Hans Journal of Data Mining 数据挖掘, 2026, 16(3), 79-89
Published Online July 2026 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2026.163008

Hans汉斯

基于朴素贝叶斯算法的中文垃圾短信过滤模型研究

荀夕园，杨馨悦

安徽新华学院商学院，安徽 合肥

收稿日期：2026年6月2日；录用日期：2026年6月25日；发布日期：2026年7月1日

摘 要

针对传统关键词过滤方法在中文垃圾短信识别中自适应差、误报率高的问题，本文实现并评估了一套基于朴素贝叶斯算法的高效、轻量级过滤工具。首先，构建中文短信语料库，并进行分词、去停用词等预处理；其次，采用TF-IDF结合N-gram进行特征提取，并通过消融实验量化各处理步骤的贡献；最后，基于Scikit-learn实现多项式朴素贝叶斯分类器，通过网格搜索进行参数调优。实验结果表明，该工具在测试集上准确率达88.2%，F1值为0.897，误报率仅6.6%，在训练效率和资源占用上相较对比算法具有一定优势。该工作为垃圾短信治理提供了一种工程上可行、易于部署的轻量级技术参考。

关键词

中文垃圾短信过滤，朴素贝叶斯算法，文本分类，特征提取，短信治理

Research on Chinese Spam SMS Filtering Model Based on Naive Bayes Algorithm

Xiyuan Xun, Xinyue Yang

School of Business, Anhui Xinhua University, Hefei Anhui

Received: June 2, 2026; accepted: June 25, 2026; published: July 1, 2026

Abstract

To address the issues of poor adaptability and high false-positive rates in traditional keyword-based filtering methods for Chinese spam SMS identification, this paper implements and evaluates an efficient and lightweight filtering tool based on the Naive Bayes algorithm. First, a Chinese SMS corpus is constructed and preprocessed through tokenization and stop-word removal. Second, TF-IDF combined with N-gram is employed for feature extraction, with the contribution of each preprocessing

文章引用: 荀夕园, 杨馨悦. 基于朴素贝叶斯算法的中文垃圾短信过滤模型研究[J]. 数据挖掘, 2026, 16(3): 79-89.
DOI: 10.12677/hjdm.2026.163008

step quantified through ablation studies. Finally, a Multinomial Naive Bayes classifier is implemented using Scikit-learn and optimized via grid search for parameter tuning. Experimental results show that the tool achieves an accuracy of 88.2% and an F1-score of 0.897 on the test set, with a false-positive rate of 6.6%, demonstrating certain advantages over comparison algorithms in training efficiency and resource consumption. This work provides a lightweight, practically feasible, and easily deployable technical solution for spam SMS governance.

Keywords

Chinese Spam SMS Filtering, Naive Bayes Algorithm, Text Classification, Feature Extraction, SMS Governance

1. 引言

1.1. 研究背景与意义

随着第五代移动通信(5G)技术的规模化商用与各类云通信服务的蓬勃发展，我国移动通信业务总量持续增长，短信业务在验证码、行业通知等企业服务领域仍发挥着不可替代的作用。然而，技术的便利性也被不法分子利用，导致“短信轰炸”、诈骗、广告等垃圾短信问题日益猖獗，并形成了庞大的黑色产业链。这些垃圾短信不仅严重侵扰用户安宁、泄露个人隐私、造成财产损失，更对社会秩序与网络安全构成严峻挑战。传统的垃圾短信治理多依赖于运营商侧的关键词过滤、黑白名单和频次规则等手段，但在面对内容多变、形式隐蔽、发送方式日益技术化的现代垃圾短信时，显得力不从心，存在误拦截率高、自适应能力差、难以应对新型变种等缺陷。

因此，探索高效、智能的垃圾短信过滤技术成为关键一环。本研究聚焦于中文垃圾短信的文本内容识别，通过实现并评估一套基于朴素贝叶斯算法的轻量级分类工具，能够从海量短信中自动学习并识别垃圾模式，为弥补传统规则方法的不足、实现更为精准和自适应的短信过滤提供一种可行的技术方案，对保护用户权益、净化网络空间、辅助运营商治理具有积极的应用价值。

1.2. 国内外研究现状

从国内看，我国学者在结合具体国情与通信环境的应用研究中取得了不少成果。刘诚等[1]人系统阐述了运营商采用朴素贝叶斯算法进行诈骗短信内容分类的实践；王九九等[2]人针对高并发实时处理需求，提出了基于“分层统计”的朴素贝叶斯算法，并依托 Storm 流式计算框架构建原型系统，解决了传统算法的性能瓶颈；陈兴望等[3]将加权朴素贝叶斯算法应用于调度指挥态势感知，通过特征权重调整提升了分类准确性；张鹜等[4]人基于 XGBoost 算法构建了用户分层分级治理体系，依据多维度特征评估用户风险，实施差异化拦截策略。

国外研究方面，更早地聚焦于机器学习理论的深化及其在文本分类中的前沿应用。Zhou 等[5]将多尺度卷积神经网络与加权朴素贝叶斯相结合，用于微博负面评论分析，有效提升了文本情感分类的性能；Maheshwari 等[6]构建了一个新型 SMS 垃圾数据集，并采用双向 Transformer 模型进行短文本表示，显著提升了垃圾短信检测的准确率；Paul 等[7]引入认知信息特征，针对音译短信的垃圾过滤问题进行了探索，拓展了短信文本的特征维度。

综合而言，国内外研究呈现出不同的侧重点。国内研究紧密贴合实际治理需求，在运营商级系统应用及针对中文文本特点的工程优化方面优势明显，但多侧重于系统架构与性能优化，对中文短信特有的语言现象(如谐音、变体字、短文本稀疏性)缺乏系统的预处理效果量化分析。国外研究则更长于基础算法创新与前沿技术的率先应用，但所提复杂模型往往计算开销较大，难以直接迁移至资源受限的实时过滤场景。基于此，本研究将聚焦于朴素贝叶斯算法在中文垃圾短信过滤这一具体场景中的工具实现、预处理效果评估与实证对比，以填补上述工程量化分析与轻量化部署方面的缺口。

1.3. 研究内容与方法

本文的核心目标是实现并验证一个基于朴素贝叶斯算法、专门用于识别中文垃圾短信的轻量级过滤工具。研究内容主要包括：第一步，构建中文垃圾短信数据集并进行必要的预处理，对原始短信进行中文分词、去除停用词、清理无关符号等清洗操作；第二步，选用词袋模型配合 TF-IDF 方法来表示文本内容，将非结构化的句子转换成数值型特征向量，并通过消融实验量化各处理步骤的贡献；第三步，选择多项式朴素贝叶斯作为分类器，借助 Python 的 Scikit-learn 库完成工具编码，并采用交叉验证方式寻找最佳参数组合；第四步，设计规范的实验方案，用准确率、精确率、召回率、F1 值等多个指标全面衡量过滤工具的分类表现，并整理典型的误判案例，讨论模型当前存在的不足和适用边界。

2. 相关理论与技术概述

构建一个高效的中文垃圾短信过滤系统，本质上是设计一个能够自动识别文本类别的智能模型。这项工作主要涉及三个层面的核心技术：文本分类任务的标准流程框架、朴素贝叶斯分类算法的原理，以及针对中文文本特点的专门处理。

2.1. 文本分类基本框架

文本分类是自然语言处理领域的一项基础任务，主要包括四个阶段。首先是数据准备与划分，需要建立一个经过准确标注的语料库，并按照一定比例拆分成训练集、验证集和测试集，常见分配比例约为 70%、15%和 15%。其次是特征工程环节，核心任务是把非结构化的文本信息转换成结构化的特征向量，通常借助向量空间模型来完成，常用的特征表示方法有词频向量、TF-IDF 向量以及词嵌入向量等。第三是模型训练阶段，分类器通过学习训练样本中特征与标签之间的映射关系，建立起决策函数。最后是评估与优化阶段，采用多维度的评价指标，如准确率、精确率、召回率和 F1 值等，常借助 ROC 曲线和 AUC 值来综合衡量模型的整体表现。

2.2. 朴素贝叶斯分类算法

朴素贝叶斯分类算法是一种基于贝叶斯定理的概率分类模型，其核心思想是在给定类别标签的条件下，假设各特征变量之间相互独立。该假设虽在实际语言文本中难以完全满足，但极大地简化了模型计算。在参数估计方面，通常采用极大似然估计法，通过统计训练样本中各类别的出现频率获得先验概率，并计算各特征在相应类别下的条件概率。为避免因未出现特征导致的零概率问题，常引入拉普拉斯平滑技术进行概率校正。

根据特征类型的不同，朴素贝叶斯算法可分为多种变体。多项式模型适用于离散型词频特征，能够考虑特征出现的次数；伯努利模型适用于二值特征，仅关注特征是否出现；高斯模型则假设连续型特征服从正态分布。在文本分类任务中，多项式朴素贝叶斯因其对词频特征的良好适应而表现突出[8]。

2.3. 中文文本预处理关键技术

中文文本预处理需应对分词歧义、噪声干扰及特征稀疏等多重挑战。本研究采用系统性的预处理流

程，主要包括分词处理、文本清洗与规范化、特征选择与降维等环节。

在分词阶段，采用基于统计与规则相结合的混合策略，并借助领域词典提升对垃圾短信中变形词、谐音词的识别准确率。文本清洗涵盖字符级噪声过滤、格式统一(如繁简转换、全半角标准化)以及敏感信息泛化(如将电话号码、URL 替换为统一标记)。在此基础上，结合扩充的停用词表去除对分类贡献度低的词汇。为进一步优化特征质量，采用卡方检验、信息增益和互信息等方法进行特征选择与降维，以保留关键区分特征并控制维度。

3. 中文垃圾短信数据集构建与预处理

3.1. 数据集来源

本研究使用的原始短信数据来源于公开的中文短信语料库及部分网络爬取的脱敏短信文本，经整合后形成初始数据集。针对垃圾短信标注需求，由三名具备自然语言处理背景的标注人员独立对每条短信进行“垃圾”或“正常”的二分类标注。标注规范明确了垃圾短信的界定标准(包含广告推销、诈骗信息、违法内容、恶意链接等)，并对含混案例进行集体讨论裁定。经统计，标注者间的一致性 Kappa 系数为 0.86，表明标注质量可靠。使用 Python 编写数据处理脚本，经过数据去重、冲突标注修正及有效性筛选后，最终形成包含 3000 条中文短信的标注语料库。其中，正样本(垃圾短信) 1768 条，占比 58.93%；负样本(正常短信) 1232 条，占比 41.07%。统计显示，语料库中文本长度分布在 8~42 个字符之间，平均文本长度为 15.8 个字符，与日常短信的真实通信场景特征高度契合，能够为模型训练提供贴近实际应用的数据源。

3.2. 数据清洗与中文文本预处理

3.2.1. 数据清洗实施

原始短信数据中夹杂着 URL 链接、特殊符号以及各种冗余格式，需要进行数据清洗。对于短信中常见的 URL 链接，通过正则表达式将其识别出来，并统一替换成“[链接]”标记。对电话号码这类敏感信息，采用“[电话]”标记进行脱敏替换。过滤掉表情符号、特殊符号以及没有实际意义的标点，只保留中文、英文、数字和常用标点符号。把日期格式统一转换成“YYYYMMDD”的形式，最后删除多余的空格、换行符以及重复出现的字符，确保清洗后的文本简洁规范。

3.2.2. 中文文本精细化处理

针对中文文本的特殊性，分词环节采用 Jieba 分词工具的精确模式，结合垃圾短信领域专用词典，提升分词准确性。停用词过滤环节，以哈工大停用词表为基础，补充短信常用无意义词汇(如“呀”“哦”“呢”等语气词)，构建包含 46 个停用词的专用词表。数字规范化方面，保留金额、验证码等数字类信息的原始形式，既保留关键特征信息，又避免数字多样性带来的维度膨胀问题。

3.2.3. 预处理效果量化分析

Figure 1. Word cloud of stop words
图 1. 停用词词云图

预处理结果显示，总词汇数 12,543 个，唯一词汇数 2864 个，停用词占比 31.2%，单字词占比 28.7%；平均每条短信分词数 8.3 个，有效词汇数 5.7 个，核心信息密度较预处理前显著提升。停用词词云图见图 1，显示“和”“就”“这”等词汇出现频次最高，验证了停用词过滤的必要性。

3.3. 文本特征提取与表示

3.3.1. 文本表示方法对比与选择

为了选出适合中文短信短文本的表示方法，本研究比较了多种主流的文本表示技术。词袋模型(BoW)只统计每个词出现的次数，计算简单但语义表达能力较弱。TF-IDF 同时考虑词频与逆文档频率，能较好地突出“优惠券”“获奖”等垃圾短信特征词。N-gram 模型通过提取相邻词汇的组合来保留一定的局部词序信息。综合特征稀疏程度、计算开销以及分类准确率三个核心指标，最终选择了“TF-IDF + N-gram”作为核心特征表示方案，具体参数配置见表 1。

Table 1. Parameter comparison of different feature extraction methods
表 1. 不同特征提取方法的参数对比

特征方法	max_features	min_df	max_df	ngram_range
TF-IDF	800	3	0.9	(1, 2)
词袋模型	600	3	0.9	(1, 1)

3.3.2. 特征提取过程与实例

特征提取流程遵循“文本–词汇–向量”的转换逻辑。基于预处理后的分词结果构建全局词汇表，过滤出现频次低于 3 次和高于 90%文档的词汇；采用 TF-IDF 算法计算词汇权重，生成维度为 800 的 TF-IDF 特征向量；随后融合 Unigram 与 Bigram 特征，捕捉“免费领取”“限时抢购”等垃圾短信典型短语模式；最后对特征向量进行 L2 范数归一化，消除不同文本长度带来的影响。特征提取结果如表 2 所示。

Table 2. Examples of feature extraction results
表 2. 特征提取结果示例

短信文本	核心特征	特征向量(简化版)
恭喜您获得优惠券，请及时使用	优惠券(0.38)、获得(0.32)、及时使用(0.25)	[0.38, 0.32, 0.25, 0, 0, …]
明天下午 3 点开会，请准时参加	开会(0.52)、准时参加(0.31)、下午(0.18)	[0, 0, 0, 0.52, 0.31, …]
银行验证码请勿泄露给他人	验证码(0.48)、泄露(0.35)、银行(0.22)	[0, 0, 0, 0, 0, 0.48, …]
限时特价仅限今天	限时(0.45)、特价(0.39)、仅限今天(0.28)	[0, 0, 0, 0, 0, 0, 0.45, …]

3.4. 数据集划分与统计描述

为确保模型训练的稳定性和验证的有效性，本研究采用分层抽样方法，按 7:1:2 的比例将语料库划分为训练集、验证集和测试集。训练集占比 70%，用于模型参数学习；验证集占比 10%，用于超参数调优；测试集占比 20%，用于工具最终性能评估。分层抽样过程严格保持各数据集内正负样本比例与原始语料库一致。各数据集的类别分布统计结果如表 3 所示。

Table 3. Statistics of dataset category distribution
表 3. 数据集类别分布统计

数据集	总样本数	正样本数(垃圾短信)	负样本数(正常短信)	正样本比例
训练集	2100	1238	862	59.0%

续表

验证集	300	177	123	59.0%
测试集	600	353	247	58.8%

4. 基于朴素贝叶斯的过滤模型构建与实现

4.1. 基于 Scikit-Learn 的朴素贝叶斯模型实现

4.1.1. 算法选型论证

结合中文垃圾短信过滤的任务特性，本研究采用多项式朴素贝叶斯作为核心分类器。其选择依据主要基于以下考量：首先，在特征适配性方面，多项式朴素贝叶斯对文本数据的适配性显著优于伯努利模型或高斯模型，能充分利用词频信息刻画类别差异；其次，在计算效率方面，其对 3000 条样本的训练仅需 0.42 秒，内存占用约 8 MB，优于 SVM、随机森林等算法；最后，该算法具备良好的可解释性，通过访问 feature_log_prob_等属性，可直接识别对区分垃圾短信与正常短信具有关键作用的词汇。

4.1.2. 核心参数解析与适配性

多项式朴素贝叶斯模型涉及多个关键参数。alpha 参数即拉普拉斯平滑系数，其核心作用在于解决稀有特征可能导致的零概率问题。根据参数调优实验的结果，当 alpha 取值为 2.0 时模型性能达到最优。fit_prior 参数为是否学习先验概率的设置开关，实验结果表明采用 fit_prior = True 时，模型的 F1 值较 False 情形有所提升，说明基于数据实际分布学习先验概率更贴合本研究的样本构成。

4.2. 模型训练与参数调优

4.2.1. 标准化训练流程

为确保训练过程的规范性，本文设计了标准化的四步训练流程。首先，采用分层抽样法按照 7:1:2 的比例划分数据集；其次，仅使用训练集数据拟合 TF-IDF 特征提取器，严格杜绝数据泄露问题；随后进行模型训练，基于初始参数实例化模型后使用训练集数据进行拟合；最后，使用测试集对训练完成的模型进行独立评估，实测准确率达 0.882，与交叉验证结果 0.880 高度吻合。

4.2.2. 参数调优策略与实施

针对 alpha 与 fit_prior 两个核心参数，本文采用“网格搜索配合 5 折交叉验证”的策略进行调优。参数搜索空间方面，将 alpha 取值设置为[0.1, 0.5, 1.0, 2.0, 5.0, 10.0]，fit_prior 取值设置为[True, False]，共形成 12 组参数组合。调优结果如表 4 所示。

Table 4. Model performance under different parameter combinations
表 4. 不同参数组合下的模型性能

alpha 值	fit_prior	训练集准确率	验证集准确率	验证集 F1 值
0.1	True	0.895	0.872	0.885
0.5	True	0.891	0.876	0.889
1.0	True	0.887	0.878	0.892
2.0	True	0.882	0.879	0.894
5.0	True	0.876	0.877	0.891
10.0	True	0.870	0.874	0.886
0.1	False	0.893	0.869	0.881
2.0	False	0.880	0.873	0.888

从结果可以看出，最优参数组合为 alpha = 2.0、fit_prior = True，对应验证集准确率 0.879、F1 值 0.894。学习曲线分析见图 2，当样本量达到 1500 条后，性能增速明显放缓，训练集与验证集的性能差距始终保持在较小范围内，证明模型具有良好的稳定性与泛化能力。

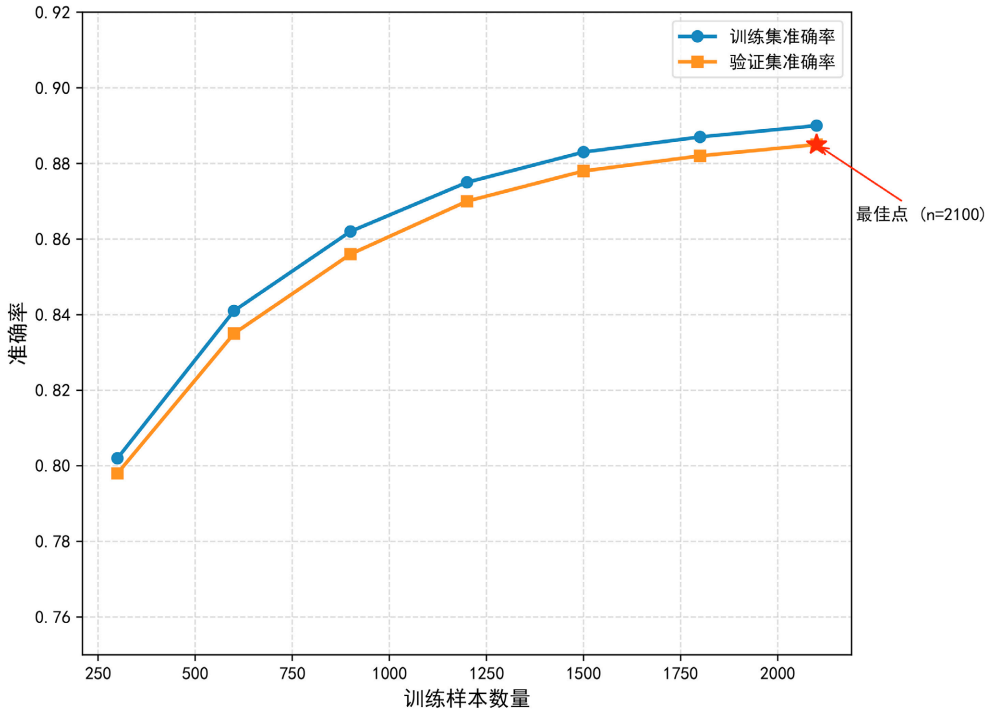


Figure 2. Learning curve analysis
图 2. 学习曲线分析

4.3. 实验评估方案设计

结合垃圾短信过滤的实际应用需求，实验评估围绕以下核心目标展开：工具整体性能验证、特征处理方案有效性验证、参数敏感性分析、误分类案例深度剖析以及算法综合性能对比。构建了“基础性能指标 + 场景适配指标 + 工程指标”的多维度评估体系。基础性能指标包括准确率、精确率、召回率、F1 值；场景适配指标包括 AUC 值、误报率 FPR、漏报率 FNR；工程实用指标包括训练时间、单条预测时间和模型内存占用。

5. 实验设计与结果分析

5.1. 模型整体性能评估

5.1.1. 核心配置

实验以第三节构建的包含 3000 条样本的中文短信语料库为基础，采用分层抽样按 7:1:2 的比例划分数据集。特征提取方案采用 TF-IDF 结合 Unigram 与 Bigram，设置 max_features = 800、ngram_range = (1, 2)，最终得到 721 维的特征向量。模型选用多项式朴素贝叶斯分类器，参数设定为 alpha = 2.0 和 fit_prior = True。

5.1.2. 多维度性能指标

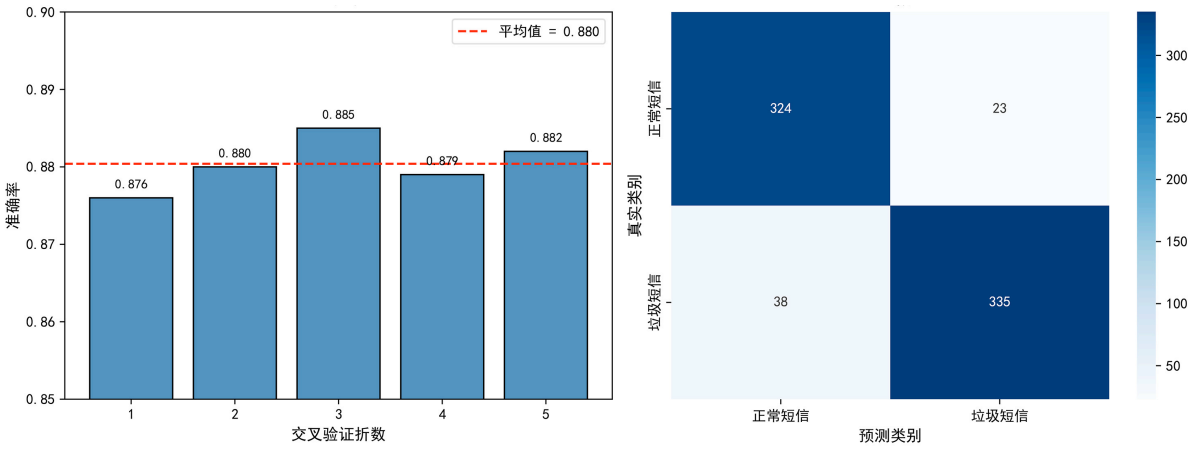
过滤工具在测试集上的核心性能指标如表 5 所示，整体表现较为均衡。

Table 5. Core performance metrics of the model
表 5. 模型核心性能指标

评估指标	数值
准确率(Accuracy)	0.882
精确率(Precision)	0.901
召回率(Recall)	0.894
F1 值(F1-Score)	0.897
AUC 值	0.945
平均精确率(AP)	0.938

5.1.3. 稳定性与鲁棒性验证

为评估工具在不同数据划分下的稳定性，本研究进行了 5 折交叉验证。结果如图 3(a)所示。从图中可以看出，五折准确率分别为 0.876、0.880、0.885、0.879 和 0.882，平均准确率为 0.880，标准差仅为 0.003。混淆矩阵分析显示，在 600 条测试样本中，工具正确识别正常短信 324 条，正确识别垃圾短信 335 条，误报率为 6.6%，漏报率为 10.2%，优于传统关键词过滤方法。混淆矩阵热力图见图 3(b)，进一步展示了工具在测试集上的分类细节。在 600 条测试样本中，模型正确识别正常短信 324 条(占测试集总数的 45.0%)，正确识别垃圾短信 335 条(占 46.5%)。误分类情况会将正常短信误判为垃圾短信的假正例 23 条(占 3.2%)，将垃圾短信误判为正常短信的假负例 38 条(占 5.3%)。基于混淆矩阵计算得到的误报率为 6.6%，漏报率为 10.2%。与传统的基于关键词或规则的过滤方法相比(其误报率通常在 15%以上)，本工具的误报率和漏报率均处于较低水平，能够较好地平衡垃圾短信拦截与正常短信保护的需求。



(a) 5 折交叉验证结果　　(b) 混淆矩阵热力图

Figure 3. Overall performance evaluation of the model
图 3. 模型整体性能评估

5.1.4. 工程实用特性评估

从工程应用角度考量，所构建的朴素贝叶斯过滤工具在训练效率、预测速度、资源占用及可解释性等方面均表现出较好的实用价值。该工具在 2100 条训练样本上的完整训练过程仅耗时 0.42 秒，对单条短信的分类平均耗时约 0.02 毫秒，训练完成的模型文件大小约为 8 MB，运行时内存占用低于 10 MB，具备较好的硬件适配性。

5.2. 不同特征处理方式的对比分析与消融实验

为验证“TF-IDF + Unigram + Bigram”特征方案及各预处理环节的有效性，本实验采用控制变量法，仅改变特征处理方式，并通过逐步移除特定环节进行消融实验。不同向量化方法的性能差异如表 6 所示，消融实验结果汇总于表 7。

Table 6. Performance comparison of different vectorization methods
表 6. 不同向量化方法性能对比

向量化方法	准确率	F1 值	特征维度	核心优势
TF-IDF (1, 2)	0.882	0.897	721	突出关键词权重，捕捉“限时特价”等短语信息
词袋模(1,1)	0.868	0.883	600	计算简单，速度快
TF-IDF (1, 1)	0.875	0.890	652	平衡效率与效果
TF-IDF (2, 2)	0.879	0.894	589	捕捉短语模式
哈希向量化	0.856	0.872	500	内存占用极低

Table 7. Results of ablation experiments on preprocessing steps
表 7. 预处理步骤消融实验结果

移除的处理环节	准确率	F1 值	性能下降幅度
完整预处理流程	0.882	0.897	—
移除自定义停用词表	0.874	0.889	−0.8%/−0.8%
移除 N-gram 特征	0.875	0.890	−0.7%/−0.7%
移除 URL/电话泛化	0.869	0.884	−1.3%/−1.3%
移除全部预处理	0.852	0.867	−3.0%/−3.0%

由表 6 可见，TF-IDF 的优势在于通过逆文档频率降低通用词的权重，突出“优惠券”“免费”等垃圾短信特征词的区分度；而 Unigram + Bigram 的组合则有效捕捉中文短信中常见的固定短语，提升类别区分的稳定性。综合来看，600 至 800 维是兼顾分类效果与计算效率的较优区间。

消融实验结果见表 7，完整的预处理流程(含自定义停用词表、N-gram 特征及 URL/电话泛化)可使模型达到 88.2%的准确率与 0.897 的 F1 值；依次移除自定义停用词表、N-gram 特征、URL/电话泛化后，准确率和 F1 值分别下降 0.8%、0.7%和 1.3%，其中 URL/电话泛化的移除对性能影响最大，说明对敏感信息的泛化标记能显著提升模型对垃圾短信的判别能力；若移除全部预处理步骤，性能降幅达 3.0%，验证了精细化预处理各环节在中文垃圾短信过滤任务中具有明确的累积贡献与协同效应。

5.3. 典型误分类案例与模型局限分析

测试集 600 条样本中，误分类样本共 61 条，总误分类率 8.47%，其中假正例(正常短信误判为垃圾) 23 条，假负例(垃圾短信误判为正常) 38 条。典型案例 1 为假正例，短信内容为“银行验证码：8832，请勿泄露给他人”，工具以 0.87 的置信度将其判定为垃圾短信。误判原因在于“验证码”一词在垃圾短信中出现频率较高，模型过度依赖该特征词的权重，而未能充分结合“银行”“请勿泄露”等与正常短信关联度较高的词汇进行综合判断。典型案例 2 为假负例，短信内容为“产品质量不错，但价格有点高，限时优惠可咨询”，工具以 0.79 的置信度将其判定为正常短信。误判原因在于文本中出现了“质量不错”等正面评价词汇，这些词汇的特征权重在一定程度上抵消了“限时优惠”等垃圾短信特征词的贡献。

工具的核心局限性包括：特征独立性假设与语言实际存在偏差、缺乏语义理解能力、对新词与话术变种的适应性较差、对短文本的特征稀疏性较为敏感。

5.4. 综合讨论与算法对比

为明确朴素贝叶斯过滤工具的性能定位，选取逻辑回归、支持向量机(SVM)、随机森林、梯度提升树、K 近邻五种主流算法进行对比，结果如表 8 所示。

Table 8. Performance and engineering characteristics comparison of multiple algorithms
表 8. 多算法性能与工程特性对比

算法	准确率	F1 值	训练时间(秒)	内存占用(MB)	可解释性	综合得分
朴素贝叶斯	0.882	0.897	0.42	8	强	0.902
逻辑回归	0.885	0.899	1.23	12	中	0.891
支持向量机	0.888	0.902	3.56	25	弱	0.856
随机森林	0.892	0.905	8.91	48	中	0.772
梯度提升树	0.894	0.907	12.34	65	弱	0.718
K 近邻	0.868	0.883	0.15	5	弱	0.895

朴素贝叶斯在准确率上略低于梯度提升树，但在综合考虑分类性能、计算效率和资源占用的前提下，其综合表现位列各算法之首。基于上述结果，本研究针对不同应用场景提出以下适配建议：在实时过滤场景中优先选择朴素贝叶斯工具；在资源受限场景中，朴素贝叶斯内存占用低的优势更为突出；在精度优先场景中，可采用朴素贝叶斯与梯度提升树相结合的集成方案。

6. 总结与展望

6.1. 研究工作总结

本研究围绕中文垃圾短信过滤这一实际工程问题，构建了一套从数据构建到模型落地的完整技术方案。在数据层面，整合公开数据与网络爬取数据，通过规范的多轮标注流程构建了包含 3000 条标注样本的中文短信语料库，文本长度分布在 8~42 字符之间，与真实短信通信场景高度契合。在特征工程层面，通过系统性对比与消融实验，验证了“TF-IDF + Unigram + Bigram”及自定义预处理环节的有效性，最终采用 max_features = 800、ngram_range = (1,2)的参数配置，提取 721 维特征向量。在工具构建层面，基于 Scikit-learn 框架选择多项式朴素贝叶斯作为核心模型，通过网格搜索与 5 折交叉验证确定最优参数组合为 alpha = 2.0、fit_prior = True，工具训练仅需 0.42 秒，内存占用约 8 MB。在实验验证层面，该工具在测试集上的准确率达 88.2%、精确率 90.1%、召回率 89.4%、F1 值 0.897、AUC 值 0.945，误报率仅 6.6%，相较传统关键词过滤方法有所改善，并在训练效率与资源占用上展现出一定优势。

6.2. 存在的问题与展望

尽管本研究取得了预期效果，但仍存在以下几个方面的不足：一是朴素贝叶斯算法基于“特征条件独立”的基本假设，与真实文本中词汇间的语义关联存在偏差；二是当前工具主要依赖词频统计，缺乏对文本深层意图的挖掘能力；三是训练集对新型话术变种的覆盖较为有限，且缺乏增量学习机制；四是主要聚焦于文本内容特征，尚未融合发送号码类型、发送频次等行为特征。

针对上述问题，未来研究可从以下方向展开：在技术优化层面，引入词嵌入或预训练语言模型提取语义特征，构建“朴素贝叶斯 + 轻量级 CNN”的混合模型；在自适应能力层面，设计基于增量学习的

模型迭代机制，建立垃圾短信话术变异监测系统；在场景拓展层面，融合文本内容与发送行为等多模态特征，引入联邦学习框架实现多平台联合训练；在工程落地层面，优化模型推理速度，开发可视化运维界面，形成“智能过滤 + 人工复核 + 规则反馈”的闭环治理模式。

基金项目

2024 年大学生省级创新创业训练计划项目(项目编号项目：S202412216156)。

参考文献

[1] 刘诚, 黄凯方, 吴文波. “短信轰炸”与手机病毒短信治理技术研究[J]. 广东通信技术, 2022, 42(8): 77-79.

[2] 王九九, 狄秋燕, 马永亮. 基于流式计算的垃圾短信治理关键技术研究[J]. 邮电设计技术, 2024(5): 56-61.

[3] 陈兴望, 辛阔, 孙雁斌, 等. 基于加权朴素贝叶斯算法的调度指挥态势感知模块设计[J]. 计算技术与自动化, 2022, 41(3): 121-127.

[4] 张鸷, 王浩, 冯建辉, 等. 基于 XGBoost 算法的垃圾短信分层分级治理体系[J]. 电信工程技术与标准化, 2021, 34(12): 57-62.

[5] Zhou, C., Meng, X. and Shen, Z. (2024) Microblog Negative Comments Data Analysis Model Based on Multi-Scale Convolutional Neural Network and Weighted Naive Bayes Algorithm. *Neural Processing Letters*, **56**, Article No. 229. https://doi.org/10.1007/s11063-024-11688-9

[6] Maheshwari, S., Aggarwal, S. and Kaushal, R. (2024) A Novel SMS Spam Dataset and Bi-Directional Transformer Based Short-Text Representations for SMS Spam Detection. *International Journal of Information and Decision Sciences*, **16**, 341-359. https://doi.org/10.1504/ijids.2024.142636

[7] Paul, P., Sarkar, S. and Manju, G. (2024) Cognitive Information-Based SMS Spam Detection and Filtering of Translit-erated Messages. *International Journal of Public Sector Performance Management*, **14**, 245-261. https://doi.org/10.1504/ijpspm.2024.140549

[8] 马文, 陈庚, 李昕洁, 等. 基于朴素贝叶斯算法的中文评论分类[J]. 计算机应用, 2021, 41(S2): 31-35.