AfroXLMR-Social: Adapting Pre-trained Language Models for African Languages Social Media Text

Tadesse Destaw Belay[1,11], **Israel Abebe Azime**[2], **Ibrahim Said Ahmad**[3,4], **David Ifeoluwa Adelani**[5,6], **Idris Abdulmumin**[7], **Abinew Ali Ayele**[8], **Shamsuddeen Hassan Muhammad**[4,9], **Seid Muhie Yimam**[10]

[1]Instituto Politécnico Nacional, [2]Saarland University, [3]Northeastern University, [4]Bayero University Kano, [5]Mila-Quebec AI Institute, McGill University, [6]Canada CIFAR AI Chair, [7]University of Pretoria, [8]Bahir Dar University, [9]Imperial College London, [10]University of Hamburg, [11]Wollo University

Contact: tadesseit@gmail.com

Abstract

Language models built from various sources are the foundation of today's NLP progress. However, for many low-resource languages, the diversity of domains is often limited, more biased to a religious domain, which impacts their performance when evaluated on distant and rapidly evolving domains such as social media. Domain adaptive pre-training (DAPT) and task-adaptive pre-training (TAPT) are popular techniques to reduce this bias through continual pre-training for BERT-based models, but they have not been explored for African multilingual encoders. In this paper, we explore DAPT and TAPT continual pre-training approaches for African languages social media domain. We introduce **AfriSocial**, a large-scale social media and news domain corpus for continual pre-training on several African languages. Leveraging AfriSocial, we show that DAPT consistently improves performance (from 1% to 30% F1 score) on three subjective tasks: sentiment analysis, multi-label emotion, and hate speech classification, covering 19 languages. Similarly, leveraging TAPT on the data from one task enhances performance on other related tasks. For example, training with unlabeled sentiment data (source) for a fine-grained emotion classification task (target) improves the baseline results by an F1 score ranging from 0.55% to 15.11%. Combining these two methods (i.e. DAPT + TAPT) further improves the overall performance. The data and model resources are available at HuggingFace[1].

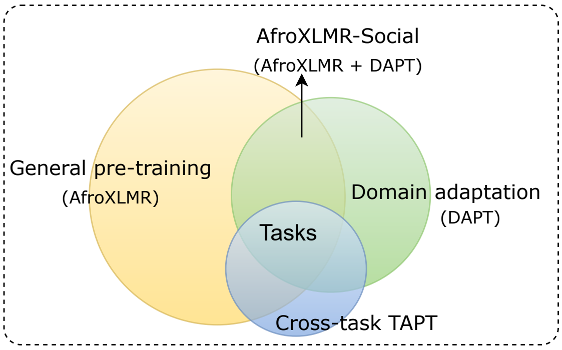

Figure 1: Continual pre-training illustration. A general-purpose pretrained model, such as AfroXLMR, is first adapted to a social domain, resulting in AfroXLMR-Social. This model then undergoes Cross-task TAPT using sentiment analysis, emotion, and hate speech data without the labels for further fine-tuning.

1 Introduction

Pre-trained language models (PLMs) are initially trained on vast and diverse corpora, including encyclopedias and web content (Conneau et al., 2020; Chiang et al., 2022). Subsequently, these pre-trained models are used in supervised training for a specific Natural Language Processing (NLP) task by further finetuning. Fine-tuned PLMs achieve strong performance across many tasks and datasets from various sources (Shi et al., 2025). However, this raises a question: Do PLMs function universally across domains, or does continual training of PLMs with domain-specific data offer better performance?

While some studies have shown the benefit of continual pre-training on a domain-specific unlabeled data (Lee et al., 2019; Gururangan et al., 2020), one domain may not be generalizable to other domains and languages. Moreover, it is unknown how the benefit of continual pre-training may vary with factors like the amount of unlabeled corpus, the source domain itself, the evaluation task, the resource richness of the target languages, and the trained target model (Gururangan et al., 2020). This raises the question of whether pre-training on a corpus more directly tied to the task can further improve performance. This work addresses these questions by continual domain adaptive pre-training (DAPT) and task adaptive pre-training (TAPT) on the downstream

[1]https://huggingface.co/tadesse/AfroXLMR-Social

subjective NLP tasks in a low-resource language setup. We consider the social media domain (X) and News for a continual pre-training from a high-performing multilingual baseline model, AfroXLMR (Alabi et al., 2022). AfroXLMR is a continually pre-trained model for African languages based on XLM-RoBERTa (Conneau et al., 2020). We explore subjective NLP task results from baseline (the base model result), DAPT, TAPT, and DAPT + TAPT, on a smaller but directly domain- and task-relevant unlabeled corpus. The results show that DAPT and TAPT highly benefit from similar source NLP tasks. Figure 1 illustrates the general high-level continual pre-training strategies. In summary, our contributions are:

- We present **AfriSocial**, a new quality domain-specific corpus for 14 African languages, collected from the social domain (X) and news.
- We perform a through analysis of domain and task adaptive continual pre-training across subjective NLP tasks for low-resource languages.
- We achieve state-of-the-art results in the evaluated NLP tasks and publicly making **AfriSocial** social-domain corpus and **AfroXLMR-Social** pretrained models for developing low-resource languages.

2 Related Work

Domain in NLP Language models (LMs) pre-trained on text from various sources are the foundation of today's NLP. Domain adaptation in NLP refers to enhancing the performance of a model using similar domain data (target domain) by leveraging knowledge from an existing domain (source domain) (Ramponi and Plank, 2020). Domain refers to different implicit clusters of text representations in pretrained LMs, such as news articles, social media posts, medical texts, or legal documents (Aharoni and Goldberg, 2020; Shi et al., 2025). Each domain has its unique characteristics, vocabulary, and writing style, which can affect the performance of NLP models when applied to new or unseen domains. Therefore, a similar domain means the text source from which the pre-trained model was made, similar to the target NLP task data source.

Continual Pre-training of LMs There are several techniques for the downstream NLP task improvements, such as general-purpose pre-training (Wang et al., 2023), language-adaptive continual pre-training (Alabi et al., 2022), domain-adaptive pre-training (Gururangan et al., 2020), task-adaptive pre-training (Alabi et al., 2022), and data augmentation (Zhang et al., 2024). Prior works have shown the benefit of continual pre-training using a domain-specific corpus (Gururangan et al., 2020) and training LMs in a specific domain from scratch (Huang et al., 2020). However, continual pre-training is arguably more cost-effective than training from scratch, since it is a continuous pre-training from the existing base language model.

Domain Adaptive Pre-training (DAPT) Domain Adaptive pre-training (DAPT) is straightforward, continuing pre-training a model on a corpus of unlabeled domain-specific text (Aharoni and Goldberg, 2020). DAPT techniques handle discrepancies between the source (pre-training) and target (fine-tuning) domains. Traditional fine-tuning often yields suboptimal results when pre-trained models encounter data that diverges significantly from their training data. In this regard, DAPT techniques reduce this mismatch by aligning data distributions, ensuring the model can generalize better in a better setup. Lee et al. (2019) considers a single domain at a time and uses a language model pretrained on a smaller and less diverse corpus than the most varied and multilingual language models.

Task Adaptive Pre-training (TAPT) Task-adaptive pre-training (TAPT) refers to pre-training on the unlabeled training set for a similar task-specific data (Gu et al., 2024). Compared to DAPT, TAPT solely leverages the training data of the similar downstream task for continuous pre-training. It uses a far smaller pre-training corpus that is much more task-relevant (assuming that the training set represents aspects of the task well). This makes TAPT much less expensive to run than DAPT. There are also a combination of various techniques, such as DAPT followed by TAPT, which is beneficial for end-task performance (Gururangan et al., 2020).

Language Adaptive Pre-training (LAPT) LAPT is also called language-adaptive fine-tuning (LAFT) (Alabi et al., 2022). It focuses on adapting a pre-trained language model to a specific language(s) using any language-specific corpus. This is done by collecting any corpus for fine-tuning language models without considering the domain (Yu and

Dataset name	Task name	# lang.	Data Sources	# Classes
AfriSenti	Sentiment analysis	14	News, social media	3
AfriEmo	Emotion analysis	17	News, social media	6
AfriHate	Hate speech detection	15	News, social media	3
AfriSocial (new domain specific corpus)	Unlabeled	14	X and news	-

Table 1: List of subjective NLP task evaluation datasets. Social media sources include posts/comments from YouTube and X. The AfriSenti dataset class labels are positive, negative, and neutral. The AfriEmo dataset labels are six basic emotions (anger, disgust, fear, joy, sadness, and surprise) in a multi-label annotation (an instance may have none, one, two, some, or all targeted emotion labels). AfriHate labels are abuse, hate, and neutral.

Joty, 2021; Alabi et al., 2022). LAPT is a vital step to improve language understanding and representation, especially for low-resource languages (Wang et al., 2023). However, in the primary studies that investigated LAPT with various language-specific corpora, the impacts of the text sources on specific NLP tasks are unexplored.

Although XLMR (Conneau et al., 2020) pretraining corpus is derived from multiple sources and languages, it has not yet been explored whether these sources are diverse enough to generalize in a specific domain and task. This leads to asking whether subjective tasks related to social media text can be understood with this generic model. Towards this end, we explore further a continual training of DAPT and TAPT from AfroXLMR-{76L} (Adelani et al., 2024) [2] and evaluate the impacts on highly subjective NLP tasks such as sentiment analysis, multi-label emotion, and hate speech classification.

3 AfriSocial: Domain Adaptation Corpus

Social Domain in our Work In computational linguistics, domain boundaries can be defined through various dimensions, including content, style, and purpose (Plank, 2016). In this work, we define the social domain based on two key factors: 1) Convergent textual characteristics, X and news sources provide a public discourse, comments, reactions, and conversational linguistic patterns that facilitate social interactions, and cultural expressions and 2) Functional similarity in downstream tasks, the selected evaluation task datasets are sourced from the two sources, shown in Table 1. This grouping demonstrates comparable performance patterns when evaluating entity recognition models across X and news data, suggesting underlying linguistic commonalities despite surface differences and prioritizing functional and distributional similarities over source platform distinctions (Derczynski et al., 2016; Ruder and Plank, 2018).

[2]While the original AfroXLMR (Alabi et al., 2022) cover 20 languages, we make use of the version with 76 languages (Adelani et al., 2024) with similar adaptation. Throughout this paper, AfroXLMR-76L is referred to as AfroXLMR.

We create AfriSocial, a social domain-specific corpus comprising X and news for 14 African languages, shown in Table 2. We select X and news because they are the most common text sources for low-resource languages to annotate a dataset for supervised NLP tasks. The motivations behind creating this domain-specific corpus are the following.

Limited coverage for African languages The available well-known compiled corpora are limited to include African languages; most of them are only English-centric, such as `fineweb` (Penedo et al., 2024) and `C4` (Raffel et al., 2020). The reasons include the extra effort required to collect data for such low-resource languages, the limited availability of text, and the challenges in detecting the language and filtering sources.

Text quality issues The quality of the available corpus is under consideration, especially for low-resource languages. For example, in the `OPUS` corpora (Lison and Tiedemann, 2016), there are Tigrinya (tir) texts under the Amharic (amh) file as both languages use the same script. Some of the available corpora are translated, such as OPUS-100 (Zhang et al., 2020), with the problem that the quality of the translator tool is still not mature enough for low-resource languages.

Non availability of social domain corpus To be specific, a domain-specific corpus is vital for adapting pre-trained language models into a specific domain, such as a health-specific domain. Likewise, the social media corpus is limited even to high-resource languages.

Lang.	X	News	Total Sent.
amh	588,154	45,480	633,634
ary	9,219	156,494	165,712
hau	640,737	30,935	671,672
ibo	15,436	38,231	53,667
kin	16,928	72,583	89,511
orm	33,587	59,429	93,016
pcm	106,577	7,781	116,358
som	144,862	24,473	169,335
swa	46,588	—	46,834
tir	167,139	45,033	212,172
twi	8,681	—	8,681
yor	26,560	49,591	76,151
xho	—	354,959	354,959
zul	12,102	854,587	866,689
Total	1.82M	1.74M	3.56M

Table 2: AfriSocial corpus statistics at language and source level, where **Total Sent.** is the number of sentences. The full names of the languages are presented in Appendix A.

3.1 Data Sources Selection

Based on our assessment, most of the NLP datasets of African languages are sourced and annotated from X and news domain, as sufficient text can be found in these two sources. As shown in Table 1, subjective NLP tasks for African languages, such as sentiment analysis, multi-label emotion, and hate speech classification datasets, are sourced from X and news. The AfriSocial corpus is sourced from similar domains to further enhance these subjective tasks. There is an X domain corpus and model for high-resource languages such as XLM-T (Barbieri et al., 2022) to evaluate and improve task datasets sourced from X. However, no available corpora specialize in the social domain for low-resource African languages. More details about the data collections are presented in Appendix I.

3.2 Pre-processing and Quality Control

We apply the following quality measures on the AfriSocial corpus.

Language Identification (LID) We apply LID tools for each language. For example, for the language mixing problem of the existing corpus mentioned in Sec §3, we used language-specific LID tools, GeezSwitch[3] to handle Ethiopic script languages (Amharic and Tigrinya) and pycld3[4] for the supported Latin and other script languages at the sentence level.

Sentence Segmentation The same approach as language identification, we used tools for each language to segment into sentences. For the Ethiopic script language, we used `amseg` tool (Yimam et al., 2021), and for other Latin script languages, we used NLTK (Bird and Loper, 2004).

Other Preprocessing We exclude sentences that contain hate/offensive words, very short sentences, only URL lines, and anonymize personally identifiable information (PII) such as usernames starting with the @ symbol, and email addresses. We pay special attention to ensuring that the available evaluation task data (sentiment, emotion, and hate speech) do not appear in the AfriSocial corpus before and after processing. De-duplication is applied if a near-similar instance is present, excluding it from AfriSocial, not from the annotated dataset. Table 2 shows the AfriSocial corpus statistics with their sources and number of sentences. We did not perform further processing on the code-switching text as we trained one single multilingual model (AfroXLMR-Social), and we need code-switching or dialectal diversity to be captured in the model.

4 Evaluation Tasks and Datasets

We select subjective NLP tasks for our evaluation based on the following reasons. 1) Subjective tasks face more disagreement during annotations, leading to less performance in the evaluation, especially for low-resource languages (Fleisig et al., 2023; Belay et al., 2025a). As we can see from the SemEval-2025 Task 11 (Muhammad et al., 2025c), an emotion detection shared task covering 32 languages, low-resource languages are not well explored, and the lowest results are from African languages. 2) Subjective NLP tasks of African languages are sourced from X and news, as shown in Table 1. These sources align with the same domain as the AfriSocial corpus. The three subjective tasks for our evaluation are sentiment analysis, multi-label emotion detection, and hate speech classification. We keep the original train-test split of all evaluation datasets throughout our experiment for proper comparison with the benchmark results.

[3] https://pypi.org/project/geezswitch/

[4] https://pypi.org/project/pycld3/

	AfriSenti			AfriEmo			AfriHate	
Language	AfroXLMR	+DAPT	Language	AfroXLMR	+DAPT	Language	AfroXLMR	+DAPT
amh	50.09	**57.22**	afr	43.66	**44.57**	amh	73.54	**78.57**
arq	52.22	**64.62**	amh	68.97	**71.67**	arq	43.41	**45.96**
ary	52.86	**62.34**	ary	47.62	**52.63**	ary	75.13	**75.6**
hau	79.34	**81.66**	hau	64.30	**70.74**	hau	**81.55**	80.78
ibo	76.92	**79.8**	ibo	26.27	**54.54**	ibo	82.78	**88.05**
kin	70.95	**72.73**	kin	52.39	**56.73**	kin	75.28	**78.75**
pcm	50.47	**52.09**	orm	52.28	**61.38**	orm	67.23	**74.11**
por	60.93	**64.81**	pcm	55.39	**59.93**	pcm	64.85	**67.61**
swa	28.26	**61.42**	ptMZ	22.09	**36.80**	som	**55.66**	55.64
tso	35.37	**38.81**	som	48.78	**54.86**	swa	91.51	**91.2**
twi	47.2	**56.00**	swa	30.74	**34.35**	tir	50.2	**55.9**
yor	72.27	**74.63**	tir	57.22	**60.71**	twi	46.89	**48.42**
orm	20.09	**24.28**	vmw	21.18	**22.08**	xho	50.91	**59.17**
tir	22.45	**24.53**	yor	28.65	**39.26**	yor	53.44	**77.9**
Avg.	51.39	**58.21**	**Avg.**	44.25	**51.45**	**Avg.**	65.17	**69.83**

Table 3: Result of baseline (AfroXLMR) and DAPT (AfroXLMR-Social) across the three datasets (AfriSenti, AfriEmo, and AfriHate). During TAPT, the text for the task-adaptive data is without the labels, and the evaluation is cross-tasked among the three target datasets. Reported results are macro-F1.

4.1 AfriSenti: Sentiment Analysis Dataset

AfriSenti (Muhammad et al., 2023) is a sentiment analysis dataset across 14 African languages. It aggragates some existing datasets such as NaijaSenti (Muhammad et al., 2022), Amharic Twitter sentiment (Yimam et al., 2020), and manually curated data. The data is sourced from X (formerly Twitter) and annotated in one of the three sentiment classes: positive, negative, and neutral. From 14 languages, the two languages, Oromo (orm) and Tigrinya (tir) have only test sets.

4.2 AfriEmo: Multi-label Emotion Dataset

SemEval-2025-Task 11 (Muhammad et al., 2025c) is an emotion dataset that covers 32 languages, from diverse domains such as social media platforms (X, Reddit, YouTube, and others) and news. The AfriSocial domain-specific corpus targets low-resource African languages; we target the African languages emotion dataset from the SemEval-2025 Task 11, specifically from BRIGHTER (Muhammad et al., 2025b) and EthioEmo (Belay et al., 2025c), which we call **AfriEmo**. It covers 17 African languages from the 32 languages. This dataset is annotated in a multi-label approach - a text might have any combination (none, one, some, or all) of emotion labels from a given set of emotions (anger, disgust, fear, joy, sadness, and surprise).

4.3 AfriHate: Hate Speech Dataset

AfriHate (Muhammad et al., 2025a) is a multilingual hate and abusive speech dataset in 15 African languages sourced from X. Each text is categorized into one of the abusive, hate, or neutral labels. The languages covered in the corresponding evaluation datasets, such as language name, ISO code, countries/regions spoken, language family, and writing script. See the details in Appendices A and B.

5 Language Models

5.1 Encoder-only Language Models

Multilingual encoder-only pretrained language models (PLMs) such as XLM-R (Conneau et al., 2020) and mBERT (Devlin et al., 2019) have shown impressive capability on many languages for a variety of downstream tasks. They are also often used to initialize checkpoints to adapt to other languages, such as AfroXLMR-76L (Adelani et al., 2024), which is initialized from XLM-R to specialize in African languages. We evaluate popular multilingual and African-centric PLMs such as AfroLM (Dossou et al., 2022) and AfriBERTa (Ogueji et al., 2021) and found that AfroXLMR is better for the targeted evaluation datasets as it covers more African languages. We make AfroXLMR our benchmarks and further train on the AfriSocial domain and the selected evaluation tasks. The AfroXLMR followed by DAPT with the AfriSocial domain-specific corpus gives us

Lang.	AfriSenti TAPT Base	AfriEmo	AfriHate	DATP + TAPT	AfriEmo TAPT Base	AfriSenti	AfriHate	DATP + TAPT	AfriHate TAPT Base	AfriSenti	AfriEmo	DATP + TAPT
amh	50.09	54.91	54.48	**55.80**	68.97	69.56	67.21	**71.04**	73.54	73.13	73.26	**78.06**
arq	52.22	62.42	59.41	**63.38**	44.93	51.25	48.95	48.72	43.41	**46.19**	43.84	44.21
ary	52.86	**64.13**	52.80	63.05	47.62	50.21	48.81	**51.63**	75.13	75.13	71.70	**77.51**
hau	79.34	80.65	80.08	**82.71**	64.30	66.93	61.16	**69.77**	81.55	77.76	81.94	**82.09**
ibo	76.92	80.01	78.29	**80.42**	26.27	53.13	51.26	**54.26**	82.78	**87.69**	86.52	**87.68**
kin	70.95	**71.47**	69.72	69.53	52.39	53.27	53.49	**54.47**	75.28	77.48	76.30	**78.36**
orm	20.09	23.53	**42.99**	28.93	52.28	56.43	52.22	**58.75**	**77.08**	69.03	67.67	71.07
pcm	50.47	50.97	50.29	**52.04**	55.39	56.71	56.60	**58.89**	64.85	67.73	66.94	**69.91**
ptMZ	60.93	**64.05**	62.80	63.75	22.09	37.20	30.99	**37.76**	—	—	—	—
som	—	—	—	—	48.78	50.33	49.63	**52.65**	55.66	**57.29**	54.97	56.75
swa	28.26	**59.33**	57.26	54.94	30.74	**33.02**	31.85	32.74	91.51	**91.91**	91.27	91.16
tir	22.45	10.81	16.22	**28.90**	**57.22**	55.72	55.84	57.12	50.20	54.21	**56.98**	32.70
twi	47.20	47.68	50.23	**54.47**	—	—	—	—	46.89	**49.30**	48.94	49.01
yor	72.27	72.22	70.90	**73.65**	28.65	34.34	32.87	**35.89**	53.44	53.69	54.51	**54.76**
Avg.	52.62	57.09	57.34	**59.35**	45.74	51.39	49.22	**52.87**	66.72	67.46	67.14	**67.77**

Table 4: Cross-TAPT results across the three datasets. **Base** column is baseline results from AfroXLMR, DATPT + TAPT column results for AfriSenti and AfriHate are TAPT from the AfriEmo dataset. DATPT + TAPT for AfriEmo results is TAPT from the AfriSenti dataset. Reported results are the Macro F1 score. Blank values (—) indicate that the specific dataset does not cover the language.

AfroXLMR-Social. The detailed hyperparameters of the continual training are shown in Appendix D.

5.2 Large Language Models

We compare our DAPT and TAPT approach results from AfroXLMR with state-of-the-art open source LLMs such as Llama 3 (Dubey et al., 2024), Gemma 2 (Riviere et al., 2024), Mistral (Jiang et al., 2024), and proprietary LLMs such as Gemini 1.5 (Reid et al., 2024) and GPT-4o (Hurst et al., 2024). For AfriSenti and AfriHate task results, we use the LLMs benchmark results from the **AfroBench** (Ojo et al., 2025) leaderboard. For the AfriEmo task, we use LLM results from the SemEval-2025 Task 11 datasets papers (Muhammad et al., 2025b; Belay et al., 2025b). The detailed versions of the LLMs are presented in the Appendix E.

6 Experiment Results

6.1 Domain Adaptive Pre-training (DAPT)

The domain-adaptive pre-training (DAPT) approach is straightforward; we continue pre-training from the strong baseline language model (AfroXLMR) using the domain-specific AfriSocial corpus in a multilingual setup. Baseline results from (AfroXLMR) and after applying DAPT are presented in Table 3.

Baseline: As a baseline, we evaluate various encoder-only models and found that AfroXLMR, which covers 76 African languages (Alabi et al., 2022), performs better than other BERT-based encoder-only models across targeted datasets since it includes more African languages during pretraining. The evaluation results of othe encoder-only multilingual and African-centric models are shown in Appendix C.

Results: The results before and after DAPT are shown in Table 3, AfroXLMR and DAPT columns, respectively. We observe that DAPT improves over the baseline in almost all languages and datasets, demonstrating the benefit of DAPT when the target domain is relevant. It consistently improves over the baseline models for each language. It suggests that continual pre-training on a small, quality, domain-relevant dataset is important for subjective tasks from the same domain.

6.2 Task Adaptive Pre-training (TAPT)

Similar to DAPT, TAPT consists of a second phase of continual pre-training. TAPT is a cross-task transfer across datasets, which refers to finetuning on the unlabeled data of the non-targeted task during evaluation. We explore the TAPT approach by directly applying it to the base model followed by DAPT. For example, if we make TAPT for the AfriSenti task, we further fine-tune the base model and DAPT model using the unlabeled data of the AfriEmo and AfriHate datasets separately. Compared to DAPT, the task-adaptive approach strikes a different trade-off: it uses a far smaller pre-training corpus, assuming the training set represents aspects

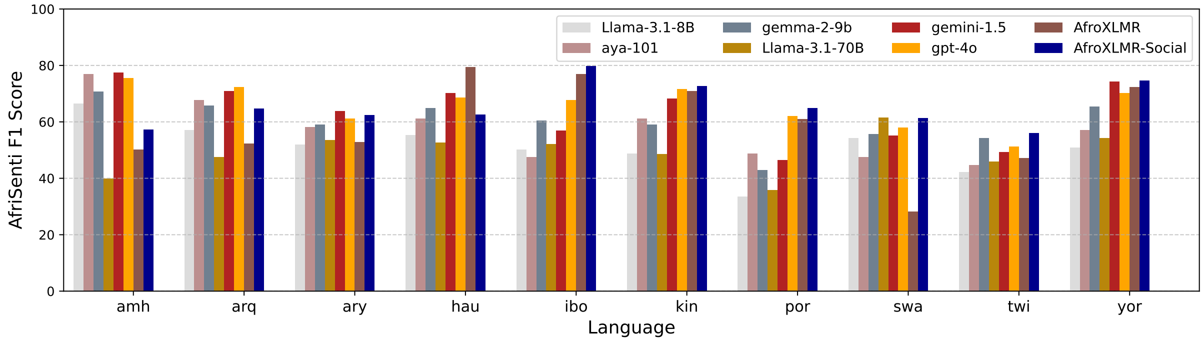

Figure 2: **AfriSenti** Macro F1 results from AfroXLMR-Social and LLMs. The results for LLMs are based on **zero-shot** evaluations, selecting the best results from five different types of prompts. The benchmark results for the LLMs are taken from the AfroBench (Ojo et al., 2025) leaderboard.

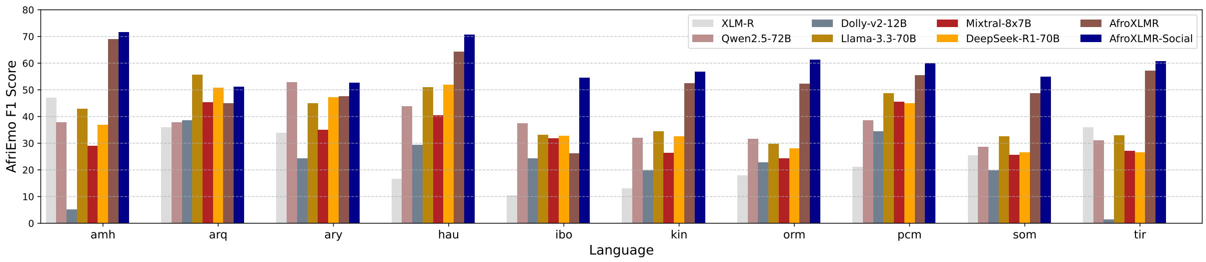

Figure 3: **AfriEmo** Macro F1 results from AfroXLMR-Social and LLMs. The results for LLMs are based on **zero-shot** evaluations, selecting the best results from five different types of prompts. The benchmark results for the LLMs are taken from the SemEval-2025 Task 11 dataset papers (Muhammad et al., 2025b; Belay et al., 2025c).

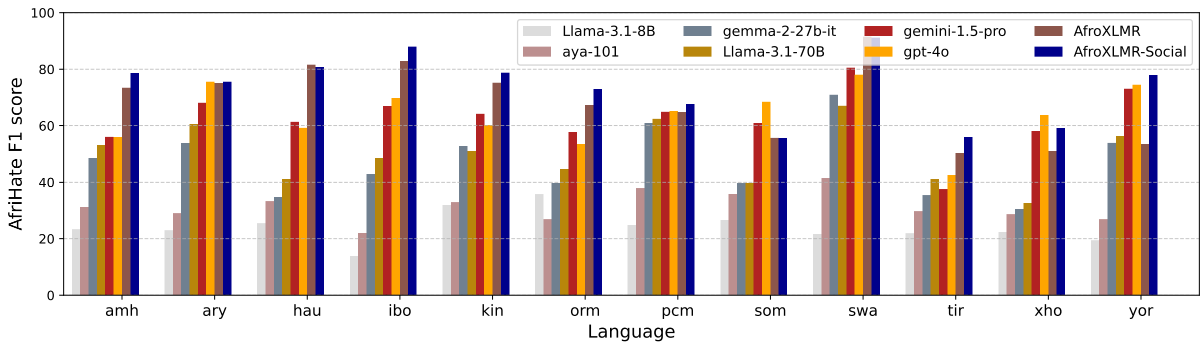

Figure 4: **AfriHate** Macro F1 results from AfroXLMR-Social and LLMs. The results for LLMs are based on **zero-shot** evaluations, selecting the best results from five different types of prompts.. The baseline results for the LLMs are taken from the AfroBench (Ojo et al., 2025) leaderboard.

of the target task. This makes TAPT much less expensive to run than DAPT. When we apply TAPT across tasks, for example, TAPT with AfriSenti for AfriHate evaluation, we ensure that we exclude any duplication from the training data (in this case, AfriSenti) to prevent data leakage.

Results Table 4 shows the TAPT results across datasets and languages. As a result, all TAPT performs better than the base model and equal or competitive results with the DAPT. This indicates we can achieve better results for our targeted evaluation task using very small, quality, and task-related data. In our case, using AfriSenti data without the labels as a TAPT is very helpful for the fine-grained multi-label emotion classification task. For the AfriSenti sentiment evaluation task, TAPT with AfriHate data achieves a better average results than TAPT with AfriEmo. For the AfriHate evaluation, TAPT in AfriSenti data has better average results than TAPT with AfriEmo. In addition to the task similarity, this improvement might be affected by the number of total instances in each dataset (the total instances of AfriSenti is 107,694, AfriEmo 70,859, and AfriHate 90,455) and the vocabulary similarity across the datasets.

6.3 Combining DAPT and TAPT

We explore the effect of both adaptation techniques by combining DAPT and TAPT. In this approach, we apply DAPT to the base model and then the TAPT training. These phases of pre-training add up to make this approach the more computationally expensive setting.

Results The results are shown in Table 4, DAPT + TAPT column. The results show that the subjective NLP tasks benefit from the combined DAPT and TAPT approaches. DAPT followed by TAPT achieves the best performance. However, first adapting the model to the domain (DAPT), then applying TAPT would be susceptible to catastrophic forgetting of the domain-relevant corpus (Yogatama et al., 2019); alternate methods of combining the procedures may result in better downstream performance. This is shown from the summarized result in Table 5 that sometimes the DAPT + TAPT result performs less than the DAPT-only results, while it is better than the baseline results.

Dataset	Models	Avg.
AfriSenti	AfroXLMR	51.39
	+ DAPT (AfroXLMR-Social)	56.85
	+ TAPT (AfriEmo)	55.72
	+ TAPT (AfriHate)	55.83
	+ DAPT + TAPT (AfriHate)	<u>56.91</u>
	+ DAPT + TAPT (AfriEmo)	**57.73**
AfriEmo	AfroXLMR	44.30
	+ DAPT (AfroXLMR-Social)	**51.48**
	+ TAPT (AfriSenti)	49.14
	+ TAPT (AfriHate)	47.12
	+ DAPT + TAPT (AfriSenti)	<u>49.84</u>
	+ DAPT + TAPT (AfriHate)	48.93
AfriHate	AfroXLMR	67.03
	+ DAPT (AfroXLMR-Social)	**70.56**
	+ TAPT (AfriSenti)	67.30
	+ TAPT (AfriEmo)	67.73
	+ DAPT + TAPT (AfriSenti)	66.55
	+ DAPT + TAPT(AfriEmo)	<u>67.18</u>

Table 5: Summary results of Table 4, the average of the Macro F1 score across languages. **Boldface** values are the overall best scores for the specific dataset, and results with <u>underlines</u> are the best cross-TAPT dataset.

Model	AfriSenti	AfriHate	Model	AfriEmo
Gemma-1.1-7B	39.7	24.3	LaBSE	35.7
Llama-2-7B	38.9	21.9	RemBERT	26.8
Llama-3-8B	41.8	27.9	XLM-R	23.4
LLaMAX3-8B	49.8	28.6	mBERT	23.0
Llama-3.1-8B	41.8	23.6	mDeBERTa	26.7
gemma-2-9B	55.5	29.9	Qwen2.5-72B	35.3
Aya-101-13B	57.1	30	Dolly-v2-12B	21.1
gemma-2-27B	58.6	45.5	Mixtral-8x7B	31.4
Llama-3.1-70B	46.9	49	Llama-3.3-70B	38.3
Gemini-1.5 pro	62.6	62.1	DeepSeek-70B	36.6
GPT-4o	62.6	63.5	AfroXLMR	44.3
AfroXLMR-Social	**57.7**	**68.8**	**AfroXLMR-Social**	**51.5**

Table 6: Summary results on fine-tuned models and LLMs. The results show the average Macro F1 score across all languages in the corresponding datasets: AfriSenti - average of 14 languages, AfriHate - average of 15 languages, and AfriEmo is the average of 17 languages.

6.4 AfroXLMR-Social Vs. LLMs

This section compares our AfroXLMR-Social results with state-of-the-art open-source and proprietary Large Language Models (LLMs). Table 6 shows the summarized average results across tasks. AfroXLMR-Social leads the performance over LLMs across the three targeted tasks. Figures 2, 3, and 4 present the comparison results for the base model (AfroXLMR), AfroXLMR-Social, and both open-source and proprietary LLMs across the AfriSenti, AfriEmo, and AfriHate datasets, respectively. The result indicates that the domain-specialized encoder-only model has better or comparable results with LLMs. Generally, it means that we still need encoder-only models for low-resourced African languages and suggests that future LLMs are expected to include more training data for underrepresented languages.

7 Conclusion

This work explored the effects of domain-adaptive pre-training (DAPT) and task-adaptive pre-training (TAPT) across three subjective tasks involving African languages. We created the AfriSocial corpus, a social domain-specific corpus sourced from X and news. Using AfriSocial, we further developed the AfroXLMR-Social language model, which specialized in the social domain. We improved the performance of evaluated tasks, sentiment analysis (AfriSenti), emotion analysis (AfriEmo), and hate speech classification (AfriHate) using DAPT and TAPT techniques. We showed that pre-training the model towards a small domain-specific corpus and related task-relevant data can provide significant improvements. While the combination of the two methods, DAPT + TAPT, also achieved better results than the baseline models, TAPT followed by DAPT would be susceptible to catastrophic forgetting of the task-relevant corpus. We achieved better state-of-the-art results using a small domain-related corpus from the encoder-only model than state-of-the-art large-language models (LLMs). AfriSocial and AfroXLMR-Social will support the development of African languages in the NLP and improve similar-sourced tasks. It opened further domain explorations as the AfriSocial X and news domains are also available separately.

Limitations

Our work is not without limitations. We identify the following limitations with its future suggestions.

Domain Coverage. This work focuses on social media data from X and news sources for downstream tasks that are inherently subjective, such as sentiment analysis, emotion recognition, and hate speech classification. Extending the evaluation to out-of-domain data (e.g., health forums, long-form blogs) and the impacts of topic variations (e.g., politics, sports, business, health) presents another promising avenue for understanding cross-domain generalization in social media–based tasks. For the domain-adaptive pre-training (DAPT) exploration, we utilized a corpus of 3.5M sentences, which exhibits substantial variation in data statistics across different languages.

Evaluation Tasks. In this work, we restrict our evaluation to three subjective tasks—sentiment analysis, emotion recognition, and hate speech classification—in order to highlight the effects of DAPT and TAPT approaches within the social domain. Future work could extend these approaches to a broader range of downstream NLP tasks, including more knowledge-intensive and objective benchmarks such as question answering and machine translation, thereby offering a more comprehensive understanding of their generalizability and impact.

Evaluation of LLMs Assessing the impact of DAPT and TAPT approaches on the latest LLM families—such as Llama, Gemini, GPT, Mistral, and others remains an open direction for future research. In-context learning evaluations, particularly in few-shot settings, provide a promising lens for understanding model behavior, while prompting strategies such as Chain-of-Thought (CoT) reasoning and in-domain prompting have demonstrated notable improvements in LLM performance across various tasks. Systematic evaluation of these techniques in combination with DAPT and TAPT may therefore yield more profound insights and potentially lead to different conclusions regarding the effectiveness and generalizability of such adaptation methods.

Imbalanced Data Across Languages. As illustrated in the AfriSocial corpus (Table 2), there exists substantial variability in the availability of domain-specific data across languages (e.g., 8.6K sentences for Twi versus 866K for Zulu). Investigating the impact of such imbalances on the effectiveness of DAPT and TAPT continual pre-training approaches could yield valuable insights into both the robustness of adaptation techniques and language-specific behaviors. Incorporating more balanced data across languages in future work may further enhance the evaluation and provide a clear understanding of the dynamics of cross-lingual adaptation.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.
Shamsuddeen acknowledges support from a Research Grant under the Nigeria Artificial Intelligence Research (NAIR) Scheme, administered by the National Information Technology Development Agency (NITDA), for developing a dataset for Nigerian languages.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects.

In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised Domain Clusters in Pretrained Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting Pre-trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. In *Proceedings of the Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Tadesse Destaw Belay, Ahmed Haj Ahmed, Alvin Grissom II, Iqra Ameer, Grigori Sidorov, Olga Kolesnikova, and Seid Muhie Yimam. 2025a. CULEMO: Cultural Lenses on Emotion - Benchmarking LLMs for Cross-Cultural Emotion Understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18894–18909, Vienna, Austria. Association for Computational Linguistics.

Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Grigori Sidorov, Dietrich Klakow, Philip Slusallek, Olga Kolesnikova, and Seid Muhie Yimam. 2025b. Evaluating the Capabilities of Large Language Models for Multi-label Emotion Understanding. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3523–3540, Abu Dhabi, UAE. Association for Computational Linguistics.

Tadesse Destaw Belay, Dawit Ketema Gete, Abinew Ali Ayele, Olga Kolesnikova, Grigori Sidorov, and Seid Muhie Yimam. 2025c. Enhancing Multi-Label Emotion Analysis and Corresponding Intensities for Ethiopian Languages.

Steven Bird and Edward Loper. 2004. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Cheng-Han Chiang, Yung-Sung Chuang, and Hung-yi Lee. 2022. Recent Advances in Pre-trained Language Models: Why Do They Work and How Do They Work. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 8–15, Taipei,China. Association for Computational Linguistics.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. *Preprint*, arXiv:2010.12821.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 18 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *ArXiv*, abs/2501.12948.

Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. Broad Twitter Corpus: A Diverse Named Entity Recognition Resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179, Osaka, Japan. International Committee on Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Emezue. 2022. AfroLM: A Self-Active Learning-based Multilingual Pretrained Language Model for 23 African Languages. In *Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 52–64, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, and 1 others. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the Majority is Wrong: Modeling Annotator Disagreement for Subjective Tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Hao Gu, Jiangyan Yi, Zheng Lian, Jianhua Tao, and Xinrui Yan. 2024. NLoPT: N-gram Enhanced Low-Rank Task Adaptive Pre-training for Efficient Language Model Adaption. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12259–12270, Torino, Italia. ELRA and ICCL.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don‘t Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *Preprint*, arXiv:2006.03654.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *Preprint*, arXiv:1904.05342.

OpenAI Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mkadry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alexander Kirillov, Alex Nichol, Alex Paino, and 2 others. 2024. Gpt-4o system card. *ArXiv*, abs/2410.21276.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of Experts. *ArXiv*, abs/2401.04088.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2019. How Language-Neutral is Multilingual BERT? *Preprint*, arXiv:1911.03310.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC‘16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Gemma Team Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, L. Sifre, Morgane Rivière, Mihir Kale, J Christopher Love, Pouya Dehghani Tafti, L'eonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Am'elie H'eliou, Andrea Tacchetti, and 13 others. 2024. Gemma: Open Models Based on Gemini Research and Technology. *ArXiv*, abs/2403.08295.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Saminu Mohammad Aliyu, Paul Röttger, Abigail Oppong, Andiswa Bukula, Chiamaka Ijeoma Chukwuneke, Ebrahim Chekol Jibril, Elyas Abdi Ismail, Esubalew Alemneh, Hagos Tesfahun Gebremichael, Lukman Jibril Aliyu, Meriem Beloucif, Oumaima Hourrane, Rooweither Mabuya, Salomey Osei, and 8 others. 2025a. AfriHate: A Multilingual Collection of Hate Speech and Abusive Language Datasets for African Languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1854–1871, Albuquerque, New Mexico. Association for Computational Linguistics.

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, Alipio Jorge, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, and 8 others. 2023. AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Sa'id Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alípio Jorge, and Pavel Brazdil. 2022. NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 590–602, Marseille, France. European Language Resources Association.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, and 29 others. 2025b. BRIGHTER: BRIdging the Gap in Human-Annotated Textual Emotion Recognition Datasets for 28 Languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8895–8916, Vienna, Austria. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Lima Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Dario Mario Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, and 2 others. 2025c. SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2558–2569, Vienna, Austria. Association for Computational Linguistics.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. AfroBench: How Good are Large Language Models on African Languages? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19048–19095, Vienna, Austria. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. *Preprint*, arXiv:2406.17557.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in NLP. *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 13–20.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1).

Alan Ramponi and Barbara Plank. 2020. Neural Unsupervised Domain Adaptation in NLP—A Survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Ben jamin Lee, and 23 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, abs/2403.05530.

Gemma Team Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L'eonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram'e, Johan Ferret, Peter Liu, Pouya Dehghani Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2408.00118.

Sebastian Ruder and Barbara Plank. 2018. Strong Baselines for Neural Semi-Supervised Learning under Domain Shift. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1044–1054, Melbourne, Australia. Association for Computational Linguistics.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. 2025. Continual Learning of Large Language Models: A Comprehensive Survey. *ACM Comput. Surv.*

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Niko lay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cris tian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas

Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Mingyang Wang, Heike Adel, Lukas Lange, Jannik Strötgen, and Hinrich Schütze. 2023. NLNDE at SemEval-2023 Task 12: Adaptive Pretraining and Source Language Selection for Low-Resource Multilingual Sentiment Analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 488–497, Toronto, Canada. Association for Computational Linguistics.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. Qwen2.5 Technical Report. *ArXiv*, abs/2412.15115.

Seid Muhie Yimam, Hizkiel Mitiku Alemayehu, Abinew Ayele, and Chris Biemann. 2020. Exploring Amharic Sentiment Analysis from Social Media Texts: Building Annotation Tools and Classification Models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1048–1060, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Seid Muhie Yimam, Abinew Ali Ayele, Gopalakrishnan Venkatesh, Ibrahim Gashaw, and Chris Biemann. 2021. Introducing Various Semantic Models for Amharic: Experimentation and Evaluation with Multiple Tasks and Datasets. *Future Internet*, 13(11).

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, and Phil Blunsom. 2019. Learning and Evaluating General Linguistic Intelligence. *Preprint*, arXiv:1901.11373.

Tao Yu and Shafiq Joty. 2021. Effective Fine-Tuning Methods for Cross-lingual Adaptation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8492–8501, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Miaoran Zhang, Mingyang Wang, Jesujoba Alabi, and Dietrich Klakow. 2024. AAdaM at SemEval-2024 Task 1: Augmentation and Adaptation for Multilingual Semantic Textual Relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 800–810, Mexico City, Mexico. Association for Computational Linguistics.

Appendix

A Languages Covered in Evaluation

Table 7 shows the 19 language details we evaluated in this work across different tasks.

Language	ISO	Subregion	Spoken in	Lang. family	Script	# Speakers
Afrikaans	afr	South Africa	South Africa, Namibia, Botswana	Indo-European	Latin	7M
Amharic	amh	East Africa	Ethiopia, Eritrea	Afro-Asiatic	Ethiopic	57M
Algerian Arabic	arq	North Africa	Algeria	Afro-Asiatic	Arabic	36M
Moroccan Arabic	ary	North Africa	Morocco	Afro-Asiatic	Arabic/Latin	29M
Hausa	hau	West Africa	Northern Nigeria, Ghana, Cameroon	Afro-Asiatic	Latin	77M
Igbo	ibo	West Africa	Southeastern Nigeria	Niger-Congo	Latin	31M
Kinyarwanda	kin	East Africa	Rwanda	Niger-Congo	Latin	10M
Oromo	orm	East Africa	Ethiopia	Afro-Asiatic	Latin	37M
Nigerian Pidgin	pcm	West Africa	Nigeria, Ghana, Cameroon	English-Creole	Latin	121M
Mozambican Portug.	ptMZ	Southeastern Africa	Mozambique	Indo-European	Latin	13M
Somali	som	East Africa	Ethiopia, Kenya, Somalia	Afro-Asiatic	Latin	22M
Swahili	swa	East Africa	Tanzania, Kenya, Mozambique	Niger-Congo	Latin	>71M
Tigrinya	tir	East Africa	Ethiopia, Eritrea	Afro-Asiatic	Ethiopic	9M
Twi	twi	West Africa	Ghana	Niger-Congo	Latin	9M
Makhuwa	vmw	East African	Mozambique, Tanzania	Niger–Congo	Latin	7M
Xitsonga	tso	Southern Africa	South Africa, Zimbabwe, Mozambique	Niger-Congo	Latin	7M
Xhosa	xho	Southern Africa	South Africa, Zimbabwe, Lesotho	Niger-Congo	Latin	19M
Yoruba	yor	West Africa	Southwestern, Central Nigeria, Togo	Niger-Congo	Latin	46M
Zulu	zul	Southern Africa	South Africa	Niger-Congo	Latin	29M

Table 7: Additional information on the African languages evaluated in this work: ISO-3 digit language code, region spoken, the family of the language, its script, and number of L1 and L2 speakers.

B Evaluation Data statistics

Table 8 shows the train-test split of the evaluation datasets AfriSenti, AfriEmo, and AfriHate.

Lang.	AfriSenti				AfriEmo				AfriHate			
	Train	Dev	Test	Total	Train	Dev	Test	Total	Train	Dev	Test	Total
afr	-	-	-	-	2107	98	1065	3270	-	-	-	-
amh	5985	1498	2000	9483	3549	592	1774	5915	3467	744	747	4958
arq	1952	415	959	3062	901	100	902	1903	716	211	323	1250
ary	5584	1216	2962	9762	1608	267	812	2687	3240	695	699	4634
hau	14173	2678	5304	22155	2145	356	1080	3581	4566	1029	1049	6644
ibo	10193	1842	3683	15718	2880	479	1444	4803	3419	774	821	5014
kin	3303	828	1027	5158	2451	407	1231	4089	3302	706	714	4722
orm	—	397	2097	2494	3442	575	1721	5738	3517	763	759	5039
pcm	5122	1282	4155	10559	3728	620	1870	6218	7416	1590	1593	10599
ptMZ	3064	768	3663	7495	1546	257	776	2579	—	—	—	—
som	—	—	—	—	3392	566	1696	5654	3174	741	745	4660
swa	1811	454	749	3014	3307	551	1656	5514	14760	3164	3168	21092
tir	—	399	2001	2400	3681	614	1840	6135	3547	760	765	5072
twi	—	—	—	—	—	—	—	—	2564	639	698	3901
vmw	—	—	—	—	— 1551	258	777	2586	—	—	—	—
xho	805	204	255	1264	—	682	1594	2276	2502	559	622	3683
yor	8523	2091	4516	15130	2992	497	1500	4989	3336	724	819	4879
zul	—	—	—	—	—	875	2047	2922	2940	640	728	4308

Table 8: Dataset distribution across different languages - AfriEmotion (train, dev, test, and total) and AfriSenti dataset. We adopt the same train-test-dev split from the data source papers: AfriEmo (Muhammad et al., 2025c), AfriSenti (Muhammad et al., 2023), and AfriHate (Muhammad et al., 2025a).

C Baseline results from encoder-only LMs

We evaluate various encoder-only language models with the more difficult multi-label emotion classification task to select the best encoder-only model for continual learning. We found that AfroXLMR is better for the low-resourced African languages, and we continue our DAPT and TAPT training settings from this model. Table 9 shows baseline results for the AfriEmo dataset from different multilingual encoder-only models.

Model	afr	amh	ary	hau	ibo	kin	orm	pcm	ptMZ	som	swa	tir	vmw	yor	xho	zul
Baselines from general Multilingual models																
LaBSE	**37.76**	63.72	45.81	58.49	45.90	**50.64**	**43.30**	51.30	36.95	41.82	**27.53**	**47.23**	**21.13**	**32.55**	**31.39**	**18.16**
RemBERT	37.14	**63.83**	**47.16**	**59.55**	**47.90**	46.29	12.63	**55.50**	**45.91**	**45.93**	22.65	46.28	12.14	9.22	12.73	15.26
mBERT	26.87	26.69	36.87	47.33	37.23	35.61	39.84	48.42	14.81	31.13	22.99	25.16	10.28	21.03	17.08	13.04
mDeBERTa	16.66	44.22	38.00	48.59	31.92	38.00	28.48	46.21	21.89	34.91	22.84	30.35	11.13	17.88	22.86	13.87
Baselines from African language-centric models																
AfroLM	21.60	54.78	30.35	57.31	42.46	38.97	41.84	47.12	17.81	32.43	20.08	38.22	15.98	24.31	13.67	10.72
AfriBERTa	22.90	60.05	30.85	61.09	**49.05**	46.35	53.69	50.29	**23.15**	44.92	24.36	49.00	**20.29**	34.52	13.86	8.50
AfroXLMR	43.66	68.97	47.62	64.30	26.27	52.39	52.28	55.39	22.09	48.78	30.74	57.22	21.18	28.65	13.52	6.90
Continual pretraining from XLM-RoBERa-Large																
XLMR-L	38.69	54.99	38.31	52.99	38.72	35.06	26.67	53.77	9.29	39.95	6.62	18.58	13.53	2.79	7.90	9.27
+ DAPT	25.94	60.08	44.41	59.81	42.91	44.61	43.61	53.74	30.57	**41.69**	27.83	**49.55**	8.44	18.05	4.75	9.27
+ TAPT	**39.65**	**62.23**	**47.61**	**63.20**	**47.56**	39.09	45.25	**57.78**	**36.68**	38.68	**30.28**	44.74	**14.89**	**26.32**	**10.03**	**12.96**
+ DAPT +TAPT	23.00	60.81	42.91	60.44	43.44	**41.53**	**45.26**	53.90	29.83	40.05	24.02	48.28	9.32	18.53	6.19	4.63
Continual Fine-tuning from AfroXLMR-large																
AfroXLMR	43.66	68.97	47.62	64.30	26.27	52.39	52.28	55.39	22.09	48.78	30.74	57.22	21.18	28.65	13.52	6.90
+ DAPT	44.57	**71.67**	**52.63**	**70.74**	**54.54**	**56.73**	**61.38**	**59.93**	36.80	**54.86**	**34.35**	**60.71**	**22.08**	**39.26**	8.54	6.72
+ TAPT	**49.61**	69.56	50.21	66.93	53.13	53.27	56.43	56.71	37.20	50.33	33.02	55.72	21.73	34.34	4.58	13.15
+ DAPT + TAPT	44.17	71.04	51.63	69.77	54.26	54.47	58.75	58.89	**37.76**	52.65	32.74	57.12	19.80	35.89	21.30	13.81

Table 9: AfriEmo detail results using AfriSocial as DAPT and AfriSenti as TAPT. *xho* and **zul** languages have no training set, and the results are in zero-shot. The results are from the average scores of five runs.

D Training Details of DAFT and TAPT

Hyper-parameters for adapters We trained the task adapter using the following hyper-parameters: batch size of 32, 10 epochs, and learning rate of 5e-5. For TAPT, the parameters are similar to those of DAPT, except that the batch size is 8. We used their tokenizer, for XLMR - XLMR tokenizer, and AfroXLMR - AfroXLMR tokenizer. PyTorch was used for fine-tuning, and pre-trained models were sourced from Hugging Face. The domain-adaptive fine-tuning training is trained on three distributed GPUs for 3 days, whereas TAPT finishes in less than one hour. Following standard practice, we pass the final layer [CLS] token representation to a task-specific feedforward layer for prediction with three epochs. The reported results from fine-tuned pre-trained language models are the average results of five runs.

E Large Language Model Details

Multilingual Encoder-only, open-source, and proprietary model names and their sources are mentioned below. The results from LLMs are used from the work (Ojo et al., 2025; Muhammad et al., 2025b). All open-source LLMs are instruction-tuned versions. The various evaluation prompts are presented in the original works mentioned above.

E.1 Encoder-only Langauge Models

- LaBSE (Feng et al., 2022) - sentence-transformers/LaBSE
- RemBERT (Chung et al., 2021) - google/rembert
- XLM-RoBERTa (Conneau et al., 2020) - FacebookAI/xlm-roberta-base (large)
- mDeBERTa (He et al., 2021) - microsoft/mdeberta-v3-base
- mBERT (Libovický et al., 2019)- google-bert_bert-base-multilingual-cased

- AfriBERTa (Ogueji et al., 2021) - castorini/afriberta_large
- AfroXLMR (Adelani et al., 2024) - Davlan/afro-xlmr-large-61L (76L)
- AfroLM (Dossou et al., 2022)- bonadossou/afrolm_active_learning

E.2 Open-source LLMs

- Aya-101 13B (Üstün et al., 2024) - CohereLabs/aya-101
- Llama 2 7B Chat (Touvron et al., 2023) - meta-llama/Llama-2-7b-chat-hf
- Llama 3 8B (Dubey et al., 2024) - meta-llama/Meta-Llama-3-8B-Instruct
- Llama 3.1 (8B, 70B) (Dubey et al., 2024) - meta-llama/Llama-3.1-8B-Instruct and meta-llama/Llama-3.1-70B-Instruct, respectively.
- Gemma 1.1 7B (Mesnard et al., 2024) - google/gemma-1.1-7b-it
- Gemma 2 (9B, 27B) (Riviere et al., 2024) - google/gemma-2-2b-it and google/gemma-2-27b-it
- DeepSeek-R1-70 (DeepSeek-AI et al., 2025) - deepseek-ai/DeepSeek-R1-Distill-Llama-70B
- Mistral-8x7B (Jiang et al., 2024) - mistralai/Mixtral-8x7B-Instruct-v0.1
- Qwen2.5-72B (Yang et al., 2024) - Qwen/Qwen2.5-72B-Instruct
- Dolly-v2-12B (Conover et al., 2023) - databricks/dolly-v2-12b

E.3 Propritary LLMs

- Gemini 1.5 Pro (Reid et al., 2024) - Gemini 1.5 Pro 002 accessed via Google API
- GPT-4o (Aug) (Hurst et al., 2024) - the August 2024 version of the model is accessed through the OpenAI API

F AfriSenti detail results

Table 10 shows all sentiment analysis results (AfriSenti dataset).

Models	Languages														Avg.
	amh	arq	ary	hau	ibo	kin	orm	pcm	por	swa	tir	tso	twi	yor	
Fine-tuned encoder-only models from the AfroXLMR baseline															
AfroXLMR	50.1	52.2	52.9	79.3	76.9	71.0	20.1	50.5	60.9	28.3	22.5	35.4	47.2	72.3	51.4
AfroXLMR-Social	57.2	64.6	62.3	62.7	79.8	72.7	24.3	52.1	64.8	61.4	24.5	38.8	56.0	74.6	56.9
TAPT-Emo	54.9	62.4	64.1	80.7	80.0	71.5	23.5	51.0	64.1	59.3	10.8	37.9	47.7	72.2	**55.7**
TAPT-Hate	54.5	59.4	52.8	80.1	78.3	69.7	43.0	50.3	62.8	57.3	16.2	36.2	50.2	70.9	55.8
DAPT+TAPT-Emo	55.8	63.4	63.1	82.7	80.4	69.5	28.9	52.0	63.8	54.9	28.9	36.7	54.5	73.7	**57.7**
DAPT+TAPT-Hate	56.3	59.7	62.1	82.0	80.2	70.1	23.6	51.2	62.1	58.5	21.9	40.0	55.8	73.4	56.9
Prompt-based proprietary models															
LLaMAX3-8B	55.2	55.5	51.0	61.7	54.6	53.2	33.6	56.0	41.3	54.1	43.5	48.0	39.0	50.4	49.8
Llama-2-7B	25.5	44.9	44.0	38.2	33.6	35.4	24.7	60.8	31.2	33.8	33.5	46.1	48.9	43.7	38.9
Llama-3.1-70B	40.0	47.5	53.5	52.6	52.2	48.5	41.4	52.6	35.9	61.5	28.2	43.3	45.8	54.3	47.0
Llama-3.1-8B	66.4	57.1	51.9	55.4	50.1	48.7	35.9	64.2	33.6	54.3	49.8	48.8	42.3	50.9	50.7
Llama-3-8B	46.3	51.0	46.1	38.5	36.1	38.4	28.2	60.2	27.9	37.8	38.0	43.3	47.7	45.1	41.8
Aya-101	76.8	67.8	58.1	61.2	47.5	61.1	37.4	70.1	48.8	47.5	71.2	50.8	44.7	57.0	57.1
Gemma-1.1-7B	24.4	43.1	42.0	37.9	34.7	32.0	25.9	66.5	37.4	37.0	32.4	50.0	48.7	43.8	39.7
Gemma-2-27B	70.7	65.8	59.0	64.8	60.4	59.1	37.3	76.0	42.8	55.6	58.9	50.0	54.3	65.5	**58.6**
Gemma-2-9B	70.1	62.0	56.4	61.4	58.2	56.1	37.9	66.8	46.6	58.7	55.4	43.7	48.1	55.4	55.5
Prompt-based proprietary models															
Gemini-1.5	77.5	70.9	63.7	70.1	56.9	68.3	42.8	74.5	46.4	55.2	70.2	55.9	49.3	74.3	**62.6**
GPT-4o	75.6	72.3	61.2	68.6	67.8	71.6	43.1	67.1	62.1	57.9	61.5	46.5	51.3	70.2	**62.6**

Table 10: AfriSenti Model Performance Across Various Languages

G AfriHate results

Table 11 shows all hate speech classification results (AfriHate dataset).

Models	Languages															Avg.
	amh	arq	ary	hau	ibo	kin	orm	pcm	som	swa	tir	twi	xho	yor	zul	
Fine-tuned encoder-only models from the AfroXLMR baseline																
AfroXLMR	73.5	43.4	75.1	81.6	82.8	75.3	67.2	64.9	55.7	91.5	50.2	46.9	50.9	53.4	54.5	64.5
DAPT	78.6	46.0	75.6	80.8	88.1	78.8	74.1	67.6	55.6	91.2	55.9	48.4	59.2	77.9	55.4	**68.9**
TAPT-Emo	73.1	46.2	75.1	77.8	87.7	77.5	69.0	67.7	57.3	91.9	54.2	49.3	55.1	53.7	56.0	66.1
TAPT-Senti	73.3	43.8	71.7	81.9	86.5	76.3	67.7	66.9	55.0	91.3	57.0	48.9	51.3	54.5	55.0	65.4
DAPT + TAPT-Emo	78.1	44.2	77.5	82.1	87.7	78.4	71.6	69.9	56.8	91.2	32.7	49.0	58.7	54.8	55.5	65.9
DAPT + TAPT-Senti	77.6	43.8	76.1	81.6	79.4	77.9	72.0	67.5	53.7	91.5	41.2	48.2	54.6	54.8	55.8	65.0
Prompt-based proprietary models																
Gemma-1.1-7B	23.0	27.4	24.5	26.0	16.7	29.9	27.9	30.2	27.2	27.4	17.3	14.2	23.3	25.0	22.5	24.3
Llama-2-7B	14.5	22.4	22.2	24.4	20.2	22.4	31.3	9.4	27.1	24.8	11.7	15.8	24.8	23.1	26.8	21.9
Llama-3-8B	26.5	31.8	28.5	24.5	19.7	36.5	37.1	38.8	17.8	34.3	28.4	14.4	25.0	25.9	28.4	27.9
LLaMAX3-8B	37.2	33.6	31.5	30.7	19.4	38.2	38.2	34.4	27.6	28.9	27.4	13.9	23.7	24.4	29.0	28.6
Llama-3.1-8B	23.3	30.7	22.9	25.4	13.9	31.9	35.7	24.9	26.7	21.7	21.9	9.9	22.4	19.4	23.3	23.6
gemma-2-9b	33.2	33.8	33.2	24.1	25.1	33.6	26.7	54.9	13.6	46.4	26.8	29.1	20.0	30.5	20.1	29.9
Aya-101-13B	31.3	32.1	28.9	33.3	22.1	32.8	26.8	37.8	35.8	41.3	29.6	13.8	28.7	26.8	29.8	30.0
Gemma-2-27B	48.4	49.1	53.8	34.8	42.8	52.7	39.8	60.9	39.6	70.9	35.4	38.1	30.6	54.0	35.0	45.5
Llama-3.1-70B	53.0	57.0	60.6	41.2	48.4	50.9	44.6	62.4	39.8	67.0	41.0	37.9	32.7	56.2	46.3	**49.0**
Prompt-based proprietary models																
Gemini-1.5 pro	56.1	70.6	68.2	61.4	66.9	64.2	57.6	65.0	60.8	80.5	37.5	50.6	58.0	73.1	55.4	62.1
GPT-4o	56.0	69.7	75.5	59.2	69.7	60.1	53.5	65.2	68.5	78.0	42.4	51.2	63.7	74.5	58.7	**63.5**

Table 11: AfriHate Model Performance Across Various Languages

H AfriEmo detail results

Table 12 shows all fine-grained multi-label emotion classification results (AfriEmo dataset).

Models	afr	amh	arq	ary	hau	ibo	kin	orm	pcm	ptMZ	som	swa	tir	vmw	yor	xho	zul	Avg.
Fine-tuned encoder-only models from the AfroXLMR baseline and others																		
AfroXLMR	43.7	69.0	44.9	47.6	64.3	26.3	52.4	52.3	55.4	22.1	48.8	30.7	57.2	21.2	28.7	13.5	6.9	40.3
+ DAPT	44.6	71.7	51.3	52.6	70.7	54.5	56.7	61.4	59.9	36.8	54.9	34.4	60.7	22.1	39.9	8.5	6.7	**46.3**
+ TAPT-Senti	49.6	69.6	49.0	50.2	66.9	53.1	53.3	56.4	56.7	37.2	50.3	33.0	55.7	21.7	34.3	14.6	13.2	45.0
+ TAPT-Hate	47.2	67.2	48.4	48.8	61.2	51.3	53.5	52.2	56.6	31.0	49.6	31.9	55.8	19.3	32.9	11.0	9.9	42.8
+ DAPT + TAPT-Senti	44.2	71.0	48.7	51.6	69.8	54.3	54.5	58.8	58.9	37.8	52.7	32.7	57.1	19.8	35.9	21.3	13.8	46.1
+ DAPT + TAPT-hate	46.2	70.8	46.6	48.8	70.0	54.2	53.5	56.8	58.2	34.8	52.7	31.4	57.1	19.7	33.0	9.4	5.0	44.0
LaBSE	35.1	63.7	35.9	42.8	38.5	18.1	30.4	43.3	33.3	31.4	41.8	21.7	47.2	9.7	11.6	31.4	18.2	32.6
RemBERT	35.0	63.8	33.8	35.5	32.0	7.5	18.4	12.6	1.0	29.7	45.9	19.0	46.3	5.2	5.3	12.7	15.3	24.7
XLM-R	41.7	46.9	35.9	33.9	16.7	10.4	13.1	17.9	21.1	7.3	25.4	16.9	35.9	12.7	6.6	11.5	10.9	21.5
mBERT	17.0	26.7	31.4	24.8	15.6	9.9	20.9	39.8	22.6	13.5	31.1	18.6	25.2	12.1	9.6	17.1	13.0	20.5
mDeBERTa	33.3	44.2	35.9	36.3	32.8	9.5	17.3	28.5	25.4	24.5	34.9	14.9	30.4	11.7	10.0	22.9	13.9	25.1
Prompt-based proprietary models																		
Qwen2.5-72B	60.2	37.8	37.8	52.8	43.8	37.4	32.0	31.6	38.7	40.4	28.6	27.4	31.1	20.4	25.0	29.6	22.0	35.1
Dolly-v2-12B	23.6	5.1	38.6	24.3	29.4	24.3	19.7	22.9	34.4	16.7	19.8	17.6	1.5	16.0	16.0	24.1	14.7	20.5
Llama-3.3-70B	61.3	42.8	55.8	45.0	50.9	33.2	34.4	29.8	48.7	34.1	32.5	29.5	32.9	19.0	23.7	30.8	21.5	**36.8**
Mixtral-8x7B	53.7	29.0	45.3	35.1	40.4	31.9	26.4	24.3	45.6	36.5	25.6	26.5	27.2	19.0	19.7	22.9	20.4	31.1
DeepSeek-R1-70B	43.7	36.9	50.9	47.2	51.9	32.9	32.5	28.2	45.0	39.6	26.6	33.3	26.5	19.1	27.4	29.1	20.4	34.8

Table 12: AfriEmo Model Performance Across Various Languages

I AfriSocial Data Sources

X (Twitter) Source There is an X domain corpus and model for high-resource languages such as XLM-T (Barbieri et al., 2022) to evaluate and improve task datasets sourced from X. However, there is a scarcity of corpora specializing in the social domain for low-resource African languages. The tweets are scraped over a different time until June 2023, before the change of an X policy that restricts their data for academic research.

News Source News platforms are the most common data source for African languages. Companies also stream their news on the X platform. While more formal, news websites also provide a platform for public discourse, comments, and reactions, often including opinion pieces and user-generated comments. The source news websites are British Broadcasting Corporation (BBC) news[5], Akan news[6], Global Voice News[7], isolezwelesixhosa[8], isolezwe[9], and others.

[5] https://www.bbc.com/
[6] https://akannews.com
[7] https://mg.globalvoices.org/
[8] https://www.isolezwelesixhosa.co.za/
[9] https://www.isolezwe.co.za/