分类号：（F830.9；O211.63）　　单位代码：10363
密　级：公开　　学　　号：2211010106

题　　目 基于网络信息文本挖掘的信用风险评估分析——以新能源上市公司为例

题　　目 Credit risk assessment and analysis based on network information text mining -- A case study of new energy listed companies

学生姓名：丁沈杰
指导教师：张玥　教授
专　　业：金融
研究方向：金融大数据挖掘

论文答辩日期：2024 年 5 月 29 日

安徽工程大学学位论文原创性声明

本人郑重声明：我恪守学术道德，崇尚严谨学风。所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已明确注明和引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品及成果的内容。论文为本人亲自撰写，我对所写的内容负责，并完全意识到本声明的法律后果由本人承担。

学位论文作者签名：丁沈杰

日期：2024 年 5 月 29 日

安徽工程大学学位论文版权使用授权书

学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅或借阅。本人授权安徽工程大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密 □，在 ____ 年解密后适用本版权书。

本学位论文属于

不保密 ☑。

学位论文作者签名：丁沈杰 指导教师签名：张琨

日期：2024 年 5 月 29 日 日期：2024 年 5 月 29 日

基于网络信息文本挖掘的信用风险评估分析

——以新能源上市公司为例

摘 要

随着全球经济明显放缓，国内经济运行面临诸多困难挑战，部分企业因经不起市场考验而陷入信用风险危机。为了保障上市公司健康发展，保持我国宏观经济的平稳运行，如何准确、高效地评估上市公司信用风险成为了亟待解决的问题。因此，本文利用深度学习模型对网络文本信息进行特征提取；其次，对文本指标和财务指标进行异质信息融合，从而构建合理且完善的信用风险评价指标体系；最后，以机器学习算法为基础依照不同的组合策略进行决策融合，构建信用风险评估组合模型对企业信用风险进行预测，并进一步对不同指标进行量化分析，以深入探讨它们对不同经营范围企业信用风险的具体影响。本研究的重要性不仅在于信用风险模型的构建，更在于所构建的信用风险指标体系，能够为相关领域的研究提供宝贵的思路与方法。本文的研究内容主要分为三个部分：

（1）构建 Word2Vec-BiLSTM-Att 文本评价指标提取模型。首先本文利用中文金融情感词典对年报中管理层讨论与分析（MD&A）以及东方财富股吧标题文本信息进行情感标注，从而确保文本数据在情感维度上的准确性。随后，采用 Word2Vec 模型对标注后的文本进行向量化表示，通过 BiLSTM 模型对文本信息进行双向特征提取，并在输出端引入注意力机制，从而更好地捕捉序列中的依赖关系。

这不仅增强了模型对复杂数据的处理能力，还提升了模型的预测准确性，为信用风险评估提供有力保障。

（2）基于异质信息融合的信用风险评价指标体系构建。由于财务指标数据存在维度多、复杂程度高等特点，为避免指标之间的冗余和重复，本文采用遗传算法对财务指标进行筛选，实现数据的降维。文本指标反映了管理层和股民对目标企业的情感倾向，这些情感因素对企业信用风险的产生有着重要影响。因此，为了更全面地评估企业风险状况，本文将文本指标纳入降维后的财务指标体系中，然后运用 XGBoost 模型对其进行重要性分析，选出能准确反映企业信用风险状况的关键指标，构建信用风险评价指标体系。

（3）信用风险评估分析。由于企业信用风险标签存在不平衡的现象，将导致后续的风险预测模型出现过拟合，为达到平衡数据和提高后续预测模型的泛化能力，本文采用 DPC-SMOTE 算法对数据集进行扩充处理。鉴于单一风险预测模型存在准确性和泛化能力不能兼顾等问题，本文通过融合多种机器学习模型来提升预测性能。具体而言，本文将以逻辑回归、支持向量机和决策树为基础模型，通过 Bagging 和 Stacking 两种不同的融合策略将机器学习模型进行决策融合，构建机器学习组合模型对企业是否具有信用风险进行预测，从而得出预测结果和最佳组合策略。最后，通过 Logit 回归估计和异质性分析进一步厘清不同指标对不同经营范围企业信用风险的具体影响，从而为企业信用风险管理提供更加全面和准确的决策依据。

本文以 Wind 数据库 151 家新能源上市公司为研究样本，选取

2022 年年度财务数据和文本数据进行实证分析。研究结果表明：基于管理层讨论与分析和东方财富股吧标题提取的文本指标可以作为企业信用风险预测指标体系的有效补充，且 MD&A 文本指标与企业是否发生信用风险显著相关，股吧文本指标则对新能源企业具有显著影响。当数据不平衡时，逻辑回归、支持向量机、决策树和 Stacking 均出现了过拟合现象，Bagging 在样本数据比例为 1:1 和 1:3 情况下预测准确率和预测效果均表现最优，模型的预测准确率最高达到了 93. 5%。

关键词：新能源上市公司，信用风险预测，文本挖掘，异质信息融合，决策融合

CREDIT RISK ASSESSMENT AND ANALYSIS BASED ON NETWORK INFORMATION TEXT MINING -- A CASE STUDY OF NEW ENERGY LISTED COMPANIES

ABSTRACT

With the global economy slowing down significantly, the domestic economic operation is facing many difficulties and challenges, and some enterprises fall into credit risk crisis because of the test of the market. In order to ensure the healthy development of listed companies and maintain the steady operation of macro-economy in our country, how to accurately and efficiently assess the credit risk of listed companies has become the problem to be solved urgently. Therefore, this paper uses the deep learning model to extract features from the web text information. Secondly, the heterogeneous information of text index and financial index is integrated to build a reasonable and perfect credit risk evaluation index system. Finally, based on the machine learning algorithm, the decision fusion is carried out according to different combination strategies, and the credit risk assessment portfolio model is constructed to predict the enterprise credit risk, and the quantitative analysis of different indicators is further carried out to deeply explore their specific impact on the credit risk of enterprises

in different business scopes. The importance of this study is not only limited to the construction of credit risk model, but also lies in the credit risk index system constructed by it, which can provide valuable ideas and methods for the research of related fields. The research content of this paper is mainly divided into three parts:

(1) Construct the Word2Vec-BiLSTM-Att text evaluation index extraction model. Firstly, this paper uses the Chinese financial sentiment dictionary to annotate the sentiment of Management Discussion and Analysis in annual reports and East Money Gub Title text information, so as to ensure the accuracy of the text data in the emotional dimension. Then, Word2Vec model is used to represent the annotated text vectorically, BiLSTM model is used to extract the bidirectional feature of the text information, and Attention mechanism is introduced in the output side to better capture the dependencies in the sequence. This not only enhances the model's ability to process complex data, but also improves the model's prediction accuracy, providing a strong guarantee for credit risk assessment.

(2) Construction of credit risk evaluation index system based on heterogeneous information fusion. Due to the characteristics of multiple dimensions and high complexity of financial index data, in order to avoid redundancy and repetition among indicators, this paper adopts genetic algorithm to screen financial indicators and achieve dimensionality reduction of data. The text indicators reflect the emotional tendency of

management and shareholders towards the target enterprise, and these emotional factors have an important impact on the enterprise credit risk. Therefore, in order to evaluate the enterprise risk status more comprehensively, this paper incorporated the text indicators into the financial indicator system after dimensionality reduction, and then used XGBoost model to analyze its importance, select the key indicators that can accurately reflect the enterprise credit risk status, and build the credit risk evaluation indicator system.

(3) Credit risk assessment and analysis. Because of the imbalance of enterprise credit risk labels, the subsequent risk prediction model will be overfit. In order to balance the data and improve the generalization ability of the subsequent prediction model, this paper uses DPC-SMOTE algorithm to expand the data set. In view of the problems of accuracy and generalization ability of a single risk prediction model, this paper integrates multiple machine learning models to improve the prediction performance. Specifically, this paper uses logistic regression, support vector machine and decision tree as the base model, and uses Bagging and Stacking as two different fusion strategies to integrate machine learning models for decision making, and constructs a machine learning combination model to predict whether an enterprise has credit risk, thereby obtaining prediction results and optimal combination strategies. Finally, Logit regression estimation and heterogeneity analysis are used to further clarify the specific

impact of different indicators on the credit risk of enterprises in different business scopes, so as to provide a more comprehensive and accurate decision basis for enterprise credit risk management.

This paper takes 151 listed new energy companies in Wind database as research samples, and selects 2022 annual financial data and text data for empirical analysis. The research results show that the text indicators extracted based on Management Discussion and Analysis and East Money Gub Title can be an effective supplement to the enterprise credit risk prediction index system, and the MD&A text indicators are significantly correlated with whether the enterprise has credit risk, while the text indicators of the stock bar have a significant impact on new energy enterprises. When data is unbalanced, logistic regression, support vector machines, decision trees, and Stacking occur overfitting. Bagging has the best prediction accuracy and effect when the ratio of sample data is 1:1 and 1:3, and the prediction accuracy of the model reaches up to 93.5%.

Ding Shenjie (Finance)

Supervised by Zhang Yue

KEY WORDS: New energy listed companies, Credit risk prediction, Text mining, Heterogeneous information fusion, Decision fusion

目 录

第 1 章 绪论

1.1 研究背景及意义

1.1.1 研究背景

中国债券市场作为金融市场的重要组成部分，随着我国金融系统的不断发展，其发行总量快速上升，已成为我国实体经济至关重要的直接融资平台，并为经济的高质量发展提供了坚实的金融基础。然而，2014 年“11 超债日”后，我国债券市场出现了第一起重大违约事件，从而打破了债券市场的刚性兑付。自此，我国债券市场的违约规模和数量总体呈现出逐年递增的趋势，债券市场违约呈常态化。据 Wind 数据库统计，2014-2023 年末，违约债券数量由 6 只累计至 939 只，违约金额由 13.40 亿元累计至 6554.59 亿元；在 2023 年，我国债券市场呈现出违约风险相对降低的态势。具体而言，新增 5 家违约发行人，到期违约债券 17 期，到期违约金额总计约为 183.35 亿元。相较于 2022 年的数据（新增违约发行人 8 家，到期违约债券 45 期，到期违约金额总计 232.62 亿元）均呈现下降趋势，这一变化表明，我国债券市场在风险管理和防范方面取得了积极的进展，但仍处于一个违约事件频发阶段。债券违约具体情况如图 1-1 所示。

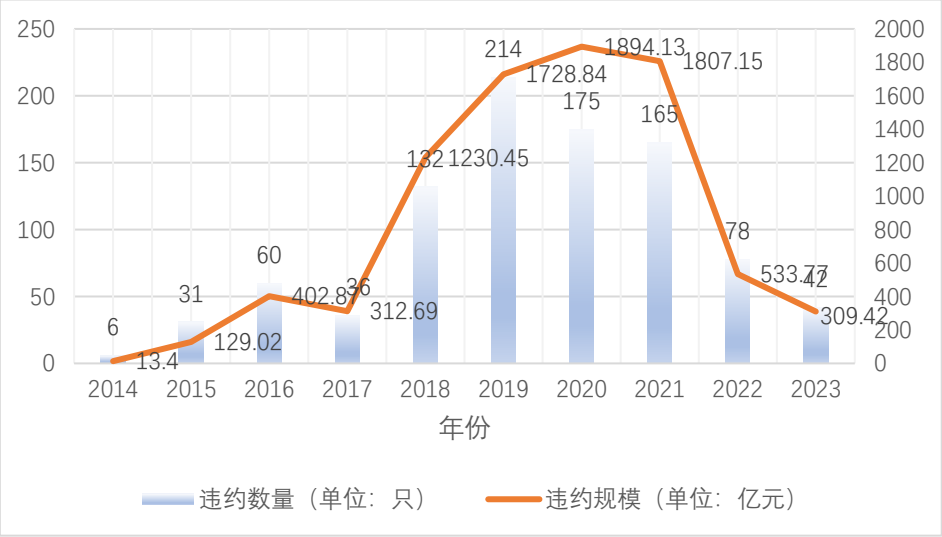

图 1-1 债券违约情况

近年来，随着中美经贸摩擦、巴以冲突等重大事件不断冲击世界贸易，全

球经济增速明显放缓，国内经济运行面临诸多困难挑战，部分企业因经不起市场考验，缺乏财务风险意识和正确的信用风险预警机制而陷入信用危机。与普通发债主体相比，上市公司发债主体具有信息披露及时、受监管程度高、融资渠道通畅等诸多优势，市场普遍认为上市公司相比于非上市公司有更低的违约可能性，但过去几年仍有部分上市公司出现债务违约，给投资者造成严重损失。同时，上市公司违约通常会导致市场风险波及范围更广，产生危机的可能性更大。因此，企业信用风险问题已经引起了整个社会的高度重视，通过采取一系列信用风险评估手段预测企业的信用风险状况，一定程度上可以让投资者规避投资风险，企业管理者借此改变经营策略，政府管理部门防范债券市场风险。

另一方面，随着大数据时代的到来和人工智能技术的持续进步，企业信息化水平不断提升，所产生的数据量呈指数级增长，这使得影响企业信用风险的因素变得愈发复杂多变，难以仅凭传统方法进行有效识别和评估。标准普尔早在 2003 年就指出，“传统研究常常忽视的定性信息中，蕴含着区分信用风险的关键信息”[1]，而早期的信用风险研究大多是以上市公司财务指标的单一数据来源，未充分挖掘非结构化的文本数据等，已经不能满足在大数据时代下的风险预测的需求，需要依靠更强有力的新技术来挖掘数据中的潜在信息。

最后，中国的高质量发展正处于一个至关重要的转折点。与此同时，碳达峰和碳中和目标的达成，更是我国在“十四五”时期必须全力推进的一项核心任务。在这一关键时期，新能源上市企业发挥着举足轻重的作用，它们不仅是推动我国经济绿色转型的微观主体，更是助力实现高质量发展目标的重要力量。

1.1.2 研究意义

（1）理论意义

影响企业信用风险因素繁多，对于非结构化的文本数据挖掘并没有统一的方法，同时有关文本数据挖掘的方法大多基于统计方法[2]-[4]，而运用深度学习方法进行文本数据分析的相关研究较少。因此，本文通过构建 Word2Vec-BiLSTM-Att 模型对文本信息进行提取，扩展了文本数据挖掘的研究手段。

在信用风险预警体系方面，通过构建 GA-XGBoost 模型对指标进行筛选，通过削减冗余指标，得到最佳风险评价指标体系。降低过拟合问题的同时还能

显著提高模型预测性能。运用不同的机器学习组合策略对企业信用风险进行预测分析，提高了模型预测的准确性和稳定性，丰富了企业信用风险预测模型的思路。

（2）现实意义

已有研究发现，企业债券的违约，影响不仅限于其自身，还会波及整个市场的隐性担保预期[5]。因此，为进一步完善企业信用风险预警问题，采用深度学习模型对非结构化的文本信息进行深度挖掘，并对指标特征进行筛选，在此基础上，构建一套全面、合理的新能源上市企业信用风险评估框架，实现对新能源上市企业信用风险的准确预测与评估。对于企业而言，这一框架能够帮助他们根据自身的经营状况，灵活调整发展战略，有效规避潜在风险；对于投资者和合作伙伴而言，该框架的应用让他们对公司的实际经营状况有更深刻的认识，从而做出更明智的投资决策，规避投资风险；对于政府管理部门而言，这一框架有助于他们及时识别并防范债券市场风险，进而促进债券市场的稳定发展，确保社会经济的平稳健康发展。

1.2 国内外研究现状

1.2.1 国外研究现状

国外在信用评估领域的研究起步较早，研究体系相对成熟和完整。早期的信用评估模式主要依靠专家的意见和经验进行定性分析，然而，由于这种模式的精度和稳定性极易受人为因素的影响，从而限制了其在大规模应用和复杂数据处理方面的有效性。在信息时代快速发展的背景下，企业所面临的信用风险因素和干扰因素日益增多，信用评估的复杂性和挑战性也随之增加。因此，越来越多的学者开始致力于探索和研究影响企业信用风险的关键指标因素，以寻求更为准确、稳定和高效的信用评估方法。

Title C 首次运用财务指标进行信用风险预测研究，结果发现净利润/股东权益和股东总权益/总负债这 2 个比率指标最能反映企业的信用状况[6]。此后，学者们在信用风险的研究领域中逐步引入更多的财务指标。Altman 选取 5 个财务比率构建多元判别分析 Z 模型对财务风险进行预测。这些定量信息之后被普遍

应用到财务风险预测研究之中[7]。Jens创新地将非财务指标纳入研究范畴，以期更全面地捕捉影响企业信用的各种因素[8]。Kim 等通过对比分析企业的核心财务指标，深入研究哪些指标能够显著影响企业信用评分，从而为信用评估提供更加准确、科学的依据[9]。Carlo 则强调了盈利能力在评估企业信用中的重要作用，他认为盈利能力不仅体现了企业的营销能力、现金获取能力、成本降低能力。还反映了企业的风险防范能力[10]。以上学者早期通过构建财务指标体系来评估分析企业信用风险，为未来的研究打下了坚实的基础。

Das 等在研究股票市场留言板的文本时首次提出情感分析概念，并将情感定义为积极观点和消极观点[11]。Dava 等针对文本中产品属性的意见提出意见挖掘概念[12]。随着大数据时代的到来，情感分析和意见挖掘也越来越广泛地被应用到企业信用风险管理方面，通过对不同文本数据进行信息挖掘，获得信用风险的影响因素指标，从而提高信用风险预测的精度和准确度。文本大数据为经典研究问题提供了新视角[13]，也可用于研究新的问题。按文本信息的来源大致可分为两类：一是企业内部文本信息，如企业年报等。Cecchini 等研究表明加入“管理层讨论与分析”可以提高预测准确率[14]，Gandhi 等发现公司发生财务困境的可能性随年报中的负面情感词汇数量增加而变大[15]，Donovan 等人发现MD&A 中的情绪指标提高了企业风险预测的准确性[16]；二是包括社交媒体、股吧评论等外部文本信息。Tetlock 等研究了《华尔街日报》专栏文章与随后股市收益和交易量之间的关系，发现消极词语频率上升时股市收益率会下降[17]。Hillert等使用1989—2010年45家美国报纸约220万条新闻数据研究了媒体关注度与股票市场动量效应的关系，发现受关注更高的公司的收益率预测性更强[18]。Lu 等和 Cerchiello 等运用文本挖掘方法，将提取和量化后的新闻报道内容应用于信用违约指标体系构建中，分类精度得到明显提升[19]-[20]。Sait 等研究结果表明社交媒体为确定公司的信誉提供了有价值的信息，且考虑社交媒体数据时，信用评级往往会下降[21]。

1.2.2 国内研究现状

相较于发达国家，我国在上市公司信用风险评估领域的起步相对较晚，基础稍显薄弱。然而，随着国内学者对企业信用风险评估技术的持续深入探究，

我国的企业信用风险评估指标体系和信用风险预测方法已经日趋成熟与完善。当前，我国不仅不断更新现有的信用评估指标体系和预测模型，还致力于提升其泛化性和预测准确率，以适应信息化时代的发展需求。

庞素琳、何毅舟等学者从风险环境的角度出发，构建了一个多层交叉信用评分模型，研究结果表明，财务净现值、财务内部收益率以及项目的抗风险能力共同构成了评估企业信用状况的关键指标[22]。田琨、庄新田等运用因子分析和 Logistic 回归建立汽车制造业上市公司信用风险评估模型，研究表明核心企业流动比率、资产负债率、净资产收益率、营业净利率为中小企业信用评级高度负相关[23]。白雪鹏和赵志冲通过对数似然函数值衡量指标及指标体系违约判别能力这一标准，建立含有 17 个指标的最优信用风险评价指标体系[24]。周志刚等通过 DBN 选取的财务指标对科创板上市公司的信用风险进行分析，结果表明偿债能力、成长能力、营运能力与信用风险显著相关，且盈利能力与信用风险密切联系[25]。

以上学者均以结构化的财务数据作为信用风险评价指标体系，随着计算机技术的不断发展，越来越多的学者开始研究非结构化的文本数据在企业信用风险预警体系中的作用。刘逸爽和陈艺云研究证明了管理层语调确实提高了信用风险预警模型的效力[26]。苗霞和李秉成研究了 MD&A 中超额乐观语调对企业财务危机预测的价值，且前瞻性信息中超额乐观语调水平越高，企业未来发生财务危机的可能性越低[27]。梁龙跃和刘波进一步提取 MD&A 和审计报告的文本特征，研究结果显示带有两种文本特征模型拥有更高的预测精度和 AUC 值[28]。李成刚等实证表明随着不同维度文本披露指标的加入，模型的准确性进一步提高[29]。宋彪、朱建明等在财务预警模型中引入全网网络数据，在短期内对预测效果有一定提高，从长期来看，对预测效果有明显提高[30]。崔炎炎和刘立新通过挖掘微博文本中的投资者情绪发现，网络舆情对金融科技股票收盘价预测具有重要作用[31]。宫晓莉等实证检验媒体情绪对企业风险承担的影响方式和途径，为企业如何利用媒体力量进行风险决策提供了有力参考[32]。范小云等提出的混合式方法应用于以新闻和股吧论坛为代表的中文金融文本，实证检验金融文本对股票市场具有一定的预测能力，且不同文本样本包含有不同的信息[33]。杨晓兰等以及段江娇等使用东方财富股吧相关企业帖子构建投资者情绪指数，并研

究投资者指数与收益率和市场变化的相关性[34]-[35]。沈艳等从媒体情绪、管理层情绪和投资者情绪三个文本情绪方面阐述了文本大数据在金融领域的重要应用[36]。以上学者研究表明文本数据可以作为风险预警体系的有效补充，为企业进行风险管理和投资者规避投资风险提供有力参考。

国内早期信用评估方法主要以单一模型为主，但单一模型容易受到自身的局限而难以通过对模型的优化同时提高精确和泛化性，因此，各种结合统计学方法和机器学习方法的组合模型应运而生。张新红和王瑞晓选择 Logistic 回归模型和 RBF 神经网络模型构建组合风险预警模型，解决了信用风险预警模型准确性和稳定性不能兼顾的问题[37]。陆健健则深入研究了随机森林（RF）、梯度提升决策树（GBDT）和极致梯度提升（XGBoost）三种集成算法在个人信用评估模型中的应用。他通过对比研究，依据相关多元评价指标的计算结果得出结论，基于 XGBoost 集成算法的个人信用评估模型在性能上表现最优[38]。迟国泰和王珊珊构建的 PSO-XGBoost 模型平均预测性能显著优于 13 种对比模型，并且增强了预测模型的可解释性[39]。吕秀梅等研究结果表明，构建的组合模型 DNN-SMOTEENN-ExtraTrees 分类能力更显著，泛化能力更加优异，更适合数据规模大、维度高的网络小贷市场评估信用风险[40]。以上学者研究表明，相较于单一模型，组合模型在分类能力、可解释性和泛化等能力上有更加优异的表现。

1.2.3 现状综述

综上所述，国内外文献在企业信用风险评价指标的构建过程中，已经逐步在结构化的财务数据基础上融入非结构化的文本数据，从而形成了更加全面、多元的风险评价指标体系。此外，原有单一预测模型也在向组合模型的方向演变，以应对复杂多变的数据处理需求。然而，当前研究依然存在一些明显的不足之处：

（1）我国目前的企业信用风险评价工作仍主要依赖已有的研究成果，以财务指标为核心构建指标体系。对于非财务指标的特征表示，往往依赖于统计方法或专家打分，且文本信息来源较为单一。因此，现有的信用风险评价指标体系在科学性和全面性方面仍有待加强。未来研究应进一步拓宽数据来源，结合先进的文本挖掘和机器学习技术，以构建更加科学、全面的企业信用风险评价

指标体系。

（2）在信用风险评估方面，目前仍以单一的预测模型为主，这在处理复杂数据和预测准确率上具有一定的局限。因此，为提高信用风险预测模型的准确率和泛化能力，组合模型得到了广泛应用。

鉴于此，本研究在财务指标选取上运用算法模型进行精细化的指标筛选。为了更全面、客观地反映评价对象的综合情况，本文选取不同来源的文本信息作为原始数据，并通过对这些文本信息进行深入的情感分析，实现定量的特征表示，从而进一步丰富评价指标体系。在构建评价指标体系的过程中，本文更多地依靠算法模型进行指标筛选，以此最大程度地降低由人为主观因素带来的影响。在风险预测中，通过模型的组合、集成等方法构建风险预测模型进行风险评估。

1.3 主要研究内容及论文结构

第一章，绪论，阐述本文的研究背景及意义，并且对国内外研究现状进行分析。

第二章，实验环境及相关理论，对相关理论知识和实验环境进行阐述。

第三章，主要内容是通过 Word2Vec-BiLSTM-Att 算法对两种来源不同的文本信息进行特征提取，获取包含文本特征的指标。

第四章，信用风险评价指标体系构建，通过将遗传算法筛选后的财务指标和文本指标相结合，并利用 XGBoost 进行特征重要性分析，从而选取信用风险关键影响指标，构建评价指标体系。

第五章，企业信用风险评估与分析，通过组合模型对企业信用风险进行预测分析，并依据 Logit 回归评估和异质性分析进一步厘清不同指标对不同经营范围企业信用风险的具体影响。

第六章，主要对本研究所获得的成果进行概括，同时，针对研究的不足之处对未来相关课题的研究方向进行展望。

第 2 章 实验环境及相关理论

2.1 实验环境

本研究的实验是在 Python3.9 上实现的，实验环境如表 2-1 所示：

表 2-1 实验环境

电脑型号：	机械革命 Code01 Ver2. 0
操作系统：	Windows 11 家庭中文版 （64 位）
处理器：	超微半导体 AMD Ryzen 7 6800H with Radeon Graphics 八核
显卡：	超微半导体 AMD Radeon(TM) Graphics （1024MB）
内存：	16GB(4800 MHz / 4800 MHz)

2.2 文本处理相关理论

2.2.1 Word2Vec 模型

Word2Vec 是语言模型中用来生成词向量工具的一种，被广泛应用于自然语言处理任务中。Word2Vec 模型结构简洁，主要包括输入层、隐藏层和输出层，其具有高效的训练过程。在 Word2Vec 中存在两种主要的训练模式：CBOW（Continuous Bag-of-Words Model）和 Skip-gram（Continuous Skip-gram Model）。这两种模式各有特点，并在不同的应用场景中展现出其独特的优势。

CBOW 模式是通过上下文来预测当前词，这种模式在处理具有相似上下文的词汇时，能够学习到它们语义的相似性，从而生成具有意义的词向量。其模型结构如图 2-1 所示：

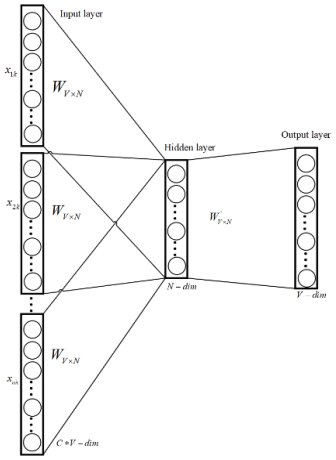

图 2-1 CBOW 结构图

Skip-gram 模式则是通过输入一个词去预测多个词的概率，这种模式在处理具有相似含义但上下文不同的词汇时，能够捕捉到它们之间的语义联系，进一步丰富词向量的语义信息。其模型结构如图 2-2 所示：

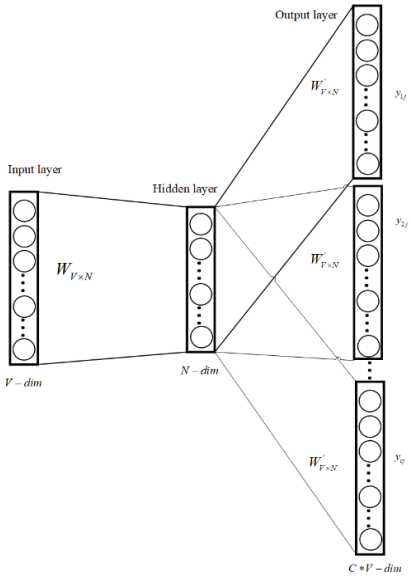

图 2-2 Skip-gram 结构图

2.2.2 长短期记忆神经网络模型

长短期记忆（Long Short-Term Memory，LSTM）网络模型解决了一般递归神经网络（Recurrent Neural Network，RNN）中普遍存在的长期依赖问题，LSTM 网络通过门控制的模型可以有效地传递和表达长时间序列中的信息，并且不会导致长时间前有用的信息被遗忘。LSTM 网络模型结构图如图 2-3 所示：

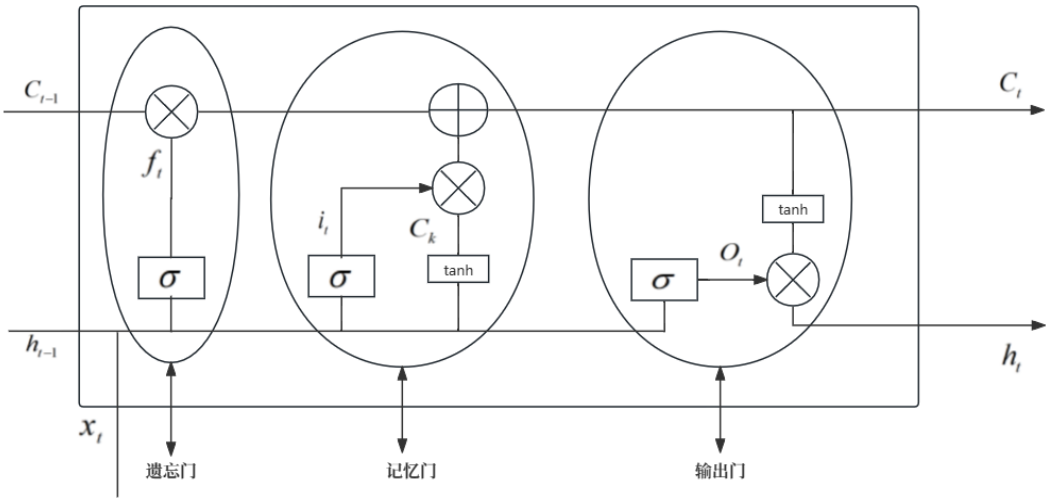

图 2-3 LSTM 神经网络结构图

首先是遗忘门的计算，公式如公式（2.1）所示

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + b_f) \tag{2.1}$$

其中，W_f ——权重向量；b_f ——遗传门参数；σ ——sigmoid 函数。

记忆门主要目的是从当前输入中提取有效信息，并对提取的有效信息做出筛选，评级越高的最后会有越多的记忆进入单元状态，如公式 （2.2）和公式（2.3）所示：

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i) \tag{2.2}$$

$$C_k = \tanh(W_C * [h_{k-1}, x_k] + b_C) \tag{2.3}$$

其中，W_i 和 W_C ——权重向量；b_i 和 b_C ——输入门参数。

输出门会先将当前输入值与上一时刻输出值整合后的向量（也就是公式中的 $[h_{t-1}, x_t]$）用 sigmoid 函数提取其中的信息，接着用 tanh 函数激活得到最终输出值，表述如公式（2.4）和公式（2.5）所示。

$$O_t = \sigma(W_o \otimes [h_{t-1}, x_t] + b_o) \tag{2.4}$$

$$h_t = O_t * \tanh(c_t) \tag{2.5}$$

2.2.3 注意力机制（Attention）

注意力机制关键是学出一个权重分布，然后作用在特征上。Attention 的引入，有效缓解了自然语言处理任务中长距离“记忆”难题。Attention 的计算可以分成以下三个阶段：阶段 1 通过打分函数，利用查询向量 q，对此时输入 Key 的部分进行打分；阶段 2 通过 SoftMax 将不同的打分做归一化处理，得到各个部分的注意力权重，也就是注意力分布；阶段 3 通过输入的 value 结合注意力分布进行信息聚合。具体流程如图 2-4 所示：

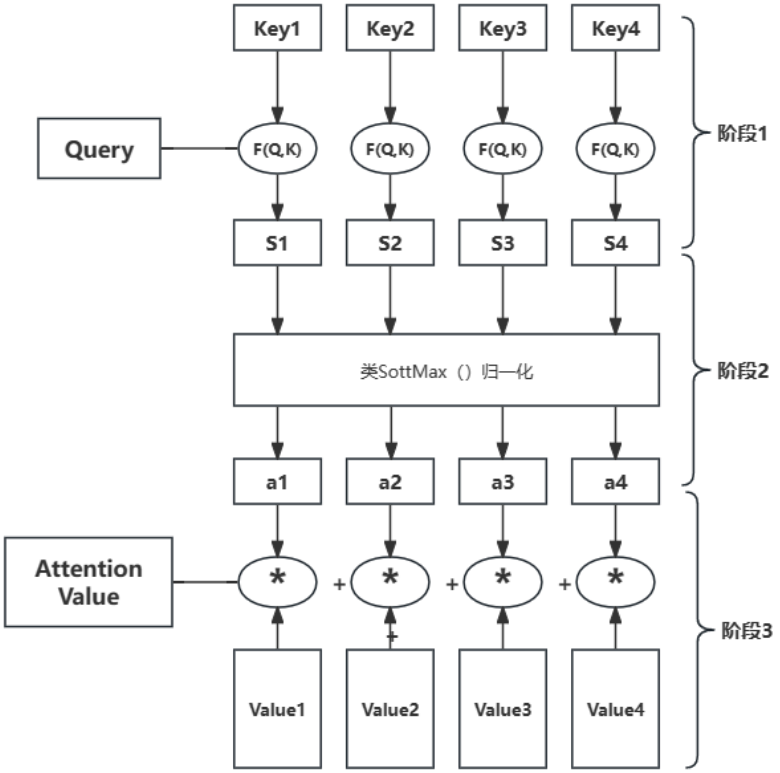

图 2-4 Attention 结构图

2.3 机器学习相关理论

2.3.1 逻辑回归模型

逻辑回归（Logistic Regression, LR）是一种常见的机器学习方法，主要用于样本分类任务。其模型表述如下：

$$S = \alpha + \sum_{i=1}^{n} \beta_i x_i \tag{2.6}$$

$$P = \frac{1}{1+e^{-s}} \tag{2.7}$$

其中，α 为常数项，β_i 为待估计系数；x_i 为财务风险的影响因素。P 为上市公司发生风险的概率值，其取值范围在 0 和 1 之间。本文借鉴吴世农和卢贤义[41]的研究，若P<0.5，代表违约概率较低，财务风险较小；反之，财务风险则较大。

2.3.2 支持向量机模型

支持向量机（Support Vector Machine, SVM）是一种二分类模型，由于其针对有限样本可以进行精确分类的特点，被广泛应用于信用风险预警领域。它的核心理念在于求解一个分离超平面作为决策曲面，这个超平面不仅能正确地将训练数据集划分开来，而且还追求最大的几何间隔，从而确保分类的准确性和稳定性。在给定的样本空间中，假设超平面的方程为：

$$\omega \bullet x + b = 0 \tag{2.8}$$

其中，ω 是法向量，b 是截距，x 是输入特征 $x = (x_1, x_2 \cdots x_n)$；而在样本空间中任意的一个样本点 $x_i, (i \in n)$ 到该超平面的距离 r 均可表示为：

$$r = \frac{\left|\omega^T x + b\right|}{\left\|\omega\right\|} \tag{2.9}$$

求解最大间隔超平面，就是要寻求能满足约束条件的参数 ω 和 b，并使得 r 值最大。

$$\min \frac{1}{2}\left\|\omega\right\|^2, s.t. \quad y(\omega^T x + b) \geq 1 \tag{2.10}$$

其中 y 的取值为 0 或 1，然后利用拉格朗日乘子法，对每一约束条件添加一个拉格朗日乘子 λ_i，从而得到 Lagrange 函数：

$$L(\omega, b, \lambda) = \frac{1}{2} \bullet 2\left\|\omega\right\|^2 - \sum_{i=1}^{n} \lambda_i [y_i(\omega^T x + b) - 1] \tag{2.11}$$

分别对 ω，b 和 λ_i 求偏导，并且使偏导值为 0 时可以得到：

$$\begin{cases} \dfrac{\partial L}{\partial \omega} = 0 \Rightarrow \omega = \sum_{i=1}^{n} \lambda_i y_i x_i \\ \dfrac{\partial L}{\partial \omega} = 0 \Rightarrow \sum_{i=1}^{n} \lambda_i y_i \\ \dfrac{\partial L}{\partial \lambda_i} = 0 \Rightarrow \lambda_i [(\omega^T x + b) - 1] = 0 \end{cases} \tag{2.12}$$

利用强对偶性对上述方程进行求解，求出 λ，ω，b 后，即得到分类函数式。

$$\max_{\lambda}[\sum_{i=1}^{n}\lambda_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j y_i y_j(x_i^T x_j)\ \ s.t.\ \ \sum_{i=1}^{n}\lambda_i y_i=0 \tag{2.13}$$

$$f(x)=\omega^T x+b=\sum_{i=1}^{n}\lambda_i y_i x_i^T x+b \tag{2.14}$$

对于非线性分类问题，可将样本从原始空间映射到一个更高维的特征空间，使得样本在这个特征空间内线性可分。

2.3.3 决策树模型

决策树（Decision Tree, DT）是一种基于树结构进行决策判断的模型，它通过一系列的条件判别过程，将数据集划分为不同的类别，进而得出所需的决策结果。通过这种方式，决策树能够有效地处理复杂的分类问题，并在信用风险评估中发挥出重要作用。从结构上来看，决策树将根节点作为起始节点，叶节点作为分类结果，而内部节点则为决策流程。首先训练样本被看作根节点，遍历其中所有变量，根据不同的划分准则将样本分为两个子集，然后分别对两个子集按最佳分割点进行划分，依次递归下去，直到所有子集足够“纯”为止。

根据节点划分依据的不同可以将决策树分为以下三种：分别是 ID3 决策树（基于信息熵和信息增益做分类）；C4.5（基于信息增益率做分类）；CART 决策树（基于基尼指数，既可以用于分类也可以用于回归）。相较于 ID3 决策树和 C4.5 决策树，CART 决策树不容易出现欠拟合现象，并且可以处理连续和离散特征。因此本文将采用 CART 决策树构建信用风险分类模型。数据 D 的基尼指数为：

$$Gini(D)=\sum_{i=1}^{n}p(x_i)*(1-p(x_i))=1-\sum_{i=1}^{n}p(x_i)^2 \tag{2.15}$$

其中，$p(x_i)$ 是分类 x_i 出现的概率，n 是分类的类别数目。$Gini(D)$ 表示随机从数据集中抽取两个样本，其类别标签不一致的概率。因此，$Gini(D)$ 越小，则分类效果越好。

2.3.4 极端梯度提升模型

极端梯度提升模型（Extreme Gradient Boosting, XGBoost）属于提升树模型，其算法思想是将许多弱分类器（即决策树）集成在一起，形成一个强分类器，这些个体学习器之间存在强烈的依赖关系，每一棵树的生成都依赖于前一棵树的输出。每一轮迭代都会通过特征分裂来生成新的决策树，新树的主要任务是拟合前一轮模型预测后剩余的残差值，即预测误差。对于每一棵生成的决策树，他都会为每一个特征样本提供一个具体的得分，最终，模型将所有决策树对应的得分进行累加，从而得到该样本的最终预测结果。

具体模型公式如（2.16）所示。

$$\hat{y}_i = \sum_{k=1}^{n} f_k(x_i),\ f_k = F \tag{2.16}$$

式中：$\hat{y}_i$ 表示对第 i 个样本的预测值；n 则表示数的数量；f_k 表示第 k 棵树的结构 q 和叶子权重 ω 相关状况；x_i 表示第 i 个点的特征向量；F 表示树的集合空间。

通过对树进行遍历迭代，极端梯度提升树的目标函数如下式所示：

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{2.17}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2 \tag{2.18}$$

其中 $l(y_i, \hat{y}^{(t-1)} + f_t(x_i))$ 为损失函数，$\Omega(f_t)$ 表示第 t 颗树的模型复杂度，这里是正则化项，通过减少树的深度和单个叶子节点的权重值来减缓过拟合。

在 $f_t = 0$ 处进行泰勒二阶展开处理，目标函数近似表示为

$$J(f_t) \approx \sum_{i=1}^{n}\left[L(y_i, \hat{y}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_i^2(x_i)\right] + \Omega(f_t) + C \tag{2.19}$$

通过公式（2.19）对 MSE 损失函数值进行计算，从而得到下式

$$J(f_t) = \sum_{j=1}^{T}[G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2] + \gamma T + C \tag{2.20}$$

其中，$g_i = L^{'}(y_i, \hat{y}^{(t-1)})$ 是第 i 个节点样本损失函数一阶梯度值，$h_i = L^{''}(y_i, \hat{y}^{(t-1)})$ 是第 i 个节点样本损失函数一阶梯度值。C 为常数；$G_j = \sum_{i \in I_j} g_i$ 表示第 j 个节点所有样本损失函数一阶梯度和；$H_j = \sum_{i \in I_j} h_i$ 表示第 j 个节点所有样本损失函数二阶梯度和。

对权重ω求偏导，并将ω的式子代回目标函数中，即可得到最优ω值以及目标函数值，如式（2.21）所示：

$$\omega_j^* = -\frac{G_j}{H_j + \lambda} \tag{2.21}$$

$$J(f_t) = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T \tag{2.22}$$

其中信息增益如公式（2.23）所示：

$$gain = \frac{1}{2}\left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{\lambda + (H_L + H_R)}\right) - \gamma \tag{2.23}$$

通过信息增益来对指标进行筛选，当信息增益越大，说明损失下降得越多，那当前的划分就越好。

第 3 章 基于 Word2Vec-BiLSTM-Att 的网络信息情感分析

网络文本信息蕴含丰富的情感倾向和观点表达，能够深刻反映出管理层和市场对企业未来发展的预期，为企业信用评价提供多维度的信息。然而，网络文本信息存在一定的噪音和不确定性。因此，本章运用自然语言处理技术，对年报中披露出的“管理层讨论与分析”（Management Discussion And Analysis，MD&A）和东方财富股吧标题（East Money Gub Title，EMGT）进行文本分析，构建 Word2Vec-BiLSTM-Att 模型得出企业经营者对其企业经营状况以及股吧吧友对有关公司预期的得分，为后续构建新能源上市公司信用风险评估指标体系奠定了坚实基础。

3.1 新能源上市公司网络信息文本数据说明

本章选择 Wind 数据库中我国 151 家新能源上市公司作为样本。管理层与讨论文本数据来自巨潮资讯网中 2022 年度报告，企业管理层讨论与分析的全部原始文本数据情况如图 3-1 所示；股吧文本数据为东方财富网 2022 年 1 月 1 日至 2022 年 12 月 31 日相关上市企业的股吧标题信息，共 3,206,831 条股吧标题数据，股吧文本信息原始截图（以东旭蓝天新能源股份有限公司 2022 年 1 月 1 日至 2022 年 12 月 31 日股吧数据为例）如图 3-2 所示。

代码	公司名称	日期	管理层讨论与分析
000040	东旭蓝天	2022-12-31	一、报告期内公司所处行业情况公司需遵守《深圳证券交易所上市公司自律监管指引
000413	东旭光电	2022-12-31	一、报告期内公司所处行业情况公司是一家以光电显示制造、新能源汽车制造为主营
000550	江铃汽车	2022-12-31	一、报告期内公司所处行业情况2022年，汽车产销分别完成2,702.1万辆和2,686.4万辆
000559	万向钱潮	2022-12-31	一、报告期内公司所处行业情况全球汽车行业周期与经济周期高度相关，一般 10 年
…	…	…	…
000591	太阳能	2022-12-31	一、报告期内公司所处行业情况根据国家能源局公开的数据，2022 年全国新增光伏装
688599	天合光能	2022-12-31	一、经营情况讨论与分析第一部分：2022 年公司总体经营业绩汇报 2022 年，全球能
688660	电气风电	2022-12-31	一、经营情况讨论与分析在“双碳”目标背景下，国家发展改革委、国家能源局等 9 部
688676	金盘科技	2022-12-31	一、经营情况讨论与分析 当前，世界百年未有之大变局加速演绎，世界之变，时代之
688819	天能股份	2022-12-31	一、经营情况讨论与分析1、业务进展：整体业绩逆势上扬，盈利能力恢复 2022 年，
835368	连城数控	2022-12-31	一、业务概要商业模式：公司是提供光伏及半导体行业所需晶体材料生长、加工设备

图 3-1 管理层讨论与分析原始文本数据

1	阅读数	评论数	标题	作者	时间
2	413	0	国电投新能源REIT和京能光	股市永动机888	12-31 12:47
3	480	0	难道侵占上市公司资产、不	股友29H5590M99	12-31 12:04
4	1843	5	东旭蓝天：公司应收的光伏	东旭蓝天资讯	12-31 10:11
5	…	…	…	…	…
6	424	0	md，东旭股票里的狗托太多	金银州鹤得段正	12-31 09:55
29775	507	0	2022年已到来，新年新气象	基民cP1Z6b	01-01 04:17
29776	544	5	出消息就代表涨到位了	有凝聚力得穆利	01-01 03:42
29777	403	1	意义非凡的一年是啥子意思	岭南私募	01-01 03:35
29778	2551	24	研判：关于补贴时间归属以	万手姨万手姨	01-01 02:56
29779	323	0	国家倡导新能源发展，长期	日夜思念上午	01-01 01:49
29780	763	1	在“碳中和”背景下，光伏	孤独的赢家	01-01 01:46
29781	719	1	截止11月末，东旭蓝天今年	孤独的赢家	01-01 01:44

图 3-2 东旭蓝天股吧原始数据截图

3.1.1 管理层讨论与分析

上市企业所披露的年报详细描绘了企业年度的生产经营与财务概况，这些信息是透明化的，旨在使投资者与股东们能对企业的运作状况有深入而全面的认识。这不仅是为了让股东们掌握企业运营脉络，更是为他们作出有益于企业发展的决策提供参考。“管理层讨论与分析”这一章在企业年报中尤为重要，此章节不仅是经营者对企业当前经营状况的客观陈述，更是他们结合外部政策环境与市场动态，对企业未来发展走向的深刻洞察与前瞻性规划。它更多地融入了经营者的主观判断和全局思考，不仅揭示了企业的发展脉络和潜在趋势，还展现了未来经营策略的制定与执行思路。因此，对于投资人、社会各界以及政府而言，这一章节不仅提供了关于企业未来的重要信息，更是建立对企业未来信心的重要依据。因此，本文将以年报中的“管理层讨论与分析”部分为研究对象，通过截取并删除其他非核心内容，精准提取出关键信息，为后续深入的分析与研究奠定基础。

3.1.2 东方财富股吧

东方财富股吧中包含了对个股进行各种各样的讨论，形式涵盖新帖发布与帖子内容的评论。当某一事件引起广泛关注时，股吧中便涌现出针对该事件的发帖与热烈讨论。股吧标题作为重点事件的互动平台，其发帖量、评论量、阅

读量等指标往往能够反映投资者对个股未来走势的情绪倾向。因此，本文聚焦于东方财富股吧标题内容的分析，旨在提取主要信息，为后续深入分析处理奠定基础。

3.2 文本数据处理

3.2.1 文本表示

鉴于企业所拥有的文本数据量过于庞大，后续的文本向量化表达无法直接进行。因此，本文运用 Python 中的 HarvestText 工具包，对文本信息进行精细的分句处理。处理过程中，我们主要依据“。！？……”等标点符号作为分句的标准，以确保分句结果的准确性和合理性。Python 中的 Jieba 分词库是一个被广泛用于自然语言处理的中文分词组件，分为全模式和精确模式两种，精确模式对于词语的切分更为准确，适合于文本分析，因此，本文采用精确模式的 Jieba 分词对文本内容进行分词处理。最后利用 Word2Vec 模型对词语进行编码，一个词语将对应一个多维向量，通过将所有的词向量化，将文本的内容转换为向量来表示。

3.2.2 文本情感标注

由于 BiLSTM 是一种有监督学习模型，它依赖于带有情感标注的文本数据进行训练，然而，当前文本数据信息中情感标注缺失，使得无法进行后续的模型训练。因此，需要对文本数据进行情感标注，这一过程旨在为每个文本数据分配具体的情感得分，为后续的模型学习训练提供必要的监督信息。为避免情感标注受研究者主观性的影响，本文将采用词袋法划分不同情感的特征词。在词典的选取方面，考虑到金融领域词语的特殊性，将选择姜富伟等[42]以 LM 英文金融情感词典和中文通用情感词典为基础构建的中文情感词典。具体的标注规则如下：遍历每个分句，并依据词汇的情感倾向进行得分计算。当检测到正面词汇时，得分增加 1 分；若出现负面词汇，则得分减少 1 分；而中性词汇对得分无影响。根据最终得分情况，为文本信息赋予相应的情感标签，若最终得

分大于 0，则标签为 2，代表正面情感；若最终得分等于 0，则标签为 1，代表中性情感；若最终得分小于 0，则标签为 0，代表负面情感。通过这一规则，最终得以计算出每个分句的情感得分。

3.3 Word2Vec-BiLSTM-Att 模型构建及情感分析

3.3.1 BiLSTM 模型

双向长短期记忆神经网络模型（Bi-directional Long Short-Term Memory，BiLSTM）是由前向 LSTM 和后向 LSTM 组合而成的。相较于单向长短期记忆神经网络模型，BiLSTM 可以同时从前向和后向进行计算，可以更好地捕捉序列中的信息，提高信息的提取能力和模型的准确性。BiLSTM 模型结构如图 3-3 所示：

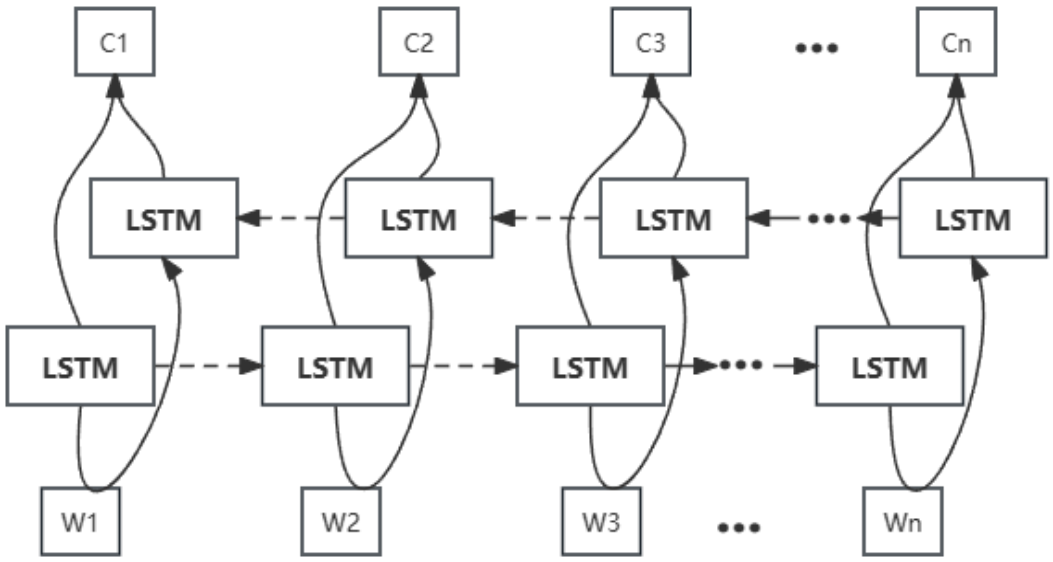

图 3-3 BiLSTM 模型结构图

3.3.2 Word2Vec-BiLSTM 模型

Word2Vec-BiLSTM 模型是一种将 Word2Vec 与双向长短期记忆网络（BiLSTM）相结合的深度学习模型，旨在通过结合两者的优势，实现对文本信息的高效特征提取和情感分析。首先，Word2Vec 模型被用于对文本数据进行

预处理和特征提取。通过训练大规模的语料库，Word2Vec 能够将文本中的单词转化为低维、稠密的向量表示。这些词向量能够捕捉到单词之间的语义和语法关系，从而有效地表示文本信息的特征。最后，这些包含丰富语义信息的词向量被作为输入，传递给 BiLSTM 模型。通过 BiLSTM 的层叠结构和门控机制，模型能够逐步提取出文本中的深层次特征，并将其整合为最终的情感得分。Word2Vec-BiLSTM 模型结构如图 3-4 所示。

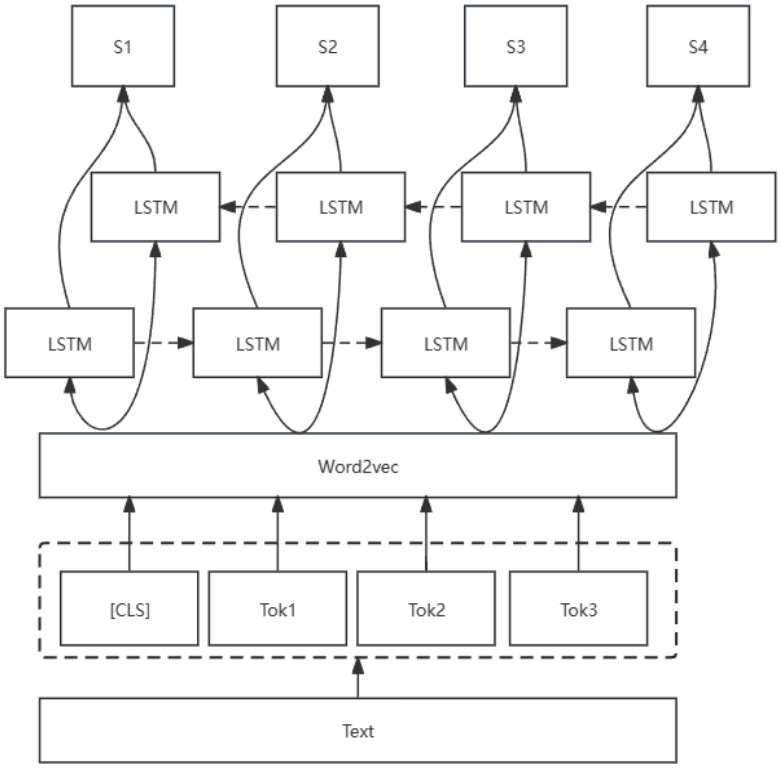

图 3-4 Word2Vec-BiLSTM 模型结构图

3.3.3 Word2Vec-BiLSTM-Att 模型

BiLSTM 在处理序列时，对于每个时间步都给予相同的权重，这可能导致模型无法专注于那些真正重要的信息。而 Attention 模型则能够根据输入数据的重要性动态地调整每个时间步的权重，使得模型更加关注那些关键信息；其次，通过引入 Attention 模型可以在处理每个时间步时，考虑到前面和后面的信息，从而更好地捕捉序列中的依赖关系。因此本文实验在 BiLSTM 模型后加入 Attention 模型以更好地捕捉序列中的依赖关系，从而提高模型的性能和准确性。Word2Vec-BiLSTM-Att 模型架构如图 3-5 所示。

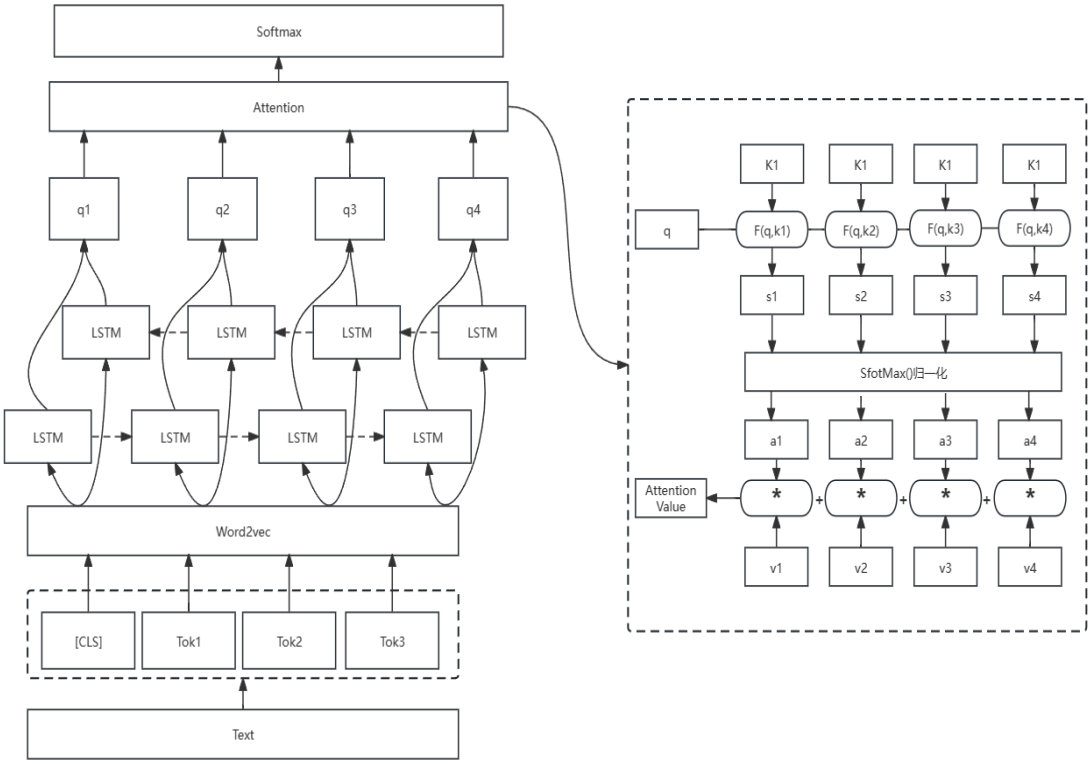

图 3-5 Word2Vec-BiLSTM-Att 模型架构图

本实验使用 python 语言作为主要编程语言，基于 keras 作为编程框架进行代码编写，其中调用了其他工具包，最终完成实验操作。具体实验过程中使用到的软硬件环境如表 3-1 所示。

表 3-1 软件版本

参数	版本
Python	3.9
Keras	2.13.1
Genism	4.3.2
Sklearn	0.0.post7
Numpy	1.24.3
Jieba	0.42.1
开发环境	Pycharm Commonity Edition

3.4 本章小结

本章文本挖掘主要实验流程首先利用爬虫技术获取新能源上市公司的年报和股吧标题信息，并利用 python 中 HarvestText 和 jieba 工具包对原始文本数据集进行预处理。其次，本研究采用以英文金融情感词典和中文通用情感词典为基础构建的 LM 中文情感词典，按照既定规则对文本进行情感标注，从而生成了一个全新的数据集，用于后续的模型学习训练。随后，利用 Word2Vec 语言模

型将文本数据转化为向量表示，并作为输入数据导入 BiLSTM 模型中，以进行更深入的情感分析。由于 BiLSTM 模型基于相同的权重可能忽视了真正重要的信息，因此本文将 BiLSTM 与注意力机制相结合，最大程度对文本信息进行挖掘，从而得到情感得分。基于 Word2Vec-BiLSTM-Att 模型进行文本信息挖掘的流程图如图 3-6 所示。

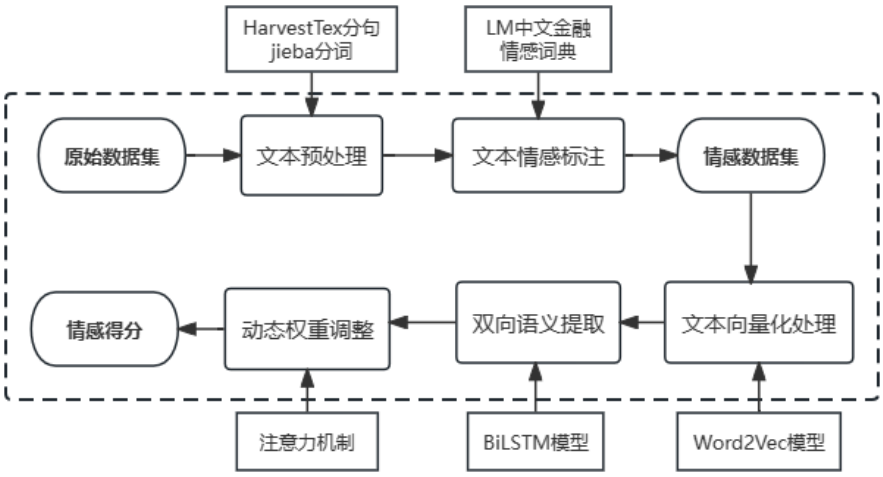

图 3-6 文本信息挖掘流程图

第 4 章 融合异质信息的信用风险评价指标体系构建

本章主要是信用风险评价指标体系的构建，自 Altman 首次将多变量分析技术应用于企业财务危机预警系统以来，目前已经形成了大量的财务指标数据。但是，数据过多容易带来噪声，导致学习算法出现过拟合，从而模型的泛化能力降低，对此，本文使用启发式随机搜索策略中最具代表性的遗传算法进行初步财务风险指标筛选，并将初步筛选后的财务指标与文本情感值相结合。其次对于新能源上市企业出现样本标签数据集的不平衡现象，本章运用基于密度峰值聚类算法的少数类样本合成过采样技术（Synthetic Minority Oversampling Algorithm Based on Density Peaks Clustering，DPC-SMOTE）[43]对数据进行补充，从而解决因数据失衡导致模型过拟合问题。最后采用 XGBoost 模型对结合后的指标进行筛选分析，选取影响程度较大的指标，构建合理的评价指标。

4.1 信用风险财务指标的来源及选取

本文将 Wind 数据库中 151 家新能源上市公司作为研究样本，选取国泰安数据库中（CSMAR）2023 年半年报的财务杠杆作为风险水平分类标签，其中有风险的公司为 38 家，无风险的公司 113 家。企业财务指标为 2022 年 12 月 31 日，均来自于国泰安数据库。结合我国新能源上市公司的具体情况，借鉴国内外有关信用风险指标选取的研究[44]，从偿债能力、发展能力、经营能力、盈利能力、现金流水平以及研发投入情况 6 个维度，共选取具有代表性的 36 个财务指标，具体如表 4-1 所示。

表 4-1 财务指标体系

评价维度	财务风险影响因素
偿债能力	流动比率(X_1)、速动比率(X_2)、现金比率(X_3)、产权比率(X_4)、资产负债率(X_5)、现金流到期债务保障倍数(X_6)、经营活动产生的现金流量净额／负债合计(X_7)
发展能力	总资产增长率(X_8)、营业总收入增长率(X_9)、可持续增长率(X_{10})、所有者权益增长率(X_{11})、每股净资产增长率(X_{12})
经营能力	应收账款周转率(X_{13})、存货周转率(X_{14})、营业周期(X_{15})、应付账款周转率(X_{16})、流动资产周转率(X_{17})、固定资产周转率(X_{18})、非流动资产周转率(X_{19})、总资产周转率(X_{20})

续表 4-1 财务指标体系

评价维度	财务风险影响因素
盈利能力	总资产净利润率(X_{21})、净资产收益率(X_{22})、长期资本收益率(X_{23})、营业毛利率(X_{24})、营业净利率(X_{25})、总营业成本率(X_{26})、财务费用率(X_{27})
现金流水平	营业收入现金含量(X_{28})、公司现金流(X_{29})、股权现金流(X_{30})、全部现金回收率(X_{31})、营运指数(X_{32})、现金适合比率(X_{33})、现金再投资比率(X_{34})
研发投入情况	研发人员数量占比(X_{35})、研发投入占营业收入比例(X_{36})

4.2 基于遗传算法的财务指标筛选

4.2.1 遗传算法

遗传算法（Genetic Algorithm，GA）是一种基于自然选择和遗传机制的随机全局搜索优化方法。该方法通过模拟生物进化过程中的复制、交叉和变异等遗传操作，从任意初始种群出发，经过一系列迭代优化过程，逐步产生出更加适应环境（即优化问题解空间）的个体。在每一轮迭代中，算法通过随机选择、交叉配对和基因突变等手段，促使种群向着搜索空间中更优的解区域进化。经过不断繁衍与进化，最终收敛至一群最适应环境的个体，从而求得问题的近似最优解。其具体步骤如下：

步骤 1：编码染色体，初始化种群 $M(t)$；

步骤 2：根据适应度函数计算种群中每条染色体的适应度值；

步骤 3：While 是否满足停止条件？

步骤 4：　基于适应值的选择；

步骤 5：　按一定的交叉概率染色体进行交叉操作；

步骤 6：　按一定的变异概率染色体进行变异操作；

步骤 7：　选择、交叉和变异操作之后。产生新的种群 $M(t+1)$，返回步骤 2；

步骤 8：获得最优全局解。

4.2.2 基于遗传算法的实证分析

本文使用遗传算法对新能源上市公司的财务指标进行了特征提取，运用遗

传算法对所选财务指标进行 100 次迭代，在第 83 次迭代时达到最佳精度（best finess）为 92.06%，通过遗传算法特征选择结果如图 4-1 所示。基于实验结果所示，本文将从 36 个财务指标中选取 19 个财务指标，分别为偿债能力（X_1、X_2、X_3、X_6）、发展能力（X_9、X_{11}、X_{12}）、经营能力（X_{13}、X_{15}、X_{16}、X_{18}、X_{19}）、盈利能力（X_{23}、X_{26}、X_{27}）以及现金流水平（X_{28}、X_{29}、X_{30}、X_{31}）。

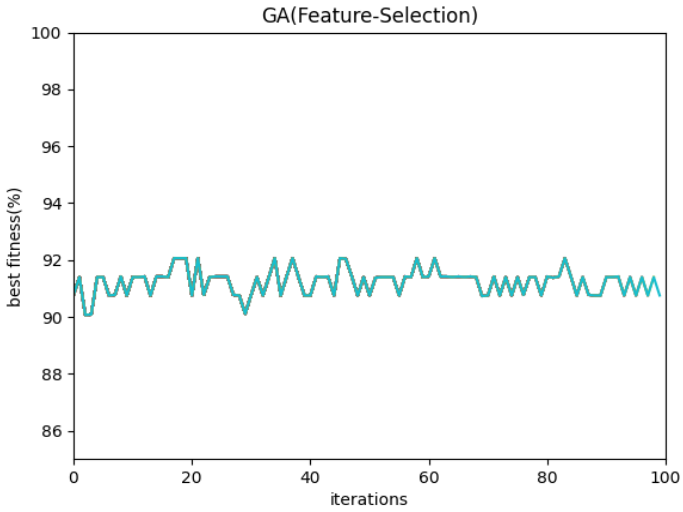

图 4-1 遗传算法特征选择实验结果

同时将管理层讨论与分析以及东方财富股吧标题文本情感得分纳入信用风险评价指标体系，不仅可以从管理层的视角洞察企业过去的经营细节和未来的战略规划，还能从广大股民的声音中捕捉到市场的真实情绪和预期。这种多维度、多角度的分析评价，有助于我们更客观、更全面地了解企业面临的经营风险。通过深入剖析管理层对企业自身情况的剖析，以及市场对企业前景的预期，我们可以更准确地把握企业信用风险的实际状况，为投资决策提供有力支持。得到的具体指标如表 4-2 所示。

表 4-2 融入文本信息后的评价指标体系

评价维度	指标名称
偿债能力	流动比率、速动比率、现金比率、现金流到期债务保障倍数
发展能力	营业总收入增长率、所有者权益增长率、每股净资产增长率
经营能力	应收账款周转率、营业周期、应付账款周转率、固定资产周转率、非流动资产周转率
盈利能力	长期资本收益率、总营业成本率、财务费用率
现金流水平	营业收入现金含量、公司现金流、股权现金流、全部现金回收率
文本信息	管理层讨论与分析、股吧标题

4.3 评价指标体系的构建

4.3.1 DPC-SMOTE 模型

由于在新能源上市企业中实际存在经营信用风险的企业在全部企业中占少数，会出现标签集不平衡现象，导致企业信用风险预测结果容易出现过拟合现象。针对分类研究中出现的数据集不平衡的问题，刘志涵等学者提出了基于密度峰值聚类算法的少数类样本合成过采样技术，并实证表明该方法能够减少样本重叠，有效避免不平衡数据集中噪声的产生，提升分类精度[43]。DPC-SMOTE 算法主要包括以下三个步骤：采用 DPC 算法对少数类进行聚类、计算每个簇的稀疏度并根据稀疏度计算采样权重，并根据采样权重计算需要合成的样本数目，采用 SMOTE 算法在簇内合成新少数类样本。其算法流程如图 4-2 所示：

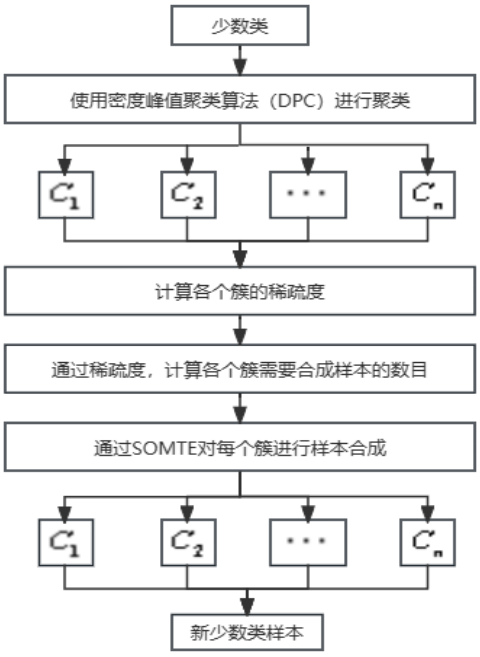

图 4-2 DPC-SMOTE 算法流程图

4.3.2 数据处理

为提高数据可比性，首先对数据进行标准化处理，通过 Z-score 标准化将不

同量级数据转化为统一度量的 Z-score 分值进行比较。标准化公式如公式（4.1）所示：

$$Z=\frac{x-\mu}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2}} \tag{4.1}$$

其中，μ 为总体数据的均值；$\sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2}$ 为总体数据的标准差。

151 家新能源上市企业中，其中企业经营情况较好的企业为 113 家，占比较大，而经营存在风险的企业占比较少，仅为 38 家，无风险与有风险企业的比例约为 3：1。为避免因样本数据不平衡导致预测结果出现过拟合的现象，本文采取 DPC-SMOTE 算法将无风险与有风险企业按照 1：1 的比例进行扩充。同时对扩充的指标进行相关性分析，并绘出指标间的相关系数热力图，如图 4-3 所示。

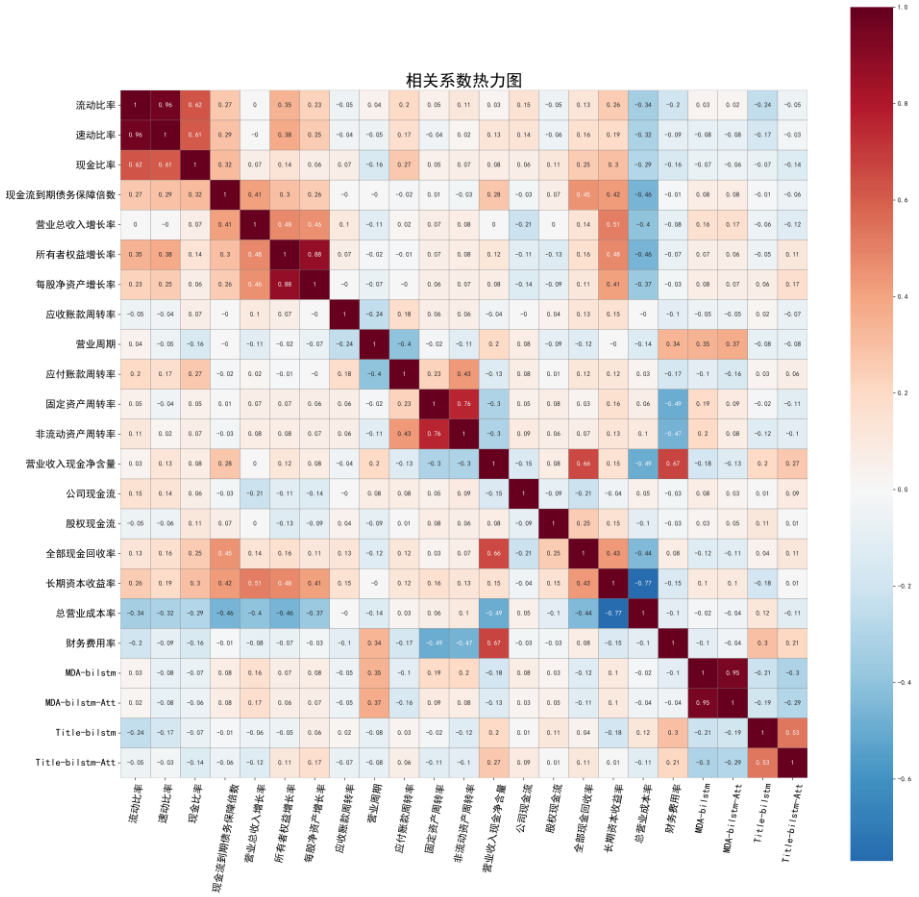

图 4-3 指标相关系数热力图

如图4-3所示，部分指标之间存在显著的相关性。为确保企业信用风险预测模型的准确性，首先需要剔除那些高度相关的风险评价指标，以避免冗余信息和潜在的预测干扰。为此，本文采用 XGBoost 模型进行数据的降维处理。这种方法不仅有效地去除了数据中的冗余信息，还保留了那些对预测结果至关重要的核心特征。最终，我们采用科学的方法，将这些经过筛选和优化的指标整合成一个完整的风险评价指标体系。

本文通过前期对数据的处理，共获取 189 条企业数据，23 个信用风险评价指标数据。最后将利用 XGBoost 模型对指标进行筛选，构建新能源上市企业信用风险评价关键指标。其中模型的初始超参数设置如下：随机取样比例为 0.8，目标函数为逻辑回归函数，学习率为 0.2，增强器选用梯度提升树。通过 XGBoost 模型，得到各个特征指标的重要性排名以及特征指标数和相应模型准确率的关系，分别如图 4-4 和图 4-5 所示。同时为了直观比较大小，输出个特征的增益、权重和重要性如表 4-3 所示。

表 4-3 特征指标重要性

特征指标	增益	权重	重要性	特征指标	增益	权重	重要性
流动比率	0.30	4	0.025	全部现金回收率	0.35	6	0.010
速动比率	0.53	10	0.027	公司现金流	1.00	18	0.072
现金比率	0.75	20	0.032	股权现金流	0.42	12	0.020
现金流到期债务保障倍数	0.47	10	0.019	营业收入现金净含量	0.63	12	0.025
营业总收入增长率	0.72	10	0.044	长期资本收益率	1.50	46	0.119
所有者权益增长率	0.95	16	0.066	总营业成本率	0.76	6	0.032
每股净资产增长率	0.62	18	0.051	财务费用率	4.13	32	0.228
应收账款周转率	0.38	4	0.006	MDA-bilstm	0.18	6	0.019
营业周期	0.24	12	0.028	MDA-bilstm-Att	0.86	20	0.037
应付账款周转率	0.36	14	0.019	title-bilstm	0.26	16	0.023
固定资产周转率	0.74	12	0.027	title-bilstm-Att	0.29	16	0.034
非流动资产周转率	0.27	12	0.038				

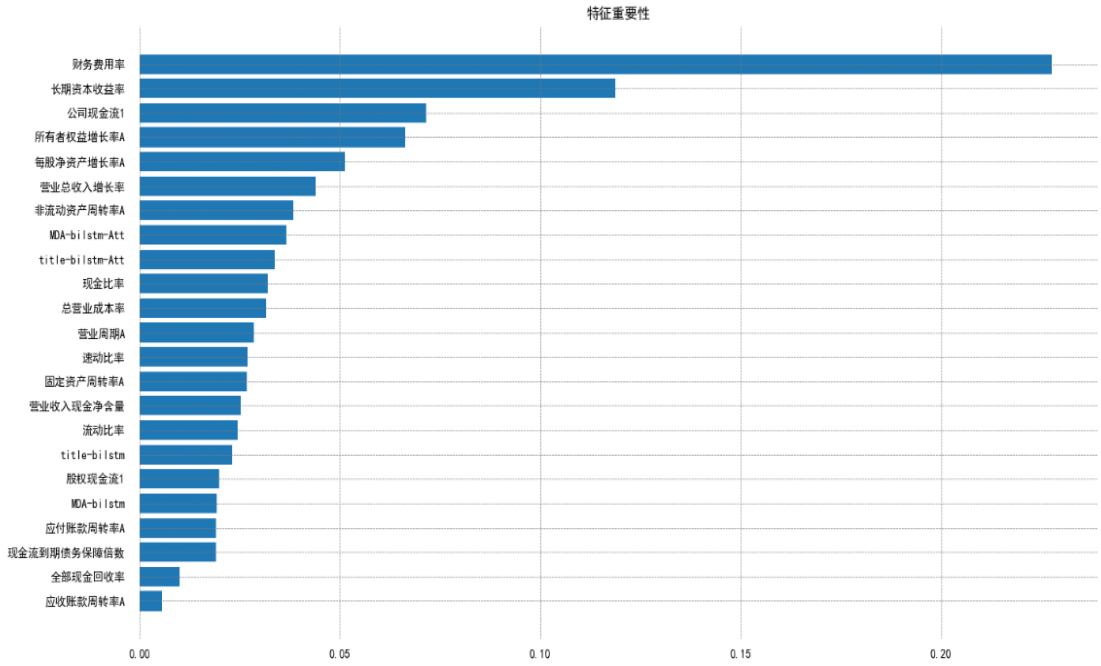

图 4-4 特征重要性

由图 4-4 可知，财务费用率、长期资本收益率、公司现金流、所有者权益增长率、每股净资产增长率、营业总收入增长率、非流动资产增长率、MDA-bilstm-Att、title-bilstm-Att 和现金比率占据了特征重要性前十的位置。而仅基于 BiLSTM 模型进行文本信息挖掘的特征指标重要性明显低于 BiLSTM-Att 模型，因此后续指标体系构建中选取以 BiLSTM-Att 模型得到的文本信息指标。

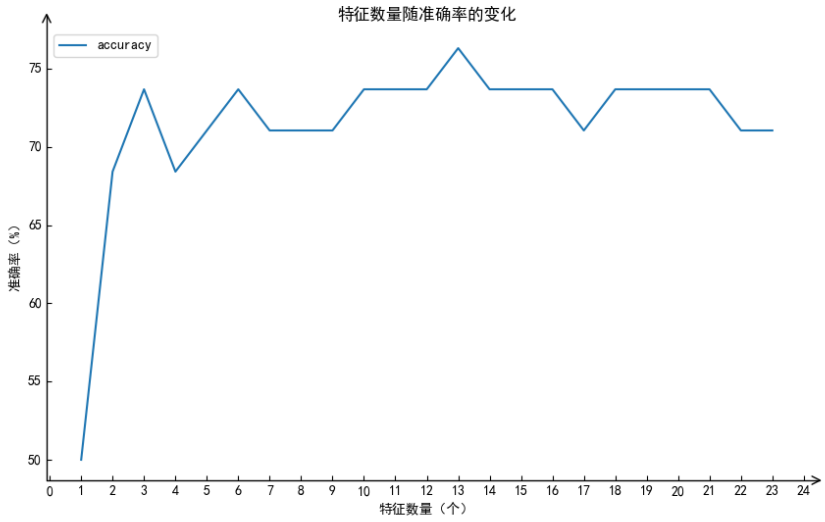

图 4-5 特征数量和准确率关系

根据图 4-5 的结果表明，当特征数量选取为 13 时，XGBoost 模型的预测准确率最高，结果更具有可靠性。因此本文将选取财务费用率、长期资本收益率、公司现金流、所有者权益增长率、每股净资产增长率、营业总收入增长率、非流动资产周转率、MDA-bilstm-Att、title-bilstm-Att、现金比率、总营业成本率、营业周期和速动比率作为新能源上市企业信用风险评价指标。

4.4 本章小结

本章首先借鉴国内外研究从六个方面共选取 36 个信用风险评价指标，利用遗传算法对风险指标进行初步筛选，并将文本信息加入筛选后的风险指标中。其次针对样本标签不平衡的问题利用 DPC-SMOTE 算法对数据进行扩充，达到数据平衡从而避免过拟合的目的。最终运用 XGBoost 模型得出各个特征指标的重要性排名，结合在不同特征数量情况下相应的准确率，最终得出前 13 个特征指标作为新能源企业信用风险的影响因素，为预测企业信用风险工作做好准备。

第 5 章 企业信用风险评估与分析

本章以第 3 章和第 4 章的研究成果为基础，选取 13 个信用风险影响指标，结合传统的机器学习模型构建决策融合模型进行风险预测研究。利用决策融合在分类任务中优势，通过结合多个模型的预测结果，减少单个模型的偏差和方差，从而提高模型稳定性和预测准确率，得到更加稳健和可靠的预测结果。最后，进一步对不同指标进行量化分析，从而分析它们对不同经营范围企业信用风险的具体影响。

5.1 决策融合

5.1.1 模型构建

决策融合可以通过结合多个模型的预测结果进行平均或投票来减少单一模型的偏差和方差。决策融合依据分类器融合的类型分为以下两种：Stacking是以不同分类器作为基分类器的叠加式决策融合方案，将基分类器对原始数据的输出按照列的方式进行堆叠，构成了（m，n）维的新特征作为输入，并以逻辑回归作为元分类器进行拟合得到预测结果，其中 m 代表样本数，n 代表基分类器个数。Bagging 是以同类分类器作为基分类器的投票式决策融合方案，首先对训练数据进行重采样（Bootstrap）并构建 n 颗决策树，同时得到 n 个 OOB（Out Of Bag）数据集，根据 OOB 误差确定决策树颗数，最后对 n 颗决策树的结果采用投票策略，得到预测结果。

根据前文的研究，已经建立了一套合理的信用风险评价指标体系。杨子晖等研究表明传统的机器学习模型可以对我国大多数财务危机进行有效预警[45]。因此，为保证所构建的评价指标体系对企业信用风险预测的稳健性，本文选择代表性较强的三种机器学习模型——逻辑回归、支持向量机和决策树模型进行风险评估。

虽然机器学习算法可以在训练数据上取得很好的性能，但他们可能无法很好地泛化到未见过的数据上，且通常对输入数据的分布非常敏感。因此，本文以基于逻辑回归、支持向量机和决策树融合的 Stacking 和基于多颗决策树融合

的 Bagging 两种决策融合策略进行信用风险预测，从而得到具有更高泛化能力和鲁棒性的预测模型。信用风险预测的具体流程图如图 5-1 所示。

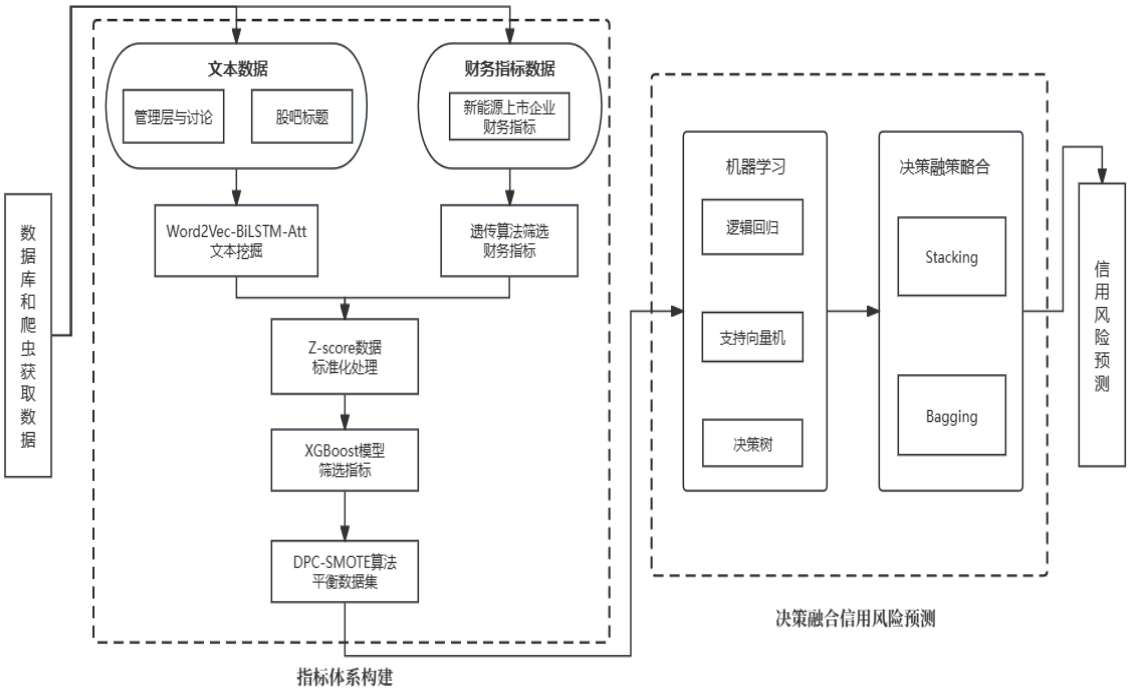

图 5-1 流程图

5.1.2 模型评价指标

为了保证模型预测能力得到客观地比较，将挑选以下指标来反映模型预测性能的优劣。数据集的混淆矩阵如表 5-1 所示：

表 5-1 混淆矩阵

真实值	预测结果	
	无风险	有风险
无风险	TP	FN
有风险	FP	TN

精确率（Precision）是指预测为无风险的样本中，真正为无风险的比例。它反映了模型对于无风险公司的预测精度。计算公式如下：

$$Precision = \frac{TP}{TP + FP} \tag{5.1}$$

召回率（Recall）是指实际为无风险中被预测为无风险的比例。它反映了模型捕获所有无风险公司的能力。计算公式如下：

$$Recall = \frac{TP}{TP + FN} \tag{5.2}$$

准确率（Accuracy）指分类模型正确预测的样本数占总样本数的比例。它是评估分类模型性能最直观的指标之一。计算公式如下：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5.3)$$

AUC 曲线（Area Under the Curve）反映了模型区分公司是否存在风险的能力，AUC 值越接近 1，表示模型区分能力越强。

K 折交叉验证（K fold Cross Validation）反映了模型的泛化能力，它通过多次训练和验证来减少单次训练和验证可能带来的偏差，从而得到更加稳定和可靠的模型性能评估结果。

5.2 信用风险评估与分析

5.2.1 信用风险评估

通过不同模型构建的新能源上市公司信用风险预测结果如表 5-2 所示。由表 5-2 可知，（1）在样本比例为 1：3 时，Bagging 模型的 AUC 值和准确率均显著高于其余模型，其值分别为 0.905 和 0.935，说明在 1：3 样本比例下 Bagging 的预测效果最好；支持向量机模型的准确率最低为 0.774，但逻辑回归、决策树以及 Stacking 模型的 AUC 值均低于支持向量机的 0.851，说明这 3 种模型的预测效果有所下降。（2）在样本比例为 1：1 时，预测准确率相较于样本比例为 1：3 时明显降低，但 AUC 值均大于 0.8，且除 Bagging 模型外，其余模型的 AUC 值均有所提高，说明在样本比例为 1：1 时的预测效果更好，且有更强的泛化能力；Bagging 在样本数量为 1：1 时整体的 AUC 值与预测准确率均为最高。综上所述，样本比例 1：1 相较于样本比例 1：3 预测准确率较低，但除 Bagging 外 AUC 值都有所提升，说明样本数量为 1：1 时模型的泛化能力更强；Bagging 在两种样本比例下均有最高的 AUC 值和预测准确率，说明整体上 Bagging 模型的预测效果最优，能更准确的对新能源上市企业进行信用风险预测。

表 5-2 不同模型的评价指标

样本比例	模型名称	精确率	召回率	AUC 值	准确率	交叉验证
1：3	LR	0.821	0.958	0.774	0.806	0.815
	SVM	0.905	0.792	0.851	0.774	0.808
	DT	0.958	0.852	0.801	0.838	0.827
	Bagging	0.908	0.958	0.905	0.935	0.874
	Stacking	0.846	0.917	0.744	0.839	0.842

续表 5-2 不同模型的评价指标

样本比例	模型名称	精确率	召回率	AUC 值	准确率	交叉验证
1: 1	LR	0759	0.917	0.827	0.763	0.798
	SVM	0.941	0.696	0.862	0.789	0.793
	DT	0.8	0.842	0.816	0.816	0.804
	Bagging	0.815	0.917	0.896	0.816	0.841
	Stacking	0.870	0.833	0.810	0.808	0.808

5.2.2 信用风险分析

为更好厘清不同指标特征对企业信用风险的影响，本文将采用 Logit 方法构建模型对指标特征与信用风险进行回归分析，模型如公式（5.4）所示：

$$CR = \alpha_0 + \sum_{i=1}^{n}(\alpha_i Feature_i) + \varepsilon \quad (5.4)$$

其中，因变量 CR（Credit Risk，CR）为企业信用风险标签，自变量 Feature 包括了上文构建信用风险指标体系中所有特征指标。

本文选取更具泛化能力的样本比例 1：1 进行回归分析。Logit 模型回归估计结果如表 5-3 所示。从表 5-3 的模型 1 和模型 2 中可以看出，财务费用率（Financial Expense Ratio，FER）和长期资本收益率（Long-term return on capital，LROC）与企业是否发生信用风险均显著相关，并且财务费用率和长期资本收益率的影响系数分别为 2.416 和-1.347，在 1%水平上显著。模型 3 的结果表明，根据管理层讨论与分析（MDA）得到的文本指标对企业信用风险具有正向影响，且在 5%水平上显著，影响系数为 0.378 低于财务费用率和长期资本收益率。综上所述，实证结果表明文本指标一定程度上包含了信用风险影响因素，且相较于股吧标题文本信息，管理层讨论与分析包含的信息更能反映企业是否发生信用风险。

表 5-3 Logit 模型回归估计结果

名称	模型 1	模型 2	模型 3
FER	2.416***		2.216***
	(0.552)		(0.543)
LROC		-1.347***	-1.466***
		(0.353)	(0.378)
MDA			0.378**
			(0.179)
Constant	-0.266	-0.567	-0.614***
	(0.192)	(0.226)	(-0.230)
R^2	0.356	0.474	0.492

注：（1）*、**、*** 分别代表在 10%、5%、1%的水平上显著；（2）括号内为标注误差（下表同）。

基于经营范围的异质性分析。目前，我国正积极推动新能源行业的发展，并出台一系列措施来支持新能源行业的发展，以实现绿色低碳转型。新能源行业的发展往往受到环境政策的影响，不同经营范围的新能源公司可能受到不同政策的支持和限制。因此，本文将依据 Wind 数据库，将新能源行业分为新能源、新能源汽车、新能源设备以及新能源整车四组进行异质性分析。如表 5-4 结果所示，无论新能源上市公司经营范围是什么，长期资本收益率对企业信用风险均有显著降低作用；财务费用率在新能源以及新能源设备经营企业中对信用风险有显著正向影响；管理层讨论与分析在新能源设备中具有显著正向影响；而股吧标题信息则在新能源对企业信用风险有显著的正向作用。原因是对于新能源汽车和新能源整车行业而言，需要大量的资金投入用于技术研发、生产设施以及市场推广等方面，因此，企业的长期资本收益率直接决定了其盈利能力和市场竞争力，从而降低信用风险。

表 5-4 异质性分析

变量	基于经营范围的异质性分析			
	新能源	新能源汽车	新能源设备	新能源整车
FER	0.119***	-0.007	0.766***	-0.382
	(0.028)	(0.110)	(0.146)	(0.347)
LROC	-0.270***	-0.126**	-0.207***	-0.349**
	(-0.080)	(0.056)	(0.060)	(0.128)
MDA	0.003	-0.043	0.081**	-0.026
	(0.068)	(0.088)	(0.037)	(0.137)
EMGT	0.169***	0.052	0.007	-0.053
	(0.061)	(0.072)	(0.041)	(0.352)
Constant	0.386***	0.283***	0.514***	-0.215
	(0.052)	(0.069)	(0.060)	(0.256)
R^2	0.464	0.148	0.548	0.421

5.3 本章小结

本章结合上一章构建的企业信用风险评价指标体系，将影响企业信用风险的指标数据作为原始数据输入到由决策融合构建的风险预测模型中。为研究不平衡数据对预测模型的影响，按样本比例的不同分别进行实证分析。同时运用 Logit 回归模型进一步分析不同指标特征对信用风险预测的影响，并研究这些指标在不同经营范围企业中的影响效果。结果发现：

本章构建组合模型对新能源上市公司信用风险预测具有良好的效果，能起

到提高预测准确性和模型泛化能力的效果。新能源上市企业应重点关注财务费用率、长期资本收益率以及管理层讨论与分析的核心指标。说明在新能源这个资本密集的行业，对资本的运用效率有着极高的要求，通过关注长期资本收益率，公司可以确保其投资决策和运营策略能够带来稳定的回报，进而提升公司的整体竞争力防止信用风险的发生；其次，新能源企业优化财务结构，提升资金使用效率，从而维护良好的信用状况；最后，新能源上市公司应密切关注管理层讨论与分析和股吧标题文本信息，以便及时发现并应对潜在的信用风险，同时，公司还应加强内部管理，提高信息披露质量，增强投资者信心，从而有效防止信用风险的发生。

第 6 章 总结与展望

企业信用风险和实现经济绿色转型是金融机构和全社会非常关注的问题。如何在海量的结构化和非结构化信息中选取信用风险关键影响因素，并构建具有预测准确率高和泛化能力强的预警体系是目前亟待解决的问题，因此，构建一个科学、准确的企业信用风险预测体系是非常有价值的。

6.1 研究的结论

本文的主要结论有以下三点：

（1）本研究对不同来源的金融文本进行分析，为避免出现人为主观影响，本文运用词典法对文本进行标注，利用 Word2Vec 模型对企业文本信息进行向量化表示。同时，通过结合双向长短期记忆算法，得以更全面地捕捉文本特征提取过程中的上下文信息，减少重要信息的遗漏，最后引入注意力机制，根据输入数据的重要性动态地调整各权重，使得模型更加关注关键信息，提高文本特征提取模型的准确率与精度，从而精确地提炼出反映企业信用风险的文本指标，并将提取的文本指标作为一个新兴指标加入信用风险评价体系中。这有助于更全面、准确地评估企业的信用风险，为金融机构的风险管理提供有力支持；

（2）本文采用遗传算法对财务数据进行初次筛选。由于标签数据集不平衡会直接导致后续信用风险预测模型出现过拟合现象，本研究采用 DPC-SMOTE 算法对企业信用风险数据集进行扩充，使风险预警体系的泛化能力得到进一步提升。同时为避免人为因素对企业信用风险指标评价指标选择的干扰，本研究采用 XGBoost 模型对包含文本指标的企业信用风险评价指标进行二次筛选，通过对评价指标的特征重要性进行分析，最终选取 13 个指标构建信用风险评价指标体系。这一体系不仅为后续的企业信用风险预测提供了研究基础，还为深入研究企业信用风险评价提供了参考依据；

（3）在构建完成的基于新能源上市企业的信用风险评价指标体系的基础上，利用传统的机器学习模型对企业信用风险进行预测分析。然而，考虑到传统机器学习模型在预测性能和泛化能力上存在的瓶颈问题，因此本文通过组合多个

单一模型来构建信用风险预测模型来提高泛化能力和鲁棒性。构建的 Stacking 和 Bagging 组合模型在泛化能力和预测准确率上均优于单一模型，且 Bagging 模型的 AUC 值为 90.5%，预测准确率达到了 93.5%，适用于企业信用风险预测，展示出了良好的预测性能。为进一步厘清不同指标特征对企业信用风险的影响，本研究采用 Logit 模型进行回归估计，结果表明 MD&A 文本指标在 10%水平上与企业信用风险显著相关。

6.2 不足及展望

由于相关知识水平的有限及技术设备的限制，本文仍然存在一些不足：

（1）影响企业信用风险因素较多，本研究参考国内外研究，遵循主流的方式选取财务指标，构建初步企业指标体系。这样的选取方式可能忽略了一些其他指标因素对于企业风险的影响。未来的研究可以运用合适的深度学习领域算法对全部财务指标进行筛选，使构建的信用风险评价指标体系更具普适性以及科学性；

（2）本文通过将管理层讨论与分析和东方财富股吧文本信息融入信用风险评价体系，从而丰富了该评价体系的指标构成，增强了评价体系的全面性和准确性。但由于年报是在管理层进行盈余管理之后对外公告[46]，同时股吧互动放大了公司传闻对投资者情绪的影响，进一步加剧了传闻对企业的影响[47]。因此，在未来的研究中，可以通过深度学习领域算法对不同来源文本特征指标的权重进行学习，同时可以考虑加入更多客观、真实的文本信息，以此获得更接近企业真实的经营状况，为后续预测企业信用风险提供依据；

（3）相较于单一的机器学习预测模型，本文构建的 Stacking 和 Bagging 组合模型对企业的信用风险具有更好的预测能力。然而，深度学习领域亦不乏众多性能卓越的预测模型，这些模型具有强大的处理复杂和高度非线性数据的能力，可以进一步提高企业信用风险预测的准确率。在未来构建企业信用风险预测模型时，可以考虑结合深度学习和其他机器学习模型。

参考文献

[1] Standard , Poor'S. Ratings and ratios: corporate ratings criteria[R]. Standard & Poor's Report, 2003.

[2] 韩嵩, 李晓俊. 融合互联网文本大数据的上市企业信用评价[J]. 统计学报, 2021, 2 (05): 72-81.

[3] 陈艺云. 基于文本信息的上市中小企业财务困境预测研究[J]. 运筹与管理, 2022, 31 (04): 136-143.

[4] 张皓楠, 张红梅. 基于文本信息的企业信用风险预测研究[J]. 中国管理信息化, 2023, 26 (02): 112-115.

[5] 柳永明, 殷越. 国企债券违约、隐性担保预期与唤醒效应-来自永煤违约事件的证据[J/OL]. 财经研究(2023-10-20)[2024-3-25].

[6] Title C. A Comparison of the Ratios of Successful Industrial Enterprises with those of Failed Companies[J]. Análise Molecular Do Gene Wwox, 1932: 598-605.

[7] Altman E I. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy[J]. Journal of Finance, 1968, 23(4):589-609.

[8] Jens G, Lars N, Martin W．The role of non-financial factors in internal credit ratings[J]. Journal of banking & finance, 2005, 29(2): 509-531.

[9] Kim K, Ahn H. A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach[J]. Computers & Operations Research, 2012, 39(8): 1800-1811.

[10] Magni, Carlo A. Economic profitability and (non) additivity of residual income[J]. Annals of Finance, 2021, 17(4): 471-499.

[11] Das S, Chen M. Yahoo! for Amazon: Extracting market sentiment from stock message boards[J]. Management Science, 2001, 53(9): 1375-1388.

[12] Dave K, Lawrence S, Pennock David M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the 12th international conference on World Wide Web (pp. 519-528). ACM.

[13] Gentzkow, Matthew, Bryan K, Matt T. Text as data[J]. Journal of Economic

Literature, 2019, 57(3): 535-574.

[14] Cecchini M, Aytug H, Koehler G J, et al. Making Words Work: Using Fina ncial Text as a Predictor of Financial Events[J]. Decision Support Systems, 2010, 50(7):164-175.

[15] Gandhi P, Loughran T, McDonald B. Using annual report sentiment as a proxy for financial distress in US banks. SSRN Scholarly Paper ID 2905225[J]. Social Science Research Network, Rochester, NY, 2019, 20(4): 424-436.

[16] Donovan J, Jennings J, Koharki K, et al. Measuring credit risk using qualitative disclosure[J]. Review of Accounting Studies, 2021, 26(2): 815-863.

[17] Tetlock, Paul C. Giving content to investor sentiment: The role of media in the stock market[J]. The Journal of finance, 2007, 62(3): 1139-1168.

[18] Hillert A, Heiko J, Sebastian M. Media makes momentum[J]. The Review of Financial Studies, 2014, 27(12): 3467-3501.

[19] Lu H M, Tsai F T, Chen H, et al. Credit rating change modeling using news and financial ratios[J]. ACM Transactions on Management Information Systems (TMIS), 2012, 3(3): 1-30.

[20] Cerchiello P, Nicola G, Rönnqvist S, et al. Assessing Banks' Distress Using News and Regular Financial Data[J]. Frontiers in Artificial Intelligence, 2022, 5: 871863.

[21] Sait G, Kabak O, Topcu I. A Multiple Criteria Credit Rating Approach Utilizing Social Media Data[J]. Data & Knowledge Engineering, 2018, (116): 80-99.

[22] 庞素琳, 何毅舟, 汪寿阳等. 基于风险环境的企业多层交叉信用评分模型与应用[J]. 管理科学学报, 2017, 20(10): 57-69.

[23] 田琨, 庄新田, 赵婉婷. 供应链金融模式下中小企业信用风险评估——基于汽车制造业数据分析[J]. 工业技术经济, 2021, 40(05): 15-20.

[24] 白雪鹏, 赵志冲. 似然函数视角下小企业信用风险最优评价指标体系的建立[J].运筹与管理, 2023, 32(04): 155-161.

[25] 周志刚, 黄运聪, 窦路遥. 科创板上市公司的信用风险体系构建、评价与组态分析——基于 DBN-SEM-fsQCA 方法[J]. 山东财经大学学报, 2024, 36 (01): 98-110.

[26] 刘逸爽, 陈艺云. 管理层语调与上市公司信用风险预警——基于公司年报文本内容分析的研究[J]. 金融经济学研究, 2018, 33(04): 46-54.

[27] 苗霞, 李秉成. 管理层超额乐观语调与企业财务危机预测——基于年报前瞻性信息的分析[J]. 商业研究, 2019, (02): 129-137.

[28] 梁龙跃, 刘波. 基于文本挖掘的上市公司财务风险预警研究[J]. 计算机工程与应用, 2022, 58(04): 255-266.

[29] 李成刚, 贾鸿业, 赵光辉, 付红. 基于信息披露文本的上市公司信用风险预警——来自中文年报管理层讨论与分析的经验证据[J]. 中国管理科学, 2023, 31(02): 18-29.

[30] 宋彪, 朱建明, 李煦. 基于大数据的企业财务预警研究[J]. 中央财经大学学报, 2015, (06): 55-64.

[31] 崔炎炎, 刘立新. 网络舆情赋能金融科技股票收盘价预测研究[J]. 统计研究, 2022, 39(06): 148-160.

[32] 宫晓莉, 徐小惠, 熊熊. 媒体情绪与企业风险承担——基于机器学习和文本分析的证据[J/OL]. 系统工程理论与实践(2024-3-7)[2024-3-25]

[33] 范小云, 王业东, 王道平, 郭文璇, 胡煊翊. 不同来源金融文本信息含量的异质性分析——基于混合式文本情绪测度方法[J]. 管理世界, 2022, 38 (10): 78-101.

[34] 杨晓兰, 沈翰彬, 祝宇. 本地偏好、投资者情绪与股票收益率:来自网络论坛的经验证据[J]. 金融研究, 2016, (12): 143-158.

[35] 段江娇, 刘红忠, 曾剑平. 中国股票网络论坛的信息含量分析[J]. 金融研究, 2017, (10): 178-192.

[36] 沈艳, 陈赟, 黄卓. 文本大数据分析在经济学和金融学中的应用:一个文献综述[J]. 经济学(季刊), 2019, 18(04): 1153-1186.

[37] 张新红, 王瑞晓. 我国上市公司信用风险预警研究[J]. 宏观经济研究, 2011, (01): 50-54.

[38] 陆健健, 江开忠. 基于 XGBoost 算法模型的金融客户信用评估研究[J]. 软件导刊, 2019, 18(04): 133-136.

[39] 迟国泰, 王珊珊. 基于 XGBoost 的中国上市公司违约风险预测模型[J]. 系统管理学报(2023-12-21)[2024-3-25].

[40] 吕秀梅, 张儒. 网络小额贷款业务个人信用风险评估——基于 DNN-SMOTEENN-ExtraTrees 组合模型[J]. 数学的实践与认识, 2023, 53 (07): 14-21.

[41] 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, (06): 46-55+96.

[42] 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(04): 1323-1344.

[43] 刘志函, 张忠林, 赵磊. 面向不平衡数据分类的 DPC-SMOTE 过采样算法[J]. 哈尔滨理工大学学报, 1-16.

[44] 吴增源, 金灵敏, 韩香丽, 王泽林, 伍蓓. 基于 SMOTE-XGBoost 的外贸企业财务危机预警模型[J]. 计算机工程与应用, 1-12.

[45] 杨子晖, 张平淼, 林师涵. 系统性风险与企业财务危机预警——基于前沿机器学习的新视角[J]. 金融研究, 2022, (08): 152-170.

[46] 严若森, 周燃. 公司年报 MD&A 中的管理层诚信承诺：真实盈余管理之后的道德掩饰行为[J]. 南开管理评论, 2023, 26 (05): 137-148.

[47] 卢锐, 张亚楠, 蔡贵龙. 社交媒体、公司传闻与股价冲击——来自东方财富股吧论坛的经验证据[J]. 会计研究, 2023, (04): 59-73.

硕士期间发表的论文

[1]丁沈杰, 张玥, 杨灿清. 基于异质特征融合的财务风险预测研究——以新能源上市公司为例[J]. 萍乡学院. （已录用）

致 谢

白驹过隙，时光荏苒，研究生学习生涯即将落下帷幕，此刻的我感慨万千。回首这三年，在师长和亲友的支持鼓励下，我历经磨砺，却也收获颇丰。这段学习经历不仅使我收获了知识，更促成了我个人的成熟与成长，对此我满怀感激。

首先，我深感庆幸能够遇到我的导师张玥老师。她不仅是我学术道路上的引路人，更是我人生中的良师益友。张老师治学严谨，为人谦逊和善，她始终强调培养我们的自主学习能力和科学思维。在她的悉心指导下，我逐渐养成了科学完善的思考模式，这不仅是一个积累知识的过程，更是一个内心成长和升华的历程。从论文选题到最终定稿，张老师始终耐心引导，以科学的方法论和严谨的态度要求我们，她的付出与教诲令我受益匪浅。在此，我向张老师表达最诚挚的感谢和崇高的敬意。

其次，我也要感谢我的同学们和师弟师妹们。感谢室友王梦旋、杨磊和陈晓东在我对论文一筹莫展之时伸出援手，并在论文修改时给予我有建设性的意见，在生活中对我也是相当包容和理解，让我深刻体会到了同窗之情的温暖；还要感谢师弟师妹们，你们的加入为科研队伍带来新的研究角度和欢快的科研氛围；感谢金融研 21 的同学们，我们在研修室共同奋斗的日子，将成为我宝贵的回忆。

最后，我要衷心感谢我的家人，正是他们一如既往的关爱和支持，让我能够坚定地走完这段求学之路。他们用无私的爱和默默地付出为我筑起了坚实的后盾，是我不断前行的动力源泉。我会用一生去回报他们的养育之恩，感谢他们为我所做的一切。

在此，我谨向所有关心、支持、鼓励和帮助过我的人表示衷心的感谢！这段宝贵的经历将永远铭刻在我的心中，激励我不断前行，追求卓越。