DuluthNLP at SemEval-2023 Task 12: AfriSenti-SemEval: Sentiment Analysis for Low-resource African Languages using Twitter Dataset

Samuel Akrah & Ted Pedersen

Department of Computer Science University of Minnesota Duluth, MN 55812 USA {akrah001, tpederse}@d.umn.edu

Abstract

This paper describes the DuluthNLP system that participated in Task 12 of SemEval-2023 on AfriSenti-SemEval: Sentiment Analysis for Low-resource African Languages using Twitter Dataset. Given a set of tweets, the task requires participating systems to classify each tweet as negative, positive or neutral. We evaluate a range of monolingual and multilingual pre-trained models on the Twi language dataset, one among the 14 African languages included in the SemEval task. We introduce TwiBERT, a new pretrained model trained from scratch. We show that TwiBERT, along with mBERT, generally perform best when trained on the Twi dataset, achieving an F1 score of 64.29% on the official evaluation test data, which ranks 14 out of 30 of the total submissions for Track 10. The TwiBERT model is released at https://huggingface.co/sakrah/TwiBERT

1 Introduction

Sentiment analysis is a well-studied problem in Natural Language Processing (NLP), at least with respect to high-resourced languages such as English. However, despite the popularity of sentiment analysis it is not well studied in languages where there are limited resources (Joshi et al., 2020). This is particularly true when discussing the languages of Africa.

Africa is the most linguistically diverse continent, and is home to more than 2,000 languages. Yet the continent is chronically underserved in NLP resources and technology, in part because there is little to no training data available for African languages in NLP research.

However, more recently there have been concerted efforts to create datasets and build models for African languages (Adelani et al., 2021; Muhammad et al., 2022). One such effort is SemEval-2023 Task 12 (Muhammad et al., 2023b), a sentiment analysis task for low-resource African Languages. This paper describes the DuluthNLP team's participation in Track 10 of this task which focused on the Twi language.

2 Task Data

The training, development and test data for the models we created were supplied by the organizers of SemEval-2023 Task 12 (Muhammad et al., 2023a). The dataset is a collection of tweets labelled with three sentiment classes (negative, positive, and neutral) from 14 African languages including Amharic, Algerian Arabic, Hausa, Igbo, Kinyarwanda, Moroccan Arabic, Mozambican Portuguese, Nigerian Pidgin, Oromo, Swahili, Tigrinya, Twi, Xitsonga, and Yorùbá. The Twi language is the focus of this paper. The sentiment of each tweet was determined by a majority vote among three annotators.

The tweets are from the the AfriSenti Corpus (Muhammad et al., 2023a) and were curated using the Twitter Academic API, a publicly available Twitter API supporting the retrieval of tweets in 70 languages. However, Amharic is the only African language included in that group of 70. The lack of support for any of the remaining 2,000 African languages by Twitter's API prompted the authors of the AfriSenti dataset to rely on novel heuristics to retrieve tweets for its 14 African languages. This resulted in a collection of more than 110,000 annotated tweets from four language families of Africa, making it the largest sentiment analysis corpus for African languages. A few examples from this data are shown in Table 1.

3 Methodology

Current state-of-the-art approaches to sentiment classification rely on using pre-trained large language models (LLMs) such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), or their variants. When the task to be carried out is in a language other than English, multilingual versions of the above have been shown to provide superior per-

Tweet	Translation into English	Sentiment
s3 gyimie paa nie nyansa baako p3 mpo sei nni wo tri mu	You are so stupid there is no wisdom in your head	negative
1 minute biaaa na broken heart nam mu	Someone's heart gets broken every minute	negative
baa bie oyerepa afutuosem εnfa bi da	Never take for a wife any woman who listens to marital advice	neutral
enfa bi da bra	Never take any, brother	negative
enti holiday yi yenfa nye deen	What should we do on this holiday?	positive
masa woy3 bi na gyae me	You also do it. Please let me be.	positive

Table 1: Tweets in Twi, their English translation, and the annotated sentiment.

formance even if it involves a language not seen during pre-training (Pfeiffer et al., 2020).

We experiment with four existing multilingual LLMs and apply those to the Twi language. These include mBERT, RoBERTa, XLM-R, and AfriXLMR-large. We also created a new pretrained Twi language model from scratch which we call TwiBERT. We fine-tuned each model on the Twi sentiment data from the task using the HuggingFace Transformer (Wolf et al., 2019)

mBERT (Devlin et al., 2018) was the first of the multilingual LLMs and is a BERT variant pre-trained on 104 languages, one of which is a Nigerian language - Yorùbá. Like BERT, mBERT was pre-trained using masked language modeling (MLM) and next sentence prediction (NSP). In MLM, a given percentage of the tokens of an input sequence is masked, and BERT is tasked to predict the original tokens. With the MLM approach, BERT was able to produce good results when transferred to downstream NLP tasks, becoming the new benchmark for other pre–trained models. We use the bert-base-multilingual-cased model from Hugging Face for the Twi dataset.

RoBERTa (Liu et al., 2019) is essentially a replication of BERT. We used it because of its overall level of success in a range of NLP problems. RoBERTa improves upon BERT since it is trained for a longer period of time using more data and larger mini-batches.

XLM-R (Conneau et al., 2019) is a multilingual variant of RoBERTa. Like RoBERTa XLM-R only employs MLM for pre-training and does not include NSP. This distinguishes it from both BERT and mBERT, which use both MLM with NSP dur-

ing pre-training. We fine-tune XLM-RoBERTa-Base on the Twi data by adding a classifier on top of the pre-trained model.

AfroXLMR-large (Alabi et al., 2022) is another multilingual variant that we used. This is an MLM adaptation of of XLM-R for 17 African languages, including languages that are linguistic cousins to Twi, such as Yorùbá. However, the Twi language is not included in the pre-trained data.

TwiBERT is the last model we used, which we pre-trained from scratch from the Asanti Twi Bible and a dataset sourced through crowdsourcing efforts. Like XLM-R, TwiBERT is a monolingual variant of RoBERTa. The model has 61 million parameters, 6 layers, 6 attention heads, 768 hidden units, and a feed-forward size of 3072. We then fine-tune TwiBERT on Twi dataset by building a classifier on top of it.

4 Development Results

To establish a traditional baseline, we trained and evaluated a Naive Bayesian classifier. It achieved an F1 score of 61%. We also implemented a random guessing model which obtained the expected F1 score of 33%.

Then to establish an LLM baseline, we finetuned a RoBERTa model by building a classifier using the Twi dataset. The model was trained using the Adam Optimizer with a linear scheduler with warmup; and a learning rate of 4e-5 for 16 epochs (see Table 2). It resulted in an F1 score of 63%.

Next, we finetune mBERT model on the task using the same hyperparameters and obtained an improved F1 score of 65%. Using these same hyperparameters we experimented with XLM-R,

Hyperparameter	Value
Learning Rate	4e-5
Adam ϵ	1e-8
Optimizer	AdamW
Learning rate decay	Linear
Weight Decay	0
Batch Size	16
Train Epochs	16

Table 2: Hyperparameter Values for Fine-Tuning.

Model	F1
TwiBERT	65.70%
mBERT	65.3%
AfriBERTa-large	63%
Afro-XLMR-base	63
RoBERTa-base	63%
XLMR-base	62.7%
Naive Bayes	61%
Random Guessing	33.3%

Table 3: Experimental Results on Development Data.

AfriBERTa, and Afro-XLMR, all of which are multilingual pre-trained LLMs that include African languages.

As our final step, we built a TwiBERT, a pretrained model learnt on the Asanti Twi Bible. When trained on the Twi sentiment data, the model gave an F1 score of 65%.

All of the results reported above were based on the development data, and are summarized in Table 3. All of these models were built using HuggingFace PyTorch implementations of the mBERT, RoBERTa, and AfriBERTa, XLM-R, and AfroX-LMR (Wolf et al., 2019) We fine-tuned our models using 2 Nvidia Quadro RTX 8000 GPUs.

5 Official Task Evaluation Results

Our official score on SemEval-2023 Task 12 Track 10 (Twi) was 64.29% with the mBERT model. The top ranked system for Track 10 obtained an F1 score of 68.28%. DuluthNLP ranked 14 among 30 systems.

6 Error Analysis

F1 scores for all the models we experimented with averaged between 61% - 65%, which suggests, in part, that our system is moderately capable of class-

sifying sentiment analysis for Twi data. What is curious, though, is that the LLMs performance are not entirely different from the baseline scores from Naive Bayes.

Part of the challenge, in part, relates to the nature and annotation of the labelled dataset for Twi.

For instance, several of the tweets omit or replace the proper Twi characters or scripts, in part because of the lack of the Twi keyboard, or, in the few instances where one exists, people will tweet without it. As an example, the first and the seventh words of the first tweet in table 1 (s3 gyimie paa nie nyansa baako p3 mpo sei nni wo tri mu) are not Twi words. Only instead of 'sɛ', the author wrote 's3', thus effectively replacing 'ɛ' with the number '3'. Other authors will replace 'ɔ' (pronounced 'or' in Twi) with the either 'c' or ')'

Another obvious challenge is the lack of a standardized way of writing Twi, at least among tweet authors. This gives people considerable latitude in writing any way they see fit. 'Gyimii', an insulting slur that has been used more than 200 times in the training data, has as many varied spellings as there are tweets ('wiegyimi', 'wegyimi', 'wagyimi' are a few of them). This presents a challenge for the model when assigning probabilities during training. Thus, without a very large Twi corpus, it will be almost impossible for any NLP model to generalize.

A further challenge is the issue of mislabeled tweets, which make up the bulk of the train set. Take, again, the word 'gyimii'. Alone, the word means just that: "foolish". But when used together with other words, its meaning may range from the hilariously funny to crass stupid. Almost everywhere the word appears, though, it is labeled as a negative tweet, which makes one wonder if the annotators assigned labels based on some key words or the immediate context of each tweet informed the labeling.

Again, as illustrated in Table 1, 'enfa bi da bra', a tweet that counsels a brother against taking or possessing unspecified object of interest, has been given a negative label. A similar tweet containing the same phrasing (obaa otie oyerepa afutuosem ɛnfa bi da) got a neutral label. The lack of consistency in labeling can impact the learning of models.

The Twi langague is among the few African languages in this SemEval Task whose data was not seen during pre-training of the LLMs. And whilst latent knowledge about multiple languages in the embedding matrix of pre-trained models is carried over when fine-tuned on downstream tasks, much of everything else is lost. To solve this, we developed TwiBERT, a pre-trained model trained from scratch on a Twi corpus. And even with that, performance is constrained by the small size and scope of the pre-trained dataset. The model was trained on a relatively limited dataset (approximately 5MB), which may hinder its ability to learn intricate contextual embeddings and effectively generalize. Additionally, the dataset's focus on the Bible could potentially introduce a strong religious bias in the model's output

7 Ethical Considerations

For the most part we relied on existing pre-trained Large Language Models, with the exception of TwiBERT which we built from scratch. While a model like RoBERTa includes multi-lingual text, whether this works or is relevant to a low-resource language like Twi is another question. In general the most popular and widely used LLMs have a very strong bias towards English and may or may not prove useful with other languages.

The fact that low-resource languages are clearly not a priority of the organizations carrying out computationally expensive pre-training of LLMs skews our field away from the languages of Africa and towards English, which is already very well represented.

In addition any biases or stereotypes that exist in these models get propagated to these new tasks and applications of these models, and we are not even fully cognizant of the harms that are contained therein.

In general the blind use of LLMs to somehow hopefully get a reasonable result with a lowresource language simply reinforces the existing hierarchy of languages in NLP, where English is at the top and low-resource languages are included in training data by accident or as a second thought. This style of approach also serves to invalidate the expertise of those who actually know and use the low-resource language.

There is a related danger with the leaderboard focused evaluation as is taken by SemEval that teams with large compute resources will be able to do well in a low-resource task like this simply by throwing hyperparameter tuning methods at the problem. Indeed it might be possible that a team with no expertise in African languages could do quite well on the leaderboard and even be declared the "winner". This would send a chilling and unintended message from SemEval that expertise in low-resource languages doesn't really matter and that compute is all you need even if you can't read nor understand the language under study. To act as a brake on this unfortunate direction, we would strongly recommend that SemEval expect and require meaningful evaluation of the results of low-resource tasks that would require expertise in the language. We would also discourage the continued use of terminology that associates the highest score with "winning" a task, particularly if said winners are unable to understand the output of their system.

Acknowledgements

The authors would like to thank the organizers for the opportunity to participate in SemEval-2023 Task 12. We also appreciate the valuable input and guidance provided by the reviewers, whose suggestions have significantly improved our work.

References

- David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named entity recognition for African languages. Transactions of the Association for Computational Linguistics, 9:1116-1131.
- Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pretrained 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.

- 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.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.
- 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. 2023a. 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. 2023b. 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.

- Jonas Pfeiffer, Ivan Vulic, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: an adapter-based framework for multi-task cross-lingual transfer. *CoRR*, abs/2005.00052.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.