分类号：O29　　单位代码：10363
密　级：公开　　学　　号：2221020104

题　　目 融合对比学习与图卷积神经网络的 MD&A 文本分类模型研究

题　　目 Research on an MD&A Text Classification Model Integrating Contrastive Learning and Graph Convolutional Neural Networks

学生姓名：杨灿清
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： 

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：杨灿清　　指导教师签名：张调

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

融合对比学习与图卷积神经网络的 MD&A 文本分类模型研究

摘 要

随着信息技术的快速发展和金融文本数据的指数级增长，文本分类技术在上市公司财务报告分析中的重要性日益凸显。管理层讨论与分析（MD&A）作为财务报告的核心组成部分，其精准分类对评估企业经营状况具有重要价值。然而，传统文本分类方法在处理 MD&A 文本时面临语义复杂性高、专业术语密集及上下文关联性强等挑战，因此，探索更为高效、精确的 MD&A 文本分类方法显得尤为重要。

对比学习作为表征学习的重要分支，通过挖掘数据内在结构信息，能够有效区分样本间的相似性与差异性。图卷积神经网络（GCN）凭借其处理图结构数据的独特优势，在捕捉复杂语义关联性方面表现卓越。将对比学习和 GCN 相结合，可以提升分类任务的准确性。因此，本文采用 BERT 获取 MD&A 文本特征，通过核支持向量机（KSVM）筛选高鉴别性样本，结合对比学习训练基于 GloVe 的正负样本对，增强类间差异感知，并将 BERT 特征和 GloVe 特征串联，利用 GCN 挖掘文本的邻接关联，实现了 MD&A 文本的高精度分类。主要研究内容如下：

（1）提出基于对比学习的 CL-KSVM 文本分类模型。首先，基于 BERT 模型提取 MD&A 文本的深层语义特征，并通过 KSVM 对样本进行归约，筛选出类间间隔最大化的高鉴别性样本筛选高鉴别性样

本，解决对比学习中表征坍塌问题。在此基础上，通过对比学习机制训练基于 GloVe 的正负样本对，将相似的样本拉近，同时将不相似的样本推远，使模型能够更好地识别关键信息的差异并捕捉到样本之间的内在关系。

（2）提出融合对比学习与 GCN 的 MD&A 文本分类模型 CG-KSVM。结合对比学习对信息间细微差异的敏感度，以及 GCN 在处理图结构数据上的独特优势，在 CL-KSVM 模型的基础上，将每个样本视为图中的一个节点，BERT 表示和 GloVe 表示的串联作为节点的特征，基于余弦相似度阈值构建邻接矩阵。然后，利用 GCN 聚合相邻节点的信息来更新节点表示，以增强文本表示的鲁棒性。GCN 通过引入图结构数据，能够有效处理文本数据中的复杂关系，捕捉文本间的语义关联性和上下文信息，从而进一步提升分类效果。

为验证模型的分类性能，本研究基于 WIND 数据库中 SW 电子板块的四个行业板块的 MD&A 文本构建实验数据集，通过与单特征模型、传统机器学习模型与深度学习模型的对比实验，验证了所提出模型的有效性。

关键词：对比学习；图卷积神经网络；核支持向量机；管理层讨论与分析；文本分类

Research on an MD&A Text Classification Model Integrating Contrastive Learning and Graph Convolutional Neural Networks

ABSTRACT

With the rapid development of information technology and exponential growth of financial textual data, text classification techniques have become increasingly crucial in analyzing listed companies' financial reports. As the core component of financial reports, Management's Discussion and Analysis (MD&A) requires precise classification for accurate business evaluation. However, traditional text classification methods face significant challenges when processing MD&A texts, including high semantic complexity, domain-specific terminology density, and strong contextual dependencies. Therefore, developing more efficient and accurate MD&A text classification methods has become particularly important.

Contrastive learning, as an important branch of representation learning, effectively distinguishes similarities and differences between samples by mining intrinsic data structures. Graph Convolutional Networks (GCN) demonstrate exceptional performance in capturing complex semantic relationships through their unique advantage in processing graph-structured data. The integration of contrastive learning

and GCN can significantly improve classification accuracy. Accordingly, this study employs BERT to extract MD&A text features, uses Kernel Support Vector Machines (KSVM) to select highly discriminative samples, and incorporates contrastive learning to train GloVe-based positive/negative sample pairs, thereby enhancing inter-class differentiation. The concatenated BERT and GloVe features are then processed through GCN to explore textual adjacency relationships, achieving high-precision MD&A text classification. The main contributions include:

(1) Propose a contrastive learning-based CL-KSVM text classification model. The framework first extracts deep semantic features using BERT, then applies KSVM for sample reduction to select margin-maximizing discriminative samples, addressing the representation collapse issue in contrastive learning. The model then trains GloVe-based sample pairs through contrastive learning, pulling similar samples closer while pushing dissimilar ones apart, enabling better recognition of key information differences and intrinsic relationships.

(2) Propose a contrastive learning and GCN-integrated CG-KSVM model for MD&A text classification. Building on CL-KSVM, it treats each sample as a graph node with concatenated BERT-GloVe features, constructing adjacency matrices using cosine similarity thresholds. GCN aggregate neighboring node information to update representations,

enhancing robustness. By introducing graph structures, the model effectively handles complex textual relationships and captures semantic associations and contextual information, further improving classification performance.

To validate the classification performance, we construct an experimental dataset using MD&A texts from four industry sectors within the SW Electronics segment of the WIND database. Through comprehensive comparisons with single-feature models, traditional machine learning models, and deep learning baselines, we systematically demonstrate the effectiveness of our proposed approach.

Yang Canqing (Mathematics)

Supervised by Zhang Yue

KEY WORDS: Contrast learning; Graph Convolutional Neural Networks; Kernel support vector machine; Management Discussion and Analysis; Text Categorization

目 录

第 1 章 绪论

1.1 研究背景与意义

1.1.1 研究背景

随着信息技术的快速发展和数据规模的急剧增长，文本分类技术在金融数据领域的应用日益广泛，尤其是在上市公司财务报告的分析中。上市公司财报中的管理层讨论与分析（Management Discussion and Analysis, MD&A）部分，作为公司经营状况的多维度展现，包含了公司经营状况、财务状况、未来战略规划及风险因素等多维度信息[1]。MD&A 文本不仅是公司过去一段时间内经营成果的总结，还展望了未来的发展战略和潜在风险，因此对其准确分类和分析，能够为评估公司经营状况、预测市场趋势提供强有力的依据[2]。

然而，随着信息多样性和异质数据多源化趋势的加剧，传统的文本分类方法在财务分析、风险预测及信用评级等领域已逐渐显现出局限性[3-4]。传统的文本分类方法通常依赖于手工特征提取和简单的机器学习模型，这些方法在处理复杂的、非结构化的文本数据时，往往难以捕捉到文本间的细微差异和复杂的关联性。特别是在 MD&A 文本中，由于具有高度的专业性和复杂性，涉及大量的专业术语、行业特定表达以及复杂的句法结构，传统方法的分类效果往往不尽如人意。因此，探索更为高效、精确的 MD&A 文本分类方法显得尤为重要。

1.1.2 研究意义

（1）理论意义

在理论层面，文本分类的研究推动了自然语言处理技术的不断进步。早期的文本分类方法主要依赖于手工设计的特征和浅层机器学习模型，如朴素贝叶斯、支持向量机等，这些方法虽然在一定程度上能够满足简单分类任务的需求，但在处理复杂文本数据时表现有限。近年来，随着深度学习技术的快速发展，基于神经网络的文本分类方法逐渐成为主流。卷积神经网络（CNN）和循环神经网络（RNN）等模型通过自动提取文本特征，显著提升了分类性能。特别是预训练

语言模型的引入，使得文本分类技术能够捕捉到更深层次的语义信息，进一步推动了该领域的发展。

然而，现有方法在处理图结构数据或需要捕捉样本间复杂关系时仍存在局限性，如何结合多种学习范式以提升分类效果，成为当前研究的重要方向。对比学习（Contrastive Learning, CL）作为一种新兴的表征学习方法，通过利用数据本身的结构信息，能够有效学习样本间的相似性或差异性[5]。对比学习的核心思想是通过构建正样本对和负样本对，使得模型在训练过程中能够区分相似和不相似的样本，从而学习到更具判别性的特征表示。这种方法在处理高维、复杂的文本数据时，表现出对信息间细微差异的敏感度，能够有效提升文本分类的准确性。

与此同时，图卷积神经网络（Graph Convolutional Neural Network, GCN）作为一种处理图结构数据的深度学习模型，凭借其卓越的学习和泛化能力，在多任务学习、大数据处理等复杂任务中展现出出色的效率和精确性[6]并广泛应用于自然语言处理、图像识别、语音识别等多个领域[7-9]。GCN 通过将图结构数据中的节点和边信息进行有效编码，能够捕捉到数据间复杂的关联性与依赖性。在文本分类任务中，GCN 可以将文本数据转化为图结构，通过节点表示文本，边表示文本间的关联关系，从而更好地捕捉文本间的语义关联和上下文信息。

将对比学习与图卷积神经网络相结合，可以充分发挥两者的优势，进一步提升文本分类的准确性[10]。对比学习能够通过构建正负样本对，学习到更具判别性的特征表示，而 GCN 则能够有效捕捉文本数据间的复杂关联性。这种结合不仅能够提升模型对文本间细微差异的敏感度，还能够增强模型对文本间复杂关系的理解能力，从而在 MD&A 文本分类任务中取得更好的效果。

（2）现实意义

在现实层面，文本分类技术的进步为多个领域提供了强有力的支持。以金融领域为例，管理层讨论与分析（MD&A）作为上市公司财报的核心部分，能够多维度展现公司的经营状况。对 MD&A 文本进行准确分类，不仅有助于评估公司的运营状况，还能够为投资者、监管机构等提供重要的决策依据。然而，随着信息多样性和异质数据多源化趋势的加剧，传统文本分类方法在财务分析及风险预测等领域已难以满足当前需求。

融合对比学习与图卷积神经网络能够显著提升 MD&A 文本信息提取效率，

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。同时，通过构建上市公司间的邻接矩阵，图卷积神经网络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。此外，该方法还能辅助风险预警与预测，通过对比学习及时发现潜在风险，为投资者提供风险参考。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。

1.2 国内外研究现状

文本分类是自然语言处理中的一项核心技术，旨在将文本文档自动归类到预定义的类别中。它广泛应用于情感分析、新闻分类、垃圾邮件过滤等领域。传统方法主要基于机器学习算法，如支持向量机（SVM）、朴素贝叶斯和 K 近邻（KNN），结合 TF-IDF 等特征提取技术。近年来，深度学习方法如卷积神经网络（CNN）、循环神经网络（RNN）以及注意力机制等显著提升了分类性能，尤其是在处理语义理解和长文本依赖关系方面。文本分类的研究不断推动着自然语言处理技术的发展，为信息检索、内容推荐等应用提供了重要支持。

近年来，基于机器学习和深度学习的文本分类模型得到了广泛研究，各类方法在不同数据集上展现了各自的优势。

（1）基于传统机器学习的文本分类研究

Joachims[11]在其文本分类研究中采用了线性支持向量机（SVM）结合词频-逆向文件频率（TF-IDF）的方法，并在 Reuters-21578 和 OHSUMED 数据集上进行了验证。实验结果显示，支持向量机在高维数据中表现出优异的泛化能力与稳定性，能够自动优化参数配置，从而显著提升文本分类模型的整体效果。Bijalwan 等人[12]对比了 Term-Graph、KNN 和朴素贝叶斯三种文本分类方法，基于 Reuters-21578 数据集证明 KNN 在分类准确性上显著优于其他两种方法。Jin 等人[13]提出了一种新型文本分类器，该分类器结合了嵌入袋（Bag of embeddings）模型和随机梯度下降（SGD）的方法。该模型通过最大化词嵌入向量的概率来预测单词所属的文本类别。实验结果表明，该方法在准确率和 F1 分数上均优于当前的主流方法。Chen 等人[14]利用 LDA 来提取主题信息，并将主题相关的特征

融入特征集中。实验结果表明，将 LDA 提取的主题信息与词频信息结合后，能够显著提升文本分类的准确性。Tellez 等人[15]提出了一种名为 μ TC 的文本分类模型，该模型不需要考虑文本领域与语言，能够更有效地管理文本分类任务。（2）基于深度学习的文本分类研究。Zheng 等人[16]提出了一个名为多变量相对歧视标准（MRDC）的新模型，用于文本分类中的特征提取。该模型通过最大化相关性和最小化冗余度来减少冗余特征。实验结果显示在多叉奈夫贝叶斯（MNB）的分类器中，该模型表现出最优的分类性能。Karim[17]等人结合 TF-IDF（词频-逆文档频率）和机器学习分类器的方法，用于判断引文对参考文献的极性。针对数据不平衡问题，采用 SMOTE 过采样技术和 uni-gram/bi-gram 特征进行实验，结果表明，分类器在 SMOTE 处理的 bi-gram 特征上表现最佳，准确率达 99%。Hussain[18]利用朴素贝叶斯（NB）算法，将从多个网站收集的信息划分为 10 个不同的类别，实验结果显示其分类准确率达到了 80%。Hu 等人[19]提出一种基于贝叶斯文本分类概率模型的监督加权方法，该方法通过匹配的分数函数为特征分配权重。实验结果显示，在大多数情况下，该方法优于其他的具有代表性加权方法。Tokgoz[20]对多种机器学习算法进行了对比研究，包括朴素贝叶斯、支持向量机、决策树以及 Rocchio 算法等，通过在 NewsGroups 数据集上的实验，他们发现支持向量机（SVM）在性能上表现最为突出。

（2）基于深度学习的文本分类研究

Liu 等人[21]提出了一种将 DBN 与 SVM 结合的文本分类方法。他们在包含 34 个类别的 7200 个中文文本数据集上进行了实验，取得了 76.66%的分类准确率。实验结果表明，这种混合方法在性能上优于单独使用 DBN 或 SVM 进行文本分类的效果。Wang 等人[22]提出了一种区域 CNN-LSTM 混合模型，用于维度情感分析。该模型将文本分割为句子级区域，通过区域 CNN 提取各句子的情感特征并加权，再通过 LSTM 跨区域整合序列依赖关系，从而同时捕捉局部情感信息和长距离上下文关联。实验表明，该方法在 VA 预测任务上优于基于词典、回归和传统神经网络的方法，实现了更细粒度的情感分析。Gu 等人[23]利用 CNN 对中文句子进行分类。他们在深度神经网络架构中嵌入了一个线性 SVM 以实现损失最小化，并采用 tanh 激活函数替代 ReLU，从而进一步提升模型性能。Kowsari 等人[24]提出了一种名为 RMDL 的深度学习方法，用于文本分类任务。他们认为，

RMDL 能够自动选择最优的深度学习架构，并通过集成多种深度学习模型来提升性能。实验表明，与其他现有方法相比，RMDL 结合并行学习架构的方式在分类准确率上达到了最高水平。Jiang 等人[25]提出了一种结合深度置信网络（DBN）和 softmax 回归的混合文本分类方法。深度置信网络主要用于处理文本数据相关的分类问题以及高维矩阵计算问题。在通过深度置信网络提取特征后，使用 softmax 回归对文本文档进行分类。他们将该方法与 SVM、KNN 等传统机器学习方法进行了对比实验，结果表明，该混合方法在性能上表现最优。Guo 等人[26]提出了一种基于 CNN-RNN 的混合神经网络，命名为 CRAN。该模型通过注意力机制将卷积神经网络（CNN）和递归神经网络（RNN）有效结合。实验在八个多标签文本分类数据集上进行，并与当前最先进的模型进行了对比。结果表明，CRAN 模型在大多数数据集上能够以更少的参数实现更优的性能表现。Qiu 等人[27]提出了一种基于词嵌入（Word Embedding）的模糊信息检索方法，结合深度学习（如 CBOW 模型）和模糊集理论，通过词向量衡量词语间的相关性（对称性），实验证明其在召回率、精确率和 F1 值上均优于传统检索方法。唐焕玲等人[28]提出了一种结合 LDA 与 word2vec 的方法，用于增强文本语义表示。该方法针对单词语义相似度计算和文本分类问题，在多个数据集上进行了实验验证。他们将所提出的方法与三种经典的文本分类算法进行了对比，结果显示，该模型在时间效率上表现更优的同时，文本分类准确率提升了 0.58%至 3.5%。Xu 等人[29]提出了一种基于 CNN-LSTM 混合模型的文本分类算法，该方法利用 word2vec 中的 Skip Gram 模型和 CBOW 模型将单词转换为向量表示。其中，CNN 用于捕捉局部特征，而 LSTM 则用于捕捉长距离依赖关系，从而有效利用上下文信息。CNN 输出的特征向量作为 LSTM 的输入，最终通过 softmax 分类器进行分类。实验在搜狗中文新闻文档集上进行，结果表明该算法能够显著提升文本分类的准确率。Li[30]等提出了一种基于词级图神经网络（TGNcl）的文本分类模型，通过构建单词依赖图并结合对比学习增强文本表征，显著优于现有方法，解决了传统 GNN 方法忽略词语间语义关联的问题。魏雯婕和张更平[31]提出了一种基于图神经网络（TextGCN）的专利文本自动分类方法，通过挖掘专利摘要的图结构特征（邻居信息与节点关系）生成文本表征向量，显著提升了专利授权预测的准确率、召回率和 F1 值，相比 Doc2Vec、TF-IDF 更具优势，为专利审查自动化提供了高

效解决方案。

1.3 论文创新点

（1）多模态特征融合：联合 BERT（上下文语义）与 GloVe（词级语义）特征，增强文本表示的全面性。

（2）对比学习优化：通过引入核支持向量机筛选高鉴别性样本，不仅有效解决了对比学习中时常遭遇的表示坍塌难题，还强化了模型对于不同类别间细微差异的感知能力，使模型能更敏锐地捕捉关键信息差异。

（3）图卷积神经网络的引入：将图卷积神经网络（GCN）引入文本分类领域，把文本数据转化为图结构数据后，借助 GCN 聚合邻接节点信息，深入挖掘文本间复杂的关联性，从而提升文本分类的精准度与鲁棒性。

1.4 论文结构

第 1 章为绪论，主要阐述了研究背景与意义、国内外研究现状以及论文结构。

第 2 章为准备工作，主要介绍实验环境配置以及数据来源。

第 3 章提出 CL-KSVM 模型，结合 BERT 提取文本特征，利用 KSVM 筛选高鉴别性样本，并通过对比学习机制增强类间差异感知能力。通过实验验证模型的可行性。

第 4 章为融合对比学习与 GCN 的 CG-KSVM 文本分类模型，在第 3 章模型的基础上，进一步引入图卷积神经网络，利用图结构数据挖掘文本间的复杂语义关联，提升 MD&A 文本分类性能，以增强 MD&A 文本分类的性能。通过对比实验验证模型的性能提升。

第 5 章为总结与展望，探讨了未来可能的研究方向和改进空间。

第 2 章 准备工作

2.1 实验环境

本文的实验是在 Python3.9 上实现的，实验环境如表 2-1 所示：

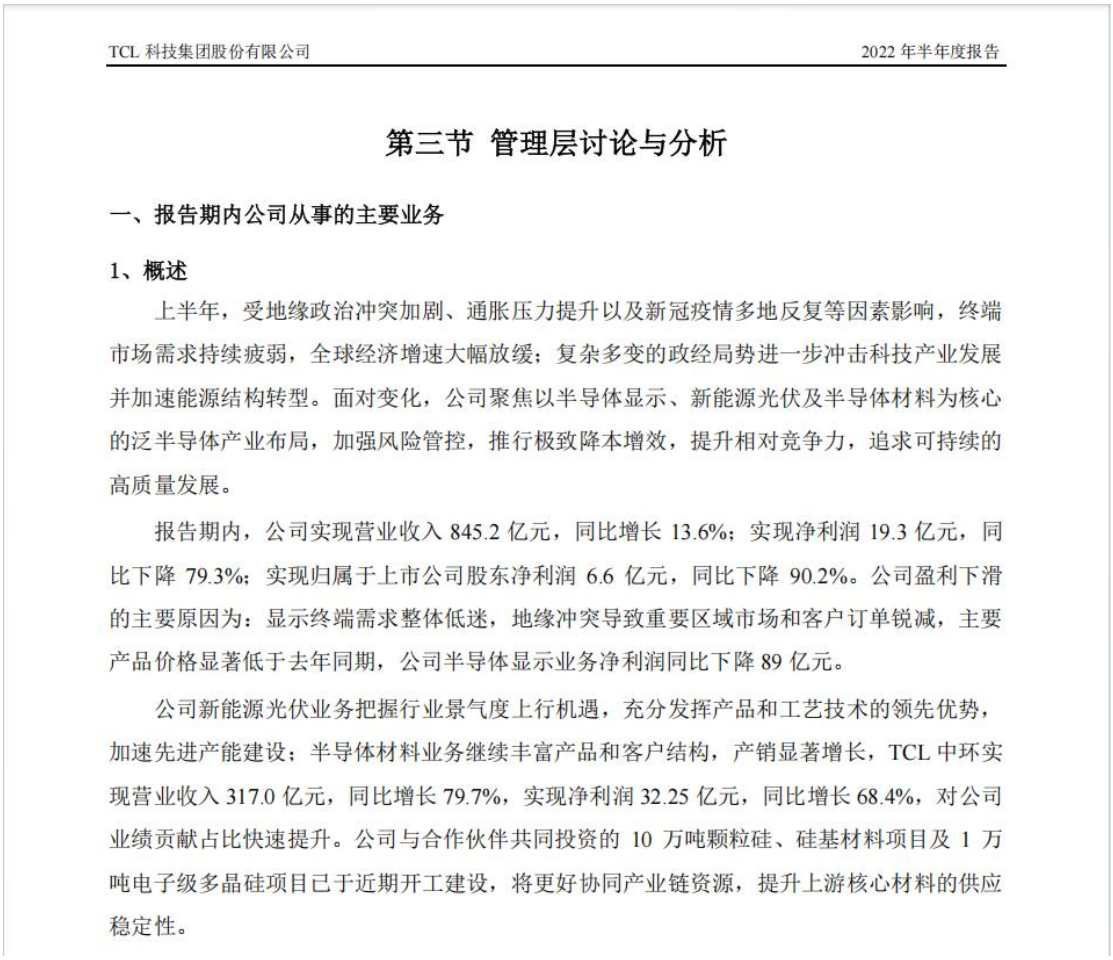
表 2-1 实验环境参数说明

名称	参数
操作系统	win11 64 位
处理器	英特尔第十二代酷睿 i5-12600KF 十核
CPU	12th Gen Intel (R) Core (TM) i5-12600KF 3.70 GHz
GPU	NVIDIA GeForce RTX4060(8G)
内存	32GB DDR4 3200MHZ
开发环境	Pycharm Commonity Edition
开发语言	Python 3.9

2.2 数据来源

电子信息制造业作为国民经济的战略性、基础性、先导性产业，是稳定工业经济增长、维护国家政治经济安全的重要领域[32-33]。由此，我们获取了 WIND 数据库中申银万国行业类的 SW 电子中的子数据库，其包含了 SW 半导体、SW 元件、SW 光学光电子和 SW 消费电子四个板块的上市公司定期报告数据。这些数据为我们提供了丰富的 MD&A 信息，为模型的测试提供了现实支撑。

选取 SW 电子中四个行业板块的上市公司 MD&A 文本数据进行 MD&A 文本分类。SW 半导体板块拥有 158 家公司含 2021-2023 的年报以及半年报共 704 份，SW 元件板块拥有 59 家公司含 2021-2023 的年报以及半年报共 314 份，SW 光学光电子板块拥有 97 家公司含 2021-2023 的年报以及半年报共 531 份，SW 消费电子板块拥有 95 家公司含 2021-2023 的年报以及半年报共 496 份。在实验中，我们将四个板块 MD&A 文本数据都按照 3：1 划分训练集与测试集。 MD&A 原始文本数据及整理如图 2-1、2-2 所示。



TCL 科技集团股份有限公司 2022 年半年度报告

第三节 管理层讨论与分析

一、报告期内公司从事的主要业务

1、概述

上半年，受地缘政治冲突加剧、通胀压力提升以及新冠疫情多地反复等因素影响，终端市场需求持续疲弱，全球经济增速大幅放缓；复杂多变的政经局势进一步冲击科技产业发展并加速能源结构转型。面对变化，公司聚焦以半导体显示、新能源光伏及半导体材料为核心的泛半导体产业布局，加强风险管控，推行极致降本增效，提升相对竞争力，追求可持续的高质量发展。

报告期内，公司实现营业收入 845.2 亿元，同比增长 13.6%；实现净利润 19.3 亿元，同比下降 79.3%；实现归属于上市公司股东净利润 6.6 亿元，同比下降 90.2%。公司盈利下滑的主要原因为：显示终端需求整体低迷，地缘冲突导致重要区域市场和客户订单锐减，主要产品价格显著低于去年同期，公司半导体显示业务净利润同比下降 89 亿元。

公司新能源光伏业务把握行业景气度上行机遇，充分发挥产品和工艺技术的领先优势，加速先进产能建设；半导体材料业务继续丰富产品和客户结构，产销显著增长，TCL 中环实现营业收入 317.0 亿元，同比增长 79.7%，实现净利润 32.25 亿元，同比增长 68.4%，对公司业绩贡献占比快速提升。公司与合作伙伴共同投资的 10 万吨颗粒硅、硅基材料项目及 1 万吨电子级多晶硅项目已于近期开工建设，将更好协同产业链资源，提升上游核心材料的供应稳定性。

图 2-1 MD&A 原始文本数据

	A	B	C	D	E	F	G	H
1	序号	证券代码	证券简称	统计截止日期	内容	标签	净资产收益率	经营等级
2	1	000020	深华发A	2023-12-31	一、报告期内公司所处行业情况 报告期内，虽然有宏观经济的不利影响，叠加注塑和视讯行业竞争	SW光学电子	0.036363	C
3	2	000020	深华发A	2022-12-31	一、报告期内公司所处行业情况报告期内，在公司领导层的正确带领下，在全体员工的齐心努力下，	SW光学电子	0.028644	C
4	3	000020	深华发A	2022-06-30	一、报告期内公司从事的主要业务公司经过多年发展，目前已逐步形成工业业务与物业经营业务两	SW光学电子	0.025001	C
5	4	000020	深华发A	2023-06-30	一、报告期内公司从事的主要业务公司经过多年发展，目前已逐步形成工业业务与物业经营业务两	SW光学电子	0.021793	C
6	5	000020	深华发A	2021-12-31	一、报告期内公司所处的行业情况随着人工智能时代的到来，智能化成为家电业发展的一大趋势，	SW光学电子	0.020969	C
7	6	000020	深华发A	2021-06-30	一、报告期内公司从事的主要业务 公司经过多年发展，目前已逐步形成工业业务与物业经营业务	SW光学电子	0.020363	C
8	7	000021	深科技	2021-12-31	一、报告期内公司所处的行业情况 1. 存储半导体行业 集成电路产业作为整个半导体产业的核	SW消费电子	0.072692	B
9	8	000021	深科技	2023-12-31	一、报告期内公司所处的行业情况1. 存储半导体行业集成电路产业作为整个半导体产业的核心，可	SW消费电子	0.064501	B
10	9	000021	深科技	2022-12-31	一、报告期内公司所处的行业情况1. 存储半导体行业 集成电路产业作为整个半导体产业的核心，	SW消费电子	0.057579	B
11	10	000021	深科技	2022-06-30	一、报告期内公司从事的主要业务公司是全球领先的专业电子制造企业，连续多年在MMI全球EMS行	SW消费电子	0.039145	C
12	11	000021	深科技	2021-06-30	一、报告期内公司从事的主要业务(一) 报告期内公司所从事的主要业务公司是全球领先的专业电子	SW消费电子	0.030563	C
13	12	000021	深科技	2023-06-30	一、报告期内公司从事的主要业务及行业情况公司所从事的主要业务公司是全球领先的专业电子制	SW消费电子	0.030254	C
14	13	000045	深纺织A	2023-12-31	一、报告期内公司所处行业情况 偏光片全称为偏振光片，可控制特定光束的偏振方向，自然光在通	SW光学电子	0.030919	C
15	14	000045	深纺织A	2021-06-30	一、报告期内公司从事的主要业务（一）公司从事的主要业务公司主要业务是以液晶显示用偏光片	SW光学电子	0.028406	C
16	15	000045	深纺织A	2022-12-31	一、报告期内公司所处行业情况 偏光片全称为偏振光片，可控制特定光束的偏振方向，自然光在通	SW光学电子	0.027733	C
17	16	000045	深纺织A	2021-12-31	一、报告期内公司所处的行业情况偏光片全称为偏振光片，可控制特定光束的偏振方向，自然光在	SW光学电子	0.01895	C
18	17	000045	深纺织A	2022-06-30	一、报告期内公司从事的主要业务（一）公司从事的主要业务公司主要业务是以液晶显示用偏光片	SW光学电子	0.017488	C
19	18	000045	深纺织A	2023-06-30	一、报告期内公司从事的主要业务（一）公司所属行业发展情况偏光片全称为偏振光片，可控制特	SW光学电子	0.012875	C
20	19	000050	深天马A	2021-12-31	一、报告期内公司所处的行业情况近年来，随着中小尺寸显示技术的发展，市场已呈现 a-Si、LTPS	SW光学电子	0.044231	B
21	20	000050	深天马A	2021-06-30	一、报告期内公司从事的主要业务报告期内，公司持续深耕中小尺寸显示领域，聚焦以智能手机、	SW光学电子	0.034436	C
22	21	000050	深天马A	2022-06-30	一、报告期内公司从事的主要业务报告期内，公司持续聚焦中小尺寸显示业务，将智能手机、车载	SW光学电子	0.012225	C
23	22	000050	深天马A	2022-12-31	一、报告期内公司所处行业情况近年来，随着中小尺寸显示技术的发展，市场已呈现 a-Si、LTPS、	SW光学电子	0.003562	D
24	23	000050	深天马A	2023-06-30	一、报告期内公司从事的主要业务天马微电子股份有限公司是一家在全球范围内提供全方位的客制	SW光学电子	-0.048749	D
25	24	000050	深天马A	2023-12-31	一、报告期内公司所处行业情况近年来，随着中小尺寸显示技术的发展，市场已呈现 a-Si、LTPS、	SW光学电子	-0.072772	D
26	25	000100	TCL科技	2021-12-31	一、报告期内公司所处的行业情况2021年，国际形势愈加复杂，区域冲突加剧，新冠韩炎度情特	SW光学电子	0.125028	A

图 2-2 MD&A 文本整理数据

净资产收益率（Return on Equity, ROE）是评估公司盈利能力的关键指标之一。 ROE 值越高，表明公司的盈利能力和资本增值潜力越强；反之，则说明其盈利能力和资本增值潜力较弱。因此，在本文的模型中，我们选择净资产收益率（ROE）作为核心财务指标，用以衡量和评估上市公司的整体经营状况。其计算公式为：

$$ROE = \text{净利润} / \text{净资产} \times 100\%$$

基于 ROE 的数值，我们对上市公司的经营状况进行了如下等级划分：

表 2-2 经营等级的划分

经营等级	ROE 范围
A	ROE>3/4 分位数
B	2/4 分位数 ＜ROE＜3/4 分位数
C	1/4 分位数 ＜ROE＜2/4 分位数
D	ROE＜1/4 分位数

当 ROE>3/4 分位数时，代表公司经营状况极佳，将其标记为 A 标签。

当 2/4 分位数 ＜ROE＜3/4 分位数时，表明公司经营状况良好，将其归类为 B 标签。

当 1/4 分位数 ＜ROE＜2/4 分位数时，意味着公司经营状况一般，将其标记为 C 标签。

而当 ROE＜1/4 分位数时，表明公司经营状况较差，将其标记为 D 标签。

这样的划分旨在通过 ROE 的量化标准，直观反映上市公司的经营优劣层次。

第 3 章 基于对比学习的 CL-KSVM 文本分类

3.1 模型框架

MD&A 文本通过对财务数据的文字解读，揭示公司经营中的风险与不确定性，增强了信息披露的有效性，为投资者提供前瞻性信息[34-35]。CL-KSVM 文本分类模型采用 BERT 预训练语言模型模型，精准捕捉 MD&A 文本上下文的深层含义和隐含关系。引入对比学习机制，通过比较不同样本之间的相似性和差异性来学习数据有效表示。在对比学习中，利用 KSVM 算法，根据样本的鉴别性信息，筛选出对 MD&A 文本分类具有高度代表性的样本，从而精简数据集并保留其核心特征。随后，将每个支持向量的 GloVe 词嵌入表示作为其正样本，并通过对比学习机制训练对应的负样本，进一步增强模型对全局信息的把握能力。CL-KSVM 文本分类模型的具体构建流程如图 3-1 所示。

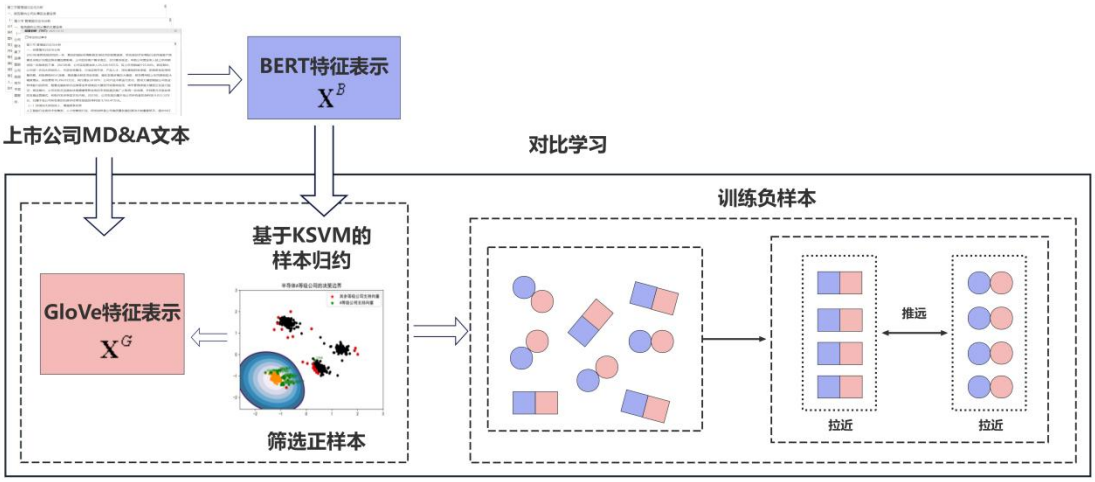


图 3-1 CL-KSVM 模型构建流程图

3.2 理论基础

3.2.1 BERT 模型

BERT（Bidirectional Encoder Representations from Transformers）是一种基于 Transformer 架构的预训练语言模型[36]。该模型通过双向编码器捕捉上下文信息，能够根据上下文动态调整词语的向量表示，从而显著提升自然语言处理任务的性能。其模型结构如图 3-2 所示。

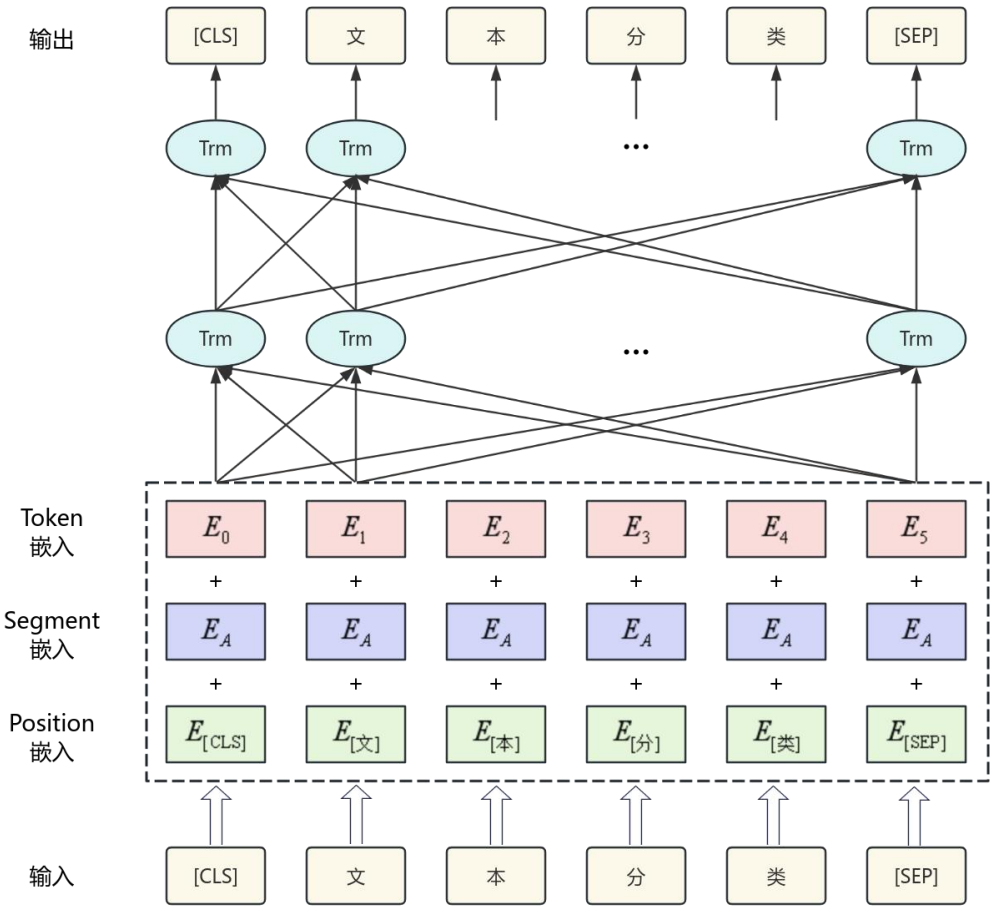


图 3-2 BERT 模型结构

在将输入传递到 BERT 之前，需要嵌入一些特殊的标记，例如[CLS]和[SEP]。[CLS]标记位于每个序列的开头，用于分类任务，而[SEP]标记用于区分句子对中的不同句子。输入表示通过对应的标记、段和位置嵌入求和来构造，随后这些标记在层叠的 Transformer 编码器中流动。

BERT 的架构与 Transformer 保持一致，但其采用了掩码语言模型（Masked Language Model, MLM）作为核心训练方法。与生成式预训练模型（GPT）和词向量表征模型（ELMO）相比，BERT 具有显著的优势。GPT 是单向语言模型，

仅利用上文信息进行预测，而 ELMO 虽然采用双向 LSTM 编码，但其双向信息是通过两个独立的 LSTM 分别获取的。相比之下，BERT 通过 Transformer 编码器和 MLM 实现了真正的双向编码，能够在语义特征上同时捕捉上下文信息，从而生成更具上下文感知能力的特征嵌入，显著提升了自然语言理解任务的性能。

3.2.2 GloVe 模型

GloVe 模型（Global Vectors for Word Representation）是一种用于生成词向量的无监督学习算法[37]。GloVe 模型的核心思想是，词语的含义可以通过其在不同上下文中的共现概率来表征，而这种共现信息可以通过对大规模语料库的统计分析来获得。GloVe 模型训练过程如下所示：

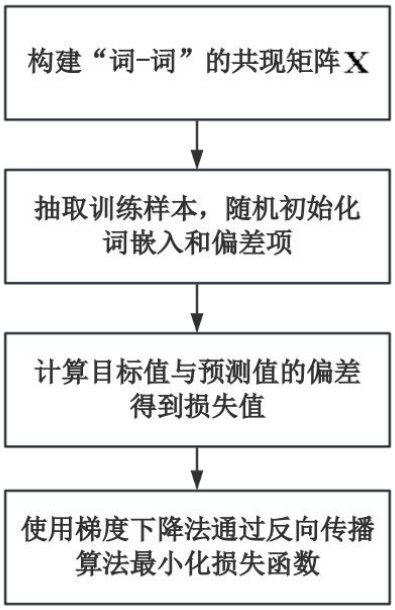


图 3-3 GloVe 模型训练过程

设定一个固定的窗口大小，在语料库中从左向右滑动，位于窗口中心的词称为中心词，窗口内位于中心词两侧的词称为上下文。共现矩阵 $\mathbf{X}$ 中 x_{ij} 表示单词 i 是中心词时，单词 j 出现在特定上下文窗口内共现的次数；$X_i = \sum_k X_{i,k}$ 表示单词 i 是中心词的总次数；$P_{ij} = P(j \mid i) = \dfrac{X_{ij}}{X_i}$ 表示单词 i 作为中心词时，单词 j 出现

的概率。构建损失函数J，使计算词向量比值的函数$F\left(w_i,w_j,w_k\right)=\frac{P_{ik}}{P_{jk}}$成立，其中$w_i$、$w_j$、$w_k$分别是单词$i$、$j$、$k$的词向量。

GloVe 模型的目标是通过学习词向量来拟合这个共现矩阵。具体来说，GloVe 模型试图找到一组词向量，使得这些向量的点积能够近似表示词语之间的共现概率的对数。为了实现这一目标，GloVe 模型定义了一个损失函数，该函数衡量了模型预测的共现概率与实际的共现概率之间的差异。通过最小化这个损失函数，GloVe 模型可以学习到能够捕捉词语之间语义关系的词向量。具体来说，GloVe 模型使用了一个加权最小二乘法的形式，其中每个共现对的权重取决于其共现次数的函数。这种设计使得模型能够更好地处理低频词和高频词之间的不平衡问题，从而提高词向量的质量。最终优化的损失函数如式 3.1 所示。

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

其中w_i、$\tilde{w}_j$分别表示单词i、j的词向量，b_i、$\tilde{b}_j$分别对应词向量w_i、$\tilde{w}_j$的偏置，$\log\left(X_{ij}\right)$是单词i、j组成词对的共现次数的对数，N表示文档集内词典的单词数。$f\left(X_{ij}\right)$是权重函数，X_{ij}用来控制不同大小的共现次数对结果的影响。该损失函数不仅考虑了共现矩阵中的非零元素，还引入了一个加权项，用于处理共现矩阵中的稀疏性问题。

GloVe 模型的训练过程通常采用随机梯度下降法（SGD）来优化损失函数。在每次迭代中，模型会随机选择一个共现对，并更新相应的词向量，以减小损失函数的值。通过多次迭代，模型逐渐收敛到一组能够较好地拟合共现矩阵的词向量。由于 GloVe 模型的训练过程是基于全局的共现矩阵，因此它能够捕捉到词语之间的全局统计信息，从而生成更具语义表达能力的词向量。

GloVe 模型的优点之一是其生成的词向量具有良好的线性特性。这意味着，通过简单的向量运算（如加法和减法），可以在词向量空间中捕捉到词语之间的语义关系。

3.2.3 核支持向量机模型

支持向量机（Support Vector Machine, SVM）是一种二分类模型[38]，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。

核支持向量机（Kernel Support Vector Machine, KSVM）通过引入核函数（Kernel Function），能够有效处理非线性分类问题。其核心思想是将原始空间中的数据映射到一个高维特征空间，使得在高维空间中数据变得线性可分，从而在该空间中构建最优分类超平面。这种方法克服了传统支持向量机在处理非线性数据时的局限性，显著提升了分类性能。

对于非线性可分的数据，支持向量机通过引入核函数将数据从原始空间映射到一个高维特征空间，使得在高维空间中数据线性可分。核函数的定义如下：

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j),$$

其中，$\phi(x_i) \cdot \phi(x_j)$ 是映射函数。

径向基核函数（Radial Basis Function, RBF Kernel）的表达式为：

$$K(x_i, x_j) = \exp(-\frac{\| x_i - x_j \|^2}{2\sigma^2}),$$

其中$\| x_i - x_j \|$表示x_i与x_j欧几里得距离。σ是一个正实数，用于控制样本点映射到高维空间的“宽度”。

在高维特征空间中，核支持向量机的优化问题可以表示为：

$$\min \frac{1}{2}\|\omega\|^2,$$

约束条件为：

$$y_i(\omega \cdot \phi(x_i) + b) \geq 1.$$

通过引入拉格朗日乘子，将优化问题转化为对偶问题：

$$\max_{\alpha}[\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j)\,,$$

约束条件为：

$$\sum_{i=1}^{n}\alpha_i y_i=0\,.$$

求解后决策函数可以表示为：

$$f(x)=sign\left(\sum_{i=1}^{n}\alpha_i y_i K\left(x_i,x\right)+b\right),$$

其中，α_i 是拉格朗日乘子，b 是偏置项。

3.2.4 对比学习

对比学习（Contrastive Learning, CL）是表征学习的重要分支[39]，专注于利用数据内在结构信息来学习样本间的相似性或差异性。其核心思想是通过构建正负样本对，使模型将相似样本在特征空间中拉近，同时将不相似样本推开，从而有效捕捉数据的本质特征。

对比学习中的正负样本是模型训练的核心概念。正样本通常是指来自同一数据实例的不同增强版本，例如同一张图像经过裁剪、旋转或颜色抖动后生成的不同视图。这些样本在语义上高度相似，模型的目标是使它们在表示空间中尽可能接近。负样本则是指来自不同数据实例的样本，例如不同图像的增强视图。负样本在语义上与正样本不相关，模型的目标是使它们在表示空间中远离。对比学习的核心在于损失函数的设计。最常用的对比损失函数是 InfoNCE（Noise Contrastive Estimation）损失，它鼓励模型将正样本对的相似度提高，同时将负样本对的相似度降低。其目标函数为：

$$L_{(u,v^+,\{v^-\})}=-\log\left(\frac{\exp\left(sim\left(u\cdot v^+/\tau\right)\right)}{\sum_{v\in\{v^+,v^-\}}\exp\left(sim\left(u\cdot v^-/\tau\right)\right)}\right),\qquad(3.1)$$

其中 u 为原样本，v^+ 为 u 的正样本，v^- 为 u 的负样本。τ 是温度超参数，τ 越大，分布中的数值越小，分布就会变得更平滑，更接近均匀分布；τ 越小，分布更集中，模型更关注那些特别困难的负样本。通过优化这一损失函数，模型能够学习

到将语义相似的样本映射到特征空间中相近的位置，而将语义不同的样本映射到远离的位置。

对比学习的优势在于其能够利用大量未标注数据进行训练，从而减少对标注数据的依赖。此外，对比学习能够学习到具有良好泛化能力的特征表示，这些表示可以迁移到下游任务中，如分类、检测和分割等。例如，在计算机视觉领域，SimCLR 和 MoCo 等对比学习框架在 ImageNet 数据集上取得了与有监督学习相媲美的性能。在自然语言处理领域，对比学习也被广泛应用于句子表示学习和预训练语言模型中，如 SimCSE 和 CLIP 等模型。

3.3 CL-KSVM 文本分类模型

3.3.1 BERT 文本表示

在对比学习中，正负样本是成对出现的[40]。正样本是指与目标样本相似或属于同类别的样本；负样本是指与目标样本不相似或不属于同类别的样本 42。采用 BERT 预训练语言模型对 MD&A 进行深入理解和挖掘，将 MD&A 文本转化为高维向量表示。一个 MD&A 的 d_b 维实向量 BERT 文本表示为 X^B，即：

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

因此，对于 N 个 MD&A 文本，就可以得到一个 $N \times d_b$ 维的矩阵表示：

$$\mathbf{X}^B = \begin{bmatrix} X_1^B \\ X_2^B \\ \vdots \\ X_N^B \end{bmatrix},$$

其中 X_i^B 为第 i 个 MD&A 文本基于 BERT 的特征表示。

3.3.2 GloVe 特征提取

在提取 MD&A 文本的 GloVe 特征时，会遇到大量如“公司”、“经营”等高频但对 MD&A 文本分类意义不大的词汇，这些冗余信息可能降低模型分析的准确性和效率。为此，我们自建停用词典，通过分词和去停用词步骤减少冗余信

息，旨在最大限度地减少术语及无意义词汇的干扰，进而提升 MD&A 文本分类的精确度和效率。

在此基础上，采用 GloVe 模型获取一个 MD&A 文本的 d_g 维实向量 GloVe 文本表示 X^G，即：

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

因此，对于 N 个 MD&A 文本，就可以得到一个 $N \times d_g$ 维的矩阵表示：

$$\mathbf{X}^G = \begin{bmatrix} X_1^G \\ X_2^G \\ \vdots \\ X_N^G \end{bmatrix},$$

其中 X_i^G 为第 i 个 MD&A 文本基于 GloVe 的特征表示。

3.3.3 基于 KSVM 的样本归约

在获得 MD&A 的 BERT 文本表示后，我们利用 KSVM 进行样本归约，旨在维持数据原有特征和信息量的基础上，最大限度地减少样本数量，从而实现数据的精简与高效处理。传统 SVM 主要作为二分类器的问题，在处理 MD&A 文本分类多分类问题时，我们可以采取“一对多”的策略。

具体地，我们针对一个行业板块，将公司的经营等级作为分类标签。在每一次分类中，我们选择其中一个类标签作为正类，而将剩余的类标签合并为一个负类，从而将问题简化为二分类问题。在此框架下，采用基于 RBF 的支持向量分类算法，将样本点映射到高维空间，使得原本非线性可分的数据在高维空间中变得线性可分。通过最大化超平面到最近样本点的间隔，进而找到一个决策边界来区分不同类别的样本点，以精准分隔正类与负类，并挖掘特征空间中各个经营等级分布边界上的高置信度支持向量，提取出对 MD&A 文本分类最具影响力的样本。

决策边界的表达式为

$$\omega \cdot X^B + b = 0,$$

ω 表示决策边界的法向量，b 表示决策边界的截距。则支持向量是那些满足

$$\omega \cdot X^{B} + b = \pm 1,$$

的样本点，则从 N 个 MD&A 文本中学习到的 M^{+} 个支持向量可以表示为

$$\hat{\mathbf{X}}^{B} = \begin{bmatrix} \hat{X}_1^{B} \\ \hat{X}_2^{B} \\ \vdots \\ \hat{X}_{M^{+}}^{B} \end{bmatrix},$$

其中 $\hat{X}_i^{B}$ 为第 i 个支持向量基于 BERT 的特征表示。

3.3.4 正负样本对的筛选

在对比学习中，正样本指与目标样本同类或相似的样本；负样本指与目标样本不相似或不同类的样本。针对每一个支持向量的 BERT 文本表示，我们将其对应的 GloVe 特征向量，视为其正样本，即

$$\hat{\mathbf{X}}^{G} = \begin{bmatrix} \hat{X}_1^{G} \\ \hat{X}_2^{G} \\ \vdots \\ \hat{X}_{M^{+}}^{G} \end{bmatrix},$$

其中 $\hat{X}_i^{G}$ 为第 i 个支持向量对应的正样本。

由（3.1），对比损失函数的数学表达式为：

$$L_{CL} = -\log\left(\frac{\exp(sim(\hat{X}_i^{B}, \hat{X}_i^{G} / \tau))}{\exp(sim(\hat{X}_i^{B}, \hat{X}_i^{G})) / \tau + \sum_{j=1}^{M} \exp(sim(\hat{X}_i^{B}, \tilde{X}_j^{G})) / \tau} \right),$$

其中，$\hat{X}_i^{B}$ 是第 i 个支持向量的 BERT 特征表示，$\hat{X}_i^{G}$ 是其正样本，$\tilde{X}_j^{G}$ 为负样本，$sim(\hat{X}_i^{B}, \tilde{X}_j^{G})$ 表示样本 $\hat{X}_i^{B}$ 与其负样本 $\tilde{X}_j^{G}$ 之间的余弦相似性度量，且 $i \neq j$。τ 为对比学习 softmax 的温度超参数。

通过最小化损失函数，模型能精准捕捉相似样本的共同特性，有效区分不同类别样本间的差异。从原始的 N 个样本中筛选出与支持向量类别相反的 GloVe 文本表示作为负样本，从而构建基于 GloVe 特征表示的正负样本对。则负样本的

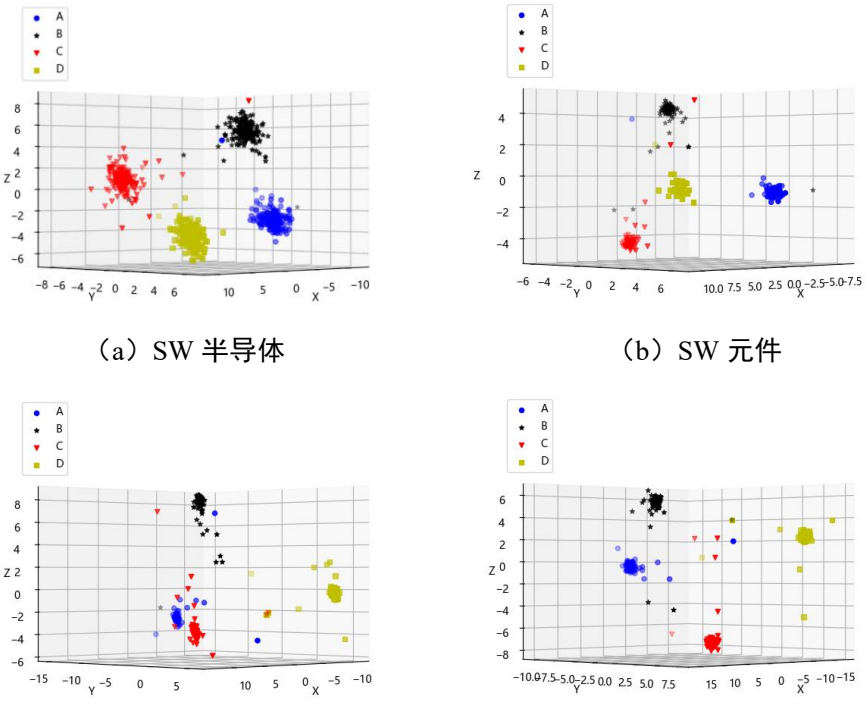
GloVe 特征表示为：

$$\tilde{\mathbf{X}}^{G}=\begin{bmatrix}\tilde{X}_{1}^{G}\\ \tilde{X}_{2}^{G}\\ \vdots\\ \tilde{X}_{M^{-}}^{G}\end{bmatrix}.$$

3.4 实验设计与结果分析

3.4.1 特征提取

CL-KSVM 模型选用 Chinese-BERT-wwm-ext 预训练模型作为基础框架，以挖掘 SW 半导体、SW 元件、SW 光学光电子以及 SW 消费电子板块 MD&A 文本的深层次语义特征。将每个 MD&A 文本编码为一个 768 维的 BERT 向量，其三维可视化如图 3-4 所示。



（a）SW 半导体　　（b）SW 元件

（c）SW 光学光电子　　（d）SW 消费电子

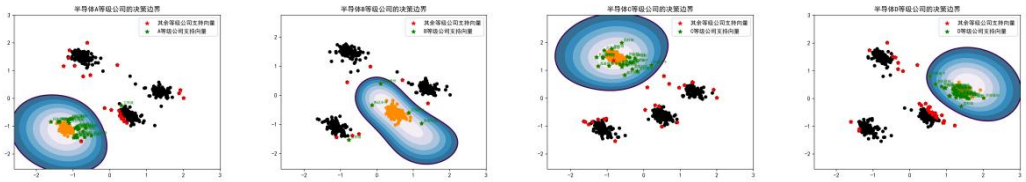
图 3-4 BERT 表示可视化

图 3-4 中（a）-（d）分别展示了这四个行业板块 MD&A 文本的聚类结果。从图中可以看出，每个板块的样本点根据其类标签形成了明显的聚类区域，不同板块之间的样本点分布相对分离，且每个板块内部的样本点也呈现出较好的聚合

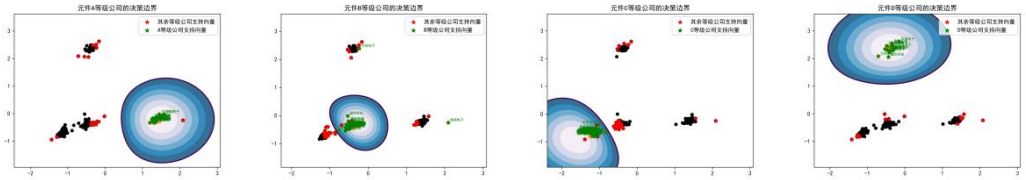
性。这充分证明了 BERT 模型在提取文本语义特征方面的有效性，并能够有效区分不同类别 MD&A 文本的特征。

3.4.2 支持向量的筛选

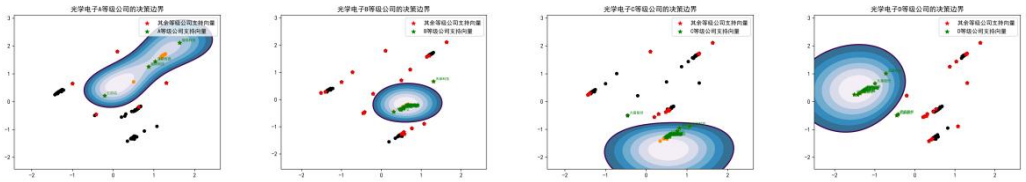
样本集合中往往蕴含着冗余信息，这对分类任务的准确性构成了挑战。特别是在对比学习的框架下，正样本的精准识别显得尤为重要。为此，我们引入了支持向量机作为筛选机制，旨在从原始数据中提炼出具备高度鉴别力的信息样本，同时明确界定对比学习中各类标签所对应的正样本集合。



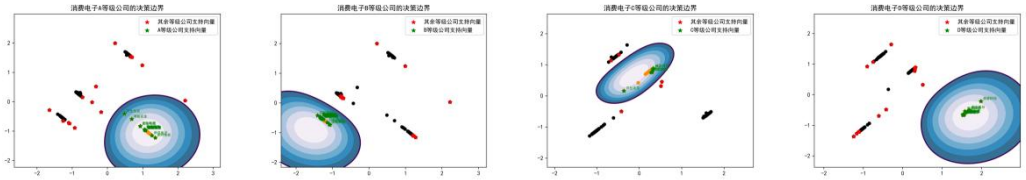
（a）SW 半导体



（b）SW 元件



（c）SW 光学光电子



（d）SW 消费电子

图 3-5 支持向量可视化

针对四行业板块，我们以公司的经营等级作为分类标签，以 RBF 为核构造的 KSVM，以此确定决策边界及其关键的支持向量。图 3-5 中的（a）-（d）分别展示了 SW 半导体、SW 元件、SW 光学光电子以及 SW 消费电子四个板块在

四个经营等级下的决策边界及其支持向量分布情况。其中，绿色标记代表目标类别的支持向量，而红色标记则代表其余类别的支持向量。

在采用 KSVM 方法进行支持向量筛选后，SW 半导体、SW 元件、SW 光学光电子以及 SW 消费电子四个板块的类标签样本的变化情况，及具有代表性的支持向量实例如表 3-1 所示。其中，士兰微作为国内领先的半导体设计与制造（IDM）企业之一，于 2021 年上半年实现营收利润大幅增长，净利暴增 13 倍；元件行业的中京电子 2021 年营业收入有所增长，净利润同比减少 8.85%；液晶薄膜材料龙头的八亿时空在 2023 上半年营收与净利润均有所下滑；以电子设备制造为核心业务领域的高新技术企业惠威科技在 2022 年上半年的营收与净利润分别同比下降 24.00%与 96.87%[1]。上述实例不仅有力地验证了本研究在支持向量选择策略上的有效性与合理性，而且所提取的支持向量在揭示企业经营状况方面展现出了极高的鉴别力与代表性。

表 3-1 部分正样本展示

		SW 半导体	SW 元件	SW 光学光电子	SW 消费电子
	样本变化	128→27	74→23	70→22	110→22
		士兰微	顺络电子	森霸传感	贝仕达克
A	正样本代表	**2021 半年报**	2022 年报	2021 年报	2021 年报
		长川科技	骏亚科技	蓝黛科技	朗特智能
		2022 年报	2021 半年报	2022 年报	2021 年报
	样本变化	139→49	75→18	80→35	87→34
		北京君正	**中京电子**	水晶光电	拓邦股份
B	正样本代表	2023 年报	**2021 年报**	2023 年报	2023 半年报
		帝奥微	铜峰电子	洲明科技	星星科技
		2022 年报	2023 年报	2023 半年报	2022 年报
	样本变化	137→23	51→16	109→37	85→51
		晶方科技	迅捷兴	**八亿时空**	智新电子
C	正样本代表	2023 半年报	2022 半年报	**2023 半年报**	2023 年报
		圣邦股份	南亚新材	爱克股份	深科技
		2023 半年报	2022 半年报	2023 半年报	2021 半年报
	样本变化	123→35	34→12	138→21	88→38
		华微电子	晶赛科技	爱克股份	**惠威科技**
D	正样本代表	2023 半年报	2023 年报	2022 半年报	**2022 半年报**
		全志科技	生益电子	汇创达	春秋电子
		2023 半年报	2023 年报	2023 半年报	2023 半年报

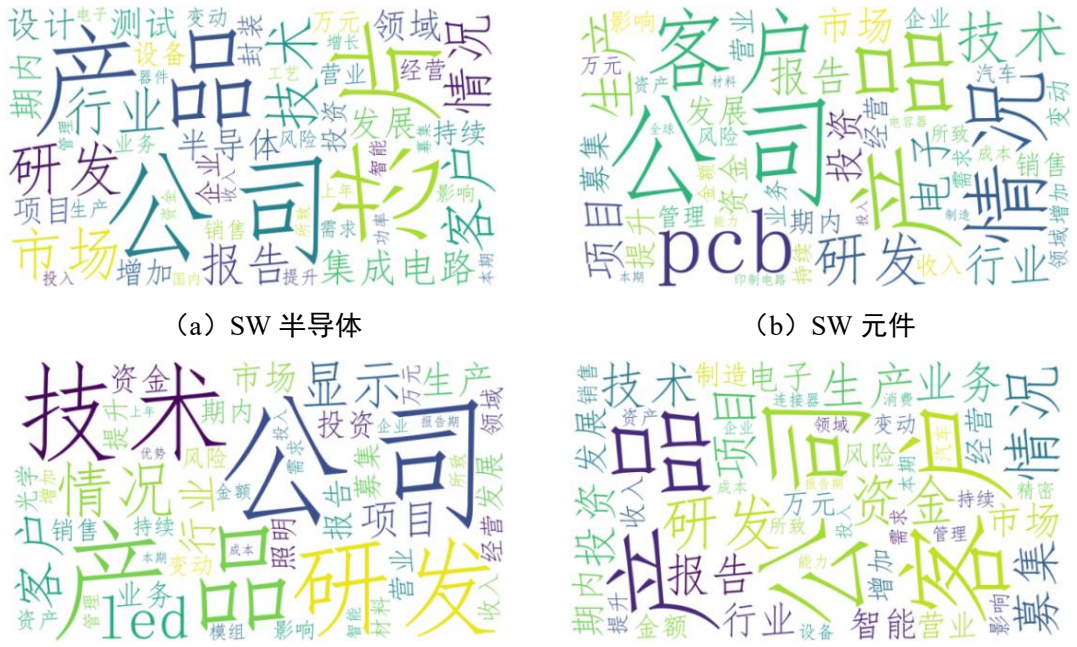
[1] 来自新浪财经

3.4.3 基于 GloVe 的对比学习

针对通过 KSVM 筛选出的支持向量，我们分别提取每个 MD&A 文本的 GloVe 特征向量，来训练正负样本对。

词云图，又称文字云，是文本数据可视化的一种创新形式，其深厚的理论基础主要根植于信息可视化和文本挖掘两大领域。在词云图中，高频词汇以大号字体和深色显示，而低频词汇则以小号字体和浅色呈现，这种设计让人能够迅速聚焦于文本的核心内容，把握关键信息。

MD&A 文本含电子、半导体等专业术语，这些术语在特征表示过程中往往会产生干扰，从而影响对公司经营状况的准确评估。因此，在获取 GloVe 特征向量前，我们借助 LDA（Latent Dirichlet Allocation）主题模型，分别提取了 SW 半导体、SW 元件、SW 光学光电子以及 SW 消费电子四个板块前 100 个高频词，并以词云图的形式直观呈现，如图 3-6（a）-（d）。



（a）SW 半导体　（b）SW 元件

（c）SW 光学光电子　（d）SW 消费电子

图 3-6 词云图

为此，我们构建停用词典，涵盖了高频但对本研究无实际意义的行业术语及无意义词汇，以提升文本特征表示的准确性，确保对公司经营状况的正确评估。

在获得在对 MD&A 文本进行分词、去停用词后，我们对四个不同板块以分类准确率为标准，探究 GloVe 特征向量的最优维度。我们利用线性判别分析分别

对四个板块的 MD&A 文本进行了分类，通过对比不同维度数的 GloVe 特征在分类任务中的表现，来评估维度选择对分类性能的具体影响。图 3-7 展示了四个板块在不同 GloVe 特征维度数下的分类准确率。

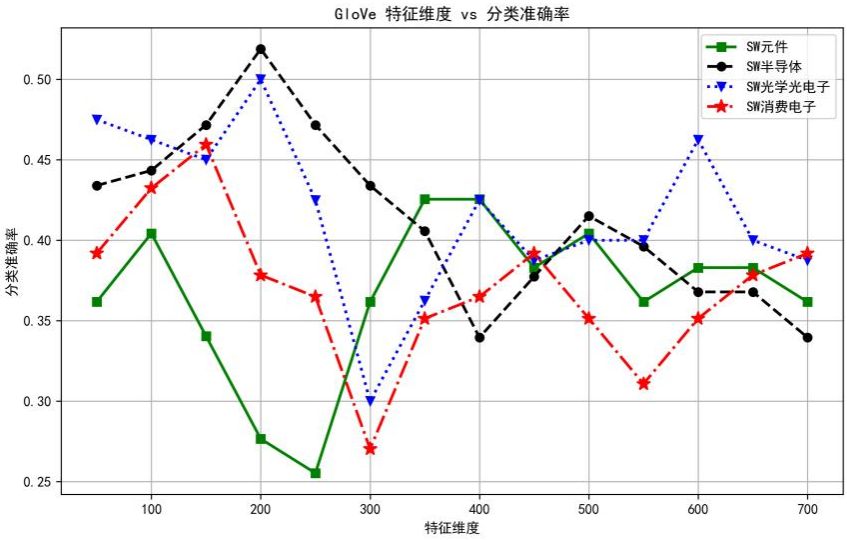


图 3-7 GloVe 特征维度的确定

由图 4-5，我们发现当 SW 消费电子的特征维度设置为 150 维，SW 半导体和 SW 光学光电子的特征维度均设置为 200 维、SW 元件的特征维度设置为 350 维时，它们的分类准确率均达到了最优水平。因此我们决定选择分类准确率达到峰值对应的维度数，作为这四个板块最终的 GloVe 向量维度。

为了保持一致性，我们对这四个板块的 BERT 向量维度进行相同设置，即与它们各自最优的 GloVe 向量维度保持一致。

在得到每个 MD&A 文本的 BERT 文本表示和 GloVe 文本表示后，对于支持向量的 BERT 文本表示，我们将与其同一个类标签内的 GloVe 文本表示作为正样本，其他类标签的 GloVe 文本表示作为负样本，以此构建基于 GloVe 的正负样本对。在对比学习实验中，使用 Adam 优化器来更新模型的参数，学习率设定为 1e-5，批次大小为 32，模型共训练 10 个轮次[42]。

通过对比学习的训练，SW 半导体、SW 元件、SW 光学光电子以及 SW 消费电子四个板块的负样本数量以及四个类标签具有代表性的两个负样本如表 3-2 所示。其中，全志科技作为 SW 半导体 A 标签公司的典型负样本，其 2022 年上半年的营业总收入及净利润同比均呈现下降；南亚新材作为 SW 元件 B 标签公

司的典型负样本，其 2023 年上半年营业收入下滑且净利润转亏；彩虹股份作为 SW 光学光电子 C 标签公司的典型负样本，其 2021 年上半年实现了显著的经营业绩增长，财务状况稳健；盈趣科技作为 SW 消费电子 D 标签公司的典型负样本，其在 2021 年主营业务表现出色，总营收与净利润均有所增长[2]。通过负样本的学习，模型能够深入学习四个板块各个类标签中上市公司经营状况的相似性和差异性特征，进而提升其在 MD&A 文本分类任务中的准确性和鲁棒性。

表 3-2 负样本的学习

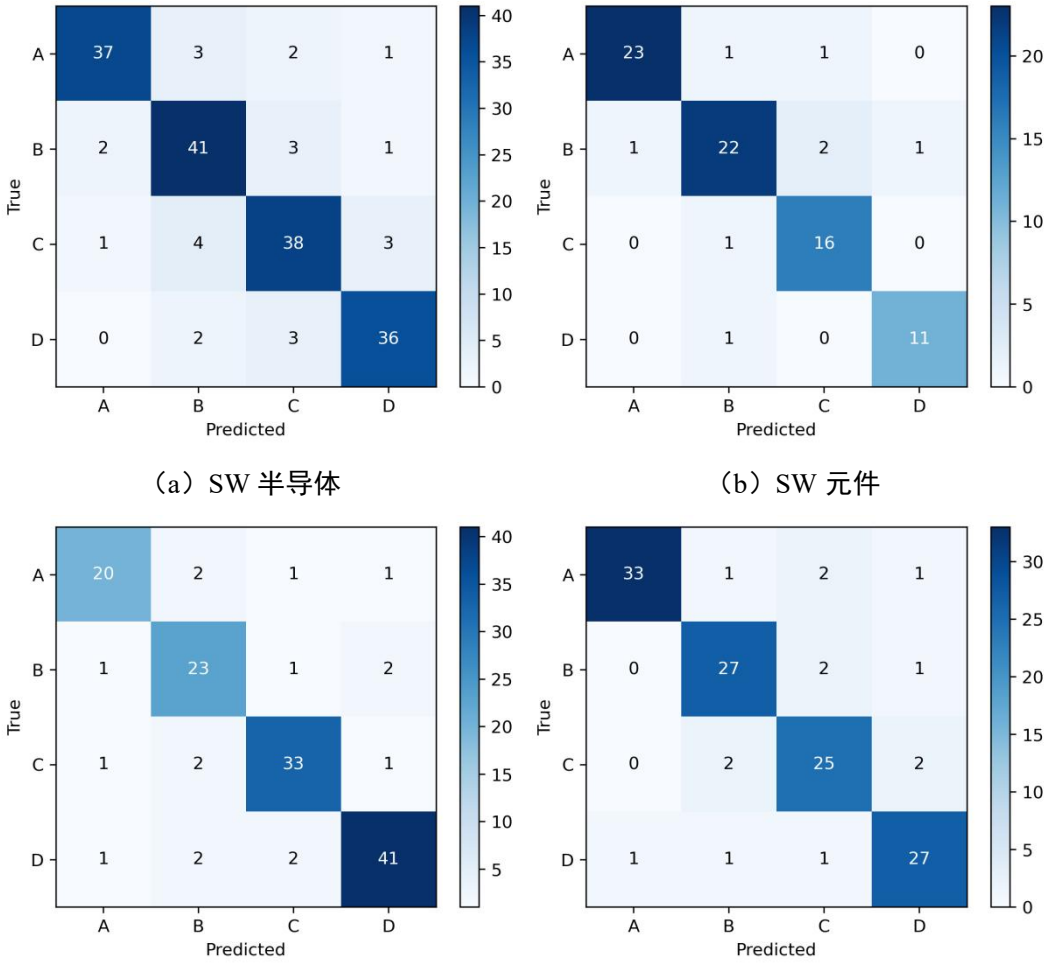
		SW 半导体	SW 元件	SW 光学光电子	SW 消费电子
A	负样本数量	399	160	327	260
	负样本代表	**全志科技 2022 半年报** 华微电子 2022 半年报	顺络电子 2022 年报 生益电子 2022 半年报	汇创达 2022 年报 久量股份 2021 半年报	海能实业 2022 半年报 星星科技 2022 年报
B	负样本数量	388	159	317	283
	负样本代表	圣邦股份 2022 半年报 中芯国际 2022 半年报	**南亚新材 2023 半年报** 惠伦晶体 2023 年报	八亿时空 2021 半年报 爱克股份 2022 半年报	春秋电子 2021 半年报 恒铭达 2022 半年报
C	负样本数量	390	183	288	285
	负样本代表	明微电子 2021 半年报 芯海科技 2021 年报	超声电子 2022 年报 鹏鼎控股 2021 年报	**彩虹股份 2021 半年报** 鸿利智汇 2021 年报	光弘科技 2021 半年报 贝仕达克 2022 半年报
D	负样本数量	404	200	259	282
	负样本代表	士兰微 2022 半年报 立昂微 2022 半年报	超声电子 2023 年报 深南电路 2022 半年报	奥来德 2021 半年报 天禄科技 2022 半年报	**盈趣科技 2021 年报** 环旭电子 2022 年报

3.4.4 模型性能评估

在这里我们通过混淆矩阵以及准确率、平均精确率、平均召回率和平均 F1

[2] 来自新浪财经

分数四个指标进一步验证 CL-KSVM 模型对经营状况的评估是有效的。对于每个板块评估模型测试性能最佳时的混淆矩阵如图 3-8 所示。



（a）SW 半导体　　（b）SW 元件

（c）SW 光学光电子　　（d）SW 消费电子

图 3-8 CL-KSVM 模型的混淆矩阵

由图 3-8（a）-（c）可知：SW 半导体 A 等级 43 家公司，其中 37 家预测正确；B 等级 47 家公司，其中 41 家预测正确；C 等级 46 家公司，其中 38 家预测正确；D 等级 41 家公司，其中 36 家预测正确。SW 元件 A 等级 25 家公司，其中 23 家预测正确；B 等级 26 家公司，其中 22 家预测正确；C 等级 17 家公司，其中 16 家预测正确；D 等级 12 家公司，其中 11 家预测正确。SW 光学光电子 A 等级 24 家公司，其中 20 家预测正确；B 等级 27 家公司，其中 23 家预测正确；C 等级 37 家公司，其中 33 家预测正确；D 等级 46 家公司，其中 41 家预测正确。SW 消费电子 A 等级 37 家公司，其中 33 家预测正确；B 等级 30 家公司，其中 27 家预测正确；C 等级 29 家公司，其中 25 家预测正确；D 等级 30 家公司，其中 27 家预测正确。

通过混淆矩阵，我们进一步得到每个板块的准确率，平均精确率，平均召回率，平均 F1 分数如表 3-3 所示。由表 3-3 可知，SW 元件分类性能最优，其准确率达到 90.00%，平均 F1 分数为 90.18%，显示出较高的分类性能和稳定性。SW 半导体与 SW 消费电子表现相近，但消费电子在准确率和 F1 分数上略胜 0.2-0.3 个百分点。分类效果均衡但略逊于 SW 元件，表明该板块在分类任务中存在一定的误判和漏判现象。总体而言，各行业 F1 分数与准确率基本吻合，说明模型在精确率和召回率间取得了平衡。

表 3-3 CL-KSVM 模型的性能评估

	SW 半导体	SW 元件	SW 光学光电子	SW 消费电子
准确率（%）	85.88	90.00	87.31	88.89
平均精确率	85.92	90.60	86.64	88.85
平均召回率	86.23	89.92	86.71	88.65
平均 F1 分数	86.03	90.18	86.64	88.69

3.5 本章小结

本章详细介绍了基于对比学习的 CL-KSVM 文本分类模型的构建过程。首先，采用 BERT 模型对 MD&A 文本进行表示，捕捉文本中的深层次语义信息。其次，引入对比学习机制，通过最大化正样本对的相似性和最小化负样本对的相似性，增强类内样本的同质性和类间样本的差异性。为了解决对比学习中的表示坍塌问题，本章采用了核支持向量机（KSVM）技术对样本进行归约，聚焦于对 MD&A 文本分类任务具有高度代表性的样本。

本章通过实验验证了 CL-KSVM 模型在 MD&A 文本分类任务中的有效性。实验结果表明，CL-KSVM 模型在四个行业板块中的分类准确率均达到了较高水平，尤其是在 SW 元件板块中表现最佳。本章的研究为后续引入图卷积神经网络奠定了基础。

第 4 章 融合对比学习与 GCN 的 CG-KSVM 文本分类

4.1 模型框架

在第三章的基础上引入了 GCN，将每个样本视为节点，BERT 表示和 GloVe 表示的串联视为节点的特征，通过设置相似度阈值来刻画节点间的邻接关系，再利用 GCN 聚合相邻节点的信息来更新节点表示，利用图结构中挖掘深层次的关联信息，以增强文本表示的鲁棒性，旨在更准确地进行 MD&A 文本分类。CG-KSVM 文本分类模型的具体构建流程如图 4-1 所示。

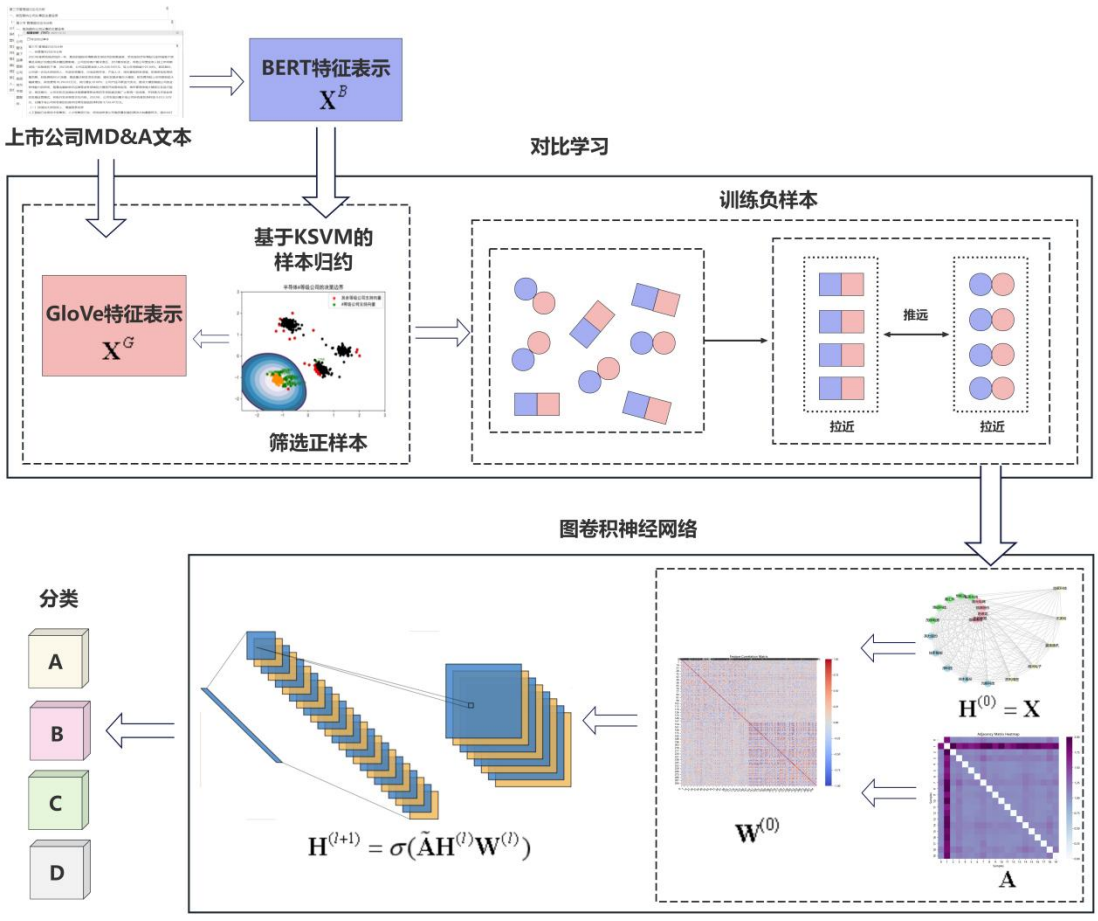


图 4-1 CG-KSVM 模型构建流程图

4.2 理论基础

随着社交网络、推荐系统、蛋白质相互作用网络等图数据的涌现，人们开始积极探索如何将深度学习模型应用于这些图数据。在这一背景下，2016 年 Thomas Kipf 和 Max Welling 提出了 GCN[41]，将卷积神经网络和图论的思想相结合，为图数据建立了一种新的深度学习模型。

GCN 的层级结构主要包括三个部分：输入层（Input Layer），图卷积层（Graph Convolutional Layer），输出层（Output Layer）。图数据主要由节点和边组成，假设它有 C 个节点（node），每个节点有 H 维特征，节点与节点之间的关系或连接称之为边。于是可以得到一个 $C\times H$ 维的特征矩阵 X，和一个反映节点之间关系的 $C\times C$ 维的邻接矩阵 A。特征矩阵 X 和邻接矩阵 A 便是 GCN 模型的输入。对于一个含有 N 个卷积层的 GCN，卷积层之间的传播方式为：

$$X^{(l+1)}=f(X^{(l)},A)=\sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X^{(l)}W^{(l)}),\quad l=0,1,2,...,N-1, \tag{4.1}$$

其中，$\tilde{A}=A+I$，I 是单位矩阵；D 是 A 的度矩阵且为一个对角矩阵，对角元素是每个节点的度数，$\tilde{D}=D+I$；$\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$ 为对称归一化拉普拉斯矩阵；若 $l=0$，$X^{(0)}$ 就是要输入到第一个卷积层的特征矩阵 X，$W^{(0)}$ 代表第一个卷积层的权重参数矩阵；若 $l=N-1, X^{(N-1)}$ 就是第 $N-1$ 个卷积层的输出，即第 N 个卷积层的输入，$W^{(N-1)}$ 是第 N 个卷积层的权重参数矩阵，$X^{(N)}$ 就是第 N 个卷积层的输出；σ 是非线性激活函数。

4.3 CG-KSVM 文本分类模型

在获得正负样本的 BERT 文本表示和 GloVe 文本表示后，将其 BERT 文本表示和 GloVe 文本表示进行串联，删除重复样本，这样 N 个 MD&A 文本所提供的类别信息就聚集在 M 个强鉴别样本中，即

$$\mathbf{X}=\left[X_i\right]=\begin{bmatrix}\hat{X}_1^B & \hat{X}_1^G\\ \vdots & \vdots\\ \hat{X}_{M^+}^B & \hat{X}_{M^+}^G\\ \tilde{X}_1^B & \tilde{X}_1^G\\ \vdots & \vdots\\ \tilde{X}_{M^-}^B & \tilde{X}_{M^-}^G\end{bmatrix}, \tag{4.2}$$

其中 X_i 为第 i 个强鉴别样本对应的 d_b+d_g 维的特征，且 $M=M^++M^-\le N$ 。

将对比学习训练得到的 M 个强鉴别样本的每个样本视为网络中的一个节点，其对应的 d_b+d_g 维的特征 X_i 就视为对应节点的特征。对于同类别标签的样本，其特征表示在特征空间内呈现出一定的相似性，从而对应的节点在网络结构中相互邻近。相反地，当两个样本的向量表示之间的相似性达到或超越某一预设的阈值 θ 时，即可判定它们拥有相同的类标签，进而将它们视为处于同一等级范畴。这种相似性程度越高，节点间的距离便越近。

为了量化这种相似性，我们采用余弦相似度作为衡量指标。在此框架下，节点间的邻接关系度量为

$$A=\frac{1}{sim\left(X_1,X_2\right)},$$

其中，$sim\left(X_1,X_2\right)=\dfrac{X_1\cdot X_2}{\left\|X_1\right\|\left\|X_2\right\|}$ 为两个强鉴别样本向量 X_1 和 X_2 的余弦相似度。当 $sim\left(X_1,X_2\right)\ge\theta$ 时，就可以判定这两个样本属于同一类别；当 $sim\left(X_1,X_2\right)<\theta$ 时，两个节点之间的邻接关系度量为 $A=0$ 。

于是 M 个强鉴别样本对应的邻接矩阵为一个 $M\times M$ 维的实对称矩阵 $\mathbf{A}$ ，

$$\mathbf{A}=\begin{bmatrix}A_{1,1} & \cdots & A_{1,M}\\ \vdots & \ddots & \vdots\\ A_{M,1} & \cdots & A_{M,M}\end{bmatrix}, \tag{4.3}$$

其中，$A_{i,j}$ 度量第 i 个强鉴别样本与第 j 个强鉴别样本间的邻接关系。

以节点特征矩阵 $\mathbf{X}$ （4.2）和邻接矩阵 $\mathbf{A}$ （4.3）为输入数据，通过 GCN 执行卷积操作，最终输出节点在高维空间中样本的特征表示。由（4.1），其信息

传播过程具体如下：

$$\mathbf{H}^{(l+1)}=\sigma(\tilde{\mathbf{A}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}),\ l=0,1, \tag{4.5}$$

若$l=0$，$\mathbf{H}^{(0)}$就是要输入到第一个卷积层的特征矩阵$\mathbf{X}$，$\mathbf{W}^{(0)}$代表第一个卷积层的权重参数矩阵；若$l=1$，$\mathbf{H}^{(1)}$就是第一个卷积层的输出，即第二个卷积层的输入，$\mathbf{W}^{(1)}$是第二个卷积层的权重参数矩阵，$\mathbf{H}^{(2)}$就是第二个卷积层的输出；$\tilde{\mathbf{A}}$ 为标准化的邻接矩阵，σ是非线性激活函数。考虑到层数过深可能会导致节点特征过度平滑，因此本模型选取 2 层 GCN 模型来进行 MD&A 文本分类[42]。

4.4 实验设计与结果分析

4.4.1 邻接关系阈值的学习

经过对比学习后的每个强鉴别样本，将其 BERT 文本表示和 GloVe 文本表示进行串联，得到一个强鉴别样本的特征表示。对于四个板块而言，SW 消费电子板块强鉴别样本的特征维度为 300 维，SW 半导体和 SW 光学光电子板块强鉴别样本的特征维度为 400 维、SW 元件板块强鉴别样本的特征维度为 700 维。

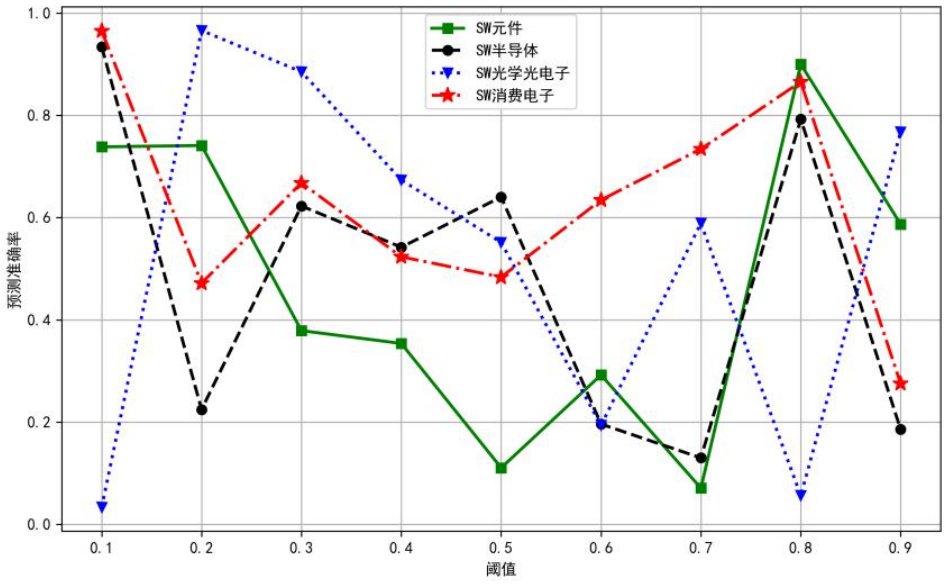
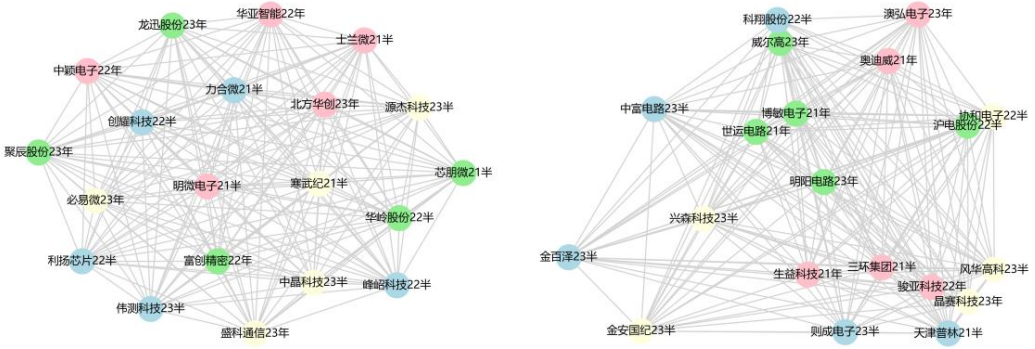


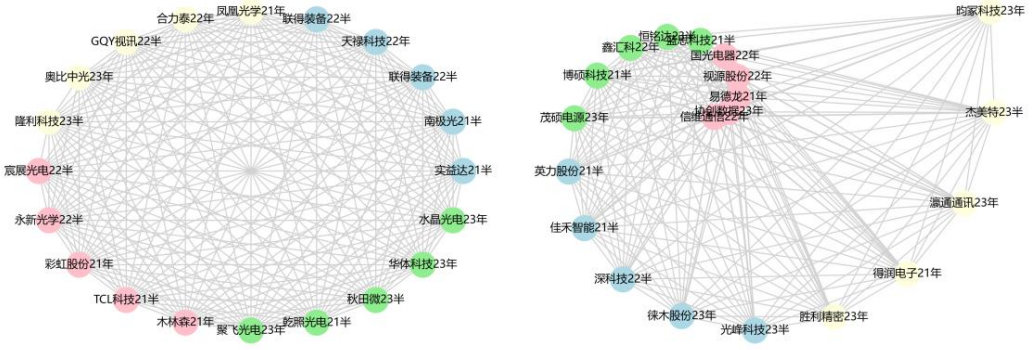
图 4-2 阈值的变化对预测准确率的影响

由图 4-2，预测准确率随着余弦相似度阈值的变化而变化。对于 SW 半导体，当余弦相似度阈值设置为 0.1 时，模型的预测性能最佳，其准确率为 93.43%；对于 SW 元件，当余弦相似度阈值设置为 0.8 时，模型的预测性能最佳，准确率为 90%；对于 SW 光学光电子，当余弦相似度阈值为 0.2 时，模型的预测性能最佳，准确率达到 96.58%；对于 SW 消费电子，当余弦相似度阈值为 0.1 时，模型的预测性能最佳，为 96.53%。综合上述分析，我们确定了不同板块的最佳余弦相似度阈值，SW 半导体和 SW 消费电子板块的阈值分别设置为 0.1，SW 光学光电子板块的阈值设置为 0.2，SW 元件板块的阈值设置为 0.8。

4.4.2 图数据的构建



（a）SW 半导体　　（b）SW 元件



（c）SW 光学光电子　　（d）SW 消费电子

图 4-3 强鉴别样本图表示可视化

将每个强鉴别样本视为一个节点，节点之间的邻接关系视为边，我们采用余弦距离来刻画样本间的相似性。当这个余弦距离超过预设阈值时，即判定两个样本处于同一类，对应节点间存在邻接关系，即节点间形成边连接。为了更直观地

呈现这一分析结果，我们分别从 SW 半导体、SW 元件、SW 光学光电子以及 SW 消费电子四个板块中的四类样本随机选择了 20 个强鉴别样本构建图数据如图 7 所示，从图 4-3（a）-（d）中可以看到每类的强鉴别样本的节点基本保持邻近关系。

4.4.3 GCN 网络结构的学习

我们通过改变权重参数矩阵 $\mathbf{W}^{(0)}$ 列维度的大小变化来观察 CG-KSVM 模型预测性能的变化如图 4-4 所示。SW 半导体板块的 $\mathbf{W}^{(0)}$ 列维度在 16 时，模型取得较好的预测效果，为 94.88%；SW 元件、SW 光学光电子与 SW 消费电子板块的 $\mathbf{W}^{(0)}$ 列维度在 32 时，模型都取得较好的预测效果。因此，对于 SW 半导体、SW 元件、SW 光学光电子与 SW 消费电子四个板块的 GCN 第一层权重参数矩阵 $\mathbf{W}^{(0)}$ 列维度分别设置为 16、32、32 和 32。

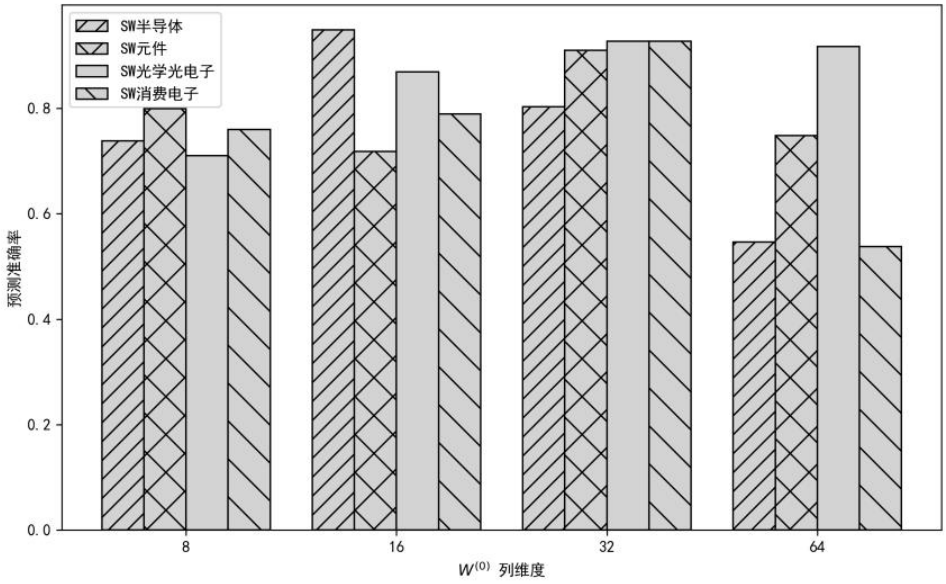


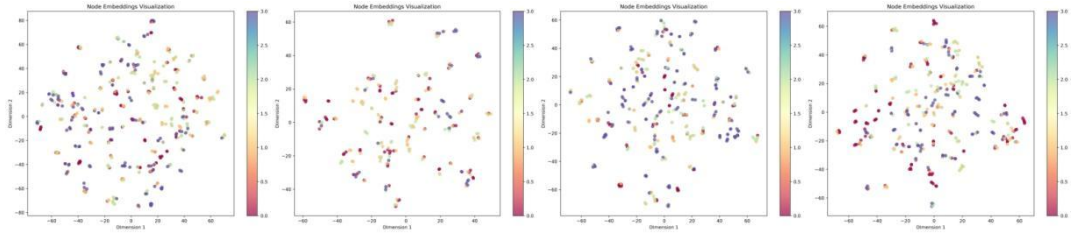
图 4-4 $\mathbf{W}^{(0)}$ 列维度的变化对预测准确率的影响

对于 $\mathbf{W}^{(1)}$ 的列维度，我们设置为类别数量 4。

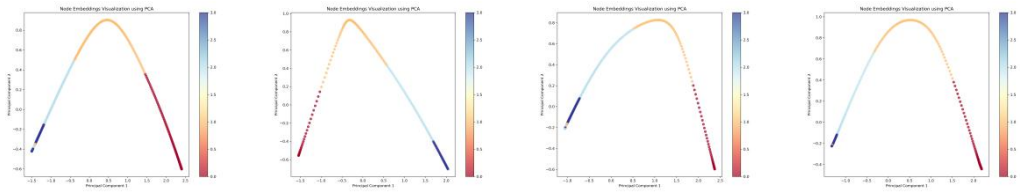
4.4.4 图嵌入表示

SW 半导体、SW 元件、SW 光学光电子以及 SW 消费电子四个板块 MD&A

文本的图嵌入可视化二维图如图 4-5 所示。图 4-5（a）的初始图展示了 MD&A 文本未经过 CG-KSVM 模型训练的图嵌入状态。此时，四个板块的 MD&A 文本在二维空间中的分布较为随机，各个类标签的 MD&A 文本之间的界限并不明显，存在一定的重叠或混杂现象。图 4-5（b）展示了通过 CG-KSVM 模型进行 1000 次迭代训练后的最终的迭代图。此时四个板块的 MD&A 文本被清晰地划分为四个类别标签，每个类标签内的 MD&A 文本较为紧密地聚集在一起，形成了明显的聚类区域。不同类标签之间的 MD&A 文本被有效地分隔开，没有出现明显的混淆或类别交叉现象，能够清晰地区分处于不同类别的 MD&A 文本。这一变化可以有力地证实了 CG-KSVM 模型已经成功地学习到了各个板块之间的特征差异和潜在关系。



（a）初始图



（b）最终迭代图

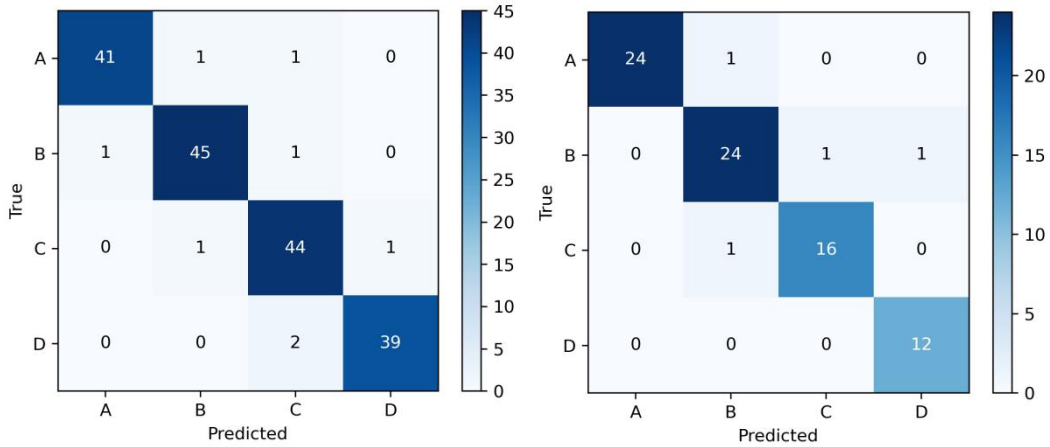
图 4-5 样本点的图嵌入表示效果图

4.4.5 分类性能评价

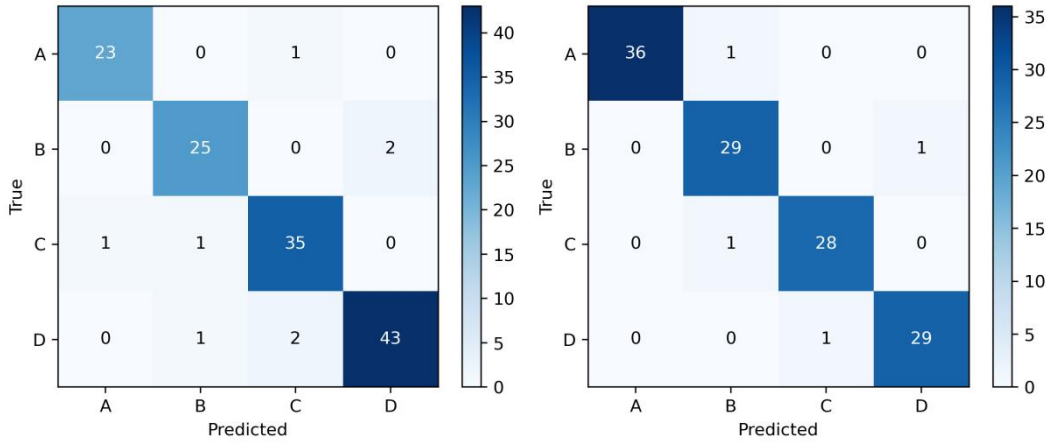
在这里我们通过混淆矩阵进一步验证 CG-KSVM 模型对经营状况的评估是有效的。对于每个板块评估模型测试性能最佳时的混淆矩阵如图 4-6 所示。

SW 半导体 A 等级 43 家公司，其中 41 家预测正确；B 等级 47 家公司，其中 45 家预测正确；C 等级 46 家公司，其中 44 家预测正确；D 等级 41 家公司，其中 39 家预测正确。SW 元件 A 等级 25 家公司，其中 24 家预测正确；B 等级 26 家公司，其中 24 家预测正确；C 等级 17 家公司，其中 16 家预测正确；D 等级 12 家公司，均预测正确。SW 光学光电子 A 等级 24 家公司，其中 23 家预测

正确；B 等级 27 家公司，其中 25 家预测正确；C 等级 37 家公司，其中 35 家预测正确；D 等级 46 家公司，其中 43 家预测正确。SW 消费电子 A 等级 37 家公司，其中 36 家预测正确；B 等级 30 家公司，其中 29 家预测正确；C 等级 29 家公司，其中 28 家预测正确；D 等级 30 家公司，均预测正确。



（a）SW 半导体　　（b）SW 元件



（c）SW 光学光电子　　（d）SW 消费电子

图 4-6 CG-KSVM 模型的混淆矩阵

和图 3-8 相比，CG-KSVM 模型在 CL-KSVM 模型的基础上进一步提升了预测性能。在 SW 半导体行业中，CG-KSVM 模型在 A、B、C、D 四个等级的预测准确率均有所提高，尤其是 A 等级和 B 等级的预测正确数分别从 40 家和 43 家提升至 41 家和 45 家。在 SW 元件行业中，CG-KSVM 模型在 D 等级的预测准确率达到 100%，优于 CL-KSVM 模型的 11 家正确预测。在 SW 光学光电子行业中，CG 模型在 A、B、C、D 四个等级的预测正确数均有所增加，其中 A 等级从 20 家提升至 23 家，B 等级从 23 家提升至 25 家。在 SW 消费电子行业中，

CG-KSVM 模型在 A、B、C、D 四个等级的预测准确率均显著提升，尤其是 D 等级的预测正确数从 28 家提升至 30 家，实现了 100%的准确率。综上所述，CG-KSVM 模型通过引入 GCN 进一步优化了分类性能，在各个行业和等级中均表现出优于 CL-KSVM 模型的预测能力。

为验证模型多特征融合的有效性，本研究选取 SW 电子行业四大板块，从准确率、平均精确率、召回率及 F1 分数四个维度，对比分析 CG-KSVM 模型与融合 GloVe 特征与对比学习的 GloVe-CL 模型、融合 BERT 特征与 GCN 的 BERT-GCN 模型的性能差异如表 4-1 所示。

表 4-1 与单特征模型的分类性能比较

	SW 半导体			SW 元件			SW 光学光电子			SW 消费电子		
	GloVe-CL	BERT-GCN	**CG-KSVM**	GloVe-CL	BERT-GCN	**CG-KSVM**	GloVe-CL	BERT-GCN	**CG-KSVM**	GloVe-CL	BERT-GCN	**CG-KSVM**
准确率（%）	76.57	80.00	**95.48**	75.95	83.75	**95.00**	78.36	84.33	**94.03**	79.37	82.54	**97.62**
平均精确率	76.52	79.95	**95.47**	75.85	84.05	**95.61**	77.86	83.74	**94.02**	79.26	82.63	**97.63**
平均召回率	76.69	80.09	**95.63**	75.66	84.83	**94.68**	77.65	83.89	**94.12**	79.44	82.50	**97.58**
平均 F1 分数	76.57	79.99	**95.53**	75.55	84.22	**95.10**	77.63	83.81	**94.07**	79.22	82.47	**97.58**

实验结果表明，CG-KSVM 在四个板块的准确率、平均精确率、平均召回率及平均 F1 分数上均显著领先。在 SW 半导体板块，CG-KSVM 的准确率达到 95.48%，相较于 GloVe-CL 和 BERT-GCN 分别提升了 18.91%和 15.48%；在 SW 元件板块，CG-KSVM 的准确率为 95.00%，较 GloVe-CL 和 BERT-GCN 分别提升了 19.05%和 11.25%，表明多特征联合优化可显著增强模型对语义差异的敏感度。此外，CG-KSVM 在平均精确率、平均召回率和平均 F1 分数上也均优于对比模型，表明该模型不仅在分类准确率上表现优异，同时在精确率和召回率之间实现了良好的平衡，具有较强的鲁棒性和泛化能力。

4.4.6 与其他模型对比

为了验证 CG-KSVM 的有效性，本文分别从传统的机器学习方法和深度学习方法中选取了多种经典算法，通过五折交叉验证对四个板块的 MD&A 文本进

行分类。其中基于传统机器学习的方法包括逻辑回归（LR）[43]、决策树（DT）[44]、随机森林（RF）[45]、K 近邻算法（KNN）[46]、支持向量机（SVM）[47]、引导聚集（Bagging）[48]和堆叠（Stacking）[49]模型。基于深度学习的方法包括图神经网络（Graph Neural Network）[50]和图注意力网络（Graph Attention Network）[51]模型。

将 CG-KSVM 与上述九种算法的分类准确率进行对比，结果如表 4-2 所示。从表 4-2 的实验结果可以看出，CG-KSVM 模型在 SW 半导体、SW 元件、SW 光学光电子以及 SW 消费电子四个板块的 MD&A 文本分类任务中均表现出显著的性能优势，其分类准确率均超过了 90%，分别为 92.41%、96.58%、92.70%和 94.96%。与其他九种经典算法相比，CG-KSVM 在四个板块上的分类准确率分别提升了 13.68%、7.70%、17.63%和 18.47%，展现了其强大的分类能力。特别是在 SW 半导体和 SW 光学光电子和 SW 消费电子板块，CG-KSVM 相较于表现次优的 GAT 模型，准确率分别提升了 17.63%和 18.47%，进一步凸显了其在处理复杂文本分类任务中的优越性。从对比结果来看，传统机器学习模型（如 LR、DT、RF 等）在四个板块上的分类准确率普遍较低，均值在 30%至 52%之间，且标准差较大，表明其分类性能不稳定。集成学习方法（如 Bagging 和 Stacking）虽然在一定程度上提升了分类性能，但其准确率仍显著低于 CG-KSVM。深度学习模型（如 GNN 和 GAT）在部分板块上表现较好，尤其是 GAT 在 SW 元件板块的准确率达到 88.88%，但其在其他板块的表现仍与 CG-KSVM 存在较大差距。

表 4-2 CG-KSVM 与经典算法分类准确率对比

方法	SW 半导体 Mean+sd（%）	SW 元件 Mean+sd（%）	SW 光学光电子 Mean+sd（%）	SW 消费电子 Mean+sd（%）
LR	41.57 ± 4.88	39.74 ± 8.97	40.30 ± 4.45	49.46 ± 4.73
DT	34.74 ± 5.14	36.72 ± 8.43	30.22 ± 5.72	37.03 ± 2.51
RF	40.99 ± 2.53	43.55 ± 7.14	48.08 ± 6.24	52.16 ± 4.57
KNN	36.23 ± 6.80	41.89 ± 4.82	42.29 ± 4.67	44.59 ± 8.24
SVM	41.56 ± 4.90	45.75 ± 9.43	44.07 ± 2.80	50.27 ± 3.95
Bagging	34.16 ± 2.23	38.45 ± 10.96	37.02 ± 4.05	46.76 ± 5.71
Stacking	42.89 ± 4.41	45.35 ± 8.21	44.84 ± 2.86	52.70 ± 5.60
GNN	44.03 ± 4.27	45.70 ± 6.07	50.61 ± 6.59	59.19 ± 2.32
GAT	78.73 ± 6.66	88.88 ± 3.69	75.07 ± 3.24	76.49 ± 2.36
CG-KSVM	**92.41 ± 3.79**	**96.58 ± 4.97**	**92.70 ± 5.76**	**94.96 ± 1.76**

4.5 本章小结

本章在第三章的基础上引入了图卷积神经网络，进一步提升了 MD&A 文本分类的准确性。通过将对比学习得到的强鉴别样本作为节点，节点间的邻接关系作为边，构建图结构数据。通过聚合邻居节点的信息来更新节点表示，GCN 捕捉文本数据间的复杂关联性与依赖性。

本章验证了 CG-KSVM 模型在 MD&A 文本分类任务中的卓越性能。实验结果表明，CG-KSVM 模型在四个行业板块中的分类准确率均超过了 90%，显著优于传统的机器学习与深度学习模型。本章的研究为 MD&A 文本分类提供了一种新的思路和方法。

第 5 章 结论与展望

5.1 结论

本研究针对上市公司 MD&A 文本分类问题，提出了一种结合对比学习与图卷积神经网络的 CG-KSVM 模型。该模型通过核支持向量机（KSVM）对样本进行归约，有效缓解了对比学习中常见的表示坍塌问题，筛选出具有高鉴别力的代表性样本；结合对比学习机制与图卷积神经网络（GCN），在捕捉类别间差异性的同时，利用图结构聚合邻域信息，深度挖掘文本的复杂语义关联。

实验结果表明，在 SW 电子行业四个板块的 MD&A 文本分类任务中，CG-KSVM 模型的分类准确率均超过 90%，显著优于传统机器学习与深度学习方法。此外，CG-KSVM 模型在平均精确率、召回率和 F1 分数等指标上均表现出良好的平衡性，标准差较低，验证了其鲁棒性与泛化能力，说明 CG-KSVM 不仅在分类准确率上表现优异，同时在高维特征学习和复杂语义关系建模方面展现出独特的优势，为 MD&A 文本提供了可靠的技术支持。

5.2 展望

尽管本文的研究取得了一定的成果，但仍存在一些值得进一步探索的问题和方向。

（1）特征提取的多样性：在提取特征时，除了 BERT 和 GloVe 模型，还可以考虑结合其他模型如生成对抗网络（GANs）、注意力机制（Attention）等模型。这些模型可以在文本表示阶段提取更多的关键特征，进一步提升模型的分类性能。例如，生成对抗网络可以通过生成新的样本来增强数据的多样性，而注意力机制则可以更好地捕捉文本中的关键信息。

（2）图池化方法与对比学习的结合：在图数据处理和图形分析中，图池化方法能有效降维图结构，提取关键特征；而对比学习则通过比较样本间的相似性与差异性，挖掘数据内在规律，且能充分利用无标签数据，避免繁琐标注。因此，未来的研究可以在 GCN 模型中加入图池化方法和对比学习，进一步增强模型的

性能。例如，图池化方法可以通过聚合节点信息来减少图的复杂度，而对比学习则可以通过构建正负样本对来提升模型的判别能力。

（3）多模态数据的融合：本文目前仅从文本角度展开研究，然而，未来的研究可以进一步拓展，将财务指标纳入分析范畴。通过结合文本数据与财务数据，我们可以对上市公司的经营状况、财务状况和市场表现进行全面而深入的了解。例如，可以将 MD&A 文本与公司的财务报表、市场表现数据等进行融合，构建多模态数据模型，进一步提升分类的准确性和鲁棒性。

参考文献

[1] 王莉, 王国军. 险资持股与上市公司经营风险[J]. 保险研究, 2024, (06): 43-54.

[2] 汪芳, 胡梦蝶, 赵玉林. 高新技术企业认定政策的高质量发展效应研究——来自中国A股上市公司面板数据的实证研究[J]. 工业技术经济, 2023, 42 (02): 79-85.

[3] 徐凯, 李东阳, 江宇. 基于MD&A文本分析的上市公司财务危机预警研究[J]. 会计之友, 2022, (17): 88-95.

[4] 卢哲, 张健. 基于SMOTETomek-RFE-MLP算法的上市公司信用风险预测[J]. 系统科学与数学, 2022, 42 (10): 2712-2726.

[5] Hakim H, Mounir G, Philippe C, et al. Negative sampling strategies for contrastive self-supervised learning of graph representations[J]. Signal Processing, 2022, 190: 108310.

[6] 谢娟英, 张建宇. 图卷积神经网络综述[J]. 陕西师范大学学报(自然科学版), 2024, 52 (2): 89-101.

[7] 檀莹莹, 王俊丽, 张超波. 基于图卷积神经网络的文本分类方法研究综述[J]. 计算机科学, 2022, 49 (08): 205-216.

[8] 李文静, 白静, 彭斌, 等. 图卷积神经网络及其在图像识别领域的应用综述[J]. 计算机工程与应用, 2023, 59 (22): 15-35.

[9] Wang H, Kim H D. Graph Neural Network-Based Speech Emotion Recognition: A Fusion of Skip Graph Convolutional Networks and Graph Attention Networks[J]. Electronics, 2024, 13 (21): 4208-4208.

[10] Xinxing F, Shuai Z, Long X, et al. Robust Classification Model for Diabetic Retinopathy Based on the Contrastive Learning Method with a Convolutional Neural Network[J]. Applied Sciences, 2022, 12 (23): 12071-12071.

[11] Joachims T. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization[J]. Carnegie-mellon univ pittsburgh pa dept of computer science, 1997, 143-151.

[12] Bijalwan V, Kumar V, Kumari P, Pascual J. KNN based machine learning approach for text and document mining[J]. International Journal of Database Theory and Application, 2014, 7(1): 61-70.

[13] Jin P, Zhang Y, Chen X, Xia Y. Bag-of-embeddings for text classification[C]. InIJCAI ,2016 (16): 2824-2830.

[14] Chen YH, Li SF. Using latent Dirichlet allocation to improve text classification performance of support vector machine[C]. 2016 IEEE Congress on Evolutionary Computation (CEC), 2016, 1280-1286.

[15] Tellez E S, Moctezuma D, Miranda-Jiménez S, et al. An automated text categorization framework based on hyperparameter optimization[J]. Knowledge-Based Systems, 2018, 149: 110-123.

[16] Zheng J, Yang W, Wang C, Jiang D, Wang D, Yang Q, Zhang Y. Research on complaint prediction based on feature weighted Naive Bayes[J]. InIOP Conference Series: Earth and Environmental Science, 2022(983): 012117.

[17] Karim M, Missen MM, Umer M, Fida A, Mohamed A, Ashraf I. Comprehension of polarity of articles by citation sentiment analysis using TF-IDF and ML classifiers[J]. PeerJ Computer Science, 2022(8): e1107.

[18] Hussain S, Malik MS, Masood N. Identification of offensive language in Urdu using semantic and embedding models[J]. PeerJ Computer Science, 2022,8: e1169.

[19] Hu W, Liu F, Peng J. An Efficient Data Classification Decision Based on Multimodel Deep Learning[J]. Computational Intelligence and Neuroscience, 2022: 2022-2023.

[20] Tokgoz E, Musafer H, Faezipour M, Mahmood A. Incorporating Derivative-Free Convexity with Trigonometric Simplex Designs for Learning-Rate Estimation of Stochastic Gradient-Descent Method[J]. Electronics,2023,12(2): 419.

[21] Liu T. A novel text classification approach based on deep belief network[C]. International Conference on Neural Information Processing. Springer, Berlin, Heidelberg, 2010: 314-321.

[22] Wang J, Yu LC, Lai KR, Zhang X. Dimensional sentiment analysis using a

regional CNN-LSTM model[J]. InProceedings of the 54th annual meeting of the association for computational linguistics, 2016(2): 225-230.

[23] Gu C, Wu M, Zhang C. Chinese sentence classification based on convolutional neural network[C]. IOP Conference Series: Materials Science and Engineering. IOP Publishing, 2017, 261(1): 012008.

[24] Kowsari K, Heidarysafa M, Brown D E, et al. Rmdl: Random multimodel deep learning for classification[C]. Proceedings of the 2nd international conference on information system and data mining. 2018: 19-28.

[25] Jiang M, Liang Y, Feng X, et al. Text classification based on deep belief network and softmax regression[J]. Neural Computing and Applications, 2018, 29(1): 61-70.

[26] Guo L, Zhang D, Wang L, et al. CRAN: a hybrid CNN-RNN attention-based model for text classification[C]. International Conference on Conceptual Modeling. Springer, Cham, 2018: 571-585.

[27] Qiu D, Jiang H, Chen S. Fuzzy information retrieval based on continuous bag-of-words model[J]. Symmetry, 2020,12(2):225.

[28] 唐焕玲,卫红敏,王育林,朱辉,窦全胜.结合 LDA 与 Word2vec 的文本语义增强方法[J]. 计算机工程与应用,2022,58(13): 135-145.

[29] Xu H, Wu L, Xiong S, Li W, Garg A, Gao L. An improved CNN-LSTM model-based state-of-health estimation approach for lithium-ion batteries[J]. Energy, 2023(276): 127585.

[30] Li X, Wang B, Wang Y, et al. Graph-based Text Classification by Contrastive Learning with Text-level Graph Augmentation[J]. ACM Transactions on Knowledge Discovery from Data, 2024, 18(4): 1-21.

[31] 魏雯婕, 张更平. 基于图神经网络的专利文本分类研究[J]. 竞争情报, 2024, 20(2): 24-34.

[32] 马艺翔,岳利媛. 电子信息制造业现况及发展前景浅析[J]. 中国国情国力, 2024, (7): 26-30.

[33] 陈飞. 中国电子信息制造业发展的内在动力[J]. 现代雷达, 2022, 44 (9): 125-126.

[34]沈隆,周颖. 管理层讨论与分析能预示企业违约吗?——基于中国股市的实证分析[J]. 系统管理学报, 2024, 33 (2): 441-459.

[35]师展,樊重俊. MD&A 文本实据性与企业权益资本成本:基于文本挖掘的方法[J]. 中国软科学, 2024, (7): 169-178.

[36]DEVLIN J, CHANG M-W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2024-09-13].

[37]Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation[C]Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014: 1532-1543.

[38]V. Vapnik, C. Cortes. Support-Vector Networks[J]. Machine Learning. 1995, 20(3): 273-297.

[39]Chen X K, Yuan Y H, Zeng G, et al. Semi-Supervised semantic segmentation with cross pseudo supervision[C] Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 2613-2622.

[40]Hu H, Wang X, Zhang Y, et al. A comprehensive survey on contrastive learning[J]. Neurocomputing, 2024, 610: 128645-128645.

[41]Kipf N T ,Welling M . Semi-Supervised Classification with Graph Convolutional Networks. [J]. CoRR, 2016, 1609.02907.

[42]You Y, Chen T, Wang Z, et al. L2-GCN: Layer-wise and learned efficient training of graph convolutional networks[J]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020, 2124-2132.

[43]周文清, 周达, 康建军. 基于逻辑回归二分类的核素识别算法研究[J]. 核电子学与探测技术, 2023, 43 (01): 12-17.

[44]赵静, 孔陶茹. 基于决策树分类算法的计算机大数据分析研究[J]. 自动化与仪器仪表, 2024, (10): 154-158.

[45]邵孟良, 齐德昱. 一种改进的随机森林 Boost 多标签文本分类算法[J]. 计算机应用与软件, 2022, 39 (11): 215-221+303.

[46]王文杰, 房海腾, 朱成杰, 等. 基于 KNN 分类算法的电力系统网络虚假数据

注入攻击防御方法[J]. 微型电脑应用, 2024, 40 (10): 130-134.

[47]吴乾生, 高健, 张揽宇, 等. 基于特征提取和 SVM 分类的 LED 芯片缺陷快速检测与实现[J]. 机械设计与制造, 2024, (06): 250-255.

[48]普运伟, 姜萤, 田春瑾, 等. 基于 MLP-Bagging 集成分类模型的在线学习行为分析[J]. 云南大学学报(自然科学版), 2024, 46 (05): 852-861.

[49]刘爱琴, 郭少鹏. 基于 Stacking 模型的学术论文多标签分类系统构建[J]. 国家图书馆学刊, 2024, 33 (02): 96-104.

[50]苏易礽, 李卫军, 刘雪洋, 等. 基于图神经网络的文本分类方法研究综述[J]. 计算机工程与应用, 2024, 60 (19): 1-17.

[51]张泽宝, 余翰男, 王勇, 等. 结合句法增强与图注意力网络的方面级情感分类[J]. 计算机科学, 2024, 51 (05): 200-207.

硕士期间发表的论文

[1] 杨灿清, 汪晓云, 张玥. 融合对比学习与图卷积神经网络的 MD&A 文本分类模型[J]. 安徽工程大学学报.（已录用）

致　谢

落笔于此，三年研究生涯即将画上句点。回首往昔，百感交集，思绪万千。三年前，我怀着对学术的憧憬与热忱踏入研究生阶段，如今站在这段旅程的终点，我深知这段时光不仅是我学术能力的提升，更是我人生中一段弥足珍贵的成长历程。

在本论文完成之际，我谨向所有在此过程中给予我帮助和支持的人致以最诚挚的谢意。首先，我要特别感谢我的导师张玥教授。从选题到实验设计，您始终以严谨的学术标准要求我，帮助我跨越了一个又一个难关。在每次实验结果不尽如人意时，您总能敏锐地指出问题所在，并引导我找到解决问题的方向。张玥老师深厚的学术造诣、对科研工作的热爱以及一丝不苟的治学态度，都深深地激励着我，让我领悟了真正的学术精神。您不仅在学术上给予我悉心指导，更在生活中给予我无微不至的关怀。每次组会后的交流时光，从全国各地的风土人情到当下生活的点滴，从学术研究到人生理想，这些轻松而深刻的对话让我感受到导师的博学与智慧，也让我在紧张的科研生活中找到了一份温暖与慰藉。张玥教授不仅是我的学术引路人，更是我人生中的良师益友，她的言传身教将永远指引着我未来的道路。

感谢我的师兄王越和丁沈杰。由于我此前未曾接触过 Python，在初入课题组时，你们从软件安装、环境配置开始，手把手教我编程，耐心解答我的每一个疑问。你们的无私帮助让我迅速融入了课题组的研究工作，并为我的后续研究打下了坚实的基础。我还要感谢我的同门杜杨，我们一起调试代码、讨论问题、分享经验，在互相鼓励中共同进步。

此外，我要感谢学校提供的良好的学术氛围和优越的办公条件，为我的研究工作提供了有力的支持。学校丰富的学术资源和舒适的研究环境，让我能够全身心地投入到科研工作中。同时，我也要感谢参与论文评审和答辩的各位专家和老师们，你们的宝贵意见和建议使我的论文更加完善。

最后，我要感谢我的父母。你们始终是我最坚强的后盾，无论我遇到什么困难，你们总是给予我无条件的支持和鼓励。在实验不顺、科研压力倍增时，你们的爱与关怀让我在学术道路上充满信心和力量，让我能够心无旁骛地追求自己的

梦想。你们的无私奉献和默默支持是我前进的动力，也是我人生中最宝贵的财富。

三年时光如白驹过隙，这段研究生生涯不仅让我在学术上有所收获，更让我学会了脚踏实地、坚持不懈与心怀感恩。未来的路还很长，我将带着这段经历中的所学所感，继续前行，迎接新的挑战。再次感谢所有帮助过我的人，愿我们在各自的人生道路上都能绽放光彩，书写属于自己的精彩篇章。