ABCD Team at SemEval-2023 Task 12: An Ensemble Transformer-based System for African Sentiment Analysis

Dang Van Thin[1,2] and **Dai Ba Nguyen**[1,2]
Dang Ba Qui[3] and **Duong Ngoc Hao**[1,2] and **Ngan Luu-Thuy Nguyen**[1,2]
[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]National University, Ho Chi Minh City, Vietnam
[3]Nong Lam University, Ho Chi Minh city, Vietnam
{thindv,ngannlt,haodn}@uit.edu.vn, 2152914@gm.uit.edu.vn, 21130500@st.hcmuaf.edu.vn

Abstract

This paper describes the system of the ABCD team for three main tasks in the SemEval-2023 Task 12: AfriSenti-SemEval for Low-resource African Languages using Twitter Dataset. We focus on exploring the performance of ensemble architectures based on the soft voting technique and different pre-trained transformer-based language models. The experimental results show that our system has achieved competitive performance in some Tracks in Task A: Monolingual Sentiment Analysis, where we rank the Top 3, Top 2, and Top 4 for the Hause, Igbo and Moroccan languages. Besides, our model achieved competitive results and ranked 14^{th} place in Task B (multilingual) setting and 14^{th} and 8^{th} place in Track 17 and Track 18 of Task C (zero-shot) setting.

1 Introduction

The AfriSenti-SemEval Shared Task 12 (Muhammad et al., 2023b) aims at building Sentiment Analysis (SA) systems for 17 African languages, including Hausa, Yoruba, Igbo, Nigerian Pidgin from Nigeria, Amharic, Tigrinya, and Oromo from Ethiopia, Swahili from Kenya and Tanzania, Algerian Arabic dialect from Algeria, Kinyarwanda from Rwanda, Twi from Ghana, Mozambique Portuguese from Mozambique and Moroccan Arabic/Darija from Morocco. This shared task has three main tasks, including two zero-shot tracks, one multilingual track and 12 monolingual tracks. The zero-shot tracks require training two zero-shot models where each model works for only one language, and each model has training from the 12 languages in the monolingual tracks. The multilingual track requires training multilingual SA models that are able to handle all languages. Twelve monolingual tracks need training individual monolingual models where each model works for only one language. The dataset involves tweets labelled with three sentiment classes (positive, negative, neutral) in 14 African languages. Each tweet is annotated by three annotators following the annotation guidelines in (Mohammad, 2016). They used a form of a majority vote to determine the sentiment of the tweet (Mohammad, 2022; Yimam et al., 2020). All the tweet of the dataset is code-mixed, which can increase the performance of the model.

In this paper, we propose an ensemble architecture for the AfriSenti-Semeval shared task. Our ensemble architecture is based on the pre-trained transformer-based language models and soft voting technique. In our case, the AfroXLMR, AfriBERTa, and LaBSE are employed as the base classifier because these models support African languages in the shared task. The final prediction is combined from the base classifier using the soft voting technique.

The rest of the paper is organized as follows. Section 2 provides the related work. The system description is presented in Section 3, followed by evaluation results in Section 5. The experimental setup and conclusion is discussed in Section 4 and Section 6, respectively.

2 Related Work

Sentiment Analysis has been studied in the NLP field for two past decades for low-resource languages. However, in the case of low-resourced languages such as African languages, the studies have yet to progress well compared to those of high-resourced languages due to the reasons such as the unavailability of annotated corpora.

In the Arabic language, a set of different machine learning and deep learning approaches has been applied to the sentiment analysis task of the language (Abdulla et al., 2013; Duwairi et al., 2014; Heikal et al., 2018; Shoukry and Rafea, 2012). In (Abdulla et al., 2013), the authors have constructed an Arabic sentiment analysis dataset extracted from Twitter and classified it as a negative and positive sentiment. The methods they used were both lexicon-based

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 324–330
July 13-14, 2023

in which they constructed a set of words labelled with some polarity indicators, and using those constructed lexicons; they implemented an algorithm that tells if the whole tweet is positive or negative depending on the number of positive or negative lexicon it contains. At the same time, they have also used supervised methods of machine learning models such as support vector machine (SVM) (Zhang, 2001), and K-nearest neighbours (KNN) (Bishop and Nasrabadi, 2006) trained on the constructed dataset. They have achieved the best result by using SVM classifiers when compared to other methods they have used.

In other Arabic sentiment analysis studies by (Heikal et al., 2018), the authors have used deep learning models to outperform the state-of-the-art result. Their approach was an ensemble of convolutional neural network (CNN) (Heaton, 2018) and long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) models through which they have achieved an F1-Score of 64.46% beating the previous state-of-the-art by more than 10%. However, Arabic has also benefited from the recent breakthrough in NLP studies related to transformer models and transfer learning techniques. A pre-trained Arabic version of BERT has achieved state-of-the-art results in many NLP downstream tasks of Arabic, including sentiment analysis, named entity recognition and question answering (Antoun et al., 2020).

In the case of Amram et al. (2018), the authors studied the effect of the input as tokens or morphemes. They concluded that for linear neural network architectures such as multi-layer perceptron (MLP) (Heaton, 2018) and (CNN), the token level granularity had achieved higher accuracy, while for LSTM architecture, the level of the morpheme achieved better results. Furthermore, they achieved state-of-the-art results for the Hebrew sentiment analysis task, using the CNN network of token level input with more than the accuracy of 89%.

With the emergence of deep learning models, Amharic sentiment analysis has also benefited from the breakthrough in the field of NLP using deep learning. In another study for Amharic sentiment classification, Yeshiwas Getachew and Abebe Alemu (Alemu, 2018) used a deep learning approach to model sentiment analysis tasks of the language. They have constructed a new dataset collected from social media platforms, Facebook, and labelled by Amrahic linguist experts. Using a fully connected neural network with different parameter tuning, including the number of neurons, they achieved an accuracy of 96.61% in the validation set.

3 System Description

3.1 Approach

The diagram in Figure 1 illustrates our ensemble approach for Task A. The framework consists of three main layers: a pre-processing layer, a layer of contextual language-based models, and a voting ensemble layer. Firstly, the input text is subject to several processing steps in the pre-processing layer. Following this, we fine-tune different pre-trained contextual language models in order to obtain probability outputs of labels. Finally, the probability outputs from individual models are combined using a soft voting technique to get the final prediction. The detailed structure of the framework is described in the following.

Pre-processing Layer: Pre-processing is one of the essential components in the classification system in the NLP field. However, the languages in the shared task are unfamiliar; therefore, we design a standard list of pre-processing steps for all languages, including:

- **Word Normalization**: We used the regular expression technique to normalize some words or phrases which are the same meanings in the sentence. For example, we replace the "URL" with the "website" word.
- **Noise Removal**: We observed that there are many noises, such as punctuation and special characters, in the dataset. We found that these noises are not necessary for the sentence-level dataset. Therefore, we remove it from the samples.
- **Emoji Replacement**: Because of the large amount of emoji in the dataset, up to 28.85% (based on our statistics), the emoji is very important for the model which can increase the performance of the model. We replace the emoji with three labels such as negative, neutral and positive.

Fine-tuning Language Model: As can be seen in Figure 1, we utilize the power of three pre-trained contextual language models, including AfroXLMR (Alabi et al., 2022), AfiBERTa (Ogueji

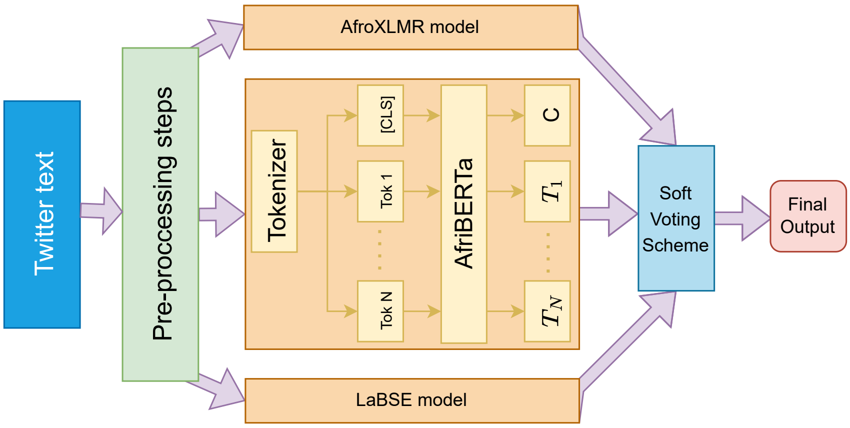

Figure 1: The soft voting ensemble architecture based on the combination of fine-tuning multilingual different contextual language models.

et al., 2021), and LaBSE (Feng et al., 2022) model as the base classifiers. To fine-tune the language models, we followed the approach of (Devlin et al., 2019a), which is presented in detail below:

Given a pre-processed Twitter with N words: $X = \{x_1, x_2, ..., x_N\}$, we first use a corresponding tokenizer to prepare the inputs. Then, we employ a pre-trained language model with L transformer layers to calculate the contextualized representations $H^L = \{h^L_{cls}, h^L_1, ..., h^L_N\} \in R^{N \times dim_h}$, where dim_h is the dimension of the representation vector. Finally, we extract the contextualized representation h^L_{cls} of [CLS] token in the last L transformer layer as the feature of the input. The obtained representations is directly fed to the linear layer with the Softmax activation function to calculate the score vector $\hat{y}^{(a)}$ for each class.

$$\hat{y} = softmax(W \cdot h^L_{cls} + b) \tag{1}$$

where W and b are the learnable parameters of the output layer. We use the Category Cross-Entropy loss to optimize the model.

Soft voting Scheme: Our motivation for applying an ensemble approach is to take advantage of the performances of various models. Given predictions $\{\hat{y}_{\theta_1}, \hat{y}_{\theta_2}, .., \hat{y}_{\theta_n}\}$ the n base classifiers. We applied the simple soft voting technique to merge the predictions of the base models. In our case, the individual classifiers are treated equally. We sum up the probability output of n classifiers and choose the sentiment class with the highest probability as the final prediction.

3.2 Pre-trained Contextual Language Models

We briefly explain the three pre-trained language models used in this paper.

- **AfroXLMR**: AfroXLMR is a large language model developed by (Alabi et al., 2022) and is released for the community of researchers in African languages. It is based on the XLM-RoBERTa architecture and applies the multilingual adaptive fine-tuning technique to 17 African languages and three high-resource languages (English, French, and Arabic) simultaneously.

- **AfriBERTa**: AfriBERTa is a transformer-based multilingual language model trained on 11 African languages, all of which are low-resource. It had developed by (Ogueji et al., 2021). The author had trained a transformer (Vaswani et al., 2017) with the standard masked language modelling objective of (Devlin et al., 2019b) without next sentence prediction. This is also the same approach used in XLM-R (Conneau et al., 2020).

- **LaBSE**: LaBSE is a language-agnostic BERT sentence embedding model supporting up to 109 languages. This language model was developed by (Feng et al., 2022). To have the best method for learning multilingual sentence embeddings, they combined the best methods for learning monolingual and cross-lingual representations, including: masked language modelling (MLM), translation language modelling (TLM) (Lample and Conneau, 2019), dual encoder translation ranking (Guo et al.,

Table 1: Results of our best system compared with two top systems on 12 tracks for Task A: Monolingual Sentiment Classification.

Track 1: Hausa		Track 2: Yoruba		Track 3:Igbo		Track 4:Nigerian Pidgin	
Team	**F1-score**	**Team**	**F1-score**	**Team**	**F1-score**	**Team**	**F1-score**
Top 1	82.62	Top 1	80.16	Top 1	82.96	Top 1	75.96
Top 2	82.04	Top 2	80.08	Top 3	81.51	Top 2	75.75
Ours (Top 3)	**81.50**	**Ours (Top 6)**	**79.73**	**Ours (Top 2)**	**82.28**	**Ours (Top 18)**	**66.30**

Track 5:Amharic		Track 6:Algerian Arabic		Track 7:Moroccan		Track 8: Swahili	
Team	**F1-score**	**Team**	**F1-score**	**Team**	**F1-score**	**Team**	**F1-score**
Top 1	78.42	Top 1	74.20	Top 1	64.83	Top 1	65.68
Top 2	72.18	Top 2	73.00	Top 2	63.54	Top 2	64.89
Ours (Top 18)	**58.05**	**Ours (Top 21)**	**63.50**	**Ours (Top 4)**	**61.54**	**Ours (Top 8)**	**63.10**

Track 9:Kinyarwanda		Track 10:Twi		Track 11:Mozambican		Track 12:Xitsonga	
Team	**F1-score**	**Team**	**F1-score**	**Team**	**F1-score**	**Team**	**F1-score**
Top 1	72.63	Top 1	68.28	Top 1	74.98	Top 1	60.67
Top 2	72.50	Top 2	67.58	Top 2	73.83	Top 2	60.32
Ours (Top 15)	**67.36**	**Ours (Top 11)**	**65.61**	**Ours (Top 19)**	**67.21**	**Ours (Top 10)**	**53.92**

2018), and additive margin softmax (Yang et al., 2019).

4 Experimental Setup

Data and Preprocessing: We utilized the official training set for training models. As the competition rules stipulated, no additional data was used during the training process (Muhammad et al., 2023a). The development set was used to optimize the hyper-parameters for each track and task.

Evaluation Metrics: The evaluation metric for three Tasks (A, B, C) is a Weighted F1-score between submission and test gold set.

Configuration Settings: We implemented our models using Trainer API from Hugging Face library (Wolf et al., 2020). The maximum input length is set as 128 tokens, and the number of epochs is set as 10 with a batch size of 32 for all languages. We used an AdamW optimizer with a linear schedule warmup technique.

Submitted Systems: We submitted the different models based on the task to the evaluation phase. For task A - Monolingual SA, we submit the performance of the ensemble soft voting model for all languages. For task B, because of the limitation of computation resource languages, we divide the training into 5 folds and train the LaBSE model on each fold to predict the test set. Then, we combine the results of 5 folds using the soft voting technique. For Task C, our strategy is to utilize the Google Translation[1] to translate the source language to a target language. In our case, we translate the test set on Track 17 and Track 18 into the Hausa language. We choose the Hause language as the source language for the following reasons: (1) the number of samples in the training set and distribution between labels.

5 Results and Discussion

In this section, we present the official results of our final submission model for three main Tasks in the AfriSenti-SemEval Shared Task competition. For task A, we only compare results with the results from the two top teams for each track. In comparison, we report the top 5 systems for task B and task C, respectively.

Task A: Monolingual Sentiment Classification Table 1 presents the performances of our ensemble model compared with two top teams for 12 tracks. Our system gives competitive results on the three tracks such as Track 1 (Hausa), Track 2 (Yoruba), Track 3 (Igbo) , Track 7 (Moroccan), and Track 8 (Swahili). Unfortunately, our submission system is not effective for the remainder Tracks. To explore the reason, we report the results of base models and ensemble systems on the development set. As seen in Table 2, we observe that the performance of the ensemble model decreases significantly on some Tracks (e.g. Track 4, Track

[1]https://translate.google.com/

Table 2: Results of the base models and the ensemble transformer-based architecture on the development set for Task A: Monolingual Sentiment Analysis.

Track 1: Hausa		Track 2: Yoruba		Track 3:Igbo		Track 4:Nigerian Pidgin	
Model	**F1-score**	**Model**	**F1-score**	**Model**	**F1-score**	**Model**	**F1-score**
AfroXLMR	78.91	AfroXLMR	74.71	AfroXLMR	80.18	AfroXLMR	75.02
AfriBERTa	80.28	AfriBERTa	76.95	AfriBERTa	81.28	AfriBERTa	73.73
LaBSE	80.61	LaBSE	77.31	LaBSE	81.86	LaBSE	74.15
Ensemble	81.30	Ensemble	78.49	Ensemble	82.64	Ensemble	73.80

Track 5:Amharic		Track 6:Algerian Arabic		Track 7:Moroccan		Track 8: Swahili	
Model	**F1-score**	**Model**	**F1-score**	**Model**	**F1-score**	**Model**	**F1-score**
AfroXLMR	57.31	AfroXLMR	65.33	AfroXLMR	74.28	AfroXLMR	60.94
AfriBERTa	58.04	AfriBERTa	45.99	AfriBERTa	58.61	AfriBERTa	59.88
LaBSE	58.73	LaBSE	64.79	LaBSE	74.49	LaBSE	57.22
Ensemble	59.45	Ensemble	64.04	Ensemble	73.89	Ensemble	61.84

Track 9:Kinyarwanda		Track 10:Twi		Track 11:Mozambican		Track 12:Xitsonga	
Model	**F1-score**	**Model**	**F1-score**	**Model**	**F1-score**	**Model**	**F1-score**
AfroXLMR	64.86	AfroXLMR	62.34	AfroXLMR	64.13	AfroXLMR	51.93
AfriBERTa	61.92	AfriBERTa	63.25	AfriBERTa	58.67	AfriBERTa	55.83
LaBSE	65.81	LaBSE	64.93	LaBSE	66.11	LaBSE	53.78
Ensemble	64.95	Ensemble	64.23	Ensemble	62.59	Ensemble	58.01

11) due to the poor performance of base models. Moreover, we noticed that our approach is effective for Tracks with a lot of samples and balance in the training data. In this study, we intend to investigate the effectiveness of soft voting techniques based on different transformer-based models for African languages. Therefore, we used the ensemble model as the final submission system instead of the best model on the development set.

Task A: Multilingual Sentiment Classification Table 3 present the results of our submission on Task B. Officially, we achieved the F1-score of 69.22% on the test set (Top 14). As mentioned in Section 4, due to the limitation of computational resources, we must split the training set as the 5 folds and use the LaBSE model as the main classifiers. Then we use the soft-voting technique to merge the prediction of 5 folds on the test set.

Task C: Zero-Shot Sentiment Classification Table 4 shows our submission results compared with the top 5 systems in Task C - Zero-shot sentiment analysis. We ranked 14^{th} and 8^{th} among all participating systems for Track 17 and Track 18, respectively. One of the reasons for the poor performance of our submission system is the error in the translation process to translate the source language to the target language.

Table 3: Results of our best system compared with five top systems on Track 16 for Task B: Multilingual Sentiment Classification.

Rank	Team	F1-score
Top 1	BCAI-AIR3	75.06
Top 2	king001	74.96
Top 3	DN	72.55
Top 4	ymf924	72.34
Top 5	mitchelldehaven	72.33
Ours (Top 14)	ABCD Team	69.22

6 Conclusion

In this paper, we presented a simple and efficient ensemble architecture for sentiment analysis tasks in the SemEval-2023 Task 12: AfriSenti-SemEval. Our system is based on fine-tuning the pre-trained transformer-based language model as the base classifiers and the soft voting technique to combine the prediction of different base classifiers. Our experiments demonstrated that it achieves competitive results on some languages in Task A: Monolingual Sentiment Analysis without relying on any additional resources. For future work, we plan to improve our system to handle the imbalanced problem in some languages. Besides, data augmentation is

Table 4: Results of our best system compared with five top systems on Track 17 and Track 18 for the Task C: Zero-Shot Sentiment Classification.

Track 17: Zero-shot Tigrinya		
Rank	**Team**	**F1-score**
Top 1	BCAI-AIR3	70.86
Top 2	king001	70.47
Top 3	ymf924	70.39
Top 4	uid	69.90
Top 5	TBS	69.61
Ours (Top 14)	ABCD Team	60.53

Track 18: Zero-shot Oromo		
Rank	**Team**	**F1-score**
Top 1	mitchelldehaven	46.23
Top 2	UCAS	45.82
Top 3	ymf924	45.34
Top 4	UM6P	45.27
Top 5	TBS	45.12
Ours (Top 8)	ABCD Team	42.64

also a promising direction to enhance the overall system's performance.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments and the AfriSenti-SemEval organization for organizing the shared task competition. This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

References

Nawaf A Abdulla, Nizar A Ahmed, Mohammed A Shehab, and Mahmoud Al-Ayyoub. 2013. Arabic sentiment analysis: Lexicon-based and corpus-based. In *2013 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT)*, pages 1–6. IEEE.

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.

Y Alemu. 2018. Deep learning approach for amharic sentiment analysis university of gondar.

Adam Amram, Anat Ben David, and Reut Tsarfaty. 2018. Representations and architectures in neural sentiment analysis for morphologically rich languages: A case study from modern hebrew. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2242–2252.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. 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.

Rehab M Duwairi, Raed Marji, Narmeen Sha'ban, and Sally Rushaidat. 2014. Sentiment analysis in arabic tweets. In *2014 5th international conference on information and communication systems (ICICS)*, pages 1–6. IEEE.

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.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith

Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 165–176, Brussels, Belgium. Association for Computational Linguistics.

Jeff Heaton. 2018. Ian goodfellow, yoshua bengio, and aaron courville: Deep learning: The mit press, 2016, 800 pp, isbn: 0262035618. *Genetic Programming and Evolvable Machines*, 19(1-2):305–307.

Maha Heikal, Marwan Torki, and Nagwa El-Makky. 2018. Sentiment analysis of arabic tweets using deep learning. *Procedia Computer Science*, 142:114–122.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Saif Mohammad. 2016. A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 174–179.

Saif M Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

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.

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.

Amira Shoukry and Ahmed Rafea. 2012. Sentence-level arabic sentiment analysis. In *2012 international conference on collaboration technologies and systems (CTS)*, pages 546–550. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. *arXiv preprint arXiv:1902.08564*.

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.

Tong Zhang. 2001. An introduction to support vector machines and other kernel-based learning methods. *Ai Magazine*, 22(2):103–103.