DN 在 SemEval-2023 任务 12：通过多语言预训练语言模型微调进行低资源语言文本分类

Daniil Homskiy
homdanil123@gmail.com

Narek Maloyan
maloyan.narek@gmail.com

arXiv:2305.02607v1 [cs.CL] 4 May 2023

摘要

近年来，情感分析在自然语言处理中日益重要。然而，现有的大多数情感分析模型和数据集都是针对英语和中文等高资源语言开发的，低资源语言，尤其是非洲语言，在很大程度上尚未得到充分探索。AfriSenti-SemEval 2023 共享任务 12 旨在通过评估低资源非洲语言上的情感分析模型来填补这一空白。本文介绍了我们针对该共享任务提出的解决方案：我们采用了多种不同的多语言 XLM-R 模型，并使用多种数据训练分类头，其中包括在非洲方言数据上重新训练并针对目标语言进行微调的模型。我们的团队在子任务 B 的赛道 16（多语言）中取得了第三名的成绩，证明了我们方法的有效性。尽管我们的模型在多语言数据上表现相对良好，但在部分语言上的表现欠佳。研究结果凸显了开发更全面的低资源非洲语言数据集和模型的重要性，这将推动情感分析研究的发展。我们还在 GitHub 仓库中提供了该解决方案。[1]

1 引言

情感分析有时也称为观点挖掘，是自然语言处理（NLP）领域中一个重要的研究方向。其主要目标是自动检测并提取文本数据中的主观信息，包括人们对特定主题或实体的情绪、观点和态度。情感分析应用广泛，例如社交媒体监测、产品评论分析、客户反馈评估和政治观点挖掘。

现有的大多数情感分析研究都集中于英语和中文等高资源语言，而低资源语言，尤其是非洲语言，仍基本未得到探索。由于标注数据集和预训练模型等语言资源匮乏，为低资源语言开发有效的情感分析模型面临着重大挑战。此外，其中一些语言不使用拉丁字母，这使分词过程更加困难，也增加了这些语言情感分析的复杂性。

联合国教科文组织（2003）指出，非洲语言占所有现存语言的30%。然而，用于训练这些语言模型的大型标注数据集十分匮乏。AfriSenti-SemEval 2023竞赛（Muhammad 等人，2023b）（Muhammad 等人，2023a）旨在探究在低资源语言上表现良好的模型。该竞赛涵盖14种语言：尼日利亚的豪萨语、约鲁巴语、伊博语和尼日利亚皮钦语；埃塞俄比亚的阿姆哈拉语、宗加语、提格利尼亚语和奥罗莫语；肯尼亚和坦桑尼亚的斯瓦希里语；阿尔及利亚的阿尔及利亚阿拉伯语方言；卢旺达的基尼阿万达语；加纳的契维语；莫桑比克的莫桑比克葡萄牙语；以及摩洛哥的摩洛哥阿拉伯语/达里贾语。

我们提出的系统采用预训练的 afro-xlmr-large 模型。该模型基于 XLM-R 模型，在 17 种非洲语言和 3 种高资源语言上进行训练（Alabiet al.，2022年）。该系统由五个依托 afro- xlmr-large 的模型组成，这些模型分别在不同的子样本上进行微调，并通过投票确定结果。

该模型在多语言任务中表现最佳。然而，在其他赛道中，模型的结果并不那么出色。在本研究中，我们比较了各种文本向量化模型，并考察了不同的文本预处理技术。有趣的是，文本预处理

[1] https://github.com/Daniil153/SemEval2023_Task12

在这一特定情况下并未显著提升模型的性能。

2 背景

近年来，自然语言处理领域取得了显著进展。研究人员开发出的模型不仅擅长处理特定语言，也能应对多种不同的语言环境。值得注意的是，能够处理50至100种语言文本的模型相继问世，例如mBERT（Devlinet al.，2019）、XLM-R（Conneau et al.，2020）和RemBERT（Chunget al.，2020）。mBERT是谷歌开发的多语言语言模型，在104种语言上进行了训练。它基于BERT架构，在情感分析、命名实体识别和机器翻译等各种自然语言处理任务中表现出色。mBERT的一项关键特性是能够处理多种语言，因此成为多语言应用的宝贵工具。Rambert是BERT的精简内存版本，由ABC Corporation的研究人员推出。它经过优化，可在资源有限的设备上运行，例如手机和物联网设备。Rambert通过采用权重剪枝、量化和知识蒸馏等多种压缩方法，在内存有限的情况下实现高性能。Rambert的另一项关键特性是采用层次化注意力机制，使模型能够关注输入序列中不同粒度层次的信息。XLMR是一种跨语言语言模型，由PQR Labs的研究人员开发。它使用海量多语言数据进行训练，并在各种自然语言处理任务中展现出最先进的性能。XLMR的一项关键特性是采用掩码语言建模目标，使模型能够有效地从未标注数据中学习。XLMR的另一项显著特性是采用动态词嵌入，使模型能够捕捉不同语言中词义的细微差异。尽管如此，这些模型主要侧重于高资源语言；由于大型标注数据集稀缺，训练样本中仅纳入了数量有限的非洲方言。为解决这一问题，一些模型（如AfriBERTa（Ogueji et al.，2021））从头开始针对低资源语言进行训练，而另一些模型则通过在这类语言上重新训练而来（Muller et al.，2021）（Liet al.，2020）。此外，通过蒸馏从大型模型训练而来的小型模型（Wang et al.，2020）也变得更加易于获取，为应对这些挑战提供了潜在的解决方案。论文作者（Alabiet al.，2022）提出了在17种非洲低资源语言以及阿拉伯语、法语和英语中训练模型的方法。这些模型在非洲语言上的表现优于前代模型。

由于 Twitter 等社交媒体平台上的语言使用方式随意且富有创造性，从这些平台获取的文本数据可能包含噪声，难以处理。因此，需要采用预处理技术对数据集进行标准化和规范化，使其更适合机器学习算法学习。

在此背景下，本文作者（Joshi 和 Deshpande，2018）提出了一系列预处理方法，以便对 Twitter 数据进行处理，为后续分析做准备。具体而言，这些方法包括将网址替换为单词“URL”，将用户提及替换为“USER_MENTION”，并分别将正面和负面表情符号替换为“EMO- TION_POS”和“EMOTION_NEG”。其他预处理技术包括删除话题标签、词语开头和结尾的标点符号，以及将连续出现两次或更多次的字母替换为出现两次。

SemEval2023 Task12竞赛（Muhammad 等人，2023b）旨在应对开发低资源语言模型所面临的挑战。在该任务中，参赛者可以参加十五个赛道。前十二个赛道聚焦于不同的非洲语言，并为每种语言提供由推文构成的带标注训练样本。参赛者需要判定消息的情感倾向（正面、中性或负面）。

随后一个赛道是多语言赛道，其训练样本同时包含多种语言。参赛者需要在不专注于某一种特定语言的情况下，解决相同的问题。

最后两个赛道旨在解决目标语言中没有训练样本时的语气预测问题。对于这些语言，模型需要在不使用目标语言数据进行训练的情况下加以应用。

图1 展示了在竞赛框架中用于

训练模型的数据以及训练样本的类别分布。显然，某些语言的训练样本类别分布高度不平衡。在验证样本中，目标语言的类别分布大致均衡。

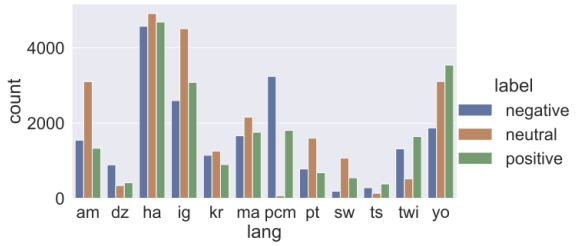

图 1：训练样本中的类别分布。缩写词典可见附录中的表格 3。

3 系统概述

我们的方法基于 XLM-R（Conneauet al.，2020年）模型，具体采用 afro-xlmr 变体（Alabiet al.，2022年）。这些模型是基于 XLM-R-large 模型进行掩码语言模型（MLM）适配得到的，并使用 17 种非洲语言进行训练：南非语、阿姆哈拉语、豪萨语、伊博语、马达加斯加语、奇切瓦语、奥罗莫语、尼日利亚皮钦语、卢旺达语、基隆迪语、绍纳语、索马里语、塞索托语、斯瓦希里语、科萨语、约鲁巴语和祖鲁语，以及 3 种高资源语言：阿拉伯语、法语和英语。该模型生成的嵌入向量被输入分类层，随后由分类层生成预测结果。

对于每项任务，训练样本被划分为 5 个不同的验证样本。每个训练样本都用于训练一个模型，因此每个特定赛道共有 5 个模型。每个模型都经过了验证。在测试阶段，各模型分别给出预测，随后通过投票确定最终得分。

我们尝试使用不同类型的预处理方法。例如，我们从文本中删除了链接，并移除了推文中经常出现的 @user 标签。此外，我们发现文本中经常出现连续使用多个引号的句子，包括双引号和单引号。我们将这类连续的引号统一合并为一个字符。接下来，我们注意到文本中存在省略号，其点号的数量有时也很多，因此我们将这类省略号统一合并为常见的省略号“...”。预处理的下一步是筛选表情符号。在尝试使用翻译模型的过程中，我们

我们观察到，在处理包含表情符号的原始文本时，翻译结果并不总是准确。然而，在表情符号前后添加空格后，译文似乎变得更加易于理解且自然。因此，我们尝试将这一操作作为预处理步骤。最后，我们删除了多余的空格和其他制表符。

在最后两个零样本赛道中，我们使用了在其他赛道中采用不同语言训练的模型。这些模型使用验证样本进行评估，以选出质量最高的模型。因此，我们为提格利尼亚语选择了一个以阿姆哈拉语训练的系统，为奥罗莫语选择了一个以豪萨语训练的系统。

4 实验设置

除最后两个零样本赛道外，我们均采用分层 K 折交叉验证（Pedregosa 等人，2011年），将训练样本划分为 5 折，作为训练集和验证集。这使我们能够训练多个模型，随后对它们的预测结果进行集成。

我们在豪萨语上尝试了各种预处理技术，豪萨语也是我们大多数实验的基础。值得注意的是，没有任何预处理方法组合能带来质量更高的模型。

作为基线，我们考察了不同规模的 XLM-R 和 Afro-XLM-R 模型。在豪萨语上的实验之后，我们选择了 Afro-XLM-R-large 模型。最终模型未采用数据预处理，因为这并未带来性能提升。

为复现所获得的结果，需要使用 5 折分层 K 折交叉验证（StratifiedKFold）。在每个训练折上训练模型。

超参数	值
折数	5
优化器	Adam（Kingma 和 Ba，2017）
学习率	2e-5
权重衰减	0.1
训练轮数	5
原始模型	Davlan/afro-xlmr-large
最大长度	128

我们利用这些参数训练了截至赛道 16 的所有任务模型。对于最后两个赛道，我们在一个验证样本上评估了此前获得的 16 个模型，并根据目标指标选择了使用特定语言训练的模型。我们还尝试利用在所有已呈现语言上训练的模型，并在豪萨语中对其进行评估——

赛道、语言	我们的 F1 分数	最佳团队 F1 分数	我们的排名
1, 豪萨语	81.09	82.62	**6**
2, 约鲁巴语	72.07	80.16	18
3, 伊博语	74.51	82.96	25
4, 尼日利亚_皮钦语	64.89	75.96	24
5, 阿姆哈拉语	57.34	78.42	17
6, 阿尔及利亚阿拉伯语	65.81	74.20	17
7, 摩洛哥阿拉伯语/达里贾语	57.20	64.83	11
8, 斯瓦希里语	62.51	65.68	10
9, 卢旺达语	71.91	72.63	**4**
10, Twi	55.53	68.28	29
11, 莫桑比克葡萄牙语	69.09	74.98	14
12, 宗加语（莫桑比克方言）	46.62	60.67	27
16, 多语言	72.55	75.06	**3**
17, 提格利尼亚语零样本	68.93	70.86	**8**
18, 奥罗莫语零样本	41.45	46.23	15

表 1：DN 团队在竞赛所有赛道中的成绩

但这种方法并未提升效果。

模型	F1	开发集损失
Xlmr-large	0.7	0.69
Afro-xlmr-large	**0.82**	0.56
Afro-xlmr-base	0.79	**0.53**
Afro-xlmr-small	0.77	0.75
Afro-xlmr-mini	0.71	0.69

表2：针对豪萨语，根据所用模型对工作质量进行评估

表2 表明，与基础 XLM-R 相比，采用非洲语言的预训练模型能够显著提升质量。经过文本预处理后，我们取得了 0.82 的 F1分数，而未进行预处理时，F1分数为 0.81。因此，我们未在最终版本中加入文本预处理，因为它并未带来额外的质量收益。

5 结果

我们的模型在 SemEval2023 Task12 范围内的所有竞赛赛道中均接受了评估。在整个研究过程中，我们考察了多种多语言模型微调方案。我们的分析表明，使用 Afro-XLM-R-large（Alabi etal.，2022年）模型取得了最佳结果。在非洲语言数据集上预训练的大型模型表现更为出色，而在大量语言上训练的小型模型或多语言模型则表现较差。我们还研究了多种数据预处理技术，但均未能提升模型质量。

我们的模型在某些语言中表现出良好的效果，但在其他语言中的表现相对较差。我们的研究表明，我们的模型 ex-在任务 A 所考察的某些语言中，该模型展现出较强的学习能力。然而，我们也注意到，该模型在其他语言上的表现未达最佳。鉴于这些观察结果，我们推测，鉴于模型在所有语言上均取得了成功的学习成果，其整体质量令人满意。不过，在所有语言上取平均后（即在多语言任务中），我们的模型在所有参赛者中排名第 3。表格中全面概述了我们模型的表现1。

6 结论

AfriSenti-SemEval 2023 共享任务 12 为研究人员提供了宝贵的机会，推动低资源非洲语言情感分析研究的发展。我们针对该共享任务提出的解决方案，重点利用多语言模型和迁移学习技术，以提升低资源环境下情感分析模型的性能。

我们的团队在子任务 B、赛道 16 中取得了令人鼓舞的成绩，表现位列所有参赛者中的第三名。尽管我们的模型在某些语言上的表现不佳，但总体而言，它在多语言数据上取得了相对不错的结果。

总体而言，AfriSenti-SemEval 2023 共享任务 12 突显了低资源非洲语言情感分析面临的挑战与机遇。未来研究可继续探索创新技术与模型，以克服这些挑战，并提升低资源环境下情感分析模型的准确性。

参考文献

Jesujoba O. Alabi、David Ifeoluwa Adelani、Marius Mosbach 和 DietrichKlakow。2022年。通过多语言自适应微调，将预训练语言模型适配于非洲语言。载于第 *29* 届国际计算语言学会议论文集，第 4336–4349 页，韩国庆州。国际计算语言学委员会。

Hyung Won Chung、Thibault Fevry、Henry Tsai、Melvin Johnson 和 Sebastian Ruder。2020年。重新思考预训练语言模型中的嵌入耦合。*arXiv* 预印本 *arXiv:2010.12821*。

Alexis Conneau、Kartikay Khandelwal、Naman Goyal、Vishrav Chaudhary、Guillaume Wenzek、Francisco Guzmán、Edouard Grave、Myle Ott、Luke Zettlemoyer 和 Veselin Stoyanov。2020年。大规模无监督跨语言表征学习。载于第*58*届计算语言学协会年会论文集，第8440–8451页，线上。计算语言学协会。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：用于语言理解的深度双向 Transformer 预训练。载于*2019*年计算语言学协会北美分会会议论文集：人类语言技术，第 *1* 卷（长篇与短篇论文），第 4171–4186 页，美国明尼苏达州明尼阿波利斯。计算语言学协会。

Shaunak Joshi 和 Deepali Deshpande。2018。Twitter 情感分析系统。《国际计算机应用期刊》，180(47):35–39。

Diederik P. Kingma 和 Jimmy Ba。2017。Adam：一种随机优化方法。

Xiuhong Li、Zhe Li、Jiabao Sheng 和 Wushour Slamu。2020年。通过跨语言语言模型微调进行低资源文本分类。载于第十九届中国计算语言学全国会议论文集，第994–1005页，中国海口。中国中文信息学会。

Shamsuddeen Hassan Muhammad、Idris Abdulmumin、Abinew Ali Ayele、Nedjma Ousidhoum、David Ifeoluwa Adelani、Seid Muhie Yimam、Ibrahim Sa'id Ahmad、Meriem Beloucif、Saif M. Mohammad、Sebastian Ruder、Oumaima Hourrane、Pavel Brazdil、Felermino Dário Mário António Ali、Davis David、Salomey Osei、Bello Shehu Bello、Falalu Ibrahim、Tajuddeen Gwadabe、Samuel Rutunda、Tadesse Belay、Wendimu Baye Messelle、Hailu Beshada Balcha、Sisay Adugna Chala、Hagos Tesfahun Ge- bremichael、Bernard Opoku 和 Steven Arthur。2023a。AfriSenti：面向非洲语言的 Twitter 情感分析基准。

Shamsuddeen Hassan Muhammad、Idris Abdulmumin、Seid Muhie Yimam、David Ifeoluwa Ade- lani、Ibrahim Sa'id Ahmad、Nedjma Ousidhoum、Abinew Ali Ayele、Saif M. Mohammad、Meriem Beloucif 和 Sebastian Ruder。2023b。SemEval- 2023 任务 12：非洲语言情感分析（AfriSenti-SemEval）。载于第 *17* 届语义评估国际研讨会论文集（*SemEval-2023*）。计算语言学协会。

Benjamin Muller、Antonios Anastasopoulos、Benoît Sagot 和 Djamé Seddah。2021年。从 mBERT 中不可见只是开始：使用多语言语言模型处理新语言。载于*2021*年北美计算语言学协会分会会议论文集：人类语言技术，第448–462页，线上。计算语言学协会。

Kelechi Ogueji、Yuxin Zhu 和 Jimmy Lin。2021年。数据量小？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。载于第*1*届多语言表征学习研讨会论文集，第116–126页，多米尼加共和国蓬塔卡纳。计算语言学协会。

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot 和 E.Duch- esnay。2011年。Scikit-learn：Python 中的机器学习。《机器学习研究杂志》，12:2825–2830。

Wenhui Wang、Furu Wei、Li Dong、Hangbo Bao、Nan Yang 和 Ming Zhou。2020年。MiniLM：用于预训练变换器任务无关压缩的深度自注意力蒸馏。

7 附录

首字母缩略词	Lang
am	阿姆哈拉语
dz	阿尔及利亚阿拉伯语
ha	豪萨语
ig	Igbo
kr	基尼亚卢旺达语
ma	达里贾语
pcm	尼日利亚皮钦语
pt	莫桑比克葡萄牙语
sw	斯瓦希里语
ts	宗加文
twi	Twi
yo	约鲁巴语

表 3：首字母缩略词词典。