AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages

Shamsuddeen Hassan Muhammad[1,2+], **Idris Abdulmumin**[3+], **Abinew Ali Ayele**[4], **Nedjma Ousidhoum**[5], **David Ifeoluwa Adelani**[6*], **Seid Muhie Yimam**[7], **Ibrahim Sa'id Ahmad**[2+], **Meriem Beloucif** [8], **Saif M. Mohammad**[9], **Sebastian Ruder**[10], **Oumaima Hourrane**[11], **Pavel Brazdil**[1], **Felermino Dário Mário António Ali**[1], **Davis David**[12], **Salomey Osei**[13], **Bello Shehu Bello**[2], **Falalu Ibrahim**[14], **Tajuddeen Gwadabe**[*+], **Samuel Rutunda**[15], **Tadesse Belay**[16], **Wendimu Baye Messelle**[4], **Hailu Beshada Balcha**[17], **Sisay Adugna Chala**[18], **Hagos Tesfahun Gebremichael**[4], **Bernard Opoku**[19], **Steven Arthur**[19]

[1]University of Porto, [2]Bayero University Kano, [3]Ahmadu Bello University, Zaria, [4]Bahir Dar University, [5]University of Cambridge, [6]University College London, [7]Universität Hamburg, [8]Uppsala University, [9]National Research Council Canada, [10]Google Research, [11]Hassan II University of Casablanca, [12]dLab, [13]University of Deusto, [14]Kaduna State University, [15]Digital Umuganda, [16]Wollo University, [17]Jimma University, [18]Fraunhofer FIT, [19]Accra Institute of Technology, [*]Masakhane NLP, [+]HausaNLP

shmuhammad.csc@buk.edu.ng

Abstract

Africa is home to over 2000 languages from over six language families and has the highest linguistic diversity among all continents. This includes 75 languages with at least one million speakers each. Yet, there is little NLP research conducted on African languages. Crucial in enabling such research is the availability of high-quality annotated datasets. In this paper, we introduce AfriSenti, which consists of 14 sentiment datasets of 110,000+ tweets in 14 African languages (Amharic, Algerian Arabic, Hausa, Igbo, Kinyarwanda, Moroccan Arabic, Mozambican Portuguese, Nigerian Pidgin, Oromo, Swahili, Tigrinya, Twi, Xitsonga, and Yorùbá) from four language families annotated by native speakers. The data is used in SemEval 2023 Task 12, the first Afro-centric SemEval shared task. We describe the data collection methodology, annotation process, and related challenges when curating each of the datasets. We conduct experiments with different sentiment classification baselines and discuss their usefulness. We hope AfriSenti enables new work on under-represented languages.[1]

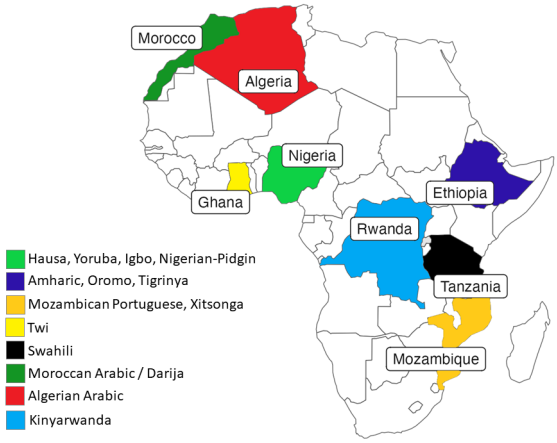

Figure 1: Countries and languages represented in the AfriSenti data collection (Amharic, Algerian Arabic, Hausa, Igbo, Kinyarwanda, Moroccan Arabic, Mozambican Portuguese, Nigerian Pidgin, Oromo, Swahili, Tigrinya, Twi, Xitsonga, and Yorùbá).

1 Introduction

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) while recent years have seen a proliferation of communication in digital and social media. Code-switching is common in these new forms of communication where speakers alternate between two or more languages in the context of a single conversation (Santy et al., 2021; Angel et al., 2020; Thara and Poornachandran, 2018). However, despite this linguistic richness, African languages have been comparatively under-represented in natural language processing (NLP) research.

An influential sub-area of NLP deals with sentiment, valence, emotions, and affect in language (Liu, 2020). Computational analysis of emotion states in language and the creation of systems that predict these states from utterances have applications in literary analysis and culturonomics (Rea-

[1]The dataset is available at https://github.com/afrisenti-semeval/afrisent-semeval-2023.

Lang.	Tweet	Sentiment
amh	ያ ጨካኝ አረመኔ ታስሮ ይኸው ካቴና ጉበቱለታል ይሉናል። ቆይ አስረው የጀበና ቡና እየጋበዙት ነው እንዴ?	negative
arq	@user .... الشروق هذه من خرجت وهي تتاع تهديل، مستوى منحط وشعبوي	negative
ary	واش بغيتوهم يبداو يتكرفسو على العادي والبادي عاد تبقاو أنتما على خاطر خاطركم	negative
ary	rabi ykhali alhbiba makayn ghir nachat o chi machat	positive
hau	@USER Aunt rahma i luv u wallah irin totally dinnan	positive
ibo	akowaro ya ofuma nne kai daalu nwanne mmadu	positive
kin	@user Ariko akokanu ngo inyebebe unyujijemo sisawa wangu	negative
orm	@user Jawaar Kenya OMN haala akkamiin argachuu dandeenya	neutral
por	Honestidade é algo que não se compra. Infelizmente a humanidade esqueceu disso por causa das suas ambições.	positive
pcm	E don tay wey I don dey crush on this fine woman ...	positive
swa	Asante sana watu wa Sirari jimbo la Tarime vijijini Huu ni Upendo usio na Mashaka kwa Mbunge wenu John Heche	positive
tir	@user ክመኽረኩም እንተኾይነ:ንሕውሓት ነዞም ውሑድ ቁጽሮም እባ ምጥፋእ ይሕሸኩም!	negative
tso	@user @user Yu , tindzava ? Tsika mbangui mpfana e nita ku desprogramara	negative
twi	messi saf den check en bp na wo kwame danso wo di twe da kor aaa na wawu	negative
yor	onírèégbè aláàdúgbò ati olójúkòkòrò	negative

Table 1: Examples of tweets and their sentiments in the different AfriSenti Languages. Note that the collected tweets in Moroccan Darija (ary) are written in both Arabic and Latin scripts.

gan et al., 2016; Hamilton et al., 2016), commerce (e.g., tracking feelings towards products), and research in psychology and social science (Dodds et al., 2015; Hamilton et al., 2016). Despite tremendous amount of work in this important space over the last two decades, there is little work on African languages, partially due to a lack of high-quality annotated data.

To enable sentiment analysis research in African languages, we present **AfriSenti**, the largest sentiment analysis benchmark for under-represented languages—covering 110,000+ annotated tweets in 14 African languages[2] (Amharic, Algerian Arabic, Hausa, Igbo, Kinyarwanda, Moroccan Arabic, Mozambican Portuguese, Nigerian Pidgin, Oromo, Swahili, Tigrinya, Twi, Xitsonga, and Yorùbá) from four language families (Afro-Asiatic, English Creole, Indo-European and Niger-Congo). We show the represented countries and languages in Figure 1 and provide examples of the data in Table 1. AfriSenti is an extension of NaijaSenti (Muhammad et al., 2022), which is a sentiment corpus in four major Nigerian languages: Hausa, Igbo, Nigerian Pidgin,and Yorùbá.

[2]For simplicity, we use the term language to refer to language varieties including dialects.

The datasets are used in the first Afrocentric SemEval shared task, *SemEval 2023 Task 12: Sentiment analysis for African languages (AfriSenti-SemEval)*. AfriSenti allows the research community to build sentiment analysis systems for various African languages and enables the study of sentiment and contemporary language use in African languages. We publicly release the corpora, which provide further opportunities to investigate the difficulty of sentiment analysis for African languages.

Our contributions are: (1) the creation of the largest Twitter dataset for sentiment analysis in African languages by annotating 10 new datasets and curating four existing ones (Muhammad et al., 2022), (2) the discussion of the data collection and annotation process in 14 low-resource African languages, (3) the release sentiment lexicons for all languages, (4) the presentation of classification baseline results using our datasets.

2 Related Work

Research in sentiment analysis developed since the early days of lexicon-based sentiment analysis approaches (Turney, 2002; Taboada et al., 2011; Mohammad et al., 2013) to more advanced machine learning (Agarwal and Mittal, 2016; Le

Language	ISO Code	Subregion	Spoken In	Script
Amharic	amh	East Africa	Ethiopia	Ethiopic
Algerian Arabic/Darja	arq	North Africa	Algeria	Arabic
Hausa	hau	West Africa	Northern Nigeria, Ghana, Cameroon,	Latin
Igbo	ibo	West Africa	Southeastern Nigeria	Latin
Kinyarwanda	kin	East Africa	Rwanda	Latin
Moroccan Arabic/Darija	ary	Northern Africa	Morocco	Arabic/Latin
Mozambican Portuguese	pt-MZ	Southeastern Africa	Mozambique	Latin
Nigerian Pidgin	pcm	West Africa	Northern Nigeria, Ghana, Cameroon,	Latin
Oromo	orm	East Africa	Ethiopia	Latin
Swahili	swa	East Africa	Tanzania, Kenya, Mozambique	Latin
Tigrinya	tir	East Africa	Ethiopia	Ethiopic
Twi	twi	West Africa	Ghana	Latin
Xitsonga	tso	Southern Africa	Mozambique, South Africa, Zimbabwe, Eswatini	Latin
Yorùbá	yor	West Africa	Southwestern and Central Nigeria	Latin

Table 2: African languages included in our study (Lewis, 2009).

and Nguyen, 2020), deep learning-based methods (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, pretrained language models (PLMs), such as XLM-R (Conneau et al., 2020), mDeBERTaV3 (He et al., 2021), AfriBERTa (Ogueji et al., 2021b), AfroXLMR (Alabi et al., 2022) and XLM-T (Barbieri et al., 2022b) provide state-of-the-art performance for sentiment classification.

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 (neuro-symbolic) (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 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 under-represented languages still presents a challenge (Yimam et al., 2020). Additionally, African languages present new challenges for sentiment analysis such as dealing with tone, code-switching, and digraphia (Adebara and Abdul-Mageed, 2022). Existing work in sentiment analysis for African languages has therefore mainly focused on polarity classification (Mataoui et al., 2016; El Abdouli et al., 2017; Moudjari et al., 2020; Yimam et al., 2020; Muhammad et al., 2022; Martin et al., 2021). We present with AfriSenti the largest and most multilingual dataset for sentiment analysis in African languages.

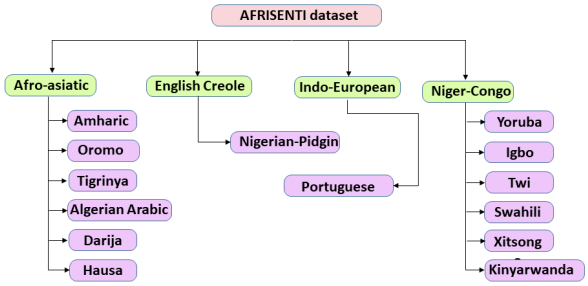

Figure 2: Language Family (shown in green) in the AfriSenti datasets.

3 Overview of the AfriSenti Datasets

AfriSenti covers 14 African languages, each with unique linguistic characteristics and writing systems, which are shown in Table 2. As shown in Figure 2, the dataset includes six languages of the Afroasiatic family, six languages of the Niger-Congo family, one from the English Creole family, and one from the Indo-European family.

Writing Systems Scripts serve not only as a means of transcribing spoken language, but also as powerful cultural symbols that reflect people's identity (Sterponi and Lai, 2014). For instance, the Bamun script is deeply connected to the identity of Bamun speakers in Cameroon, while the Geez/Ethiopic script (for Amharic and Tigrinya) evokes the strength and significance of Ethiopian

culture (Sterponi and Lai, 2014). Similarly, the Ajami script, a variant of the Arabic script used in various African languages such as Hausa, serves as a reminder of the rich African cultural heritage of the Hausa community (Gee, 2005).

African languages, with a few exceptions, use the Latin script, written from left to right, or the Arabic script, written from right to left (Gee, 2005; Meshesha and Jawahar, 2008), with the Latin script being the most widely used in Africa (Eberhard et al., 2020). Ten languages out of fourteen in AfriSenti are written in Latin script, two in Arabic script, and two in Ethiopic (or Geez) script. On social media, people may write Moroccan Arabic (Darija) and Algerian Arabic (Darja) in both Latin and Arabic characters due to various reasons including access to technology, i.e., the fact that Arabic keyboards were not easily accessible on commonly used devices for many years, code-switching, and other phenomena. This makes Algerian and Moroccan Arabic digraphic, i.e., their texts can be written in multiple scripts on social media. Similarly, Amharic is digraphic and is written in both Latin and Geez script (Belay et al., 2021). This constitutes an additional challenge to the processing of these languages in NLP.[3]

Geographic Representation AfriSenti covers the majority of African sub-regions. Many African languages are spoken in neighbouring countries within the same sub-regions. For instance, variations of Hausa are spoken in Nigeria, Ghana, and Cameroon, while Swahili variants are widely spoken in East African countries, including Kenya, Tanzania, and Uganda. AfriSenti also includes datasets in the top three languages with the highest numbers of speakers in Africa (Swahili, Amharic, and Hausa). We show the geographic distribution of languages in AfriSenti in Figure 1.

New and Existing Datasets AfriSenti includes existing and newly created datasets as shown in Table 3. For the existing datasets whose test sets are public, we created new test sets to further evaluate their performance in the AfriSenti-SemEval shared task.

[3]Table 1 shows an example of Moroccan Darija tweets written in Latin and Arabic script. For Algerian Arabic/Darja and Amharic, AfriSenti includes data in only Arabic and Geez scripts.

Lang.	New	Existing	Source
ama	test	train, dev	Yimam et al. (2020)
arq	all	✗	-
ary	all	✗	-
hau	✗	all	Muhammad et al. (2022)
ibo	✗	all	Muhammad et al. (2022)
kin	all	✗	-
orm	all	✗	-
pcm	✗	all	Muhammad et al. (2022)
pt-MZ	all	✗	-
swa	all	✗	-
tir	all	✗	-
tso	all	✗	-
twi	all	✗	-
yor	✗	all	Muhammad et al. (2022)

Table 3: The AfriSenti datasets. We show the new and previously available datasets (with their sources).

4 Data Collection and Processing

Twitter's Limited Support for African Languages Since many people share their opinions on Twitter, the platform is widely used to study sentiment analysis (Muhammad et al., 2022). However, the Twitter API's support for African languages is limited, which makes it difficult for researchers to collect data. Specifically, the Twitter language API currently supports only Amharic out of more than 2000 African languages[4]. This disparity in language coverage highlights the need for further research and development in NLP for low-resource languages.

4.1 Tweet Collection

We used the Twitter Academic API to collect tweets. However, as the API does not provide language identification for tweets in African languages, we use location-based and vocabulary-based heuristics to collect the datasets.

4.1.1 Location-based data collection

For all languages except Algerian Arabic and Afan Oromo, we used a location-based collection approach to filter out results. Hence, tweets were collected based on the names of the countries where the majority of the target language speakers are located. For Afaan Oromo, tweets were collected globally due to the small size of the data collected from Ethiopia.

[4]https://blog.twitter.com/engineering/en_us/a/2015/evaluating-language-identification-performance

Lang.	Manually	Translated	Source
ama	✓	✗	Yimam et al. (2020)
arq	✓	✗	-
hau	✓	✓	Muhammad et al. (2022)
ibo	✓	✓	Muhammad et al. (2022)
ary	✗	✗	-
orm	✓	✗	Yimam et al. (2020)
pcm	✓	✗	Muhammad et al. (2022)
pt-MZ	✓	✗	-
kin	✗	✓	-
swa	✗	✗	-
tir	✓	✗	Yimam et al. (2020)
tso	✗	✓	-
twi	✗	✗	-
yor	✓	✓	Muhammad et al. (2022)

Table 4: Manually collected and translated lexicons in AfriSenti.

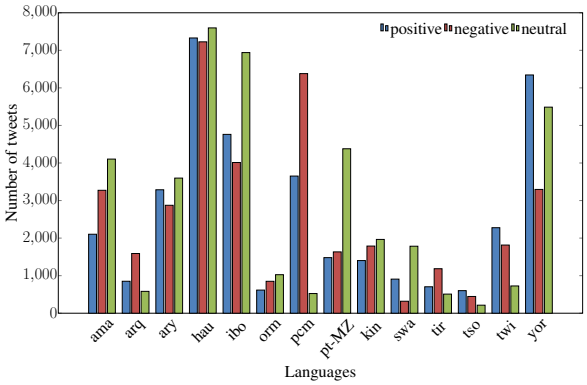

Figure 3: Label distributions for the different AfriSenti datasets.

4.1.2 Vocabulary-based Data Collection

As different languages are spoken within the same region in Africa (Amfo and Anderson, 2019), the location-based approach did not help in all cases. For instance, searching for tweets from *"Lagos"* (Nigeria) returned tweets in multiple languages, such as *Yorùbá*, *Igbo*, *Hausa*, *Pidgin*, *English*, etc.

To address these challenges, we combined the location-based approach with vocabulary-based collection strategies. These included the use of stopwords, sentiment lexicons, and a language detection tool. For languages that used the Geez script, we used the Ethiopic Twitter Dataset for Amharic (ETD-AM), which includes tweets that were collected since 2014 (Yimam et al., 2019).

Data collection using stopwords Most African languages do not have curated stopword lists (Emezue et al., 2022). Therefore, we created stopword lists for some AfriSenti languages and used them to collect data. We used corpora from different domains, i.e. news data and religious texts, to rank words based on their frequency (Adelani et al., 2021). We filtered out the top 100 words by deleting domain-specific words (e.g., the word *God* in religious texts) and created lists based on the top 50 words that appeared across domains.

We also used a word co-occurrence-based approach to extract stopwords (Liang et al., 2009) using text sources from different domains. We lower-cased and removed punctuation symbols and numbers, constructed a co-occurrence graph, and filtered out the words that occurred most often. Native speakers verified the generated lists before use. This approach worked the best for Xistonga.

Data collection using sentiment lexicons As data collection based on stopwords sometimes results in tweets that are inadequate for sentiment analysis (e.g., too many neutral tweets), we used a sentiment lexicon—a dictionary of positive and negative words—for tweet collection. This allows for a balanced collection across sentiment classes (positive/negative/neutral). For Moroccan Darija, we used emotion word list curated by Outchakoucht and Es-Samaali (2021).

Table 4 provides details on the sentiment lexicons in AfriSenti and indicates whether they were manually created or translated.

Data collection using mixed lists of words Besides stopwords and sentiment lexicons, native speakers provided lists of language-specific terms including generic words. For instance, this strategy helped us collect Algerian Arabic tweets, and the generic terms included equivalents of words such as "الغاشي" *(the crowd)* and names of Algerian cities.

4.2 Language Detection

As we mainly used heuristics for data collection, the result included tweets in a language that is different from the target one. For instance, when collecting tweets using lists of Amharic words, some returned tweets were in Tigrinya, due to Amharic–Tigrinya code-mixing. Similarly, we applied an additional manual filtering step in the case of Tunisian, Moroccan, and Modern Standard Arabic tweets that were returned when searching for Algerian Arabic ones due to overlapping terms.

Hence, we used different techniques for language detection as a post-processing step.

Language detection using existing tools Few African languages have pre-existing language de-

tection tools (Keet, 2021). We used Google CLD3[5] and the Pycld2 library[6] for the supported AfriSenti languages (Amharic, Oromo and Tigrinya).

Manual language detection For languages that do not have a pre-existing tool, the detection was conducted by native speakers. For instance, annotators who are native speakers of Twi and Xitsonga manually labeled 2,000 tweets in these languages. In addition, as native speakers collected the Algerian Arabic tweets, they deleted all possible tweets that were expressed in another language or Arabic variation instead.

Language detection using pre-trained language models To reduce the effort spent on language detection, we also used a pretrained language model fine-tuned on 2,000 manually annotated tweets (Caswell et al., 2020) to identify Twi and Xitsonga.

Despite our efforts to detect the right languages, it is worth mentioning that as multilingualism is common in African societies, the final dataset contains many code-mixed tweets.

4.3 Tweet Anonymization and Pre-processing

We anonymized the tweets by replacing all *@mentions* by *@user* and removed all URLs. For the Nigerian language test sets, we further lower-cased the tweets (Muhammad et al., 2022).

5 Data Annotation Challenges

Tweet samples were randomly selected based on the different collection strategies. Then, with the exception of the Ethiopian languages, each tweet was annotated by three native speakers. We followed the sentiment annotation guidelines by Mohammad (2016) and used majority voting (Davani et al., 2021) to determine the final sentiment label for each tweet (Muhammad et al., 2022). We discarded the cases where all annotators disagree. The datasets of the three Ethiopian languages (Amharic, Tigriniya, and Oromo) were annotated using two independent annotators, and then curated by a third more experienced individual who decided on the final gold labels.

Prabhakaran et al. (2021) showed that the majority vote conceals systematic disagreements between annotators resulting from their sociocultural backgrounds and experiences. Therefore, we release all the individual labels to the research community. We report the free marginal multi-rater kappa scores (Randolph, 2005) in Table 5 since chance-adjusted scores such as Fleiss-κ can be low despite a high agreement due to the imbalanced label distributions (Randolph, 2005; Falotico and Quatto, 2015; Matheson, 2019). We obtained intermediate to good levels of agreement ($0.40 - 0.75$) across all languages, except for Oromo where we obtained a low agreement score due the annotation challenges that we discuss in Section 5.

Table 6 shows the number of tweets in each of the 14 datasets. The Hausa collection of tweets is the largest AfriSenti dataset and the Xitsonga dataset is the smallest one. Figure 3 shows the distribution of the labeled classes in the datasets. We observe that the distribution for some languages such as `ha` is fairly equitable while in others such as `pcm`, the proportion of tweets in each class varies widely. Sentiment annotation for African languages presents some challenges (Muhammad et al., 2022) that we highlight in the following.

Twi A significant portion of tweets in *Twi* were ambiguous, making it difficult to accurately categorize sentiment. Some tweets contained symbols that are not in the Twi alphabet, which is a frequent occurrence due to the lack of support for certain Twi letters on keyboards (Scannell, 2011). For example, "ɔ" is replaced by the English letter "c", and "ɛ" is replaced by "3".

Additionally, tweets are more often annotated as negative (cf. Figure 3). This is due to some common expressions that can be seen as offensive depending on the context. For instance, "*Tweaa*" was once considered an insult but has become a playful expression through trolling, and "*gyae gyimii*" is commonly used by young people to say "stop" while its literal meaning is "stop fooling".

Mozambican Portuguese and Xitsonga One of the significant challenges for the Mozambican Portuguese and Xitsonga data annotators was the presence of code-mixed and sarcastic tweets. Code-mixing in tweets made it challenging for the annotators to determine the intended meaning of the tweet as it involved multiple languages spoken in Mozambique that some annotators did not understand. Similarly, the presence of two variants of Xitsonga spoken in Mozambique (Changana and Ronga) added to the complexity of the annotation

[5] https://github.com/google/cld3
[6] https://pypi.org/project/pycld2/

	3-way										2-way		
Lang.	arq	ary	hau	ibo	kin	pcm	pt-MZ	swa	tso	twi	yor	ama	orm	tir
κ	0.41	0.62	0.66	0.61	0.43	0.60	0.50	–	0.50	0.51	0.65	0.47	0.20	0.51

Table 5: Inter-annotator agreement scores using the free marginal multi-rater kappa (Randolph, 2005) for the different languages.

	ama	arq	hau	ibo	ary	orm	pcm	pt-MZ	kin	swa	tir	tso	twi	yor
train	5,985	1,652	14,173	10,193	5,584	-	5,122	3,064	3303	1,811	-	805	3,482	8,523
dev	1,498	415	2,678	1,842	1,216	397	1,282	768	828	454	399	204	389	2,091
test	2,000	959	5,304	3,683	2,962	2,097	4,155	3,663	1027	749	2,001	255	950	4,516
Total	9,483	3,062	22,155	15,718	9,762	2,494	10,559	7,495	5,158	3,014	2,400	1,264	4,821	15,130

Table 6: Sizes and splits of the AfriSenti datasets. We do not allocate training splits for Oromo (orm) and Tigrinya (tir) due to the limited size of the data and only evaluate on them in a zero-shot transfer setting in §6.

task. Additionally, sarcasm was a source of disagreement among annotators, leading to the exclusion of many tweets from the final dataset.

Ethiopian languages For Oromo and Tigrinya, challenges included finding annotators and the lack of a reliable Internet connection and access to personal computers. Although we trained the Oromo annotators, we observed severe problems in the quality of the annotated data which led to a low agreement score.

Algerian Arabic For Algerian Arabic, the main challenge was the use of sarcasm. When this caused a disagreement among the annotators, the tweet was further labeled by two additional annotators. If all the annotators did not agree on one final label, we discarded it. As Twitter is also commonly used to discuss controversial topics in the region, we removed offensive tweets.

6 Experiments

6.1 Setup

For our baseline experiments, we considered three settings: (1) monolingual baseline models based on multilingual pre-trained language models for 12 AfriSenti languages with training data, (2) multilingual training of all 12 languages, and their evaluation on a combined test of all 12 languages, (3) Zero-shot transfer to Oromo (orm) and Tigrinya (tir) from any of the 12 languages with available training data.

Monolingual baseline models We *fine-tune* massively multilingual pre-trained language models (PLMs) trained on 100 languages from around the world and Africa-centric PLMs trained exclusively on languages spoken in Africa. For the massively multilingual PLMs, we selected two representative PLMs: XLM-R-{base & large} (Conneau et al., 2020) and mDeBERTaV3 (He et al., 2021). For the Africa-centric models, we make use of AfriBERTa-large (Ogueji et al., 2021a) and AfroXLMR-{base & large} (Alabi et al., 2022) — an XLM-R model adapted to African languages. AfriBERTa was pre-trained from the scratch on 11 African languages including nine of the AfriSenti languages while AfroXLMR supports 10 of the AfriSenti languages. Additionally, we fine-tune XLM-T (Barbieri et al., 2022a), an XLM-R model adapted to the multilingual Twitter domain, supporting over 30 languages but fewer African languages due to a lack of coverage by Twitter's language API (cf. §4).

6.2 Experimental Results

Table 7 shows the results of the monolingual baseline models on AfriSenti. AfriBERTa obtained the worst performance on average (61.7) especially for languages it was not pre-trained on, e.g., < 50 for the Arabic dialects. However, it achieved a good results for languages it has been pre-trained on, such as hau, ibo, swa, yor. XLM-R-base led to a performance that is comparable to AfriBERTa on average, was worse for most African languages, but better for Arabic dialects and pt-MZ. On the other hand, AfroXLMR-base and mDeBERTaV3 achieve similar performance, although AfroXLMR-base performs slightly better for kin and pcm compared to other models. Overall, considering models with up to 270M parameters, XLM-T achieves

Lang.	In XLM-R or mDeBERTa?	In AfriBERTa	In AfroXLMR	In XLM-T	AfriBERTa large	XLM-R base	AfroXLMR base	mDeBERTa base	XLM-T base	XLM-R large	AfroXLMR large
amh	✓	✓	✓	✓	56.9	60.2	54.9	57.6	60.8	**61.8**	61.6
arq	✓	✗	✓	✓	47.7	65.9	65.5	65.7	**69.5**	63.9	68.3
ary	✓	✗	✓	✓	44.1	50.9	52.4	55.0	**58.3**	57.7	56.6
hau	✓	✓	✓	✗	78.7	73.2	77.2	75.7	73.3	75.7	**80.7**
ibo	✗	✓	✓	✗	78.6	75.6	76.3	77.5	76.1	76.5	**79.5**
kin	✗	✓	✓	✗	62.7	56.7	67.2	65.5	59.0	55.7	**70.6**
pcm	✗	✓	✓	✗	62.3	63.8	67.6	66.2	66.6	67.2	**68.7**
pt-MZ	✓	✗	✗	✓	58.3	70.1	66.6	68.6	71.3	71.6	**71.6**
swa	✓	✓	✓	✗	61.5	57.8	60.8	59.5	58.4	61.4	**63.4**
tso	✗	✗	✗	✗	51.6	47.4	45.9	47.4	**53.8**	43.7	47.3
twi	✗	✗	✗	✗	**65.2**	61.4	62.6	63.8	65.1	59.9	64.3
yor	✗	✓	✓	✗	72.9	62.7	70.0	68.4	64.2	62.4	**74.1**
AVG	-	-	-	-	61.7	61.9	63.9	64.2	64.7	63.1	**67.2**

Table 7: Accuracy scores of monolingual baselines for AfriSenti on the 12 languages with training splits. Results are averaged over 5 runs.

Model	F1
AfriBERTa-large	64.7
XLM-R-base	64.3
AfroXLMR-base	68.4
mDeBERTaV3-base	66.1
XLM-T-base	65.9
XLM-R-large	66.9
AfroXLMR-large	**71.2**

Table 8: Multilingual training and evaluation on combined test sets of all languages. Average over 5 runs.

the best performance, which highlights the importance of domain-specific pre-training. XLM-T performs particularly well on Arabic and Portuguese dialects, i.e., arq, ary and pt-MZ, where it outperforms AfriBERTa by 21.8, 14.2, and 13.0 and AfroXLMR-base by 4.0, 5.9, and 4.7 F1 points respectively. AfroXLMR-large achieves the best overall performance and improves over XLM-T by 2.5 F1 points, which highlights the benefit of scaling for large PLMs. Scaling is of limited use for XLM-R-large, however, as it has not been pre-trained on many of the African languages. Overall, our results demonstrate the importance of both language and domain-specific pre-training as well as the benefits of scale for appropriately pre-trained models.

Table 8 shows the performance of multilingual models that were fine-tuned on the combined training data and evaluated on the combined test data of all languages. Similar to before, AfroXLMR-large achieves the best performance, outperforming AfroXLMR-base, XLM-R-large, and XLM-T-base by more than 2.5 F1 points.

Finally, Table 9 shows the zero-shot cross-lingual transfer performance from models trained on different source languages with available training data to the test-only languages orm and tir. The best source languages are Hausa or Amharic for orm, and Hausa or Yorùbá for tir. Hausa even outperforms a multilingually trained model. The impressive performance for transfer between Hausa and Oromo may be because both are from the same language family and share a similar Latin script. In addition, Hausa has the largest training dataset in AfriSenti. Both linguistic similarity and size of source language data have been shown to correlate with successful cross-lingual transfer (Lin et al., 2019). However, it is unclear why Yorùbá performs particularly well for tir despite the difference in script. One hypothesis is that Yorùbá may be a good source language in general, as shown in Adelani et al. (2022) where Yorùbá is the second best source language for named entity recognition in African languages.

7 Conclusion and Future Work

We presented AfriSenti, a collection of sentiment Twitter datasets annotated by native speakers in 14 African languages used in the first Afro-centric SemEval shared task—SemEval 2023 Task 12: Sentiment analysis for African languages (AfriSenti-SemEval). We reported the challenges faced during data collection and annotation as well as experimental results using state-of-the-art pre-trained language models. We release the datasets, and data resources to the research community. AfriSenti opens up new avenues for sentiment analysis research in under-represented languages. In the future, we plan to extend *AfriSenti* to more African languages and different sentiment analysis sub-tasks.

Source Lang.	Target Lang. orm	tir	AVG
amh	46.5	62.6	54.6
arq	27.5	56.0	41.8
ary	42.5	58.6	50.6
hau	**47.1**	**68.6**	**57.9**
ibo	41.7	39.8	40.8
kin	43.6	64.8	54.2
pcm	26.7	58.2	42.5
por	28.7	21.5	25.1
swa	36.8	26.7	31.8
tso	21.5	15.8	18.7
twi	9.8	15.6	12.7
yor	39.2	67.1	53.2
multilingual	42.0	66.4	54.2

Table 9: Zero-shot evaluation on orm and tir. All SRC LANGs are trained on AfroXLMR-large

8 Ethics Statement

Sentiment and emotions are complex and nuanced mental states. Additionally, each individual expresses sentiment differently through language, which results in large amounts of variation. Therefore, several ethical considerations should be accounted for when working on sentiment analysis. See Mohammad (2022, 2023) for a comprehensive discussion of ethical considerations relevant to sentiment and emotion analysis.

Acknowledgements

We thank all the volunteer annotators involved in this project. Without their support and valuable contributions, this project would not have been possible. This research was partly funded by the Lacuna Fund, an initiative co-founded by The Rockefeller Foundation, Google.org, and Canada's International Development Research Centre. The views expressed herein do not necessarily represent those of Lacuna Fund, its Steering Committee, its funders, or Meridian Institute. We are grateful to Adnan Ozturel for helpful comments on a draft of this paper. We thank Tal Perry for providing the LightTag (Perry, 2021) annotation tool. We also thank the Language Technology Group, University of Hamburg, for allowing us to use the WebAnno (Yimam et al., 2013) annotation tool for all the Ethiopian languages annotation tasks. David Adelani acknowledges the support of DeepMind Academic Fellowship Programme. Finally, we are grateful for the support of Masakhane.

References

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

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. 2022. MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition. In *Proceedings of the 2022 Con-*

ference 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.

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.

Nana Aba Appiah Amfo and Jemima Anderson. 2019. Multilingualism and language policies in the african context: lessons from ghana.

Jason Angel, Segun Taofeek Aroyehun, Antonio Tamayo, and Alexander Gelbukh. 2020. NLP-CIC at SemEval-2020 task 9: Analysing sentiment in code-switching language using a simple deep-learning classifier. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 957–962, Barcelona (online). International Committee for Computational Linguistics.

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. 2022a. 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.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022b. 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.

Tadesse Destaw Belay, Abinew Ali Ayele, Getie Gelaye, Seid Muhie Yimam, and Chris Biemann. 2021. Impacts of homophone normalization on semantic models for amharic. In *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 101–106.

Soumia Bougrine, Aicha Chorana, Abdallah Lakhdari, and Hadda Cherroun. 2017. Toward a web-based speech corpus for algerian dialectal arabic varieties. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 138–146.

Erik Cambria, Qian Liu, Sergio Decherchi, Frank Xing, and Kenneth Kwok. 2022. Senticnet 7: a commonsense-based neurosymbolic ai framework for explainable sentiment analysis. *Proceedings of LREC 2022*.

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.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on 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.

Aida Mostafazadeh Davani, Mark D'iaz, and Vinodkumar Prabhakaran. 2021. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

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, et al. 2015. Human language reveals a universal positivity bias. *Proceedings of the national academy of sciences*, 112(8):2389–2394.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2020. Ethnologue: Languages of the World. Twenty-third edition. Dallas, Texas: SIL International. Url: `http://www.ethnologue.com`.

Abdeljalil El Abdouli, Larbi Hassouni, and Houda Anoun. 2017. Sentiment analysis of moroccan tweets using naive bayes algorithm. *International Journal of Computer Science and Information Security (IJCSIS)*, 15(12).

Abdou Elimam. 2009. Du punique au maghribi trajectoires d'une langue sémito-méditerranéenne. *Synergies Tunisie*, 1:25–38.

Chris Chinenye Emezue, Hellina Hailu Nigatu, Cynthia Thinwa, Helper Zhou, Shamsuddeen Hassan Muhammad, Lerato Louis, Idris Abdulmumin, Samuel Gbenga Oyerinde, Benjamin Ayoade Ajibade, Olanrewaju Samuel, et al. 2022. The african stopwords project: Curating stopwords for african languages. In *3rd Workshop on African Natural Language Processing*.

Rosa Falotico and Piero Quatto. 2015. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49(2):463–470.

Quintin Gee. 2005. Review of script displays of african languages by current software. *New Review of Hypermedia and Multimedia*, 11:247 – 255.

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*.

Salima Harrat, Karima Meftouh, Mourad Abbas, Walid-Khaled Hidouci, and Kamel Smaili. 2016. An algerian dialect: Study and resources. *International journal of advanced computer science and applications (IJACSA)*, 7(3):384–396.

Martin Haspelmath and Uri Tadmor. 2009. *Loanwords in the world's languages: a comparative handbook*. Walter de Gruyter.

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

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.

Dr Kaur et al. 2022. Incorporating sentimental analysis into development of a hybrid classification model: A comprehensive study. *International Journal of Health Sciences*, 6:1709–1720.

C Maria Keet. 2021. Natural language generation requirements for social robots in sub-saharan africa. In *2021 IST-Africa Conference (IST-Africa)*, pages 1–8. IEEE.

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.

Wei Liang, Yuming Shi, Chi K. Tse, Jing Liu, Yanli Wang, and Xunqiang Cui. 2009. Comparison of co-occurrence networks of the chinese and english languages. *Physica A: Statistical Mechanics and its Applications*, 388(23):4901–4909.

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

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

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

Gati L Martin, Medard E Mswahili, and Young-Seob Jeong. 2021. Sentiment classification in swahili language using multilingual bert. *arXiv preprint arXiv:2104.09006*.

M'hamed Mataoui, Omar Zelmati, and Madiha Boumechache. 2016. A proposed lexicon-based sentiment analysis approach for the vernacular algerian arabic. *Research in Computing Science*, 110(1):55–70.

Granville J Matheson. 2019. We need to talk about reliability: making better use of test-retest studies for study design and interpretation. *PeerJ*, 7:e6918.

Million Meshesha and C. V. Jawahar. 2008. Indigenous scripts of african languages. *Indilinga: African Journal of Indigenous Knowledge Systems*, 6:132–142.

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, San Diego, California. Association for Computational Linguistics.

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.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiao-Dan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *International Workshop on Semantic Evaluation*.

Leila Moudjari, Karima Akli-Astouati, and Farah Benamara. 2020. An Algerian corpus and an annotation platform for opinion and emotion analysis. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1202–1210, Marseille, France. European Language Resources Association.

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.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021a. 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.

Kelechi Ogueji, Yuxin Zhu, and Jimmy J. Lin. 2021b. 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*.

Aissam Outchakoucht and Hamza Es-Samaali. 2021. Moroccan dialect -darija- open dataset.

Tal Perry. 2021. LightTag: Text annotation platform. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 20–27, Online and Punta Cana, Dominican Republic. 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*.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark D'iaz. 2021. On releasing annotator-level labels and information in datasets. *ArXiv*, abs/2110.05699.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Justus J Randolph. 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*.

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.

Sebastin Santy, Anirudh Srinivasan, and Monojit Choudhury. 2021. BERTologiCoMix: How does code-mixing interact with multilingual BERT? In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 111–121, Kyiv, Ukraine. Association for Computational Linguistics.

Kevin P Scannell. 2011. Statistical unicodification of african languages. *Language resources and evaluation*, 45(3):375–386.

Laura Sterponi and Paul F. Lai. 2014. Culture and language development. In Farzad Sharifian, editor, *The Routledge Handbook of Language and Culture*, pages 339–356. Routledge, London, UK.

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

S Thara and Prabaharan Poornachandran. 2018. Code-mixing: A brief survey. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2382–2388.

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*.

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*.

Seid Muhie Yimam, Abinew Ali Ayele, and Chris Biemann. 2019. Analysis of the Ethiopic Twitter Dataset for Abusive Speech in Amharic. In *In Proceedings of International Conference On Language Technologies For All: Enabling Linguistic Diversity And Multilingualism Worldwide (LT4ALL 2019)*, pages 1–5, Paris, France.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. WebAnno: A flexible, web-based and visually supported system for distributed annotations. In

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 1–6, Sofia, Bulgaria. Association for 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.

A Focus Languages

Afaan Oromo Afaan Oromo is spoken by more than 37 million speakers and is written in the Latin script (Eberhard et al., 2020). It is the most widely spoken language in Ethiopia and the third most widely spoken language in Africa next to Arabic and Hausa languages. In the Horn of Africa including Ethiopia, Kenya, and Somalia alone, there are over 45 million native Afaan Oromo speakers.

Algerian Arabic/ Darja Algerian Arabic/Darja is the Arabic "dialect" spoken in Algeria. It varies across the Algerian region (Bougrine et al., 2017) and is mastered by almost all Algerians (more than 40 million people). It has mostly an Arabic vocabulary but it also contains Berber (Amazigh), French, Andalusian Arabic, Turkish and Spanish influences and loanwords (Elimam, 2009; Haspelmath and Tadmor, 2009; Harrat et al., 2016).

Amharic Amharic is an Ethio-Semitic and Afro-Asiatic language. It is spoken in Ethiopia, Israel, and America (Eberhard et al., 2020). It has about 57 million speakers, where 32 million of them are native speakers and uses Ge'ez or Fidel script for writing.

Kinyarwanda Kinyarwanda is a language spoken in Central and East Africa, it is the official language of Rwanda but is also spoken in Uganda, D.R.C, Burundi and Tanzania, it is spoken by over 13 million people. It is one of the major Bantu languages, and it is mutually intelligible with Kirundi. Kinyarwanda uses the Latin Alphabet, composed of 24 letters used in English excluding x and q.

Moroccan Darija Moroccan Arabic Darija is the dialect of Arabic spoken in Morocco. It is a mixture of classical Arabic, Berber, and French with some Spanish and Portuguese influences. According to the 2014 general census [7], 92% of the Moroccan population speak Arabic Darija. This dialect retains many of the characteristics that make it unique among other dialects. Its phonology and syntax are quite different from other forms of spoken Arabic. However, Darija is not widely used outside Morocco and, therefore, may be difficult to find resources either online or in print, which makes it severely low-resourced like the majority of African vernaculars. Its written form has only started appearing in social media using either Arabic script or a mix of numbers and the Latin alphabet.

Mozambican Portuguese The Portuguese spoken in Mozambique is called Mozambican Portuguese, commonly referred to as the Portuguese of Mozambique. It differs from other Portuguese variants in a few ways, such as the lexicon, which incorporates many African terms and expressions that are common in Mozambique and other Portuguese-speaking nations. Additionally, it has a distinctive accent and rhythm that are affected by the Mozambican languages used locally in its pronunciation. Some grammar forms and structures in Mozambican Portuguese are different from those used in European Portuguese. Additionally, there are loan terms that were acquired from Mozambican and other African languages.

Tigrinya Also spelt Tigringa, is a Semitic language spoken in the Tigray region and Eritrea. The language uses Geez script with some additional Tigrinya alphabets and is closely related to Geez, and Amharic. The language has around 10 million speakers and 6.4 million are found in the Ethiopian Tigray region.

Xitsonga/Tsonga Xitsonga is a Bantu language originally from Mozambique but also spoken in different southern African countries. In Mozambique, the same language is referred to as Changana. It is part of the Tswa-Ronga language group, which also includes Tshwa and Rhonga. These three languages are mutually intelligible, meaning speakers of one can understand the other two languages in the same group. In addition to Mozambique, Xitsonga is also spoken in South Africa, Eswatini (formerly Swaziland), and Zimbabwe. In Mozambique, the following dialectal variants of Changana are recognized: Xihlanganu, Xidzonga, Xin'walungu, Xibila, and Xihlengwe. According to Omniglot [8] there are about 8.9 million speakers of Xitsonga, including 5.68 million in South Africa (in 2013), 3.1 million speakers in Mozambique (in

[7] http://rgphentableaux.hcp.ma/Default1/

[8] https://omniglot.com/writing/tsonga.php

2016), 100,000 speakers in Zimbabwe (in 1998), and 20,000 speakers in Eswatini (in 2010).

B Sentiment Class Distribution

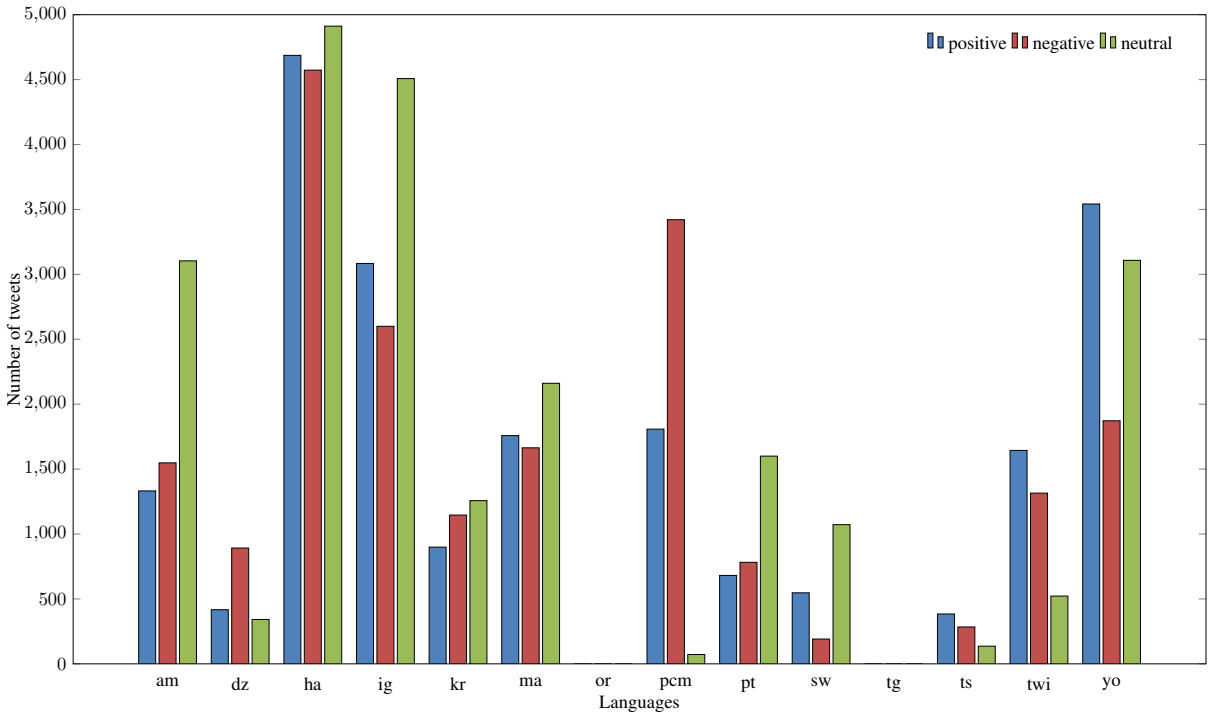

Figure 4: Training Datasets Sentiment Class Distribution

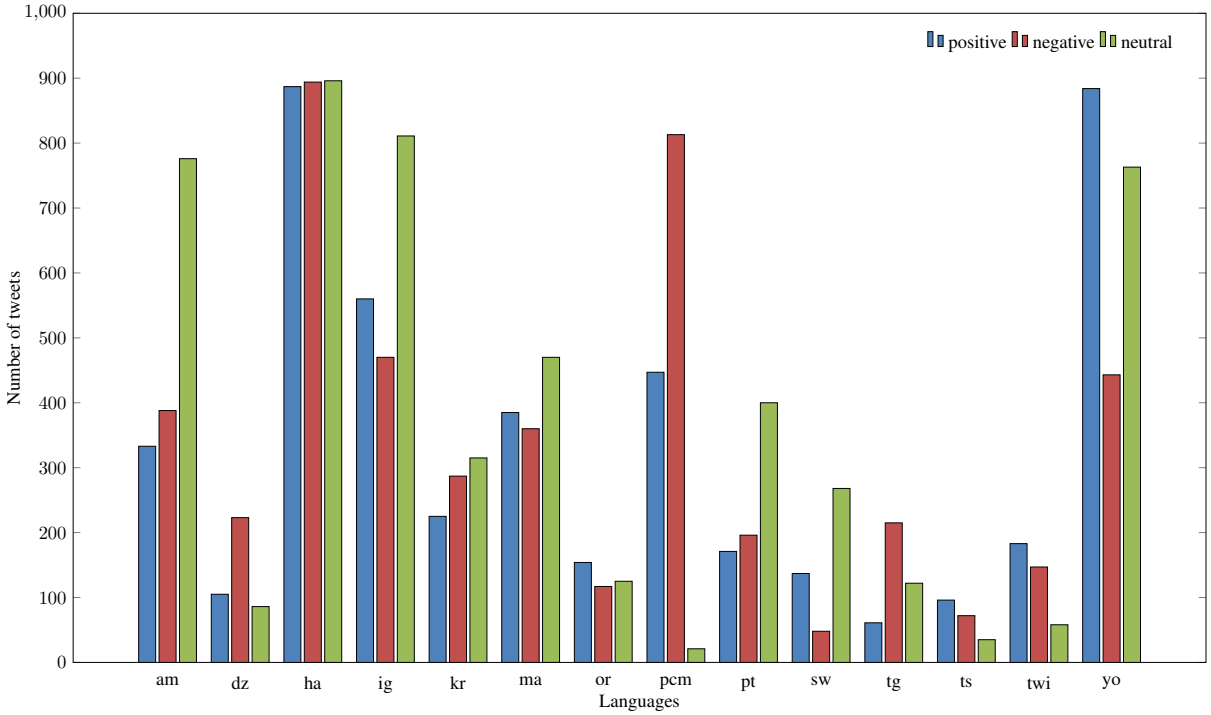

Figure 5: Development Datasets Sentiment Class Distribution

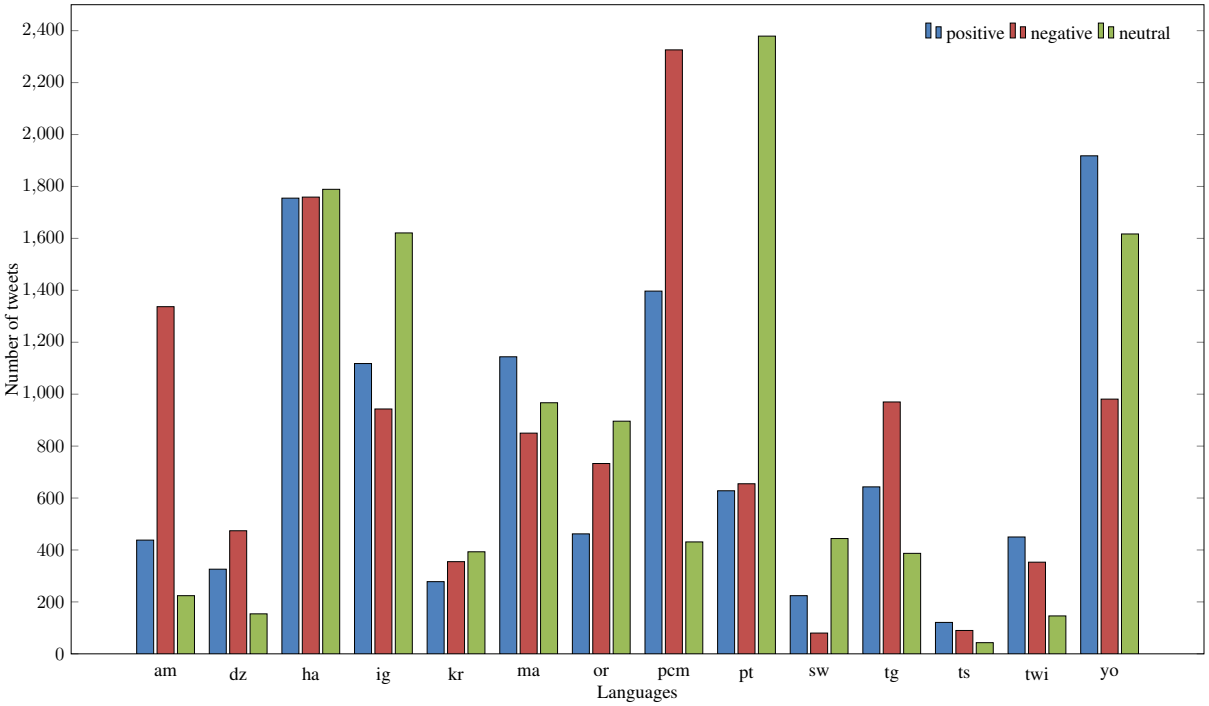

Figure 6: Test Datasets Sentiment Class Distribution