基于金融产品数字营销场景的自监督文本分类方法研究

统计学院

武宁

2025-11-08

中国人民大学
RENMIN UNIVERSITY OF CHINA

目 录

01

研究背景及意义

研究背景与意义

海量文本数据挑战

- ✓ 随着信息技术的爆炸式发展，有效分类成NLP核心挑战；
- ✓ 传统学习依赖人工标注，成本高，效率低，难以满足实际需求；
- ✓ 金融数字营销场景中涉及大量的文本数据，标注数据十分稀少；

自监督学习优势

- ✓ 解决标注问题；
- ✓ 挖掘数据深层含义；
- ✓ 学习通用表示；

自监督学习挑战

- ✓ 资源消耗大、领域适应差，推理成本高、解释性不足等挑战；
- ✓ 研究尚处早期；

研究意义与贡献

- ✓ 实际应用提升业务行业竞争力；
- ✓ 促进MLM和CL理论整合；
- ✓ 提出多重优化策略，提升模型准确率和泛化性。

文献综述

01

无标注文本分类方法

在20世纪90年代，基于规则和统计的方法，依赖大量人工标注；

K-means等无标注方法，但聚类结果不稳定；

Word2Vec实现了词语的向量化表示；

Seq2Seq模型通过编码器-解码器结构处理文本，但易丢信息；

02

无标注文本分类方法研究缺口

高度依赖人工特征工程，难以适应复杂语义表达

03

自监督学习发展

SSL通过自监督任务从未标注数据中学习特征和结构，降低标注成本，在CV和NLP领域展示强大性能；

04

Transformer与BERT

Transformer用自注意力机制克服RNN等方法在长序列处理方面的局限；

DBN、GAN推动了生成式模型和特征提取的进一步发展；

BERT、GPT-3的广泛应用展示了SSL的强大泛化性。

05

对比学习进展

MoCo、SimCLR等方法在CV领域通过不同机制解决负样本有限问题，提升模型表示能力；

DINO模型通过自蒸馏方式展示无监督对比学习的潜力；

06

自监督学习分类方法研究缺口

自监督学习集中于通用视觉与语言任务，对特定领域（如金融文本）的复杂语义结构建模与特征迁移能力的探索仍然有限

对比学习方法集中于视觉领域应用，在NLP领域应用有限

创新点

01

端到端金融模型创新

打通从数据到决策的实际应用路径，为金融机构智能营销提供新的技术手段。

语义增强范式创新

02

- ✓ 提出融合掩码语言建模与对比学习的多任务学习框架
- ✓ 提出动态权重机制克服传统单任务模型的视角局限

03

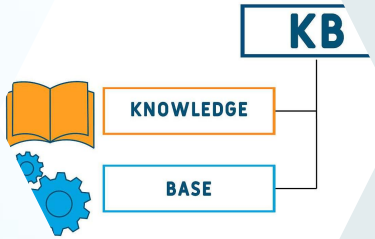
模型优化创新

- ✓ 提出多重优化机制
- ✓ 实现了表示学习与分类任务的协同优化

02

研究目标与内容

研究思路



统一框架

将生成式与对比式方法整合于统一框架，提升模型对多领域、多任务的泛化能力。

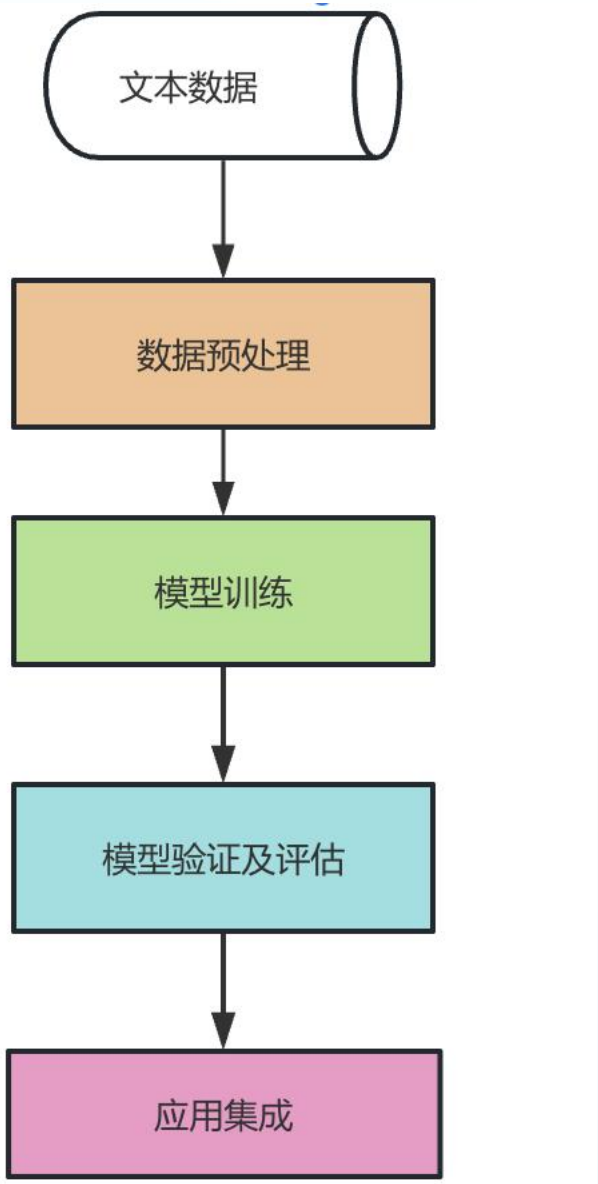
模型优化

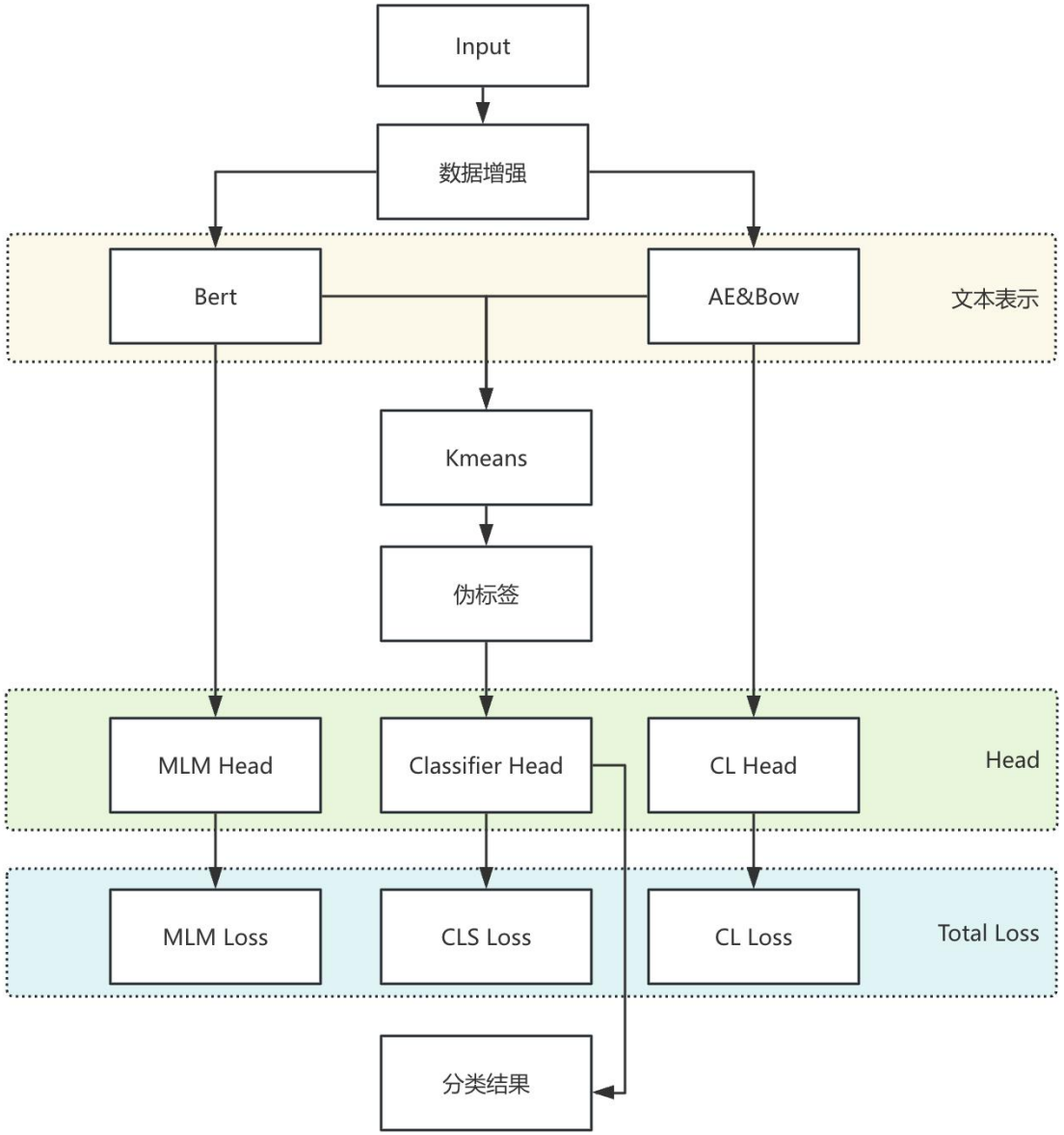
引入了五重机制来优化伪标签质量，提高模型精度。包括：双模态特征融合、动态聚类、置信度筛选、迭代优化、复合损失函数优化策略。

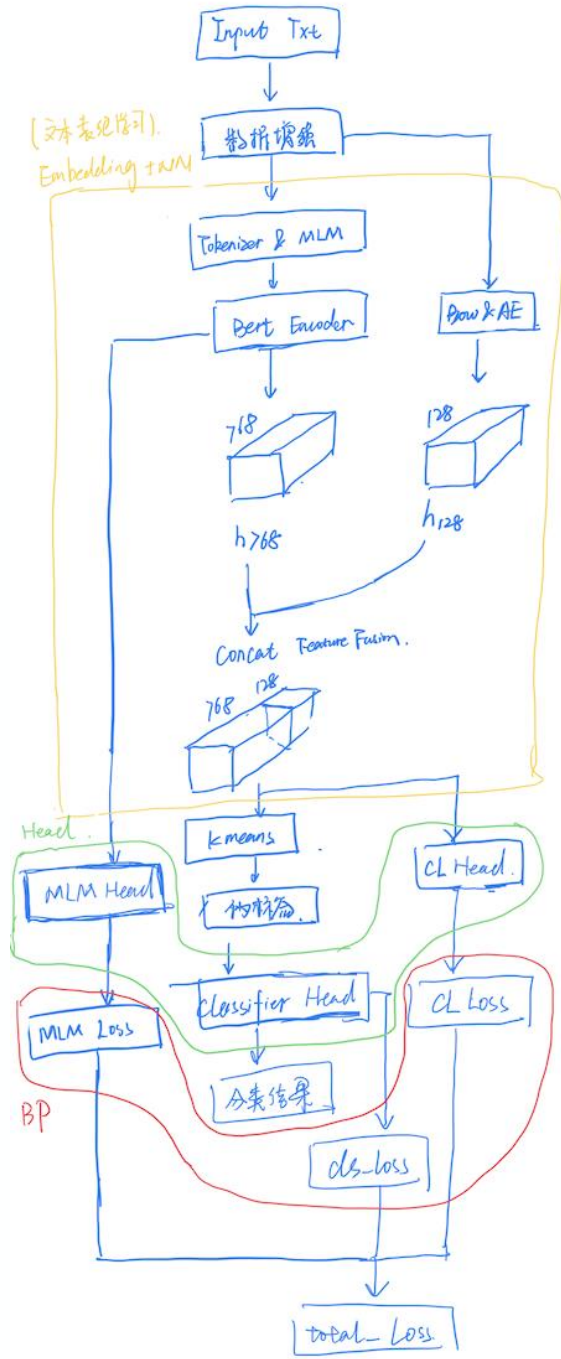
形成结论

总结了自监督学习在无标签文本分类中的应用机制，为无标签文本分类学习提供理论支持。

流程与框架


文本数据
数据预处理
模型训练
模型验证及评估
应用集成


Input
数据增强
Bert
AE&Bow
文本表示
Kmeans
伪标签
MLM Head
Classifier Head
CL Head
Head
MLM Loss
CLS Loss
CL Loss
Total Loss
分类结果


Input Txt
数据增强
(文本表现增对).
Embedding +NN
Tokenizer & MLM
Bert Encoder
Bow & AE
768
128
h768
h128
Concat Feature Fusion.
768
128
kmeans
MLM Head
CL Head.
MLM Loss
Classifier Head
CL Loss
BP
cls_loss
total_Loss

03

研究思路与方法

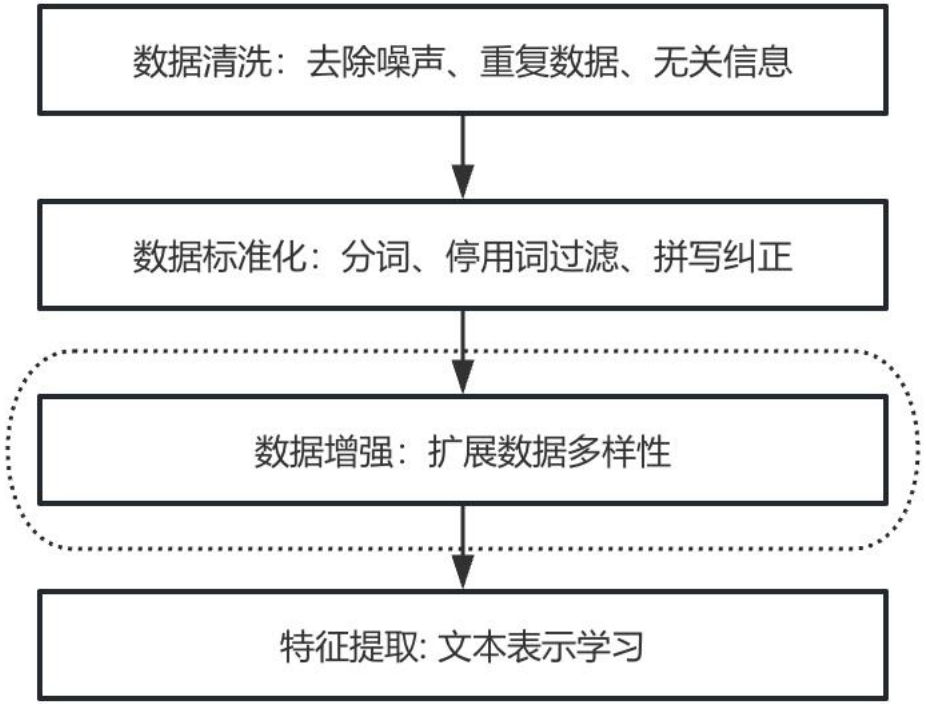
一、数据集及文本增强策略

伪标签生成数据集

构建包含金融评论、新闻文章、产品评论等数据的金融领域数据集；
由自监督学习生成伪标签；

文本增强

- ✓ 多种增强策略：同义词替换、随机插入删除、句子顺序打乱和噪声注入
- ✓ 解决多类别不平衡问题的增强策略：设计增加少数类别的样本数量的策略，避免模型对主流类型过度偏向。



二、文本表示学习方法

文本表示学习的挑战

传统BoW与TF-IDF无法捕捉语义关系与上下文信息；

词嵌入与BERT模型

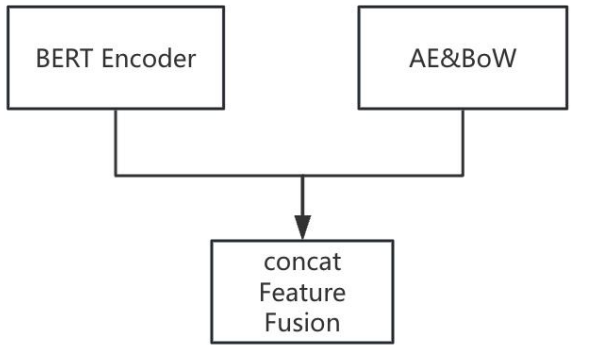
Word2Vec模型通过预测上下文来学习词的分布式表示；
BERT基于Transformer，使用掩码语言模型任务捕捉上下文信息；

自编码器

自编码器由编码器与解码器组成，学习数据低维表示，捕捉潜在语义结构。

BERT与AE双通道特征提取机制

BoW显式词频与AE隐式语义特征拼接，通过与BERT并行通道分别提取语义特征；
融合生成统一特征表示矩阵，生成多粒度更丰富的语义表示特征。



三、伪标签生成机制

添加项标题

01

动态聚类：在模型的初始阶段，伪标签通过聚类算法（K—means）生成，并进行结果比对，使用改进的Calinski—Harabasz（ICH）指数动态确定最佳聚类数量K，避免人为设定。

02

置信度筛选：根据样本与聚类中心的距离，动态设定置信度阈值，选取置信度较高的样本作为伪标签生成的基础，有效避免噪声干扰。

03

迭代优化：在多轮迭代中，利用模型自身的预测结果和动态阈值更新伪标签，形成“模型优化—伪标签质量提升—模型进一步优化”的良性循环。

四、动态掩码语言建模策略

掩码策略设计背景

传统的掩码语言建模任务（选择一定比例的词汇进行遮蔽）：

1.过度掩码导致模型无法有效学习上下文信息；

2.单一掩码策略无法有效平衡模型对局部及全局语义信息的学习。

掩码策略设计

改进的动态掩码策略：

1.逐步加深掩码深度，初期以较低掩码比例开始训练10%，随着训练过程的深入15%，逐步增加掩码比例。

2.在保留足够上下文前提下，将不同形式的掩码策略动态结合：单一词汇掩码；掩码句子重要成分；动态调整掩码位置、比例。

掩码策略方法对比

对比分析，论文中表4-1所示，动态掩码策略在准确率、F1值、AUC值均表现优异，证明动态掩码策略在文本分类任务中有显著的优越性，能有效捕捉文本深层语义，提高模型泛化能力。

五、对比学习策略

对比学习方法设计

定义合理的正负样本对

基于InfoNCE定义损失函数，目标是最大化正样本对的相似度并最小化负样本对的相似度

加入正则项来增强模型的泛化能力

$$\mathcal{L}_{\mathrm{CL}} = \mathcal{L}_{\mathrm{InfoNCE}} + \mathcal{L}_{\mathrm{reg}}$$

对比学习方法性能评估

实验证明对比学习能显著提升模型的分类性能，平均准确率提升2.0%，验证了方法的有效性。

对比学习损失函数对模型性能的影响

表 4-4 对比学习损失函数对模型性能的影响

数据集	原始准确率	基于对比学习损失函数的准确率	提升百分比
金融评论	85.3%	87.8%	2.5%
新闻文章	81.2%	83.1%	1.9%
产品评论	79.8%	81.4%	1.6%
总计	82.1%	84.1%	2.0%

对比学习损失函数

$$\mathcal{L}_{\text{InfoNCE}} = -\log\left(\frac{\exp\left(\text{sim}(x_i, x_j)/\tau\right)}{\sum_{k=1}^{N}\exp\left(\text{sim}(x_i, x_k)/\tau\right)}\right)$$

$$\mathcal{L}_{\text{reg}} = \lambda \cdot \|\theta\|^2$$

$$\mathcal{L}_{\text{CL}} = \mathcal{L}_{\text{InfoNCE}} + \mathcal{L}_{\text{reg}}$$

- $\text{sim}(x_i, x_j)$表示样本之间的相似度，通常通过余弦相似度来计算:
- N是批量中的样本数。
- τ 是一个温度超参数，用于调节分布的平滑度。
- lamda是正则化超参数
- $\|\theta\|$: 是模型参数的L2范数

六、分类器设计方法

01

综合损失函数

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \alpha \cdot \mathcal{L}_{mlm} + \beta \cdot \mathcal{L}_{cl}$$

融合主分类任务、对比学习、掩码语言建模的综合损失函数；

平衡系数

02

设置平衡系数，通过动态权重机制实现多任务协同优化；

03

Focal Loss缓解不平衡

采用Focal Loss来缓解分类过程中的类别不平衡问题，通过降低易分类样本的权重（γ=2），提高分类的精度。

分类器设计方法

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \alpha \cdot \mathcal{L}_{mlm} + \beta \cdot \mathcal{L}_{cl}$$

- L_class 是主分类任务的损失。
- L_contrast 是基于InfoNCE的对比学习损失，负责拉近正样本、拉远负样本。
- L_mlm 是掩码语言建模损失，负责重构被掩码的词汇。
- α 和 β 是超参数，用于平衡三个任务的贡献度。

$$\mathcal{L}_{\text{cls}} = -\sum_{i=1}^{N} (1 - p_{i,y_i})^{\gamma} \log(p_{i,y_i})$$

MLM损失函数

$$\mathcal{L}_{MLM} = -\sum_{i \in M} \log P_\theta(x_i \mid x_{masked})$$

计算预测分布与真实标签之间的 交叉熵（Cross Entropy）。只在被 [MASK] 的 token 上计算交叉熵损失，从而得到 MLM 的训练目标。

04

实验与结果分析

实验设置

01 实验数据集

金融领域真实应用场景的混合数据集：
金融评论数据集、
新闻文章数据集、
产品评价数据集。

02 数据集划分详情

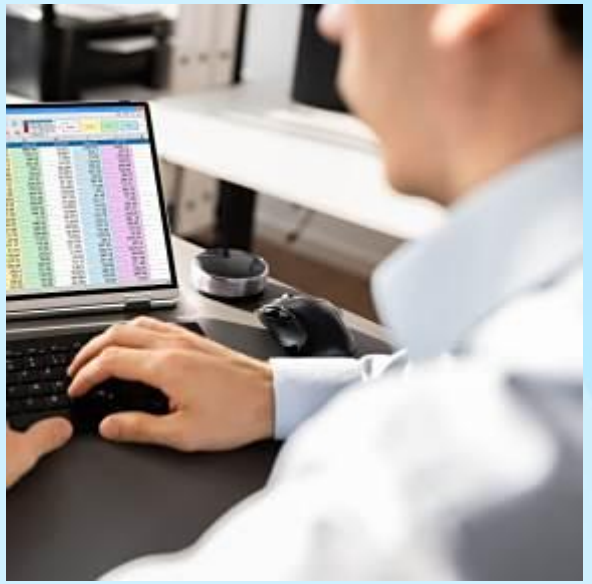
实验数据集划分为训练、验证、测试集，比例7:1:2。

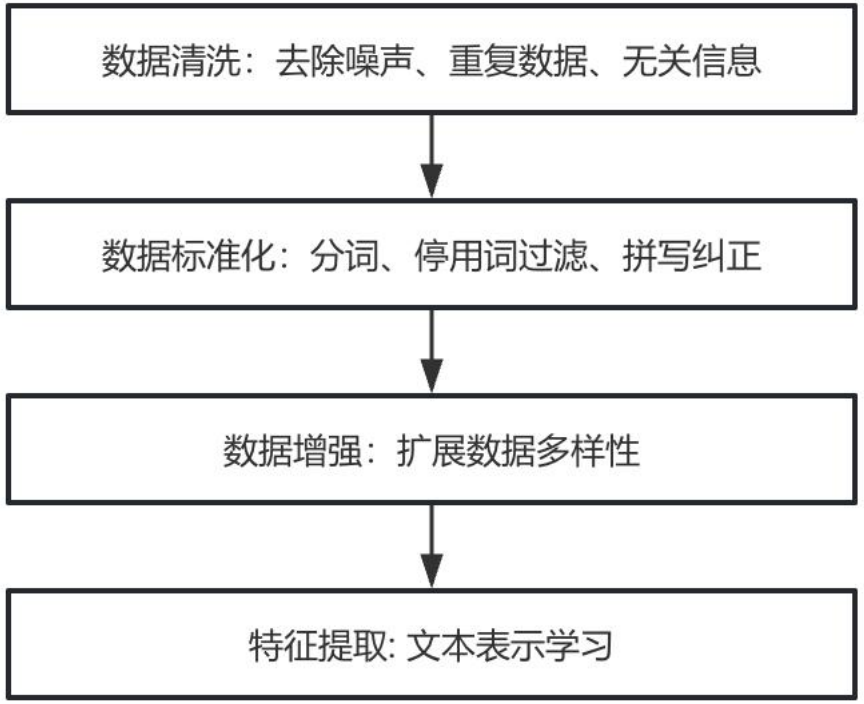
03 分类任务训练参数

采用Adam优化器训练，学习率0.001，批大小32，训练50轮，融合L2正则化防过拟合，对比学习损失优化文本相似度，提升模型性能。

04 性能评估指标

实验设定分类准确率、F1分数、AUC值等指标评估模型性能，准确率衡量整体正确率，F1分数综合精确与召回率，AUC值评估模型区分能力。



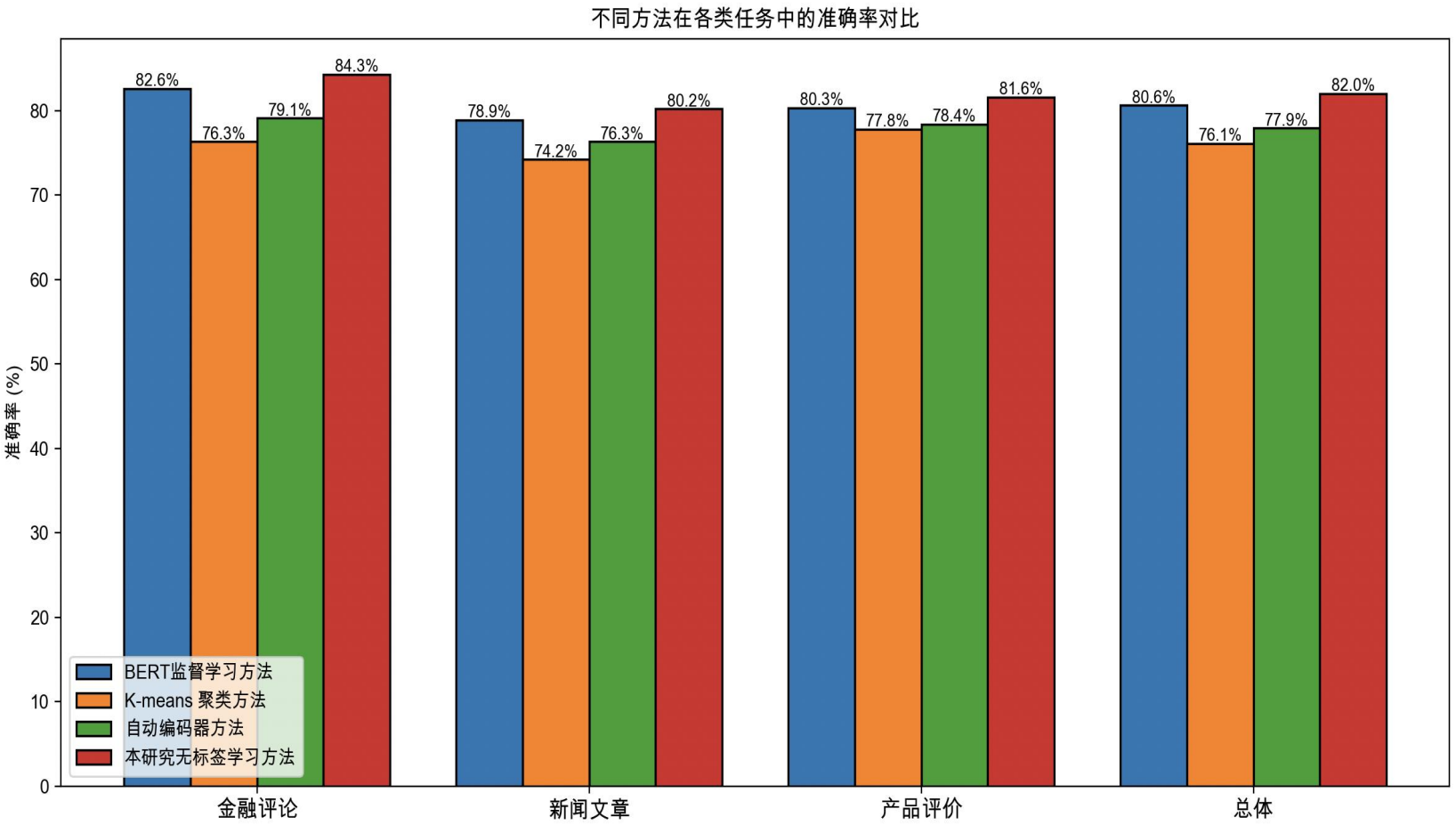


数据主要通过爬虫、API方式获取。

公司内部的数据集

开源数据集

模型整体性能分析



整体性能评估

各任务上综合准确率均优于基线模型，总体准确率达82.0%；

金融数据集准确率达84.3%（附录表1-1）

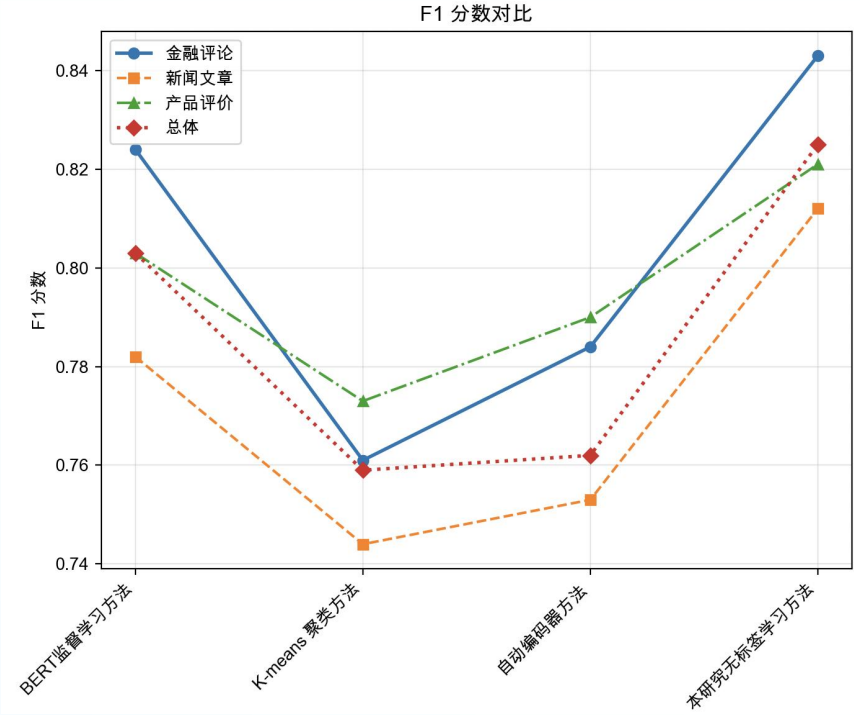
表 1-1 文本分类任务不同方法性能对比

方法名称	金融评论准确率	新闻文章准确率	产品评价准确率	总体准确率
BERT 监督学习方法	82.6%	78.9%	80.3%	80.6%
K-means 聚类方法	76.3%	74.2%	77.8%	76.1%
自动编码器方法	79.1%	76.3%	78.4%	77.9%
本研究无标签学习方法	84.3%	80.2%	81.6%	82.0%

模型整体性能分析

F1值与AUC值对比图

在F1值与AUC值上表现优异，金融评论集F1值达0.843，AUC为0.917，证明方法在精确度、召回率、判别能力具有显著优势。（附录表1-2）



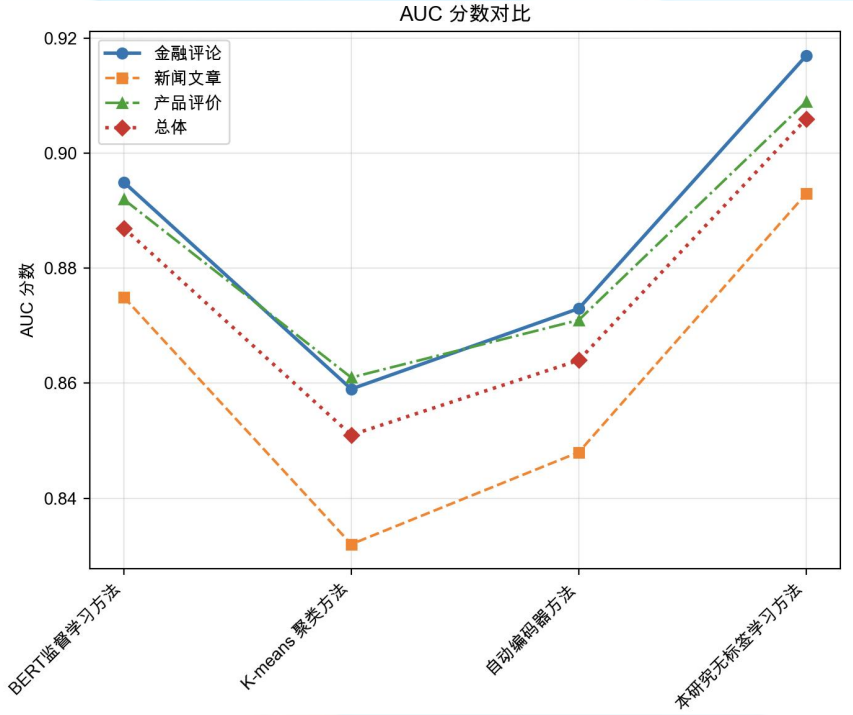


表 1-2 文本分类任务不同方法 F1 分数和 AUC 值对比

方法名称	金融评论 F1 分数	新闻文章 F1 分数	产品评价 F1 分数	总体 F1 分数	金融评论 AUC	新闻文章 AUC	产品评价 AUC	总体 AUC
BERT 监督学习方法	0.824	0.782	0.803	0.803	0.895	0.875	0.892	0.887
K-means 聚类方法	0.761	0.744	0.773	0.759	0.859	0.832	0.861	0.851
自动编码器方法	0.784	0.753	0.790	0.762	0.873	0.848	0.871	0.864
本研究无标签学习方法	0.843	0.812	0.821	0.825	0.917	0.893	0.909	0.906

模型整体性能分析

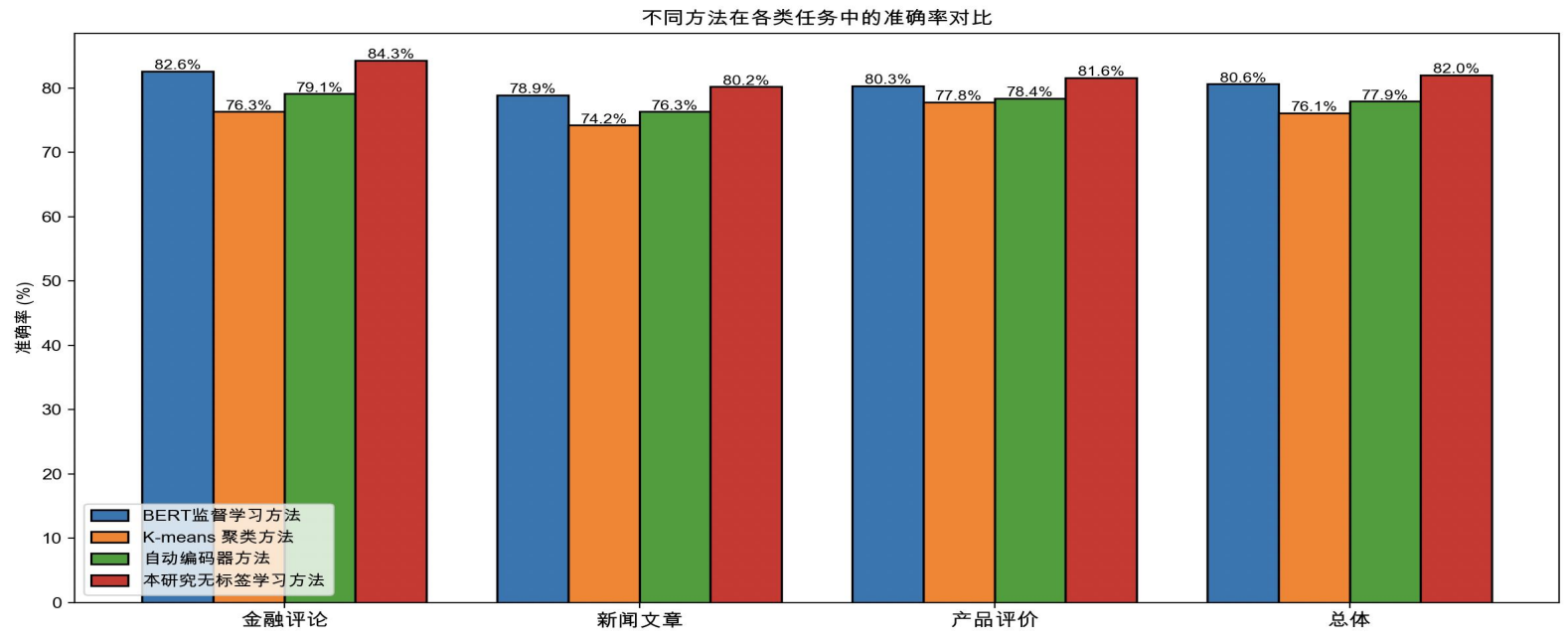


表 5-4 各方法在测试集上的性能对比

方法名称	准确率(%)	F1 分数	AUC 值	性能下降率
K-means 聚类	75.8	0.742	0.823	4.2%
AutoEncoder	78.2	0.768	0.845	3.6%
BERT 监督学习	84.5	0.831	0.901	1.2%
本文方法	82.3	0.812	0.887	1.7%

泛化能力评估

测试集准确率82.3%，性能下降率1.7%，远低于其他方法；(表5-4)

泛化能力较强

样本外预测能力表现优异。

泛化误差与交叉K折验证结果对比

表 5-8 各方法泛化误差对比

方法名称	训练误差	测试误差	泛化误差	过拟合风险
K-means 聚类	0.285	0.382	0.097	高
AutoEncoder	0.253	0.324	0.071	中
BERT 监督学习	0.170	0.185	0.015	低
本文方法	0.186	0.218	0.032	低

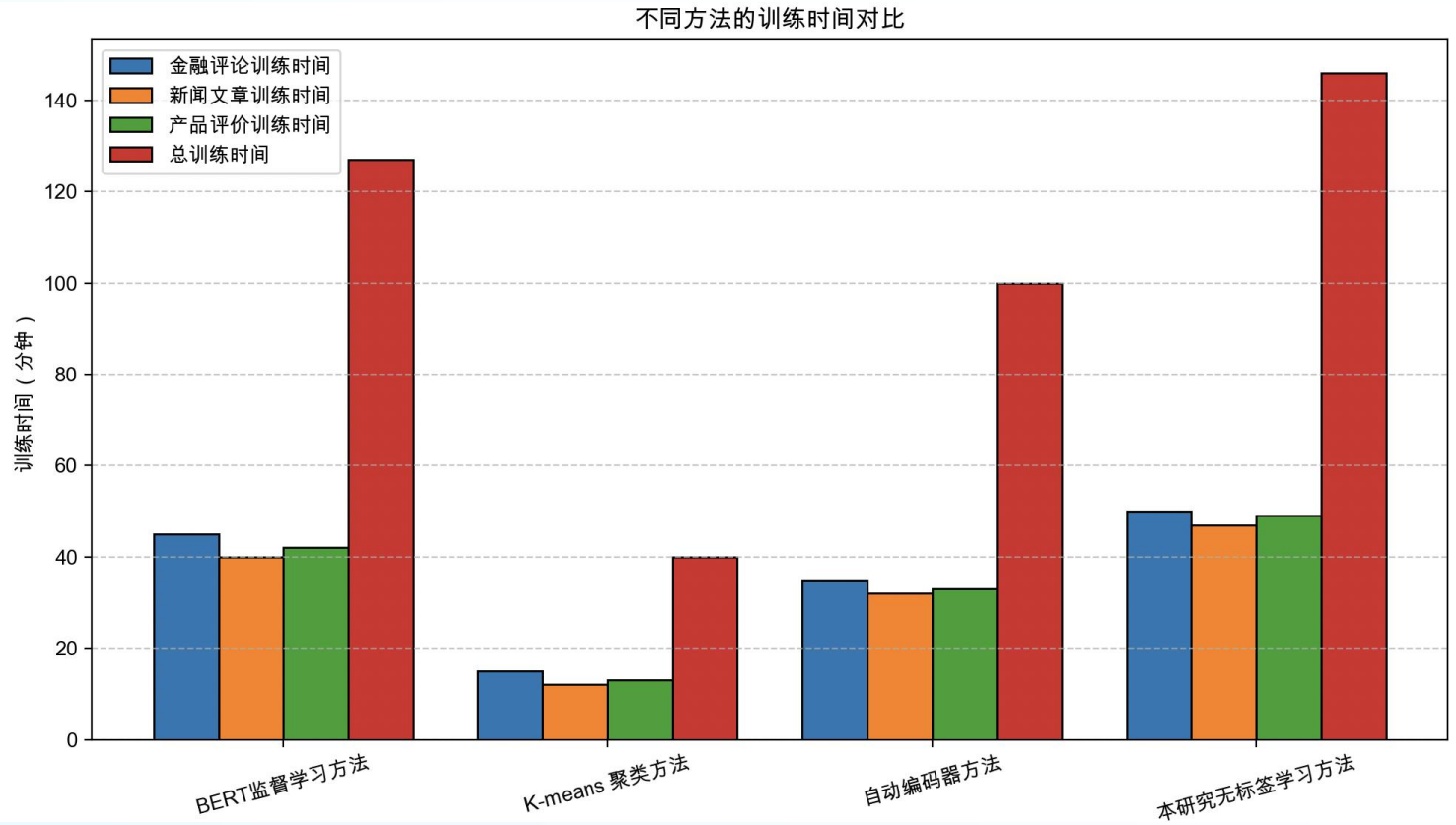
表 5-9 5 折交叉验证结果对比

方法名称	折 1	折 2	折 3	折 4	折 5	平均值	标准差
K-means 聚类	75.3	76.8	74.2	77.1	75.9	75.86	0.031
AutoEncoder	78.4	79.2	77.8	78.9	79.1	78.68	0.025
BERT 监督学习	84.2	84.8	83.9	84.5	85.1	84.48	0.010
本文方法	82.1	83.2	81.8	82.9	83.1	82.62	0.018

我的方法泛化误差最低（0.032，表5-10），证明其过拟合风险小，稳定性强。

5折交叉验证的标准差仅为0.018（表5-11），远低于对比方法，证明了模型的鲁棒性。

模型整体性能分析



训练时间与效率图

研究方法虽训练时间较长，但准确率和F1分数优势显著；
性能提升使得额外训练时间投入是值得的。

其他影响因素分析

数据集规模影响

随着数据规模的增加，模型分类性能提升，表明大数据集增强模型鲁棒性与泛化能力，验证了模型在无标注文本分类中的有效性与适应性。

多任务学习提升

采用多任务学习框架提升模型性能，分类准确率、F1分数、AUC值均显著提升，特别是分类准确率从86.4%提升至88.0%，验证了学习框架的有效性。

置信阈值影响分析

对比不同置信阈值对伪标签生成的影响，发现较低的置信阈值能生成更多高质量伪标签，优化伪标签质量，进一步提升分类准确率至88.5%。

05

模型在实际应用的实例分析

应用实例

单击添加标题

识别明令禁止的违规用语，及时触发预警（如："保本保收益""零风险"等）

单击添加标题

识别用户兴趣类别（如“基金”、“保险”、“数字货币”），实现精准的广告推送和产品推荐。

识别用户风险偏好类别（如“‘理财型”“投机型”或“稳健型”），提升广告转化率和客户粘性

对金融新闻进行自动分类，快速掌握政策动向、市场热点和竞争对手的动态，为营销策略调整提供数据支持。

06

总结与展望

工作总结

理论方面

- ✓ 分析无标签文本分类现状
- ✓ 提出结合MLM与CL任务的无标签文本分类多任务框架
- ✓ 解决特定领域缺少数据标注文本分类问题

模型训练

- ✓ 提出多重优化策略
- ✓ 有效捕获文本中的深层语义信息
- ✓ 提高分类精度和模型的泛化能力

实验测试

- ✓ 通过系统的全方面的实验，验证方法在多个数据集上均表现良好，证明所提方法的优越性

应用方面

- ✓ 从金融产品个性化推荐、合规监控、营销决策场景三个典型应用实例分析方法的实用价值

局限性与改进计划

长文本处理优化

长文本处理可以结合目前最新的研究成果：探索Transformer变体（如Longformer） 或结合图神经网络来更好地建模长文本场景。

结合GPT大模型优化架构

基于GPT进行posttraining 训练领域内的小模型。

利用Prompt工程生成伪标签

将源领域和目标领域的样本输入GPT，提示其生成风格接近目标领域的数据。

模型蒸馏

所提方法作为知识迁移的“教师模型”，在少量标注样本或无标签数据上生成伪标签；
再用这些高质量伪标签来训练一个更小的模型。

未来展望

模型推理能力

目标是让模型不仅“读懂”文字，还能像金融分析师一样进行简单的因果推断和逻辑推理，从而实现更深层次的意图理解、风险预警和投资决策支持。

与结构化数据融合

未来的研究可以探索如何将金融知识图谱、宏观经济指标、公司财报数据等结构化信息，与文本的非结构化信息进行融合预训练。

轻量级模型

在保证学习准确度的前提下，加快模型学习效率，为业务推动提供更快速的数据保障。

具有金融常识与推理能力的自监督模型

致谢

汇报人：武宁