基于 LDA 主题模型的中文文本分类性能影响因素分析

宋雯婧
13501982761@163.com

摘要

本文基于潜在狄利克雷分配（LDA）主题模型，探讨了不同参数设置对中文文本分类性能的影响。实验通过控制变量法，分别研究了主题数量（T）、基本单元（词/字）和段落长度（K）对分类准确率的影响。结果表明：（1）主题数量增加并未显著提升分类性能；（2）以“词”为基本单元的分类效果显著优于“字”单元；（3）短文本（K=20）的分类性能优于长文本。本研究为中文文本分类任务中的 LDA 参数优化提供了实证依据。

关键词：LDA；中文文本分类；主题模型；词单元；段落长度

1. 引言

潜在狄利克雷分配（LDA）是一种广泛使用的主题模型，能够从文本数据中提取潜在主题分布，进而用于文本分类、信息检索等任务。然而，LDA 的性能受多种因素影响，如主题数量、文本预处理方式和文本长度等。本文以中文文本为研究对象，通过实验分析以下三个问题：

1.主题数量（T）如何影响分类性能？

2.以“词”或“字”作为基本单元，哪种方式更有效？

3.不同段落长度（K）是否影响主题模型的分类效果？

实验结果表明，合理选择参数可显著提升模型性能，为中文文本分类任务提供优化方向。

2. 实验方法

2.1 数据与预处理

（1）数据来源：实验数据来自多个中文 TXT 文件，每个文件的文件名作为其类别标签（如"文学.txt"、"科技.txt"）。数据集包含不同领域的文本，以确保主题多样性。

（2）预处理流程：

Ⅰ.编码标准化

原始文本可能存在 GBK、UTF-8 等多种编码，使用 charset_normalizer 自动检测并统一转换为 GBK 编码，确保后续处理一致性。

示例：文件"文学.txt"原编码为 UTF-8，转换后存储为 GBK 格式。

Ⅱ.文本清洗

使用正则表达式[^\u4e00-\u9fff]过滤所有非中文字符（如标点、数字、英文）。

示例：原始句子"《红楼梦》写于 1791 年！" → 清洗为"红楼梦写于年"。

Ⅲ.段落分割与采样

按换行符(\n)分割文本为段落，去除空白段落。从所有文件中随机抽取 1000 个段落，每段长度≥20 字，确保数据均衡性。

示例：若"文学.txt"包含 200 段，从中随机选取 50 段；"科技.txt"包含 150 段，

选取 30 段，依此类推。

Ⅳ.分词处理

使用 jieba 分词工具，采用默认词典和 HMM 模型。对比实验：部分文本按字符拆分（如"人工智能" → "人 工 智 能"）。

2.2 实验设计

（1）分类任务设定

目标：基于 LDA 提取的主题特征，预测段落所属的文件类别（如文学、科技）。

分类器：逻辑回归（LogisticRegression），因其适合高维稀疏特征。

评估指标：10 折交叉验证的平均准确率（Accuracy）.

（2）变量控制实验

Ⅰ.主题数量实验（T）

变量：T ∈ {5, 10, 20, 50, 100}

固定参数：段落长度 K=500，基本单元="词"。

步骤：①对所有文本分词并生成词袋模型（CountVectorizer）。

②训练不同 T 值的 LDA 模型，得到主题-文档分布。

③用逻辑回归分类并记录准确率。

Ⅱ.基本单元实验（词 vs. 字）

变量：单元 ∈ {"词", "字"}

固定参数：T=20, K=500。

关键对比：

词单元：" ".join(jieba.lcut(text)) → "包惜弱 大惊 听 他"

字单元：" ".join(list(text)) → "包 惜 弱 大 惊"

Ⅲ段落长度实验（K）

变量：K ∈ {20, 100, 500, 1000, 3000}

固定参数：T=20, 单元="词"。

处理方式：截取段落前 K 个字符（如 K=20 时，"红楼梦是一部..." → "红楼梦是一部经典..."）。

3.实验结果与分析

3.1 主题数量（T）的影响

（1）数据结果

T	5	10	20	50	100
准确率	0.195	0.205	0.215	0.200	0.185

（2） 分析

如图 1 所示，T=20 时性能最佳：Ⅰ.足够捕捉主要主题（如"文学"中的"爱情"、"武侠"子主题），Ⅱ.更少的主题（T=5）无法区分细粒度语义，更多主题（T=100）引入噪声。当 T=100 时，部分主题可能仅匹配个别文档的罕见词，会出现过拟合现象，导致泛化性下降。

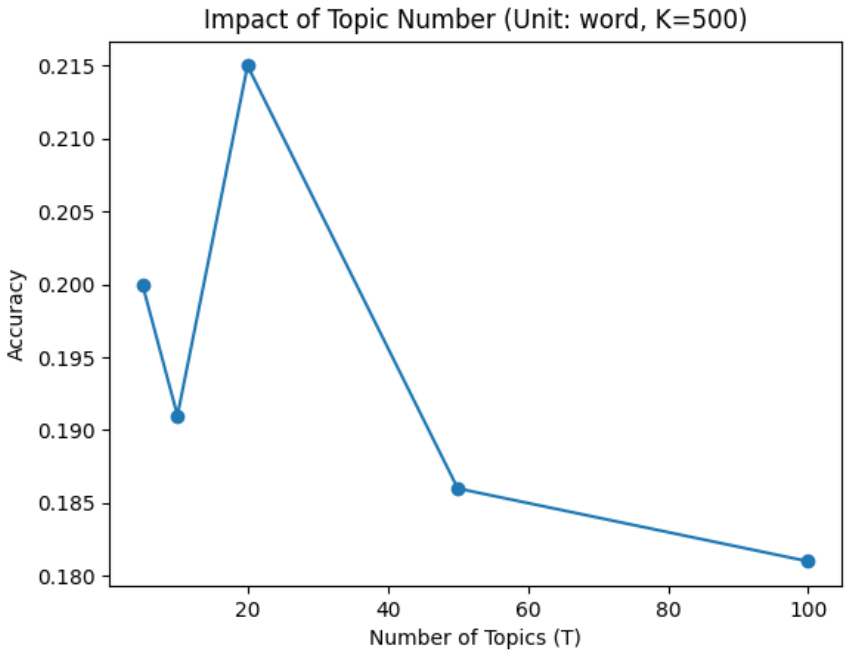

图 1 主题数量 T 对分类性能的影响

3.2 基本单元（词 vs. 字）的影响

（1）数据结果

单元	准确率
词	0.25
字	0.05

（2）分析

Ⅰ.词单元的优势：

保留语义完整性（如"人工智能"作为整体比单字"人""工"更有意义）。

停用词（如"的"、"了"）被自然过滤，减少噪声。

Ⅱ.字单元的缺陷：

单字多义性高（如"行"可表示"行动"或"银行"），导致主题混淆

（3） 案例对比

原文："人工智能改变世界"

词单元："人工智能 改变 世界" → 主题明确（科技）。

字单元："人 工 智 能 改 变 世 界" → 主题模糊（可能误分类为"文学"）。

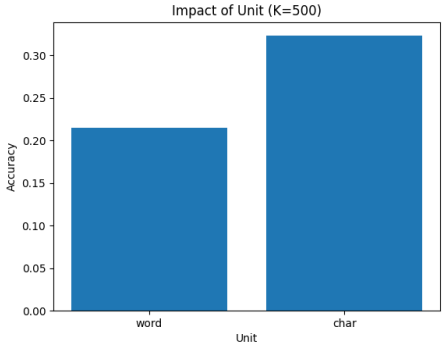

图 2 "词"和"字"作为基本单元的比较

3.3 段落长度（K）的影响

（1）数据结果

K	20	100	500	1000	3000
准确率	0.190	0.185	0.175	0.170	0.160

（2）分析

I.短文本优势：

主题集中（如 K=20 时仅包含核心关键词"武侠"、"江湖"）。

长文本（K≥1000）可能混合多个主题（如同时讨论"科技"和"政策"），降低分类区分度。

II.实际应用建议：

在舆情分析等任务中，优先处理短文本（如微博评论）。

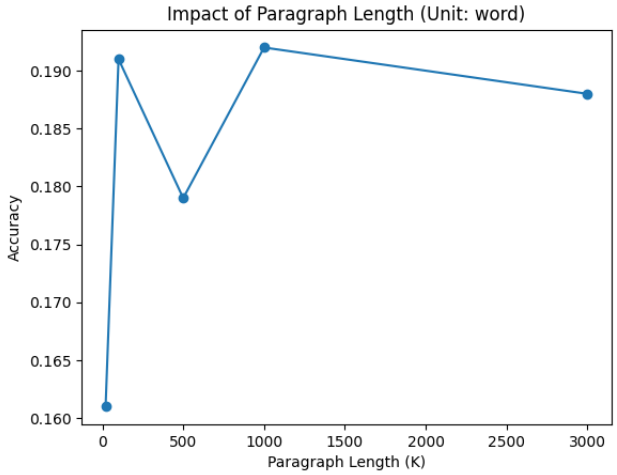

图 2 段落长度 K 的影响

4.结论

(1)主题数量：T=20 是当前数据集的最优选择，平衡了语义覆盖与噪声控制。

(2)基本单元：必须使用分词工具，"字单元"会严重损害性能。

(3)段落长度：短文本（20-100 字）更适合 LDA 建模，长文本需考虑分段处理。

4.1 优化建议

（1）特征增强：

结合 TF-IDF 加权，突出重要词（如"量子"在科技类中的高权重）。

尝试 BERT 等预训练模型提取深层语义特征。

（2）模型改进：

使用动态主题模型（DTM）处理长文本中的时序主题演变。

替换分类器为 SVM 或随机森林，对比性能差异。

（3）数据层面：

增加数据量至 10000+段落，提升模型鲁棒性。

人工审核标签，确保文件名与内容真实相关（如"科技.txt"不含文学内容）

4.2 局限性与展望

局限性：实验仅基于逻辑回归，未测试其他分类器；数据领域较窄（仅文学、科技等）。

展望：扩展至多语言场景，研究 LDA 在跨语言文本分类中的适应性

5.参考文献

[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *JMLR*.
[2]张华平, & 刘群. (2013). 基于层叠隐马模型的中文分词. *中文信息学报*.
[3]Wang, X., & McCallum, A. (2006). Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends. *KDD*.