分类号：029　　单位代码：10363
密　级：公开　　学　　号：2221020117

题　　目　融合大语言模型的上市公司经营状况研究

英文并列

题　　目　Research on the Management Status of Listed Companies with Large Language Model

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

论文答辩日期：**2025 年 5 月 14 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：杜楠

日期： 2025 年 5 月 14 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：杜构　　指导教师签名：张珊

日期： 2025 年 5 月 14 日　　日期：2025 年 5 月 14 日

融合大语言模型的上市公司经营状况研究

摘 要

在信息化与数据化深度融合的时代背景下，文本数据作为信息传递的核心载体，已全面渗透至社会生活的各个领域。作为资本市场的重要主体，上市公司披露的年报、公告及新闻稿等文本数据，不仅直观反映了企业的经营与财务状况，更为洞察其战略布局、市场定位及未来发展方向提供了关键依据。因此，对上市公司文本数据进行深度挖掘与分析，已成为信息时代的重要研究课题。

管理层讨论与分析（Management's Discussion and Analysis，MD&A）作为上市公司信息披露体系的重要组成部分，在资本市场的信息传递中发挥着至关重要的作用。然而，MD&A 内容结构复杂，涵盖多个层面的信息，使得精准挖掘其中蕴含的经营状况面临诸多挑战。针对这一问题，本文提出了一种基于 MD&A 的融合大语言模型，以弥补传统方法在处理复杂语境及捕捉文本深层关联性等方面的不足，从而提升模型对企业经营状况的识别能力与评估精度。鉴于信息技术在战略性新兴产业发展中的举足轻重的地位，我们选取了 CN+WIND 互联网软件与服务Ⅲ、WIND 信息技术服务与 WIND 软件三个板块的上市公司 MD&A 文本数据进行深入的文本挖掘，主要研究内容如下：

（1）构建了上市公司 MD&A 的特征融合表示。在经营状况自定义停用词典的基础上，使用基于全局共现统计计数方式的 GloVe（Global Vectors for Word Representation）将三个板块的 MD&A 分别构造成 564×100 维的文档-词项矩阵，708×100 维的文档-词项矩阵，以及 849×100 维的文档-词项矩阵。与此同时，考虑到 MD&A 中通常包含复杂的上下文关系，单纯依赖 GloVe 获取词与词间的全局关系难以精准捕捉多义词在不同语境中的含义。于是我们使用基于上下文相关的动态表示计数方式的 BERT（Bidirectional Encoder Representations from Transformers）将三个板块的 MD&A 分别构造成 564×200 维的文档-词项矩阵，708×100 维的文档-词项矩阵，以及 849×200 维的文档-词项矩阵。为使模型在处理复杂语义信息时既能保持一致性，又能灵活应对上下文语义关系，我们将 GloVe 表示和 BERT 表示串联,将三个板块的 MD&A 分别对应成 564×300 维的文档-词项矩阵，708×200 维的文档-词项矩阵，以及 849×300 维的文档-词项矩阵。对融合特征表示的各板块文档-词项矩阵进行文本聚类可视化分析，并与单一特征表示（GloVe 或 BERT）的可视化效果进行对比。结果表明，融合特征的可视化表现最佳，不仅能够清晰展现不同经营等级的分布情况，还能确保各等级之间边界清晰，避免过多重叠或混杂。在三维可视化图中，各类别的区分效果尤为显著，进一步验证了融合特征表示在文本结构化分析中的优势。

（2）构建了基于 PGBG-MD&A 的上市公司经营状况评估模型。在完成 MD&A 文本数据的特征融合表示后，我们将每个 MD&A 文本视为图结构数据中的一个节点，其融合表示则作为该节点的特征。同时，通过设置基于 GloVe 和 BERT 的余弦相似度阈值构建度量文档间相似度的邻接矩阵，并利用 Poisson 核函数对分别基于 GloVe 和 BERT 的邻接关系进行关联性融合，作为图结构数据中的边权，从而增强图中边的表达能力。明确 GCN（Graph Convolutional Network, GCN）中各卷积层权重参数矩阵的列维维度，得到了 PGBG-MD&A 经营状况评估模型。通过与一些主流的机器学习和深度学习模型进行比较，PGBG-MD&A 展现了更强的预测性能、鲁棒性以及适应性，能够为投资者和利益相关方提供更全面、可靠的上市公司经营状况评估。

关键词：Poisson 核；关联性融合；融合大语言模型；图卷积网络（GCN）；经营状况评估

RESEARCH ON THE MANAGEMENT STATUS OF LISTED COMPANIES WITH LARGE LANGUAGE MODEL

ABSTRACT

Under the background of the deep integration of informatization and data, text data, as the core carrier of information transmission, has fully penetrated into every field of social life. As an important subject of the capital market, the text data disclosed by listed companies such as annual reports, announcements and press releases not only directly reflect the operation and financial status of the company, but also provide a key basis for insight into its strategic layout, market positioning and future development direction. Therefore, the in-depth mining and analysis of the text data of listed companies has become an important research topic in the information age.

As an important part of the information disclosure system of listed companies, Management's Discussion and Analysis (MD&A) plays a vital role in the information transmission of the capital market. However, the complex structure of the MD&A content, which covers multiple layers of information, makes it challenging to accurately discover the business conditions contained therein. To solve this problem, this paper proposes a fusion large language model based on MD&A to make up for the shortcomings of traditional methods in dealing with complex context and capturing the deep relevance of text, so as to improve the model's ability to identify and evaluate the business situation of enterprises. In view of the pivotal position of information technology in the development of strategic emerging industries, we selected the MD&A text data of listed companies in the three sections of CN+WIND Internet software

and services III, WIND information technology services and WIND software for in-depth text mining. The main research contents are as follows:

(1) Construction of Feature Fusion Representations for MD&A of Listed Companies.Based on the customized stop-word dictionary for business conditions, we utilized GloVe (Global Vectors for Word Representation), which employs a global co-occurrence statistics-based counting method, to construct document-term matrices for the MD&A of three different sectors. These matrices were structured as 564×100, 708×100, and 849×100 dimensions, respectively.Meanwhile, considering that MD&A texts typically contain complex contextual relationships, relying solely on GloVe to capture global word relationships may not precisely represent the meaning of polysemous words in different contexts. Therefore, we employed BERT (Bidirectional Encoder Representations from Transformers), which utilizes a context-dependent dynamic representation counting method, to construct document-term matrices of dimensions 564×200, 708×100, and 849×200 for the three sectors' MD&A texts.To ensure that the model maintains consistency while flexibly adapting to contextual semantic relationships when processing complex semantic information, we concatenated the GloVe and BERT representations. As a result, the MD&A of the three sectors were respectively represented as document-term matrices of dimensions 564×300, 708×200, and 849×300.A text clustering visualization analysis was conducted on the document-term matrices of each sector using the fused feature representation. The visualization results were then compared with those derived from single-feature representations (either GloVe or BERT). The findings demonstrated that the fused feature representation yielded the best visualization performance. It not only

clearly illustrated the distribution of different business condition levels but also ensured well-defined boundaries between categories, thereby minimizing excessive overlap or mixing. In the three-dimensional visualization plots, the distinction among different categories was particularly pronounced, further validating the advantages of the fused feature representation in textual structural analysis.

(2) The operating status assessment model of listed companies based on PGBG-MD&A is constructed. After the feature fusion representation of MD&A text data is completed, each MD&A text is regarded as a node in the graph structure data, and its fusion representation is regarded as the feature of this node. At the same time, an adjacency matrix to measure the similarity between documents is constructed by setting the cosine similarity threshold based on GloVe and BERT, and Poisson kernel function is used to combine the adjacency relationships based on GloVe and BERT respectively as edge weights in the graph structure data, thus enhancing the expression ability of edges in the graph. The levy-dimension of the weight parameter matrix of each Convolutional layer in the Graph Convolutional Network (GCN) is clarified, and the PGBG-MD&A business situation evaluation model is obtained. By comparing with some mainstream machine learning and deep learning models, PGBG-MD&A shows stronger predictive performance, robustness and adaptability, and can provide investors and stakeholders with a more comprehensive and reliable assessment of the operating status of listed companies.

Du Yang

(Applied Mathematics)

Surpervised by Zhang Yue

Key words: Poisson kernel; Associative fusion; Integration of large

language models; Graph convolutional Network (GCN); Business condition assessment

目 录

第 1 章 绪论

1.1 研究背景与意义

1.1.1 研究背景

2024 年党的二十届三中全会明确指出，在复杂多变的国际经济形势与国内经济转型的大背景下，必须同步推进高质量发展与高水平安全的双重目标。这一战略不仅强调提升经济的内生动力，还要求有效防范潜在风险，确保经济在波动中保持韧性与活力。上市公司作为资源配置、产业竞争力和经济韧性的重要支撑，其经营状况的监测与评估构成了政策执行中的关键环节。上市公司的稳健运营不仅对社会福利、民生改善以及社会秩序的维护具有积极推动作用，还在国家经济转型和高质量发展中扮演着重要角色[1-2]。

随着数据科学和人工智能的不断发展，文本分析作为一种新兴的分析方法，为企业的经营状况分析提供了新的视角[3-4]。上市公司定期报告是评价上市公司经营状况优劣的重要信息来源之一，提供了关于上市公司财务状况、业务运营和市场发展等方面的详细数据和分析。而其中的管理层讨论与分析部分是一份定期报告的核心组成部分，它包含了企业过去一段时间的经营状况的描述，以及企业对未来战略、经营计划和风险等方面的阐述[5-6]。MD&A 文本不仅反映了企业的历史表现，还提供了对未来发展的前瞻性信息，因此，深入挖掘 MD&A 文本中的信息对于评估上市公司的经营状况具有重要意义。

随着互联网的快速发展和技术的进步，大量的文本数据得以记录和积累。同时，计算机的计算能力和深度学习算法的发展也为处理和分析这些海量文本数据提供了强大的支持。在此背景下，大语言模型应运而生，人们可以利用其强大的语言生成和理解能力更好地处理和分析文本数据[7-9]。在大语言模型的发展过程中，深度学习发挥着关键作用。它的强大表达能力、数据驱动的学习和预训练与微调方法，使得大语言模型能够更好地理解和生成自然语言，推动了自然语言处理技术的快速发展。其中，词嵌入模型 GloVe[10-11]、预训练语言模型 BERT[12-13]以及图卷积神经网络（GCN）[14-15]作为深度学习领域的重要模型，

在文本数据处理中展现出强大的能力。这些模型不仅能够全面捕捉文本数据中的静态全局信息，还能有效提取上下文关联性和语义特征。基于此，本文提出利用 Poisson 核函数[16-17]将 GloVe 词向量表示模型、BERT 大语言模型以及 GCN 模型进行深度融合，旨在充分聚合 MD&A 文本中词汇的语义相似性与文本间的关联性，从而显著提升语义传递效率和信息整合的准确性。这一融合方法为文本数据的深度分析与语义建模提供了更为全面和精准的技术支持。

1.1.2 研究意义

（1）理论意义

基于 Poisson 核的 GloVe、BERT 与 GCN 融合模型在上市公司经营状况评估中具有重要的理论意义。首先，该模型通过融合 GloVe 和 BERT 的文本表示，能够同时捕捉文本的全局静态语义信息和动态上下文语义信息，克服了单一模型在处理复杂文本数据时的局限性。GloVe 模型通过全局词共现矩阵捕捉词汇的静态语义信息，而 BERT 模型则通过上下文感知的动态语义表示捕捉词汇的多义性和复杂语境下的语义变化。通过 Poisson 核函数，模型能够有效融合这两种表示，进一步提升语义传递效率和信息整合的准确性。

其次，该模型通过引入图卷积神经网络，能够捕捉文本数据中的上下文关联性和语义信息，从而显著提高文本挖掘的准确性和效率。GCN 通过图结构的学习，能够自动捕捉文本中的全局和局部信息，提升模型的泛化能力和鲁棒性。此外，该模型能够深入挖掘上市公司之间的关联性和影响，为金融领域的决策制定者提供有价值的参考信息，促进市场透明度的提升。上市公司之间的关联性不仅体现在财务数据的相关性上，还体现在业务合作、市场竞争、供应链关系等多个方面。通过 GCN 模型，可以构建上市公司之间的关联网络，揭示公司间的潜在关系，为投资者、分析师和监管机构提供更全面的市场洞察。

最后，该研究不仅有助于推动自然语言处理技术的发展，还能为其他相关领域的文本挖掘任务提供借鉴。随着大语言模型的广泛应用，文本挖掘技术在金融、医疗、法律等多个领域的应用前景广阔。基于 Poisson 核的融合模型为这些领域的智能化提供了重要支持，有助于提升决策效率和准确性，推动行业的创新与发展。

（2）现实意义

基于 Poisson 核的 GloVe、BERT 与 GCN 融合模型在上市公司经营状况评估中具有重要的现实意义。首先，该模型能够显著提升信息提取效率，自动化地处理和分析上市公司发布的繁多文本信息，快速提取关键信息，克服传统人工分析方法的局限性。传统的文本分析方法往往依赖于人工阅读和主观判断，效率低下且容易受到人为因素的影响。而基于 Poisson 核的融合模型能够快速处理大量文本数据，提取出有价值的信息，为决策者提供及时、准确的支持。

其次，该模型通过构建上市公司间的邻接矩阵，能够深入捕捉公司间的复杂关系，增强市场分析能力，为投资者、分析师及公司管理层提供重要决策支持。上市公司之间的关联性不仅体现在财务数据的相关性上，还体现在业务合作、市场竞争、供应链关系等多个方面。通过 GCN 模型，可以构建上市公司之间的关联网络，揭示公司间的潜在关系，为投资者、分析师和监管机构提供更全面的市场洞察。

此外，该模型还能辅助风险预警与预测，通过挖掘文本信息及时发现潜在风险，为投资者提供风险参考。MD&A 文本中包含了企业对未来风险的评估和应对策略，通过 GCN 模型的分析，可以提前识别出企业的潜在风险，帮助投资者做出更明智的投资决策。最重要的是，该研究推动了金融科技创新的进程，通过优化模型结构和算法，开发出更智能、高效的金融产品和服务，满足市场多样化需求。

综上所述，基于 Poisson 核融合 GloVe、BERT 与 GCN 模型不仅提升了信息处理的效率和准确性，还深化了市场分析和风险预警能力，为金融科技创新提供了强大动力，具有长期的发展潜力和深远的社会影响。通过该研究，可以为金融市场的参与者提供更全面、准确的信息支持，促进市场的健康发展，推动经济的高质量增长。

1.2 国内外研究现状

国外学术界在基于文本挖掘的上市公司经营状况研究领域已形成了较为完善的研究体系。早期的经营状况评估主要依赖专家经验和主观判断进行定性分析，但由于该方法受主观因素影响较大，难以保证评估结果的稳定性和广泛适

用性，尤其是在大规模数据处理和复杂场景应用中存在明显局限性。随着信息技术的快速发展和企业经营环境的日益复杂，传统评估方法在应对不确定性和风险评估方面的不足愈发凸显。在此背景下，学者们逐渐将研究重心转向企业经营风险的核心驱动因素，致力于探索更为精准、稳定且高效的评估方法。

学术界对文本挖掘的定义呈现多元化特征。Sullivan[18]提出，文本挖掘的核心在于实现文本的高效检索，并借助语言学理论与自动化管理技术对文本进行处理，这突显了其在信息检索与计算语言学领域的重要价值。Thuraisingham 和 Maning[19]指出，文本挖掘依托计算语言学的理论与方法，从海量文本中提取有价值的信息，这充分展现了计算语言学在该领域的基础性作用。Bhattacharya 和 Getoor[20]则认为，文本挖掘是一种以非结构化文本为研究对象的数据分析方法，旨在从大规模文本数据中挖掘潜在的模式与关联，从而为决策提供科学支持。文本挖掘技术整合了信息检索、计算语言学、信息抽取及数据挖掘等多领域的理论与方法，其核心在于从非结构化或半结构化的文本中提炼新知识，为各学科提供创新的分析工具。与传统的信息抽取相比，文本挖掘更注重揭示未知的规律与关联，具有跨领域的普适性，能够广泛应用于多种研究场景。

当前，文本挖掘领域的研究重心正逐步向文本表示方法、模型构建、特征提取以及多维特征融合等方向倾斜。这些研究不仅致力于解决文本数据高维度带来的复杂性问题，还力求通过更全面的表征方式捕捉文本的深层次信息。与此同时，文本分类、关联规则挖掘以及文本聚类等技术也日益成为学界关注的焦点，为文本数据的深度解析与实际应用开辟了新的路径。例如，Jamar Nina39[21]通过对比分析图书馆与信息科学领域中有影响因子和无影响因子期刊摘要的可读性，揭示了文本可读性在学术交流中的重要性。Zhang Kunli[22]等设计了一种基于图结构的知识感知网络（GSKN）模型，通过融合电子病历（EMRs）与医学知识图谱，显著提高了诊断结果的准确性，为知识图谱与文本数据的高效融合提供了创新性解决方案。Zheng[23]等通过提示工程技术优化 ChatGPT 的文本挖掘能力，成功从多种格式与风格的科学文献中提取金属-有机框架（MOF）的合成条件，有效缓解了大语言模型生成虚假信息的局限性。Lowri Williams[24]等探讨了层次化分类方法在情感文本分类中的应用，显著提升

了分类性能，尤其是在测试与验证数据中引入深度方法时表现尤为突出。Gurcan 和 Cagiltay[25]对过去十年内发表的 27,735 篇远程教育领域期刊文献进行了系统性分析，采用基于 N-gram 的文本分类技术对语义内容进行深入挖掘，识别出 10 个核心研究主题，为远程教育领域的未来研究方向与实践应用提供了重要依据。M Torabi[26]等提出基于 BERT 预训练模型的深度学习模型，解决了抗癌化合物活性预测中数据稀缺与实验成本高的问题，该模型在大数据集上高效且可扩展，在多种癌细胞系上预测性能优于传统方法。Chunlei Zhou[27]等提出了一种基于关联规则的高维时空大数据分层挖掘算法，通过构建时空事务数据库并进行频繁谓词搜索，显著提高了数据挖掘效率。

与发达国家相比，我国在上市公司经营状况评估领域的研究起步较晚，研究基础相对薄弱。然而，近年来，国内学者围绕企业经营状况评估技术展开深入研究，推动了相关方法体系的持续完善。当前，研究重点不仅聚焦于优化现有的评估与预测模型，还致力于提升模型的泛化能力与预测精度，以更好地适应信息化时代的需求。在这一过程中，研究范式逐步从单一模型优化转向多维度、多方法的融合创新，力求在复杂多变的商业环境中实现更精准的风险评估与决策支持。

朱芷璇[28]系统梳理了中文分词技术的研究进展，重点探讨了优化算法在神经网络中的应用，并提出了一种融合互信息、朴素贝叶斯分类器与改进优化算法的混合分词方法。此外，她对情感分析模型的输入向量进行了优化设计，有效提升了中文分词的准确性及情感分析的精度。陈加元和刘彦[29]利用 LDA 主题模型对成果导向教育（OBE）相关文献进行了分析，揭示了该领域的研究热点与发展趋势。研究表明，当前学术关注点主要集中在 OBE 理念的应用、效果评估及课程改革三个方面，高频关键词包括“out-based education”“competence”“curriculum”，反映出课程设计、能力培养与效果评估是该领域的核心议题，为文本挖掘技术在教育研究中的应用提供了借鉴。周中雨[30]针对网络舆情话题分析的关键技术展开研究，提出了一种结合二进制蜉蝣优化的特征选择与文本聚类算法，并改进了基于双向量模型的 K 近邻话题跟踪算法，显著提升了话题检测与跟踪的精度及效率。实际应用验证表明，该方法能够有效增强舆情监测

能力。汪克尧[31]提出了一种融合多特征的超图卷积神经网络模型，通过构建语序超图、句法超图和语义超图，弥补了传统图神经网络在文本分类任务中无法充分表示多元高阶关系及上下文语义信息的不足。此外，他针对短文本分类中的数据稀疏问题，提出了一种基于知识增强的异构图注意力网络模型，结合BERT 词嵌入与双向长短期记忆网络（BiLSTM），显著提升了短文本分类的性能。文飞[32]通过实证对比研究探讨了大语言模型在中文文本分类中的应用，采用传统统计学习方法、经典深度学习算法与大语言模型（如 BERT、LLM）进行实验对比，结果表明，大模型在语言上下文理解和泛化能力方面具有显著优势，错误率普遍降低 10%以上。同时，该研究指出，在特定应用场景下，传统方法仍然具有独特价值，并强调工程优化和模型微调技术在提升模型性能方面的有效性。陈美和赵子莜[33]探讨了开放政府数据政策与数字经济的协同效应，运用 LDA 模型对政策文本进行分析，并结合共现网络与关联规则提取高频关键词。研究发现，市级政策在政策执行主体、政策目标及政策工具三个方面存在一定协同性，但在具体措施上存在显著差异，而省级政策则表现出较强的纵向与横向协同特征，为政策优化设计提供了理论支持。唐媛[34]等提出了一种基于多尺度混合注意力卷积神经网络的关系抽取模型，该模型通过多尺度特征提取与融合获取细粒度语义信息，并结合通道注意力与空间注意力机制对特征图进行自适应细化，使模型能够聚焦关键上下文信息。此外，该研究采用双预测器结构共同预测实体关系，在 SemEval-2010 task 8、TACRED、Re-TACRED 和 SciERC 数据集上的 F1 值分别达到 90.32%、70.74%、85.71%和 89.66%，显著提升了关系抽取的准确性。尽管财务信息研究已较为成熟，但对年报文本信息的探索仍处于较早阶段。李成刚[35]等采用文本挖掘技术对上市公司年报中的 MD&A 文本信息进行了深入分析，并在此基础上构建了信用风险预警模型。顾水彬和罗佳敏[36]基于信息经济学、制度经济学和认知经济学理论，探讨了资本市场信息披露改革，提出 MD&A 应从“报表解说”向“价值发现”转型，构建以价值创造为核心的信息披露体系，以引导长期资本流向科技创新，从而助力新质生产力发展，为我国新型信息披露机制的构建提供了重要的理论和实践参考。

综上所述，国内外学者们在文本挖掘、深度学习、表示学习及风险预警等多个领域的研究，为上市公司经营状况分析奠定了坚实的理论基础与技术支撑。这些研究不仅拓宽了文本挖掘技术的应用范围，也为上市公司经营状况的评估与风险预测提供了新的思路，对提升企业管理与投资决策的科学性具有重要意义。

1.3 主要研究内容及论文结构

第 1 章，绪论。主要围绕研究背景、研究意义及国内外相关研究进展展开论述。

第 2 章，准备工作。主要介绍实验的具体实施环境，包括硬件与软件配置，并详细说明数据的获取过程、预处理方法及实验数据的划分策略。通过对数据处理流程的描述，为后续实验分析奠定基础。

第 3 章，基于 GloVe 和 BERT 模型，构建针对上市公司经营状况的 MD&A 文本融合表示，以提升文本特征的表达能力和语义理解效果。

第 4 章，主要研究融合大语言模型 PGBG 对上市公司 MD&A 文本进行经营状况评估和图嵌入表示。

第 5 章，对全文内容进行梳理与总结，归纳研究成果的核心要点，并分析其中的局限性。最后，针对当前研究的不足，提出未来的改进方向和研究展望，以期为后续相关研究提供参考与借鉴。

第 2 章 数据处理与预处理

2.1 实验环境

本研究的实验是在 Python3.8 上实现的，实验环境如表 2-1 所示：

表 2-1 实验环境

设备名称	LAPTOP-KT8OE4QK
处理器	AMD Ryzen 5 5600H with Radeon Graphics 3.30GHz
机带 RAM	16.0GB（15.9GB 可用）
设备 ID	8DD14EE5-3268-4AF6-8411-C2B7555756D8
产品 ID	00342-30505-48997-AAOEM
系统类型	64 位操作系统, 基于 x64 的处理器

2.2 数据来源与处理

当前，国家经济正处于转型升级的关键阶段，数字经济逐渐成为推动经济高质量发展的重要引擎。在这一背景下，信息技术的应用不仅能够优化企业管理流程、提升工作效率，还能有效应对市场需求的变化，推动行业创新与发展。为此，本章从 WIND 数据库中选取了 WIND 软件与服务子库的数据，该子库涵盖了 WIND 互联网软件与服务Ⅲ、 和 WIND 软件等板块的上市公司定期报告。然而，WIND 互联网软件与服务Ⅲ板块的公司数量相对较少，无法满足实验数据的全面性要求。为确保数据的完整性与代表性，本章进一步从 WIND 数据库中的国证指数行业类中选取了 CN 信息技术板块下的 CN 互联网软件与服务Ⅲ子版块，并将其与 WIND 互联网软件与服务Ⅲ板块的公司合并，形成了一个新的综合板块——CN+WIND 互联网软件与服务Ⅲ。

实验的数据集源自 Wind 数据库中 380 家上市公司的 2021-2023 的年报以及半年报，共 2121 份。其中，CN+WIND 互联网软件与服务Ⅲ板块包含 104 家公司，报告数为 564 份；WIND 信息技术服务板块包含 126 家公司，报告数为 708 份；WIND 软件板块则包含 150 家公司，报告数为 849 份。

我们针对这三个板块，选取净资产收益率（ROE）作为评估上市公司经营

状况的评价指标，计算公式如（2.1）所示：

$$ROE = 净利润 / 净资产 \times 100\% \quad (2.1)$$

根据 ROE 的值对上市公司经营状况进行等级划分：当 ROE > 3 / 4 分位数时，设定为 A 等级；当 2 / 4 分位数 < ROE < 3 / 4 分位数时,设定为 B 等级；当 1 / 4 分位数 < ROE < 2 / 4 分位数时,设定为 C 等级；当 ROE < 1 / 4 分位数时，设定为 D 等级。经过等级划分后的三个板块企业管理层讨论与分析的文本数据情况如图 2-1 所示：

代码	简称	日期	管讨论与分析	ROE	经营等级
000004	国华网安	2021-06-3(	一、报告期内公司从事的主要业务报告期内，公司主要从	0.004191	C
000004	国华网安	2021-12-3	一、报告期内公司所处的行业情况公司需遵守《深圳证券	-0.536333	D
000004	ST 国华	2022-06-3(	一、报告期内公司从事的主要业务报告期内，公司主要从	-0.039447	D
000004	ST 国华	2022-12-3	一、报告期内公司所处行业情况公司需遵守《深圳证券交	-1.690548	D
...	...	...	...	...	...
839680	广道数字	2023-12-3	一、 业务概要商业模式报告期内变化情况：（一）	0.066285	A
900901	云赛B股	2021-06-3	会计数据和经营情况一、主要会计数据和财务指标（一）	0.025826	B
900901	云赛B股	2021-12-3	会计数据、经营情况和管理层分析一、主要会计数据和财	0.062408	A
900901	云赛B股	2022-06-3	会计数据和经营情况一、主要会计数据和财务指标（一）	0.017835	C
900901	云赛B股	2022-12-3	会计数据、经营情况和管理层分析一、 主要会计数据和	0.039781	B
900901	云赛B股	2023-12-3	第二节 会计数据、经营情况和管理层分析一、 业务概要	0.048333	B

（a）

代码	简称	日期	管讨论与分析	ROE	经营等级
000032	深桑达A	2021-06-30	一、报告期内公司从事的主要业务报告期内，公司实现营业收	0.045556	B
000032	深桑达A	2021-12-31	一、报告期内公司所处的行业情况（一）全球形势全球从工业	0.116801	A
000032	深桑达A	2022-06-30	一、报告期内公司从事的主要业务公司是中国电子旗下的重要	0.003343	C
000032	深桑达A	2022-12-31	一、报告期内公司所处行业情况（一）宏观政策及发展形势20	0.070468	A
...	...	...	...	...	...
900901	云赛B股	2021-06-30	第三节管理层讨论与分析一、报告期内公司所属行业及主营业	0.025826	B
900901	云赛B股	2021-12-31	第三节管理层讨论与分析一、经营情况讨论与分析数字化转型	0.062408	A
900901	云赛B股	2022-06-30	第三节管理层讨论与分析一、 报告期内公司所属行业及主营	0.017835	C
900901	云赛B股	2022-12-31	第三节 管理层讨论与分析一、经营情况讨论与分析2022年，	0.039781	B
900901	云赛B股	2023-06-30	第三节 管理层讨论与分析一、 报告期内公司所属行业及主营	0.020387	B
900901	云赛B股	2023-12-31	第三节 管理层讨论与分析一、经营情况讨论与分析在世界经	0.048333	B

（b）

代码	简称	日期	管讨论与分析	ROE	经营等级
000004	国华网安	2021-06-30	一、报告期内公司从事的主要业务报告期内，公司主要从事	0.004191	C
000004	国华网安	2021-12-31	一、报告期内公司所处的行业情况公司需遵守《深圳证券交	-0.536333	D
000004	ST 国华	2022-06-30	一、报告期内公司从事的主要业务报告期内，公司主要从事	-0.039447	D
000004	ST 国华	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	-1.690548	D
...	...	...	...	...	...
900926	宝信B	2021-06-30	第三节管理层讨论与分析一、报告期内公司所属行业及主营	0.1251	A
900926	宝信B	2021-12-31	第三节管理层讨论与分析一、经营情况讨论与分析2021年度	0.198756	A
900926	宝信B	2022-06-30	第三节管理层讨论与分析一、 报告期内公司所属行业及主	0.093664	A
900926	宝信B	2022-12-31	第三节 管理层讨论与分析一、经营情况讨论与分析2022年	0.211191	A
900926	宝信B	2023-06-30	第三节 管理层讨论与分析一、 报告期内公司所属行业及主	0.113787	A
900926	宝信B	2023-12-31	第三节 管理层讨论与分析一、 报告期内公司所属行业及主	0.215996	A

（c）

图 2-1 MD&A 文本。（a）CN+WIND 互联网软件与服务III （b）WIND 信息技术服务 （c）WIND 软件

2.3 实验数据划分

我们从 CN+WIND 互联网软件与服务Ⅲ、WIND 信息技术服务以及 WIND 软件三个板块中的各经营等级随机抽取 75%的样本作为训练集，用于训练 GloVe 和 BERT 分类性能最优的 MD&A 文本表示维度，剩余的 25%作为测试集用于评估 PGBG-MD&A 模型的泛化性能。在模型训练过程中为了学习模型的参数，在训练集中再次随机选取各经营等级的 75%作为学习样本用于参数调优，而剩下的 25%作为回测样本用于模型分类性能评估。

第 3 章 基于上市公司经营状况的 MD&A 文本融合表示

在数字化时代，文本数据已成为企业经营状况评估的关键资源。其中，MD&A 文本承载着运营细节、市场洞察及未来战略，能全面反映企业经营状况。然而，其内容复杂、涉及多维度信息，使得有效提取经营信号充满挑战。为精准表征 MD&A 文本，我们融合静态词嵌入与动态语言模型。GloVe 通过词共现关系捕捉全局信息，构建稳固的词向量；BERT 则依赖双向注意力机制理解动态上下文语义，增强文本的深度解析。二者串联融合，使文本表示兼具全局稳定性与上下文适应性，确保信息的完整性与精确度。这一方法旨在突破单一嵌入模型的局限，使 MD&A 文本的深层语义得以充分挖掘，为上市公司经营状况评估提供更加精准的分析支撑。

3.1 GloVe 文本表示

3.1.1 理论基础

GloVe 是一种基于全局词语共现统计的稠密向量表示方法。与词袋模型不同，GloVe 不仅考虑了词语的局部上下文信息，还利用了全局的词语共现统计信息，从而生成能够捕捉语义关系的词向量。在 GloVe 模型中，文本被看作是一个词语共现的统计分布，每个词语的向量表示是通过优化全局共现矩阵得到的。

GloVe 模型的处理流程大致包括：

（1）分词：将文本按照一定的规则或算法进行分词，将其划分为词语的序列；对语料库中的所有文本进行分词处理，构建词语序列集合。

（2）构建共现矩阵：定义一个滑动窗口（如窗口大小为 5），统计语料库中每对词语在窗口内共同出现的次数；构建共现矩阵 X ,其中 X_{ij} 表示词语 i 和词语 j 在窗口内共现的次数。

（3）定义损失函数：GloVe 通过优化以下损失函数来学习词向量：

$$J=\sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log\left(X_{ij}\right)^2\right) \tag{3.1}$$

其中：X_{ij} 表示词 w_i 和 w_j 的共现次数；w_i，$\tilde{w}_j$ 是词向量；b_i，$\tilde{b}_j$ 是偏置项；$f\left(X_{ij}\right)$ 是一个权重函数，用于减少高频词对损失函数的过度影响。

（4）训练词向量：使用随机梯度下降（SGD）或其他优化算法，最小化损失函数，得到每个词语的稠密向量表示。

（5）向量化：将文本中的每个词语替换为其对应的 GloVe 词向量；对于整个文本的表示，可以通过词向量的加权平均、求和或其他聚合方式得到。

GloVe 的计数方式基于全局共现统计，生成稠密词向量。具体规则如下：

（1）共现矩阵：统计词语 i 和词语 j 在在滑动窗口内共同出现的次 X_{ij}。例如，词语 "apple" 和 "fruit" 在语料库中共同出现了 50 次，则 $X_{apple,fruit}=50$。

（2）加权共现统计：使用 $f\left(X_{ij}\right)$ 对共现矩阵中的值进行加权，减少高频词对损失函数的过度影响。权重函数的常见形式为：

$$f\left(X_{ij}\right)=\begin{cases}\left(\dfrac{X_{ij}}{x_{\max}}\right)^{\alpha}, 若X_{ij}<x_{\max}\\ 1 \qquad\quad ,其他情况\end{cases} \tag{3.2}$$

其中，$x_{\max}$ 和 α 是超参数。

（3）通过优化损失函数，学习词向量表示。

3.1.2 经营状况自定义停用词典构建

（1）理论基础

词云图，又称文字云，是文本数据可视化的一种创新形式，它将信息可视化技术与文本挖掘方法深度融合，使文本分析更具直观性与可读性。通过对输入文本进行词频统计，词云图以大小、颜色、布局等多维视觉元素展现词汇的重要性。高频词以更大的字体、深色调突出，而低频词则相对缩小、颜色淡化，从而使核心内容一目了然，信息重点直观可感。

然而，词云图的设计绝非随意堆砌，而是遵循美观性、可读性、紧凑性等多项原则。首先，字体清晰、结构有序，确保读者能迅速捕捉关键信息。其次，

词汇的排列既需紧凑避免凌乱，又需讲究空间均衡，营造层次感。此外，色彩搭配应富有逻辑性，增强视觉吸引力，而词语的关联性亦不可忽视，以更好地揭示文本主题与潜在结构。交互式词云更是近年来的趋势，使用户能够动态探索文本数据，提升分析体验。

词云图的应用领域广阔，涵盖舆情分析、市场调研、用户画像、教育教学等多个方向。在教育领域，词云图能够帮助学生快速提炼文本主旨，直观理解核心概念；在文化研究中，它提供关键词索引，辅助读者高效获取信息；在商业智能方面，企业借助词云分析客户反馈，精准洞察市场需求。同时，得益于计算机技术的发展，Wordle、Tagxedo 等工具的涌现，让词云图的生成变得更加便捷高效。作为数据可视化的一种重要形式，词云图不仅是文本的艺术表达，更是信息探索的强大工具。

（2）停用词典的构建

我们对获取的 CN+WIND 互联网软件与服务Ⅲ、WIND 信息技术服务、WIND 软件三个板块的上市公司 MD&A 文本进行了 jieba 分词，并统计了三个板块中出现频次最高的前 100 个词，生成了词云图（图 3-1）。从图 3-1 中可以看出，这些高频词主要集中在与经营领域相关的内容上，但对于经营状况分析而言，并未提供有效的鉴别信息。为此，我们构建了停用词典，并成功去除了这些高频冗余词汇。

（a）　　（b）　　（c）

图 3-1 MD&A 文本词云图。（a）CN+WIND 互联网软件与服务III （b）WIND 信息技术服务 （c）WIND 软件

3.1.3 GloVe 特征维度确定

在去除停用词后，我们将每个板块的文本转化为了高维的特征词序列，形成了数十万维的词空间。基于这些高维特征，我们利用训练样本，采用线性判别分析（LDA）作为分类器，因其最大化类间区分度的特性能够有效识别词向

量的最优表示维度。在此过程中，我们学习了 GloVe 在 1 到 500 维特征维度范围内的分类准确率，并绘制了分类准确率随 GloVe 特征维度递增的变化曲线图（图 3-2）。从图 3-2 可以看出，当 GloVe 嵌入为 100 维时，三个板块的分类准确率均达到了峰值。为了避免冗余信息过多，我们在相同分类准确率的情况下，选择了较低的维度。因此，每个板块 MD&A 文本的 GloVe 嵌入表示均选取为 100 维。

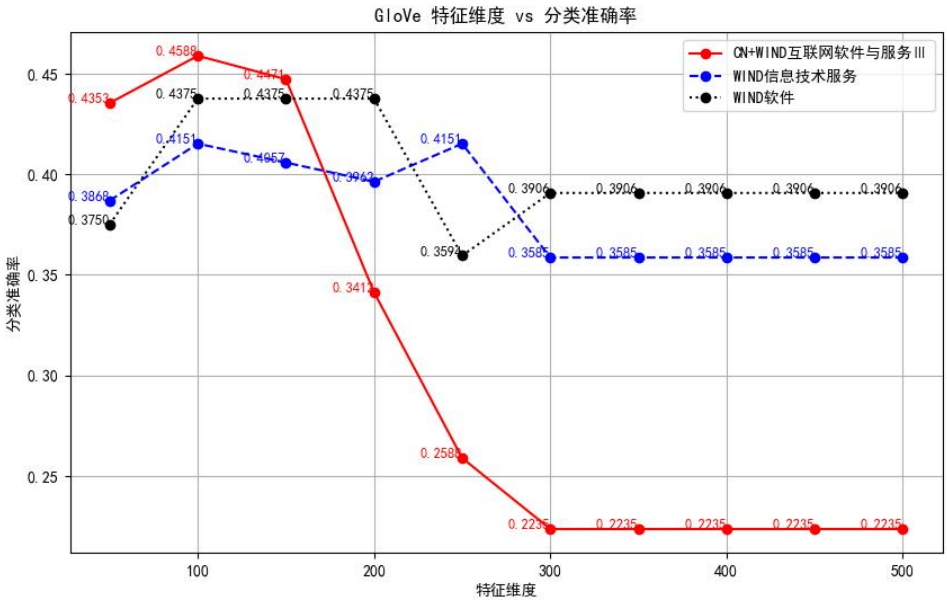
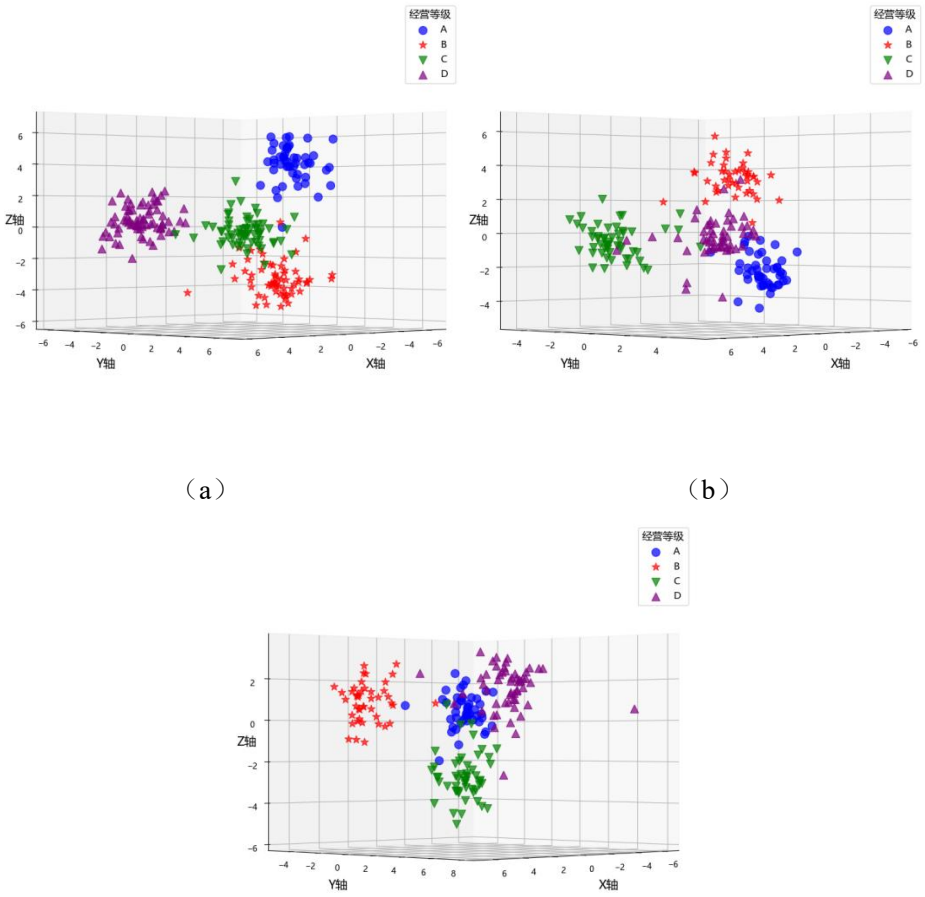


图 3-2 GloVe 特征维度 vs 分类准确率

3.1.4 GloVe 词嵌入的聚类可视化

为验证 CN+WIND 互联网软件与服务III、WIND 信息技术服务及 WIND 软件三个板块的 MD&A 文本在 100 维 GloVe 嵌入表示下经营等级评估的有效性，我们通过领域聚类图进行了可视化分析。结果表明，嵌入维度的设定具有较强的合理性，同时也进一步验证了我们构建的自定义经营状况停用词典的有效性。通过精准剔除冗余信息，我们成功强化了文本的核心语义，使经营状况的分类更加稳定，并显著提升了模型的泛化能力。更重要的是，在以分类准确率为优化目标的过程中，我们确定了最低可接受的嵌入维度。该设定不仅有效削减了无效信息的干扰，还确保了全局语义特征的充分利用。对于大规模语料而言，这一策略不仅能够高效捕捉词汇间的隐含关系，还能生成密集且富含语义的信息向量，从而广泛适用于语义分析、文本分类等任务。在这一优化框架下，我

们观察到，不同经营等级的 MD&A 文本在语义空间中的聚类效果尤为明显。这一发现不仅进一步验证了 GloVe 嵌入在经营状况分析中的适用性，也为后续基于语义特征的深入研究奠定了坚实基础。如图 3-3 所示：



（a）　　（b）

（c）

图 3-3 GloVe 文本表示的经营等级聚类图。（a）CN+WIND 互联网软件与服务III （b）WIND 信息技术服务 （c）WIND 软件

3.2 BERT 文本表示

3.2.1 理论基础

BERT 是一种基于 Transformer 的预训练语言模型，是一种上下文相关的文本表示方法。与词袋模型和 GloVe 词嵌入模型不同的是，BERT 能够根据词语在句子中的具体位置和上下文生成动态的词向量，从而捕捉词语在不同语境中

的语义变化。在 BERT 模型中，文本被看作是一个词语序列，每个词语的向量表示是通过多层 Transformer 编码器生成的。

BERT 的处理流程大致包括：

（1）使用 BERT 的分词器（如 WordPiece 或 Byte-Pair Encoding）将文本划分为子词单元；在输入序列的开头添加特殊标记 [CLS]（用于分类任务），在句子对之间添加特殊标记 [SEP]。

（2）输入表示：将每个词语转换为输入向量，输入向量由三部分组成：词嵌入（Token Embedding）：将词语映射为稠密向量；位置嵌入（Position Embedding）：表示词语在句子中的位置信息；段嵌入（Segment Embedding）：用于区分句子对中的不同句子（如问答任务中的问题和答案）。

（3）Transformer 编码器：将输入向量序列输入到多层 Transformer 编码器中。Transformer 的核心是自注意力机制（Self-Attention），它能够捕捉词语之间的长距离依赖关系。通过多层 Transformer 块，生成每个词语的上下文相关向量表示。

（4）预训练任务：掩码语言模型（Masked Language Model, MLM）：随机掩码输入序列中的部分词语（如 15%），并预测被掩码的词语。下一句预测（Next Sentence Prediction, NSP）：预测两个句子是否是连续的。

（5）微调：在特定任务（如文本分类、命名实体识别等）上，加载预训练的 BERT 模型。根据任务需求，在 BERT 的输出层上添加任务特定的分类器或回归器。使用任务数据对模型进行微调。

（6）对于分类任务，使用 [CLS] 标记的向量作为整个文本的表示。对于其他任务（如序列标注），使用每个词语的上下文相关向量作为输出。

BERT 的计数方式基于上下文相关的动态表示，不直接使用显式的计数方法。具体规则如下：

对于文本中的每个词语，通过 BERT 模型生成其上下文相关向量。对整个文本的 BERT 特征进行聚合（如取 [CLS] 标记的向量或词向量的平均值）：

$$BERT_{text} = BERT\left(\left[CLS\right]\right) \tag{3.3}$$

或

$$BERT_{text} = \frac{1}{n}\sum_{i=1}^{n} BERT(w_i) \tag{3.4}$$

其中，$BERT_{text}$ 是词语 w_i 的 $BERT$ 上下文相关向量。

3.2.2 BERT 特征维度确定

类似于前实验 GloVe 的特征维度确定，我们同样采用线性判别分析（LDA）作为分类器，来学习在 1 到 500 维特征维度范围内，分类性能最佳的 BERT 表示维度。BERT 可以从原始文本中提取更加丰富的动态语义信息，增强分类的表现。从三个板块的分类准确率随 BERT 表示维度变化而变化的曲线图（图 3-4）中可以看出，在 CN+WIND 互联网软件与服务Ⅲ板块，BERT 文本表示达到 200 维时，准确率最高；在 WIND 信息技术服务板块，准确率在 100 维时最高；而在 WIND 软件板块，200 维的 BERT 文本表示效果最佳。因此，针对 CN+WIND 互联网软件与服务Ⅲ板块、WIND 信息技术服务板块和 WIND 软件中的每篇 MD&A 文本 BERT 嵌入表示分别选择 200 维、100 维和 200 维。

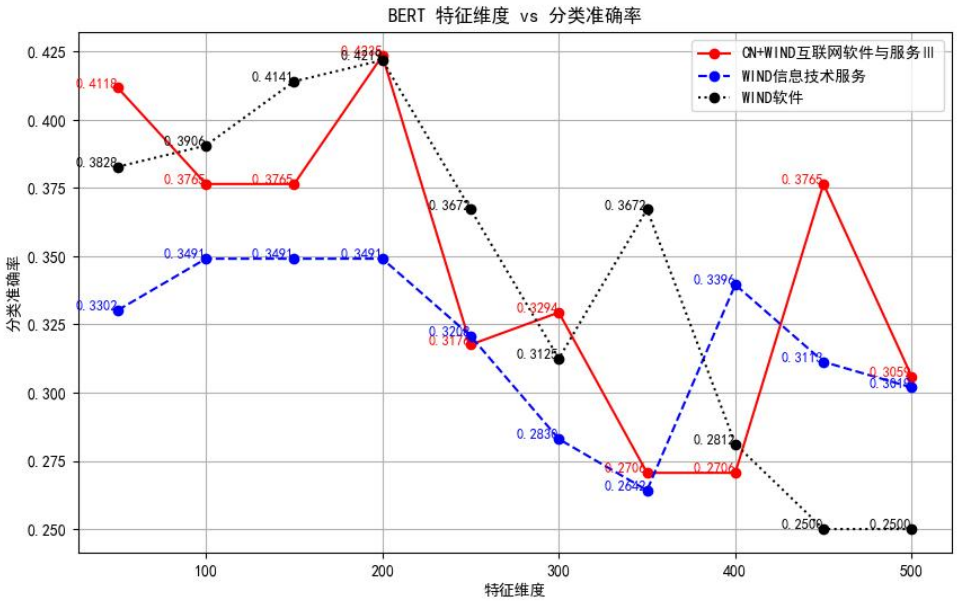


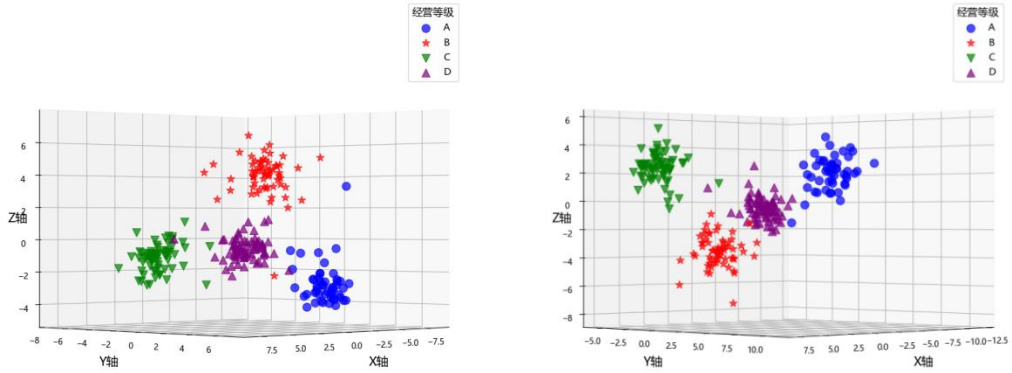
图 3-4 BERT 特征维度 vs 分类准确率

3.2.3 BERT 词嵌入的聚类可视化

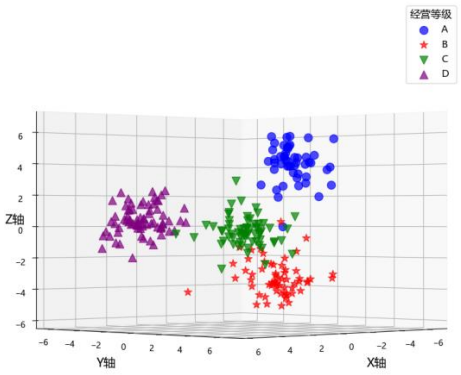
在对 CN+WIND 互联网软件与服务Ⅲ、WIND 信息技术服务、WIND 软件三个板块的 MD&A 文本分别采用 200 维、100 维和 200 维 BERT 嵌入后，我们可以发现，相较于 GloVe 词嵌入，BERT 在语义空间中的边界划分更加清晰，

板块式结构愈发鲜明。这一优势得益于其独特的双向注意力机制，能够跨越局部上下文，精准捕捉深层语义关联，使文本的语义表达更加细腻且富有层次感。换言之，BERT 不仅关注单个词的静态含义，更在动态上下文中挖掘语义演化轨迹，从而实现更加精准的语境感知与信息提炼。

更值得关注的是，在这一优化框架下，MD&A 文本的聚类特性尤为突出，各经营等级的语义边界逐渐分明，形成高度结构化的分布模式。这种聚类效果不仅印证了 BERT 在文本表征上的卓越性能，也凸显了其在企业经营状况分析中的应用潜力。相较于 GloVe，BERT 能够捕捉更精细的上下文依赖关系，有效区分经营状况的微妙差异，使文本分类更加稳定可靠。此外，BERT 的深度语义建模能力还赋予了文本更强的可解释性，使得经营状况的评估在复杂环境下仍能保持高效适用，为后续语义分析与文本分类任务提供了更坚实的技术支撑。如图 3-5 所示：



（a）　　（b）



（c）

图 3-5 BERT 文本表示的经营等级聚类图。（a）CN+WIND 互联网软件与服务III （b）WIND 信息技术服务 （c）WIND 软件

3.3 语义融合表示

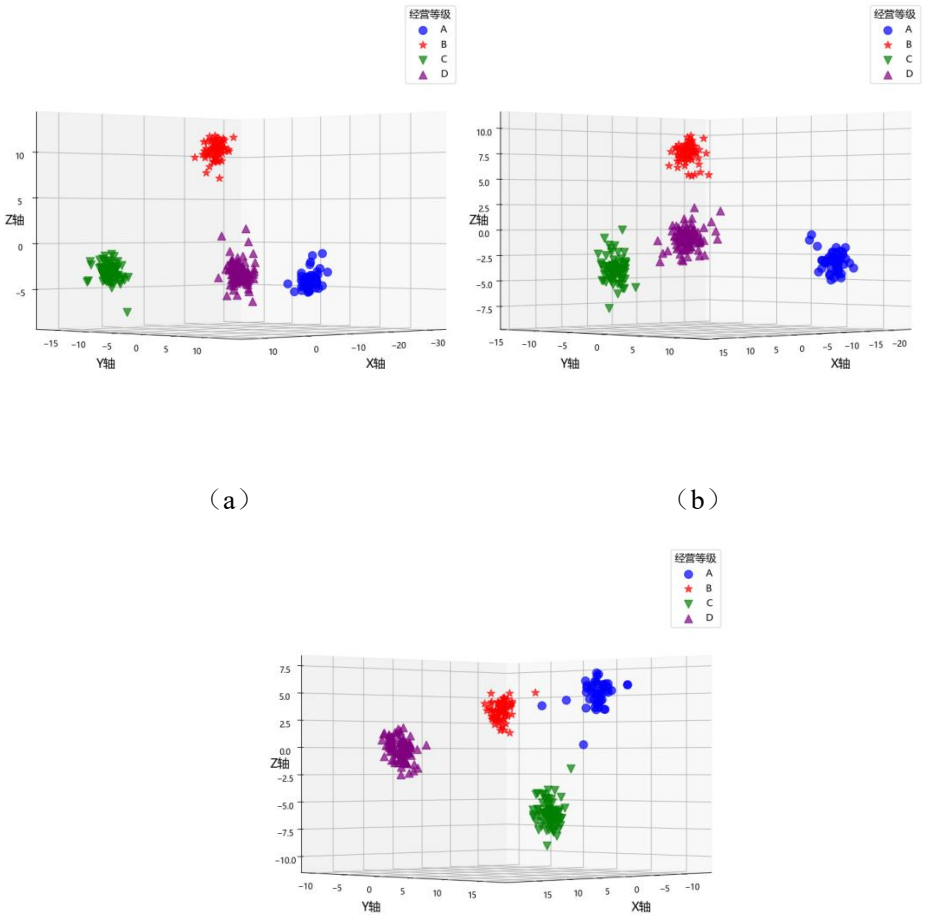
3.3.1 理论基础

通过拼接方式融合 GloVe 与 BERT 的文本特征是一种常见的特征融合策略，旨在综合利用 GloVe 提供的全局语义信息与 BERT 捕捉的上下文依赖关系，以增强文本表示的丰富性和表达能力，从而提升模型的语义理解能力与分类性能。假设 GloVe 向量的维度为d_g，BERT 向量的维度为d_b，则拼接后的特征向量维度为$d_g + d_b$。因此，对于 CN+WIND 互联网软件与服务III板块和 WIND 软件板块的每篇 MD&A 文本，我们为其生成了一个$100+200$维的融合表示；而对于 WIND 信息技术服务板块的每篇 MD&A 文本，则生成了一个$100+100$维的融合表示。基于此，CN+WIND 互联网软件与服务III、WIND 信息技术服务、WIND 软件三个板块的 MD&A 文本分别构建了564×300维的文档-词项矩阵，708×200维的文档-词项矩阵，849×300维的文档-词项矩阵。该融合策略有效整合了全局语义稳定性与局部上下文的敏感性，从而提升了文本表征的准确性和表达能力，为后续的文本分析与评估任务提供了更具鲁棒性的特征支持。

3.3.2 融合表示的聚类可视化

为验证这一融合特征的优势，我们生成了以下经营等级聚类图，展示了不

同经营等级的 MD&A 文本在语义空间中的清晰分布，进一步证明了融合表示在分类与聚类任务中的优越性能。如图 3-6 所示：



（a）　　（b）

（c）

图 3-6 融合表示的经营等级聚类图。（a）CN+WIND 互联网软件与服务III （b）WIND 信息技术服务 （c）WIND 软件

由图 3-6 可知，三个板块各经营等级的聚类效果相比于 GloVe 和 BERT 的单特征嵌入表示，表现出极为明显的提升。各经营等级的 MD&A 文本在语义空间中的边界划分非常清晰，没有任何经营等级出现混杂的情况。这一现象显著证明了我们采用的 MD&A 文本融合表示的有效性，进一步验证了通过融合 GloVe 和 BERT 嵌入所带来的全局稳定性和上下文适应性，使得文本在分类和聚类任务中的表现更加精确和可靠。

3.4 本章小结

本章通过对 CN+WIND 互联网软件与服务III、WIND 信息技术服务以及 WIND 软件三个板块的 MD&A 文本，分别进行 GloVe 和 BERT 词嵌入提取，并借助线性判别分析（LDA）作为分类器，系统地探索了在 1 到 500 维特征维度范围内，GloVe 和 BERT 的最佳分类性能。通过对 GloVe 与 BERT 表示维度的串联融合，我们构建了三个板块各自的 MD&A 文本融合表示：CN+WIND 互联网软件与服务III板块 564×300 维的文档-词项矩阵，WIND 信息技术服务板块 708×200 维的文档-词项矩阵，以及 WIND 软件板块 849×300 维的文档-词项矩阵。至此，文本的融合特征表示已然完成，成为文本挖掘中不可或缺的核心环节。这一过程至关重要，它通过将原始文本数据转化为便于计算机处理与分析的数值或向量形式，为后续分析打下了坚实的基础。

更为重要的是，随着融合特征表示的引入，三个板块中不同经营等级的 MD&A 文本聚类效果明显优于单一特征的聚类结果，这进一步证明了融合特征表示的高效性和准确性。在此基础上，下一章将基于本章的特征矩阵与构建的邻接矩阵，输入到图卷积网络（GCN）中进行上市公司 MD&A 经营状况的全面评估。这一跨越，不仅强化了分类和聚类任务的效果，也为经营状况的深度解析提供了更高层次的支持。

第四章 基于 PGBG-MD&A 的上市公司经营状况评估模型

MD&A 数据能够全面反映企业的经营状况，但由于其内容复杂且涉及多个维度的信息，使得挖掘其蕴含的经营信息会面临着诸多挑战。因此在所提出的评估模型 PGBG-MD&A 中，利用 Poisson 核将词向量表示模型 GloVe、大语言模型 BERT 以及 GCN 模型相融合，旨在有效聚合 MD&A 中词汇的语义相似性与文本之间的关联性，从而进一步提升语义传递效率和信息整合的准确性。具体实现流程如图 4-1 所示。

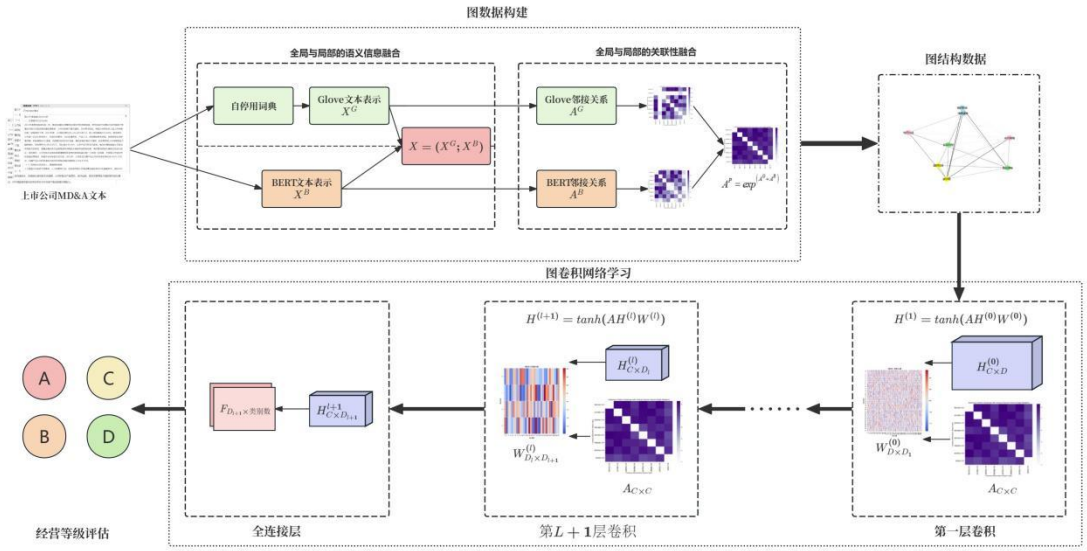


图 4-1 构建经营状况评估模型流程图

4.1 基于 PGBG-MD&A 的模型构建

4.1.1 全局与局部的语义信息融合

MD&A 提供了经营管理层对公司财务状况、经营成果以及未来展望等方面的信息。但其含有大量如“行业”、“报告”、“数字”等冗余信息，因此在获取反映上市公司经营状况的 MD&A 表示之前自建了停用词典。在此基础上，通过 GloVe 静态词嵌入模型获得 MD&A 的 GloVe 文本表示，一个 d_g 维的实向量 X^G，即：

$$X^G=\left[x_1^g,x_2^g,\cdots,x_{d_g}^g\right] \tag{4.1}$$

考虑到 MD&A 中通常包含复杂的上下文关系，单纯依赖 GloVe 获取词与词间的全局关系难以准确捕捉多义词在不同语境中的含义。为克服这一限制，又通过 BERT 预训练语言模型获得 MD&A 的 BERT 文本表示，一个 d_b 维的实向量 X^B，即：

$$X^B=\left[x_1^b,x_2^b,\cdots,x_{d_b}^b\right] \tag{4.2}$$

为使模型在处理复杂语义信息时既能保持一致性，又能灵活应对上下文语义关系，我们将 GloVe 表示 X^G 和 BERT 表示 X^B 串联得到 MD&A 的一个 d_g+d_b 维的实向量表示 X，即：

$$X=\left[X^G;X^B\right] \tag{4.3}$$

于是，N 个 MD&A 文本的融合表示为 $N\times\left(d_g+d_b\right)$ 维的矩阵：

$$\mathbf{X}=\begin{bmatrix}X_1\\X_2\\\vdots\\X_N\end{bmatrix}=\left[\begin{array}{ccc:ccc}x_{11}^g&\cdots&x_{1d_g}^g&x_{11}^b&\cdots&x_{1d_b}^b\\x_{21}^g&\cdots&x_{2d_g}^g&x_{21}^b&\cdots&x_{2d_b}^b\\\vdots&\ddots&\vdots&\vdots&\ddots&\vdots\\x_{N1}^g&\cdots&x_{Nd_g}^g&x_{N1}^b&\cdots&x_{Nd_b}^b\end{array}\right]_{N\times\left(d_g+d_b\right)} \tag{4.4}$$

如果将每篇 MD&A 文本视为一个节点，那么 $\mathbf{X}$ 中的 $X_i,i=1,2,\cdots,N$ 就可以作为第 i 个节点的融合了全局与局部语义信息的特征。

4.1.2 全局与局部的关联性融合

当两家上市公司的经营状况属于同一级别时，它们的 MD&A 文本往往表现出较强的关联性。反之，当两个 MD&A 文本的关联性较强时，我们也可以推测两家公司具有同级别的经营状况。为此，我们设置了一个关联性度量阈值 θ，用于刻画文本间关联性的强弱。如果每个 MD&A 文本对应一个节点，那么两个节点之间的邻近关系就能够衡量对应的两个 MD&A 文本之间的关联性强弱。邻近关系越近，说明两个 MD&A 文本之间的关联性越强；反之，关联性越强，两个 MD&A 文本对应的节点就更邻近。因此，当两个文本之间的关联性超过阈值

θ时，将邻近关系设置为关联性的倒数；否则，邻近关系设置为零。于是两个节点间的邻近关系A的计算公式如下：

$$A=\begin{cases}\dfrac{1}{\cos\_sim}, & if\ \cos\_sim>\theta\\ 0, & otherwise\end{cases} \tag{4.5}$$

其中$\cos_sim=\dfrac{\vec{\alpha}\cdot\vec{\beta}}{\|\vec{\alpha}\|\|\vec{\beta}\|}$，当$\vec{\alpha}$和$\vec{\beta}$为基于 Glove 的文本表示时，阈值$\theta$记为$\theta_g$，其对应的关联性记为$A^G$；当$\vec{\alpha}$和$\vec{\beta}$为基于 BERT 的文本表示时，阈值$\theta$记为$\theta_b$，其对应的关联性记为$A^B$。

Poisson 核能够有效平滑数据、降低噪声，在文本分类任务中能够提高刻画词嵌入表示间关联性强弱的度量的计算精度。因此，在评估模型 PGBG-MD&A 中通过引入 Poisson 核分别将基于 GloVe 的关联性和基于 BERT 的关联性相融合，从而增强评估性能。具体计算公式如下：

$$A^P=exp^{\left(A^G+A^B\right)} \tag{4.6}$$

于是，对于 N 个 MD&A 文本间的关联性的存在与否以及强弱可以用一个$N\times N$维的邻接矩阵$\mathbf{A}$来度量，即：

$$\mathbf{A}=\begin{bmatrix}A^P_{1,1} & \cdots & A^P_{1,N}\\ \vdots & \ddots & \vdots\\ A^P_{N,1} & \cdots & A^P_{N,N}\end{bmatrix} \tag{4.7}$$

其中$A^p_{i,j}$为第i个与第j个 MD&A 文本间融合了全局与局部关联性的邻近关系的度量。

4.1.3 图卷积网络学习

在我们的融合方案中，将每篇 MD&A 文本视为一个节点，并通过节点间边的存在与否来表示 MD&A 文本间经营状况关联性的存在与否，边的权重反映了 MD&A 文本间经营状况关联性的强弱。通过这种方式，我们能够有效地刻画上市公司之间的经营状况及其潜在的共性关系。为提升评估性能，将所学习到的文本的融合表示$\mathbf{X}$与融合了关联性的邻接矩阵$\mathbf{A}$作为输入，通过 GCN 进行两

层卷积操作[37-38]。在这一过程中，GCN 通过节点间的消息传递，增强了节点特征表达能力的同时又有效利用邻接矩阵中编码的图结构信息，生成了低维的图嵌入表示，深入挖掘了 MD&A 文本中潜藏的深层信息。其层间信息传播过程可表示为：

$$H^{(l+1)} = \sigma(\hat{\mathbf{A}}H^{(l)}W^{(l)}), \quad l = 0,1 \tag{4.8}$$

其中，$\hat{\mathbf{A}}$ 为标准化的邻接矩阵，W^l 是第 $l+1$ 层的权重矩阵，σ 是激活函数（在本模型中采用 Tanh 激活函数）。若 $l=0$，$H^{(0)}$ 就是要输入到第一个卷积层的特征矩阵 $\mathbf{X}$，$W^{(0)}$ 代表第一个卷积层的权重参数矩阵；若 $l=1$，$H^{(1)}$ 就是第一个卷积层的输出，即第二个卷积层的输入，$W^{(1)}$ 是第二个卷积层的权重参数矩阵，$H^{(2)}$ 就是第二个卷积层的输出，其维度设定为经营等级类别数，以便更精确地区分 MD&A 文本的经营状况。最后通过全连接层对 $H^{(2)}$ 进行分类映射，从而获得最终的经营等级分类。

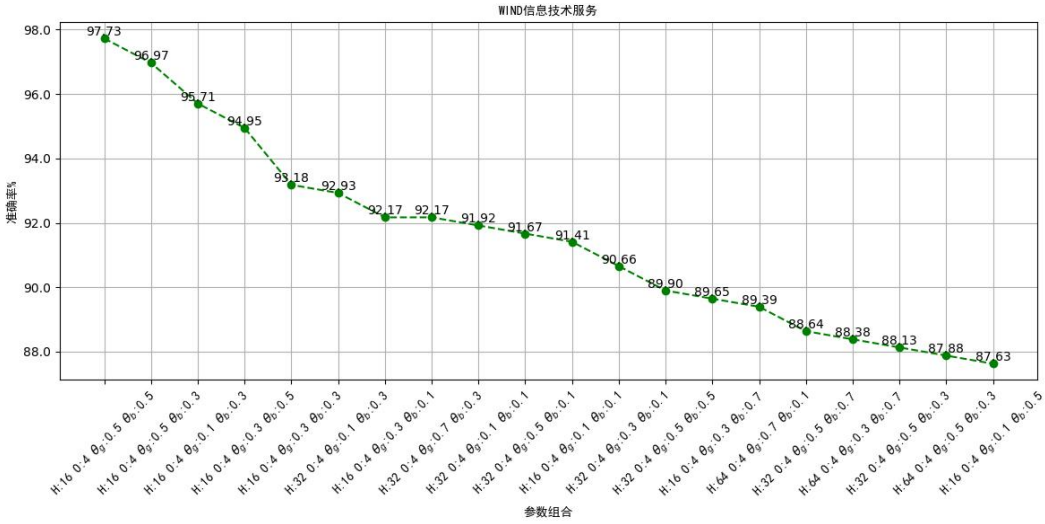
4.2 图卷积网络的构建

图卷积网络能够通过多层卷积操作逐步聚合节点及其邻居节点的信息，实现了局部与全局语义的高效融合，使得 PGBG-MD&A 模型不仅能够精准评估单个公司的经营状况，还能从整体网络视角揭示公司间复杂的关联关系。在这里，我们将每篇 MD&A 文本视为图卷积网络中的一个节点，并以该文本的融合表示作为节点的特征。通过 Poisson 核函数融合分别基于 GloVe 和 BERT 生成的关联性邻接矩阵，能够有效地聚合并挖掘 MD&A 文本中关于经营状况的深层信息。

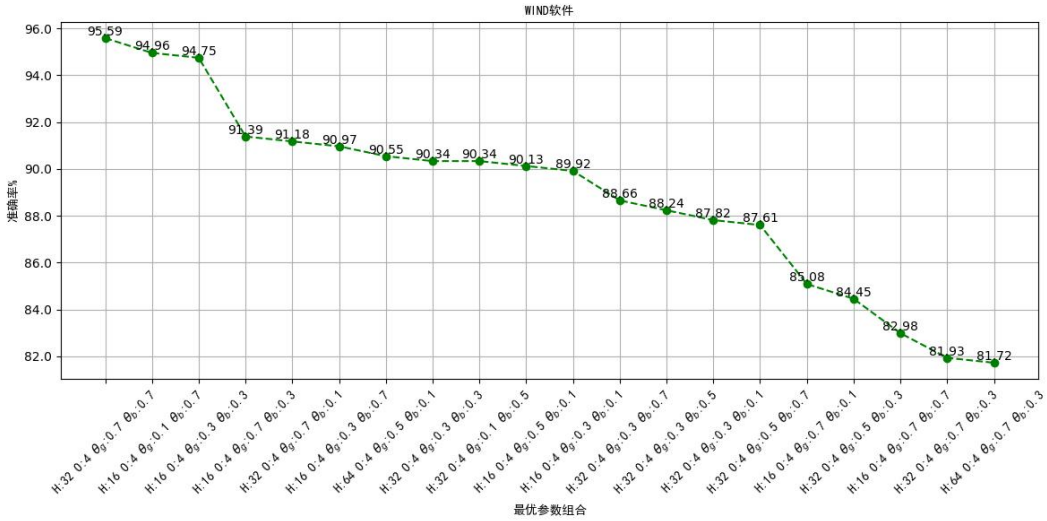
本节旨在探索以下两个关键问题：一是反映 GloVe 和 BERT 关联性强弱的阈值设定，二是卷积网络中第一层卷积层输出维度对经营状况评估准确性的影响。因此，我们采用了网格搜索这一超参数优化方法[39]，以经营状况评估的最高预测准确率为标准，系统地探索了 GloVe 与 BERT 的关联度阈值，以及第一层卷积层输出维度的最佳参数组合。我们利用训练集中的学习样本进行实验，统一了交叉熵损失迭代次数，并分别绘制了三个板块分类准确率从高到低随不

同参数组合变化的曲线，展示了预测准确率最高的前 20 个参数组合的变化趋势图（图 4-2），并以此来确定各板块的 GloVe 和 BERT 的关联度阈值以及图卷积网络结构。

（a）



（b）



（c）

图 4-2 超参数确定。（a）CN+WIND 互联网软件与服务III （b）WIND 信息技术服务 （c）WIND 软件

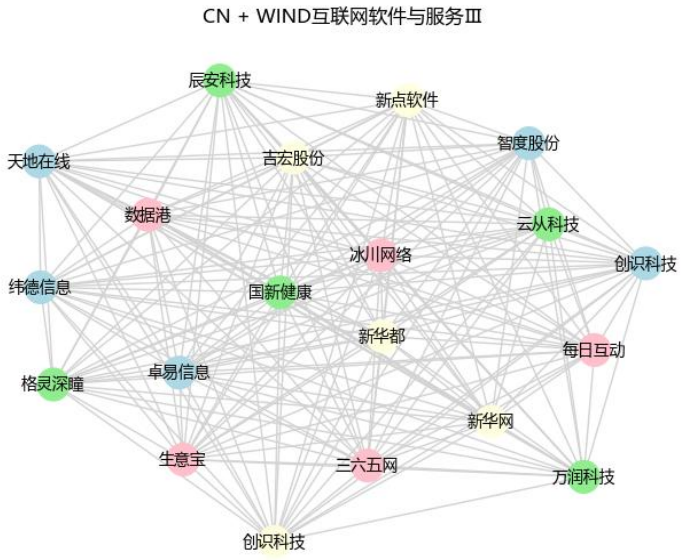
从子图（a）中可以看出，在 CN+WIND 互联网软件与服务III板块中，当 GloVe 的相似度阈值 θ_g 为 0.7，BERT 的相似度阈值 θ_b 为 0.5，且卷积层第一层的输出维度为 32 维时，预测准确率达到了最高值 99.68%。因此，对于 CN+WIND 互联网软件与服务III板块的 MD&A 文本所学习到的网络结构为：每个节点的特征维度为 300，第一个卷积层的权重参数矩阵的维度为 300×32，即权重参数 $W^{(0)}_{300\times32}$，第二个卷积层的权重参数矩阵维度为 32×4，即权重参数 $W^{(1)}_{32\times4}$。

从子图（b）中可以看出，在 WIND 信息技术服务板块中，当 GloVe 的相似度阈值 θ_g 和 BERT 的相似度阈值 θ_b 均为 0.5，且卷积层第一层的输出维度为 16 维时，预测准确率达到了最高值 97.73%。因此，对于 WIND 信息技术服务板块的 MD&A 文本所学习到的网络结构为：每个节点的特征维度为 200，第一个卷积层的权重参数矩阵的维度为 200×16，即权重参数 $W^{(0)}_{200\times16}$，第二个卷积层的权重参数矩阵维度为 16×4，即权重参数 $W^{(1)}_{16\times4}$。

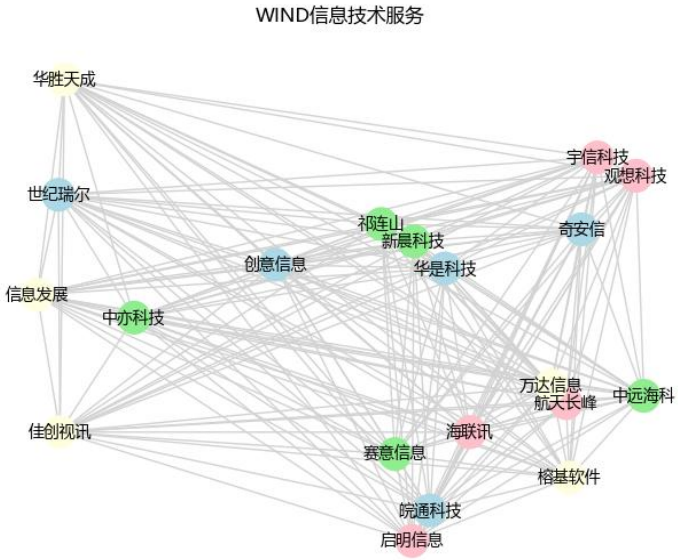
从子图（c）中可以看出，在 WIND 软件板块中，当 GloVe 的相似度阈值 θ_g 和 BERT 的相似度阈值 θ_b 均为 0.7，且卷积层第一层的输出维度为 32 维时，预测准确率达到了最高值 95.59%。因此，对于 WIND 软件板块的 MD&A 文本所学

习到的网络结构为：每个节点的特征维度为 300，第一个卷积层的权重参数矩阵的维度为 300×32，即权重参数 $W^{(0)}_{300\times32}$，第二个卷积层的权重参数矩阵维度为 32×4，即权重参数 $W^{(1)}_{32\times4}$。

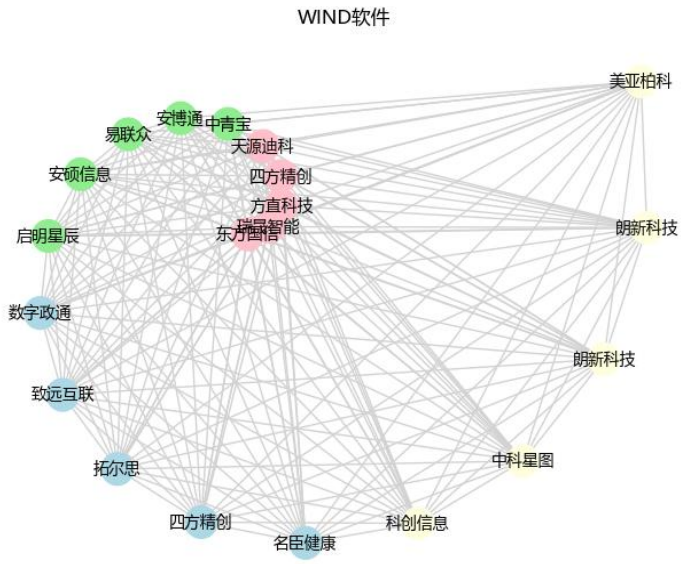
为了更直观地呈现这一分析结果，我们分别从 CN+WIND 互联网软件与服务Ⅲ，WIND 信息技术服务，WIND 软件三个板块中的四种经营状况随机选择了 5 家上市公司构建图数据如图 4-3 所示，从图 4-3（a）-（c）中可以看到每个板块的上市公司同种经营状况的节点基本保持邻近关系。



（a）



（b）



（c）

图 4-3 图数据结构。（a）CN+WIND 互联网软件与服务III （b）WIND 信息技术服务 （c）WIND 软件

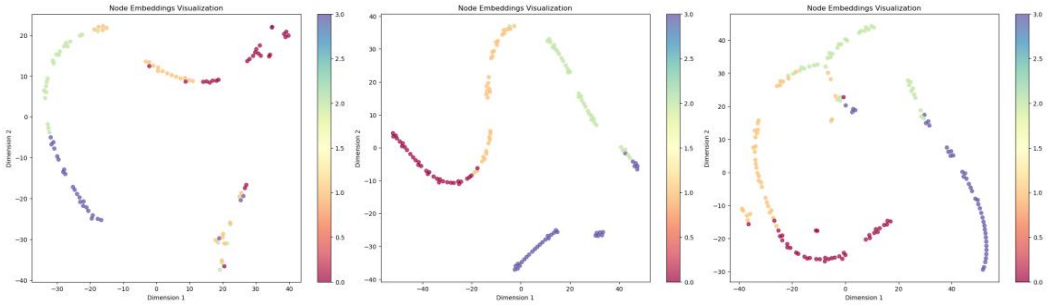
4.3 实验结果与分析

在本节中，通过图嵌入表示的可视化，与单特征模型的分类性能对比以及与经典评估模型的分类性能对比三个方面来验证所提出的 PGBG-MD&A 经营状况评估模型的有效性。

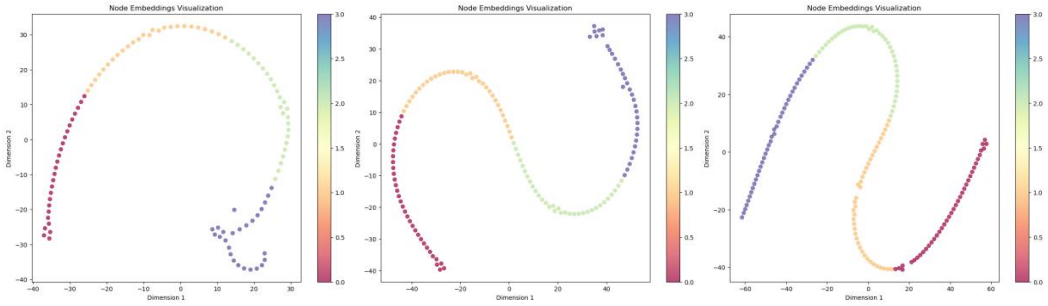
4.3.1 图嵌入表示的可视化

图嵌入作为一种将高维数据映射到低维空间的有效方法，能够帮助我们直观地分析文本数据的内部结构及潜在关系，从而揭示不同模型所学习到的融合表示之间的差异。因此，本节分别对基于 GloVe 和 BERT 的单特征模型 GG-MD&A、BG-MD&A 和融合模型 PGBG-MD&A 展开了三个板块训练集中的回测样本的图嵌入对比分析（图 4-4）。如图 4-4 所示：其中子图（a）展示了在单特征模型 GG-MD&A 中，嵌入点呈现出明显的散乱与重叠现象；子图（b）展示了在单特征模型 BG-MD&A 的嵌入点相对集中，重叠现象有所减少，但仍未能实现精细的区分；相比之下，子图（c）展示的融合模型 PGBG-MD&A 文本的图嵌入点在二维平面上聚集在紧密区域内，且不同经营等级的嵌入点显著

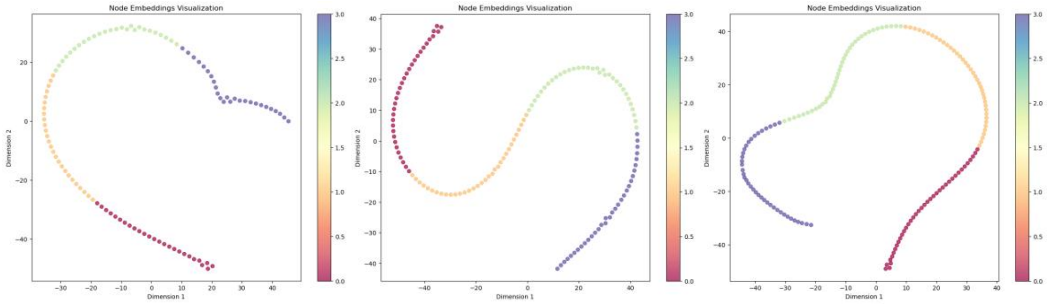
分开，呈现出明显的同质性与异质性，三个板块中不同经营等级的 MD&A 文本，在经过两层图卷积神经网络（GCN）处理后，呈现出显著的嵌入分布差异。



（a）



（b）



（c）

图 4-4 样本点的图嵌入效果图。（a）GG-MD&A 的词嵌入 （b）BG-MD&A 的词嵌入 （c）PGBG-MD&A 的词嵌入

4.3.2 与单特征模型的对比

在交叉熵损失迭代次数相同的情况下，对 PGBG-MD&A 模型与单特征模型（GG-MD&A 和 BG-MD&A）进行了经营状况等级预测准确率的对比分析，如图 4-5 所示。实验结果表明：在 CN+WIND 互联网软件与服务Ⅲ板块中，PGBG-MD&A 模型的预测准确率高达 98.08%，相较于 GG-MD&A 模型提升了 16.35%，相较于 BG-MD&A 模型提升了 5.77%；在 WIND 信息技术服务板块中，该模型的预测准确率达到 99.24%，分别比 GG-MD&A 模型和 BG-MD&A 模型高出 10.60% 和 2.57%；而在 WIND 软件板块，PGBG-MD&A 模型同样表现卓越，其预测准确率为 98.72%，比 GG-MD&A 模型提升了 13.46%，比 BG-

MD&A 模型提升了 6.41%。由此我们可以看出融合模型 PGBG-MD&A 比单特征模型在经营状况评估中更显优势。

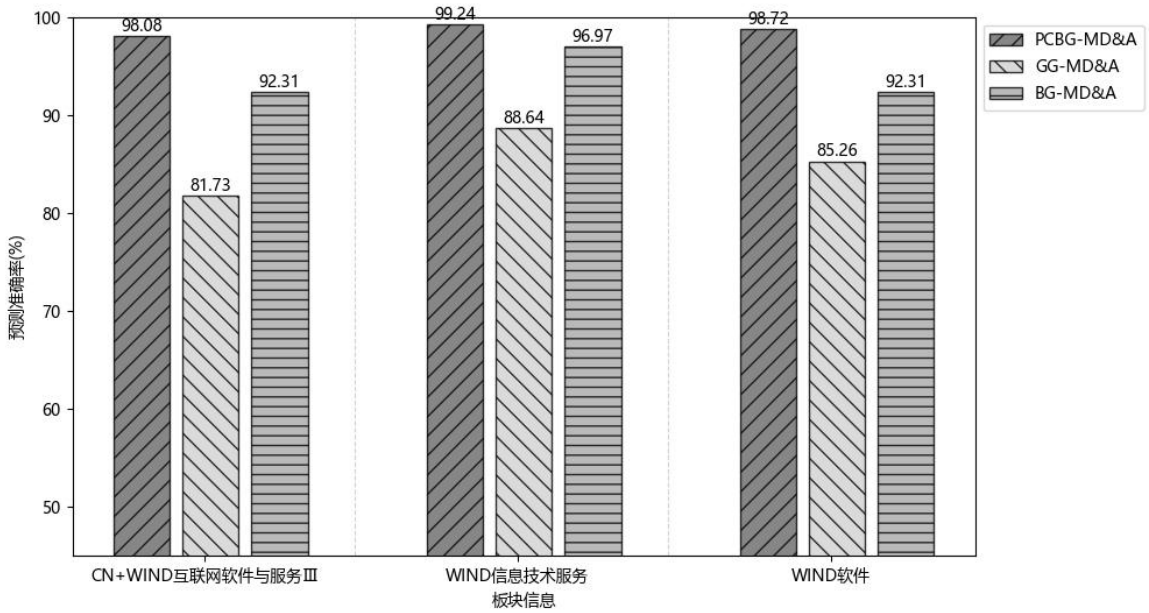


图 4-5 单特征模型与融合模型的预测准确率对比

接着，通过混淆矩阵对模型 PGBG-MD&A 的预测性能进行评估，如图 4-6 所示：

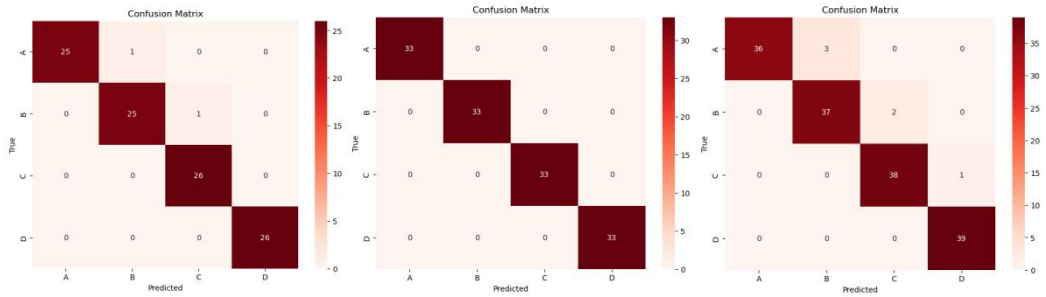


图 4-6 PCBG-MD&A 模型的混淆矩阵。（a）CN+WIND 互联网软件与服务III （b）WIND 信息技术服务 （c）WIND 软件

子图（a）展示了在 CN+WIND 互联网软件与服务III板块中，A、B、C、D 四个等级各有 26 家公司，其中 A 等级与 B 等级均为 25 家预测正确；C 等级与 D 等级公司预测结果与实际等级完全匹配。子图（b）展示了在 WIND 信息技术服务板块的 A、B、C、D 等级预测结果均与实际等级完全一致。子图（c）展示的 WIND 软件板块，A、B、C、D 四个等级各有 39 家公司，其中 A 等级 36 家公司预测准确；B 等级 37 家公司预测准确；C 等级 38 家公司预测准确；D 等级的预测则完全准确。

进一步，通过准确率、平均精确率、平均召回率和平均 F1 分数四个指标对模型 PGBG-MD&A 的预测性能进行评估，如表 4-1 所示：

表 4-1 每个板块的准确率、平均精确率、平均召回率和平均 F1 分数

	CN+WIND 互联网软件与服务III	WIND 信息技术服务	WIND 软件
准确率（%）	98.08	100	96.15
平均精确率（%）	98.11	100	96.25
平均召回率（%）	98.08	100	96.15
平均 F1 分数（%）	98.08	100	96.15

在 WIND 信息技术服务领域，所提出模型的准确率、平均精确率、平均召回率以及平均 F1 分数均达到了完美的 100%，这一结果彰显了模型在此领域内的卓越性能和高度准确性。同时，在 CN+WIND 互联网软件与服务III以及 WIND 软件这两个板块，PGBG-MD&A 模型的上述四项关键性能指标也均超过了 96%的优异水平。

4.3.3 与经典评估模型的对比

在本节中，利用测试集上的五折交叉验证评估模型的泛化性能，并将融合模型 PGBG-MD&A 与八种经典分类学习模型进行经营状况预测准确率的对比。这八种模型包括：

逻辑回归（Logistic Regression, LR）[40]是一种广义线性回归模型，尽管名称中带有“回归”，其主要用于二分类任务。该方法利用 sigmoid 函数将线性回归的输出映射至（0,1）区间，并通过设定阈值实现分类。由于 sigmoid 函数的性质，使得逻辑回归的输出可解释为概率，提高了模型的可理解性。其优化过程依赖于构建代价函数，并借助优化算法迭代求解参数，同时使用对数损失函数评估分类效果。逻辑回归因其计算效率高、实现简单、可解释性强，在数据挖掘、疾病预测和经济分析等领域广泛应用。

K 近邻算法（K-Nearest Neighbor, KNN）[41]是一种基于实例的学习方法，适用于分类和回归问题。其核心思想是基于待预测样本与 K 个最邻近样本的类别进行投票或加权平均，从而确定最终分类结果。该算法无需显式训练过程，只需计算样本间的距离或相似度，因此实现简便。在处理非线性分类任务时，KNN 具有良好的性能，广泛应用于数据挖掘和模式识别领域。

决策树（Decision Tree, DT）[42]采用树状结构进行分类，每个内部节点对应一个属性测试，分支表示测试结果，叶节点则代表最终类别或预测值。决策树可进一步分为分类树和回归树，分别用于处理离散型和连续型变量。构建过程中，采用贪心算法逐层划分数据，同时运用剪枝策略防止过拟合。决策树的优势在于结构直观、计算高效、易于解释，尤其适用于大规模数据集的分类任务。

支持向量机（Support Vector Machine, SVM）[43]是一种经典的监督学习方法，主要用于分类与回归任务。其核心原理是通过构造最优超平面，将数据进行分类，并最大化不同类别之间的间隔。SVM 依赖于少量支持向量来决定决策边界，具有良好的泛化能力，尤其适用于高维数据。此外，通过核函数，SVM 能够有效处理非线性分类问题，在数据挖掘、文本分类和生物信息学等领域有广泛应用。

随机森林（Random Forest, RF）[44]是一种集成学习方法，通过构建多个决策树并结合投票或平均策略提高分类精度。其核心思想在于“随机性”：即在构建每棵树时随机选取样本和特征，从而形成一片“森林”。随机森林具备较强的抗过拟合能力，并能在高维数据场景下保持稳定的分类性能。同时，它对缺失值与异常值具有良好的鲁棒性，被广泛用于金融分析、医学诊断和市场预测等领域。

堆叠模型（Stacking）[45]作为一种集成学习策略，通过组合多个基学习器的预测结果，优化最终分类性能。Stacking 采用多层结构，首先由多个基础模型对数据进行预测，再将预测结果作为新特征输入到元学习器进行训练。通常，k 折交叉验证用于生成训练集和测试集，以防止过拟合。该方法能够充分结合不同模型的优势，提升预测准确性，适用于复杂的非线性任务，如图像识别、自然语言处理和金融风控等。

图神经网络（Graph Neural Network, GNN[46]是一种专门用于处理图结构数据的深度学习方法，它结合了图论与神经网络技术，通过递归方式聚合邻居节点的特征，从而不断优化节点表示。GNN 能够有效捕捉数据中的复杂关系，在社交网络分析、推荐系统和知识图谱等领域表现出卓越的性能。相比传统图算

法，GNN 具有更强的表达能力和适应性，不仅可以端到端训练，还能灵活应用于不同类型的任务。

图注意力网络（Graph Attention Network, GAT）[47]是图神经网络（GNN）的改进版本，它引入了注意力机制，使得节点在信息传递过程中能够动态调整邻居节点的权重。通过学习不同邻居的重要性，GAT 自适应地优化特征聚合方式，无需预设图结构或进行复杂的矩阵计算。多层 GAT 能够充分挖掘图数据的复杂关系，提升分类性能。在多个基准数据集上的实验结果表明，GAT 的表现优于传统 GNN 模型，已广泛应用于文本分析、知识图谱构建和生物信息学等领域。

进行经营状况预测准确率的对比如表 4-2 所示：

表 4-2 提出的模型与经典模型预测性能的对比

方法	CN+WIND 互联网软件与服务III	WIND 信息技术服务	WIND 软件
	准确率（%） Mean±SD	准确率（%） Mean±SD	准确率（%） Mean±SD
LR	32.96±10.50	43.57±8.27	39.04±8.92
KNN	31.26±9.41	42.14±6.14	29.63±8.48
DT	36.48±9.51	35.71±4.52	38.97±9.14
SVM	49.13±8.80	45.71±7.95	43.76±11.12
RF	44.58±9.58	43.57±5.25	40.18±6.32
Stacking	34.82±5.75	33.57±6.62	32.50±5.67
GNN	51.43±4.29	43.29±10.69	43.26±1.86
GAT	54.29±5.71	43.84±9.75	46.05±5.58
PGBG-MD&A	**95.14±2.84**	**96.59±1.87**	**95.91±2.28**

根据表 4-2 可知，本文提出的 PGBG-MD&A 模型在三个板块的经营状况预测中始终表现出最优的性能。在 CN+WIND 互联网软件与服务III板块，尽管图注意力网络（GAT）在所有机器学习和深度学习模型中表现最佳，其平均准确率为 54.29%，标准差为 5.71%，但 PGBG-MD&A 模型的预测准确率高达 95.14%，标准差为 2.84%，相较于 GAT 提升了 40.85%。在 WIND 信息技术服务板块，支持向量机（SVM）表现优异，平均准确率为 45.71%，标准差为 7.95%。然而，PGBG-MD&A 模型的预测准确率达到 96.59%，标准差仅为

1.87%，相比 SVM 提升了 50.88%。在 WIND 软件板块，尽管图注意力网络（GAT）以 46.05%的准确率和 5.58%的标准差位居所有模型中的首位，PGBG-MD&A 模型的预测准确率依然高达 95.91%，标准差为 2.28%，相较于 GAT 提升了 49.86%。

4.4 本章小结

随着大模型技术的飞速发展，尤其是大语言模型在各领域的广泛应用，文本分析领域迎来了新的机遇。大语言模型凭借其强大的语义理解和文本生成能力，在多个应用场景中展现了巨大潜力。本文通过 Poisson 核将词嵌入表示模型 GloVe、大语言模型 BERT 以及图卷积网络（GCN）模型相融合，构建了一种基于 MD&A 文本的上市公司经营状况评估模型 PGBG-MD&A。实验结果表明：

（1）对于 CN+WIND 互联网软件与服务III板块，当 GloVe 特征维度为 100，余弦相似度阈值设为 0.7；BERT 特征维度为 200，余弦相似度阈值设为 0.5；权重参数矩阵 $W^{(0)}$ 列维度 $D_1=32$，$W^{(1)}$ 列维度 $D_2=4$ 时预测准确率最高为 97.98%。

（2）对于 WIND 信息技术服务板块，当 GloVe 特征维度为 100，余弦相似度阈值设为 0.5；BERT 特征维度为 100，余弦相似度阈值设为 0.5；权重参数矩阵 $W^{(0)}$ 列维度 $D_1=16$，$W^{(1)}$ 列维度 $D_2=4$ 时预测准确率最高为 98.46%。

（3）对于 WIND 软件板块，当 GloVe 特征维度为 100，余弦相似度阈值设为 0.7；BERT 特征维度为 200，余弦相似度阈值设为 0.7；权重参数矩阵 $W^{(0)}$ 列维度 $D_1=32$，$W^{(1)}$ 列维度 $D_2=4$ 时预测准确率最高为 98.19%。

这些实验结果充分验证了经营状况评估模型 PGBG-MD&A 在文本分析中的可行性与有效性。

第 5 章 总结与展望

5.1 结论

本文创新性地融合了大语言模型 PGBG-MD&A，对上市公司的经营状况进行了精准而全面的评估。首先，构建了一套定制化的经营状况停用词典，有效减少了冗余信息，显著降低了文本特征的维度，使分析更加高效。随后，借助 GloVe 与 BERT 词嵌入模型，将文本转换为高维语义向量，深度挖掘词汇间的关联性，从而增强模型的表达能力。值得注意的是，这种多层次优化不仅精炼了文本信息，还在特征降维的同时保留了关键语义，使数据处理更具针对性。在此基础上，模型进一步通过图卷积网络（GCN）提取文本的全局和局部特征，实现跨文档的语义关联建模。相比传统方法，本研究所构建的框架兼具广度与深度，为企业经营状况的智能分析提供了一种高效且精准的技术路径。其中：

（1）对于 CN+WIND 互联网软件与服务III板块，当 GloVe 特征维度为 100，余弦相似度阈值设为 0.7；BERT 特征维度为 200，余弦相似度阈值设为 0.5；权重参数矩阵 $W^{(0)}$ 列维度 $D_1=32$，$W^{(1)}$ 列维度 $D_2=4$ 时预测准确率最高为 97.98%。

（2）对于 WIND 信息技术服务板块，当 GloVe 特征维度为 100，余弦相似度阈值设为 0.5；BERT 特征维度为 100，余弦相似度阈值设为 0.5；权重参数矩阵 $W^{(0)}$ 列维度 $D_1=16$，$W^{(1)}$ 列维度 $D_2=4$ 时预测准确率最高为 98.46%。

（3）对于 WIND 软件板块，当 GloVe 特征维度为 100，余弦相似度阈值设为 0.7；BERT 特征维度为 200，余弦相似度阈值设为 0.7；权重参数矩阵 $W^{(0)}$ 列维度 $D_1=32$，$W^{(1)}$ 列维度 $D_2=4$ 时预测准确率最高为 98.19%。

且模型随着训练样本的变化，鲁棒性较好，在与主流的 LR、KNN、DT、SVM、RF、Stacking 六个机器学习模型与两个 GNN、GAT 深度学习模型相比，其预测性能始终保持最优。三个板块的三维图嵌入表示明确地区分了不同经营等级的 MD&A 文本。

5.2 不足与展望

1. 数据集的局限性

不足：本文的实验数据主要来源于 Wind 数据库中的特定板块（如 CN+WIND 互联网软件与服务III、WIND 信息技术服务和 WIND 软件），虽然这些板块具有一定的代表性，但数据集的覆盖范围仍然有限，未能涵盖更多的行业和板块。此外，数据的时间跨度较短（2021-2023 年），可能无法充分反映不同经济周期下的企业经营状况变化。

展望：未来的研究可以扩展数据集的覆盖范围，纳入更多的行业和板块，并增加数据的时间跨度，以验证模型在不同经济环境下的适应性和稳定性。此外，可以考虑引入更多的外部数据源（如新闻、社交媒体等），以丰富模型的输入信息。

2. 多模态数据的融合

不足：本文主要基于文本数据（MD&A）进行经营状况评估，未能充分利用其他类型的数据（如图像、音频、视频等）。随着多模态数据的广泛应用，单一文本数据的分析可能无法全面反映企业的经营状况。

展望：未来的研究可以探索如何将多模态数据（如企业发布的视频、图片等）与文本数据相结合，构建更加全面的经营状况评估模型。通过多模态数据的融合，模型可以更好地捕捉企业的多维信息，提升评估的准确性。

尽管 PGBG-MD&A 模型在上市公司经营状况评估中表现出色，但仍存在一些不足之处。未来的研究可以从数据集扩展、模型可解释性、泛化能力、实时性、多模态数据融合、模型优化和应用场景拓展等方面进行改进和拓展，进一步提升模型的性能和应用价值。

参考文献

[1] 路世昌, 黄贺楠, 李丹. 环境信息披露、技术创新投入及企业高质量发展的相互关系研究——基于制造业上市公司的经验证据[J]. 科技促进发展, 2021, 17(8): 1590-1599.

[2] 杨杰, 陈隆轩. 高质量发展背景下高管内部薪酬差距与公司绩效关系研究——来自我国 A 股制造业上市公司的新证据 [J]. 哈尔滨商业大学学报(社会科学版), 2021, (05): 3-14+30.

[3] 蔚莹. 基于语义网的“网易云”评论文本分析 [J]. 科技与创新, 2024, (22): 181-184.

[4] 蒲平, 石烨烨. 网络流行语近 20 年的情感转场研究——基于 2004—2023 年网络流行语的文本分析 [J]. 新闻爱好者, 2024, (11): 39-42.

[5] 邱吉福, 杨锋源, 张航. 审计能抑制管理层 MD&A 语调操纵吗——基于关键审计事项披露视角[J]. 会计之友, 2023, (22): 20-27.

[6] Choi J, Suh Y, Jung N. Predicting corporate credit rating based on qualitative information of MD&A transformed using document vectorization techniques[J]. Data Technologies and Applications, 2020, 54(2): 151-168.

[7] 高国静, 麦沛添, 刘桂金. 基于人工智能的大语言模型在中医领域的应用 [J]. 现代医院, 2025, 25 (02): 282-287+292.

[8] 王浩, 陈广磊, 王涵, 等. 海关知识问答场景下的大语言模型应用研究 [J]. 中国口岸科学技术, 2025, 7 (02): 4-9.

[9] Ding S, Ye J, Hu X, et al. Distilling the knowledge from large-language model for health event prediction [J]. Scientific Reports, 2024, 14 (1): 30675.

[10] Dewi C, Dai G, Christanto J H. Analysis of Internet Movie Database with Global Vectors for Word Representation[J]. Vietnam Journal of Computer Science, 2024, 11(3): 343-362.

[11] 张怀博. 基于 Glove 模型的文本分类方法与应用研究[D]. 华东师范大学, 2024.

[12] Eang C, Lee S. Improving the Accuracy and Effectiveness of Text Classification Based on the Integration of the Bert Model and a Recurrent Neural Network (RNN Bert Based)[J]. Applied Sciences, 2024, 14(18): 8388.

[13] Bo P, Tao Z, Kundong H, et al. BVMHA: Text classification model with variable multi head hybrid attention based on BERT[J]. Journal of Intelligent & Fuzzy Systems, 2024, 46(1): 1443-1454.

[14] Yinbin P, Wei W, Jiansi R, et al. Novel GCN Model Using Dense Connection and Attention Mechanism for Text Classification[J]. Neural Processing Letters, 2024, 56(2): 1-17.

[15] 刘赏, 陈浩, 陈小玉, 等. 面向交通流预测的双分支时空图卷积神经网络[J]. 信息与控制, 2023, 52(3): 391-404+416.

[16] Ziqing Y, Shoudong H, Jun Z. Poisson kernel: Avoiding self-smoothing in graph convolutional networks[J]. Pattern Recognition, 2022, 124:108443.

[17] 杨赞伟, 郑亮亮, 曲宏松, 等. 联合均值滤波与泊松核双边滤波降噪算法研究[J]. 计算机仿真, 2020, 37 (09): 460-463+468.

[18] Sullivan D. The need for text mining in business intelligence[J]. DM REVIEW, 2000, 10: 12-16.

[19] Thuraisingham B, Maning D. technologies, techniques, tools, and Trends[M]. CRC press, 1999.

[20] Bhattacharya I, Getoor L. A latent dirichlet model for unsupervised entity resolution[C]//Proceedings of the 2006 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2006: 47-58.

[21] Nina J. The readability of abstracts in library and information science journals [J]. Journal of Documentation, 2022, 79 (7): 1-11.

[22] Kunli Z, Bin H, Feijie Z, et al. Graph-based structural knowledge-aware network for diagnosis assistant. [J]. Mathematical biosciences and engineering : MBE, 2022, 19 (10): 10533-10549.

[23] Zheng Z, Zhang O, Borgs C, et al. ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis[J]. Journal of the American Chemical Society, 2023, 145(32): 18048-18062.

[24] Williams L, Anthi E, Burnap P. Comparing Hierarchical Approaches to Enhance Supervised Emotive Text Classification [J]. Big Data and Cognitive Computing, 2024, 8 (4): 38.

[25] Gurcan F, Cagiltay N E. Research trends on distance learning: a text mining-based literature review from 2008 to 2018[J]. Interactive Learning Environments, 2023, 31(2): 1007-1028.

[26] Torabi M, Haririan I, Foroumadi A, et al. A deep learning model based on the BERT pre-trained model to predict the antiproliferative activity of anti-cancer chemical compounds. [J]. SAR and QSAR in environmental research, 2024, 35 (11): 21-22.

[27] Zhou C, Dong X, Ji L , et al. Hierarchical mining algorithm for high dimensional spatiotemporal big data based on association rules [J]. E3S Web of Conferences, 2021, 256 02040.

[28] 朱芷璇. 基于 FoolNLTK 的中文分词改进研究与应用[D]. 西安: 陕西科技大学, 2023.

[29] 陈加元, 刘彦. 基于 LDA 模型的以成果为导向教育研究热点主题分析 [J]. 电脑知识与技术, 2024, 20 (36): 137-140+143.

[30] 周中雨. 基于文本挖掘的网络舆情话题分析方法研究[D]. 大庆: 东北石油大学, 2024.

[31] 汪克尧. 基于图神经网络的文本分类算法研究[D]. 江南大学, 2023.

[32] 文飞. 传统与大模型并举：中文文本分类技术对比研究 [J]. 智能计算机与应用, 2024, 14 (06): 88-94.

[33] 陈美, 赵子莜. 基于文本挖掘的开放政府数据与数字经济政策协同研究[J].情报杂志, 2024, 43(4): 184-191+88.

[34] 唐媛, 陈艳平, 扈应, 等. 基于多尺度混合注意力卷积神经网络的关系抽取[J/OL].计算机应用, 1-7.

[35] 李成刚, 贾鸿业, 赵光辉, 等. 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(2): 18-29.

[36] 顾水彬, 罗佳敏. 中国资本市场信息披露改革研究——基于管理层讨论与分析视角 [J]. 财会月刊, 2025, 46 (03): 61-66.

[37] 王婷, 朱小飞, 唐顾. 基于知识增强的图卷积神经网络的文本分类[J]. 浙江大学学报(工学版), 2022, 56 (2): 322-328.

[38] Guoshu L, Li Y, Sichang B, et al. BIKAGCN: Knowledge-Aware Recommendations Under Bi-layer Graph Convolutional Networks[J]. Neural Processing Letters, 2024, 56(1): 20.

[39] Pillai R, Sharma N, Gupta S, et al. Enhanced skin cancer diagnosis through grid search algorithm-optimized deep learning models for skin lesion analysis[J]. Frontiers in Medicine, 2024, 11: 1436470.

[40] Balboa A, Cuesta A, Villa G J, et al. Logistic regression vs machine learning to predict evacuation decisions in fire alarm situations [J]. Safety Science, 2024, 174: 106485.

[41] Wang B X, Feng K Z, Wang C H, et al. Model for quality classification of dam foundation rock mass based on Gaussian function weighted KNN algorithm and its application[J]. Bulletin of Engineering Geology and the Environment, 2024, 83(12): 510.

[42] Hafezi G S, Sahranavard T, Kooshki A, et al. Predicting high sensitivity C-reactive protein levels and their associations in a large population using decision tree and linear regression[J]. Scientific reports, 2024, 14(1): 30298.

[43] Haddoud M, Mokhtari A, Lecroq T, et al. Combining supervised term-weighting metrics for SVM text classification with extended term representation[J]. Knowledge and Information Systems, 2016, 49(3): 909-931.

[44] Fang W, Ren K, Liu T, et al. An evaluation of random forest based input variable selection methods for one month ahead streamflow forecasting[J]. Scientific reports, 2024, 14(1): 29766.

[45] 关慧, 许航, 蔡丽娥. 基于词性和改进 Stacking 模型的需求依赖关系提取[J]. 计算机工程与设计, 2024, 45(11): 3345-3351.

[46] Sejan S A M, Rahman H M, Aziz A M, et al. Powerful graph neural network for node classification of the IoT network[J]. Internet of Things, 2024, 28: 101410.

[47] 施荣华, 金鑫, 胡超. 基于图注意力网络的方面级别文本情感分析[J]. 计算机工程, 2022, 48(2): 34-39.

硕士期间发表的论文

[1]杜杨, 张玥，史龙梅，邹建.A Graph Convolutional Classification Model for MD&A Text Based on GloVe-BERT Feature Fusion and Poisson Correlation Enhancement.（第一作者 ISF 2025 已录用）

致谢

岁月如梭，光阴的河流悄无声息地流淌，转瞬之间，研究生的三年生涯已悄然落幕。在这段充满磨砺与成长的旅程中，我的心灵经历了深刻的蜕变，对人生的真谛与学术的殿堂有了更为透彻的领悟。在此，我满怀感激之情，向那些在求知路上为我点亮明灯的人们致以最诚挚的谢意。

首先，我要向我的导师张玥教授表达深深的敬意与感激。张老师，这位治学严谨、责任心强的学者，以她无尽的耐心与智慧，引领我在学术的海洋中破浪前行。面对我薄弱的学术根基，张老师总是慷慨地牺牲个人宝贵时间，以细致入微的指导和深入浅出的解答，为我扫清研究路上的重重障碍，指引我明确研究方向。她深厚的学术底蕴、对科研的执着热爱以及那份对完美的不懈追求，不仅为我的学术生涯奠定了坚实的基础，更在无形中塑造了我的人格魅力，教会了我诸多人生的哲理。

接着，我要感谢实验室中的同桌杨颖。在我为海量数据所困时，是她那精湛的 Excel 技艺，如同春风化雨，极大地提升了我的工作效率，让我的数据整理之路变得顺畅无阻。此外，我还要向我的同门王越、丁沈杰、杨灿清、张娜表达由衷的谢意。他们不仅在论文的行文逻辑与排版布局上给予了我宝贵的建议，更以其独到的见解为我的论文增色不少，使之更加完善。

同时，我深感荣幸能置身于这所学术氛围浓厚、设施完备的学府之中。学校为我们提供了良好的研究与学习环境，为我的学术探索之路铺设了坚实的基石。对于参与论文评审与答辩的各位专家与教授，我同样心怀感激。他们的真知灼见与宝贵意见，如同璀璨星辰，照亮了我前行的道路，使我的论文得以更加精进。

最后，我要将最深的感谢献给我的父母。他们是我生命中永恒的港湾，无论风雨兼程，始终给予我最坚实的支持与最温暖的关怀。在我埋头于论文撰写的日子里，是他们的默默付出与无私奉献，让我感受到了家的温馨与力量，成为我不断前行的强大动力。

在此，我再次向所有给予我帮助与支持的人们表达最真挚的感谢！愿这份

感激之情如繁星点点，照亮我前行的道路；愿我的文字如潺潺流水，虽历经岁月洗礼，仍能传递出那份不变的感恩之心。