推进使用 Shona 俚语的对话式 AI：促进数字包容的数据集和混合模型

Happymore Masoka
Pace University, Seidenberg School of Computer Science and Information Systems
Advisor: Krishna Bathula, Ph.D.

2025 年 9 月

摘要

非洲语言在自然语言处理（NLP）中仍然代表性不足，大多数语料库局限于正式用法，无法捕捉日常交流的活力。本研究通过引入一个从匿名社交媒体对话整理的新绍纳语—英语俚语数据集来解决这一问题。该数据集标注了意图、情感、对话行为、混码和语气，并公开发布在 https://github.com/HappymoreMasoka/Working_with_shona-slang。我们微调了一个多语言 DistilBERT 分类器用于意图识别，实现了 96.4%的准确率和 96.3%的 F1 值，模型托管在 https://huggingface.co/HappymoreMasoka。该分类器被整合进一个混合聊天机器人中，结合了基于规则的响应与检索增强生成（RAG），以处理特定领域的查询，并通过一个使用案例展示，在帕克大学协助意向学生获取研究生项目信息。定性评估显示，混合系统在文化相关性和用户参与度方面优于仅使用 RAG 的基础模型。通过发布数据集、模型和方法论，这项工作推进了非洲语言的 NLP 资源，促进了包容且文化共鸣的对话 AI 的发展。

1 介绍

人工智能（AI）系统，从虚拟助手 [Kepuska and Bohouta, 2018] 到推荐引擎 [Gomez-Uribe and Hunt, 2015] 和自动驾驶汽车 [Shladover, 2018] 的普及已经重塑了人机交互。然而，非洲语言，在整个大陆有超过 2,000 种 [Eberhard et al., 2023]，由于其低资源状态 [Ahia and Boakye, 2023, Nekoto et al., 2020]，在自然语言处理（NLP）中仍然严重代表性不足。这种排除风险加剧了数字鸿沟，限制了教育、医疗保健和治理等关键领域的 AI 驱动服务的访问 [Ndichu et al., 2024, Joshi et al., 2020]。

尚加语，一种在津巴布韦和南部赞比亚由数百万人使用的班图语系语言，体现了这一挑战。现有的尚加语文本资料主要包含正式文本，如新闻文章或宗教文献 [Eberhard et al., 2023]，而日常交流，特别是在年轻使用者中，则主要是俚语、与英语混用以及非正式表达 [Eisenstein, 2013]。基于正式数据训练的标准自然语言处理模型难以处理这些动态的语用模式，阻碍了具有文化相关性的对话式人工智能的发展。

本研究介绍了一个从匿名社交媒体对话中整理的尚瓦语-英语俚语数据集，并对其进行了意图、情感、对话行为、代码混合和语气的标注，公开获取地址为 https://github.com/

HappymoreMasoka/Working_with_shona-slang。我们对一个多语言 DistilBERT 模型 [Devlin et al., 2019] 进行了微调以进行意图分类，在非正式输入上实现了稳健的表现，该模型托管于 https://huggingface.co/HappymoreMasoka。分类器被整合进一个混合聊天机器人中，该聊天机器人结合了用于可预测查询的基于规则的响应与检索增强生成（RAG）[Lewis et al., 2020] 以提供特定领域的信息，在帮助佩斯大学潜在学生的用例中进行了评估。我们的贡献包括：

- 一个公开的 Shona-英语俚语数据集，标注了多种语言特征。
- 在意图识别上达到 96.4%准确率的微调 DistilBERT 分类器。
- 一种结合意图分类和 RAG 的混合聊天机器人架构。
- 文化相关性和用户参与度的用例评估。

这些贡献推动了低资源非洲语言的自然语言处理，促进了数字包容和文化共鸣的人工智能系统。

2 相关工作

推动非洲自然语言处理的进步已通过 Masakhane 等倡议取得进展，这些倡议促进协作资源创建 [Nekoto et al., 2020]。多语言模型如 BERT[Devlin et al., 2019]、XLM-R[Conneau et al., 2020] 和 AfriBERTa[Ogueji et al., 2021] 显示出潜力，但受限于在正式语料库上的训练，难以处理俚语和混码输入 [Eisenstein, 2013]。意图分类已经在英语 [Larson et al., 2019] 和代码切换环境中 [Ladhak et al., 2021, Nguyen et al., 2020] 进行了探索，但非洲语言的标注数据集仍然稀缺 [Ahia and Boakye, 2023]。检索增强生成（RAG）通过整合外部知识 [Lewis et al., 2020] 提升了对话式 AI，如乌干达的 WASHtsApp 系统 [Kloker et al., 2024] 和最近的开放领域对话模型 [Wang et al., 2024] 所示。据我们所知，这是首个针对绍纳语俚语进行对话式 AI 研究的工作，为非正式的非洲沟通提供了定制的数据集和模型。

3 数据集和方法论

3.1 数据集

我们整理了一个非正式的绍纳语和绍纳语-英语混合文本的数据集，数据来源于匿名的社会媒体对话，可在 https://github.com/HappymoreMasoka/Working_with_shona-slang 获取。该数据集包含约 34000 个表达，手动标注了以下内容：

- **意图**：类别包括问候、感谢、请求、宗教查询、金融、教育和告别。
- **情感**： 正面、负面或中性。
- **对话行为**：问题、陈述或命令。
- **代码混合特征**：词级或短语级的语言切换（例如，绍纳语到英语）。

- **语气**：友好、正式或幽默。

示例注释：

消息："Hie swit mom"
规范化："Hi sweet mom"
意图：问候
情感：正面
对话行为：陈述
代码混用：英语形容词（sweet），绍纳语问候（Hie）
语气：友好

为了解决类别不平衡问题，我们使用了 `sklearn.utils.resample` [Larson et al., 2019] 混合过采样和欠采样方法。数据集被转换成 Hugging Face Dataset 对象，使用 `distilbert-base-multilingual-cased` 分词器进行分词（应用了截断和填充），并以固定随机种子分成 80% 的训练集和 20% 的验证集，以便于复现。

3.2 意图分类管道

我们对一种多语言 DistilBERT 模型（*用于序列分类的自动模型*）[Devlin et al., 2019] 进行了意图分类的微调，并将其托管在 `https://huggingface.co/HappymoreMasoka`。该流程包括：

- **标签编码**：意图被映射到数值索引。
- **训练**：使用 Hugging Face Trainer API 进行，学习率为 2×10^{-5}，批量大小为 4，3 个纪元，权重衰减为 0.1，混合精度（FP16），以及提前停止（耐心值为 2）。
- **评估**：度量包括准确率、加权 F1 分数、精度和召回率，使用 `sklearn.metrics` 计算。

训练在 Google Colab 上进行，最佳模型检查点是根据验证损失选择的。

3.3 混合聊天机器人架构

聊天机器人集成了意图分类、基于规则的响应和 RAG[Lewis et al., 2020]，使用 Python 通过 Hugging Face Transformers、Sentence-Transformers 和 ChromaDB 实现。组件包括：

- **意图分类器**：微调后的 DistilBERT 模型处理输入，返回意图标签和置信度分数。基于规则的后处理器将标签映射为人类可读的意图。
- **基于规则的响应**：预定义的绍纳语回应，用于问候或告别等意图，融入文化细微差别。
- **RAG 模块**：对于特定领域的查询（例如大学项目），使用 `all-MiniLM-L6-v2` 嵌入查询，并从包含佩斯大学研究生项目数据的 ChromaDB 知识库中检索出最相关的前 5 份文档。响应是使用谷歌/`flan-t5-`小模型生成的。

- **交互式工作流**： 与应用程序相关的查询触发多轮对话以收集用户详细信息（例如，姓名、教育背景）。

系统在一个实时循环中运行，处理退出命令并记录响应。

4 结果与评估

4.1 定量结果

微调后的 DistilBERT 分类器在验证集上表现出了强大的性能：

- 准确率：96.48%
- F1 分数: 96.39%
- 准确率：96.40%
- 回忆：96.48%
- 评估损失：0.387

指标的高一致性表明在不平衡意图类别中具有稳健的泛化能力。

4.2 定性评估

聊天机器人在实时对话中进行了评估，展示了有效意图识别和响应生成：

- **问候**：输入: “wadii” → 意图: 问候（置信度 1.00）→ 响应: “Hesi shamwari! Uri sei hako?”
- **宗教查询**：输入: “mune mufundisi here” → 意图: 宗教 → 响应: 部门联系方式。
- **金融查询**：输入: “pace inoita mari?” → 意图: 财务 → 因训练数据有限，无法回复。
- **应用查询**： 输入：“mune ma program api pa Pace” → 意图：教育 → 互动对话收集用户详细信息。
- **退出**：输入: “exit” → 响应: “好的，再见！”

4.3 基线比较

混合系统与仅使用 RAG 的基线（不进行意图分类）进行了比较。混合模型提供了具有文化相关性的回复（例如，自然的尚比亚问候语），而基线则生成了通用或离题的回答，特别是在非教育查询方面。

5 讨论

本工作解决了低资源非洲语言中需要了解俚语的自然语言处理的关键需求 [Ahia and Boakye, 2023, Nekoto et al., 2020]。标注的数据集使在混用语境中的意图和情感检测变得稳健 [Eisenstein, 2013]，而混合架构则平衡了对可预测查询的精确性和对开放式查询的灵活性 [Wang et al., 2024, Kloker et al., 2024]。关键影响包括：

- **增强的语言理解能力**：社交媒体分析和用户互动中非正式输入的改进处理 [Nguyen et al., 2020]。
- **文化相关的对话**：能够与讲尚比亚语的用户产生共鸣的情境感知型响应 [Kepuska and Bohouta, 2018]。
- **数字包容**：扩大低资源语言社区对人工智能的访问权限 [Joshi et al., 2020, Ndichu et al., 2024]。

5.1 用例：教育助理

聊天机器人被部署以协助佩斯大学的潜在学生，处理绍纳语俚语和英语的查询。它支持问候、课程咨询和申请流程，提高绍纳语使用者的可访问性 [Ahia and Boakye, 2023]。

5.2 限制

该数据集规模较小（约 34 条语句），限制了覆盖范围，尤其是在金融相关意图方面。计算约束限制了在 Google Colab 上的超参数调整。伦理考量包括确保数据匿名化以及社区验证文化适宜性 [Ndichu et al., 2024, Joshi et al., 2020]。

6 结论与未来工作

本研究提出了一套绍纳语-英语俚语数据集、一个微调的 DistilBERT 分类器和一个混合聊天机器人，可在 https://github.com/HappymoreMasoka/Working_with_shona-slang 和 https://huggingface.co/HappymoreMasoka 处获得。该系统准确率达到 96.4%，通过解决非正式沟通问题，推动了非洲自然语言处理的发展 [Ahia and Boakye, 2023, Joshi et al., 2020]。未来的工作包括：

- 扩大数据集以包含更多意图和语言。
- 集成音频输入以实现多模态对话。
- 增强域自适应检索的 RAG。
- 进行人机回路评估。

通过释放这些资源，我们促进了包容性的自然语言处理，弥合了正式模型与非洲沟通动态现实之间的差距。

致谢

本工作在佩斯大学进行，在 Krishna Bathula 博士的指导下，位于赛登伯格计算机科学与信息系统学院。感谢佩斯大学提供必要的资源和支持。

参考文献

Orevaoghene Ahia and Edward Boakye. Bridging the digital language divide: African languages in nlp. In *Proceedings of the 4th Workshop on African Natural Language Processing (AfricaNLP 2023)*, pages 1–10, 2023.

Alexis Conneau, Kushal Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, et al. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, 2019.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. *Ethnologue: Languages of the World*. SIL International, 26th edition, 2023.

Jacob Eisenstein. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, 2013.

Carlos A. Gomez-Uribe and Neil Hunt. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems*, 6(4):Article 13, 2015.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, 2020.

Veton Kepuska and Gamal Bohouta. Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa, and google home). In *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 99–103. IEEE, 2018.

Robert Kloker, Sarah Kizza, and Ruth Babirye. Washtsapp: A retrieval-augmented conversational agent for health information in uganda. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2041–2053, 2024.

Faisal Ladhak, Mikel Artetxe, Xiang Li, Shruti Bhosale, Sebastian Ruder, Francisco Guzmán, and Phil Lewis. On the effectiveness of cross-lingual transfer for intent detection and slot filling in low-resource languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2576–2587, 2021.

Stefan Larson, Anish Mahendran, Joseph Peper, Christopher Clarke, Andrew Lee, Adam Hill, et al. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1311–1316, 2019.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, pages 9459–9474, 2020.

Michael Ndichu, Cynthia Oduor, and Ife Orife. African languages and the global ai race: Risks and opportunities. *AI and Society*, 39(2):341–354, 2024.

Wilhelmina Nekoto, Vukosi Marivate, Muhammad Abdul-Mageed, et al. Participatory research for low-resourced machine translation: A case study in african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, 2020.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, 2020.

Kelechi Ogueji, David Adelani, and Orevaoghene Ahia. Afriberta: Transformer-based language models for african languages, 2021. arXiv preprint arXiv:2106.01547.

Steven E. Shladover. Connected and automated vehicle systems: Introduction and overview. *Journal of Intelligent Transportation Systems*, 22(3):190–200, 2018.

Tianyu Wang, Haoran Chen, and Zhiyuan Yu. Rethinking retrieval-augmented generation for open-domain dialogue. In *Proceedings of the 2024 Annual Conference of the Association for Computational Linguistics*, pages 310–325, 2024.