DOI:10.16452/j.cnki.sdkjzk.2021.01.011 文章编号:1672-3767(2021)01-0091-09

# 基于新闻情感的上市公司财务造假识别方法研究

张春梅[1],赵明清[2],吴学子[2]
(1.青岛黄海学院 大数据学院,山东 青岛 266000;2.山东科技大学 数学与系统科学学院,山东 青岛 266590)

**摘 要**:将大数据技术应用于上市公司财务造假识别是一有益探索。选取某段时期内存在财务造假的 39 家上市公司和与之配对的财务诚信度良好的 39 家上市公司作为研究样本,根据筛选后的 2 个反映新闻情感的指标与 5 个反映上市公司财务状况的指标,分别建立了基于财务指标、基于新闻情感、基于财务指标和新闻情感的三种上市公司财务造假 Logistic 识别模型,并对这三种模型进行了对比分析。结果表明:综合新闻情感与财务指标的财务造假 Logistic 识别模型的准确率最高,是上市公司财务造假识别的一种更有效方法。

**关键词**:财务造假;上市公司;财务指标;新闻情感;Logistic 模型
**中图分类号**:F275 **文献标志码**:A

## Financial fraud identification method for listed companies based on news sentiment

ZHANG Chunmei[1], ZHAO Mingqing[2], WU Xuezi[2]
(1. Big Data College, Qingdao Huanghai University, Qingdao, Shandong 266000, China;
2. College of Mathematics and Systems Science,
Shandong University of Science and Technology, Qingdao, Shandong 266590, China)

**Abstract**: Applying big data technology to the identification of financial fraud in listed companies can be of significant importance in stock market. In this paper, 39 companies having history of financial fraud in a certain period and 39 companies having history of good financial integrity were selected as research samples. According to the two indicators reflecting news sentiment and five indicators reflecting the companies' financial status, three Logistic financial fraud identification models of listed companies created. Those models are based on financial indicators, news sentiments, and financial indicators and news emotions respectively. Afterwards comparative analysis of the three models was achieved. The results show that the Logistic model that integrates news sentiment and considered as highly effective financial indicators has the highest accuracy and considered as highly effective financial fraud identification method for listed companies.

**Key words**: financial fraud; listed company; financial indicators; news sentiment; Logistic model

现实中,财务造假行为屡见不鲜,涉及到各行各业,造假花样层出不穷,其中不乏相关行业的龙头上市公司因存在严重的虚假财务报表而被披露,这给股票市场带来了极为不利的影响,也给股民带来了严重的经济

**收稿日期**:2020-06-01
**基金项目**:国家自然科学青年基金项目(61502280);山东科技大学研究生导师指导能力提升计划项目(KDYC17018);山东省高等学校人文社科计划项目(J17RB115)
**作者简介**:张春梅(1979—),女,山东即墨人,副教授,硕士,研究方向为统计建模与数据挖掘.E-mail:158434851@qq.com
赵明清(1963—),男,山东临朐人,教授,博士,研究方向为统计建模与大数据分析、保险精算与风险管理,本文通信作者.E-mail:zhaomq64@163.com

损失。因此财务造假问题是政府、证监会、上市公司和股民都十分关注的热点问题。对上市公司财务数据进行分析,构造准确率高的财务造假识别模型,对上市公司的造假行为进行有效识别非常必要。

值得注意的是,互联网产业使得各种信息呈现出爆炸式增长,有关上市公司的信息数量也在惊人地增长。为方便用户获取相关财经信息,众多财经新闻媒体、客户端如雨后春笋般出现在公众的视野。这些来自不同渠道、拥有不同形式的新闻数据构成了大数据,为财务造假识别提供了丰富的数据资源。

关于财务造假识别模型的研究,Beaver[1]首先通过对比 79 组上市公司 1954—1964 年间的财务数据,选取了 30 个财务指标作为分析变量,然后运用单变量分析模型从中筛选出 5 个对判别具有显著性影响的指标进行分析。尽管分析出的结果从某种程度上是相对可信和准确的,但是该模型的稳定性不高,如若选取不同的变量,甚至会得到相反的结论。针对以上不足,文献[2-3]提出了采用多元判别分析的方法构建 Z-Score 模型,并在此基础上构建了 Zeta 模型,在该模型中选取了 7 个财务指标作为分析的对象。目前,被广泛认可和应用的是 Probit 模型和 Logistic 模型,这两个模型在识别和预测效果上不相上下,但后者比前者更为方便,主要原因是前者对模型中变量的选择和数据的要求较为严格,使得确定指标数据较为困难。Martin[4]最早利用 logistic 进行银行的财务造假问题的识别,在接下来的几年里,Logistic 模型被广泛应用到上市公司财务舞弊问题的识别中。洪荭等[5]利用 Logistic 模型建立了上市公司财务造假与上市公司的监事会规模、董事长是否变更等公司内部治理指标的识别模型,效果良好。

大数据在财经领域的应用,早在二十世纪末就有了雏形。Nagar 等[6]通过对 50 家上市公司股票交易信息公布栏中企业消息的活跃程度进行分析。结果表明:相关企业股票的走势和交易健康状况与公告消息活跃程度紧密相关,从而提出了由公告栏的内容对公司股票走势的预测方法,这在当时的金融研究领域引起了很大的轰动。Antweiler 等[7]从雅虎和 Ragingbull.com 等网站获取数据,通过使用 Naive Bayes 分类法,以买入、持有和卖出的分类标准对从网站上获取的数据进行了分类,将最终经过统计量化得到的数据与企业的股票收益率、企业的股票波动性、企业的股票交易量等建立联系。结果表明:网页上提取的信息与企业的财务信息相关,但并没有指出相关性的强弱和具体的系数值。宋彪[8]详细阐述了基于大数据的财务预警的研究,将计算机分析技术、财务报表分析和数学模型有机结合,利用支持向量机进行实证分析,最后指出新闻文本数据可以大大提高公司财务预警的准确度。Jones[9]分析了在雅虎建立消息公告栏对股票价格波动的影响,并在结论中指出建立互联网信息公告栏会使得上市公司的股票价格波动性变小,其中很主要的原因是公告栏使得信息渐渐公开透明化,不同的投资者对这些信息会做出不同的反应,从而降低了风险,增加了交易量。Tetlock 等[10]以 500 家上市公司的新闻报道为研究对象,通过对 1980—2004 年间所有新闻中的消极词汇进行统计量化,结合了新闻情感分析与财务分析,研究其与公司收益率和股票回报率之间的相关性。结果表明:上市公司的盈利状况可以通过公司的新闻报道中的负面词汇进行预测。陈海文等[11]通过挖掘财经新闻网站中的具有情感倾向性的信息,在历史金融产品价格数据的基础上,组合多元线性回归模型和差分自回归滑动平均模型,提出一种新的预测金融产品价格的模型,实证研究表明该预测模型具有较高的准确率。戴德宝等[12]采用文本挖掘技术从网络故事论坛中挖掘投资者情绪综合指数,并利用支持向量机和神经网络预测股市价格,以提高预测的精度和有效性。赵杰[13]采用 ChiMerge 算法和决策树算法建立了财务舞弊的识别模型。易珩等[14]利用语义分析技术对创业板的上市公司的首次披露招股说明书和年度财务报表进行了情感分析,进而对创业板企业的风险进行了分类。宋英慧等[15]使用文本挖掘工具对 A 股上市公司财务报告的财务报表附注进行了分析,以发现财务报表的显著特征。王泽霞等[16]提出了专业性、真诚性、前瞻性深度和情感性四项 MD&A 语言,并以创业板制造业上市公司为研究对象,结果发现四项 MD&A 语言与公司未来财务业绩存在显著的正相关关系。

财务信息和财经新闻都是财务造假识别的重要信息源。目前,将财经新闻情感用于上市公司财务造假识别的研究尚未见到。本研究尝试利用文本挖掘技术,在上市公司财务造假识别中加入新闻情感要素,建立基于新闻情感的上市公司财务造假识别方法。具体来说:利用网易新闻中和上市公司相关的新闻建立新闻“数据库”,并对这些新闻进行情感分析,得到积极情感、中性情感和消极情感三个新闻情感指标,再与财务指标一起构建财务造假识别的 Logistic 模型,并与不加入新闻情感的识别模型进行对比分析。

## 1 研究设计

### 1.1 研究框架

研究框架如图 1 所示。

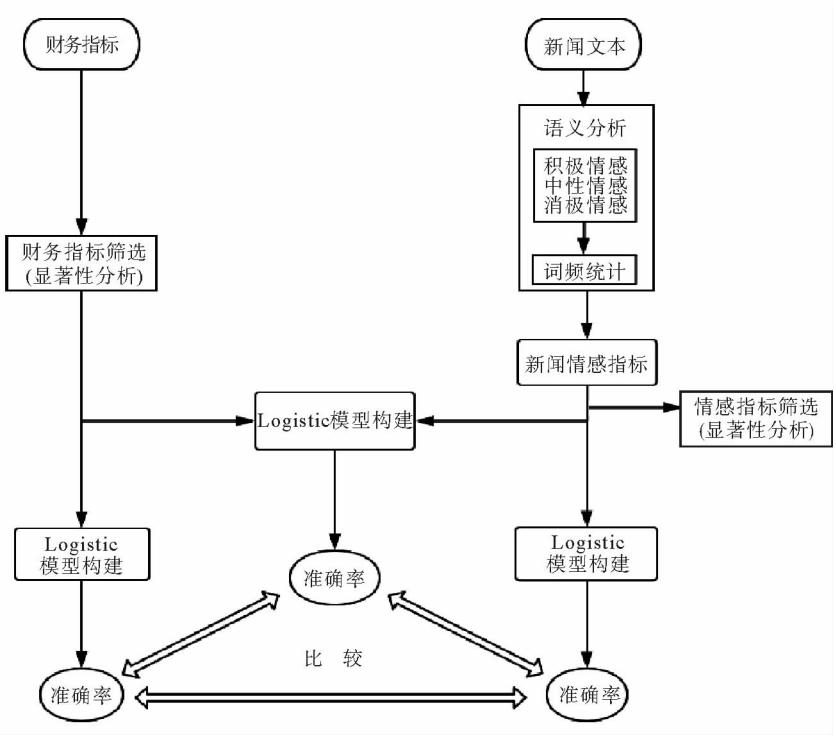


**图 1 研究框架**

Fig.1 Research frame

### 1.2 样本选取

本研究选取了 2001—2016 年被中国证监会、上交所、深交所公开进行行政处罚的 217 家上市公司为初始样本,从中剔除因没有及时披露信息被行政处罚的、以 IPO 为目的进行财务造假而被处罚的、关键性财务数据缺失的上市公司,最终选择了 39 家上市公司作为研究对象。对于连续多年都出现年度财务报表数据造假行为的上市公司,选择其首次造假的年份进行研究分析。在确定好造假上市公司的样本之后,再选择在相关年份与该上市公司处于同一行业、公司规模相近的,并且没有被中国证监会等监督机构处罚过的 39 家财务良好的上市公司作为配对样本。所选取的上市公司覆盖各行业,其中制造业、批发和零售业以及农、林、牧、渔业占比重较大。所有相关上市公司名单和股票代码均来自于中国证监会官网中有关“行政处罚”的公示信息。

从 39 对上市公司样本中选取前 28 对作为训练样本集,建立识别模型,剩余的 11 对作为测试样本集,验证识别模型的准确率。

### 1.3 财务指标选取

本研究从反映上市公司资产总体状况、偿债能力、盈利水平、现金流量和资本流通水平等五个方面共选取 18 个财务指标作为初始解释变量,如表 1 所示。按照确定的上市公司股票代码,在中经网查找其相关造假年份的财务报表,获取财务指标数据。

选取初始解释变量后,对配对的 39 对样本即可进行 T 检验,该检验统计量为 $T=\dfrac{\bar{d}-(\mu_1-\mu_2)}{S/\sqrt{n}}\sim t(n-1)$。其中,$\mu_1$ 和 $\mu_2$ 分别为第一个总体和第二个总体的均值,$d_i$ 为配对样本差值,$\bar{d}$ 为 $d_i$ 的均值,$S$ 为 $d_i$ 的标准差,$n$ 为样本数,其结果如表 2 所示。由此可知,在 10% 的显著性水平上,$X_1$、$X_2$、$X_5$、$X_6$、$X_7$、$X_{10}$、$X_{11}$、$X_{12}$、$X_{13}$、$X_{14}$、$X_{15}$ 和 $X_{17}$ 对财务造假或非造假不存在显著性影响,将其剔除,只保留具有显著影响的 $X_3$、$X_4$、$X_8$、$X_9$ 和 $X_{16}$。

## 2 新闻文本数据的获取与处理

### 2.1 数据的来源与获取

本研究采用网易新闻网站(https://news.163.com/)涉及样本上市公司的新闻。

### 2.2 数据的清洗

将所获取的数据导出至本地文件,并根据时间列表选取所需年份的不重复“新闻报道”保存至txt文件,

**表 1 财务造假识别的初始财务指标**

Tab. 1 Initial financial indicators of financial fraud identification

| 指标类型 | 指标符号 | 指标名称 | 计算公式 |
|---|---|---|---|
| 资产总体状况 | $X_1$ | 存货周转率 | 营业成本/平均存货余额 |
| | $X_2$ | 资产周转率 | 营业收入/平均资产总额 |
| | $X_3$ | 应收账款周转率 | 营业收入总额/应收账款平均占用额 |
| | $X_4$ | 资产质量 | 资产减值准备总额/资产总额 |
| 偿债能力 | $X_5$ | 资产负债率 | 总负债/总资产 |
| | $X_6$ | 速动比率 | (流动资产—存货)/流动负债 |
| | $X_7$ | 现金流量比率 | 经营性现金流量/平均流动负债 |
| 盈利水平 | $X_8$ | 财务杠杆系数 | (利润总额+财务费用)/利润总额 |
| | $X_9$ | 资产回报率 | 净利润/平均资产总额 |
| | $X_{10}$ | 净资产回报率 | 净利润/平均净资产 |
| | $X_{11}$ | 收益质量 | 现金和现金等价物的净增加额/净利润 |
| | $X_{12}$ | 收入成本配比性 | 主营业务成本/主营业务收入 |
| | $X_{13}$ | 税金及附加比例 | 主营业务税金及附加/主营业务收入 |
| | $X_{14}$ | 实际所得税率 | 所得税/利润总额 |
| 现金流量 | $X_{15}$ | 每股现金流量 | (经营活动产生的现金流量净额—优先股股利)/流通在外的普通股股数 |
| 资本流通水平 | $X_{16}$ | 利息资本化率 | 财务费用/总负债 |
| | $X_{17}$ | 费用资本化率 | 除固定资产外的非流动资产/总资产 |

**表 2 财务指标配对样本 T 检验**

Tab. 2 Paired samples T-test of financial indicators

| 指标符号 | 均值 | | $T$ | $Sig.$ |
|---|---|---|---|---|
| | 财务诚信度良好的上市公司 | 财务造假的上市公司 | | |
| $X_1$ | 7.969 49 | 16.036 15 | −0.809 | 0.424 |
| $X_2$ | 0.569 23 | 0.614 36 | −0.616 | 0.542 |
| $X_3$ | 13.493 59 | 7.958 46 | 1.731 | 0.091 |
| $X_4$ | −0.023 41 | 0.067 75 | −1.740 | 0.090 |
| $X_5$ | 1.776 31 | 1.684 58 | 0.312 | 0.756 |
| $X_6$ | 1.352 69 | 1.269 28 | 0.317 | 0.753 |
| $X_7$ | 0.064 77 | −0.008 14 | 0.827 | 0.414 |
| $X_8$ | 0.457 18 | 0.553 47 | −1.992 | 0.054 |
| $X_9$ | 0.152 80 | 0.037 38 | 2.065 | 0.046 |
| $X_{10}$ | 0.030 59 | 0.012 80 | 0.576 | 0.568 |
| $X_{11}$ | −5.360 02 | −0.050 75 | −0.672 | 0.506 |
| $X_{12}$ | 0.066 09 | 1.202 98 | −1.236 | 0.224 |
| $X_{13}$ | 0.013 46 | 0.010 26 | 1.008 | 0.320 |
| $X_{14}$ | 0.144 15 | 0.158 92 | −0.370 | 0.713 |
| $X_{15}$ | 0.254 10 | 0.298 718 | −0.172 | 0.865 |
| $X_{16}$ | 0.014 44 | 0.239 34 | −1.764 | 0.086 |
| $X_{17}$ | 0.174 10 | 0.192 33 | −0.582 | 0.564 |

将 txt 文档导入 R 中进行分词,将句子、各式标点符号拆分成单独的个体形式,同时清除标点、空格、URL 以免对 R 程序进行后续分析产生干扰。

### 2.3 新闻情感指标的构建

在构建用于分析上市公司的新闻文本情感词典时采用了徐娅[17]公布的财经情感词典,并参考了中文分词词典。由于在财经词典尚未有较为全面的中性词汇总结,本研究在选取积极词汇和消极词汇时采用了财经情感词典,而中性词汇选取了中文分词词典中的内容,如表 3 所示。

**表 3 财经情感词典**

Tab. 3 Financial sentiment dictionary

| 情感属性 | 关键词 |
|---|---|
| 积极词汇 | 反弹,发展,增长,上涨,新,高,多,涨幅,超过,有望,买入,提升,涨停,增持,利好,增速,较大,最大,建设,收益,重组,推进,回升,超,积极,上升,涨,看好,大涨,信心,解禁,有效,偏好,修复,推荐,活跃,升级,新增,牛市,加强,回暖,爆发,分红,乐观,大量,有利于,促进,开展,降准,恢复,优质,暴涨 |
| 中性词汇 | 观望,考虑,谨慎,平稳,比较,选择,可能,也许,未来,以后,后来,产生,变化,变动,停,不变,不久 |
| 消极词汇 | 下跌,震荡,超过,没有,减持,暴跌,波动,回落,大跌,亏损,谨慎,通胀,不足,过剩,弱势,冲击,回调,恐慌,低迷,债转股,跳水,担忧,严重,暂停,悲观,跌幅,下降,低,跌,下行,贬值,减少,降低,下滑,跌破,退市,跌停,杀跌,股灾,压力,炒作,债务,熊市,萎缩,无法,退出,下调,缺乏,难,泡沫,减仓,过度 |

将实现分词后的文本与词典在 R 中调用 Rwordseg 程序包进行词频统计。在初始设置时自动过滤掉出现少于 5 次的词汇,将统计出的词与财经情感词典中的词进行分类匹配,相应的情感词占总词数的百分比作为最后的新闻情感指标,即积极情感指标($E_1$)、中性情感指标($E_2$)和消极情感指标($E_3$):

$$E_1=\frac{\sum N_{\text{poslarity}}}{N},E_2=\frac{\sum N_{\text{neutral}}}{N},E_3=\frac{\sum N_{\text{negative}}}{N}。$$

其中,$N_{\text{poslarity}}$、$N_{\text{neutral}}$、$N_{\text{negative}}$ 分别表示财经新闻中某一积极词汇、中性词汇、消极词汇的词频。每家上市公司选取 50 篇新闻,共计1 950篇,前 10 家上市公司的积极、中性和消极词频统计结果如表 4 所示。

**表 4 词频统计结果(部分)**

Tab. 4 Word frequency statistics (section)

| 公司简称 | $E_1$ | $E_2$ | $E_3$ |
|---|---|---|---|
| 兴发集团 | 0.358 9 | 0.284 3 | 0.356 8 |
| 启迪古汉 | 0.299 5 | 0.431 6 | 0.268 9 |
| 广汇物流 | 0.351 1 | 0.292 8 | 0.356 1 |
| 广东榕泰 | 0.300 7 | 0.411 2 | 0.288 1 |
| 中南建设 | 0.400 2 | 0.322 4 | 0.277 4 |
| 北方导航 | 0.233 6 | 0.422 3 | 0.344 1 |
| 广电网络 | 0.200 4 | 0.548 2 | 0.251 4 |
| 财信发展 | 0.269 7 | 0.410 3 | 0.320 0 |
| 宁波韵升 | 0.352 4 | 0.437 6 | 0.210 0 |
| 太龙药业 | 0.425 1 | 0.419 3 | 0.155 6 |

### 2.4 显著性分析

将统计出的 39 组公司的各情感指标数据剔除异常值后,利用 SPSS 软件进行单样本 T 检验,检验结果如表 5 所示。可以看出,$E_1$ 和 $E_3$ 通过了显著性检验,即积极情感指标($E_1$)和消极情感指标($E_3$)数据之间存在显著性差异,而中性情感指标($E_2$)数据之间无显著性差异。由此可知,不存在财务造假的上市公司相应的新闻报道中积极情感指标和消极情感指标所占比重都比较低,且数值近乎相当;相反地,存在财务造假的上市公司特别是连续多年造假的公司造假年份的新闻文本中消极情感指标所占比重远远高于积极情感指标。中性情感指标,在一定程度上不能代表大众的偏向性,在舆情中更多的是起到过渡的作用,这一类公众对

**表 5 情感指标显著性分析**

Tab. 5 Sentiment indicators T-test

| 变量 | $T$ | $DF$ | $Sig.$(双侧) |
|---|---|---|---|
| $E_1$ | 4.455 | 77 | 0.000 |
| $E_2$ | 1.265 | 60 | 0.210 |
| $E_3$ | 2.829 | 74 | 0.006 |

现状持观望和谨慎的态度,无法明确地判断出他们未来的观点和态度,将其剔除。

## 3 识别模型的构建

为了说明新闻情感是否在财务造假识别中起到有效的作用,本研究分别建立基于财务指标、基于新闻情感、基于财务指标和新闻情感的三种上市公司财务造假 Logistic 识别模型,以便于比较。

设 $P$ 为上市公司财务造假的概率,则可建立如下财务造假 Logistic 识别模型

$$\ln\left(\frac{P}{1-P}\right)=\beta_0+\beta_1 x_1+\cdots+\beta_k x_k\ ,$$

式中:$\beta_0,\beta_1,\cdots,\beta_k$ 是待估计的未知参数;$x_0,x_1,\cdots,x_k$ 是财务造假识别变量。由上式,可得

$$P=\frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}。$$

### 3.1 基于财务指标的财务造假 Logistic 识别模型

利用 5 个具有显著性差异的财务指标,对训练样本集进行 Logistic 回归,结果如表 6 所示。

**表 6 财务指标回归结果**

Tab. 6 Regression results of financial indicators

| 变量 | *B* | *S.E.* | *Wals* | *Sig.* |
|---|---|---|---|---|
| $X_3$ | −0.062 | 0.044 | 1.972 | 0.160 |
| $X_4$ | 113.512 | 63.741 | 3.171 | 0.075 |
| $X_8$ | 2.169 | 1.899 | 1.305 | 0.253 |
| $X_9$ | 6.152 | 8.021 | 0.588 | 0.443 |
| $X_{16}$ | 46.844 | 20.855 | 5.045 | 0.025 |
| 常量 | −2.421 | 1.422 | 2.899 | 0.089 |

由此,可得识别模型

$$\hat{P}=\frac{e^{-2.421-0.062X_3+113.512X_4+2.169X_8+6.152X_9+46.844X_{16}}}{1+e^{-2.421-0.062X_3+113.512X_4+2.169X_8+6.152X_9+46.844X_{16}}},$$

其中,$\hat{P}$ 表示上市公司财务造假的概率估计值。利用该模型对测试样本集进行识别,结果如表 7 所示。

**表 7 基于财务指标的 Logistic 模型识别准确率**

Tab. 7 Logistic model recognition accuracy based on financial indicators

| 已观测 | | 已识别 | | |
|---|---|---|---|---|
| | | 是否造假 | | 识别率/% |
| | | 0 | 1 | |
| 是否造假 | 0 | 9 | 2 | 81.8 |
| | 1 | 6 | 5 | 45.5 |
| 平均识别率 | | | | 63.6 |

### 3.2 基于新闻情感的财务造假 Logistic 识别模型

利用积极情感指标($E_1$)和消极情感指标($E_3$)对训练样本集进行 Logistic 回归,结果如表 8 所示。

**表 8 情感指标回归结果**

Tab. 8 Regression results of sentiment indicators

| 变量 | *B* | *S.E.* | *Wals* | *Sig.* |
|---|---|---|---|---|
| $E_1$ | −24.019 | 12.384 | 3.762 | 0.042 |
| $E_2$ | 55.849 | 29.054 | 3.695 | 0.045 |
| 常量 | −13.776 | 7.683 | 3.215 | 0.063 |

由此,可得识别模型

$$\hat{P}=\frac{e^{-13.776-24.019E_1+55.849E_3}}{1+e^{-13.776-24.019E_1+55.849E_3}}。$$

利用该模型对测试样本集进行识别,结果如表 9 所示。

**表 9 基于情感指标的 Logistic 模型识别准确率**

Tab. 9 Logistic model recognition accuracy based on sentiment indicators

| 已观测 | | 已识别 | | |
|---|---|---|---|---|
| | | 是否造假 | | 识别率/% |
| | | 0 | 1 | |
| 是否造假 | 0 | 7 | 4 | 63.6 |
| | 1 | 2 | 9 | 81.8 |
| 平均识别率 | | | | 72.7 |

### 3.3 基于新闻情感和财务指标的财务造假 Logistic 识别模型

财务指标中有 5 个财务指标通过了显著性检验,在新闻情感指标中有 2 个情感指标通过了显著性检验,将这 7 个指标:$X_3$、$X_4$、$X_8$、$X_9$、$X_{16}$、$E_1$、$E_3$ 进行标准化处理后,对训练样本集进行 Logistic 回归,结果如表 10 所示。

**表 10 财务指标 & 情感指标回归结果**

Tab. 10 Regression results of financial indicators & sentiment indicators

| 变量 | *B* | *S.E.* | *Wals* | *Sig.* |
|---|---|---|---|---|
| $X_3$ | −0.930 | 1.480 | 0.395 | 0.530 |
| $X_4$ | 15.825 | 20.924 | 0.572 | 0.449 |
| $X_8$ | 0.611 | 1.425 | 0.184 | 0.668 |
| $X_9$ | 0.549 | 1.710 | 0.103 | 0.748 |
| $X_{16}$ | −0.633 | 2.093 | 0.091 | 0.762 |
| $E_1$ | −3.975 | 3.084 | 1.661 | 0.197 |
| $E_2$ | 17.464 | 12.941 | 1.821 | 0.177 |
| 常量 | 5.858 | 5.026 | 1.358 | 0.244 |

由此,可得识别模型

$$\hat{P}=\frac{e^{5.858-0.930X_3+15.825X_4+0.611X_8+0.549X_9-0.633X_{16}-3.975E_1+17.464E_3}}{1+e^{5.858-0.930X_3+15.825X_4+0.611X_8+0.549X_9-0.633X_{16}-3.975E_1+17.464E_3}}。$$

利用该模型对测试样本集进行识别,结果如表 11 所示。

**表 11 基于财务指标和情感指标的 Logistic 模型识别准确率**

Tab. 11 Logistic model recognition accuracy based on financial indicators & sentiment indicators

| 已观测 | | 已识别 | | |
|---|---|---|---|---|
| | | 是否造假 | | 识别率/% |
| | | 0 | 1 | |
| 是否造假 | 0 | 9 | 2 | 81.8 |
| | 1 | 1 | 10 | 90.9 |
| 平均识别率 | | | | 86.4 |

### 3.4 模型的对比分析

由表 7、表 9 和表 11 可以看出:

1) 基于财务指标的模型识别率较低。由于分析的样本不够充足,选取的解释变量可能存在偏差,以及财务造假上市公司在被行政处罚后及时更改了相关年度财务报表的关键数据等原因,使该模型的预测效果没有达到预期,尽管如此,仅通过 5 项关键的财务指标对上市公司是否存在财务造假进行识别的准确度高于 60%,足以说明由公司的财务指标情况可以较有效地识别财务造假。

2) 基于新闻情感的模型识别率高于基于财务指标的识别率。新闻可以反映在某一阶段媒体、公众对上市公司的评价,这些评价体现出的情感倾向可以较为准确地反映出上市公司的经营状况和盈利水平。反常的财务数据会引起财经新闻媒体的关注,从而进行相关的报道。上市公司也无法阻止、篡改新闻网站出现的负面新闻报道,所以新闻文本数据比财务数据的真实性会更高。

3) 基于财务指标和新闻情感的识别模型结合了两个模型的优点,在三个模型中识别率最高,说明新闻文本数据样本可以提高对上市公司是否存在造假的识别能力。财务数据的滞后性和难以判断真实性的特点,使得仅通过财务指标建立识别模型的准确性无法保证,但是新闻文本数据弥补了财务数据的不足。

## 4 结束语

将新闻情感用于上市公司财务造假的识别,结果表明:融合新闻情感与财务指标的财务造假 Logistic 识别模型是有效的,新闻情感在财务造假识别中有重要影响,能够弥补财务数据的不足,体现了大数据的价值所在。研究所使用的模型是在实际中得到了广泛应用的 Logistic 回归模型,而且在应用前首先利用 T 检验法进行了变量筛选,可进一步考虑利用基于 Lasso 的 Logistic 回归模型,识别效果可能会更好,这是下一步要研究的内容。

**参考文献:**

[1]BEAVER W H.Financial ratios as predictors of failure[J].Journal of Accounting Research,1966,4(3):71-111.

[2]ALTMAN E I.Financial ratios,discriminant analysis and the prediction of corporate bankruptcy[J].The Journal of Finance,1968,23(4):589-609.

[3]ALTMAN E I,HALDEMAN R G,NARAYANAN P.ZETA™ analysis:A new model to identify bankruptcy risk of corporations[J].Journal of Banking & Finance,1977,1(1):29-54.

[4]MARTIN D.Early warning of bank failure[J].Journal of Banking & Finance,1977,1(3):249-276.

[5]洪荭,胡华夏,郭春飞.基于 GONE 理论的上市公司财务报告舞弊识别研究[J].会计研究,2012(8):84-90.
HONG Hong,HU Huaxia,GUO Chunfei.Research on the identification of listed companies' financial reporting fraud:Based on GONE theory[J].Accounting Research,2012(8):84-90.

[6]NAGAR V,NANDA D,WYSOCKI P.Discretionary disclosure and stock-based incentives[J].Journal of Accounting and Economics,2003,34(1/2/3):283-309.

[7]ANTWEILER W,FRANK M Z.Is all that talk just noise? The information content of Internet stock message boards[J].The Journal of Finance,2004,59(3):1259-1294.

[8]宋彪.基于大数据的企业财务预警理论与方法研究[D].北京:中央财经大学,2015.
SONG Biao.Research on theory and approach of enterprise financial early warning based on big data[D].Beijing:Central University of Finance & Economics,2015.

[9]JONES A L.Have Internet message boards changed market behavior? [J].Info,2006,8(5):67-76.

[10]TETLOCK P C,SAAR-TSECHANSKY M,MACSKASSY S.More than words:Quantifying language to measure firms′ fundamentals[J].The Journal of Finance,2008,63(3):1437-1467.

[11]陈海文,蔡志平,方峰.应用财经新闻挖掘的金融品种价格走势预测[J].计算机工程与科学,2016,38(9):1909-1916.
CHEN Haiwen,CAI Zhiping,FANG Feng.Financial price trend forecast using financial news mining[J].Computer Engineering & Science,2016,38(9):1909-1916.

[12]戴德宝,兰玉森,范体军,等.基于文本挖掘和机器学习的股指预测与决策研究[J].中国软科学,2019(4):166-175.
DAI Debao,LAN Yusen,FAN Tijun,et al.Stock forecast with investors sentiment by text mining and machine learning[J].China Soft Science,2019(4):166-175.

[13]赵杰.数据挖掘在识别财务舞弊中的研究与应用[D].北京:首都经济贸易大学,2017.
ZHAO Jie.Research and application of data mining in identifying financial fraud[D].Beijing:Capital University of Economics and Business,2017.

[14]易珩,马琪琪,章惟一.基于语义分析方法的创业板风险信息披露研究[J].商业会计,2019(2):74-77.
YI Heng,MA Qiqi,ZHANG Weiyi.A study on risk information disclosure of GEM listed companies based on semantic analysis method[J].Commercial Accounting,2019(2):74-77.

[15]宋英慧,黄麒.基于文本挖掘技术的财务报表附注披露研究[J].会计之友,2019(1):142-147.

[16]王泽霞,潘梦雪,郜鼎.MD&A语言特征与公司未来财务业绩—基于中国创业板制造业上市公司的实证研究[J].财会月刊,2019(2):78-87.

[17]徐娅.基于词典的财经微博信息的情感态度挖掘[D].金华:浙江师范大学,2014.
XU Ya.Mining of the financial micro-blog emotion attitude based on dictionary[D].Jinhua:Zhejiang Normal University,2014.

(责任编辑:刘西奎)