

上市公司信息披露的文字主观叙述与盈余操纵

——基于机器学习法提取主观模式的研究

毛怡欣

(厦门大学 管理学院, 福建 厦门 361005)

内容提要: 选取 2006–2016 年沪深 A 股上市公司样本, 研究上市公司是否会操纵文字信息披露, 以更主观的方式叙述来配合上市公司的盈余操纵行为。利用机器学习法提取 2-POS 主观模式, 构建了可量化的文字叙述的主观性指标, 通过实证检验发现, 处于财务重述前, 即在盈余操纵隐藏期间, 上市公司年报中使用的主观句比例更高、整体主观性分数也更高; 进一步的研究发现, 良好的外部市场环境和内部治理环境会抑制上市公司利用主观色彩更浓的信息叙述来配合其盈余操纵的行为。本文的研究既为上市公司操纵文字信息披露的动机提供了新的证据, 补充了财务数据与文字信息披露关系的相关研究, 也为监管部门加强信息披露监管以及外部信息使用者判断信息质量提供了新的实证依据。

关键词: 文字叙述; 主观性; 盈余操纵

中图分类号: F275 **文献标识码:** A **文章编号:** 1001-148X (2019) 01-0117-10

DOI:10.13902/j.cnki.syyj.2019.01.014

一、引言

公司年报中有大量的文字叙述, 对财务状况和其他重要的非财务状况进行说明和解释, 这些文字叙述是非常重要的展示上市公司的方式 (Li, 2010)。虽然监管部门要求上市公司进行信息披露时必须保证其内容客观真实, 但是上市公司在进行相关事实的文字叙述时, 叙事者是难以跳脱出自身的认知模式、完全不带任何自身的态度和感情色彩的 (沈家煊, 2001), 并且由于语言的复杂性, 相关监管部门也难以制定出系统的、可量化的标准来对上市公司的相关文字叙述进行统一的规范。因此, 上市公司年报的文字叙述很难达到完全客观、不带任何上市公司自身的立场和感情色彩 (Baginski 等, 2016)。尽管上市公司在年报的文字叙述难以摆脱主观性印记, 主观性在各上市公司年报中普遍存在, 但是各公司年报主观性程度是不同的。有的上市公司在叙事时尽量的还原事实, 使文字叙述成为向外部人员传递信息的重要渠道, 从而使外部投资者掌握更多有效信息做出更准确的决策; 但是有的上市公司会利用更主观的叙述来模糊或转移焦点, 故意夸大对上市公司有利的

事实以及故意削弱对上市公司不利的事实, 误导外部信息使用者。上市公司的财务数据与文字信息是相辅相成的, 上市公司披露的相关财务数据需要利用文字进行进一步说明和解释, 管理状况、发展前景等其他重要的非财务状况的文字叙述又是财务信息的重要支撑, 所以, 文字信息和财务数据不是相互独立的 (Loughran 和 McDonald, 2016)。那么上市公司若进行了盈余操纵, 就会有强烈的动机操纵文字叙述, 对相关事实进行更主观的叙述, 运用词汇和句法等方式削弱某些事实的真实性, 以此来配合盈余操纵, 以避免相关操纵行为被外部信息使用者发现。因此, 构建一套可量化的指标衡量上市公司年报文字的主观程度, 对上市公司信息披露进行实证检验是有必要的。本文通过机器学习法对年报语料进行训练提取 2-POS 主观模式对年报文字叙述的主观性特征进行度量后, 对上市公司文字叙述是否配合盈余操纵进行的实证检验。

二、文献回顾和研究假设

(一) 文本挖掘方法的使用及研究

近年来, 随着机器学习、自然语言处理技术的

收稿日期: 2018-08-01

作者简介: 毛怡欣 (1989-), 女, 四川崇州人, 厦门大学管理学院博士研究生, 研究方向: 公司治理与财务。

发展,文本挖掘方法越来越多地被应用到对上市公司会计文本的定量分析中。其中关于对文本主观性的研究起始于对英文语料的研究,早期的研究认为形容词的使用是判断主观性的重要特征,因此利用标识形容词来计算文本的主观性(Hatzivassiloglou和Wiebe,2000),这也激发了学者对其他词性与主观性关系的研究。Riloff等(2003)发现名词也与语句的主观性存在很强的相关性,通过Bootstrapping算法得出1052个主观性的名词,然后将这些主观性名词作为判断语句主观性的特征。随后,有学者在前人的基础上又突破了仅以单个的词语作为标识语句主观性的依据,叶强等(2007)基于中文的语料提出了利用连续双词词类模式(2-POS模式)作为判断语句主观性的依据。目前国内已经有较多学者将2-POS模式进一步应用到对中文的主观性判断中,并且验证了利用2-POS模式判断中文主观性的可行性。张文文和王挺(2013)在比较了6种不同的主客观分类方法后,发现基于2-POS模式的方法效果是最好的。虽然大多关于文本主观性研究的语料是基于互联网的评论、微博和新闻等,这些语料与上市公司披露的会计文本的语料有一定的差异。但段钊等(2017)也将2-POS模式应用到上市公司的社会责任报告的文本分析中,发现利用2-POS主观模式对主观句的查全率和查准率可以达到90%以上,这表明了利用2-POS模式对上市公司披露的会计文本进行主观性分析是可行的。

(二) 文字主观性与盈余操纵

许多研究显示上市公司对文字信息披露进行的策略性管理更多的是为了转移外部信息使用人员对不利消息的关注或模糊上市公司的不利消息,即出于信息模糊动机而进行文字信息披露的操纵。例如,Bonsall等(2013)发现当上市公司面临的不确定越高,就会披露更多与盈余无关的前瞻性信息;Merkley(2014)也发现当上市公司的业绩较差时会披露更多关于研发的信息,并且当公司绩效较差时,上市公司年报的可读性也会越差。上市公司年报中的财务数据和文字信息之间并不是相互独立的,财务数据由准确的数字构成可以提供更为客观的公司状况,年报中的文字叙述可以反映结构化的财务数据无法反映的更多细节信息,例如财务数据相关的会计准则和管理者动机,外部信息使用者在了解了相关会计准则后,可以更准确地理解财务数据的含义,上市公司对外公告的财务数据与文字叙述之间是相互联系和相互印

证的。相关的研究也显示了上市公司的会计文本信息对财务数据起到补充作用而不是替代作用,例如Bochkay和Levine(2013)发现结合财务数据和管理层讨论与分析的文本信息才可以做出更准确的盈余预测,上市公司的文本信息也会影响财务数据与市场反应之间的关系。因此,为了隐藏上市公司的盈余操纵行为,上市公司就会有强烈的动机采取对文字信息也进行操纵的手段,提供更模糊的相关事实的信息以避免相关盈余操纵行为被外部人员发现。虽然利用文字对相关事实进行叙述时难以避免叙述者所带的立场和感情从而难以达到绝对客观(沈家煊,2001),上市公司年报中的文字叙述可能普遍存在着一定程度上的主观性,但是各上市公司的文字叙述所带的主观色彩程度并不是完全一致的,其中也大量存在有主观目的性的。

对前期财务报告中的财务数据进行更正的财务报告重述(即“财务重述”)行为经常性地一定程度上隐藏了上市公司的盈余操纵行为(Richardson等,2002;Lev,2003;周晓苏和周琦,2011)。通过以我国上市公司为研究样本的实证检验发现,我国上市公司财务重述泛滥的主要原因就在于上市公司为了掩盖其不良经营状况进行的盈余操纵(于鹏,2007);上市公司及其管理层盈余操纵行为一旦实施,财务重述就会随即发生(Lo等,2017)。魏志华等(2009)通过梳理我国上市公司年报重述情况后也认为监管部门需要重点关注上市公司的财务重述中可能存在的盈余操纵。许多学者也在实证研究中把财务重述作为上市公司进行盈余操纵的表征,例如,程新生等(2015)利用财务重述作为上市公司盈余操纵的表征,实证检验盈余操纵与MD&A中的非财务信息披露之间的配合;张璇等(2016)也把财务重述作为上市公司盈余质量的代理变量。因此,本文选择在财务信息披露以后进行过财务重述,但该年报还未被重述,公司进行了盈余操纵但还处于隐藏期间的年报作为样本,来研究上市公司是否会在该财务报告中以更主观的文字叙述来配合其盈余操纵行为。若上市公司的年报处于财务重述前的状态,这时上市公司处于盈余操纵隐藏期间,上市公司是具有为了隐藏和配合其盈余操纵行为,策略性披露更主观的文字叙述的动机的。因此,本文提出了以下假设。

假设1:相较于其他没有进行过财务重述的上市公司,进行过财务重述的上市公司在财务重述

公告前,即在隐藏期间,其尚未被更正的年报中会使用更多的主观句式。

假设2:相较于其他没有进行过财务重述的上市公司,进行过财务重述的上市公司在财务重述公告前,即在隐藏期间,其尚未被更正的年报文字部分的整体主观性更高。

三、研究设计

(一) 样本选择和数据来源

本文选取2006-2016年沪深两市的A股上市公司作为研究样本,剔除了金融行业样本、上市时间不超过两年的样本以及相关变量缺失的样本后,一共得到了11715个样本数据,其中在年报披露后进行过财务重述但当前处于还未被重述的状态的年报样本共有655个。由于在进行进一步分析时涉及了不同变量,导致变量缺失情况不同,因此样本数略有不同。上市公司的年报数据来自于Wind数据库,财务重述公告数据来自于国泰安数据库,内部控制指数数据来自于迪博数据库,其他相关变量数据来自于锐思数据库。为了消除极端异常值的影响,本文对所有连续变量都进行了1%和99%水平下的Winsorize缩尾处理。本文的行业分类标准为证监会门类行业。

(二) 变量度量

1. 被解释变量。本文的被解释变量为主观句占比(Sub_1)和主观性分数(Sub_2)。本文采用2-POS模式(将句子按照词性进行分词,2个连续词的顺序组合就构成1个2-POS模式)来对语句主观性进行度量。提取2-POS主观模式以及计算主观句占比(Sub_1)、主观性分数(Sub_2)这两个变量的具体步骤如下所示:

第一步,在上市公司年报中的文本中随机抽取2000个句子,请3位志愿者分别对这2000个句子属于主观句还是客观句进行标注,若有两位志愿者以上将语句标注为主观句,那么将该句划分为主观句,否则划分为客观句,一共得到744句主观句和1256句客观句,再随机抽取500句主观句和500句客观句,构建训练年报语料的样本。

第二步,利用LTP(Language Technology Platform)分词系统对训练样本进行分词和词性的标注,将所有主观句和客观句中的按顺序组合的连续双词模式(2-POS模式)提取出来,计算所有主观句中的2-POS模式的 χ^2 值并排序,选取前20个2-POS模式作为2-POS主观模式。 χ^2 值的计算公式如下:

$$\chi^2 = \frac{N \times (A \times D - B \times C)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)} \quad (1)$$

其中, N 表示主观句和客观句总数, A 表示主观句中包含该2-POS模式的句数, B 表示客观句中不包含该2-POS模式的句数, C 表示主观句中不包含该2-POS模式的句数, D 表示客观句中不包含该2-POS模式的句数。

第三步,计算前20个2-POS主观模式的查准率和查全率,即按照利用该2-POS模式判断出的主观句的准确率和比例作为该2-POS模式的查准率和查全率,具体计算公式为:

$$\text{查准率} = \frac{\text{识别出的准确的主观句数量}}{\text{识别出的总主观句数量}} \quad (2)$$

$$\text{查全率} = \frac{\text{识别出的准确的主观句数量}}{\text{语料库中的总主观句数量}} \quad (3)$$

将2-POS主观模式的查准率作为主观性分数的权重。表1报告了提取的20个2-POS主观模式的查准率和查全率。

表1 2-POS主观模式

编号	2-POS 主观模式	首词	尾词	查准率	查全率
1	D-U	副词	助词	0.85	0.08
2	U1-D	助词(的)	副词	0.82	0.07
3	U-A	助词	形容词	0.81	0.06
4	V-M	动词	量词	0.80	0.68
5	D-A	副词	形容词	0.78	0.09
6	A-U1	形容词	助词(的)	0.77	0.11
7	V-U	动词	助词	0.76	0.29
8	D-N	副词	名词	0.74	0.09
9	A-A	形容词	形容词	0.73	0.05
10	R-U1	代词	助词(的)	0.72	0.06
11	V-D	动词	副词	0.71	0.11
12	R-V	代词	动词	0.68	0.14
13	C-R	连词	代词	0.66	0.06
14	V-A	动词	形容词	0.64	0.18
15	U1-V	助词(的)	动词	0.63	0.28
16	R-N	代词	名词	0.60	0.25
17	V-R	动词	代词	0.58	0.13
18	C-D	连词	副词	0.58	0.03
19	V-V	动词	动词	0.53	0.64
20	D-V	副词	动词	0.49	0.32

第四步,利用LTP(Language Technology Platform)分词系统将所有样本的年报进行分词和词性标注后,提取每一份年报中每句话中包含的2-POS模式,然后标识年报中包含的2-POS主观模式,再加总每句话中各自包含的2-POS主观模式与其主观权重乘积。当按主观和客观进行分类设置阈值时,发现将阈值设置为0时,主客观分类的查准率和查全率都达到90%以上,因此本文将判断主客观句的阈值设置为0,即加总该句话中包含的2-POS主观模式与其主观权重乘积大于0,那么将该句分类为主观句,否则将该句分类为客观句。

因此可计算出年报中的主观句占比 (Sub_1) 和主观性分数 (Sub_2) , Sub_1 和 Sub_2 的具体计算公式分别如下所示:

$$Sub_1 = \frac{sub_count}{(ob_count + sub_count)} \quad (4)$$

$$Sub_2 = \frac{\sum n_i \times weight_i}{2 - POS_count} \quad (5)$$

公式 (4) 中的 sub_count 表示年报中包含的主观句数 , ob_count 表示年报中包含的客观句数; 公式 (5) 中的 n_i 表示年报中包含的第 i 个 2-POS 主观模式的频数 , $weight_i$ 表示第 i 个 2-POS 主观模式的主观权重 , $2 - POS_count$ 表示包含的所有 2-POS 模式个数。

2. 解释变量。本文的解释变量为财务重述前

($Restate_pre$) ,若上市公司在披露年报之后对该年报进行了财务重述,那么在该年报还未进行财务重述之时,变量取值为 1,否则取值为 0。在经过处理后,一共得到了 655 个财务重述前的样本。

3. 控制变量。本文选取了公司规模 ($Size$)、资产负债率 (Lev)、总资产报酬率 (ROA)、公司成长性 ($Growth$)、分析师关注度 ($Follow$)、机构持股比 (Ins)、董事会规模 (BDS)、独董比例 ($Indep$)、是否国有控股 ($State$) 这些对上市公司信息披露可能会有显著影响的因素作为控制变量。除此以外,为了控制上市公司年报披露在不同年份和不同行业之间的差异,还控制了行业虚拟变量 ($Industry$) 和年份虚拟变量 ($Year$)。表 2 报告了本文涉及的主要变量的定义和计算方法。

表 2 变量定义以及计算方法

变量符号	变量名称	变量定义和计算
Sub_1	主观句占比	年报中文字描述使用的主观句所占主观句和客观句之和的比例
Sub_2	主观性分数	利用年报中包含的 2-POS 主观模式,评估年报整体主观性分数
$Restate_pre$	财务重述前	以披露的年报是否进行过财务重述并且是否处于还未被重述的状态设置虚拟变量,若披露之后进行过重述但当前还处于未被重述的状态,取值为 1,否则取值为 0
$Size$	公司规模	年末公司总资产进行对数化处理
Lev	资产负债率	负债总额 / 资产总额
ROA	总资产报酬率	净利润 / 资产总额
$Growth$	公司成长性	营业总收入增长率 (本年营业总收入 - 上一年营业总收入) / 上一年营业总收入
$Follow$	分析师关注度	分析师跟踪人数
Ins	机构持股比	公司所有机构股东所持股份比例之和
BDS	董事会规模	董事会人数
$Indep$	独董比例	独立董事占董事人数比例
$State$	是否国有	以上市公司是否为国有控股设置虚拟变量,若为国有控股,取值为 1,否则取值为 0
$Industry$	行业	以上市公司所处证监会门类行业设置虚拟变量
$Year$	年份	以上市公司所在年份设置虚拟变量

(三) 模型设定

为了检验假设 1 和假设 2,构建了如下回归模型:

$$Sub_1_{i,t} = \alpha_0 + \beta_1 Restate_pre_{i,t} + \beta_2 Size_{i,t} + \beta_3 Lev_{i,t} + \beta_4 ROA_{i,t} + \beta_5 Growth_{i,t} + \beta_6 Follow_{i,t} + \beta_7 Ins_{i,t} + \beta_8 BDS_{i,t} + \beta_9 Indep_{i,t} + \beta_{10} State_{i,t} + \sum Industry + \sum Year + \varepsilon_{i,t} \quad (I)$$

$$Sub_2 = \alpha_0 + \beta_1 Restate_pre_{i,t} + \beta_2 Size_{i,t} + \beta_3 Lev_{i,t} + \beta_4 ROA_{i,t} + \beta_5 Growth_{i,t} + \beta_6 Follow_{i,t} + \beta_7 Ins_{i,t} + \beta_8 BDS_{i,t} + \beta_9 Indep_{i,t} + \beta_{10} State_{i,t} + \sum Industry + \sum Year + \varepsilon_{i,t} \quad (II)$$

分别以 Sub_1 和 Sub_2 作为回归模型 (I) 和回归模型 (II) 的被解释变量; 回归模型 (I) 和回归模型 (II) 的被解释变量为 $Restate_pre$, 控制变量为 $Size$ 、 Lev 、 ROA 、 $Growth$ 、 $Follow$ 、 Ins 、 BDS 、 $Indep$ 、 $State$; 采用行业和年份固定效应模型; 回归模型中各变量的具体定义和计算方法如表 2 所示。若回归模型 (I) 和回归模型 (II) 中解释变量 $Restate_pre$ 的回归系数 β_1 显著为正,

那么假设 1 和假设 2 成立。为了解决内生性问题以及进行稳健性检验,本文进一步采用倾向性评分匹配 (PSM) 方法为处于财务重述前的样本进行 1:1 匹配选取控制组样本,利用匹配后的样本再次对回归模型 (I) 和回归模型 (II) 进行回归,以及构建双重差分模型进行检验,除此以外还采用修正 Jones 模型计算上市公司的盈余管理作为盈余操纵代理变量重新进行回归。

四、实证结果分析

(一) 主要变量描述性统计和相关性分析

表 3 报告了本文涉及的主要变量的描述性统计结果。其中:上市公司年报中的主观句占比的平均值为 0.3444,表示年报中的语句中平均有大约 34% 比例的为主观句;年报的整体得分平均为 0.1157,而主观性分数最低为 0.088,最高得分为 0.1831。从主观句占比和主观性分数的描述性统计结果来看,最小值都大于 0 并且都没有趋近于 0,

这表明上市公司在年报中的文字叙述中普遍都带有主观性色彩，但是不同上市公司的年报主观性程度是有差异的；财务重述前的平均值为 0.0559，这表示约有 5.6% 的上市公司年报处于盈余操纵隐藏期间。根据是否国有的平均值来看，样本中国有控股公司的比例接近 30%。其他公司特征在不同上市公司之间也存在着一定的差异，资产负债率、总资产报酬率、公司成长性、分析师关注都有较大的差异；资产负债率最小值仅有 4.78%，而最大值达到了 107.44%；总资产报酬率最低为 -7.87%，最高为 19.85%；公司成长性最小值为 -0.616，而成长性最高的公司的营业总收入增长近 2 倍；有的上市公司没有分析师关注，而有的上市公司多达 36 位分析师对其进行跟踪；上市公司最低机构持股比例为 0，最高机构持股比例达 84.25%。

表 4 报告了本文涉及的主要变量之间的 Pearson 相关系数。其中：年报的主观句比例与主观性分数之间的相关系数为 0.7407，在 1% 水平下显著，主观句比例和主观性分数有较强的正相关关系；财务重述前与主观句比例、主观性分数的相关

系数分别为 0.0905 和 0.0836，并且在 1% 水平下显著，解释变量与被解释变量之间都有显著的正相关关系。支持假设 1 和假设 2。解释变量与其他控制变量之间的相关系数都不超过 0.3，虽然在控制变量中，公司规模与分析师关注度、公司规模与机构持股比、分析师关注度与机构持股比、董事会规模与独董比例的相关系数超过了 0.3，但在进行回归检验时进行了多重共线性检验，方差膨胀因子 (vif 值) 都小于 5，因此，本文的回归模型不存在多重共线性问题。

表 3 主要变量描述性统计结果

变量名	样本数	平均值	标准差	最小值	中位数	最大值
Sub_1	11715	0.3444	0.0385	0.2416	0.3443	0.4391
Sub_2	11715	0.1157	0.0140	0.0880	0.1144	0.1831
Restate_pre	11715	0.0559	0.2298	0	0	1
Size	11715	21.4324	1.2474	18.8362	21.3590	25.0083
Lev	11715	0.4463	0.2147	0.0478	0.4418	1.0744
ROA	11715	0.0422	0.0417	-0.0787	0.0355	0.1985
Growth	11715	0.1712	0.3568	-0.6160	0.1213	1.9825
Follow	11715	5.8540	8.0986	0	2	36
Ins	11715	0.2786	0.2504	0.0000	0.2407	0.8425
BDS	11715	8.7132	1.7130	5	9	15
Indep	11715	0.3666	0.0506	0.3077	0.3333	0.5714
State	11715	0.2976	0.4572	0	0	1

表 4 主要变量 Pearson 相关系数

	Sub_1	Sub_2	Restate_pre	Size	Lev	ROA	Growth	Follow	Ins	BDS	Indep	State
Sub_1	1											
Sub_2	0.7407***	1										
Restate_pre	0.0905***	0.0836***	1									
Size	0.1679***	0.1864***	0.1538***	1								
Lev	-0.0036	-0.0099	0.0903***	0.2562***	1							
ROA	-0.1159***	-0.1254***	-0.0931***	-0.2373***	-0.2049***	1						
Growth	-0.0668***	-0.0718***	-0.0401***	0.0142	0.0353***	0.2737***	1					
Follow	-0.0026	0.0101	0.0046	0.4268***	-0.1344***	0.1099***	0.0908***	1				
Ins	0.1212***	0.1309***	0.1261***	0.5111***	0.0292***	-0.1850***	-0.0177*	0.4124***	1			
BDS	-0.0189*	-0.0216**	0.0434***	0.2062***	0.1044***	0.0155	0.0130	0.1100***	0.1252***	1		
Indep	0.0333*	0.0282*	0.0326*	0.0851*	-0.0201*	-0.1040*	-0.0291	0.0632*	0.0607*	-0.4547*	1	
State	0.0556***	0.0462***	0.1399***	0.2927***	0.1726***	-0.1525**	-0.0741***	0.0357***	0.2817***	0.2095***	-0.0269**	1

注：* 表示在 10% 水平下显著，** 表示在 5% 水平下显著，*** 表示在 1% 水平下显著 (下同)。

表 5 主要变量是否处于财务重述前的差异

变量名	非财务重述前样本		财务重述前样本		均值差 (t 检验)
	样本数	平均值	样本数	平均值	
Sub_1	11060	0.3436	655	0.3587	-0.0152***
Sub_2	11060	0.1154	655	0.1205	-0.0051***
Size	11060	21.3857	655	22.2208	-0.8351***
Lev	11060	0.4416	655	0.526	-0.0844***
ROA	11060	0.0431	655	0.0262	0.0169***
Growth	11060	0.1747	655	0.1124	0.0623***
Follow	11060	5.8449	655	6.0076	-0.1627
Ins	11060	0.2709	655	0.4083	-0.1374***
BDS	11060	8.6894	655	9.1151	-0.4257***
Indep	11060	0.3662	655	0.3734	-0.0072***
State	11060	0.2820	655	0.5603	-0.2783***

本文进一步以上市公司年报是否处于财务重述前为划分标准将样本划分为两组子样本进行了描述性统计，表 5 报告了两组子样本主要变量的样本数和平均值、两组子样本主要变量之间的均值

差以及对两组子样本主要变量之间的平均值是否有显著差异进行的 t 检验结果。根据表 5 报告的数据，除了分析师关注度以外，其他主要变量的平均值在两组样本之间都有显著的差异。其中，没有处于财务重述前的年报的主观句比例平均为 34.36%，处于财务重述前的年报的主观句比例平均为 35.87%，经过 t 检验，没有处于财务重述前的年报的主观句比例显著低于处于财务重述前的年报。没有处于财务重述前的年报的主观性分数平均为 0.1154，而处于财务重述前的年报的主观性分数平均为 0.1205，根据 t 检验结果，没有处于财务重述前的年报的主观性分数也显著低于处于财务重述前的年报。从分组描述性统计结果来看，处于财务重述前的年报的文字叙述的主观性程度更高。

表6 财务重述前与年报主观句比例、主观性分数

	Sub_1	Sub_2
<i>Restate_pre</i>	0.0059*** (4.17)	0.0014*** (2.87)
<i>Size</i>	-0.0002 (-0.44)	0.0001 (0.60)
<i>Lev</i>	0.0029* (1.66)	0.0010 (1.58)
<i>ROA</i>	-0.0100 (-1.17)	-0.0055* (-1.81)
<i>Growth</i>	-0.0012 (-1.23)	-0.0004 (-1.33)
<i>Follow</i>	-0.0009** (-2.45)	-0.0002* (-1.90)
<i>Ins</i>	-0.0002 (-0.14)	-0.0009 (-1.51)
<i>BDS</i>	0.0030 (1.47)	0.0006 (0.84)
<i>Indep</i>	0.0037 (0.51)	-0.0021 (-0.81)
<i>State</i>	0.0042*** (5.34)	0.0013*** (4.71)
<i>_cons</i>	0.3112*** (33.00)	0.0961*** (28.57)
行业	控制	控制
年份	控制	控制
<i>N</i>	11715	11715
<i>adj. R²</i>	0.2208	0.2525
<i>F</i>	90.7095	107.9668

(二) 财务重述前与年报主观性

表6报告了回归模型(I)和回归模型(II)的回归结果。其中:第2列报告了以年报主观句比例(*Sub_1*)作为被解释变量的回归结果,控制了相关影响因素后,解释变量财务重述前(*Restate_pre*)的回归系数为0.0059,在1%水平下显著,财务重述前与主观句占比显著正相关;第3列报告了以年报主观性分数(*Sub_2*)为被解释变量的回归结果,加入控制变量后,解释变量财务重述前(*Restate_pre*)的回归系数为0.0014,也在1%的水平下显著,财务重述前与主观性分数显著正相关;回归结果表明处于财务重述前状态的年报中使用的主观句的比例更高、年报的整体主观性分数也更高,上市公司在盈余操纵隐藏期间,会使用更主观的文字叙述配合其盈余操纵行为,与本文提出的假设1和假设2的预期是一致的。

(三) 上市公司外部市场环境的影响

不同地区的上市公司所面临的外部市场化环境是不同的。在市场化程度较高的地区,上市公司面临更为有效的外部治理机制,上市公司与外部人员之间的信息传递效率可能会更高,这可能会使上市公司操纵年报文字信息披露隐藏盈余操纵行为的动机减弱,因此上市公司面临的外部市场环境可能会影响上市公司年报中的主观文字叙述与盈余操纵之间的关系。为了进一步检验上市公司所处的外部市场环境是否影响盈余操纵对主观

文字叙述的影响,本文根据王小鲁等(2016)编制的中国分省份市场化指数,将样本划分为市场化程度较低(MKT-L)和市场化程度较高(MKT-H)两组,进行分组回归检验。计算各地区的市场化指数中位数,上市公司所在地区的市场化程度低于中位数,那么将其划分为市场化程度较低组,反之则将样本划分为市场化程度较高组。除了利用王小鲁等(2016)编制的市场化指数进行分组以外,东部和中西部地区之间的市场化发展程度有显著差异,因此本文还将样本以上市公司所在省份属于东部还是中西部划分为东部和中西部两组,进行分组回归检验。

表7报告了外部市场环境影响的检验结果,其中,第2列至第5列报告了市场化程度较高组和市场化程度较低组的分组回归结果,第6列至第9列报告了东部组和中西部组的分组回归结果。从第2列和第4列的回归结果来看,在市场化程度较高的情况下,是否处于财务重述前(*Restate_pre*)对主观句比例(*Sub_1*)和主观性分数(*Sub_2*)没有显著的影响;而第3列和第5列的回归结果显示在市场化程度较低的情况下,财务重述前(*Restate_pre*)对主观句比例(*Sub_1*)和主观性分数(*Sub_2*)的影响显著为正。第6列和第8列显示,在东部地区,财务重述前(*Restate_pre*)对年报中的主观句比例(*Sub_1*)和年报的主观性分数(*Sub_2*)没有显著的影响;而第7列和第9列的回归结果显示,在中西部地区,财务重述前(*Restate_pre*)与主观句比例(*Sub_1*)、年报的主观性分数(*Sub_2*)都是在1%水平下显著正相关。综合表7的分组回归结果,上市公司的外部市场环境发展程度越好,上市公司操纵年报文字叙述以掩盖盈余操纵的动机就会减弱,高程度的市场化环境可以抑制上市公司在年报中采用较多的主观叙述进行策略性的披露;而上市公司在面临较差的市场化环境时,更容易利用主观叙述配合其盈余操纵,在外部市场化程度发展较低时,盈余操纵与主观叙述之间的关系更为显著。

(四) 上市公司内部治理环境的影响

除了外部市场环境会影响上市公司的信息披露行为以外,前人的研究也显示了公司内部的治理环境也会影响上市公司信息披露行为(李慧云等,2013)。若上市公司内部治理环境良好,可能会约束管理层对信息披露的操纵,因此内部治理环境更好的上市公司利用主观性叙述掩盖盈余操纵的可能性会降低。本文选取了是否两职合一、董事会规模、

独董比例、监事会规模、前十大股东持股比例、机构持股比例这些公司治理指标，利用主成分分析法构建了上市公司的公司治理指数，将公司治理指数高于中位数的上市公司划分为公司治理环境较好的一组 (GI_H)，低于中位数的上市公司划分为公司治理环境较差的一组 (GI_L)，分别对两组样本进

行回归检验。本文还利用了迪博数据库的内部控制指数将上市公司划分为内部控制环境较好 (IC_H) 和内部控制环境较差 (IC_L) 两组样本，内部控制指数高于中位数则划分为内部控制环境较好一组，低于中位数，则划分为内部控制环境较差一组，然后也对两组样本进行分组回归检验。

表 7 外部市场环境影响

	Sub_1		Sub_2		Sub_1		Sub_2	
	MKT - H	MKT - L	MKT - H	MKT - L	东部	中西部	东部	中西部
Restate_pre	0.0023 (0.98)	0.0084*** (4.82)	0.0009 (1.06)	0.0020*** (3.09)	0.0039 (1.48)	0.0079*** (4.22)	0.0011 (1.49)	0.0020*** (2.81)
Size	-0.0006 (-1.01)	0.0012** (2.17)	0.0002 (1.04)	0.0003 (1.31)	0.0002 (0.39)	0.0007 (1.18)	0.0003 (1.46)	0.0002 (1.00)
Lev	-0.0013 (-0.49)	0.0059** (2.51)	-0.0000 (-0.01)	0.0014 (1.64)	-0.0003 (-0.13)	0.0070*** (2.66)	0.0003 (0.32)	0.0015 (1.48)
ROA	0.0006 (0.05)	-0.0266*** (-2.21)	-0.0002 (-0.04)	-0.0117*** (-2.63)	0.0068 (0.61)	-0.0382*** (-2.80)	0.0010 (0.25)	-0.0176*** (-3.38)
Growth	-0.0011 (-0.84)	-0.0017 (-1.25)	-0.0005 (-1.00)	-0.0007 (-1.45)	-0.0021* (-1.69)	0.0003 (0.20)	-0.0006 (-1.53)	-0.0003 (-0.61)
Follow	-0.0017*** (-3.37)	-0.0005 (-1.09)	-0.0006*** (-3.23)	-0.0000 (-0.03)	-0.0012*** (-2.58)	-0.0012** (-2.08)	-0.0004*** (-2.62)	-0.0002 (-0.88)
Ins	0.0027 (1.23)	0.0006 (0.26)	-0.0002 (-0.23)	-0.0003 (-0.35)	0.0030 (1.51)	-0.0014 (-0.53)	-0.0003 (-0.41)	-0.0000 (-0.02)
BDS	0.0062** (2.05)	0.0016 (0.58)	0.0022** (2.10)	-0.0003 (-0.30)	0.0076*** (2.81)	-0.0013 (-0.40)	0.0021** (2.29)	-0.0008 (-0.64)
Indep	0.0145 (1.33)	-0.0107 (-1.10)	-0.0016 (-0.42)	-0.0032 (-0.89)	0.0105 (1.08)	-0.0048 (-0.44)	-0.0017 (-0.49)	-0.0014 (-0.34)
State	0.0023 (0.98)	0.0084*** (4.82)	0.0009 (1.06)	0.0020*** (3.09)	0.0039* (1.88)	0.0079*** (4.22)	0.0011 (1.49)	0.0020*** (2.81)
_cons	0.3269*** (18.44)	0.2923*** (23.54)	0.1052*** (17.08)	0.0946*** (20.62)	0.3029*** (24.74)	0.2972*** (21.14)	0.0913*** (21.58)	0.0951*** (17.66)
行业	控制							
年份	控制							
N	6286	5429	6286	5429	7740	3975	7740	3975
adj. R ²	0.2121	0.2368	0.2604	0.2466	0.2031	0.2649	0.2538	0.2516
F	49.3533	49.1078	64.2086	51.7565	55.7772	43.1095	74.1266	40.2911

表 8 内部治理环境的影响

	Sub_1		Sub_2		Sub_1		Sub_2	
	GI - H	GI - L	GI - H	GI - L	IC_H	IC_L	IC_H	IC_L
Restate_pre	0.0044 (1.39)	0.0058*** (3.32)	0.0004 (0.48)	0.0018*** (2.92)	0.0025 (1.11)	0.0073*** (3.82)	0.0011 (1.57)	0.0016** (2.05)
Size	0.0005 (0.95)	-0.0012** (-2.12)	0.0002 (1.26)	-0.0001 (-0.65)	0.0013* (1.83)	-0.0018*** (-2.98)	0.0005** (2.01)	-0.0003 (-1.37)
Lev	0.0022 (0.80)	0.0039 (1.64)	0.0000 (0.01)	0.0018** (2.11)	0.0012 (0.33)	0.0068*** (2.88)	-0.0006 (-0.52)	0.0014* (1.66)
ROA	0.0134 (0.87)	-0.0150 (-1.34)	0.0020 (0.38)	-0.0098** (-2.38)	0.0320 (1.55)	-0.0069 (-0.49)	0.0092 (1.29)	-0.0072 (-1.43)
Growth	-0.0022 (-1.58)	-0.0006 (-0.47)	-0.0012*** (-2.62)	0.0003 (0.54)	-0.0031* (-1.93)	0.0006 (0.45)	-0.0011** (-1.99)	-0.0002 (-0.41)
Follow	-0.0012** (-2.39)	-0.0010* (-1.80)	-0.0002 (-1.08)	-0.0005** (-2.26)	-0.0022*** (-3.92)	-0.0011* (-1.89)	-0.0005** (-2.30)	-0.0003 (-1.24)
Ins	-0.0079*** (-3.31)	0.0142*** (3.39)	-0.0023*** (-2.76)	0.0022 (1.43)	-0.0046* (-1.85)	-0.0043* (-1.70)	-0.0016* (-1.87)	-0.0028*** (-3.13)
BDS	0.0020 (0.73)	0.0070** (2.21)	-0.0001 (-0.08)	0.0021* (1.79)	0.0050 (1.54)	0.0060* (1.90)	0.0020* (1.73)	0.0001 (0.08)
Indep	-0.0069 (-0.72)	0.0130 (1.11)	-0.0065** (-1.97)	0.0030 (0.70)	0.0062 (0.55)	-0.0046 (-0.41)	-0.0015 (-0.37)	-0.0047 (-1.18)
State	0.0033*** (3.22)	0.0053*** (4.15)	0.0010*** (2.84)	0.0019*** (4.08)	0.0045*** (3.69)	0.0052*** (4.42)	0.0018*** (4.28)	0.0013*** (3.15)
_cons	0.3032*** (23.06)	0.3201*** (22.18)	0.0947*** (20.86)	0.0971*** (18.26)	0.2811*** (17.47)	0.3426*** (23.52)	0.0868*** (15.46)	0.1083*** (21.03)
行业	控制	控制	控制	控制	控制	控制	控制	控制
年份	控制	控制	控制	控制	控制	控制	控制	控制
N	5857	5858	5857	5858	4480	4473	4480	4473
adj. R ²	0.2354	0.2029	0.2504	0.2487	0.2003	0.2449	0.2111	0.2581
F	49.7259	41.3044	53.8576	53.4107	32.1556	40.1919	34.2928	43.0433

表 8 报告了上市公司内部治理环境对年报主观叙述的影响。其中：第 2 列至第 5 列报告了以公司治理指数分组的回归结果，回归结果显示只有在公司治理环境较差的情况下 (GI_L)，财务重述前 (Restate_pre) 的回归系数显著为正，在公司治理环境较好的情况下 (GI_H)，财务重述前 (Restate_pre) 的回归系数并不显著。第 6 列至第 9 列中报告了以内部控制指数分组的回归结果，只有在内部控制环境较差的情况下 (IC_L)，财务重述前 (Restate_pre) 的回归系数显著为正，当上市公司的内部控制较好时，财务重述前 (Restate_pre) 对年报的主观句比例 (Sub_1) 和主观性分数 (Sub_2) 没有显著影响。因此根据表 8 报告的分组回归结果，上市公司良好的内部治理环境会抑制操纵文字信息披露配合盈余操纵的行为，而较差的内部治理环境会使上市公司操纵文字叙述，以带有更多主观色彩的描述与盈余操纵相配合。

表 9 PSM 匹配样本和 DID 模型回归结果

	PSM 匹配样本		DID 模型	
	Sub_1	Sub_2	Sub_1	Sub_2
Restate_pre	0.0110*** (5.07)	0.0026*** (3.38)		
Res			0.0073*** (4.05)	-0.0002 (-0.25)
Pre			-0.0055*** (-3.03)	-0.0021*** (-3.37)
Res* Pre			0.0049** (2.05)	0.0035*** (4.30)
Size	-0.0036*** (-3.35)	-0.0011*** (-2.81)	-0.0023*** (-3.36)	-0.0006*** (-2.58)
Lev	-0.0054 (-0.95)	-0.0011 (-0.52)	-0.0004 (-0.11)	0.0006 (0.53)
ROA	0.0195 (0.62)	0.0116 (1.03)	0.0257 (1.33)	0.0086 (1.31)
Growth	-0.0007 (-0.25)	-0.0009 (-0.90)	-0.0006 (-0.39)	-0.0003 (-0.52)
Follow	-0.0011 (-1.11)	0.0001 (0.20)	-0.0003 (-0.50)	-0.0000 (-0.18)
Ins	0.0075 (1.60)	0.0015 (0.86)	0.0014 (0.48)	-0.0004 (-0.43)
BDS	0.0083 (1.34)	0.0028 (1.29)	0.0102*** (2.65)	0.0031** (2.33)
Indep	0.0197 (0.95)	0.0003 (0.04)	0.0151 (1.12)	-0.0014 (-0.31)
State	-0.0001 (-0.05)	0.0008 (0.93)	0.0001 (0.05)	-0.0000 (-0.01)
_cons	0.3721*** (13.26)	0.1176*** (11.69)	0.3552*** (18.21)	0.1127*** (16.91)
行业	控制	控制	控制	控制
年份	控制	控制	控制	控制
N	1264	1264	3410	3410
adj. R ²	0.2174	0.1923	0.2210	0.2030
F	11.0269	9.5908	27.1324	24.4714

(五) 稳健性检验

1. 倾向性评分匹配 (PSM)。在前文的回归检验中，采用了整体混合样本进行检验，本文在稳健性检验中进一步采用倾向性评分匹配方法进行 1:1 匹配，为发生过财务重述的上市公司在没有发生过财务重述的上市公司中寻找配对样本，以配对后得到的实验组和控制组的样本重新对回归模型 (I) 和回归模型 (II) 进行回归。首先将进行过财务重述的上市公司的公司年标记为实验组 (treat = 1)，没有发生过财务重述的上市公司的公司年标记为控制组 (treat = 0)；选取的匹配变量为：行业 (Industry)、上市年份 (Age)、公司规模 (Size)、资产负债率 (Lev)、总资产报酬率 (ROA) 和营业总收入增长率 (Growth)，然后进行 Logit 回归；根据倾向性评分对每一个财务重述前的样本进行最近邻一对一匹配，最终得到 632 组 (1264 个) 配对样本。表 9 中的第 2 列和第 3 列报告了 PSM 匹配样本的回归结果，财务重述前 (Restate_pre) 的回归系数仍然显著为正，利用 PSM 方法得到匹配样本进行的稳健性检验得到的回归结果与前文的回归结果是一致的，仍然支持本文提出的假设 1 和假设 2，上市公司会在年报中采用更主观的叙述配合盈余操纵行为。

2. 双重差分模型。本文进一步设计了双重差分模型 (DID) 对内生性问题进行检验，双重差分模型如下所示：

$$Sub_1_{i,t} = \alpha_0 + \beta_1 Res_{i,t} + \beta_2 Pre_{i,t} + \beta_3 Res_{i,t} \times Pre_{i,t} + \beta_4 Size_{i,t} + \beta_5 Lev_{i,t} + \beta_6 ROA_{i,t} + \beta_7 Growth_{i,t} + \beta_8 Follow_{i,t} + \beta_9 Ins_{i,t} + \beta_{10} BDS_{i,t} + \beta_{11} Indep_{i,t} + \beta_{12} State_{i,t} + \sum Industry + \sum Year + \varepsilon_{i,t} \quad (III)$$

$$Sub_2 = \alpha_0 + \beta_1 Res_{i,t} + \beta_2 Pre_{i,t} + \beta_3 Res_{i,t} \times Pre_{i,t} + \beta_4 Size_{i,t} + \beta_5 Lev_{i,t} + \beta_6 ROA_{i,t} + \beta_7 Growth_{i,t} + \beta_8 Follow_{i,t} + \beta_9 Ins_{i,t} + \beta_{10} BDS_{i,t} + \beta_{11} Indep_{i,t} + \beta_{12} State_{i,t} + \sum Industry + \sum Year + \varepsilon_{i,t} \quad (IV)$$

在回归模型 (III) 和回归模型 (IV) 中，Res 定义为是否发生过财务重述的虚拟变量，若上市公司发生过财务重述，那么取值为 1，否则取值为 0。Pre 定义为财务重述隐藏期间的虚拟变量 (隐藏期可能大于一年)，在上市公司还未发布财务重述公告，取值 1，发布了财务重述公告后，取值 0；若是 PSM 配对样本，若处于其配对样本还未进行财务重述公告前的期间，Pre 取值 1；若处于其配对样本进行财务重述公告后的期间，Pre 则取值 0。回归模型 (III) 和回归模型 (IV) 中的 Res 和 Pre 的

交乘项的回归系数 β_3 是本文关注的双重差分的估计量,表示盈余操纵对主观句比例和主观性分数的净影响。表9的第5列和第6列报告了DID模型的回归结果,根据表中报告的数据,交乘项的回归系数都显著为正,这表明相较于没有进行过财务重述的上市公司,进行过财务重述的上市公司在进行财务重述公告前披露的年报的主观句比例更高、主观性分数也更高,这与前文的结论也是一致的。

3. 其他稳健性检验。本文还选取了盈余管理(DA)作为盈余操纵的代理变量再次进行回归检验,本文采用修正Jones模型计算上市公司的盈余管理,首先估计回归模型:

$$TA_{i,t}/A_{i,t-1} = \alpha_1 * 1/A_{i,t-1} + \alpha_2 * (\Delta REV_{i,t} - \Delta REC_{i,t})/A_{i,t-1} + \alpha_3 * PPE_{i,t}/A_{i,t-1} + \varepsilon_{i,t} \quad (V)$$

其中, $TA_{i,t}$ 为第t年净利润与经营活动净现金流之差, $A_{i,t-1}$ 为第t-1年年末的资产总额, $\Delta REV_{i,t}$ 为第t年与第t-1年营业收入之差, $\Delta REC_{i,t}$ 为第t年与第t-1年应收账款之差, $PPE_{i,t}$ 为第t年年末固定资产。将模型(V)分别按行业和年份进行回归,回归模型的残差就是上市公司的盈余管理 $DA_{i,t}$ 。由于本文不考虑盈余管理方向,因此采用绝对值 $|DA_{i,t}|$ 作为盈余管理指标。表10中的第2列和第3列报告了以 $|DA|$ 为解释变量的回归结果,从回归系数来看, $|DA|$ 的回归系数都显著为正。除此以外本文还构建了高盈余管理($|DA|_H$)虚拟变量,若上市公司的盈余管理高于中位数, $|DA|_H$ 取值1,否则取值0,表9中的第4列和第5列报告了以 $|DA|_H$ 作为解释变量的回归结果,结果显示高盈余管理与主观句比例和主观性分数显著正相关。选取上市公司的盈余管理作为代理变量进行稳健性检验结果仍然显示上市公司会利用主观性叙述配合上市公司的盈余操纵行为。

五、结论与启示

本文利用机器学习方法对上市公司年报语料进行训练,通过提取2-POS主观模式,构建了度量年报文字叙述主观程度的指标,即年报的主观句比例(Sub_1)和整体主观性分数(Sub_2)。对上市公司年报的主观性指标进行分组描述性统计后,发现处于进行财务重述前状态的上市公司年报的主观句比例、主观性分数都显著高于不处于财务重述前的上市公司年报。通过回归检验发现,年报中的主观句比例、主观性分数与年报处于财务重述前的状态是显著正相关的。经过进一步

的分组回归检验发现,在较差的外部市场环境和内部治理环境下,主观句比例、主观性分数与年报处于财务重述前的状态显著正相关;而在较好的外部市场环境和内部治理环境下,主观句比例、主观性分数与年报是否处于财务重述前之间并不显著相关。本文的实证结果显示,我国上市公司进行盈余操纵的同时,会采取主观的文字叙述策略模糊其盈余操纵行为,但是外部的高市场化环境和内部的良好治理环境会抑制上市公司利用主观文字叙述隐藏盈余操纵行为的问题。

表10 盈余管理与年报主观句比例、主观性分数

	Sub_1	Sub_2	Sub_1	Sub_2
$ DA $	0.0040** (2.06)	0.0009*		
$ DA _H$			0.0013* (1.87)	0.0001* (1.74)
Size	0.0000 (0.07)	0.0002 (1.07)	-0.0000 (-0.02)	0.0001 (0.97)
Lev	0.0030* (1.68)	0.0011* (1.77)	0.0032* (1.80)	0.0012* (1.92)
ROA	-0.0073 (-0.83)	-0.0052* (-1.66)	-0.0070 (-0.80)	-0.0051 (-1.63)
Growth	-0.0013 (-1.36)	-0.0005 (-1.41)	-0.0014 (-1.46)	-0.0005 (-1.46)
Follow	-0.0010*** (-2.77)	-0.0003** (-2.00)	-0.0010*** (-2.72)	-0.0003** (-1.97)
Ins	-0.0003 (-0.19)	-0.0009 (-1.64)	-0.0002 (-0.14)	-0.0009 (-1.61)
BDS	0.0030 (1.45)	0.0005 (0.67)	0.0031 (1.48)	0.0005 (0.67)
Indep	0.0062 (0.84)	-0.0017 (-0.66)	0.0064 (0.87)	-0.0017 (-0.64)
State	0.0045*** (5.66)	0.0014*** (4.91)	0.0045*** (5.66)	0.0014*** (4.90)
_cons	0.3054*** (31.71)	0.0945*** (27.53)	0.3056*** (31.74)	0.0948*** (27.61)
行业	控制	控制	控制	控制
年份	控制	控制	控制	控制
N	11508	11508	11508	11508
adj. R ²	0.2187	0.2523	0.2187	0.2522
F	88.0795	105.9269	88.0528	105.8764

针对我国上市公司披露的主观叙述问题,由于中文语言的复杂性以及监管部门目前对上市公司文字信息披露的监管力度仍然较弱,为上市公司对文字信息披露进行操纵提供了可乘之机。我国监管部门应该加强对上市公司文字信息披露的监管力度,进一步规范上市公司会计文本的用语。美国证监会(SEC)在1998年制定了“Plain English”的规则,并且发布了指导上市公司如何使用简洁和规范用语的相关手册,我国监管部门也可以借鉴国外相关成功经验制定相关会计文本披露规范对我国上市公司的文字信息披露问题进行监管。对于外部信息使用者而言,应该注意上市公司文字叙述

所带的主观色彩,当文字叙述的主观色彩较浓时,应该仔细辨别相关文字信息的质量,考虑上市公司是否具有信息披露的操纵行为,以避免被低质量的信息误导,从而做出错误的决策。

参考文献:

- [1] Bochkay K and Levine C B. Using MD&A to Improve Earnings Forecasts [R]. Working Paper 2013.
- [2] Bonsall SB, Bozanic Z and Fischer P E. What Do Management Earnings Forecasts Convey About the Macroeconomy? [J]. Journal of Accounting Research, 2013, 51(2): 225-266
- [3] Hatzivassiloglou V and Wiebe J M. Effects of adjective orientation and gradability on sentence subjectivity [J]. Proceedings of Coling, 2000, 30(3): 299-305.
- [4] Lo K, Ramos F, Rogo R. Earnings management and annual report readability [J]. Journal of Accounting and Economics, 2017, 63(1): 1-25.
- [5] Loughran T and McDonald B. Textual Analysis in Accounting and Finance: A Survey [J]. Journal of Accounting Research, 2016, 54(4): 1187-1230.
- [6] Merkley K J. Narrative Disclosure and Earnings Performance: Evidence from R&D Disclosures [J]. Accounting Review, 2014, 89(2): 725-757.
- [7] Riloff E, Wiebe J and Wilson T. Learning subjective nouns using extraction pattern bootstrapping [C]. Proceedings of HLT-NAACL'03 2003: 25-32.
- [8] Richardson S A, Tuna A I, Wu M. Predicting Earnings Management: The Case of Earnings Restatements [R]. Working Paper, University of Pennsylvania 2002.
- [9] 程新生, 刘建梅, 程悦. 相得益彰抑或掩人耳目: 盈余操纵与 MD&A 中非财务信息披露 [J]. 会计研究, 2015(8): 11-18.
- [10] 段钊, 何雅娟, 钟原. 企业社会责任信息披露是否客观——基于文本挖掘的我国上市公司实证研究 [J]. 南开管理评论, 2017, 20(4): 62-72.
- [11] 李慧云, 郭晓萍, 张林, 等. 自愿性信息披露水平高的上市公司治理特征研究 [J]. 统计研究, 2013, 30(7): 72-77.
- [12] 沈家焯. 语言的“主观性”和“主观化” [J]. 外语教学与研究, 2001, 33(4): 268-275.
- [13] 魏志华, 李常青, 王毅辉. 中国上市公司年报重述公告效应研究 [J]. 会计研究, 2009(8): 31-39.
- [14] 叶强, 张紫琼, 罗振雄. 面向互联网评论情感分析的中文主观性自动判别方法研究 [J]. 信息系统学报, 2007(1): 79-91.
- [15] 于鹏. 股权结构与财务重述: 来自上市公司的证据 [J]. 经济研究, 2007(9): 134-144.
- [16] 张文文和王挺. 不规范文本的无监督观点句抽取 [J]. 计算机与数字工程, 2013, 41(1): 64-68.
- [17] 张璇, 周鹏, 李春涛. 卖空与盈余质量——来自财务重述的证据 [J]. 金融研究, 2016(8): 175-190.
- [18] 周晓苏, 周琦. 基于盈余管理动机的财务重述研究 [J]. 当代财经, 2011(2): 109-117.

Subjective Textual Description in Listed Companies' Information Disclosure and Earnings Manipulation: Research based on Machine Learning Method for Extracting Subjective Patterns

MAO Yi-xin

(School of Management, Xiamen University, Xiamen 361005, China)

Abstract: Selecting the sample of Shanghai and Shenzhen A-share listed companies from 2006 to 2016, this paper examines whether listed companies would manipulate text information disclosure and use more subjective textual description to cover up their earnings manipulation behaviors. Machine learning is used to extract 2-POS subjective model and construct quantifiable subjective indicators. Empirical results show that before financial restatement, that is, when earnings manipulation is concealed, listed companies use more subjective sentences in their annual reports and have higher scores with regard to overall subjectivity; developed external market environment and strong corporate governance effectively prevent listed companies taking advantage of subjective text information to work for earnings manipulation. This paper provides a new evidence for motives of listed companies to manipulate narrative disclosure, enriching the research of the relationship between financial data and text information, and also provides a new empirical basis for the supervision department to strengthen the supervision of information disclosure and external information users to judge the quality of information.

Key words: textual description; subjectivity; earnings manipulation

(责任编辑: 李江)