都进行了描述性统计。

**表 5　财务变量描述性统计**

| 变量 | 类别 | 样本量 | 最小值 | 25% 分位 | 中位数 | 75% 分位 | 最大值 | 均值 | 方差 |
|---|---|---|---|---|---|---|---|---|---|
| *Size* | 0 | 147 | 25.122 | 27.621 | 28.619 | 29.598 | 31.036 | 28.529 | 2.269 |
| | 1 | 68 | 25.225 | 27.225 | 28.359 | 29.452 | 30.952 | 28.224 | 2.536 |
| *ΔReceivables* | 0 | 147 | -0.098 | 0 | 0 | 0.010 | 0.084 | 0.005 | 0 |
| | 1 | 68 | -0.035 | 0 | 0 | 0.003 | 0.211 | 0.007 | 0.001 |
| *RSS Taccruals* | 0 | 147 | -0.109 | -0.004 | 0 | 0.010 | 0.111 | 0.004 | 0.001 |
| | 1 | 68 | -0.012 | 0 | 0.002 | 0.012 | 0.050 | 0.008 | 0 |
| *% Soft assets* | 0 | 147 | 0.677 | 0.911 | 0.934 | 0.951 | 0.976 | 0.925 | 0.001 |
| | 1 | 68 | 0.882 | 0.937 | 0.952 | 0.967 | 0.977 | 0.948 | 0.001 |
| *ΔCash sales* | 0 | 147 | -2.203 | 0 | 0 | 0.162 | 1.757 | 0.048 | 0.132 |
| | 1 | 68 | -28.165 | -0.011 | 0 | 0 | 0.601 | -0.434 | 11.695 |
| *ΔReturn on assets* | 0 | 147 | -0.002 | 0 | 0 | 0.001 | 0.004 | 0 | 0 |
| | 1 | 68 | -0.002 | 0 | 0 | 0 | 0.001 | 0 | 0 |
| *Book-to-market* | 0 | 147 | 0 | 0.831 | 0 | 1.349 | 2.796 | 1.119 | 0.257 |
| | 1 | 68 | 0 | 0.967 | 1.178 | 1.506 | 3.624 | 1.304 | 0.400 |
| *Lag(Mkt-AdjReturn)* | 0 | 147 | -0.824 | -0.287 | 0 | 0.213 | 1.877 | 0.061 | 0.240 |
| | 1 | 68 | -0.671 | -0.115 | 0.017 | 0.259 | 1.001 | 0.050 | 0.129 |
| *Merger* | 0 | 147 | 0 | 0 | 0 | 0 | 0 | 0.313 | 0.216 |
| | 1 | 68 | 0 | 0 | 0 | 0 | 0 | 0.353 | 0.232 |
| *Big4* | 0 | 147 | 0 | 0 | 0 | 0 | 0 | 0.905 | 0.087 |
| | 1 | 68 | 0 | 0 | 0 | 0 | 0 | 0.838 | 0.138 |
| *Tot Financing* | 0 | 147 | -0.014 | 0 | 0.002 | 0.007 | 0.066 | 0.006 | 0 |
| | 1 | 68 | -0.019 | 0 | 0.003 | 0.014 | 0.057 | 0.007 | 0 |
| *Restructure* | 0 | 147 | 0 | 0 | 0 | 0 | 0 | 0.197 | 0.159 |
| | 1 | 68 | 0 | 0 | 0 | 0 | 0 | 0.044 | 0.043 |

注:类别中的 0、1 分别指正常样本、违规样本。

#### 2. 文本特征变量描述性统计

本文文本特征变量的描述性统计如表 6 所示。表中的文本特征变量是对 36 家上市银行 2010—2019 年的年报进行文本分析后计算得到的。各个文本特征变量的具体定义如上文表 4 所示。为了对比两类样本在同一文本特征变量上的差别,本文在表 6 中对每一类样本都进行了描述性统计。

表 6　文本特征变量描述性统计

| 变量 | 类别 | 样本量 | 最小值 | 25% 分位 | 中位数 | 75% 分位 | 最大值 | 均值 | 方差 |
|---|---|---|---|---|---|---|---|---|---|
| Log(*Bullets*) | 0 | 147 | 0 | 0 | 0.693 | 1.609 | 4.248 | 1.117 | 1.341 |
| | 1 | 68 | 0 | 0 | 0.693 | 1.488 | 4.691 | 1.023 | 1.227 |
| *Newlines* | 0 | 147 | 658 | 48160 | 66960 | 8959 | 173130 | 7311 | 8947646 |
| | 1 | 68 | 6650 | 41990 | 6064 | 8769 | 166970 | 6866 | 12960384 |
| *Tags* | 0 | 147 | 602430 | 183416 | 2391380 | 282560 | 367538 | 235770 | 344068437 |
| | 1 | 68 | 67362 | 182705 | 224900 | 296058 | 385119 | 239733 | 506347955 |
| *Sentence Length* | 0 | 147 | 69.165 | 73.823 | 76.708 | 81.186 | 97.167 | 77.750 | 28.307 |
| | 1 | 68 | 68.653 | 75.527 | 77.853 | 80.421 | 124505 | 1908 | 227674877.485 |
| *Word Stddev* | 0 | 147 | 0.854 | 1.025 | 1.071 | 1.127 | 1.423 | 1.085 | 0.008 |
| | 1 | 68 | 0.899 | 1.010 | 1.043 | 1.096 | 1.287 | 1.061 | 0.006 |
| *Paragraph-Stddev* | 0 | 147 | 586.435 | 1272 | 1638 | 2183 | 4297 | 1761 | 511636 |
| | 1 | 68 | 765.598 | 1138 | 1507 | 2077 | 2835 | 1602 | 312459 |
| *Repetitions* | 0 | 147 | 0.005 | 0.032 | 0.048 | 0.100 | 0.569 | 0.089 | 0.011 |
| | 1 | 68 | 0.000 | 0.031 | 0.046 | 0.070 | 0.610 | 0.074 | 0.009 |
| *Sentence-Stddev* | 0 | 147 | 54.530 | 90.149 | 119.210 | 207.936 | 481.703 | 158.233 | 8916 |
| | 1 | 68 | 54.372 | 89.179 | 109.532 | 132.666 | 38866 | 694.158 | 22076290 |
| *Type Token Ratio* | 0 | 147 | 0.032 | 0.046 | 0.059 | 0.077 | 0.148 | 0.062 | 0.000 |
| | 1 | 68 | 0.031 | 0.041 | 0.055 | 0.078 | 0.119 | 0.061 | 0.000 |
| *Coleman-Liau Index* | 0 | 147 | -0.616 | 3.530 | 3.990 | 4.821 | 6.845 | 4.102 | 1.415 |
| | 1 | 68 | -1.579 | 3.286 | 3.669 | 4.316 | 6.233 | 3.749 | 1.452 |
| *Fog* | 0 | 147 | 9.118 | 10.068 | 10.501 | 11.025 | 13.706 | 10.566 | 9.118 |
| | 1 | 68 | 9.304 | 10.307 | 10.539 | 11.019 | 10307 | 162.107 | 15593306 |
| *% Active Voice* | 0 | 147 | 0.893 | 0.951 | 0.962 | 0.970 | 0.997 | 0.960 | 0.000 |
| | 1 | 68 | -30.000 | 0.944 | 0.956 | 0.968 | 0.998 | 0.499 | 14.091 |
| *% Passive Voice* | 0 | 147 | 0.003 | 0.030 | 0.038 | 0.049 | 0.107 | 0.040 | 0.000 |
| | 1 | 68 | 0.002 | 0.032 | 0.044 | 0.056 | 31.000 | 0.501 | 14.091 |
| *% Negative* | 0 | 147 | 0.443 | 2.086 | 2.409 | 2.596 | 3.265 | 2.284 | 0.260 |
| | 1 | 68 | 0.413 | 2.117 | 2.367 | 2.644 | 3.148 | 2.310 | 0.255 |
| *% Positive* | 0 | 147 | 2.836 | 4.668 | 5.030 | 5.632 | 8.806 | 5.182 | 0.975 |
| | 1 | 68 | 2.461 | 4.656 | 5.195 | 5.672 | 7.773 | 5.255 | 0.856 |

注：类别中的 0、1 分别指正常样本、违规样本。

（二）主题变化情况

本文采用滚动回归验证通过 LDA 主题模型所提取的年报主题指标与违规之间的相关性以及在不同时间段内与违规显著相关的主题的变化情况。本文的样本区间为 2010—2019 年，将每五年作为一个时间窗口，最终将样本区间划分为五个窗口，分别是 2010—2014 年、2011—2015 年、2012—2016 年、2013—2017 年、2014—2018 年，在五个窗口上分别运行 LDA 模型，提取每个窗口年报文本的潜在主题。为了便于展示，本文计算所有主题对应的词语权重向量间的余弦相似度，基于相似度将所有窗口内的单个主题聚合为组合主题，最终在整个样本区间内生成 14 个组合主题。

为了呈现与违规显著相关的主题在不同时间段内的变化情况，本文分别对每个时间窗口上的主题指标构建逻辑回归模型，根据回归系数的 $z$ 值判断组合主题的显著性。下图 3 描述了在样本预测年份 2015－2019 年上每个组合主题是否存在以及是否与违规显著相关。图中正方形表示在 50% 的置信水平下，该组合主题中至少一个子主题与违规显著负相关且其他子主题与违规均不显著相关；菱形表示在 50% 的置信水平下，该组合主题中至少一个子主题与违规显著正相关且其他子主题与违规均不显著相关；三角形表示该组合主题中所有子主题与违规相关性均不显著或多个子主题显著但显著性符号相反。

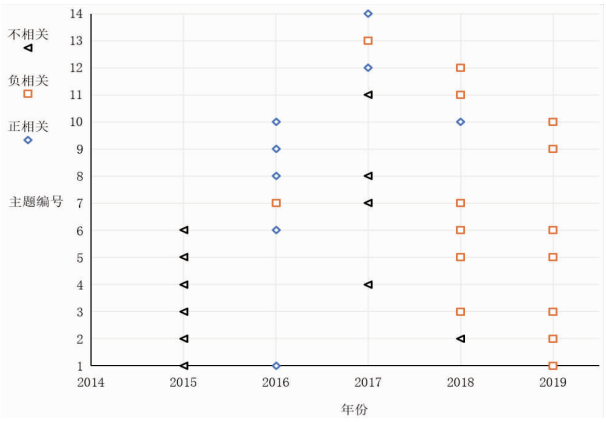


图 3　组合主题显著性与主题变化情况

通过图 3 可以发现存在多个主题与违规显著相关。随着时间的推移，可以观察到某些主题的变化情况，如组合主题 2 在 2015 年和 2018 年中与违规均无显著相关

关系,但在 2019 年与违规呈现出负相关关系;同时随着时间变化,也出现了一些之前未出现过的且与违规显著相关的主题,如组合主题 12、13、14,在前两年中未曾出现,在 2017 年、2018 年出现并与违规呈显著相关关系。因此本文认为对于一些新出现的违规行为,尽管手段更加隐蔽且复杂,我们依然可以通过主题指标找到与违规之间的相关关系,从而可有效识别违规。

(三)评估指标

在机器学习算法中,常用的评价指标有准确率(*Accuracy*)、精确率(*Precision*)、召回率(*Recall*)、F1 分数和 F2 分数。为了直观地解释以上指标,本文基于混淆矩阵对以上指标进行定义如表 7 所示。

**表 7　二分类混淆矩阵**

| 混淆矩阵 | | 真实值 | |
|---|---|---|---|
| | | *Positive* | *Negative* |
| 预测值 | *Positive* | TP | FP |
| | *Negative* | FN | TN |

基于此,本文将准确率、精确率、召回率、F1 分数、F2 分数定义如下。

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$P = \frac{TP}{TP + FP} \tag{7}$$

$$R = \frac{TP}{TP + FN} \tag{8}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{9}$$

$$F2 = \frac{5 \times P \times R}{4 \times P + R} \tag{10}$$

式中,$A$ 表示准确率,$P$ 表示精确率,$R$ 表示召回率,F1 表示 F1 分数,F2 表示 F2 分数。

由于在本文中违规样本与正常样本存在不平衡现象,有些机器学习算法会将样本全部预测为无违规从而造成准确率很高,但显然此时准确率指标已经失去参考意义。因此本文选取精确率、召回率、F1 分数和 F2 分数作为分类的评价指标。在审计工作中,重要的是尽可能将违规样本识别出来,因此召回率与 F2 分数更为关键。由于本文将全部样本划分成到五个时间窗口,并在每个时间窗口上都建立违规识别模型,在每个窗口后一年上进行预测得到模型预测结果,因此下文中的精确率、召回率、F1 分数和 F2 分数都是取五个时间窗口上的平均值。

(四)主题指标与财务指标预测效果对比

为了验证本文构建的主题指标的有效性以及假设 H2,本文先将主题指标 *Topic* 和财务指标 *F-score* 进行对比。除此之外,本文还将主题指标和财务指标进行合并,同时作为模型的解释变量,来研究主题指标是否可以弥补财务指标中的信息损失,使基于合并指标的模型与基于财务指标的模型相比,预测效果有进一步的提高。因此本文将对 *Topic*、*F-score*、*Topic* + *F-score* 三种指标体系的预测效果进行比较,并分别构建多种机器学习模型进行对比分析。基于不同指标的模型预测结果如表 8 所示。

**表 8　主题指标和财务指标下的预测效果对比**

| 指标体系 | 评估指标 | Logistic | KNN | SVM | RF | AdaBoost | MLP |
|---|---|---|---|---|---|---|---|
| *F-score* | 精确度 | 52.40% | 49.44% | 53.81% | 51.31% | 57.68% | 52.81% |
| | 召回率 | 73.84% | 59.41% | 65.59% | 58.05% | 53.37% | 63.61% |
| | F1 分数 | 57.86% | 50.17% | 58.47% | 49.96% | 53.93% | 51.42% |
| | F2 分数 | 65.40% | 54.13% | 62.29% | 53.32% | 53.31% | 57.12% |
| *Topic* | 精确度 | 46.44% | 53.46% | 49.84% | 50.79% | 59.26% | 65.80% |
| | 召回率 | 67.00% | 63.72% | 59.09% | 53.44% | 45.41% | 65.80% |
| | F1 分数 | 53.89% | 56.34% | 52.45% | 47.34% | 63.17% | 58.03% |
| | F2 分数 | 60.63% | 60.19% | 55.78% | 49.52% | 50.82% | 59.61% |
| *Topic + F-score* | 精确度 | 60.80% | 48.11% | 61.09% | 48.74% | 62.10% | 66.45% |
| | 召回率 | 64.29% | 76.53% | 68.13% | 65.90% | 60.08% | 68.22% |
| | F1 分数 | 59.51% | 58.10% | 63.34% | 54.05% | 52.73% | 59.26% |
| | F2 分数 | 61.30% | 67.44% | 65.88% | 59.78% | 54.53% | 62.97% |

在违规识别中,我们着重关注模型的召回率以及 F2 分数。从实验结果中可以看出,当采用单一财务指标作为输入指标时,逻辑回归模型 Logistic 的召回率和 F2 分数最高,分别为 73.84% 和 65.40%,其次是支持向量机 *SVM* 和多层感知器 *MLP*,召回率分别达到 65.59% 和 63.61%。且与单一财务指标相比,基于单一主题指标的违规识别模型召回率普遍较低,但其中基于主题指标的多层感知机 *MLP* 的精确率和召回率都较高,说明基于主题指标的 *MLP* 模型可以即准确又尽可能多地识别出上市公司违规。除此之外,基于单一主题指标的逻辑回归模型 Logistic 和 K - 近邻模型 *KNN* 的召回率和 F2 分数也较高。为了探究主题指标的加入是否可以提升财务指标

的识别性能,我们着重对比财务指标与主题指标 + 财务指标的召回率和 F2 分数,发现与单一财务指标相比,除逻辑回归模型 Logistic 外,其余五个模型基于合并指标的召回率和 F2 分数都有较大提升,其中 *K* - 近邻模型 *KNN* 基于合并指标的召回率提升最为明显,相较于单一财务指标,召回率提高了 17.12%。实验结果表明,主题指标可以弥补财务指标的不足,提升了财务指标的违规识别性能。

(五)主题指标与文本特征指标预测效果对比

为了进一步验证假设 H2,本文将主题指标 *Topic* 和文本特征指标 *Style* 的预测效果进行对比。同样地,本文也构建主题指标和文本特征指标的合并指标 *Topic* + *Style* 来探究主题模型是否可以弥补文本特征模型中缺少的语义信息,使得合并指标的预测效果优于文本特征指标的预测效果。基于不同指标的模型预测结果如表 9 所示。

**表 9　主题指标与文本特征指标下的预测效果对比**

| 指标体系 | 评估指标 | Logistic | KNN | SVM | RF | AdaBoost | MLP |
|---|---|---|---|---|---|---|---|
| *Style* | 精确度 | 57.67% | 49.21% | 44.29% | 55.02% | 53.54% | 55.66% |
| | 召回率 | 63.76% | 47.66% | 51.50% | 68.77% | 62.36% | 60.32% |
| | F1 分数 | 54.99% | 46.30% | 47.15% | 51.06% | 50.14% | 52.71% |
| | F2 分数 | 58.11% | 46.71% | 49.53% | 58.05% | 55.54% | 56.23% |
| *Topic* | 精确度 | 46.44% | 53.46% | 49.84% | 50.79% | 45.41% | 65.80% |
| | 召回率 | 67.00% | 63.72% | 59.09% | 53.44% | 63.17% | 61.96% |
| | F1 分数 | 53.89% | 56.34% | 52.45% | 47.34% | 50.82% | 58.03% |
| | F2 分数 | 60.63% | 60.19% | 55.78% | 49.52% | 56.78% | 59.61% |
| *Topic + Style* | 精确度 | 53.54% | 43.57% | 53.86% | 57.42% | 54.55% | 64.53% |
| | 召回率 | 63.31% | 65.71% | 55.94% | 73.33% | 69.91% | 68.72% |
| | F1 分数 | 54.06% | 51.46% | 51.93% | 56.02% | 52.39% | 59.18% |
| | F2 分数 | 57.61% | 58.77% | 53.11% | 64.52% | 59.76% | 63.33% |

从实验结果中可以看出,当采用单一文本特征指标进行违规识别时,随机森林 RF 的召回率最高,达到了 68.77%,其次是逻辑回归 Logistic 和多层感知器 MLP,召回率分别为 63.76%、60.32%。通过对比单一文本特征指标和主题指标的识别性能可以发现,除随机森林 RF 外,其余五个模型基于主题指标的召回率都高于基于文本特征指标的召回率,说明在利用主题指标进行识别违规时,集成学习模型可能并不适用。为了探究主题指标的加入是否可以提升文本特征指标的识别性能,我们着重对比文本特征指标与主题指标 + 文本特征指标的召回率和 F2 分数。通过对比发现,相较于单一的文本特征指标,除逻辑回归模型 Logistic 外,其余五个机器学习模型基于合并指标的召回率都有较大提升,其中提升最多的是 K - 近邻模型 KNN,相较于单一文本特征指标,召回率提升了 18.05%。实验结果表明,主题指标可以弥补文本特征指标的不足,提升文本特征指标的违规识别率。

## 六、研究结论

本文基于 A 股上市银行年度报告的文本数据和相关财务数据,构建了财务指标、文本特征指标,并运用 LDA 主题模型对年报文本建模构建主题指标,并在不同指标下分别建立机器学习预测模型,以发现上市银行是否存在违规行为。研究发现:第一,基于年报文本所构建的主题指标可有效预测上市银行违规。第二,本文提取出与违规显著相关的主题指标后发现,可用于识别违规的主题并不是一成不变的,呈现出迭代更新的现象。第三,将主题指标与财务指标、文本特征指标合并后共同构建的违规模型的预测效果优于仅使用财务指标、文本特征指标构建的违规识别模型,说明主题指标可提供财务指标和文本特征指标中所缺少的语义信息,能够提升财务指标和文本特征指标违规识别的性能。

**参考文献:**


[1]Dechow P M,Ge W,Larson C R,et al. Predicting material accounting misstatements[J]. Contemporary accounting research,2011,28(1):17 - 82.

[2]钱苹,罗玫. 中国上市公司财务造假预测模型[J]. 会计研究,2015(7):18 - 25.

[3]杨贵军,周亚梦,孙玲莉,等. 基于 Benford 律的 Logistic 模型及其在财务舞弊识别中的应用[J]. 统计与信息论坛,2019(8):50 - 56.

[4]Brown N C,Crowley R M,Elliott W B. What are you saying? Using topic to detect financial misreporting[J]. Journal of Accounting Research,2020,58(1):237 - 291.

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

[6]Purda L,Skillicorn D. Accounting variables,deception,and a bag of words:Assessing the tools of fraud detection[J]. Contemporary Accounting Research,2015,32(3):1193 - 1223.

[7]黄志刚,刘佳进,林朝颖. 基于机器学习的上市公司财报舞弊识别前沿方法比较研究[J]. 系统科学与数学,2020(10):1882 - 1900.

[8]赵纳晖,张天洋. 基于 MD&A 文本和深度学习模型的财务报告舞弊识别[J]. 会计之友,2022(8):140 - 149.

[9]程新生,谭有超,刘建梅. 非财务信息、外部融资与投资效率——基于外部制度约束的研究[J]. 管理世界,2012(7):137 - 150.
[10]Hoberg G,Lewis C. Do fraudulent firms produce abnormal disclosure? [J]. Journal of Corporate Finance,2017,43(4):58 - 85.
[11]Goel S,Gangolly J. Beyond the numbers:Mining the annual reports for hidden cues indicative of financial statement fraud[J]. Intelligent Systems in Accounting,Finance and Management,2012,19(2):75 - 89.
[12]崔雪. 基于情感分析的上市公司年报舞弊预警研究[D]. 哈尔滨:东北林业大学,2020.
[13]Blei D M,Ng A Y,Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research,2003,3(5):993 - 1022.
[14]关鹏,王曰芬,傅柱. 不同语料下基于 LDA 主题模型的科学文献主题抽取效果分析[J]. 图书情报工作,2016(2):112 - 121.
[15]贺亮,李芳. 科技文献话题演化研究[J]. 现代图书情报技术,2012(4):61 - 67.
[16]李梦杰,刘建国,郭强,等. 基于文本挖掘的互联网教育课程主题发现与聚类研究[J]. 上海理工大学学报,2018(3):259 - 266.
[17]荀静. 基于 LDA 模型的微信图书馆热点话题检测[J]. 软件工程与应用,2017(6):145.
[18]杨金庆,吴乐艳. 中美科技文献时滞差异及主题演化对比分析研究——以农业领域为例[J]. 情报理论与实践,2021(8):167 - 172.
[19]Mahajan A,Dey L,Haque S M. Mining financial news for major events and their impacts on the market[C]. 2008 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology,2008,1(1):423 - 426.
[20]傅魁,鲁冬,覃桂双. 基于 SGC-LDA 模型的财经文本主题研究[J]. 计算机工程与应用,2022(15):285 - 283.
[21]龙文,毛元丰,管利静,等. 财经新闻的话题会影响股票收益率吗? ——基于行业板块的研究[J]. 管理评论,2019(5):18 - 27.
[22]Gray G L,Debreceny R S. A taxonomy to guide research on the application of data mining to fraud detection in financial statement audits[J]. International Journal of Accounting Information Systems,2014,15(4):357 - 380.
[23]文炳洲,焦少杰. 利益驱使、中介背书与上市公司财务舞弊——基于 2008—2017 年证监会处罚公告书[J]. 财会通讯,2020(23):96 - 100.
[24]Loughran T,McDonald B. Textual analysis in accounting and finance:A survey[J]. Journal of Accounting Research,2016,54(4):1187 - 1230.
[25]Bushee B J,Gow I D,Taylor D J. Linguistic complexity in firm disclosures:Obfuscation or information? [J]. Journal of Accounting Research,2018,56(1):85 - 121.
[26]Douglas K M,Sutton R M. Effects of communication goals and expectancies on language abstraction[J]. Journal of Personality and Social Psychology,2003,84(4):682.
[27]王泽贤. 主题模型在财经文本主题演化中的应用[D]. 厦门:厦门大学,2018.

[责任编辑:杨志辉]

# Fraud Detection of Listed Companies Based on LDA Topic Model: An Empirical Study of China A Share Listed Banks

ZHANG Yi, XU Yang, LI Weiping
(School of Information Engineering, Nanjing Audit University, Nanjing 211815, China)

**Abstract**: This paper takes the annual reports of A-share listed banks in China from 2010 to 2019 as the research sample, by using the LDA topic model to deeply mine the semantic information of Chinese annual reports and construct the topic measure of the banks' annual reports, and compare the performance of topic measure with commonly used financial measure, text feature measure and their combined measure with topic measure in detecting frauds of listed banks on a variety of machine learning models. This paper found that the topic content of the Chinese annual report has a certain predictive effect on the frauds of listed banks, and compared with a single traditional indicator, the topic measure can improve the fraud detection performance of the traditional indicators. The results of the study provide direct evidence for the effectiveness of using annual report topic content information and machine learning methods to detect listed banks' frauds, build a more effective fraud detection measure system for the Chinese market, and find a more efficient method for auditors, which is conducive to further avoiding and preventing audit risks.

**Key Words**: fraud detection of listed companies; annual report; LDA topic model; machine learning; fraud prediction; financial statements