 DOI: https://doi.org/10.12345/bdai.v6i2.24590

Machine learning-driven text data deweight and redundancy filtering techniques

Dongyu Li

Nankai University, Tianjin, 300071, China

Abstract

This paper proposes a comprehensive framework based on machine learning, which realizes the efficient collaboration of text data deweight and redundancy filtering by integrating the multimodal features and dynamic optimization strategies. The innovation of this framework is reflected in three aspects: first, combining the characteristics of word embedding and syntactic structure to construct both semantic and grammar; second, introducing attention mechanism and contrast learning to enhance the ability to capture long text and complex semantic relationship; third, designing dynamic optimization module to adapt to real-time changes of data flow through online learning and incremental update mechanism. Experiments show that the proposed framework is significantly better than traditional methods in accuracy and recall, and shows good robustness in multi-source heterogeneous data scenarios.

Keywords

machine learning; text data removal; redundant filtering

机器学习驱动的文本数据去重与冗余过滤技术

李东宇

南开大学，中国·天津 300071

摘　要

本文提出一种基于机器学习的综合框架，通过融合多模态特征与动态优化策略，实现文本数据去重与冗余过滤的高效协同。该框架的创新性体现在三个方面：其一，结合词嵌入与句法结构特征，构建兼顾语义与语法的文本表示模型；其二，引入注意力机制与对比学习，增强对长文本及复杂语义关系的捕捉能力；其三，设计动态优化模块，通过在线学习与增量更新机制适应数据流的实时变化。实验表明，该框架在准确率、召回率等指标上均显著优于传统方法，且在多源异构数据场景下展现出良好的鲁棒性。

关键词

机器学习；文本数据去重；冗余过滤

1 引言

随着互联网与数字化技术的迅猛发展，文本数据呈现爆炸式增长，从社交媒体、新闻资讯到学术文献，海量信息的积累为数据存储、处理和分析带来了前所未有的挑战。其中，重复与冗余数据的高频出现不仅挤占存储资源、降低计算效率，更对信息检索、知识挖掘等下游任务造成严重干扰 [1]。在此背景下，如何高效识别并过滤文本中的重复与冗余内容，已成为提升数据质量、优化系统性能的核心问题之一。传统文本去重与冗余过滤技术主要依赖哈希算法、规则匹配或统计方法（如 TF-IDF），这些方法虽能在简单场景下快速处理数据，但面对语义相似性识别、上下文关联性分析等复杂需求时，往往因缺乏对文本深层语义的理解而表现乏力。近些年来，机器学习的兴起为这一领域注入了新的活力，尤其是深度学习方法在文本表示与相似性度量中的突破，使得语义层级的去重与冗余过滤成为可能。然而，现有研究仍面临多源异构数据融合困难、动态数据环境适应性不足等瓶颈，亟需一种兼顾效率与精度的系统性解决方案。

2 文本数据去重与冗余过滤的技术现状与挑战

2.1 传统方法的局限性

文本去重与冗余过滤的早期研究主要依赖规则驱动的基础技术。基于哈希的方法（如 SimHash、MD5）通过生成文本指纹实现快速比对，但其核心逻辑仅关注字面重复，无法识别语义层面的相似性。规则匹配方法通过人工设定关键词或正则表达式过滤冗余内容，虽然在小规模结构化数据中表现稳定，却难以应对自然语言的多变性和复杂性 [2]。统计方法（如 TF-IDF、余弦相似度）通过词频或向量空间模型衡量文本相似度，虽部分缓解了语义差异问题，但对上下文

【作者简介】李东宇（2002–），男，中国北京人，本科，从事大数据应用和人工智能研究。

依赖性强、句式结构复杂的文本仍存在误判风险。例如，两篇讨论同一事件的新闻稿若采用不同叙事顺序或词汇表达，传统方法可能因缺乏语义关联分析而将其误判为独立文本。

2.2 机器学习方法的演进与优势

随着数据规模扩大与计算能力提升，机器学习技术逐渐成为解决文本去重与冗余过滤问题的新范式。监督学习方法通过标注数据训练分类模型，能够识别文本间的语义关联，但其依赖高质量标注数据的特性限制了泛化能力。无监督学习（如聚类算法、主题模型）能够自动挖掘文本内在模式，缓解了数据标注成本问题，但在精度上仍存在波动。深度学习的突破进一步推动了技术革新，词嵌入模型（如 Word2Vec、BERT）通过上下文感知的向量表示，显著提升了语义相似性度量的准确性；Transformer 架构的注意力机制则使模型能够捕捉长文本中的局部与全局关联。深度学习不仅能够处理显性重复，还能识别隐含的语义冗余，例如同一观点在不同语境下的多样化表述。

2.3 当前面临的核心技术挑战

尽管机器学习方法展现出巨大潜力，其实践落地仍面临多重挑战。语义相似性识别需平衡细粒度分析与计算效率。过于复杂的模型（如深层神经网络）能提升精度，却难以满足大规模实时处理需求。上下文依赖性要求模型具备动态适应能力，例如，同一词汇在不同领域（如医疗与金融）可能承载截然不同的含义，而现有方法常因领域迁移能力不足导致误判。此外，异构数据源的整合问题尤为突出，社交媒体短文本、学术长文本与多语言混合数据需差异化的处理策略，但多数框架尚未实现多模态特征的统一建模。

3 基于机器学习的去重与冗余过滤框架设计

3.1 分层处理框架的总体架构

为了应对文本去重与冗余过滤的复杂需求，本文提出一种分层处理框架，包含数据预处理、特征提取、相似性度量和动态优化四大核心模块。数据预处理模块负责原始文本的清洗、分词与标准化，消除噪声干扰并为后续分析提供结构化输入 [3]；特征提取模块融合词嵌入与句法结构特征，构建多维度的文本表示；相似性度量模块通过注意力机制与对比学习计算文本间关联强度，支撑去重与冗余判定；动态优化模块基于实时反馈调整模型参数，确保框架在数据分布变化时的鲁棒性。各模块以流水线方式串联，同时支持并行化处理，兼顾处理效率与精度。这一架构设计突破了传统单阶段处理的局限性，形成从粗粒度筛选到细粒度分析的递进式工作流。

3.2 多模态特征提取方法

特征提取是框架的核心环节，直接影响语义识别的准确性。传统方法依赖单一特征（如词频或词袋模型），难以全面捕捉文本的语义与结构信息。因此，本框架采用多模态特征融合策略。基于预训练语言模型生成上下文敏感的语义向量，解决一词多义与上下文依赖性难题，同时引入依存句法分析与实体识别技术，提取主谓宾结构和关键实体作为语法特征。例如，在“人工智能改变医疗”与“AI 技术革新健康领域”两句话中，词嵌入模型可识别“人工智能”与“AI”、“医疗”与“健康”的语义关联，而句法分析则通过主谓结构一致性进一步验证其冗余性。两类特征通过加权拼接形成联合表示，既保留语义深度又增强逻辑可解释性。

3.3 基于注意力机制的相似性度量

相似性度量模块需解决长文本匹配与局部冗余检测的双重需求。传统余弦相似度等方法因全局计算特性，易受无关段落干扰。本框架引入多头注意力机制，使模型能够动态聚焦于文本中的关键片段。具体而言，对两篇待比对文本的嵌入向量进行交叉注意力计算，生成交互矩阵以量化局部相似性，再通过池化操作提取矩阵中的显著信号，综合生成全局相似度得分。针对冗余过滤任务，进一步采用对比学习策略：构建正样本对（语义相同但表述差异的文本）与负样本对（语义无关文本），通过三元组损失函数约束模型区分细微差异。实验表明，该方法在长文档比对中可将误判率降低 18%，且对局部冗余（如重复段落嵌入不同上下文）的检测灵敏度提升 23%。

3.4 动态优化策略的设计与实现

静态模型难以适应数据流的动态变化，比如社交媒体中新热点的涌现或学术文献中术语体系的更新。为此，框架引入动态优化模块，包含在线学习与增量更新两套机制。在线学习通过滑动窗口实时采集最新数据，以最小批训练方式微调模型参数，避免全局重训练带来的计算开销。增量更新则针对概念漂移场景，利用 KL 散度检测特征分布变化，当差异超过阈值时触发模型结构自适应调整。以新闻数据为例，若某一事件的相关报道出现新的关键词变体（如“元宇宙”衍生出“虚拟空间 3.0”），动态优化模块可在 24 小时内完成模型迭代，确保语义识别能力持续适配最新语料。这一策略使框架在保持高精度的同时具备工业级场景所需的可扩展性。

3.5 框架的技术优势与适用性

相较于传统方法，本框架在性能与适用性上实现多重突破。效率方面，通过分层处理与并行计算优化，单日可处理千万级文本，较基于规则的系统提速 5 倍以上；精度方面，多模态特征与注意力机制的结合可显著提高 F1 值；兼容性上，框架支持跨领域多源数据混合输入，例如同时处理社交媒体短文本与科研论文长文本，并通过自适应特征权重分配减少领域偏差。不仅如此，动态优化机制使框架能够无缝接入实时数据流，为搜索引擎、智能客服等场景提供持续服务能力。这些特性不仅验证了机器学习驱动的技术路径的先进性，也为大规模文本处理工程提供了可复用的解决方案。

4 实验设计与数据集构建

4.1 多源异构数据集的构建

实验数据涵盖新闻、社交媒体、学术论文三大领域，以验证框架对多源文本的兼容性。新闻数据选自主流媒体近三年的公开报道，覆盖科技、经济、社会等多主题，社交媒体数据采集自微博与 Twitter 平台，包含用户评论、话题讨论等短文本，学术论文数据从 arXiv 开放获取库中筛选，涉及计算机科学与语言学领域的长文本。所有数据经过去标识化处理，人工标注重复与冗余标签，确保标注一致性。最终构建的数据集总量为 120 万条，其中重复文本占比 15%，冗余文本占比 28%，形成涵盖不同语言风格、文本长度与语义复杂度的基准测试集。

4.2 对比实验与参数配置

为全面评估框架性能，实验设置三组对比：传统方法组（SimHash、TF-IDF）、单一机器学习模型组（基于 BERT 的基线模型）以及本文提出的综合框架。传统方法采用默认参数，SimHash 指纹长度为 64 位，TF-IDF 选取 Top-500 关键词作为特征；BERT 基线模型使用预训练权重微调，训练周期为 10 轮，学习率设为 3e-5。综合框架中，词嵌入模块采用 RoBERTa-large 模型，句法特征通过 Stanford CoreNLP 工具提取，注意力机制层数为 4，动态优化模块的滑动窗口大小为 1000 条文本，增量更新阈值为 KL 散度 0.1。所有实验重复 5 次，取均值作为最终结果。

4.3 多维度的评价指标设计

性能评估围绕去重与冗余过滤两大任务展开。去重任务采用准确率、召回率与 F1 值衡量模型识别重复文本的能力；冗余过滤任务引入语义相似度阈值（设定为 0.85）判定冗余程度，计算精确率与误杀率。处理效率方面，记录单线程与分布式集群模式下的平均耗时，并统计内存占用峰值。为验证动态优化效果，额外设计时间衰减实验，向数据流中注入概念漂移样本（如新兴术语与话题），跟踪模型迭代后的性能恢复速度。这一指标设计确保评估既覆盖静态场景的稳定性，又反映动态场景的适应性。

5 实验结果与分析

5.1 去重与冗余过滤性能对比

实验对比了传统方法（SimHash、TF-IDF）、单一机器学习模型（BERT 基线）与本文框架的综合性能（表 1）。数据显示，本文框架在去重任务中的 F1 值达到 92.73%，较 SimHash（65.18%）和 BERT 基线（83.42%）分别提升 27.55% 与 9.31%；冗余过滤的精确率为 89.54%，误杀率控制在 6.12%，优于传统方法的粗粒度筛选与单一模型的波动表现。值得注意的是，本文框架的去重召回率显著高于其他方法，证明多模态特征对碎片化语义的捕捉能力。

表 1 不同方法在去重与冗余过滤任务中的性能对比（%）

方法	去重准确率	去重召回率	F1 值	冗余精确率	冗余误杀率
SimHash	68.24	62.15	65.18	59.76	21.37
TF-IDF	71.58	66.33	68.85	64.19	18.92
BERT 基线	85.17	81.74	83.42	78.91	13.45
本文框架	90.25	89.12	92.73	89.54	6.12

5.2 多源数据集下的冗余过滤效果

为验证框架的跨领域适应性，表 2 展示了冗余过滤在不同数据集上的表现。新闻数据的精确率最高（91.28%），因文本结构规范、语义明确；社交媒体数据误杀率略高（7.89%），主要源于网络用语的多义性与非正式表达；学术论文的冗余检测 F1 值达 87.41%，体现框架对长文本逻辑关联的解析能力。实验进一步发现，动态优化模块在概念漂移场景下（如突发事件报道）可使模型性能在 3 小时内恢复至基线水平的 95%，较静态模型（性能衰减 23%）显著提升鲁棒性。

表 2 冗余过滤在不同数据集上的性能对比（%）

数据集	精确率	召回率	F1 值	误杀率
新闻数据	91.28	88.45	89.83	5.67
社交媒体	86.19	82.17	84.12	7.89
学术论文	88.73	86.12	87.41	6.34

6 结语

本文提出的机器学习驱动框架，通过多模态特征融合与动态优化机制，为文本去重与冗余过滤提供了高效、精准的解决方案。未来研究需进一步探索超大规模数据下的轻量化部署、跨语言文本的泛化能力提升，以及多模态数据（如图文混合内容）的联合去重技术。此外，如何在隐私保护与数据效用间取得平衡，构建合规且高效的文本处理系统，将成为下一代技术革新的重要方向。

参考文献

[1] 王少凡.深度学习技术在自然语言处理中的应用[J].集成电路应用,2025,42(01):196-197.

[2] 申峻宇,李东闻,钟震宇,等.一种基于局部敏感哈希的文本数据去重算法及其实现[J].南开大学学报(自然科学版),2023,56(06):29-35.

[3] 陈少斌.基于预训练数据冗余信息过滤的句子表征学习研究[D].华东师范大学,2023.