跨文化内容生成模型优化与实践——基于 BERT-Crosslingual 的应用

秦雅欣 周晟竹

黑龙江外国语学院，黑龙江哈尔滨，150500；

摘要：针对中国企业出海跨文化内容本土化适配不足的核心痛点，本研究聚焦 BERT-Crosslingual 模型的优化与实践应用。通过构建 50 万条多语种跨文化语料库，融入人工标注的文化禁忌与语言变体信息，对模型进行跨文化适配微调，开发出支持多场景的跨文化内容生成系统。实证测试表明，优化后的模型文化误判率降至 4.7%，支持 15 种英语变体生成，文案本土化适配满意度达 89.6%，较传统翻译与生成工具效率提升 8 倍，成本降低 60%。该模型在巴西新能源汽车市场试点应用中，成功产出符合当地文化语境的西班牙语、葡萄牙语营销内容，有效降低文化冲突风险，为企业出海跨文化内容生产提供了高效可行的技术方案，同时为“语言+技术”跨学科融合提供了实践范例。

关键词：BERT-Crosslingual；跨文化内容生成；语料库构建；模型优化；出海营销

DOI：10.69979/3029-2735.26.02.074

1 引言

1.1 研究背景与意义

中国品牌出海过程中，跨文化内容本土化适配不足成为关键瓶颈，大量企业因内容文化不适配、缺乏高效工具而面临营销周期长、成本高的问题。传统模式依赖人工团队，存在效率低、成本高、适配不准等问题；现有 AI 工具多局限于单一语种，难以适配多元文化禁忌与语言变体，无法满足多场景营销需求。

“语数领航·智链全球”项目针对该痛点，开展 BERT-Crosslingual 模型跨文化优化与应用研究。该模型虽具备较强跨语言理解与生成能力，但在特定文化场景适配、小语种变体处理上存在短板。项目通过构建专属跨文化语料库、优化训练策略、强化文化规则嵌入，提升模型生成的准确性与实用性，其研究意义体现在三方面：技术上完善跨文化 NLP 模型优化路径,增强小语种及语言变体处理能力；商业上为出海企业提供高效低成本的内容生产工具，降低营销风险；教育上探索 “语言 + 技术” 跨学科融合路径，为新文科人才培养提供支撑。

1.2 国内外研究现状

1.2.1 国外研究现状

海外自然语言处理技术成熟,BERT-Crosslingual、XLM-R 等跨语言预训练模型已广泛应用于多语言翻译、内容生成领域。海外研究多聚焦单一语言生成效率，缺乏“语言变体 + 文化场景 + 行业需求”三维适配探索，无法满足企业定制化营销内容生产需求。

1.2.2 国内研究现状

国内出海营销研究多集中于策略层面，技术实践相对滞后。少数基于 NLP 的跨文化内容生成研究存在短板：现有研究未形成“语料构建-模型优化-场景应用”闭环，技术成果难以转化为实用工具，与市场需求脱节。

1.3 研究内容与创新点

1.3.1 核心研究内容

本研究聚焦 BERT-Crosslingual 模型跨文化优化与实践，核心内容有三：一是构建多语种跨文化语料库，整合 50 万条覆盖 30 个重点出海国家的语言样本，标注文化禁忌、语言变体、行业术语等关键信息；二是优化模型训练策略，通过跨文化预训练、文化规则嵌入、行业场景微调提升生成准确性；三是开展实证测试与应用，在巴西新能源汽车市场试点，验证模型实际效果并形成可复制技术方案。

1.3.2 创新点

语料库创新：搭建“语言 + 文化 + 行业”三维语料库，标注 15 种英语变体、50 + 宗教文化禁忌、3000 + 行业专属术语，填补通用语料库文化细节短板；

模型优化创新：提出“预训练 + 文化适配 + 场景微调”三级优化策略，嵌入动态文化规则库实现禁忌实时校验，将文化误判率控制在 5% 以下；

应用模式创新：开发轻量化跨文化内容生成系统，支持营销文案、短视频脚本等多场景输出，适配中小企业低成本、高效率的生产需求。

2 跨文化内容生成模型优化设计

2.1 模型优化总体框架

本研究基于 BERT-Crosslingual 预训练模型，构建“语料库支撑-模型优化-规则嵌入-应用输出”总体框架（图 1）。

图 1 总体框架

2.2 跨文化语料库构建

语料库是模型优化的核心基础，其构建遵循“全面性、精准性、针对性”原则，采用“公开数据采集 + 人工标注 + 行业补充”模式，总规模达 50 万条样本。

2.2.1 语料来源

语料来源包括公开多语言数据库、目标市场社媒内容、行业专业语料及人工标注数据。

2.2.2 语料标注体系

构建“语言特征-文化属性-行业标签”三维标注体系，标注工具为 LabelStudio，经“双人标注 + 交叉校验 + 专家审核”保障质量，标注一致率超 92%。

2.2.3 语料库管理与更新

采用 MySQL 数据库存储，设“语言-国家-行业-场景”四级分类索引；建立动态更新机制，每月新增 2000-3000 条相关语料，确保时效性与适用性。

2.3 模型优化训练策略

基于 BERT-Crosslingual 预训练模型，采用“预训练-文化适配-场景微调”三级训练策略，逐步提升模型的跨文化内容生成能力：

2.3.1 跨文化预训练

将 50 万条语料输入模型，优化词嵌入层与注意力机制，强化模型对语言变体、文化词汇的识别理解。

2.3.2 文化适配微调

新增文化分类损失函数；嵌入 3000+条文化禁忌规则作为约束，实时校验生成内容，降低文化误判风险。

2.3.3 行业场景微调

针对汽车、美妆、3C 行业，覆盖多类场景；引入行业术语词典与场景模板库，提升生成内容的行业专业性与实用性。

2.4 跨文化内容生成系统开发

基于优化模型搭建“前端交互+后端推理+数据库支撑”架构的生成系统：前端以 Vue.js 开发界面，支持用户录入生成需求；后端依托 Python Flask 框架集成模型实现推理，对接语料库与规则库完成校验；数据库采用 MySQL 与 Redis 协同方案保障效率。系统支持多格式输出，提供内容修改与二次生成功能，满足定制化需求。

3 实证测试与结果分析

3.1 测试方案设计

3.1.1 测试指标设定

设定文化误判率、生成时间、用户满意度等核心指

标。

3.1.2 测试数据与对象

测试数据分两类：一是从跨文化语料库抽取 1 万条未参与训练的语料，覆盖 15 种英语变体、30 个国家文化场景、3 个重点行业；二是国内某新能源汽车品牌巴西市场的 100 项实际营销任务。测试对象为优化后的 BERT-Crosslingual 模型（实验组）、传统人工翻译 + 本地化团队（对照组 1）、通用 GPT 模型（对照组 2）。

3.1.3 测试流程

先通过测试集开展实验室测试，统计三组对象的准确性与效率指标；再将三组对象应用于巴西新能源汽车营销内容生产，收集实地应用反馈与效果数据；最后通过对比分析验证优化后模型的优势。

3.2 测试结果与分析

3.2.1 准确性测试结果

实验室测试显示，实验组文化误判率仅 4.7%，远低于对照组 1 的 8.3%和对照组 2 的 15.6%；语言变体适配准确率 91.2%，高于对照组 1 的 85.7% 和对照组 2 的 72.3%；行业术语使用准确率 93.5%，略高于对照组 1 的 92.1%，远高于对照组 2 的 78.6%，模型跨文化及行业适配优化成效显著。

表 1 准确性测试结果对比表

测试指标	实验组（优化后模型）	对照组 1（人工团队）	对照组 2（通过 GPT 团队）
文化误判率	4.7%	8.3%	15.6%
语言变体适配准确率	91.2%	85.7%	72.3%
行业术语使用准确率	93.5%	92.1%	78.6%

3.2.2 效率与成本测试结果

实验组单条文案平均生成时间 30 秒，较对照组 1 的 4 小时效率提升 480 倍，较对照组 2 的 2 分钟提升 4 倍；1000 字文案生成成本 1.2 元，较对照组 1 的 600 元降低 99.8%，较对照组 2 的 10 元降低 88%，兼具高效与低成本优势，适配中小出海企业预算。

表 2 效率与成本测试结果对比表

时间指标	实验组（优化后模型）	对照组 1（人工团队）	对照组 2（通过 GPT 模型）
单条文案生成时间（平均）	30 秒	240 分钟	2 分钟
单条内容生成时间（1000 字）	1.2 秒	600 元	10 元

3.2.3 实地应用测试结果

实验组完成巴西市场 100 项多语种、多场景营销内容生成任务，企业营销团队对内容本土化适配满意度达 89.6%，修改率仅 12.3%，87 项内容经简改后即可投用。应用后品牌巴西社媒互动量提升 45.3%，文化相关投诉归零，模型实用性与可靠性得到验证。

3.3 结果讨论

测试结果表明，本研究优化后的 BERT-Crosslingual 模型在跨文化内容生成方面具备显著优势。

同时，测试过程中也发现模型存在一定局限性：一是对部分小众语言变体的支持能力不足，对非洲部分小众语言变体的处理效果有待提升；二是在极端复杂的文化场景（如宗教仪式相关营销内容）中，生成内容的细腻度仍需优化；三是对最新文化热点的响应存在延迟，需进一步提升语料库的动态更新速度。这些问题将作为后续研究的重点优化方向。

4 讨论

4.1 模型优化的核心价值

4.1.1 技术价值

本研究提出“语料库构建-模型微调-规则嵌入”的跨文化 NLP 模型优化路径，完善了跨语言生成模型的文化适配方法。模型支持 15 种英语变体生成且文化误判率降至 4.7%，填补了国内相关工具在语言变体处理与文化适配准确性上的短板，推动 NLP 技术在跨文化营销领域的深度落地。

4.1.2 商业价值

优化后的模型及生成系统，为出海企业提供“高效、低成本、低风险”的跨文化内容生产方案，企业无需组建专业团队即可快速生成符合目标市场文化语境的营销内容，大幅缩短周期、降低成本。巴西新能源汽车市场试点显示，系统可提升内容本土化适配性、降低文化冲突风险，助力企业开拓海外市场，具备广泛推广前景。

4.1.3 教育价值

本研究是“语言 + 技术”跨学科融合的成功实践，整合了外语专业文化研究与计算机专业技术开发优势，为新文科建设提供范例。本研究为‘语言+技术’跨学科融合与新文科人才培养提供了实践范例。

4.2 研究局限与未来展望

4.2.1 研究局限

研究在语料覆盖广度、动态热点响应及创意内容生成方面仍存在局限。

4.2.2 未来展望

未来将从三方面深化研究：一是扩展语料库，新增 10 种小众出海语言语料并强化小语种变体标注训练；二是优化模型动态更新机制，接入海外主流社媒实时数据接口，实现文化热点与语言变体的实时捕捉和模型微调；三是引入生成式对抗网络（GAN）搭建创意生成模块，提升创意营销内容产出能力，开发文本+图像建议的多模态跨文化内容生成系统，拓展应用场景与价值。

5 结论

本研究聚焦 BERT-Crosslingual 模型的跨文化优化与应用，通过构建“语言 + 文化 + 行业”三维跨文化语料库，实施“预训练 + 文化适配 + 场景微调”三级训练策略，嵌入动态文化规则库，成功研发高效精准的跨文化内容生成模型与系统。实证测试与实地应用表明，优化后模型文化误判率降至 4.7%，在生成效率、成本控制、文化适配准确性上优势显著，可有效破解出海企业跨文化内容本土化适配不足的核心痛点。

后续通过扩展语料覆盖、优化动态更新机制、提升创意生成能力等持续优化，模型将具备更广泛应用场景与更高实用价值，为中国品牌出海提供更强技术支撑。

参考文献

[1]刘群,李艳翠.自然语言处理技术在跨文化沟通中的应用研究[J].计算机工程与应用,2021,57(12):1-8.

[2]张卫山,王晨.跨境电商中的跨文化营销风险及应对策略[J].国际经贸探索,2020,36(7):89-102.

[3]李强,陈丽.基于 BERT 的跨语言文本分类模型优化研究[J].中文信息学报,2021,35(6):56-64.

[4]王健,赵亮.多模态舆情分析技术研究进展[J].模式识别与人工智能,2021,34(8):721-732.