SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval)

Shamsuddeen Hassan Muhammad[1,2+*]**, Idris Abdulmumin**[3+*]**, Seid Muhie Yimam**[4*]**,
David Ifeoluwa Adelani**[5*]**, Ibrahim Sa'id Ahmad**[2+*]**, Nedjma Ousidhoum**[6]**,
Abinew Ayele**[4,7]**, Saif M. Mohammad**[8]**, Meriem Beloucif**[9]** , Sebastian Ruder**[10]

[1]University of Porto, [2]Bayero University Kano, [3]Ahmadu Bello University, Zaria, [4]Universität Hamburg,
[5]University College London, [6]University of Cambridge, [7]Bahir Dar University, [8]National Research Council Canada,
[9]Uppsala University, [10]Google Research, *Masakhane NLP, +HausaNLP

shmuhammad.csc@buk.edu.ng

Abstract

We present the first Africentric SemEval Shared task, Sentiment Analysis for African Languages (AfriSenti-SemEval)[1]. AfriSenti-SemEval is a sentiment classification challenge in 14 African languages (Amharic, Algerian Arabic, Hausa, Igbo, Kinyarwanda, Moroccan Arabic, Mozambican Portuguese, Nigerian Pidgin, Oromo, Swahili, Tigrinya, Twi, Xitsonga, and Yorùbá) (Muhammad et al., 2023), using data labeled with 3 sentiment classes. We present three subtasks: (1) Task A: monolingual classification, which received 44 submissions; (2) Task B: multilingual classification, which received 32 submissions; and (3) Task C: zero-shot classification, which received 34 submissions. The best performance for tasks A and B was achieved by NLNDE team with 71.31 and 75.06 weighted F1, respectively. UCAS-IIE-NLP achieved the best average score for task C with 58.15 weighted F1. We describe the various approaches adopted by the top 10 systems and their approaches.

1 Introduction

Sentiment Analysis is a prominent sub-field of Natural Language Processing that focuses on the automatic identification of sentiments or opinions expressed through online content, such as social media posts, blogs, or reviews (Liu, 2020; Mohammad, 2021; Nakov et al., 2016). Example applications are the computational analysis of emotions in language, which has been applied to literary analysis and culturonomics (Mohammad, 2011; Reagan et al., 2016; Hamilton et al., 2016); commercial use (e.g., tracking opinions towards products); and research in psychology and social science (Dodds et al., 2015; Mohammad et al., 2016). Despite tremendous amount of work in sentiment analysis
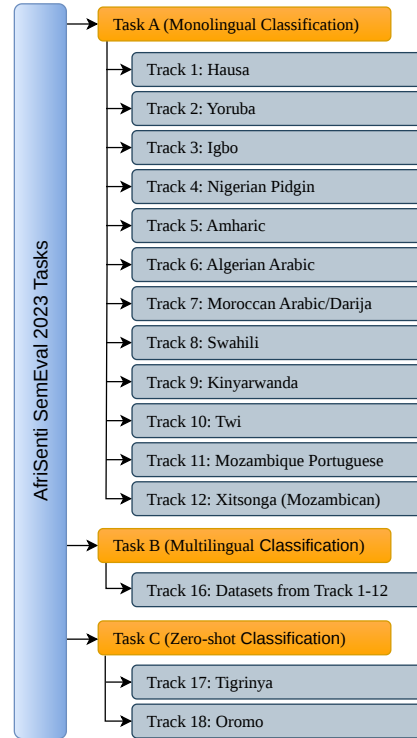
over the last two decades, little work has been conducted on under-represented languages in general and African languages in particular.

Africa has a long and rich linguistic history, experiencing language contact, language expansion, development of trade languages, language shift, and language death, on several occasions. The continent is incredibly linguistically diverse and home to over 2000 languages. This includes 75 languages with at least one million speakers each. Africa has a rich tradition of storytelling, poems, songs, and literature (Carter-Black, 2007; Banks-Wallace, 2002). Yet, it is only in recent years that there is nascent interest in NLP research for African languages, including Named Entity Recognition (NER; Adelani et al., 2021, 2022c; Jibril and Tantuğ, 2023),

Figure 1: The AfriSenti-SemEval 2023 Shared Task tracks.

---

[1]The dataset is available at https://github.com/afrisenti-semeval/afrisent-semeval-2023.
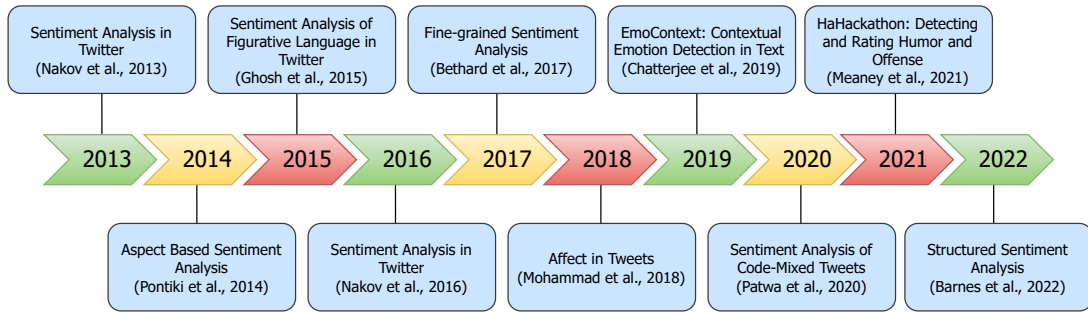
Figure 2: A timeline of SemEval Shared Tasks from 2013 to 2022 with examples of sentiment analysis tasks.

Machine Translation (MT; Nekoto et al., 2020; Abdulmumin et al., 2022; Adelani et al., 2022b; Belay et al., 2022), and Language Identification (LID) for African languages (Adebara et al., 2022a). However, African sentiment analysis has not yet received comparable attention. Similarly, although sentiment analysis is a common task in SemEval (see tasks examples in Figure 2), previous tasks have mainly focused on high-resource languages.

To this end, we present the *AfriSenti-SemEval*, a shared task in the 2023 edition of the Semantic Evaluation workshop (Ojha et al., 2023). *AfriSenti-SemEval* targets sentiment analysis in low-resource African languages. We provide researchers interested in African NLP with 110K sentiment-labeled tweets that were collected using the Twitter API. These tweets are in 14 languages (Amharic, Algerian Arabic, Hausa, Igbo, Kinyarwanda, Moroccan Arabic, Mozambican Portuguese, Nigerian Pidgin, Oromo, Swahili, Tigrinya, Twi, Xitsonga, and Yoruba) from four language families (Afro-Asiatic, English Creole, Indo-European and Niger-Congo). The annotations were conducted by native speakers of the respective languages. Besides making the annotated dataset public, we also share sentiment lexicons in most of the languages.

AfriSenti-SemEval 2023 consists of 15 tracks from three sub-tasks on the 14 collected datasets as illustrated in Figure 1. We received submissions from 44 teams, with 29 submitting a system description paper. The top-ranked teams for the different subtasks used pre-trained language models (PLMs). In particular, AfroXLMR (Alabi et al., 2022), an Africa-centric model, was the best performing model in both Tasks A (monolingual) and B (multilingual) with an average weighted F1 score of 71.30% and 75.06% respectively. For Task C, sentiment lexicons were used to build a lexicon-based multilingual BERT, which performed best in this setting with an average weighted F1 of 58.15%.

2 Background and Related Tasks

Early work in sentiment analysis relied on lexicon-based approaches (Turney, 2002; Taboada et al., 2011; Mohammad et al., 2013). Subsequent work employed more advanced machine learning (Agarwal and Mittal, 2016; Le and Nguyen, 2020), deep learning (Zhang et al., 2018; Yadav and Vishwakarma, 2020), and hybrid approaches that combine lexicon and machine learning-based approaches (Gupta and Joshi, 2020; Kaur et al., 2022). Nowadays, pre-trained language models (PLMs), such as XLM-R (Conneau et al., 2020), mDeBER-TaV3 (He et al., 2021), AfriBERTa (Ogueji et al., 2021), AfroXLMR (Alabi et al., 2022) and XLM-T (Barbieri et al., 2022) provide state-of-the-art performance for sentiment classification in different languages.

Recent work in sentiment analysis focused on sub-tasks that tackle new challenges, including aspect-based (Chen et al., 2022), multimodal (Liang et al., 2022), explainable (Cambria et al., 2022), and multilingual sentiment analysis (Muhammad et al., 2022). On the other hand, standard sentiment analysis sub-tasks such as polarity classification (positive, negative, neutral) are widely considered saturated and almost solved (Poria et al., 2020), with an accuracy of 97.5% in certain domains (Raffel et al., 2020; Jiang et al., 2020). However, while this may be true for high-resource languages in relatively clean, long-form text domains such as movie reviews, noisy user-generated data in low-resource languages still presents a challenge (Yimam et al., 2020). Additionally, African languages exhibit new challenges for sentiment analysis such as dealing with tone, code-switching, and digraphia (Adebara and Abdul-Mageed, 2022). Thus, further research is necessary to assess the efficacy of existing NLP techniques and present solutions that can solve language-specific challenges

in African contexts. SemEval, with its widespread recognition and popularity, is an ideal venue to conduct a shared task in sentiment analysis in African languages.

The SemEval competition has become the de facto venue for sentiment analysis shared tasks, featuring at least one task per year as shown in Figure 2. Some tasks focused on three-way sentiment classification—positive, negative, or neutral—while others explored more fine-grained aspect-based sentiment analysis (ABSA; Rosenthal et al., 2014, 2015; Nakov et al., 2016; Patwa et al., 2020). Additionally, there are other closely related tasks, including the *Affect in Tweets* task, which involves inferring perceived emotional states of a person from their tweet (Mohammad et al., 2018) and stance detection (Mohammad et al., 2016), which refers to the automatic identification of stance of an author towards a target from text, where the stance can be in favor, against, or neutral. Finally, structured sentiment analysis requires participants to predict the sentiment graphs present in a text. Each sentiment graph comprises a sentiment holder, a target, an expression, and a polarity (Barnes et al., 2022).

3 Task Description and Settings

The AfriSenti-SemEval shared task consists of three sub-tasks: A) monolingual sentiment classification, B) multilingual sentiment classification, and C) zero-shot sentiment classification. As shown in Figure 1, each sub-task also includes one or more tracks depending on the languages involved. Participants were free to participate in one or more sub-tasks and one or more tracks for each chosen subtask.

Task A: Monolingual Sentiment Classification

Given a training set in a language, determine the polarity (positive/negative/neutral) of tweets in the same language. If a tweet conveys both positive and negative sentiments, the strongest sentiment should be chosen. This sub-task involves 12 tracks (all languages except Oromo and Tigrinya), with one track per language.

Task B: Multilingual Sentiment Classification

Given the combined training sets from Task A, determine the polarity (positive/negative/neutral) of tweets in the test sets of the languages. This subtask has only one track with tweets from 12 languages (Hausa,

Yoruba, Igbo, Nigerian_Pidgin, Amharic, Algerian Arabic, Moroccan Arabic Darija, Swahili, Kinyarwanda, Twi, Mozambican Portuguese, and Xitsonga).

Task C: Zero-Shot Sentiment Classification

Given unlabelled tweets in two African languages (Tigrinya and Oromo), use any of the training datasets of Task A to determine the sentiment of a tweet in the two target languages. This sub-task has two tracks (Tigrinya and Oromo).

3.1 Pilot Dataset

We released the pilot datasets for our SemEval shared task one month before the start of the shared task. The pilot datasets allowed the participants to have a better understanding of the shared task (i.e., the datasets, the languages involved, and the labels).

3.2 Task Settings

The AfriSenti-SemEval shared task consisted of two phases: (1) the development phase and (2) the evaluation phase. In the development phase, we released a training set with gold labels and a development set without gold labels. Participants trained their models on the training set, tested on the development set, and submitted their predictions on the CodaLab competition page for evaluation. The task offers a prize to the best-performing team in each of the three sub-tasks (A, B, and C) based on the following criteria[2]: (1) African League—for teams with at least one African member to encourage African participation; (2) Students League—for Master's and Undergraduate students only; and (3) Worldwide League—open to all participants.

4 Dataset and Lexicon

The AfriSenti collection covers 14 African languages, each with unique linguistic characteristics, writing systems, and language families, as shown in Table 1. The dataset covers four of the five African sub-regions and includes the top three languages with the largest number of speakers in Africa (Swahili, Amharic, and Hausa). The datasets include tweets collected using location-based and vocabulary-based (i.e., stopwords, sentiment lexicons, or language-specific terms) heuristics. Figure 3 shows the label distribution for the

---

[2]https://github.com/afrisenti-semeval/afrisent-semeval-2023

Language	Subregion	Script	Dataset			Manual Lexicon
			Train	**Dev**	**Test**	
Amharic(`amh`)	East Africa	Ethiopic	5,985	1,498	2,000	✓
Algerian Arabic/Darja (`arq`)	North Africa	Arabic	1,652	415	959	✗
Hausa (`hau`)	West Africa	Latin	14,173	2,678	5,304	✓
Igbo (`ibo`)	West Africa	Latin	10,193	1,842	3,683	✓
Kinyarwanda (`kin`)	East Africa	Latin	3,303	828	1,027	✗
Moroccan Arabic/Darija (`ary`)	North Africa	Arabic/Latin	5,584	1,216	2,962	✗
Mozambican Portuguese (`pt-MZ`)	Southeast Africa	Latin	3,064	768	3,663	✗
Nigerian Pidgin (`pcm`)	West Africa	Latin	5,122	1,282	4,155	✓
Oromo (`orm`)	East Africa	Latin	-	397	2,096	✓
Swahili (`swa`)	East Africa	Latin	1,198	454	749	✗
Tigrinya (`tir`)	East Africa	Ethiopic	-	399	2,001	✓
Twi (`twi`)	West Africa	Latin	3,482	389	950	✗
Xitsonga (`tso`)	South Africa	Latin	805	204	255	✓
Yorùbá (`yor`)	West Africa	Latin	8,523	2,091	4,516	✓

Table 1: African languages included in our study (Lewis, 2009; Muhammad et al., 2023).
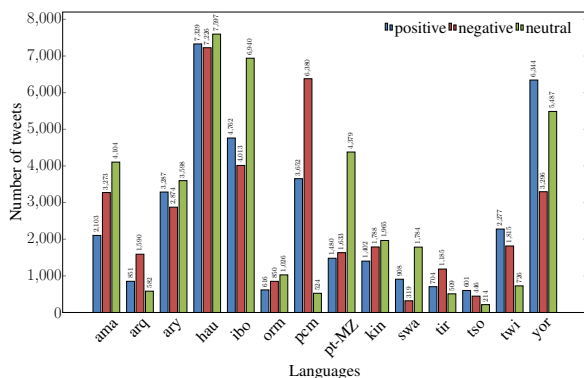

Figure 3: Label distributions for the AfriSenti datasets.

different languages in the AfriSenti dataset. For more information on the AfriSenti dataset collection, annotation, and linguistic challenges, please refer to the AfriSenti dataset paper (Muhammad et al., 2023).

Early work on sentiment analysis in SemEval showed that sentiment lexicons can be leveraged and combined with training data and machine learning algorithms to obtain marked improvements in accuracy (Mohammad et al., 2013; Kiritchenko et al., 2014). Therefore, we also provide manually annotated sentiment lexicons in African languages[3]. For languages that do not have manually curated lexicons, we translated existing lexicons into the target languages. Table 1 provides details on the sentiment lexicons in AfriSenti and indicates whether they were manually created or translated.

5 Evaluation

All three tasks in the AfriSenti-SemEval shared task required participants to perform a sentiment (negative, neutral, positive) classification. To evaluate the performance of the systems submitted by the participating teams, we used weighted F1 as the evaluation metric. For each label, weighted F1 measure its performance and weight it by the number of actual instances it has. This adjusts the 'macro' method to deal with uneven labels. We also provided the evaluation script to the participants to ensure consistency in the evaluation process.

We also created baseline systems for all three sub-tasks using multilingual pre-trained language models (PLMs). The baseline systems are: (1) monolingual baseline models based on multilingual PLMs for the 12 AfriSenti languages with training data; (2) models with multilingual training on all 12 languages, and evaluation on the combined test data of all 12 languages; and (3) zero-shot transfer of models to Oromo (`orm`) and Tigrinya (`tir`) from any of the 12 languages with available training data. Our best baseline model is shown in Table 2 and is based on *fine-tuning* AfroXLMR-large[4] (Alabi et al., 2022) in all the three sub-tasks. For more information on the baseline experimental results, please refer to the AfriSenti dataset paper (Muhammad et al., 2023).

---

[3]Dataset, scripts, lexicons available here:
https://github.com/afrisenti-semeval/
afrisent-semeval-2023

[4]https://huggingface.co/Davlan/
afro-xlmr-large

#	Sub-task A (monolingual)		Sub-task B (multilingual)		Sub-task C (zero-shot)	
	Team	F1	Team	F1	Team	F1
*	AfriSenti baseline	67.20	AfriSenti baseline	71.20	AfriSenti baseline	57.90
1	NLNDE	71.31	NLNDE	75.06	UCAS-IIE-NLP	58.15
2	PALI	70.28	DN	72.55	NLNDE	57.92
3	UM6P	69.54	UM6P	71.95	UM6P	57.40
4	NLP-LISAC	67.93	GMNLP	71.24	DN	55.19
5	GMNLP	67.65	Masakhane-Afrisenti	70.34	GMNLP	51.70
6	ABCD Team	67.51	UCAS-IIE-NLP	70.25	ABCD Team	51.59
7	UCAS-IIE-NLP	64.48	Witcherses	69.56	Masakhane-Afrisenti	50.04
8	HausaNLP	66.31	HausaNLP	69.54	UBC-DLNLP	49.41
9	UIO	65.41	ABCD Team	69.22	NLP-LISAC	47.33
10	Masakhane-AfriSenti	65.27	NLP-LISAC	67.30	FIT BUT	45.33

Table 2: Top 10 submissions for Tasks A, B and C. We only ranked systems with corresponding paper submissions. See Table 6 in Appendix C for paper information and teams' affliations. * show the baseline result from the AfriSenti dataset paper (Muhammad et al., 2023).

6 Participating Systems and Results

The AfriSenti-SemEval competition had 213 registered participants on the CodaLab competition website. Of these, 44 teams submitted their systems during the evaluation phase. Out of the 44 submissions, 29 submitted system-description papers. As participants could participate in one or more tasks, certain tasks received more submissions than others. Specifically, Task A (monolingual classification) had the highest number of participants with 44 submissions from different teams, followed by Task C (zero-shot classification) with 34 submissions and Task B (multilingual classification) with 33 submissions from different teams.

The majority of the teams participated in all tracks of each task, with 24 teams participating in at least 13 out of 15 tracks. For example, team NLNDE participated in all 12 tracks in Task A, one track in Task B, and two tracks in Task C. To rank the best-performing teams in each task and provide a comparison for future work, we rank each of the top-10 teams that participated in all tracks in a given task based on their average performance, as shown in Table 2.

Table 3, Table 4, and Table 5 present the overall results of participating systems for Task A (Monolingual), Task B (Multilingual), and Task C (Zero-shot), respectively. Table 6 in Appendix C presents information regarding the teams and their affiliations. In the following sections, we describe the best systems in each subtask.

6.1 Subtask A: Monolingual Sentiment Classification Systems

We describe the top-10 teams that submitted system description papers as highlighted in Table 2.

NLNDE (Wang et al., 2023) used language adaptive pre-training (LAPT) and task adaptive pre-training (TAPT) as an additional pre-training step on AfroXLMR-large. The LAPT approach involved continued pre-training of the PLM on the monolingual portion of the Leipzig Corpus Collection (Goldhahn et al., 2012) (covering Wikipedia, Community, Web, and News corpora) for the target language. TAPT involved continued pre-training on the AfriSenti training data of the target language. By leveraging LAPT followed by TAPT, they achieved significant improvements over fine-tuning AfroXLMR-large directly. **NLNDE ranked first in 7 out of 12 languages, and first in sub-task A**.

PALI (Jin et al., 2023) used weighted fusion of several PLMs such as *naija-twitter-sentiment-afriberta-large*[5], an AfriBERTa PLM trained on the NaijaSenti dataset (Muhammad et al., 2022), TwHIN-BERT (Zhang et al., 2022), and mDeBER-TaV3 (He et al., 2021) trained on AfriSenti corpora translated to English. The performance of the PLMs varied across different languages. For example, AfriBERTa fine-tuned on NaijaSenti provided better results on Nigerian languages. **PALI**

---

[5] https://huggingface.co/Davlan/naija-twitter-sentiment-afriberta-large

#	Team	ama	arq	ary	hau	ibo	kin	pcm	pt-MZ	swa	tso	twi	yor	Avg.
1	NLNDE	64.04	69.99	**64.83**	**82.62**	**82.96**	**72.63**	71.94	72.90	**65.68**	**60.67**	67.51	79.95	**71.31**
2	PALI	65.56	72.62	55.92	81.10	81.30	69.61	75.16	73.83	64.37	56.26	67.58	80.06	70.28
3	king001	69.77	73.00	57.94	81.11	81.39	60.25	75.75	73.53	64.89	56.26	**68.28**	**80.16**	70.20
4	stce	65.56	71.72	55.42	80.99	81.37	69.61	75.30	73.57	64.37	56.26	67.58	80.08	70.15
5	UM6P	72.18	72.02	60.15	82.04	81.51	70.71	69.14	67.35	60.26	56.13	66.98	76.01	69.54
6	mitchelldehaven	60.83	69.64	63.54	78.75	78.99	72.48	69.17	72.28	63.01	60.32	65.65	77.25	69.35
7	ymf924	69.83	67.83	57.12	75.88	72.49	70.88	72.48	73.80	64.23	55.38	62.56	75.94	68.20
8	NLP-LISAC	64.71	**74.20**	62.11	79.74	79.66	65.59	68.00	65.90	59.69	53.72	66.38	75.42	67.93
9	GMNLP	**78.42**	67.99	55.23	79.56	75.34	71.80	68.84	71.90	63.70	51.67	56.46	70.84	67.65
10	ABCD Team	58.05	63.50	61.54	81.50	82.28	67.36	66.30	67.21	63.10	53.92	65.61	79.73	67.51
11	tmn	63.11	70.54	57.22	72.93	72.66	69.82	69.63	73.09	61.66	53.66	67.06	75.08	67.21
*	AfriSenti paper baseline	61.60	68.30	56.60	80.70	79.50	70.60	68.70	71.60	63.40	47.30	64.30	74.10	67.20
12	PA14	64.52	72.01	55.34	77.17	80.29	61.16	75.53	**74.98**	61.73	50.46	67.39	65.65	67.19
13	UCAS-IIE-NLP	67.45	63.31	55.36	80.79	78.51	71.47	66.63	65.13	60.73	51.09	60.15	77.13	66.48
14	HausaNLP	57.30	65.12	58.49	80.97	76.96	70.61	68.48	68.37	63.23	50.27	64.07	71.88	66.31
15	UIO	56.86	69.21	57.46	74.53	77.58	61.66	67.08	71.64	56.69	60.07	65.02	66.99	65.40
16	Masakhane-Afrisenti	68.85	64.35	55.50	73.12	73.75	64.14	66.88	70.26	60.24	54.33	62.83	68.98	65.27
17	DN	57.34	65.81	57.20	81.09	74.51	71.91	64.89	69.09	62.51	46.62	55.53	72.07	64.88
18	UMUTeam	55.39	68.52	54.75	73.92	76.78	65.03	65.55	71.14	60.52	54.89	63.01	66.10	64.63
19	Howard University CS	61.68	62.49	47.34	77.68	78.02	66.50	66.12	67.72	52.31	48.80	62.84	74.72	63.85
20	FIT BUT	65.10	62.15	52.08	72.56	75.64	67.25	65.00	63.54	57.90	52.86	63.41	67.93	63.79
21	UBC-DLNLP	56.88	64.02	53.06	79.37	77.52	62.02	65.57	61.98	58.60	45.49	65.14	71.02	63.39
22	Witcherses	39.09	57.55	50.68	79.65	80.87	62.69	64.44	65.03	58.91	52.82	66.47	78.44	63.05
23	Sefamerve	70.18	71.74	52.54	78.59	73.24	63.15	63.67	67.94	63.94	48.27	65.66	25.33	62.02
24	FUOYENLP	54.25	68.14	52.78	73.17	72.22	54.20	68.01	72.88	54.31	51.38	57.94	64.31	61.97
25	efrat4050	31.99	55.40	54.43	75.65	76.28	59.56	53.17	62.75	51.76	54.10	64.18	72.93	59.34
26	JCT	53.76	24.59	43.95	79.50	77.08	58.22	62.52	62.96	58.98	54.10	61.82	69.98	58.96
27	JacobLevy248	21.55	54.71	46.32	73.38	78.14	57.92	57.47	62.67	55.97	51.26	64.06	64.83	57.36
28	jacklight971	53.76	24.59	43.95	79.57	75.79	51.41	62.17	55.65	58.98	42.74	60.37	63.96	56.08
29	MaChAmp	2.26	32.87	12.79	17.02	26.91	17.79	40.20	51.17	44.22	26.01	20.17	18.87	25.86
NR	TechSSN	-	-	-	80.32	-	-	-	-	-	-	-	-	80.32
NR	KINLP	-	-	-	-	-	72.50	-	-	-	-	-	-	72.50
NR	afrisenti23kb	-	-	-	-	-	71.00	-	-	-	-	-	-	71.00
NR	foul	-	70.36	-	-	-	-	-	-	-	-	-	-	70.36
NR	Bhattacharya_Lab	-	-	-	-	-	47.10	**75.96**	-	-	-	-	79.86	67.64
NR	uid	-	-	55.29	80.45	78.51	70.99	-	-	-	46.91	59.63	76.98	66.97
NR	TBS	-	-	55.29	80.67	78.47	70.98	-	-	-	49.40	55.11	76.57	66.64
NR	DuluthNLP	-	-	-	-	-	65.85	-	-	-	-	64.29	-	65.07
NR	Uppsala University	-	-	-	-	-	64.69	-	-	-	-	-	-	64.69
NR	Seals_Lab	54.67	-	-	80.85	80.82	-	-	52.43	52.18	-	-	-	64.19
NR	NLP-LTU	-	-	60.27	-	-	-	-	-	-	-	-	-	60.27
NR	Trinity	-	-	-	76.53	-	-	-	-	47.55	-	-	50.00	58.03
NR	ronaharo	14.98	-	39.72	72.84	73.21	57.87	62.52	-	53.35	-	-	-	53.50
NR	GunadarmaxBRIN	-	-	-	-	-	-	-	-	-	50.11	-	-	50.11
NR	aptxaaaa	-	-	51.01	-	-	-	-	-	-	48.97	-	-	49.99

Table 3: Task A Results. The ranking is based on the average of the scores. Partial submissions were not included in the final ranking. (**NR** - No Ranking.)

ranked first in 2 out of 12 languages, and second for sub-task A.

UM6P (El Mahdaouy et al., 2023) combined MARBERT (Abdul-Mageed et al., 2021) and an adapted AfroXLMR using a projection and a residual layer. First, AfroXLMR was first fine-tuned on the AfriSenti training data and web resources like MAFAND-MT (Adelani et al., 2022a) and WebCrawl African multilingual parallel corpora (Vegi et al., 2022) using a whole-word masking objective (Cui et al., 2021). It was then fine-tuned on the labeled AfriSenti training data and combined with MARBERT. The **UM6P team ranked 5th for sub-task A.**

NLP-LISAC (Benlahbib and Boumhidi, 2023) used different PLMs for different languages. For Hausa, Yoruba, Igbo, Swahili, and Kinyarwanda, they used mBERT (Devlin et al., 2019) that was adapted to African languages with continued pre-training. The checkpoints[6] were based on models released by Adelani et al. (2021). For Mozambican Portuguese, Xitsonga, Nigerian-Pidgin, and Amharic, they used mDeBERTaV3. They used DziriBERT (Abdaoui et al., 2021), a model specifically trained for the Algerian dialect, for Algerian Arabic and AfriBERTa-large for Twi. **NLP-LISAC ranked first for Algerian Arabic**.

GMNLP (Alam et al., 2023) used phylogeny-based adapter-tuning (Faisal and Anastasopoulos, 2022) of AfroXLMR-large. They used a dictionary and machine translation-based data augmentation strategies to increase the amount of training data.

[6]https://huggingface.co/models?search=davlan/xlm-roberta-base-finetuned

The dictionary approach used PanLex (Kamholz et al., 2014) while the MT approach was based on the WMT Shared task model (Alam and Anastasopoulos, 2022). However, this excluded some languages (Twi, Mozambican Portuguese, Nigerian Pidgin, Moroccan Arabic, and Algerian Arabic). Their findings suggest that using phylogeny adapters (language adapter, genus adapter, and family adapter) can lead to better performance. For example, on Amharic, Hausa, Kinyarwanda, Nigerian-Pidgin, and Xitsonga, all the phylogeny adapters were useful. **GMNLP ranked first for Amharic**.

ABCD Team (Dang et al., 2023) used soft voting ensemble of three PLMs: AfroXLMR, AfriBERTa, and LaBSE – a multilingual sentence transformer model that supports over 100 languages, and is also popular for mining parallel corpus for machine translation. The **ABCD team ranked 10th in the sub-task A competition**.

UCAS-IIE-NLP (Hu et al., 2023) used a lexicon-based multilingual transformer model based on AfroXLMR-base to facilitate language adaptation and sentiment-aware representation learning. Additionally, they applied a supervised adversarial contrastive learning strategy to improve the sentiment-spread representations and enhance model generalization. On average, their approach was worse than the AfriSenti baseline likely because they used AfroXLMR-base rather than the large version. Interestingly, they achieved much better results than the baseline on Amharic, Xitsonga, and Yorùbá by over 3 F1 points.

HausaNLP (Abdullahi et al., 2023) used two BERT-based models: AfroXLMR-large and an Arabic BERT (Inoue et al., 2021) fine-tuned on a sentiment corpus[7]. They used AfroXLMR-large for all languages except the Arabic dialects. On average across 12 languages, the HausaNLP system ranked lower than the AfriSenti paper baseline.

UIO (Rønningstad, 2023) compared four different PLMs: AfroXLMR-mini[8] (based on language adaptation of MiniLMv2 (Wang et al., 2021), a compressed XLM-R-large model), mpnet-base-v2 sentence transformer (Reimers and Gurevych, 2019), and XLM-T (Barbieri et al., 2022) and

XLM-T fine-tuned on a multilingual sentiment corpus[9]. They chose the latter as the best model since it was pre-trained both on the Twitter domain and a sentiment classification task. This aligns with the findings of (Muhammad et al., 2023) where XLM-T was the second-best model because of in-domain pre-training. However, UIO's system underperformed the AfriSenti baseline as making use of a large Afro-centric PLM like AfroXLMR-large outperforms the use of domain and task-specific data with smaller models.

Masakhane-AfriSenti (Azime et al., 2023) used five PLMs: LabSE, AfriBERTa, AfroXLMR-base, XLM-T, and Bernice (DeLucia et al., 2022), a multilingual Twitter RoBERTa model. The performance of their best systems varied by language. For example, they reported AfroXLMR-base to be better for Hausa and Swahili, while LaBSE gave better results for Kinyarwanda and Mozambique Portuguese. In general, their best systems ranked lower than the AfriSenti baseline.

Apart from the top-10 teams that submitted their papers, there were other teams that only worked on one or few languages, and achieved excellent rankings. For instance, **KINLP** (Nzeyimana, 2023) only attempted the task for the Kinyarwanda language. Their approach was based on KinyaBERT (Nzeyimana and Niyongabo Rubungo, 2022), a Kinyarwanda PLM that incorporates morphological features of the language during pre-training. **KINLP** ranked second for Kinyarwanda. **Bhattacharya_Lab** (Hughes et al., 2023) only worked on Nigerian-Pidgin and Yorùbá. They pre-trained a RoBERTa-style (Liu et al., 2019) transformer-based architecture jointly on the two languages using the AfriBERTa training corpus and AfriSenti data. **Bhattacharya_Lab ranked first for Nigerian-Pidgin, and fifth for Yorùbá**.

6.2 Subtask B: Multilingual Sentiment Classification Systems

Most teams used the same model as in sub-task A for sub-task B with minor changes. We highlight here the teams that used strategies apart from jointly training a PLM on the concatenation of all 12 sub-task A languages. We describe the top-10 teams with corresponding system description papers as shown in Table 2. The complete results for this sub-task are shown in Table 4.

---

[7]https://huggingface.co/CAMeL-Lab/
bert-base-arabic-camelbert-da-sentiment

[8]https://huggingface.co/Davlan/afro-xlmr-mini

[9]https://huggingface.co/cardiffnlp/
twitter-xlm-roberta-base-sentiment

#	Team	F1
1	NLNDE	**75.06**
2	king001	74.96
3	DN	72.55
4	ymf924	72.34
5	mitchelldehaven	72.33
6	UM6P	71.95
7	GMNLP	71.24
*	AfriSenti baseline	71.20
8	PA14	70.81
9	Masakhane-Afrisenti	70.34
10	UCAS-IIE-NLP	70.25
11	tmn	70.05
12	Witcherses	69.56
13	HausaNLP	69.54
14	ABCD Team	69.22
15	iREL	68.23
16	BERT 4EVER	67.46
17	NLP-LISAC	67.30
18	Howard University CS	66.60
19	ISCL_WINTER	66.38
20	UIO	66.29
21	FIT BUT	66.02
22	FUOYENLP	65.51
23	UMUTeam	65.47
24	GunadarmaxBRIN	65.23
25	saroyehun	64.34
26	JacobLevy248	64.04
27	efrat4050	63.85
28	Sefamerve	63.40
29	DuluthNLP	61.49
30	jacklight971	59.75
31	JCT	59.50
32	Hercules	59.35
33	ronaharo	57.79

Table 4: AfriSenti-SemEval Task B results.

NLNDE For each target language, they first chose the best source languages for multilingual training to prevent harmful interference from dissimilar languages. For selecting the source language set, they performed forward and backward source language selection, similar to feature selection approaches (Tsamardinos and Aliferis, 2003; Borboudakis and Tsamardinos, 2019). Forward feature selection starts with an empty set of languages and adds languages to it, while backward feature selection starts with a complete set of languages and then excludes languages from it. For example, the best source languages for multilingual training for Hausa using Forward selection are Kinyarwanda, Twi, Algerian Arabic and Nigerian-Pidgin. For Yorùbá, the best source languages according to the backward selection are Kinyarwanda, Xitsonga, Twi, and Algerian-Arabic. NLNDE used multiple models for this task rather than a single one like most teams did. The target language of the tweet would determine the corresponding multi-

ple training strategy. **NLNDE ranked first in this sub-task**.

DN (Homskiy and Maloyan, 2023) They fine-tuned AfroXMLR-large on the 12 languages available in the training data. They performed additional pre-processing on the tweets before training, i.e., they removed links, hashtags, and @mentions, which boosted the performance of their system over those trained on a single multilingual model on all 12 languages. **DN ranked third in this sub-task**.

GMNLP , unlike in sub-task A, did not use a phylogeny-based adapter fine-tuning for this sub-task due to the absence of language ID information. They only performed task adapter training.

Witcherses (Gokani et al., 2023) used an ensemble of both multilingual PLMs and classical ML models: Logistic Regression, Random Forest, Support Vector Machine, AfriBERTa, AfroXLMR, and AfriBERTa fine-tuned on the NaijaSenti corpus.

Hausa NLP used the same approach as in sub-task A. They used AfroXLMR-large for multilingual training, which was previously fine-tuned on MasakhaNER 2.0[10] (Adelani et al., 2022c).

NLP-LISAC used the same approach described in sub-task A. They chose the mDeBERTaV3 PLM to fine-tune on the multilingual corpus.

The other teams i.e. **UM6P**, **Masakhane-AfriSenti**, **UCAS-IIE-NLP**, and **ABCD Team** used the same approach as the one they used for sub-task A. The only difference was that they trained a PLM on all 12 languages instead of training a monolingual sentiment model.

6.3 Subtask C: Zero-Shot Sentiment Classification Systems

We provide an overview of the top-10 submissions with system description papers in Table 2 and show the complete results for this sub-task in Table 5.

UCAS-IIE-NLP used the same approach described in sub-task A. They used additional lexicon information for zero-shot transfer to both Oromo and Tigrinya. **UCAS-IIE-NLP ranked first for sub-task C, first for Oromo, and second for Tigrinya.** Surprisingly, their best performance is below the AfriSenti baseline for Oromo (-1.28

---

[10]https://huggingface.co/masakhane/afroxlmr-large-ner-masakhaner-1.0_2.0

#	Team	orm	tir	Avg.
1	UCAS-IIE-NLP	45.82	70.47	**58.15**
2	NLNDE	44.97	**70.86**	57.92
*	AfriSenti baseline	**47.10**	68.60	57.90
3	ymf924	45.34	70.39	57.87
4	UM6P	45.27	69.53	57.40
5	TBS	45.12	69.61	57.37
6	uid	44.75	69.90	57.33
7	mitchelldehaven	46.23	66.96	56.60
8	tmn	41.97	69.09	55.53
9	DN	41.45	68.93	55.19
10	GMNLP	41.87	61.52	51.70
11	ABCD Team	42.64	60.53	51.59
12	BERT 4EVER	42.32	60.37	51.35
13	Masakhane-Afrisenti	42.09	57.99	50.04
14	UBC-DLNLP	41.79	57.03	49.41
15	NLP-LTU	36.89	61.48	49.19
16	NLP-LISAC	42.02	52.64	47.33
17	FIT BUT	38.24	52.42	45.33
18	PA2023	27.76	62.27	45.02
19	saroyehun	37.79	45.19	41.49
20	MaChAmp	38.02	44.86	41.44
21	FUOYENLP	35.57	47.18	41.38
22	Sefamerve	39.35	34.16	36.76
23	PA14	39.86	29.49	34.68
NR	yg.ren	-	60.87	60.87
NR	hk001	-	60.35	60.35
NR	abc1234	-	59.07	59.07
NR	abc123	-	55.49	55.49
NR	abc1	-	42.42	42.42
NR	Bhattacharya_Lab	40.61	-	40.61
NR	Seals_Lab	40.61	-	40.61
NR	Snarci	33.98	-	33.98
NR	abc12	-	33.33	33.33
NR	GunadarmaxBRIN	32.05	-	32.05
NR	HausaNLP	25.43	-	25.43

Table 5: Task C Results. The ranking is based on the average of the scores. Partial submissions are not included in the final ranking. (**NR** - No Ranking.)

F1), which is based on choosing the best language for zero-shot transfer. Muhammad et al. (2023) identified Hausa and Amharic as the best source languages for Oromo and Hausa and Yorùbá as the best source languages for Tigrinya. Co-training on the two languages led to a better performance.

NLNDE used the same approach as in sub-task B. They used forward and backward language selection to decide the best source languages to transfer from. For Oromo, the best source languages they identified were Kinyarwanda, Hausa, Yorùbá, and Xitsonga using forward selection, and Yoruba, Mozambique Portuguese, and Xitsonga using backward selection. Similarly, for Tigrinya, they identified Hausa, Kinyarwanda, Amharic, Moroccan Arabic, and Mozambique Portuguese in the forward selection, and Mozambican Portuguese, Yorùbá, and Hausa. The languages selected were similar to

those identified by Muhammad et al. (2023). **NLNDE ranked second on sub-task C and first for Tigrinya**.

Masakhane-AfriSenti used the multilingual model they introduced in sub-task B based on AfroXLMR-base and AfriBERTa. They also tried adapter-based training based on MAD-X (Pfeiffer et al., 2020). Their final result is based on an ensemble of the three methods.

UBC-DLNLP (Bhatia et al., 2023) compared about 10 PLMs for all the sub-tasks including XLM-R, AfriBERTa, mBERT, AfroXLMR, Serengeti (Adebara et al., 2022b), Serengeti_ft (Serengeti + TAPT), and AfroXLMR-base_ft (AfroXLMR-base + TAPT). Their best-reported result is based on AfroXLMR-base_ft. Surprisingly, AfroXLMR-base_ft and AfroXLMR-base, which were pre-trained on 17 African languages outperformed Serengeti_ft & Serengeti, which were pre-trained on 500+ African languages. This indicates that training on a large number of languages may suffer from the curse of multilinguality (Conneau et al., 2020) without appropriate scaling since it is difficult to learn a good representation for 500 languages with a small PLM of 270M parameters.

FIT BUT (Aparovich et al., 2023) used AfroXLMR-small with additional adversarial training but they only achieved average performance. This is probably due to the use of a small PLM for training.

Other teams like **NLP-LISAC**, **UM6P**, **DN**, **GMNLP**, and **ABCD** adopted approaches similar to those of sub-task B: they trained on all multilingual datasets and performed zero-shot evaluation on Oromo and Tigrinya.

7 Discussion

We summarize some of the approaches that led to the best results in different sub-tasks.

Sub-task A All of the top-10 best-performing teams with systems description papers employed multilingual pre-trained models, especially Afrocentric models. For example, eight of the ten teams make use of AfroXLM—one of the best-performing PLM for African languages. AfroXLMR-large with additional pre-training often led to the best results while multilingual PLMs like mDeBERTaV3 and LaBSE led to competitive

results. A few teams used other PLMs, specifically trained on Arabic variants such as DziriBERT. Some teams also reported significant language-specific improvements using further domain and task-specific pre-training. For instance, the NL-NDE team, which ranked first, used both language and task adaptive pre-training. UIO and Masakhane-AfriSenti also demonstrated the benefit of domain adaptive pre-training. In addition, PALI and Masakhane-AfriSenti showed that using a PLM that has already been trained on sentiment classification can help. Interestingly, other teams using an ensemble of different fine-tuned PLMs tended to perform worse, which highlights that the quality of individual models is important.

Sub-task B Most teams used a single multilingual PLM and fine-tuned it on all languages. In fact, most of the best-ranking teams used AfroXLMR-large as it performed well on sub-task A. The best-performing team for this task, NLNDE, chose to select the most appropriate languages to co-train for each language before performing multilingual training, highlighting the importance of the choice of source languages.

Sub-task C UCAS-IIE-NLP ranked first and used a lexicon-based multilingual BERT. This shows the usefulness of leveraging sentiment lexicons as side information in building language models. However, their best performance was below the AfriSenti paper baseline for Oromo (-1.28 F1).

The top-performing teams in each subtask were not affiliated with African institutions. They developed the best models despite a lack of language expertise. This highlights both the generality of existing models and adaptation paradigms as well as the need for a more collaborative approach to building more effective and inclusive solutions for Africa-centric sentiment analysis.

8 Conclusion

We presented the *SemEval-2023 Task 12: Sentiment Analysis for African Languages*, the first SemEval shared task that focuses on sentiment analysis for African languages. The task included monolingual classification (in Amharic, Algerian Arabic, Hausa, Igbo, Kinyarwanda, Moroccan Arabic, Mozambican Portuguese, Nigerian Pidgin, Swahili, Twi, Xitsonga, and Yorùbá), multilingual classification, and zero-shot classification (in Oromo and Tigrinya). We described the task settings, datasets,

and baselines.

We discussed the main findings of the 44 participating teams who submitted their systems, based on their system description papers (i.e., 29 papers) as well as our observations and analysis of some common errors. Overall, the best ranking teams used pre-trained language models (PLMs), with Africa-centric models such as AfroXLMR performing the best in the Task A (monolingual) and Task B (multilingual) classification tasks, with an average weighted F1 of 71.3%, and 75.06%, respectively. For Task C (zero-shot), the top team used lexicon-based multilingual BERT and achieved an average weighted F1 of 58.15.%. These scores indicate that there is still room for improvement in polarity classification in low-resource settings.

By sharing our insights, we aim to encourage researchers to work on under-resourced and under-studied African languages and help them improve the performance of current sentiment analysis systems. In the future, we will extend our task to more languages by building additional datasets.

9 Ethics Statement

People often express sentiment in unique and interesting ways. Thus, there is large amounts of person–person variation. Therefore, any automatic method for sentiment analysis will achieve different results on data from different people, from different domains, etc. We do not recommend the use of automatic methods of sentiment analysis (based on individual instances of text) to make important decisions that can impact an individual. Instead, it is often better to use automatic sentiment analysis to determine broad trends of sentiment across large amounts of data. Sentiment analysis, like many other AI technologies, can be used not just for beneficial purposes, but also to cause harm such as using it to identify and suppress dissent. There are several such ethical considerations that should be accounted for when developing and deploying sentiment analysis systems. We refer to Mohammad (2022, 2023) for a comprehensive discussion of ethical considerations relevant to sentiment and emotion analysis.

References

Amine Abdaoui, Mohamed Berrimi, Mourad Oussalah, and Abdelouahab Moussaoui. 2021. Dziribert: a pre-trained language model for the algerian dialect. *arXiv preprint arXiv:2109.12346*.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Saheed Abdullahi, FALALU LAWAN, Ahmad Wali, Amina Imam Abubakar, Musa Bello, Aliyu Rabiu Shuaibu, Saminu Aliyu, Murja Gadanya, Nur Rabiu, Yusuf Aliyu, Sanah Abdullahi Muaz, Mahmoud Said Ahmad, Abdulkadir Abdullahi, and Abdulmalik Yusuf Jamoh. 2023. HausaNLP at SemEval-2023 Task 12: Leveraging African Low Resource TweetData for Sentiment Analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Idris Abdulmumin, Michael Beukman, Jesujoba Alabi, Chris Chinenye Emezue, Everlyn Chimoto, Tosin Adewumi, Shamsuddeen Muhammad, Mofetoluwa Adeyemi, Oreen Yousuf, Sahib Singh, and Tajuddeen Gwadabe. 2022. Separating grains from the chaff: Using data filtering to improve multilingual translation for low-resourced african languages. In *Proceedings of the Seventh Conference on Machine Translation*, pages 1001–1014, Abu Dhabi. Association for Computational Linguistics.

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.

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2022a. Afrolid: A neural language identification tool for african languages. *arXiv preprint arXiv:2210.11744*.

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2022b. Serengeti: Massively multilingual language models for africa. *ArXiv*, abs/2212.10785.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo,

Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022a. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

David Adelani, Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Akshita Bhagia, Marta R. Costa-jussà, Jesse Dodge, Fahim Faisal, Christian Federmann, Natalia Fedorova, Francisco Guzmán, Sergey Koshelev, Jean Maillard, Vukosi Marivate, Jonathan Mbuya, Alexandre Mourachko, Safiyyah Saleem, Holger Schwenk, and Guillaume Wenzek. 2022b. Findings of the WMT'22 shared task on large-scale machine translation evaluation for African languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 773–800, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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.

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O. Alabi, Shamsuddeen H. Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure

F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire M. Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Elvis Mboning, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo L. Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Adeyemi, Gilles Q. Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu Ngoli, and Dietrich Klakow. 2022c. MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Basant Agarwal and Namita Mittal. 2016. Machine learning approach for sentiment analysis. In *Prominent feature extraction for sentiment analysis*, pages 21–45. Springer.

Samuel Akrah and Ted Pedersen. 2023. DuluthNLP at SemEval-2023 Task 12: AfriSenti-SemEval: Sentiment Analysis for Low-resource African Languages Using Twitter Dataset. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

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.

Md Mahfuz Ibn Alam and Antonios Anastasopoulos. 2022. Language adapters for large-scale MT: The GMU system for the WMT 2022 large-scale machine translation evaluation for African languages shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1015–1033, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Md Mahfuz Ibn Alam, Ruoyu Xie, Fahim Faisal, and Antonios Anastasopoulos. 2023. GMNLP at SemEval-2023 Task 12: Sentiment Analysis with Phylogeny-based Adapters. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Maksim Aparovich, Santosh Kesiraju, Aneta Dufkova, and Pavel Smrz. 2023. FIT BUT at SemEval-2023 Task 12: Sentiment without Borders - Multilingual Domain Adaptation for Low-resource Sentiment Classification. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Novitasari Arlim, Slamet Riyanto, Rodiah Rodiah, and Al Hafiz Akbar Maulana Siagian. 2023. GunadarmaXBRIN at SemEval-2023 Task 12: Utilization of SVM and AfriBERTa for Monolingual, Multilingual, and Zero-shot Sentiment Analysis in African Languages. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Saurav Aryal and Howard Prioleau. 2023. Howard University Computer Science at SemEval-2023 Task 12: A 2-Step System Design for Multilingual Sentiment Classification with Language Identification. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Israel Abebe Azime, Sana Sabah Al-Azzawi, Atnafu Lambebo Tonja, Iyanuoluwa Shode, Jesujoba O. Alabi, Ayodele Awokoya, Mardiyyah Oduwole, Tosin P. Adewumi, Samuel Fanijo, Oyinkansola Awosan, and Oreen Yousuf. 2023. Masakhane-afrisenti at SemEval-2023 Task 12: Sentiment Analysis using Afro-centric Language Models and Adapters for Low-resource African Languages. *CoRR*, abs/2304.06459.

JoAnne Banks-Wallace. 2002. Talk that talk: Storytelling and analysis rooted in african american oral tradition. *Qualitative Health Research*, 12(3):410–426. PMID: 11918105.

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 Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Jeremy Barnes, Laura Oberlaender, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. SemEval 2022 task 10: Structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1280–1295, Seattle, United States. Association for Computational Linguistics.

Tadesse Destaw Belay, Atnafu Lambebo Tonja, Olga Kolesnikova, Seid Muhie Yimam, Abinew Ali Ayele, Silesh Bogale Haile, Grigori Sidorov, and Alexander Gelbukh. 2022. The effect of normalization for bi-directional amharic-english neural machine translation. In *2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 84–89.

Faiza BELBACHIR. 2023. Foul at SemEval-2023 Task 12 : MARBERT Language Model and Lexical Filtering for Sentiments Analysis of Tweets in Algerian Arabic. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Abdessamad Benlahbib and Achraf Boumhidi. 2023. NLP-LISAC at SemEval-2023 Task 12: Sentiment Analysis for Tweets Expressed in African Languages via Transformer-based Models. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens, editors. 2017. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada.

Gagan Bhatia, Ife Adebara, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2023. UBC-DLNLP at SemEval-2023 Task 12: Impact of Transfer Learning on African Sentiment Analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Giorgos Borboudakis and Ioannis Tsamardinos. 2019. Forward-backward selection with early dropping. *J. Mach. Learn. Res.*, 20(1):276–314.

Erik Cambria, Qian Liu, Sergio Decherchi, Frank Xing, and Kenneth Kwok. 2022. SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3829–3839, Marseille, France. European Language Resources Association.

Jan Carter-Black. 2007. Teaching cultural competence: An innovative strategy grounded in the universality of storytelling as depicted in african and african american storytelling traditions. *Journal of Social Work Education*, 43(1):31–50.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Chenhua Chen, Zhiyang Teng, Zhongqing Wang, and Yue Zhang. 2022. Discrete opinion tree induction for aspect-based sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2051–2064.

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.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3504–3514.

Thin Dang, Dai Ba Nguyen, Dang Ba Qui, Duong Ngoc Hao, and Ngan Nguyen. 2023. ABCD Team at SemEval-2023 Task 12: An Ensemble Transformer-based System for African Sentiment Analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Selman Delil and Birol Kuyumcu. 2023. Sefamerve at SemEval-2023 Task 12: Semantic Evaluation of Rarely Studied Languages. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Alexandra DeLucia, Shijie Wu, Aaron Mueller, Carlos Aguirre, Philip Resnik, and Mark Dredze. 2022. Bernice: A multilingual pre-trained encoder for Twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6191–6205, Abu Dhabi, United Arab Emirates. Association for 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.

Peter Sheridan Dodds, Eric M. Clark, Suma Desu, Morgan R. Frank, Andrew J. Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M. Kloumann, James P. Bagrow, Karine Megerdoomian, Matthew T. McMahon, Brian F. Tivnan, and Christopher M. Danforth. 2015. Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, 112(8):2389–2394.

Abdelkader El Mahdaouy, Hamza Alami, salima lamsiyah, and Ismail Berrada. 2023. UM6P at SemEval-2023 Task 12: Out-Of-distribution Generalization Method for African Languages Sentiment Analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Fahim Faisal and Antonios Anastasopoulos. 2022. Phylogeny-inspired adaptation of multilingual models to new languages. 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 (Volume 1: Long Papers)*, pages 434–452, Online only. Association for Computational Linguistics.

José Antonio García-Díaz, Camilo Caparros-Laiz, Ángela Almela, Gema Alcaráz-Marbol, María José Marín-Pérez, and Rafael Valencia-García. 2023. UMUTeam at SemEval-2023 Task 12: Ensemble Learning of LLMs Applied to Sentiment Analysis for Low-resource African Languages. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. SemEval-2015 task 11: Sentiment analysis of figurative language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 470–478, Denver, Colorado. Association for Computational Linguistics.

Monil Gokani, KV Aditya Srivatsa, and Radhika Mamidi. 2023. Witcherses at SemEval-2023 Task 12: Ensemble Learning for African Sentiment Analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Itisha Gupta and Nisheeth Joshi. 2020. Enhanced twitter sentiment analysis using hybrid approach and by accounting local contextual semantic. *Journal of intelligent systems*, 29(1):1611–1625.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Alina Hancharova, John Wang, and Mayank Kumar. 2023. Team ISCL_WINTER at SemEval-2023 Task 12:AfriSenti-SemEval: Sentiment Analysis for Low-resource African Languages Using Twitter Dataset. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Daniil Homskiy and Narek Maloyan. 2023. DN SemEval2023 Task12. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Dou Hu, Lingwei Wei, Yaxin Liu, Wei Zhou, and Songlin Hu. 2023. UCAS-IIE-NLP at SemEval-2023 Task 12: Enhancing Generalization of Multilingual BERT for Low-resource Sentiment Analysis.

In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Nathaniel Hughes, Kevan Baker, Aditya Singh, Aryavardhan Singh, Tharalillah Dauda, and Sutanu Bhattacharya. 2023. Bhattacharya_Lab at SemEval-2023 Task 12: A Transformer-based Language Model for Sentiment Classification for Low Resource African Languages: Nigerian Pidgin and Yoruba. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190, Online. Association for Computational Linguistics.

Ebrahim Chekol Jibril and A. Cüneyd Tantuğ. 2023. Anec: An amharic named entity corpus and transformer based recognizer. *IEEE Access*, 11:15799–15815.

MeiZhi Jin, Cheng Chen, MengYuan Zhou, MengFei Yuan, XiaoLong Hou, XiYang Du, LianXin Jiang, and JianYu Li. 2023. PingAnLifeInsurance at SemEval-2023 Task 12: Sentiment Analysis for Low-resource African Languages with Multi-model Fusion. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a resource for pan-lingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3145–3150, Reykjavik, Iceland. European Language Resources Association (ELRA).

Chamandeep Kaur, Mawahib Sharafeldin Adam Boush, Samar Mansoor Hassen, Wafaa Abushmlah Hakami, Mohammed Hassan Osman Abdalraheem, Najla Mohammed Galam, Nedaa Abdulaziz Hadi, Nedaa Abdulaziz Hadi, and Atheer Omar S Benjeed. 2022. Incorporating sentimental analysis into development of a hybrid classification model. *Int. J. Health Sci. (IJHS)*, 6(S1):1709–1720.

Ron Keinan and Yaakov HaCohen-Kerner. 2023. JCT at SemEval-2023 Tasks 12a and 12B: Sentiment Analysis for Tweets Written in Low-resource

African Languages Using Various Machine Learning and Deep Learning Methods, Resampling, and HyperParameter Tuning. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M. Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research (JAIR)*, 50:723–762.

Annika Kniele and Meriem Beloucif. 2023. Uppsala University at SemEval-2023 Task12: Zero-shot Sentiment Classification for Nigerian Pidgin Tweets. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Hoai Bac Le and Huy Nguyen. 2020. Twitter sentiment analysis using machine learning techniques. In *International Conference on Computer Science, Applied Mathematics and Applications*.

Paul M. A. Lewis. 2009. Ethnologue : languages of the world.

Yunlong Liang, Fandong Meng, Jinan Xu, Yufeng Chen, and Jie Zhou. 2022. Msctd: A multimodal sentiment chat translation dataset. *arXiv preprint arXiv:2202.13645*.

Bing Liu. 2020. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press.

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. *ArXiv*, abs/1907.11692.

J. A. Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 105–119, Online. Association for Computational Linguistics.

Saif Mohammad. 2011. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114, Portland, OR, USA. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 321–327, Atlanta, Georgia, USA. Association for Computational Linguistics.

Saif M. Mohammad. 2021. Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text. In Herbert L. Meiselman, editor, *Emotion Measurement (Second Edition)*, pages 323–379. Woodhead Publishing.

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

Saif M. Mohammad. 2023. Best practices in the creation and use of emotion lexicons. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia. 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, 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. 2023. AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages.

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.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on*

Semantic Evaluation (SemEval-2016), pages 1–18, San Diego, California. Association for Computational Linguistics.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. SemEval-2013 task 2: Sentiment analysis in Twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA. Association for Computational Linguistics.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Antoine Nzeyimana. 2023. KINLP at SemEval-2023 Task 12: Kinyarwanda Tweet Sentiment Analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Antoine Nzeyimana and Andre Niyongabo Rubungo. 2022. KinyaBERT: a morphology-aware Kinyarwanda language model. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5347–5363, Dublin, Ireland. Association for Computational Linguistics.

Kelechi Ogueji, Yuxin Zhu, and Jimmy J. Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. *Proceedings of the 1st Workshop on Multilingual Representation Learning*.

Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, and Ritesh Kumar, editors. 2023. *Proceedings of the 17th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Toronto, Canada.

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2020. SemEval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 774–790, Barcelona (online). International Committee for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2020. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Transactions on Affective Computing*.

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. *J. Mach. Learn. Res.*, 21(1).

Shashank Rathi, Siddhesh Pande, Harshwardhan Atkare, Rahul Tangsali, Aditya Vyawahare, and Dipali Kadam. 2023. Trinity at SemEval-2023 Task 12: Sentiment Analysis for Low-resource African Languages Using Twitter Dataset. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Nilanjana Raychawdhary, Amit Das, Gerry Dozier, and Cheryl D. Seals. 2023. Seals_Lab at SemEval-2023 Task 12: Sentiment Analysis for Low-resource African Languages, Hausa and Igbo. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter S. Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):1–12. Copyright - EPJ Data Science is a copyright of Springer, 2016; Last updated - 2017-02-06.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*

and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SemEval-2015 Task 10: Sentiment Analysis in Twitter.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. SemEval-2014 Task 9: Sentiment Analysis in Twitter. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80.

Egil Rønningstad. 2023. UIO at SemEval-2023 Task 12: Multilingual Fine-tuning for Sentiment Classification in Low-resource Languages. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Rajalakshmi Sivanaiah, Nishaanth Ramanathan, Angel Deborah S, and Mirnalinee T T. 2023. TechSSN at SemEval-2023 Task 12: Monolingual Sentiment Classification in Hausa Tweets. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly D. Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37:267–307.

Ioannis Tsamardinos and Constantin F. Aliferis. 2003. Towards principled feature selection: Relevancy, filters and wrappers. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, volume R4 of *Proceedings of Machine Learning Research*, pages 300–307. PMLR. Reissued by PMLR on 01 April 2021.

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Annual Meeting of the Association for Computational Linguistics*.

Rob van der Goot. 2023. MaChAmp at SemEval-2023 Tasks 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12: On the Effectiveness of Intermediate Training on an Uncurated Collection of Datasets. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul, Abhinav Mishra, Prashant Banjare, Prasanna K R, and Chitra Viswanathan. 2022. WebCrawl African : A multilingual parallel corpora for African languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1076–1089, Abu Dhabi, United Arab Emirates (Hybrid). 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*, Toronto, Canada. Association for Computational Linguistics.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151, Online. Association for Computational Linguistics.

Ashima Yadav and Dinesh Kumar Vishwakarma. 2020. Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, 53(6):4335–4385.

Seid Muhie Yimam, Hizkiel Mitiku Alemayehu, Abinew Ali Ayele, and Chris Biemann. 2020. Exploring amharic sentiment analysis from social media texts: Building annotation tools and classification models. In *International Conference on Computational Linguistics*.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2022. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations. *arXiv preprint arXiv:2209.07562*.

A Algorithms Used

Algorithms used by the participants for data pre-processing and for building the classification systems are shown in Figure 4.
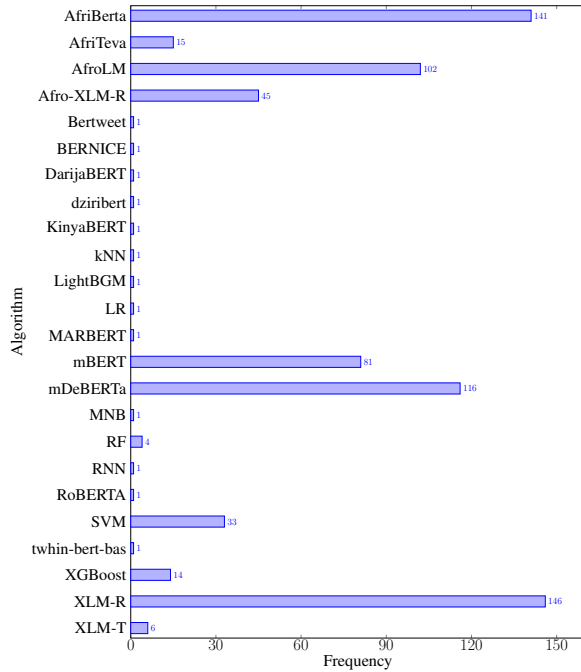
(a) Classification
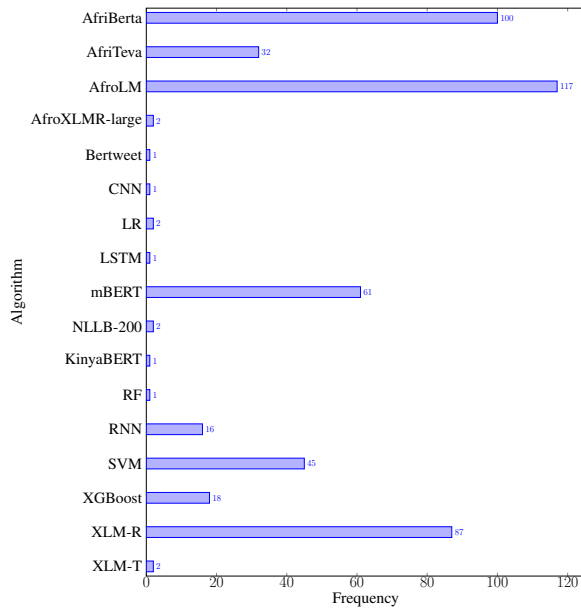
(b) Pre-processing

Figure 4: Algorithms used for pre-processing and final tweet classification.

B Tools Used

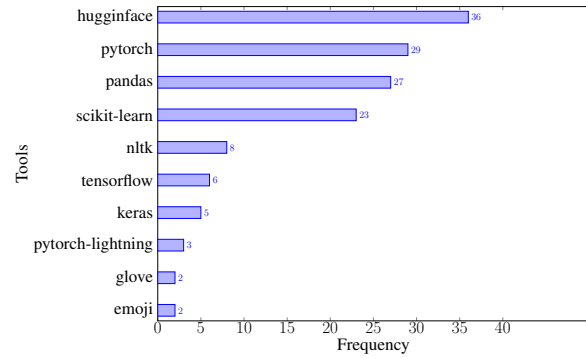Tools used by the participants to implement their systems are shown in Figure 5.

Figure 5: Tools Used

C Affiliation

Table 6 shows the participating teams, the tasks they made submissions for, their system description paper, and their affiliations.

Tasks	Team	System paper	Affiliation
A,B,C	ABCD Team	(Dang et al., 2023)	University of Information Technology - Ho Chi Minh City, Vietnam; National University, Ho Chi Minh City, Vietnam; Nong Lam University - Ho Chi Minh City, Vietnam
B,C	BERT 4EVER	—	Guangdong University of Technology; Guangdong University of Foreign Studies
A,C	Bhattacharya_Lab	(Hughes et al., 2023)	Auburn University at Montgomery; Florida Polytechnic University
A,B,C	DN	(Homskiy and Maloyan, 2023)	none
A,B	DuluthNLP	(Akrah and Pedersen, 2023)	University of Minnesota, Duluth, USA
A,B,C	FIT BUT	(Aparovich et al., 2023)	Brno Technical University, Czech Republic
A	Foul	(BELBACHIR, 2023)	IPSA - Ecole d'ingenieurs Aéronautique et Spatiale Paris
A,B,C	FUOYENLP	—	Federal University Oye-Ekiti Ekiti, Nigeria
A,B,C	GMNLP	(Alam et al., 2023)	George Mason University
A,B,C	GunadarmaxBRIN	(Arlim et al., 2023)	Gunadarma University, Indonesia; National Research and Innovation Agency, Indonesia
A,B,C	HausaNLP	(Abdullahi et al., 2023)	Kaduna State University; University of Bucharest; University of Abuja; Nile University; Universiti Teknologi PETRONAS; Bayero University Kano; Shehu Shagari College of Education; Ahmadu Bello University; Masakhane; HausaNLP
A,B	Howard University CS	(Aryal and Prioleau, 2023)	Howard University
B	iREL	—	IIIT-H
B	ISCL_WINTER	(Hancharova et al., 2023)	Eberhard Karls Universität Tübingen
A,B	JCT	(Keinan and HaCohen-Kerner, 2023)	Jerusalem College of Technology
A	KINLP	(Nzeyimana, 2023)	University of Massachusetts Amherst
A,B	king001	—	PingAnLifeInsurance
A,C	MaChAmp	(van der Goot, 2023)	none
A,B,C	Masakhane-Afrisenti	(Azime et al., 2023)	Saarland Universty; Luleå University of Technology; Instituto Politécnico Nacional; Masakhane; Montclair State University; University of Ibadan; Iowa State University
A,B,C	mitchelldehaven	—	Information Sciences Institute (ISI)
A,B,C	NLNDE	(Wang et al., 2023)	Bosch Center for Artificial Intelligence, Renningen, Germany; Center for Information and Language Processing (CIS), LMU Munich, Germany
A,B,C	NLP-LISAC	(Benlahbib and Boumhidi, 2023)	LISAC Faculty of Sciences Dhar EL Mehraz (F.S.D.M), Fez, Morocco
C	PA2023	—	PingAn
A,B,C	PA14	—	PA Sales Support Development Co., Ltd.
A,B,C	PingAnLifeInsurance	(Jin et al., 2023)	Ping An Life Insurance Company of China, Ltd.
B,C	saroyehun	—	University of Konstanz
A	Seals_Lab	(Raychawdhary et al., 2023)	Auburn University
A,B,C	Sefamerve	(Delil and Kuyumcu, 2023)	Sefamerve R&D Center
C	Snarci	—	University of Cagliari
A	TechSSN	(Sivanaiah et al., 2023)	Sri Sivasubramaniya Nadar College of Engineering, Chennai, India
A,B,C	tmn	—	University of Tyumen, Russia
A	Trinity	(Rathi et al., 2023)	Pune Institute Of Computer Technology
A,C	UBC-DLNLP	(Bhatia et al., 2023)	The University of British Columbia; Mohamed Bin Zayed University of Artificial Intelligence
A,B,C	UCAS-IIE-NLP	(Hu et al., 2023)	Chinese Academy of Sciences; University of Chinese Academy of Sciences
A,B	UIO	(Rønningstad, 2023)	University of Oslo
A,B,C	UM6P	(El Mahdaouy et al., 2023)	Mohammed VI Polytechnic University, Morocco; University of Luxembourg, Luxembourg
A,B	UMUTeam	(García-Díaz et al., 2023)	University of Murcia; University of Castilla-La Mancha
A	Uppsala University	(Kniele and Beloucif, 2023)	Uppsala University
A,B	Witcherses	(Gokani et al., 2023)	Language Technologies Research Center (LTRC); Kohli Center on Intelligent Systems; International Institute of Information Technology, Hyderabad
A,B,C	ymf924	—	The Ohio State University

Table 6: Participating teams, their affiliations and the tasks they participated in.