Sustainable Development 可持续发展, 2024, 14(5), 1099-1109
Published Online May 2024 in Hans. https://www.hanspub.org/journal/sd
https://doi.org/10.12677/sd.2024.145124

Hans汉斯

基于文本挖掘和二元Logistics回归的生物医药行业上市公司财务预警研究

曹芷萱，宁雯峰，尚　可，周嘉仪，朱鹏霖

中国矿业大学(北京)，管理学院，北京

收稿日期：2024年3月18日；录用日期：2024年4月17日；发布日期：2024年5月14日

摘　要

本研究论文主要探讨了基于文本挖掘和二元logistics回归的生物医药行业上市公司财务预警研究。通过分析生物医药行业上市公司的财务数据和管理层讨论与分析(MD&A)文本，利用二元logistics回归模型和文本挖掘技术，研究构建了一套财务预警模型，以预测企业是否可能发生财务危机。研究选取了现金比率、现金流量负债比率、资产负债率、净资产收益率、营业毛利率、息税前营业利润率、总资产增长率、营业收入增长率、资本保值增率、每股经营活动产生的净流量增长率等财务指标，以及基于MD&A文本信息构建的语调指标作为模型的关键预警指标。这些指标涵盖了企业的多个财务维度，旨在全面评估公司的财务健康状况。模型在非ST公司的判断准确率达到了87.50%，在ST公司的判断准确率为80.00%，整体准确率为85.29%，显示了该模型对于早期识别财务风险具有一定的有效性。

关键词

生物医药上市公司，财务预警，文本挖掘，Logistic回归

Research on Financial Early Warning of Listed Companies in Biopharmaceutical Industry Based on Text Mining and Binary Logistic Regression

Zhixuan Cao, Wenfeng Ning, Ke Shang, Jiayi Zhou, Penglin Zhu

School of Management (Beijing), China University of Mining and Technology, Beijing

Received: Mar. 18th, 2024; accepted: Apr. 17th, 2024; published: May 14th, 2024

文章引用: 曹芷萱, 宁雯峰, 尚可, 周嘉仪, 朱鹏霖. 基于文本挖掘和二元Logistics回归的生物医药行业上市公司财务预警研究[J]. 可持续发展, 2024, 14(5): 1099-1109. DOI: 10.12677/sd.2024.145124

Abstract

This research paper mainly discusses the research on financial early warning of listed companies in the biomedical industry based on text mining and binary logistic regression. By analyzing the financial data and management discussion and analysis (MD&A) texts of listed companies in the biopharmaceutical industry, and using binary logistic regression models and text mining technology, a set of financial early warning models were constructed to predict whether a company may have a financial crisis. The study selected cash ratio, cash flow-liability ratio, asset-liability ratio, return on net assets, operating gross profit margin, operating profit margin before interest and taxes, total asset growth rate, operating income growth rate, capital preservation growth rate, and operating activities per share. Financial indicators such as the generated net traffic growth rate, as well as tone indicators constructed based on MD&A text information are used as key early warning indicators of the model. These indicators cover multiple financial dimensions of a business and are designed to provide a comprehensive assessment of a company's financial health. The model's judgment accuracy in non-ST companies reached 87.50%, in ST companies it was 80.00%, and the overall accuracy was 85.29%, showing that the model has certain effectiveness in early identification of financial risks.

Keywords

Listed Biopharmaceutical Companies, Financial Warning, Text Mining, Logistic Regression

1. 引言

生物医药行业成为我国新一轮竞争和技术革命的焦点领域。国家发改委也在“十四五”规划中提出了对生物医药行业的发展要求。不仅是国家方面重视，大量投资者也在涌入该行业。在此背景下，投资者和管理者们更需要警惕并避免出现盲目的投资或者管理决策错误而不自知的情况。财务危机并非没有预兆，企业出现财务危机往往是经过一个逐步渐进的过程，可以根据相关指标情况被预测，以保护投资者和债权人的权益，以及为管理者的决策提供参考警示。

国内外对于财务危机预警的研究已有多年历史，但大部分研究是面对整体市场企业，缺少对于生物医药行业的具体研究。本研究建立在基于 logistics 回归预警模型的基础上还引入了文本挖掘技术，本文将主要从公司的 MD&A 文本等非结构化数据中提取关键信息。这使模型能够更全面地获取和分析生物医药行业上市公司的财务情况，并运用二元 logistics 回归模型对这些指标进行建模分析，以识别出对公司财务状况产生重要影响的关键因素，并据此进行风险预测和预警。财务危机存在着两种最终结果，分别是破产和化解财务危机。如果企业能够及时合理地处置财务危机，也可能顺利渡过难关，重新恢复到日常经营[1]。

本研究的意义主要体现在如下几个方面：首先，通过引入文本挖掘技术，我们能够更全面地获取和分析生物医药行业上市公司的财务情况，提高预警模型的准确性和有效性。在生物医药行业范围内，将文本挖掘和财务预警相结合的相关研究还相对较少，本论文可为相关研究丰富研究内容；其次，采用二

元 logistics 回归模型，可以对财务预警指标进行定量化分析，为企业管理者提供科学依据和决策支持；最后，本研究的结果不仅可以帮助生物医药行业上市公司及时预警财务风险，还可以为投资者提供更准确的投资建议，促进行业的健康发展和资本市场的稳定。

2. 文献综述

2.1. 财务危机概念与预警模型概述

在财务危机概念的界定方面，国内外因为经济与市场环境的不同，对财务危机的界定有着很大的差异。国外的资本市场发展相对更加成熟，多以公司是否破产为标准来对财务危机进行界定。而我国资本市场破产企业相对较少，导致直接预测破产缺少研究样本，实证难度大，故研究中往往以被特别处理(ST或*ST)的上市公司作为财务危机的标志。其中吴世农[2]在构建财务危机预警模型以我国 A 股市场上的 ST 公司作为研究对象。但综合所有，公司财务危机的内涵具体是指企业盈利能力和偿债能力的不足,明显地已经陷入债务危机或具备不能够按时偿还到期的特征。即如果企业处于筹融资困难、经营不善等状态时，企业的营运能力会随之下降，进而影响投资收益，那么这种情况的发生就表明企业可能面临了财务风险。如果情况不得好转、逐渐恶化，最后将会陷入财务危机。

目前国内外的财务预警研究，主要集中在以下三个方面：如何界定财务危机程度以确定出现财务危机的研究对象；如何进行财务危机预警指标选择以构建危机指标体系；如何建立有效的财务危机预警模型以识别出潜在的财务风险。李莉[3]指出财务预警的分析方法分为定性与定量两种，定性分析方法客观、易懂,但也存在准确性不足的问题,目前大多数实证研究都是基于定量分析方法进行的。本文也基于定量分析方法建立模型进行预警。在财务危机预警指标的选取方面，国内外学者都从最初考虑财务指标，如偿债能力、现金流量、营运能力和盈利能力，拓展到把公司治理、股权结构、审计意见等非财务指标纳入预警指标评价体系，从更为广阔的视角为财务危机预警和防范提供可能。宋慧斌[4]提出相应财务预警的理解：“财务预警系统是以企业财务信息数据为基础，以财务指标体系为中心，通过对财务指标的综合分析、预测、及时反映企业经营情况和财务状况的变化，并对企业各环节发生或将可能发生的经营风险发出预警信号，为管理当局提供决策依据的监控系统”。学者王永明提到通过文本挖掘，可以及时捕捉到企业经营环境的变化、市场情绪的波动以及管理层对未来的预期等信息，这些都是传统财务指标难以快速反映的。[5]综上，在财务危机预警模型方面，运用的模型指标越趋于复杂化。将这文本挖掘技术融入财务预警模型，可以有效地提高预警的准确性。

2.2. 医药生物行业财务危机的研究进展

2.2.1. 医药生物行业面临财务风险现状

后疫情时代伴随着人口老龄化加剧的来临，人们对健康的重视和需求与日俱增，医药生物行业的重要性日益扩大。邱会[6]结合当下社会环境，分析医药生物行业发展前景与风险。医药生物行业在全球范围内都是一个重要的经济支柱产业，由于其特殊性质，医药行业面临着许多财务风险和挑战，例如研发成本极高、新药研发周期长、严格的审批和监管、专利保护期限过期后的仿冒竞争、市场不确定性等。因此，如何及早预警医药生物行业的财务危机，成为了学术界和业界关注的热点问题。

2.2.2. 医药生物行业财务危机研究进度及本文研究方向

医药生物行业财务危机预警是一个重要的研究领域，旨在通过对企业财务数据和相关指标的分析，提前预警潜在的财务风险。刘抗英[7]、黄世忠[8]等人深刻剖析在当今大数据、碳中和等大环境下的财务危机指标评估，将之与医药生物行业前景相结合，其相关研究为本文提供了医药生物行业财务危机预警

相关的研究进度及主要解决方向：1) 指标模型：研究者通过对医药生物企业的财务指标进行建模和分析，开发了各种财务危机预警模型。常用的模型包括多元判别分析、逻辑回归模型、人工神经网络等；2) 基于机器学习的预测方法：通过将机器学习算法应用于财务危机预警中。通过利用大量的历史数据和特征工程技术，机器学习模型能够更精确地识别潜在的财务风险；3) 财务比率分析：通过分析医药生物企业的财务比率，如偿债能力、盈利能力、运营能力等指标，来判断企业是否存在财务风险并进行预警；4) 基于文本挖掘的市场数据的预测：分析 MD&A 文本、市场评级和投资者情绪等指标，可以为财务预警模型提供额外的预警变量，为提高财务预警模型预测准确率提供了一条可行的思路。

本文采用多种方式，从“搜狐证券”和国泰安数据库获取数据源，计算所选 ST 及非 ST 企业的财务比率，选择合适指标，运用逻辑回归模型进行财务危机预警，同时通过文本挖掘技术分析企业年报管理层语调情感构建辅助指标提高模型预测准确率。

3. 方法和数据

3.1. 方法

3.1.1. 二元 logistic 回归模型

模型选用二元 logistic 回归模型，是一种广义的线性回归分析模型，常用与数据挖掘、疾病自动诊断、经济预测等领域。本文选取该模型主要有以下原因：1) 简单高效。二元 logistic 回归模型实现简单，分类时计算量小，运行速度快，对存储资源的需求低，适合较大数据处理；2) 概率预测。模型直接提供了观测样本属于某一类别的概率分数；3) 可以克服多重共线性问题。二元 logistic 回归能够较好地处理自变量之间存在的多重共线性问题，且计算代价不高，易于理解实现。4) logistics 模型特别适合用于处理响应变量为二分类的预测问题，在财务预警系统中预测企业是否可能发生财务危机方面得到了广大学者的认可。

参考田青、马越越主编的《大数据计量经济分析》做法，将线性组合的结果通过逻辑函数转换为概率，这一步骤将线性组合的结果映射到[0, 1]区间，得到事件发生的概率。[9]本文使用的二元 logistic 回归模型公式的基本形式见式(1)。

其主要预测判别过程是通过设置一个阈值，通常为 0.5，将 P_i 和阈值进行比较得出预测结果。如果计算出的 $P_i \geq 0.5$，则预测该公司会陷入财务危机；若 $P_i \leq 0.5$，则预测公司财务状态正常。为简化模型解释，使模型变得更加直观，本文在实证部分将用二进制变量代表企业是否存在财务危机(1 代表存在财务危机，0 代表不存在财务危机)。

$$L_i = \ln\frac{P_i}{1-P_i} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} \tag{1}$$

其中，k 是变量数量，即本文选取的指标总数；i 是训练样本的样本序列号；X_{ki} 为自变量。是 P_i 是第 i 个公司的财务预警概率，β_i 为模型各自变量的回归系数。

3.1.2. 文本挖掘

文本挖掘在本文的优势以及作用文本挖掘技术在金融领域的应用日益受到重视，尤其在分析和预测公司的财务状况方面展现出独特的价值。本文通过对上市公司年度财报中的管理层讨论与分析(MD&A)文本信息进行深入挖掘，构建情绪指标，进一步证实了文本挖掘技术在理解和预测公司经营表现方面的有效性。文本挖掘不仅增强了对传统财务指标分析的补充，而且还为财务预警模型提供了一种新的视角和分析维度。文本挖掘技术的核心优势在于其能够将大量非结构化的文本信息转化为可量化的数据，这对于深入理解企业的经营状态、市场情绪及未来发展趋势具有重要意义。特别是在分析 MD&A 文本时，

通过提取和量化其中的情感表达，可以更准确地捕捉到管理层对于企业未来发展的看法和态度，从而为投资者和监管者提供更为丰富和直观的信息。

本文主要通过文本挖掘技术构建语调指标，旨在丰富财务预警模型的类型完善模型预测的全面性、提高预测准确率。部分样本公司的语调指标实例如表 1 所示，本文运用文本挖掘技术的主要步骤如下。

在挖掘样本的选择方面，上市公司年度财报中 MD&A (管理层讨论与分析)文本信息用于构建情绪指标有较好的适用性。因此本文对样本公司的 MD&A 文本信息进行文本挖掘，用以构建情绪指标[10]。在挖掘和处理方面，本文采用分词方法和词典法来识别词汇并进行词频统计，主要统计语调计算公式所需变量的具体数据。本文采用的文本分析处理的主要方式如下。我们对年报文本进行分词处理，并去除停用词。我们利用情感词典筛选出所有的情感词汇。筛选分词后，采用词典法对文本进行匹配识别。这种方法主要通过词典对文本中的词频、词汇情感及词汇词性进行识别与统计，将词典中的情感词汇与文本内容相匹配，最终将非结构化的文本信息转化为结构化数据[11]。在运用词典法时，需考虑不同词典的适用领域和专业性。本文采用了经过英译汉处理的 LM 词典进行文本信息的分析处理[12]。

情感指标，亦称语调指标，主要检测和衡量文本信息中表达者的主观情感、喜好。在分析 MD&A 的语调时，通常需要拆解文本中的积极与消极信息，涉及对文本中的积极和消极词汇进行词频统计，最终采用公式计算语调指标。

参照邱静、杨妮的研究[13]，两种语调计算公式见式(2)和式(3)。由此公式计算得出的语调指标，范围区间均为[−1, 1]，当语调指标趋近于 1 时，MD&A 中的文本信息具有正面倾向；反之当语调指标趋近于−1 时，MD&A 中的文本信息具有负面倾向。

$$\text{Tone1} = \frac{\text{positive} - \text{negative}}{\text{total}} \tag{2}$$

$$\text{Tone2} = \frac{\text{positive} - \text{negative}}{\text{positive} + \text{negative}} \tag{3}$$

其中 Tone1 为语调指标一；Tone2 为语调指标二；positive 为积极词汇出现频次；negative 为消极词汇出现频次；total 为总词汇量(MD&A 文本中除积极消极词汇外还包含其他词汇)

Table 1. Examples of intonation indicator data for some sample companies
表 1. 部分样本公司语调指标数据实例

证券代码	positive	negative	total	Tone1	Tone2
000028	505	270	5720	0.0411	0.3032
000403	490	210	4506	0.0621	0.4000
002435	318	236	4352	0.0188	0.1480
300030	474	232	5755	0.0421	0.3428

3.2. 变量选择和数据来源

3.2.1. 样本范围

本文选取我国 2023~2024 年被实施 ST 或者*ST 的 18 家 A 股医药生物行业上市公司作为研究样本，并按照大约 1:5 的比例选取同行业的 91 家非 ST 公司作为配对样本。当公司出现“最近连续三个会计年度扣除非经常性损益前后净利润孰低者均为负值，且最近一个会计年度财务会计报告的审计报告显示公

司持续经营能力存在不确定性”的情形时，对其股票交易实施退市风险警示。因此，本文主要分析、研究 T-3 年的样本数据(T 为被实施 ST 或 ST*年份)。将这 109 家公司设置为建模组，对建模组样本前 3 年的财务数据利用二元 Logistic 回归分析方法建立财务危机预警模型。然后选取 2022~2023 年被实施 ST 或者*ST 的 10 家 A 股医药生物行业上市公司作为研究样本，选取同行业的 24 家非 ST 公司作为配对样本，将这 34 家公司设置为检验组，利用检验组样本前 3 年的财务数据验证模型的准确度。样本数据取自“搜狐证券”和国泰安数据库；样本年报的 MD&A 文本取自“东方财富网股吧论坛”。

3.2.2. 变量选择

本部分旨在探讨将财务指标和情绪指标整合到逻辑回归模型中的合理性，并阐述这种整合如何提高模型的预测准确性和应用价值。

对于财务指标，本文选取的财务指标应当全面反映企业的财务状况和经营状况，包括企业的偿债能力、营运能力、盈利能力和发展能力。偿债能力关注短期内的财务安全，营运能力和盈利能力反映当前的经营效率和盈利水平，而发展能力则关注企业的长期成长和扩张潜力。通常认为在这四个方面都表现良好的企业，才能被视为财务健康且具有良好的市场竞争力和发展前景。本文在财务预警指标选取时，主要参考学者杜明蔚的研究，筛选了 25 个财务指标。[14]

对于情绪指标，情绪指标能够较好地提高模型的预测准确性、增强模型的适应性，并优化决策支持系统。[15]在现有研究成果中，对于生物医药行业公司财务状况的分析并没有形成较为统一的情绪指标。因此将在模型构建时参考业内学者做法，使用上文构建的 Tone1 和 Tone2 语调指标。

将财务指标和情绪指标整合到逻辑回归模型中是一个合理且有价值的方法。这种综合模型将在金融分析、市场预测和消费者行为研究等多个领域发挥重要作用。综上，本文确定的初步指标集合如表 2 所示。

Table 2. Initial financial indicators for financial warning of listed companies in the biopharmaceutical industry
表 2. 生物医药行业上市公司财务预警的初始财务指标

一级指标	符号	二级指标	符号	二级指标
企业偿债能力	X_1	流动比率	X_5	资产负债率
	X_2	速动比率	X_6	长期债务与运营资金比率
	X_3	现金比率	X_7	产权比率
	X_4	现金流量负债比率	X_8	权益乘数
企业营运能力	X_9	应收账款周转率	X_{12}	总资产周转率
	X_{10}	存货周转率	X_{13}	资产报酬率
企业盈利能力	X_{14}	资产收益率	X_{18}	成本费用利润率
	X_{15}	净资产收益率	X_{19}	息税前营业利润率
	X_{16}	投入资本回报率	X_{20}	总资产增长率
	X_{17}	营业毛利率		
企业发展能力	X_{21}	净利润增长率	X_{24}	资本积累率
	X_{22}	营业收入增长率	X_{25}	每股经营活动产生的净流量增长率
语调指标	X_{26}	Tone1	X_{27}	Tone2

4. 实证结果和讨论

4.1. 指标筛选与模型构建

本研究通过结合文本挖掘技术和二元 logistic 回归模型，探讨了生物医药行业上市公司的财务预警模型构建问题。研究首先选取了包括现金比率、现金流量负债比率、资产负债率等在内的多个财务指标，以及通过文本挖掘技术实现并基于公司年度报告中管理层讨论与分析(MD&A)文本信息构建的语调指标。这些指标旨在全面评估公司的财务健康状况，并作为模型的关键预警指标。其中情绪指标对于财务指标的影响具有前瞻性特征。例如，正面情绪指标的提高往往预示着企业未来一段时间内财务表现的改善，而负面情绪的增加则可能是财务风险提升的前兆。这为管理者提供了重要的决策支持，提示他们应更加关注公开文本数据中的情绪变化，作为衡量企业财务健康状况的补充指标。将这些指标纳入二元 logistic 回归模型中。构建财务预警模型具体步骤如下。

指标筛选方面，借鉴徐小灿学者的研究做法[16]，本文将显著性检验作为指标筛选的关键步骤，主要原因如下：1) 确认变量有效性。这一步骤确保了模型仅包含对预测目标有实质性贡献的变量；2) 防止过拟合。有助于简化模型结构，减少模型复杂度；3) 提高模型稳健性。确保模型结果是稳健的，而不是随机变化的产物。

在确定了初步指标集合后，我们对逻辑回归模型进行了第一轮的显著性检验。本研究在保证模型稳健性的同时，尽可能地捕获与公司财务状况相关的所有潜在因素，因此设置了较为宽松的 P 值标准，即将 $P \leq 0.10$ 的指标视为在模型中具有统计学意义的显著变量。[17]初步的回归分析结果如表 3 所展示，反映了各指标在模型中的显著性水平。

Table 3. First regression index significance table
表 3. 第一次回归指标显著性表

指标代码	显著性	指标代码	显著性
X_1	0.994	X_{15}	0.385
X_2	0.954	X_{16}	0.618
X_3	0.478	X_{17}	0.045
X_4	0.104	X_{18}	0.398
X_5	0.040	X_{19}	0.269
X_6	0.928	X_{20}	0.037
X_7	1.000	X_{21}	0.828
X_8	1.000	X_{22}	0.179
X_9	0.375	X_{23}	0.067
X_{10}	0.314	X_{24}	0.977
X_{11}	0.825	X_{25}	0.023
X_{12}	0.898	X_{26}	0.200
X_{13}	0.098	X_{27}	0.325
X_{14}	0.134		

基于初次回归分析的结果，本文采取了逐步剔除法对模型进行迭代优化。具体来说，我们在每轮迭代中手动剔除显著性最低(即 P 值最大)的一个模型变量，然后重新建立回归模型。这一过程重复进行，尽可能保留了更多的显著变量，直至所有变量的显著性都小于等于 0.10 的标准。经过多轮迭代后，我们得到了优化后的模型指标集，其显著性结果如表 4 所示，表中所有指标显著性均已达标。特别是，企业管理者较为关注的现金流量负债比率和资产负债率等，在模型中的显著性水平低于 0.05，表明它们对于预测生物医药行业上市公司的财务危机具有重要影响。此外，基于文本挖掘技术提炼的情绪指标，其主要体现管理层语调情绪正负面强度和情绪变化趋势，也在模型中表现出显著性，这强调了非结构化文本数据在财务预警中的价值。

Table 4. Index significance table after optimization
表 4. 优化后指标显著性表

指标代码	显著性	指标代码	显著性
X_3	0.046	X_{20}	0.013
X_4	0.035	X_{22}	0.079
X_5	0.014	X_{23}	0.019
X_{15}	0.007	X_{25}	0.007
X_{17}	0.012	X_{26}	0.088
X_{19}	0.035		

在完成指标筛选与优化后，本文根据最终筛选出的指标构建了逻辑回归模型。选定的指标包括 X_3，X_4，X_5，X_{15}，X_{17}，X_{19}，X_{20}，X_{22}，X_{23}，X_{25}，X_{26}作为最终指标加入我们的模型，其他指标均剔除。这些指标覆盖了公司的不同财务维度，为我们提供了关于公司财务健康状况的全面视角。回归系数表如表 5 所示。基于这些系数，我们构建了以下回归方程模型，如公式(4)所示。

Table 5. Indicator B value (regression coefficient) table
表 5. 指标 B 值(回归系数)表

指标代码	B	指标代码	B
X_3	−3.496	X_{20}	−7.870
X_4	−7.793	X_{22}	3.382
X_5	−13.839	X_{23}	−5.977
X_{15}	−9.310	X_{25}	0.734
X_{17}	−6.215	X_{26}	58.109
X_{19}	4.746		

基于回归系数，我们构建了以下回归方程模型，如公式(4)所示。其中各自变量代表的具体指标含义已在表中详细说明，反映了从不同角度衡量公司财务健康状况的能力。该逻辑回归模型旨在通过分析这

些关键指标以预测公司是否会被特别处理(ST)，进而为投资这和管理者提供重要决策依据。

$$\ln\frac{P}{1-P}=14.908-3.496*X_3-7.793*X_4-13.839*X_5-9.31*X_{15}-6.215*X_{17}+4.746*X_{19}-7.87*X_{20}+3.382*X_{22}-5.977*X_{23}+0.734*X_{25}+58.908*X_{26} \quad (4)$$

4.2. 模型拟合优度检验和模型准确性检验

对于逻辑回归模型，进行拟合优度检验和模型准确性检验是评估模型性能的关键环节。这些检验帮助判断模型对数据的拟合程度以及预测的准确性，从而确保模型在实际应用中的有效性和可靠性。拟合优度检验可以量化模型解释变量对因变量的解释程度，帮助评估模型的解释力。本文选择适用性广且易于理解实施的霍斯曼－勒梅肖检验(简称 HL 检验)对模型进行拟合优度检验。通常认为检验显著性较大(通常大于 0.05)时，表明模型拟合程度良好。[18]如表 6 所示，模型的 HL 检验显著性达标，说明模型拟合程度良好。

Table 6. Haussmann-Lemecho test
表 6. 霍斯曼－勒梅肖检验

卡方	自由度	显著性
1.102	8	0.778

带入检验样本的各指标实际数据，检验模型准确度。其中 ST 公司 10 家，非 ST 公司 24 家。模型对于非 ST 公司判断 21 个为非 ST、3 个为 ST，判断准确率为 87.50%，对于 ST 公司判断 2 个为非 ST、8 个为 ST，判断准确率为 80.00%，模型整体准确率为 85.29%，这一结果表明，通过引入文本挖掘技术和财务指标的综合模型，对于生物医药行业上市公司的财务危机风险有一定的预警作用。财务指标能够较为直观的体现公司的具体财务状况，而情绪指标能够提现公司财务状况的潜在危机，起到较好的辅助参考作用。

财务预警模型旨在预测企业可能面临的财务困境或破产风险，传统模型多依赖于财务比率和指标进行分析。文本挖掘技术的引入，为财务预警模型增添了新的维度，即通过分析 MD&A 等文本资料中的情绪和语调，能够提前捕捉到企业潜在的风险信号。具体而言，负面情绪的增加往往预示着企业可能遇到的问题或挑战，而这些细微的变化可能在传统财务指标中并不明显。整体而言，本文通过将文本挖掘和逻辑回归模型相结合的方法，整体准确率达到 85.29%，说明可以帮助企业在初期预测公司财务状况并采取相应措施，是一个较为有效的财务危机预测方法。将文本挖掘与传统财务分析相结合，在一定程度上可以使财务预警模型的预测更为准确和全面。

[19]模型检验具体结果如表 7 所示，结果保留两位小数。

Table 7. Financial crisis early warning model prediction results table
表 7. 财务危机预警模型预测结果表

	ST 公司(预测)	非 ST 公司(预测)	正确率(%)	误判率(%)
ST 公司(真实)	8	2	80.00	20.00
非 ST 公司(真实)	3	21	87.50	12.50
总计百分比			85.29	14.71

综上所述，文本挖掘在本文的应用不仅证实了其在分析 MD&A 文本信息中构建情绪指标的有效性，而且还展示了其在丰富财务预警模型分析维度方面的独特价值。通过综合运用文本挖掘技术和传统财务分析方法，可以更全面地评估企业的财务健康状况，为投资决策和企业管理提供更为深入和多元化的视角。

5. 结论

5.1. 研究结论和优势

本研究基于文本挖掘和二元 logistics 回归模型，对生物医药行业上市公司的财务预警进行了较为深入地探讨。在经过筛选后，我们最终确定了现金比率、现金流量负债比率、资产负债率、净资产收益率、营业毛利率、息税前营业利润率、总资产增长率、营业收入增长率、资本保值增率、每股经营活动增长的净流量涨率等财务指标作为预警模型的关键指标，涵盖了多个方面，能够比较全面、多角度的评估公司的财务风险。本文还通过文本挖掘技术引入了语调指标，基于 MD&A 的文本情绪可以较显著的影响投资者对于宏观经济的预期，而投资者会根据预期调整金融市场参与程度，从而让市场回报产生相应的反应[20]。这使得模型能够充分利用非结构化数据，提高了预警模型的准确性和综合性。通过对模型的性能评估，我们发现该模型在非 ST 公司的判断准确率达到了 87.50%，在 ST 公司的判断准确率为 80.00%，整体准确率为 85.29%。说明可以帮助企业在初期观察财务数据变化趋势并采取相应的管理办法，是一个比较有效的财务危机预防方法[19]。通过文本挖掘和二元 logistic 回归模型的结合应用，本研究成功构建了生物医药行业上市公司的财务预警模型，不仅在理论上丰富了财务预警领域的研究，也在实践中提供了有效的财务风险管理工具。本研究突显了文本挖掘在提升模型预测性能方面的重要价值，展现了大数据技术与传统财务分析相结合的潜力。

5.2. 局限性

本研究提供了一种结合财务指标和文本挖掘技术的财务预警模型，具有一定的预测准确性和应用价值，但仍存在一些局限性。首先，由于数据完整性和实际操作的限制，样本量相对较小，这可能影响模型的泛化能力。其次，非结构化数据的来源相对单一，主要依赖于公司年报中的 MD&A 文本，未来的研究可以探索更多样化的数据来源，以进一步提高模型的预测能力和准确性。此外，生物医药行业的复杂性和市场环境的不断变化也为模型的长期有效性和准确性带来挑战，需要定期更新和调整模型以适应新的市场环境。

5.3. 对策建议

生物医药企业一般体量较大，产生财务危机的原因也复杂多样，比如销售收入较少但是成本规模大，盈利空间较小，或营销收入少并且现金短缺，导致企业偿债支付困难导致的亏损型财务困境，以及通过资产扩张来追求超常发展的盈利型财务困境等[1]。应该建立起持续监测财务风险的机制，及时更新预警模型，以应对生物医药行业市场环境的变化，确保模型的长期有效性。根据本研究模型的预警指标，公司可以对相应方面进行更密切的监控，比如维持适度的现金比率确保公司具备足够的流动性以应对突发情况。在面对不稳定的内外部环境时，随之调整经营策略，保持对财务危机的敏感性，提高核心竞争力，才能确保公司的稳健经营和可持续发展。

参考文献

[1] 李娜. 企业财务危机预警研究[J]. 现代商业, 2023(23): 189-192.

[2] 吴世农, 卢贤义. 我国上市公司财务困境的预测模型研究[J]. 经济研究, 2001(6): 46-55.
[3] 李莉. 财务预警研究的若干问题探讨[J]. 财会学习, 2018(26): 14-15.
[4] 宋慧斌. 企业财务预警系统的设计与实施[J]. 财会通讯, 2001(9): 26-27.
[5] 王永明. 基于文本挖掘的投资者情绪指标构建及其在量化投资策略中的应用[D]: [硕士学位论文]. 长春: 吉林大学, 2022.
[6] 邱会. 基于 PCA-Cox 模型的上市医药生物公司财务危机预警研究[D]: [硕士学位论文]. 南京: 南京信息工程大学, 2023.
[7] 刘抗英. 大数据背景下的财务信息管理系统风险估计[J]. 现代电子技术, 2020, 43(11): 79-82.
[8] 黄世忠, 叶丰滢, 李诗. 碳中和背景下财务风险的识别与评估[J]. 财会月刊, 2021(22): 7-11.
[9] 田青, 马越, 主编. 大数据计量经济分析[M]. 北京: 北京高等教育出版社, 2023: 4.
[10] 李东阳. 考虑文本信息的数字经济企业财务危机预警研究[D]: [硕士学位论文]. 成都: 成都大学, 2023.
[11] 甘翔翀. 引入文本信息指标的财务危机预警模型应用探究[D]: [硕士学位论文]. 南昌: 江西财经大学, 2022.
[12] 付梦瑶. 年报文本挖掘视角下的公司财务预警研究[D]: [硕士学位论文]. 北京: 华北电力大学, 2022.
[13] 邱静, 杨妮. 情感语调信号传递与企业融资约束[J]. 中南财经政法大学学报, 2021(5): 75-88.
[14] 杜明蔚. 京能电力财务风险管理研究[D]: [硕士学位论文]. 大庆: 东北石油大学, 2023.
[15] 犹梦洁. 基于文本挖掘的煤矿安全风险识别与评价研究[D]: [博士学位论文]. 徐州: 中国矿业大学, 2022.
[16] 徐小灿. 基于功效系数法的 A 煤炭企业财务风险预警体系构建研究[D]: [硕士学位论文]. 济南: 山东财经大学, 2023.
[17] 段文翔. 不同隧道路段下示廓灯与制动信号对车辆显著性影响研究[D]: [硕士学位论文]. 泉州: 华侨大学, 2022.
[18] 毛明春. 基于核拟合优度统计法的车道荷载极值模型中的阈值选取[J]. 中国市政工程, 2022(2): 66-71.
[19] 康楠. 制造业上市公司财务危机预警研究[D]: [硕士学位论文]. 兰州: 兰州交通大学, 2020.
[20] 姜富伟, 孟令超, 唐国豪. 媒体文本情绪与股票回报预测[J]. 经济学(季刊), 2021, 21(4): 1323-1344.