从本土走向全球：在非洲语境中应对语言多样性

Nelson Ndugu *
nelson.ndugu@mak.ac.ug

Rashmi Margani *
rashmi.margani@gmail.com

arXiv:2305.01427v1 [cs.CL] 2 May 2023

摘要

本文聚焦自然语言处理领域中与非洲大陆语言多样性及其变异相关的关键问题，特别是长期以来鲜少受到关注的非洲地方方言和阿拉伯语方言。本文提出的方法将多种词嵌入技术与 *LightGBM* 相结合，以应对准确捕捉 *TUNIZI TUNIZI* 中所表达语境的挑战。我们评估了多种方法，证明了其有效性，同时指出该方法可能对寻求改善客户体验并以非洲地方方言开展产品开发的企业产生影响。将该模型用作基于产品的教学工具颇具意义，因为它有望激发学习者的兴趣，并促进科技创业。总体而言，我们改进后的方法为应对非洲地方方言（尤其是阿拉伯语方言）相关挑战提供了颇具前景的分析，这可能对寻求改善客户体验和产品开发的企业产生重大影响。

1. 引言

近年来，社交媒体平台已成为非洲普遍使用的沟通方式，每天都会产生海量数据 [5, 17, 40]。其中许多数据包含用户生成内容，例如状态更新、评论和评价，这些内容能够为了解消费者行为以及他们对产品或服务的情感提供宝贵见解 [25, 26, 29]。然而，分析此类数据可能颇具挑战，尤其是涉及使用范围不广或具有独特特征的非洲地方方言时 [34, 30, 2]。我们可以从运用自然语言处理技术进行简单的情感分析入手，克服语言复杂性带来的障碍 [32], ，进而推动电子商务、社交媒体网络、银行业、医疗保健及其他领域中人工智能产品的有效开发。这将促进全球范围内的多元化与无障碍化，尤其惠及那些在自然语言处理领域鲜少受到关注的地区。

*马凯雷雷大学，坎帕拉，乌干达

非洲地方方言数据资源库 [34]。本文聚焦于非洲地方方言的情感分析，尤其关注分析来自社交媒体平台的文本数据所面临的挑战。我们提出一种利用自然语言处理从各种非洲地方方言交流内容中提取含义的新方法，并凸显情感分析在从社交媒体平台文本中提取消费者评价方面的有效性。我们的方法展现了人工智能帮助企业和组织改善非洲市场客户体验与产品开发的潜力。

本文结构如下。第 2 节着重介绍非洲语境中的自然语言处理资源，并从商业视角进行探讨。第 3 节介绍所使用的数据和计算模型，包括多种词嵌入技术以及作为分类器的 LightGBM。此外，我们还展示所采用的实验设置，以及在原始数据集中使用不同平衡技术所产生的影响。第 4 节和第 5 节分别阐述所构建模型的应用及其面临的挑战。第 6 节总结该模型，并展望未来的研究方向。

2. 问题背景

非洲大陆拥有 2,000 多种不同的语言和方言，其中一些源于与外来文化的融合 [9, 10, 7]。例如，突尼斯阿拉伯语（以下简称 TUNIZI）和斯瓦希里语方言分别广泛通行于北非和东非，它们融合了阿拉伯语、法语和拉丁语[18]；东非沿海地区的班图语，以及阿拉伯商人 [31]。这些新兴语言独具特色，拥有大量非洲使用者 [18, 31]。然而，非洲语言的语言多样性给自然语言处理带来了重大挑战，尤其是在本土数据集相关知识有限的领域 [3]。

除了语言复杂性之外，词嵌入技术作为一种预处理步骤，依赖大型数据集来生成有意义的向量表示 [28, 36, 11]。然而，非洲方言的数据集可用性有限，这使得创建准确的词表示变得颇具挑战。此外，预训练模-

型通常基于无关的语言数据集进行训练，因此可能无法准确捕捉非洲语言的细微差异和复杂性 [27]。例如，某些词嵌入技术对于具有类型学复杂性的语言可能无法充分发挥作用 [14]。一些具有复杂声调系统的非洲语言，或具有独特句子结构、会根据所使用的声调传达不同含义的语言，可能无法在未经相应训练的词嵌入算法上取得良好表现，而另一些具有独特句子结构的语言也可能存在同样的问题。为应对这些挑战，开发专门针对非洲语言和方言的词嵌入算法至关重要。由于研究表明，不同的词嵌入技术在不同方言上的表现各不相同 [14], ，我们将在第3.3小节中探讨最适合本文所使用的 TUNIZI 数据集的词嵌入技术。

然而，开发此类算法需要克服诸多挑战。首先，非洲方言的数据集有限，难以生成准确的词语表示。此外，人工标注产生的人为错误可能会进一步降低词嵌入的准确性。缺乏非洲语言方面的专业人才，也给这些方言的自然语言处理模型开发带来了重大挑战。

在当今时代，许多非洲人更喜欢在社交媒体上使用当地方言，或将方言与法语、英语等外语混合来表达自己，这种现象通常称为语码转换 [38]。然而，上文讨论的挑战，包括缺乏合适的自然语言处理工具，使银行业 [33],保险公司 [39], 和社交媒体影响者 [4]难以从客户对其产品或服务的评价、点赞和不喜欢中获得有价值的洞见 [8, 35]。尽管存在这些复杂性，本文展示了如何通过收集非洲当地语言方言的客户评价，变革银行业、保险、医疗保健、社交媒体及其他领域

3. 数据与方法

3.1. 数据

本研究以从社交媒体平台获取的 TUNIZI 方言书面短信作为主要数据来源 [19]。数据通过 iCompass 平台收集，该平台汇集了表达对热门话题情感的评论。为确保数据可靠，预处理时移除了链接、表情符号和标点符号。随后，使用网络爬虫从社交媒体平台收集的最新数据对这些新数据进行了扩充，并进一步采用标准的自然语言处理技术进行预处理，包括分词、停用词移除和词干提取。

需要注意的是，TUNIZI 方言指的是使用拉丁字符和数字书写的突尼斯方言，而非使用阿拉伯字母书写的方言。由于数据质量会直接影响机器学习模型 [6], 的性能，我们对预处理后的数据采取了一系列诊断措施，包括检查是否存在数据不平衡。我们的分析发现，新兴数据存在显著的不平衡现象（见图 1），对此我们通过平衡数据进行了解决。

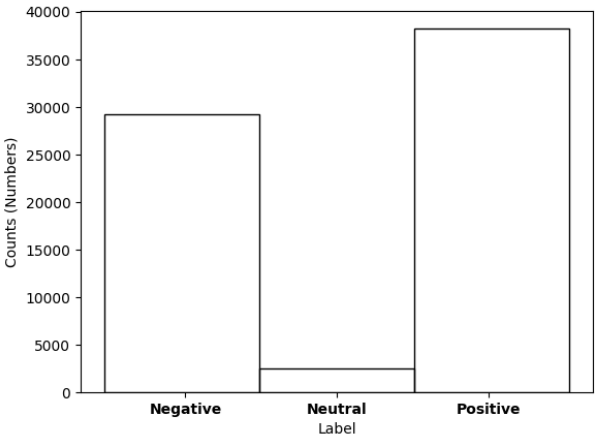

图1. 标注数据集的比例。正面情感的占比高于负面和中性标签对应的情感类别（按比例递减排序），这清楚地表明本文数据集存在类别不平衡问题。

不平衡数据会导致机器学习模型偏向多数类，从而显著影响其性能 [15, 22]。为缓解这一问题，有多种处理不平衡数据的技术，各有优缺点。最常用的技术包括随机欠采样、随机过采样、合成少数类过采样技术（SMOTE） [15], 、ADASYN（自适应合成采样） [21], 和代价敏感学习[16]。我们在第 3.3 节详细介绍了这些技术对新兴模型性能的影响。

3.2. 计算语言学模型

我们的情感分析方法采用自然语言处理（NLP），并结合了两种机器学习模型，即 Facebook 的 FastText[11, 23, 20][1] 和 LightGBM [24]。FastText 用于文本特征提取，而 LightGBM 用于情感——

[1]提醒读者，后续阶段曾考虑多种词嵌入技术，其对模型性能的影响见表2。

分类。FastText 是一个轻量、高效的自然语言处理库，可用于快速训练文本分类模型。LightGBM 是一种梯度提升框架，经过优化以实现高性能，并能高效处理大规模数据。

令 $\mathcal{D} = (x_i,\ y_i)_{i=1}^{n}$ 表示训练数据集，其中x_i 是 TUNIZI 方言中的第 i 条文本消息，$y_i \in -1,\ 0,\ 1$ 是其情感标签（-1 表示负面，0 表示中性，1 表示正面）。令 $f_{\text{fastText}}(\cdot)$ 表示使用 Facebook 的 fastText 算法将文本消息映射为固定长度向量表示的函数，令 $f_{\texttt{LightGBM}}(\cdot)$ 表示使用 LightGBM 将向量表示映射为情感预测的函数。那么，我们的情感分析方法可表述如下：

$$f_{\text{sentiment}}(\cdot) = f_{\texttt{LightGBM}}(f_{\text{fastText}}(\cdot)) \tag{1}$$

该方法的目标是在训练数据集$\mathcal{D}$上学习 $f_{\text{fastText}}(\cdot)$ 和 $f_{\texttt{LightGBM}}(\cdot)$ 的最优参数，使所得的 $f_{\text{sentiment}}(\cdot)$ 能够准确预测 TUNIZI 方言中新文本消息的情感。这可以通过最小化以下损失函数 [13] 来实现：

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell(y_i, f_{\text{sentiment}}(x_i;\theta)), \tag{2}$$

其中，θ 表示 f_{fastText}（·）和$f_{\texttt{LightGBM}}$（·）的参数，ℓ（·）为二元交叉熵损失函数，$f_{\text{sentiment}}$（x_i；θ）表示给定参数 θ时 x_i 的预测情感。最优参数 θ^* 可通过求解以下优化问题得到：

$$\theta^* = \arg\min_{\theta} \mathcal{L}(\theta) \tag{3}$$

获得最优参数后，可以按如下方式预测新文本消息 x 的情感：

$$\hat{y} = f_{\text{sentiment}}(x;\theta^*) \tag{4}$$

其中，$\hat{y}$ 是情感类别的预测概率（-1 表示负面，0 表示中性，1 表示正面）。多极性分类需要一个阈值概率参数 t。如果$\hat{y} \geq t$，则文本被分类为正面；如果$\hat{y} \leq 1-t$，则被分类为负面；如果$1-t < \hat{y} < t$，则被分类为中性。阈值 t 的选择会影响预测模型的性能。较低的阈值会将更多文本分类为正面，而较高的阈值会将更多文本分类为负面。可以通过选择使 F1 分数（衡量精确率和召回率之间平衡的指标）达到最大值的取值，来确定最佳阈值 [1]。此模型未指定阈值，因此 LightGBM 会根据概率最高的类别标签来分配预测类别标签。例如，如果某个给定文本的预测概率为输入文本为 [0.2、0.3、0.5] ，对于多类别标签 [0, -1, 1]，LightGBM 会将类别标签 1 分配为预测结果，因为它的概率最高。

3.3. 实验设置

我们使用 Python 以及 scikit-learn 和 LightGBM 库实现了所提出的方法。模型在一台配备英特尔酷睿 i7 处理器、16GB 随机存取存储器和 NVIDIA GeForce GTX 1050Ti 图形处理器的计算机上进行训练和评估。我们采用 70:30 的比例分别划分训练集和测试集。模型的超参数通过结合 5 折交叉验证的网格搜索进行调优。我们的模型使用 F1 分数进行评估。

在非洲地方方言语言情感分析中，由于语言中某些情感类别的出现频率存在差异，数据集可能会出现不平衡。例如，某些情感类别的表达频率可能高于其他类别，从而导致数据集不平衡。F1 分数同时考虑精确率和召回率，这两项指标在情感分析中都非常重要。精确率衡量模型所做出的所有正面分类中真正例分类所占的比例，而召回率则衡量数据集中所有实际正样本中真正例分类所占的比例。在情感分析中，较高的精确率意味着模型能够在正面或负面情感出现时正确识别它们，而较高的召回率则意味着模型能够正确识别数据集中的所有正面或负面情感。因此，通过使用兼顾精确率和召回率的 F1 分数 [37, 12], ，我们可以更全面地衡量情感分析模型在非洲地方方言语言学领域的整体表现。它有助于衡量模型在所有类别（而不仅仅是多数类）上的表现，这在处理不平衡数据集时尤为重要。

在名义模型上进行了参数扫描，通过对 50 组参数组合执行五折交叉验证，共完成 250 次拟合。经过这些操作，我们为名义模型取得了 0.8065 的最佳 F1 分数，其模型参数如下：

{’subsample’: 0.6, ’reg_lambda’: 1.0, ’reg_alpha’: 0.1, ’num_leaves’: 64, ’n_estimators’: 1000, ’min_child_weight’:10, ’max_depth’: 8, ’learning_rate’: 0.1, ’colsample_bytree’: 1.0}

由于自然语言处理模型对训练数据很敏感，而平衡有助于提升数据集多样性，因此我们测试了在原始不平衡数据上采用不同平衡技术对模型性能的影响。我们比较了不同平衡技术，尤其是欠采样、过采样和 ADASYN 的性能

与未采用任何平衡技术的基准模型进行了比较。为评估模型在不同平衡技术（即不同数据集多样性）下的性能，我们使用了 F1分数，结果汇总于表 1。

表 1. 我们的模型对不同数据平衡技术的敏感性。US、OS 和 ADASYN 分别表示欠采样、过采样和自适应合成采样技术。

技术	学习率	最大深度	# 估计器	F1 分数
US	0.01	6	1000	0.7036
OS	0.1	8	1000	**0.9071**
ADASYN	0.1	8	1000	0.8640

本分析旨在确定最适合用于我们的自然语言处理模型的类别平衡技术。表中汇总的结果表明，过采样是解决类别不平衡问题并提升模型在所有类别上表现的最有效方法。过采样取得了最高的 F1 分数**0.9071**，这表明模型能够更好地捕捉数据集的细微特征和复杂性，并得出更准确的结果。

结果还表明，ADASYN 通过生成合成样本来解决类别不平衡问题，其表现优于欠采样；欠采样则只是移除多数类中的样本。这说明，生成合成样本可以有效解决 TUNIZI 数据集中的类别不平衡问题，并提升模型表现。

此外，为每种技术选择的超参数也影响了所得结果。较高的学习率和更大的深度有助于模型更好地捕捉数据集的复杂性，而增加估计器的数量则可以提高模型的准确率。不过，最佳超参数会因数据集和所解决的具体问题而异。

总体而言，结果表明，谨慎选择合适的技术和超参数对于取得良好的模型性能至关重要。尤其是在对合适的词嵌入技术知之甚少，且模型是在复杂、不平衡的数据集（如 TUNIZI 数据集）上训练时。本研究的发现对于指导从业者采用最合适的平衡技术，以提升自然语言处理模型在不平衡的非洲地方方言数据集上的性能，具有重要意义。

除 FastText（本文所称的名义词嵌入技术）外，本研究考察的不同词嵌入技术还包括上下文词嵌入和字节对编码（BPE）。FastText、BERT 和字节对编码（BPE）词嵌入技术属于基于特征数量的方法。我们的分析结果见下方表2。

表2. 不同词嵌入技术对模型性能的影响。CWE 表示上下文词嵌入。

模型	学习率	最大深度	# 估计器	F1 分数
FastText	0.1	8	1000	**0.8065**
CWE	0.1	8	1000	0.7019
BPE	0.05	8	500	0.6944

该表展示了三种不同词嵌入技术——FastText、上下文词嵌入和 BPE——在 TUNIZI 数据集上的结果。

这些技术分别使用不同的超参数进行训练，具体参数如表所示。具体而言，不同模型的学习率、最大深度和估计器数量有所不同。

结果显示，FastText 的整体表现最佳，F1 分数为 0.8065。这表明，与其他模型相比，FastText 能够捕捉 TUNIZI 数据集更多的细微差异和复杂性。

上下文词嵌入和 BPE 的表现均不如 FastText，F1分数分别为 0.7019 和 0.6944，这是因为像 BERT 这样的上下文词嵌入模型，以及像 GPT-2、RoBERTa、BART 和 DeBERTa 这样的字节对嵌入模型面临诸多挑战，例如：

(1) 数据不足：这些模型的训练数据主要为英语，因此由于非洲语言的训练数据有限，模型表现不佳。

(2) 词汇量有限：许多非洲语言的形态学复杂，屈折和派生形式丰富，因此词汇量庞大。然而，这些模型的词汇量有限，难以表达非洲语言中丰富多样的含义。

(3) 语言多样性：非洲语言丰富多样，拥有不同的方言、口音和书写系统，因此难以训练出能够处理这些差异的模型。相比之下，FastText 使用子词单元来表示词语，因此更适用于形态学复杂且训练数据有限的语言。与 BERT 和其他模型相比，FastText 也能更好地处理语言多样性，因为它不依赖预先训练好的语言模型，而是直接从训练数据中学习嵌入。

总体而言，这些结果凸显了选择合适超参数、尝试不同词嵌入技术和模型的重要性，以及

使用合适的指标（如 F1 分数）评估这些模型性能的必要性。

4. 我们模型的应用

通过情感分析，可以使用开发的模型衡量消费者对社交媒体平台上内容的评论。情感分析涉及分析给定文本的极性；在本例中，极性分为正面、中性或负面。开发的模型可用于将文本归入这些类别。假设某家公司想衡量消费者对其最近推出的某款产品在社交媒体平台上的评论。该公司可以部署开发的模型，对情感分析进行测试。

给定一条新的社交媒体文本，x，模型会输出一个预测概率，$\hat{y}$，表示该文本属于某一特定情感类别的概率。随后，可根据阈值 t，使用预测概率 $\hat{y}$ 将文本归入三个类别之一。

企业可以利用这种情感分析监测消费者对其产品的评论，并根据反馈作出必要的改进。它们还可以利用情感分析，将自家产品的情感与竞争对手的产品进行比较，从而作出明智决策，保持竞争优势。开发的模型可以成为一种强大的工具，帮助企业衡量消费者对社交媒体平台及其他领域的评论，并据此作出数据驱动的决策。

5. 挑战

本文中，我们开发了一个情感分析模型，并将其应用于衡量消费者在社交媒体平台上的评论。然而，开发此类模型的过程并非没有挑战。本节将讨论我们在开发模型时遇到的一些挑战：

(1) 获取高质量的训练数据；我们遇到的第一个挑战是获取高质量的训练数据。为了训练模型，我们需要一个规模庞大的社交媒体文本语料库，并由人工为其标注相应的情感。这一过程既耗时又费力。此外，标注人员的可靠性和一致性也存在问题，这可能会影响最终模型的准确性。为应对这一挑战，我们采用了多种策略，例如使用多名标注人员，并实施质量控制检查，以确保标注的准确性和一致性。

（2）处理自然语言固有的歧义性和复杂性，是我们项目面临的一项重大挑战，尤其是在突尼斯语境中。突尼斯人倾向于在社交媒体平台上使用非正式语言，这为情感分析任务增添了额外的复杂性。

在突尼斯，不同的人可能会因其文化和社会背景不同而使用相同的词语或短语表达不同的评论。此外，语境在判断文本的情感方面起着至关重要的作用。例如，突尼斯阿拉伯语中的“hlib”意为“面包”，但也可以作为俚语指代金钱。因此，要正确理解一个词所表达的情感，必须考虑其使用语境。

为说明这一点，请看以下两个突尼斯阿拉伯语句子：

(a) ”我的三明治很好吃，但服务不好。” (b) ”我的三明治不好吃，但服务很好。”

两个句子都包含正面和负面词语，但句子的整体情感并不相同。第一个句子的意思是“我的三明治很好吃，但服务不好”，而第二个句子的意思是“我的三明治不好吃，但服务很好”。这凸显了考虑语境和文本整体含义对于准确判断其情感的重要性。

为克服这些挑战，我们谨慎地运用了特征工程和模型选择技术，以确保模型能够捕捉突尼斯阿拉伯语的细微差异，并作出准确预测。我们还使用了大量标注数据构成的语料库来训练和测试模型，这有助于提高模型的准确率。

总的来说，自然语言固有的歧义性和复杂性是我们项目面临的一项重大挑战，在突尼斯语境中尤其如此。不过，通过谨慎运用特征工程和模型选择技术，并使用大量标注数据构成的语料库，我们成功开发出一个能够准确预测突尼斯阿拉伯语社交媒体文本情感的模型。

(3) 与实现和部署相关的技术挑战：最后，模型的实现和部署也面临着技术挑战。训练数据规模庞大，模型复杂度高，因此需要大量计算资源；此外，在优化模型效率和可扩展性方面也存在问题。另外，模型还面临着与现有软件系统集成的挑战

，以及将其部署到生产环境中的挑战。解决这些挑战需要在机器学习、软件工程和 DevOps 方面具备丰富的专业知识，这对缺乏这些资源的组织而言构成了显著的采用障碍。

总而言之，针对非洲地方方言数据开发情感分析模型是一项颇具挑战性的任务，需要仔细考量数据质量、模型复杂度和部署要求等多种因素。通过应对本文讨论的挑战，我们成功开发出一个能够将社交媒体文本准确分类为正面、负面或中性类别的模型。然而，值得注意的是，我们遇到的挑战并非个例，其他尝试开发类似模型的人很可能也会遇到。因此，了解这些挑战并制定恰当的应对策略至关重要。

6. 结论

本文全面分析并阐述了如何应对 TUNIZI 及其他非洲方言数据集中的自然语言处理问题。这有望革新非洲商业市场，并提升透明度与多样性，因为它使人们能够通过语言翻译、聊天对话等方式，在与世界其他地区交流时使用非洲地方方言。我们运用成熟的自然语言处理技术，以社交媒体平台上的最新文本扩充 TUNIZI 数据，构建了一个全面的数据集。我们使用 Facebook 的 FastText 对模型进行嵌入，并采用自主开发的 LightGBM 分类器进行分类。经过全面的参数实验，我们的模型取得了令人瞩目的 0.8065 F1 分数。

尽管 TUNIZI 数据集本身存在复杂性，我们的模型在高准确率预测客户评论方面展现出良好前景。我们相信，我们的方法能够在产品开发过程中为决策提供数据驱动的洞见，助力企业更有效地满足使用 TUNIZI 方言的客户的需求和期望。此外，经过针对多种非洲方言的进一步训练后，我们的模型不仅能够预测 TUNIZI 方言的客户评论，还能预测其他几种非洲方言的客户评论。

本文提出了三个未来方向，以进一步改进多样化非洲方言语境中的情感分析模型：

(1) 根据方言相似性对多样的非洲语言进行聚类，并为每个类别创建专门的词嵌入技术。(2) 收集多样的非洲方言，并在当地方言专家的指导下进行标注。(3) 尝试不同的词嵌入技术，并在不了解合适的词嵌入技术时，采用能优化模型性能的方法。

我们相信，这种适度的创新有望对教育领域产生积极影响。例如，它有能力激发基于产品的教学方法，让 STEM 学习者从我们模型的卓越性能中获得动力，并以此激发所在社区的创新。

致谢

本文的研究工作部分得到了 KaggleXBIPOCmentorship 项目的资助。

参考文献

[1] A. Abbasi、A. Hassan 和 M. Dhar。Twitter 情感分析工具基准测试。载于第九届国际语言资源与评估会议论文集（*LREC'14*），第823–829页，冰岛雷克雅未克，2014年5月。欧洲语言资源协会（ELRA）。[2] I. Adebara 和 M. Abdul-Mageed。面向非洲语言的非洲中心主义自然语言处理：现状与未来方向，2022年。[3] I. Adebara 和 M. Abdul-Mageed。面向非洲语言的非洲中心主义自然语言处理：现状与未来方向。载于计算语言学协会第*60*届年会论文集（第*1*卷：长篇论文），第3814–3841页，爱尔兰都柏林，2022年5月。计算语言学协会。[4] L. Ao、R. Bansal、N. Pruthi 和 M. B. Khaskheli。社交媒体影响者对客户参与度和购买意愿的影响：一项荟萃分析。《*Sustainability*》，15(3)，2023年。[5] G. Barbier 和 H. Liu。《社交媒体中的数据挖掘》，第327–352页。Springer US，美国波士顿，马萨诸塞州，2011年。[6] G. E. A. P. A. Batista、R. C. Prati 和 M. C. Monard。机器学习训练数据平衡的多种方法行为研究。《*SIGKDD Explor. Newsl.*》，6(1):20–29，2004年6月。[7] R. M. Beck。非洲的城市语言。《*Africa Spectrum*》，45(3):11–41，2010年。[8] A. Bilgihan、J. Kandampully 和 T. C. Zhang。打造统一的在线购物客户体验：前因与结果。《*International Journal of Quality and Service Sciences*》，8(1):102–119，2016年。[9] J. Blommaert。*4.* 语言多样性：非洲，第123–150页。2007年。[10] J. Blommaert。语言多样性：非洲，第123–149页。10 2007年。[11] P. Bojanowski、E. Grave、A. Joulin 和 T. Mikolov。利用子词信息丰富词向量。《*Transactions of the Association for Computational Linguistics*》，5 146，2017 135: – 。

[12] S. Boughorbel、F. Jarray 和 M. El-Anbari。采用 Matthews 相关系数指标处理不平衡数据的最优分类器。*PloSone*，12(6):e0177678，2017年。[13] C. Broyden、R. Fletcher 和 D. Goldfarb。非线性规划。电气工程师学会会刊，117(12):2157–2163，1970年。[14] C. P. Chai。文本预处理方法比较。自然语言工程，第1–45页，2022年。[15] N. V. Chawla、K. W. Bowyer、L. O. Hall 和 W. P. Kegelmeyer Jr.。合成少数类过采样技术。人工智能研究杂志，16:321–357，2002年。[16] C. Drummond 和 R. Holte。C4.5、类别不平衡与成本敏感性：为何欠采样优于过采样。载于不平衡数据集学习研讨会 *II*，第1–8页，2003年。[17] G. Duffett、Rodney。社交媒体营销传播对年轻消费者态度的影响。*Young Consumers*，18(3):19–39，2017年。[18] C. Fourati、A. Messaoudi 和 H. Haddad。TUNIZI：突尼斯阿拉伯语拉丁转写情感分析数据集。*arXiv*，abs/2004 14303，2020年。[19] C. Fourati、A. Messaoudi 和 H. Haddad。TUNIZI：突尼斯阿拉伯语拉丁转写情感分析数据集。载于*AfricaNLP* 研讨会：将非洲置于自然语言处理版图。*ICLR 2020*，线上活动，卷 arXiv:3091079，2020年。[20] E. Grave、P. Bojanowski、P. Gupta、A. Joulin 和 T. Mikolov。为157种语言学习词向量。载于第十一届国际语言资源与评估会议论文集（*LREC 2018*），日本宫崎，2018年5月。欧洲语言资源协会（ELRA）。[21] H. He、Y. Bai、E. A. Garcia 和 S. Li。ADASYN：针对不平衡学习的自适应合成采样方法。载于*2008*年 *IEEE* 国际神经网络联合会议（*IEEE* 世界计算智能大会），第1322–1328页，2008年。[22] H. He 和 E. A. Garcia。从不平衡数据中学习。*IEEE* 知识与数据工程汇刊，21(9):1263–1284，2009年。[23] A. Joulin、E. Grave、P. Bojanowski 和 T. Mikolov。高效文本分类技巧集。载于欧洲计算语言学协会第*15*届会议论文集：第*2*卷，短论文，第427–431页，西班牙瓦伦西亚，2017年4月。计算语言学协会。[24] G. Ke、Q. Meng、T. Finley、T. Wang、W. Chen、W. Ma、Q. Ye 和 T.-Y. Liu。LightGBM：一种高效的梯度提升决策树。载于 I. Guyon、U. V. Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan 和 R. Garnett 编，神经信息处理系统进展，第30卷。Curran Associates, Inc.，2017年。[25] C. S. Lee 和 L. Ma。社交媒体中的新闻分享：满足感与先前经验的影响。人类行为中的计算机，28(2):331–339，2012年。[26] S. W. Litvin、R. E. Goldsmith 和 B. Pan。酒店与旅游管理中的电子口碑。旅游管理，29(3):458–468，2008年。

[27] M. C. Mafunda、M. Schuld、K. Durrheim 和 S. Mazibuko。基于南非新闻数据训练的词嵌入。*TheAfrican Journal of Information and Communication (AJIC)*，(30)，2022年12月。[28] T. Mikolov、K. Chen、G. Corrado 和 J. Dean。高效估算向量空间中的词表示，2013年。[29] M. M. . S. E. G. Muntinga、D. G. 引入 cobras：探索品牌相关社交媒体使用的动机。*In- ternational Journal of Advertising*，30(1):13–46，2011年。[30] W. Nekoto、V. Marivate、T. Matsila、T. Fasubaa、T. Ko- lawole、T. Fagbohungbe、S. O. Akinola、S. H. Muham- mad、S. Kabongo、S. Osei、S. Freshia、R. A. Niy- ongabo、R. Macharm、P. Ogayo、O. Ahia、M. Meressa、M. Adeyemi、M. Mokgesi-Selinga、L. Okegbemi、L. J. Mar- tinus、K. Tajudeen、K. Degila、K. Ogueji、K. Siminyu、J. Kreutzer、J. Webster、J. T. Ali、J. Abbott、I. Orife、I. Ezeani、I. A. Dangana、H. Kamper、H. Elsahar、G. Duru、G. Kioko、E. Murhabazi、E. van Biljon、D. Whitenack、C. Onyefuluchi、C. Emezue、B. Dossou、B. Sibanda、B. I. Bassey、A. Olabiyi、A. Ramkilowan、A. ¨Oktem、A. Akinfaderin 和 A. Bashir。面向低资源机器翻译的参与式研究：非洲语言案例研究，2020年。[31] D. Nurse、T. J. Hinnebusch 和 G. Philippson。斯瓦希里语与萨巴基语：语言史。1993年。[32] K. Ogueji、Y. Z hu 和 J. Lin。小数据？没问题！探索预训练多语言语言模型在低资源语言中的可行性。载于*Proceedingsof the 1st Workshop on Multilingual Representation Learn- ing*，第116–126页，蓬塔卡纳，多米尼加共和国，2021年11月。计算语言学协会。[33] V. O. A. Okolo、K. P. Ikpo 和 C. O. Obikeze。客户反馈对尼日利亚东南部存款银行客户留存的影响。*British Journal of Management and Marketing Studies*，4(1):1–16，2021年。[34] T. Olagunju、O. Oyebode 和 R. Orji。探讨影响非洲移动电子商务应用的关键问题：情感与主题分析。*IEEE Access*，8:114475–114486，2020年。[35] J. Palmer、K. Revis 和 Y. Romain。社交媒体分析如何助力企业。*SSRNElectronic Journal*，2020年10月。[36] J. Pennington、R. Socher 和 C. Manning。GloVe：词表示全局向量。载于*Proceedings of the2014 Conference on Empirical Methods in NaturalLan-guage Processing (EMNLP)*，第1532–1543页，多哈，卡塔尔，2014年10月。计算语言学协会。[37] Y. Sasaki。F度量的真相。*Journal of Machine Learning Research*，8:1001–1015，2007年。[38] B. Sutrisno 和 Y. Ariesta。超越 Instagram 社交媒体影响者的语码混用。*Advances inLanguage and Literary Studies*，2019年。[39] Y. Wang、X. Zhao 和 J. Wu。客户评价如何影响在线保险市场中的购买决策。*Journal of Business Research*，115:1–11，2020年。[40] A. Wong、S. Ho、O. Olusanya、M. V. Antonini 和 D. Ly- ness。疫情期间社交媒体和在线通信的使用。*Journal of the Intensive Care Society*，22(3):255–260，2021年。PMID: 34422109。