UBC-DLNLP at SemEval-2023 Task 12: Impact of Transfer Learning on African Sentiment Analysis

Gagan Bhatia[1,⋆] **Ife Adebara**[1,⋆] **AbdelRahim Elmadany**[1] **Muhammad Abdul-Mageed**[1,2]
[1]Deep Learning & Natural Language Processing Group, The University of British Columbia
[2]Department of Natural Language Processing & Department of Machine Learning, MBZUAI
{gagan30@student.,ife.adebara@,a.elmadany@,muhammad.mageed@}ubc.ca

Abstract

We describe our contribution to the SemEVAl 2023 AfriSenti-SemEval shared task, where we tackle the task of sentiment analysis in 14 different African languages. We develop both monolingual and multilingual models under a full supervised setting (subtasks A and B). We also develop models for the zero-shot setting (subtask C). Our approach involves experimenting with transfer learning using six language models, including further pretraining of some of these models as well as a final finetuning stage. Our best performing models achieve an F_1-score of 70.36 on development data and an F_1-score of 66.13 on test data. Unsurprisingly, our results demonstrate the effectiveness of transfer learning and finetuning techniques for sentiment analysis across multiple languages. Our approach can be applied to other sentiment analysis tasks in different languages and domains.

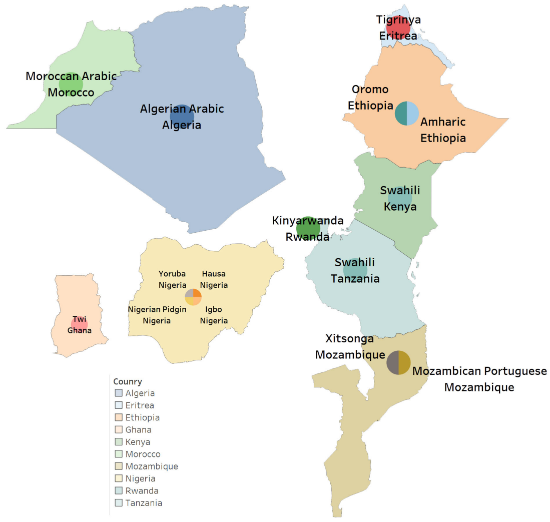

Figure 1: A map showing the countries where each language in the shared task is spoken in Africa.

1 Introduction

Sentiment Analysis, also referred to as opinion mining, is a Natural Language Processing (NLP) technique that aims to identify, extract, and evaluate opinions, attitudes, perceptions, and sentiments towards topics, products, services, and individuals from textual data (Birjali et al., 2021). With the increased accessibility of the internet, people are increasingly sharing their opinions on various platforms, such as forums, blogs, wikis, websites, and social media pages. Consequently, there is a need for the automatic extraction of sentiments to gain valuable insights into user perception, popular opinion, and trends (Georgiadou et al., 2020; Ramírez-Tinoco et al., 2018).

Despite the increasing popularity of sentiment analysis, its application in low-resource African languages is still under-explored (Shode et al., 2022; Diallo et al., 2021). This is because many African languages have limited digital resources, such as annotated data and lexical resources, which can hinder the development and evaluation of sentiment analysis models. So far, only a handful of African languages have few datasets for sentiment analysis (Imam Abubakar et al., 2021; Ogbuju and Onyesolu, 2019; Muhammad et al., 2023c,a; Oyewusi et al., 2020; Muhammad et al., 2022). Furthermore, African languages often exhibit complex morphology, syntax, semantics, stylistics, pragmatic and orthographic conventions including the use of diacritics, and code-mixing that can make it difficult to accurately identify and extract sentiment from text data (Muhammad et al., 2023c,a; Orimaye et al., 2012). For instance, for some African languages, a single change in tone assignment can change the sentiment of a text (Adebara and Abdul-Mageed, 2022).

In this task, we conduct sentiment analysis on 14 African languages including Algerian Arabic, Amharic, Hausa, Igbo, Kinyarwanda, Moroccan Arabic, Mozambican Portuguese, Nigerian Pidgin, Oromo, Swahili, Tigrinya, Twi, and Yoruba. The

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 246–255
July 13-14, 2023 ©2023 Association for Computational Linguistics

sentiment analysis data used for this shared task is the largest and most multilingual dataset for sentiment analysis for African languages to date (Muhammad et al., 2023c,a).

Our contribution is as follows:

1. We show the utility of finetuning six language models for sentiment analysis on 14 African languages.

2. We show the utility of further pretraining two language models for sentiment analysis on 14 African languages.

3. We show the performance of our models in zero-shot settings.

The rest of this paper is organized as follows: We discuss existing literature in Section 2, and provide background information in Section 3. Section 4 has details about the models we develop. In Section 5 we describe each experiment performed and show the results on Dev. and Test sets in Section 6. We conclude in Section 7.

2 Literature Review

2.1 Sentiment Analysis

Sentiment analysis can be conceptualized as a text classification problem, where the sentiment of the text is classified into one of three categories: negative, neutral, or positive. Different levels of sentiment analysis include document level (Behdenna et al., 2016), sentence level, and aspect level (Do et al., 2019; Xue and Li, 2018). Document level analysis focuses on the overall sentiment of a text, whereas sentence level analysis evaluates sentiment on a more fine-grained level. Aspect level analysis focuses on specific features in the text.

The methods for sentiment analysis have evolved rapidly, from rule-based approaches (Turney, 2002) to machine learning, deep learning, and hybrid methods (Akhtar et al., 2016). Rule-based methods rely on identifying polarity items (Wilson et al., 2005; Medhaffar et al., 2017), punctuation, and other linguistic features to determine sentiment. Although these methods are easy to interpret and implement, developing rules can be tedious, expensive, and lack scalability. Machine learning approaches like support vector machines and Naive Bayes learn from labeled data to predict sentiment in new, unlabeled text. Deep learning methods, including convolutional neural networks (dos Santos and Gatti, 2014; Xue and Li, 2018), transformers, and transfer learning approaches (Baert et al., 2020; Sun et al., 2019; Hosseini-Asl et al., 2022), have achieved state-of-the-art performance in sentiment analysis. In hybrid methods (Akhtar et al., 2016), two or more of the aforementioned methods are combined for sentiment analysis. Hybrid methods and Transfer learning methods are able to achieve high accuracy in low resource scenarios.

2.2 Transfer Learning

Transfer learning (Raffel et al., 2020; He et al., 2022; Ruder et al., 2019; Ruder, 2022) is an integral part of modern NLP systems. Transfer learning attempts to transfer knowledge from other sources to benefit a current task; based on the premise that previous knowledge may improve solutions for a current task (Pan and Yang, 2010). It allows the domains, tasks, and distributions used in training and testing to be different, enabling a new task to leverage previously acquired domain knowledge. Potential benefits include faster learning, better generalization, and a more robust system. It has significantly improved state of the art in natural language generation (NLG) and natural language understanding (NLU) tasks of which Sentiment Analysis is one. Transfer learning, through the use of large transformer models have enabled the use of low-resource languages through finetuned on various NLP tasks.

In monolingual settings, transfer learning involves using pre-trained models on data in one language while multilingual transfer learning involves using pre-trained models on large datasets in multiple languages (Přibáň and Steinberger, 2021). The multilingual transfer learning approach takes advantage of the fact that many languages share similar structures and patterns, which can be leveraged to improve performance in low resource languages (Ruder et al., 2019; Ruder, 2022). In this work, we experiment with language models that have representations of some African languages to transfer representations for our sentiment analysis task. We also experiment with monolingual and multilingual settings. In addition, we perform two experiments in zero-shot settings.

2.3 African NLP

Africa is home to over 2,000 Indigenous languages, which represents about one-third of all languages spoken globally (Eberhard et al., 2021). Despite this, most of these languages have not received much attention in the field of Natural Language

Processing (NLP). Unfortunately, the majority of NLP research has focused on higher-resource languages, which are typologically distinct from Indigenous African languages. The methods used to develop NLP technologies for these languages have been Western-centric, making them challenging to apply directly to African languages (Adebara and Abdul-Mageed, 2022). Additionally, existing NLP technologies function within the context of Western values and beliefs, which poses unique challenges when these technologies are applied within African communities.

To address this language bias problem, an Afrocentric approach to technology development is crucial for African languages. Such an approach would entail developing technologies that meet the needs of local African communities (Adebara and Abdul-Mageed, 2022). Several NLP It would involve not only deciding what technologies to build but also determining how to build, evaluate, and deploy them (Adebara et al., 2022a). By adopting an Afrocentric approach, NLP researchers and practitioners can help to bridge the digital divide and ensure that language technologies are accessible to African communities.

3 Approach

We perform sentiment analysis on three different subtasks with 14 languages spoken across Africa. The languages are quite diverse belonging to four different language families and written in different scripts including Arabic, Ethiopic, and Latin scripts. We provide details about the languages and the datasets.

3.1 Datasets

This study utilizes Twitter datasets provided for the SemEVAl 2023 AfriSenti-SemEval shared task (Muhammad et al., 2023b). The dataset comprises three subtasks, each with a different focus on sentiment analysis. **Subtask A** consists of monolingual datasets for 12 different languages, each labeled as positive, negative, or neutral. **Subtask B** involves a multilingual sentiment analysis system, with multilingual data for the 12 languages in Task A. **Subtask C** provides unlabeled data for two African languages (Tigrinya and Oromo), and participants are expected to develop a zero-shot model for sentiment analysis in these languages. The dataset statistics for each language are presented in detail in Table 1. The use of Twitter datasets enables the evaluation of sentiment analysis models on real-world data, providing insights into the effectiveness of different approaches for sentiment analysis in a multilingual context. We provide details of each language in Table 3 and Section A. For preprocessing, we remove all URLs and tokenize with wordpiece.

Subsets	Subtask	Train	Dev	Test
am	A	8,978	1,498	2,000
dz	A	2,479	415	959
ha	A	19,526	2,678	5,304
ig	A	13,874	1,842	3,683
kr	A	4,956	828	1,027
ma	A	8,013	495	2,962
sw	A	2,716	454	749
pcm	A	7,683	1,282	4,155
pt	A	4,597	768	3,663
ts	A	1,210	204	255
twi	A	4,257	389	950
yo	A	12,702	2,091	4,516
multilingual	B	90,991	13,654	30,212
or	C	——	397	2,097
tg	C	——	399	2,001

Table 1: Statistics of data for each language across the three tasks. **am**: Amharic, **dz**: Algerian Arabic, **ha**: Hausa, **ig**: Igbo, **kr**: Kinyarwanda , **ma**: Darija, **sw**: Swahili, **pcm**: Nigerian Pidgin, **pt**: Mozambican Portuguese, **ts**: Xitsonga (Mozambique Dialect), **twi**: Twi, **yo**: Yoruba, **or**: Oromo, **tg**: Tigrinya.

3.2 Code and Script Switching

We found examples of code-switching and script switching in the data used for training. Moroccan Arabic data for instance had both Arabic and Latin scripts examples. We also found code-mixing with English in the Hausa, Igbo, Twi, Swahili, and Yoruba and code-mixing with French in the Algerian Arabic examples.

4 System Overview

In order to identify the best-performing model for our datasets, we first finetuned 6 LMs on the data from sub tasks A and B. Specifically, we finetuned mBERT, XLM-R, Afro-XLMR, AfriBERTa, AfriTEVA, and Serengeti. We also further pretrained Afro-XLMR and Serengeti. We refer to the pre-trained models as Afro-XLMR-LM and Serengeti-LM, respectively. We provide further details for each of the LMS in what follows.

4.1 Models

4.1.1 XLM-R

XLM-R (Conneau et al., 2019) is an encoder-only model based on RoBERTa. It was pretrained on a corpus of 100 languages, of which only 8 were African. Namely Afrikaans, Amharic, Hausa, Oromo, Somali, Swahili, Xhosa, out of which Oromo, Hausa and Swahili are part of the shared task. We use finetune both base and large models.

4.1.2 mBERT

mBERT (Devlin et al., 2018) is a multilingual variant of BERT pretrained on 104 languages. Out of these 104 languages only 4 languages are African out of which Swahili and Yoruba are part of this shared task. mBERT was pre-trained using masked language modeling (MLM) and next-sentence prediction task. We finetune the base model.

4.1.3 Afro-XLMR

Afro-XLM-R (Alabi et al., 2022) uses language adaptation on the 17 most-resourced African languages and three other high-resource foreign languages widely used in Africa – English, French, and Arabic – simultaneously to provide a single model for cross-lingual transfer learning for African languages. Afro-XLM-R has Afrikaans, Amharic, Hausa, Igbo, Malagasy, Chichewa, Oromo, Nigerian Pidgin, Kinyarwanda, Kirundi, Shona, Somali, Sesotho, Swahili, isiXhosa, Yoruba, and isiZulu. Out of which we have Amharic, Hausa, Igbo, Oromo, Nigerian Pidgin, Kinyarwanda, Swahili and Yoruba are in the shared task. We finetune the base and large models.

4.1.4 AfriBERTa

AfriBERTa is a language model that supports 11 African languages, including Afaan Oromoo, Amharic, Gahuza (a code-mixed language of Kinyarwanda and Kirundi), Hausa, Igbo, Nigerian Pidgin, Somali, Swahili, Tigrinya, and Yoruba (Ogueji et al., 2021). The pretraining corpus for this model is small (only 108.8 million tokens), when compared to many other language models). AfriBERTa is trained using a Transformer with the standard masked language modelling objective. The AfriBERTa model uses 6 attention heads, 768 hidden units, 3072 feed forward dimensions, and a maximum length of 512 for the 3 configurations of the model. We finetune the base and large models.

4.1.5 Serengeti

Serengeti is an XLM-R based model on pretrained on 517 African languages, the largest number of African languages in a single model (Adebara et al., 2022b).

4.1.6 Afro-XLMR$_{\text{ft}}$

Afro-XLMR$_{\text{ft}}$ is further pretrained using MLM objective on the training data for all tasks. We pretrain for 75 epochs to improve the performance on the sentiment analysis task.

4.1.7 Serengeti$_{\text{ft}}$

Serengeti$_{\text{ft}}$is further pretrained using MLM objective on the training data for all tasks. We pretrain for 75 epochs to improve the performance on the sentiment analysis task.

5 Experimental Setup

All our models are implemented using the PyTorch framework and the open-source Huggingface Transformers libraries. All the models were trained on a single Nvidia A100. All our models are trained using Adam optimizer with a linear learning rate scheduler. After hyperparameter tuning using Optuna, it was found that the optimum learning rate, batch size and number of epochs is $5 * e^{-}5$, 16 and 50 respectively. For the focal loss, the hyperparameters γ and α are set to 2 and 0.8, respectively. All models are evaluated on the Weighted F_1 Metric which was also used the objective for fine-tuning.

For further pretraining of Serengeti and Afro-XLMR we used a more aggressive learning rate of $4 * e^4$ using a batch size of 16 for 75 epochs.

6 Results

We show the results on the Dev. set for each model in Table 4 and results on the Test set in Table 2. The official results from the shared task is labelled as M11 in Table 2. Afro-XLMR-base$_{\text{ft}}$ (M9) outperforms other models on 5 languages with an average F_1 score of 70.36 in the Dev. set. Serengeti$_{\text{ft}}$ (M10) has the second highest performance with an average F_1 score of 69.59 and achieving best performance on 3 languages on Dev. set. For the Test set, Afro-XLMR-base$_{\text{ft}}$ (M9) outperforms other models on 9 languages with an average F_1 score of 66.13 while Serengeti$_{\text{ft}}$ (M10) has the second highest performance with an average F_1 score of 64.97 and best performance on 1 language.

Lang.	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	Rank
yo	61.65	25.33	65.17	25.33	71.02	72.53	73.88	69.63	**75.060**	74.82	71.02	20th
twi	49.58	30.51	60.18	30.51	63.46	65.74	65.24	46.86	**65.950**	65.73	65.14	12th
ts	35.42	30.74	51.05	30.74	45.49	53.07	49.82	35.18	51.62	**54.970**	45.49	28th
sw	45.02	44.22	51.95	44.22	58.60	**62.820**	60.58	60.87	62.09	60.40	58.60	20th
pt	67.37	51.17	63.39	51.17	65.64	57.19	58.37	61.07	**70.670**	61.98	61.98	27th
pcm	66.24	40.20	40.20	40.20	67.68	64.22	62.99	61.93	**69.500**	65.57	65.57	21st
ma	52.75	21.53	45.14	21.53	48.11	40.60	45.24	42.67	**59.520**	53.06	53.06	22nd
kr	53.80	21.22	57.27	21.22	67.56	64.12	62.02	65.24	**69.590**	64.94	62.02	23rd
ig	75.89	26.91	75.79	26.91	77.52	78.41	79.24	71.87	**79.630**	79.31	77.52	17th
ha	73.49	17.02	73.18	17.02	77.60	79.37	78.00	77.30	**79.380**	79.37	79.37	18th
dz	59.30	32.87	61.45	32.87	64.02	44.35	35.96	37.27	**66.570**	60.45	64.02	20th
am	60.47	2.26	2.36	2.26	56.88	**61.630**	60.62	53.77	43.95	59.02	56.88	19th
Average	58.42	28.66	53.93	28.66	63.63	62.00	61.00	56.97	**66.13**	64.97	63.39	
multilingual	61.43	17.06	17.06	17.06	68.69	64.84	65.64	65.60	**69.030**	67.89	69.03	-
or	36.00	15.15	15.15	15.15	43.97	**50.720**	49.78	38.20	44.98	45.27	41.79	14th
tg	38.91	14.38	14.38	14.38	54.38	40.70	45.24	**57.720**	56.64	45.73	57.03	19th

Table 2: Results of Model Performance and Rank on Test Set. **M1:** xlmr-base, **M2:** xlmr-large, **M3:** mbert-base-cased, **M4:** afro-xlmr-large, **M5:** afro-xlmr-base, **M6:** afriberta_large, **M7:** afriberta_base, **M8:** serengeti, **M9:** afro-xlmr-base$_{\text{ft}}$, **M10:** serengeti$_{\text{ft}}$, **M11:** Official shared task results with Serengeti model

6.1 Further-Pretraining

We find significant improvement in model performance after pre-training when compared to fine-tuning. For all but two languages, the further pre-trained LMs - Afro-XLMR-base$_{\text{ft}}$ (M9) and Serengeti$_{\text{ft}}$ (M10) outperform their fine-tune counterparts Afro-XLMR-base (M5) and Serengeti (M8). Our findings corroborates research that further pretraining encodes shallow domain knowledge that has influence in low resource scenarios. This is said to be beneficial for providing task specific knowledge for fine-tuning (Zhu et al., 2021).

6.2 Multi-Lingual Settings

In multilingual settings, we find that each model achieves F_1 scores higher than the average on individual languages. Our finding corroborates research that multilingual training can even achieve better performance than monolingual training, especially for low-resource languages (Přibáň and Steinberger, 2021).

6.3 Zero Shot Settings

In the zero-shot settings with Oromo and Tigrinya, AfriBERTa-large outperforms other models on Oromo while Serengeti outperforms other models on Tigrinya. In both languages, the further-pretrained models do not achieve best performance. Although further-pretraining improves the performance on Oromo, further-pretraining hurt Serengeti's the performance on Tigrinya.

7 Conclusion

We reported our participation in the three sub-stacks for the AfriSenti-SemEval 2023 shared task. We described our transfer learning approaches using finetuning and further pretraining of existing LMs. We show the performance of our models across the 14 languages in the three subtask.

8 Acknowledgement

We gratefully acknowledge support from the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 435-2018-0576; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Compute Canada (CC),[1] UBC ARC-Sockeye,[2] and Advanced Micro Devices, Inc. (AMD). Any opinions, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSERC, SSHRC, CFI, CC, AMD, or UBC ARC-Sockeye.

References

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric NLP for African languages: Where we are and where we can go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841, Dublin, Ireland. Association for Computational Linguistics.

[1]https://www.computecanada.ca
[2]https://arc.ubc.ca/ubc-arc-sockeye

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Inciarte. 2022a. AfroLID: A neural language identification tool for African languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1958–1981, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2022b. Serengeti: Massively multilingual language models for africa.

Md Shad Akhtar, Ayush Kumar, Asif Ekbal, and Pushpak Bhattacharyya. 2016. A hybrid deep learning architecture for sentiment analysis. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 482–493, Osaka, Japan. The COLING 2016 Organizing Committee.

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.

Gaétan Baert, Souhir Gahbiche, Guillaume Gadek, and Alexandre Pauchet. 2020. Arabizi language models for sentiment analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 592–603, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Salima Behdenna, Fatiha Barigou, and Ghalem Belalem. 2016. Sentiment analysis at document level. *Communications in Computer and Information Science*, 628 CCIS:159 – 168. Cited by: 16.

Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Mountaga Diallo, Chayma Fourati, and Hatem Haddad. 2021. Bambara language dataset for sentiment analysis.

Hai Ha Do, P.W.C. Prasad, Angelika Maag, and Abeer Alsadoon. 2019. Deep learning for aspect-based sentiment analysis: A comparative review. *Expert Systems with Applications*, 118:272 – 299. Cited by: 262.

Cícero dos Santos and Maíra Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 69–78, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

David M Eberhard, F Simons Gary, and Charles D Fennig (eds). 2021. Ethnologue: Languages of the world. *Twenty-fourth edition*, Dallas, Texas: SIL International.

Elena Georgiadou, Spyros Angelopoulos, and Helen Drake. 2020. Big data analytics and international negotiations: Sentiment analysis of brexit negotiating outcomes. *International Journal of Information Management*, 51:102048.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*.

Ehsan Hosseini-Asl, Wenhao Liu, and Caiming Xiong. 2022. A generative language model for few-shot aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 770–787, Seattle, United States. Association for Computational Linguistics.

Amina Imam Abubakar, Abubakar Roko, Aminu Bui, and Ibrahim Saidu. 2021. An enhanced feature acquisition for sentiment analysis of english and hausa tweets. *International Journal of Advanced Computer Science and Applications*, 12.

Salima Medhaffar, Fethi Bougares, Yannick Estève, and Lamia Hadrich-Belguith. 2017. Sentiment analysis of Tunisian dialects: Linguistic ressources and experiments. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 55–61, Valencia, Spain. 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 Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino Dário Mário António Ali, Davis Davis, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Steven Arthur. 2023a. AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages.

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 Gebremichael, Bernard Opoku, and Steven Arthur. 2023b. AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Sa'id Ahmad, Nedjma Ousidhoum, Abinew Ali Ayele, Saif M. Mohammad, Meriem Beloucif, and Sebastian Ruder. 2023c. SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. 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.

Emeka Ogbuju and Moses Onyesolu. 2019. Development of a general purpose sentiment lexicon for Igbo language. In *Proceedings of the 2019 Workshop on Widening NLP*, page 1, Florence, Italy. 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.

Sylvester Olubolu Orimaye, Michael Rabinovich, Saadat Mehmood Alhashmi, Steffen Staab, and Eu-Gene Siew. 2012. Sentiment analysis amidst ambiguities in youtube comments on yoruba language (nollywood) movies. pages 583 – 584. International World Wide Web Conference 2012, WWW 2012 ; Conference date: 16-04-2012 Through 20-04-2012.

Wuraola Fisayo Oyewusi, Olubayo Adekanmbi, and Olalekan Akinsande. 2020. Semantic enrichment of nigerian pidgin english for contextual sentiment classification.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Pavel Přibáň and Josef Steinberger. 2021. Are the multilingual models better? improving czech sentiment with transformers. *CoRR*, abs/2108.10640.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Francisco Javier Ramírez-Tinoco, Giner Alor-Hernández, José Luis Sánchez-Cervantes, Beatriz Alejandra Olivares-Zepahua, and Lisbeth Rodríguez-Mazahua. 2018. A brief review on the use of sentiment analysis approaches in social networks. In *Trends and Applications in Software Engineering*, pages 263–273, Cham. Springer International Publishing.

Sebastian Ruder. 2022. The State of Multilingual AI. `http://ruder.io/state-of-multilingual-ai/`.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Iyanuoluwa Shode, David Ifeoluwa Adelani, and Anna Feldman. 2022. YOSM: A new Yorùbá Sentiment Corpus for Movie Reviews. *AfricaNLP 2022 @ICLR*.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. 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 380–385, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2514–2523, Melbourne, Australia. Association for Computational Linguistics.

Qi Zhu, Yuxian Gu, Lingxiao Luo, Bing Li, Cheng Li, Wei Peng, Minlie Huang, and Xiaoyan Zhu. 2021. When does further pre-training MLM help? an empirical study on task-oriented dialog pre-training. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 54–61, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Appendix

A.1 Hausa

Hausa is a Chadic language spoken by over 50 million people in West Africa. It is tonal, with a diverse vocabulary influenced by Arabic, Fula, and English. Hausa has a long literary tradition, written in a modified Arabic script. It is an important lingua franca and cultural language in West Africa.

A.2 Yoruba

Yoruba is a tonal, complex language spoken in Nigeria by over 20 million people. It has a rich vocabulary, oral tradition, and unique script. It conveys meaning through three distinct tones and a noun class system.

A.3 Igbo

Igbo is a tonal language spoken in Nigeria by over 20 million people. It has a rich oral tradition, expressive vocabulary and unique writing system. It conveys meaning through tone variation and has complex sentence structure.

A.4 Nigerian Pidgin

Nigerian Pidgin is a creole language that blends English with African languages. It's widely spoken in Nigeria as a lingua franca and has its own unique grammar, vocabulary and pronunciation.

A.5 Amharic

Amharic is a Semitic language spoken in Ethiopia by over 22 million people. It uses the Ethiopian script and is characteristically known for its unique sounds and tonal patterns.

A.6 Tigrinya

Tigrinya is a Semitic language spoken in Eritrea and Ethiopia by over 6 million people. It uses a unique script called "Ge'ez" and has a rich oral tradition. Tigrinya is characterized by its distinctive vowel harmonies and use of suffixes.

A.7 Oromo

Oromo is a Cushitic language spoken in Ethiopia and Kenya by over 30 million people. It has a unique alphabet called "Qubee" and a rich oral tradition, including folktales and traditional songs. Oromo is characterized by its tonal system and use of suffixes to convey grammatical relationships.

A.8 Swahili

Swahili is a Bantu language widely spoken in East Africa, particularly in Kenya and Tanzania. It uses the Latin script and has loanwords from Arabic, Portuguese, and English. Swahili has many variations and dialects, with a rich oral tradition of poetry and song. It is a tonal language, with two distinctive tones that change the meaning of words.

A.9 Algerian Arabic

Algerian Arabic is a dialect of Arabic spoken in Algeria. It is characterized by its unique vocabulary, pronunciation, and grammar, as well as the influence of Berber and French. It is written in the Arabic script.

A.10 Moroccan Arabic

Moroccan Arabic, also known as Darija, is a Arabic dialect spoken in Morocco. It has Berber, French, and Spanish influences and uses the Arabic script. Darija is known for its unique pronunciation, vocabulary, and grammar, making it distinct from Standard Arabic.

A.11 Kinyarwanda

Kinyarwanda is a Bantu language spoken in Rwanda and Uganda. It uses a unique script called "Kirundi" and has a complex noun class system. It also has a rich oral tradition, with proverbs playing a significant role in the language and culture. Kinyarwanda is characterized by its use of tone to convey meaning and its distinct vowel harmony.

A.12 Twi

Twi is a Kwa language spoken in Ghana by over 9 million people. It is tonal and has a rich vocabulary with loanwords from various African and European languages. Twi uses the Latin script and has a long history of oral tradition, including proverbs and folktales.

A.13 Mozambican Portuguese

Mozambican Portuguese is a Portuguese dialect spoken in Mozambique. It is characterized by African influences and has evolved differently from European Portuguese. It uses the Latin alphabet and has unique vocabulary and pronunciation.

Language	Code	Classification	Script
Algerian Arabic	dz	afro-asiatic, semitic, west semitic, central semitic, arabian, Arabic, north African Arabic, Algerian Arabic	Latin, Arabic
Amharic	am	Afro-asiatic, Semitic, South, Ethiopian, South, Transversal, Amharic-argobba	Ethiopic
Hausa	ha	Afro-asiatic, Chadic, west, A, A.1	Latin
Igbo	ig	Niger-congo, Atlantic congo, volta-congo, benue-congo, igboid, igbo	Latin
Kinyarwanda	kr	Niger-congo, Atlantic congo, volta-congo, benue-congo, bantoid, southern, narrow bantu, central, J, Ruanda-rundi	Latin
Moroccan Arabic	ma	afro-asiatic, semitic, west semitic, central semitic, arabian, Arabic, north African Arabic, Moroccan-Andalusian Arabic, Moroccan Arabic	Arabic
Mozambican Portuguese	pt	Indo-European, classical Indo-European, Italic, Latino-Faliscan, Latinic, Imperial Latin, Romance, Italo-Western Romance, Western Romance, Shifted Western Romance, Southwestern Shifted Romance, West Ibero-Romance, Galician Romance, Macro-Portuguese, Brazil-Portugal Portuguese, Portuguese, Nigerian Pidgin	Latin
Nigerian Pidgin	pcm	Creole-English, English based, Atlantic, Krio	Latin
Oromo	or	Afro-asiatic, Cushitic, East, Oromo	Latin
Swahili	sw	Niger-congo, Atlantic congo, volta-congo, benue-congo, bantoid, southern, narrow bantu, central, G, swahili	Latin
Tigrinya	tg	Afro-asiatic, Semitic, South, Ethiopian, North	Ethiopic
Twi	twi	Niger-congo, Atlantic congo, Volta-congo, Kwa, Nyo, Potou-tano, Tano, Central, Akan	Latin
Yoruba	yo	Niger-congo, Atlantic congo, volta-congo, benue-congo, defoid, yoruboid, edekiri	Latin

Table 3: Details about each language in Afri-Senti Data

Lang.	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
yo	70.74	25.14	71.49	25.14	76.43	78.07	76.74	73.34	78.15	**78.72**
twi	47.95	30.23	58.55	30.23	63.30	66.80	62.96	44.90	67.72	**67.88**
ts	34.80	30.37	50.85	30.37	48.02	57.76	57.25	34.54	51.09	59.33
sw	44.95	43.98	51.70	43.98	61.33	**61.62**	60.61	58.27	59.73	60.44
pt	68.72	35.75	63.71	35.75	67.37	59.04	58.91	59.73	**70.40**	66.03
pcm	74.15	49.28	49.28	49.28	75.90	72.55	73.36	70.79	**76.27**	75.19
ma	**82.39**	15.25	85.29	15.25	74.72	74.13	64.76	71.45	75.92	75.22
kr	56.05	21.01	56.15	21.01	68.92	64.60	**81.41**	67.98	68.86	67.01
ig	78.48	26.94	78.18	26.94	78.92	80.58	79.93	73.97	80.80	**80.82**
ha	76.58	16.79	74.84	16.79	79.69	79.59	43.91	78.09	**81.56**	79.10
dz	54.97	37.71	64.41	37.71	65.41	48.08	43.91	45.06	**70.75**	65.94
am	59.81	35.39	38.69	35.39	62.53	60.98	61.44	59.64	**63.05**	59.38
Average	62.46	30.65	61.93	30.65	68.55	66.98	66.71	61.48	**70.36**	69.59
multilingual	68.28	68.28	68.28	68.28	73.89	71.67	71.40	72.88	**75.57**	73.40
or	36.00	15.15	15.15	15.15	43.97	**50.72**	49.78	38.20	44.98	45.27
tg	38.91	14.38	14.38	14.38	54.38	40.70	45.24	**57.72**	56.64	45.73

Table 4: Results of Model Performance on Dev Set. **M1:** xlmr-base, **M2:** xlmr-large, **M3:** mbert-base-cased, **M4:** afro-xlmr-large, **M5:** afro-xlmr-base, **M6:** afriberta_large, **M7:** afriberta_base, **M8:** serengeti, **M9:** afro-xlmr-base$_{\text{ft}}$, **M10:** serengeti$_{\text{ft}}$.