NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis

Shamsuddeen Hassan Muhammad[1,2*+], **David Ifeoluwa Adelani**[3*], **Sebastian Ruder**[4]
Ibrahim Sa'id Ahmad[5+], **Idris Abdulmumin**[6*+], **Bello Shehu Bello**[5+], **Monojit Choudhury**[7]
Chris Chinenye Emezue[8*], **Saheed Salahudeen Abdullahi**[10+]
, Anuoluwapo Aremu[11*], **Alipio Jeorge**[1,2] **Pavel Brazdil**[1]
[1] LIAAD - INESC TEC, [2]Faculty of Sciences-University of Porto, Portugal,
[3]Spoken Language Systems Group (LSV), Saarland University, Germany, [4]Google Research
[5]Faculty of Computer Science and Information Technology, Bayero University, Kano, Nigeria
[6]Department of Computer Science, Ahmadu Bello University, Zaria, Nigeria, [7]Microsoft Research India,
[8]Technical University of Munich, Germany, [9]Clear Global, [10]Kaduna state University
[*]MasakhaneNLP ,[+]HausaNLP
{shmuhammad.csc,isahmad.it, bsbello.cs}@buk.edu.ng

Abstract

Sentiment analysis is one of the most widely studied applications in NLP, but most work focuses on languages with large amounts of data. We introduce the first large-scale human-annotated Twitter sentiment dataset for the four most widely spoken languages in Nigeria—Hausa, Igbo, Nigerian-Pidgin, and Yorùbá—consisting of around 30,000 annotated tweets per language (and 14,000 for Nigerian-Pidgin), including a significant fraction of code-mixed tweets. We propose text collection, filtering, processing and labeling methods that enable us to create datasets for these low-resource languages. We evaluate a range of pre-trained models and transfer strategies on the dataset. We find that language-specific models and language-adaptive fine-tuning generally perform best. We release the datasets, trained models, sentiment lexicons, and code to incentivize research on sentiment analysis in under-represented languages.

Keywords: sentiment analysis, low-resource, twitter corpus, natural language processing

1. Introduction

Sentiment analysis (SA) deals with the detection and classification of sentiment in texts (Pang and Lee, 2007). In recent years, SA has attracted a considerable amount of interest, which can be attributed to its many vital applications. However, most work on SA focuses on high-resource languages such as English (Yimam et al., 2020) while languages with a limited amount of data remain poorly represented (Nasim and Ghani, 2020). This problem is not unique to sentiment analysis but affects NLP research as a whole (Joshi et al., 2020). Recently, ∀ et al. (2020) and Adelani et al. (2021) examined how socio-cultural factors hinder NLP for low-resource languages, potentially resulting in economic inequities (Weidinger et al., 2021).

With over 200 million people and 522 native languages, Nigeria is the most populous and linguistically diverse country in Africa as well as the third most multilingual country in the world.[1] However, due to a lack of training data for many NLP applications, these languages are underserved by digital technology. Therefore, a concerted effort is required to create resources for such languages (Adelani et al., 2021).

In this paper, we present *NaijaSenti*[2]—an open-source Twitter sentiment dataset for the four most widely spoken languages in Nigeria—Hausa, Igbo, Pidgin, and Yorùbá. This is the largest labelled sentiment dataset in these languages to date. As the Twitter API does not support these languages, we propose methods to enable the collection, filtering, and annotation of such low-resource language data. Overall, we annotated around 30,000 tweets in Hausa, Igbo, and Yorùbá as well as 14,000 tweets in Nigerian Pidgin (also known as Naija). The data highlights the challenges of sentiment analysis in these languages. For instance, the absence of diacritics makes some tweets ambiguous in Yorùbá and Igbo. In addition, code-mixing is a common occurrence, with about 43% of Igbo tweets code-mixing between Igbo and English.

We conduct extensive experiments demonstrating that state-of-the-art pre-trained multilingual models achieve strong performance on sentiment classification on NaijaSenti. The best models have been explicitly trained on unlabelled data in African languages during pre-training such as AfriBERTa (Ogueji et al., 2021) or using language-adaptive fine-tuning (Pfeiffer et al., 2020).

Contributions The main contributions of this paper are:

[1] We curate large-scale manually annotated code-mixed and monolingual sentiment datasets for Hausa, Igbo, Yorùbá and Pidgin languages.

[2] We built a manually annotated sentiment lexi-

[1] https://www.ethnologue.com/guides/countries-most-languages

[2] https://github.com/hausanlp/NaijaSenti

Dataset	Language	Open-source	Annotated/translated	Code-mixed	Source
Abubakar et al. (2021)	Hausa	✗	annotated	✓	Twitter
Ogbuju and Onyesolu (2019)	Igbo	✗	translated	✗	General
Umoh et al. (2020)	Igbo	✗	annotated	✗	General
Oyewusi et al. (2020)*	Pidgin	✗	annotated/translated	✓	Twitter
Orimaye et al. (2012)	Yorùbá	✓	annotated	✓	Youtube
Iyanda and Abegunde (2019)	Yorùbá	✗	annotated	✗	General
Ours	Hausa, Igbo, Yorùbá, Pidgin	✓	annotated	✓	Twitter

Table 1: Summary of datasets used in six existing datasets on sentiment analysis in four major Nigerian languages in comparison to ours. *: The provided URL is no longer accessible.

con in Hausa, Igbo, and Yorùbá. We also semi-automatically develop translated emotion and sentiment lexicons in these languages.

[3] We curate the largest Twitter corpus in each language that can be useful for other NLP downstream tasks.

[4] We present several benchmark experiments on sentiment analysis in Hausa, Igbo, Yorùbá, and Pidgin languages.

[5] We make the datasets and code freely available to foster further research in the NLP community.

2. Related Work

SA for low-resource languages Sentiment analysis for low-resource languages has recently gained popularity (Yimam et al., 2020; Xia et al., 2021; Jovanoski et al., 2021) due to the availability of relatively large amounts of tweets in such languages. Several studies have investigated using Twitter for sentiment analysis—by either automatically building a Twitter corpus or manually annotating one. Notable studies that automatically built Twitter corpora include Go et al. (2009), Pak and Paroubek (2010), and Wicaksono et al. (2014). More recently, Kwaik et al. (2020) automatically built an Arabic Twitter sentiment analysis corpus using distant supervision and self-training. In contrast, other studies, such as Refaee and Rieser (2014a), Brum and Nunes (2017), Mozetič et al. (2016), Nakov et al. (2019), and Moudjari et al. (2020) employed native speakers or expert annotators to manually annotate the corpus. Our work is most closely related to Al-Twairesh et al. (2017) and Kwaik et al. (2020), as it involves both the use of emoji as a distantly supervised approach for tweet extraction and the use of a translated sentiment lexicon to filter tweets before manual annotation (Nakov et al., 2019)

Despite advances in sentiment analysis for low-resource languages, indigenous Nigerian languages have received scant attention. This is mostly owing to the absence of a freely accessible dataset in these languages. Nevertheless, there has been a significant amount of studies on Nigerian code-mixed English (Nwofe, 2017; Olaleye et al., 2018; Oyebode and Orji, 2019; Kolajo et al., 2019; Rakhmanov, 2020; Olagunju et al., 2020; Onyenwe et al., 2020; Honkanen and Müller, 2021). Most work on Nigerian languages has relied on automatically generated data, including the following:

Hausa Abubakar et al. (2021) built a Twitter corpus and introduced combined Hausa and English features in a classifier.

Igbo Ogbuju and Onyesolu (2019) translated an English sentiment lexicon (Hu and Liu, 2004) and manually added Igbo native words to create IgboSentiLex. Umoh et al. (2020) analysed Igbo emotion words using Interval Type-2 Fuzzy Logic.

Yorùbá Orimaye et al. (2012) built a Yorùbá corpus from YouTube and applied a translated SentiWordNet for the sentiment analysis task. Iyanda and Abegunde (2019) created a multi-domain corpus (health, business, education, politics) and used different classic ML classifiers such as SVM to predict sentiment in text.

Pidgin Oyewusi et al. (2020) built a Pidgin tweet corpus and used a translated VADER English lexicon for sentiment analysis.

Table 1 summarizes the existing datasets for Nigerian languages; only two datasets are freely available, indicating that more work is needed to make indigenous datasets accessible and to stimulate research in these languages. To the best of our knowledge, ours is the first publicly available large-scale manually annotated dataset for sentiment analysis research in the following Nigerian languages: Hausa, Igbo, Yorùbá and Nigerian Pidgin (see Appendix A for the language description and characteristics.)

3. Data Collection and Cleaning

3.1. Data Collection

Twitter provides easy access to a large amount of domain-independent and topic-independent public opinionated user-generated data. We collected tweets using the Twitter Academic API [3], which provides real-time and historical tweet data. The Twitter API supports retrieving tweets in 32 high-resource languages

[3] https://developer.twitter.com/en/products/twitter-api/academic-research

Language	Tweet	Translation into English	Sentiment
Hausa (hau)	@USER Aunt rahma i luv u wallah irin totally dinnan	@USER Aunty rahma I swear I love you very much	positive
Igbo (ibo)	akowaro ya ofuma nne kai daalu nwanne mmadu we go dey alright las las	they told it well my fellow sister well done at the end we will be all right	positive
Naija (pcm)	E don tay wey I don dey crush on this fine woman . . .	I have had a crush on the beautiful woman for a while . . .	positive
Yorùbá (yor)	oníréégbè aláàdúgbò ati olójúkòkòrò	mischievous and coveteous neighbour	negative

Table 2: Examples of tweets, their English translation, and sentiment in different Nigerian languages. The Hausa and Igbo examples are code-switched with Naija. We highlight sentiment-bearing words in blue (positive) and red (negative).

using language parameters. This makes extracting a tweet in these languages easy. In contrast, none of the languages considered in this work are supported by the API. We therefore considered different heuristic approaches for crawling tweets.

Stopwords, emoji, and sentiment words Caswell et al. (2020) have shown that token-based filtering is a useful processing step for automatic language identification. Hence, we automatically built lists of common words (stopwords), which are verified by native speakers and used them to query the Twitter API to retrieve tweets in each language. Go et al. (2009) used emoticons and Kwaik et al. (2020) used emojis as a distantly supervised approach to automatically classify subjective tweets as positive or negative. Using a similar approach, we used happy and sad emojis (Kralj Novak et al., 2015) in combination with stopwords to query the Twitter API to extract tweets that contain both stopwords and emojis. In addition, we used the Google Language API to translate the Affin lexicon (Årup Nielsen, 2011) to each of the languages (Hausa, Igbo, Yorùbá), except for Pidgin. We then filtered the tweets using the translated Affin sentiment lexicon to improve the likelihood of annotating sentiment-bearing tweets (UzZaman et al., 2013).

Hashtags and Handles We used Twitter hashtags to crawl tweets from trending issues (e.g., #Yorubaday) and news handles (e.g., @bbchausa)—which are factual and nonsubjective. We selected handles that tweet frequently in each language from the Indigenous Tweets[4] website.

One downside of this approach is that Twitter conversations with a popular Twitter handle may dominate the dataset and may introduce a bias towards certain topics. For example, a Hausa Twitter conversation that involves the handle `@bbchausa` and another conversation involving the handle `@Rahmasadau` make up 54% and 14% respectively of collected tweets associated with Hausa handles. Limiting the number of tweets per conversation mitigates this problem.

[4] http://indigenoustweets.com/

3.2. Language Detection and Data Cleaning

Stopwords overlap across indigenous languages in a multilingual society such as Nigeria (Caswell et al., 2020). This results in tweets being collected in a language that differs from the query language. For example, using the stop word "*nke*" to crawl tweets in Igbo produces tweets in Hausa, such as "*amin ya rabbi godiya nke*". To mitigate this, we collected tweets based on locations where a language is predominantly spoken, using the location, longitude, latitude, and radius parameters (25 miles) to specify a circular geographic area.

We also used Google CLD3[5] and Natural Language API[6] to detect the language of the collected tweets. Pidgin is not supported by the API, so we used the stopword list to build an n-gram language detection tool to detect Pidgin. Before annotation, we cleaned the tweets. Retweets and duplicates were removed. We removed URLs and mentions as well as trailing and redundant white spaces, converted all tweets to lowercase, and removed tweets with less than three words as they may contain insufficient information for sentiment analysis (Yang et al., 2018).

4. Annotation and the NaijaSenti Dataset

4.1. Annotation Guidelines

Our annotation guidelines focus on the classification of subjective tweets. A subjective tweet has a positive or negative emotion, opinion, or attitude (Refaee and Rieser, 2014b). The guidelines define five classes: positive (POS), negative (NEG), neutral (NEU), mixed (MIX), and indeterminate (IND) (see Appendix B for a detailed explanation of these classes).

4.2. Annotation Process

Annotators training and preparation: For each language, with the exception of Pidgin, we recruited three native speakers as annotators. For Pidgin, we recruited nine annotators. We also recruited a coordinator for each language to supervise and ensure the qual-

[5] https://github.com/google/cld3

[6] https://cloud.google.com/natural-language/docs

ity of the annotation task. The annotators and coordinators have backgrounds in either computer science or linguistics and were trained on the annotation task using the LightTag annotation tool (Perry, 2021).
Data annotation is not a one-off process; it requires an agile approach with many iterations, collecting feedback from the annotators during the pilot stage, and refining the annotation guide to ensure the annotators can achieve a reasonable performance before moving to the next stage. We performed three iterations of training and annotation practice of 100 tweets. For the first two iterations, the agreement among the annotators was poor. We asked the annotators for feedback and adapted a simplified sentiment questionnaire annotation guide (Mohammad, 2016).

Tweets annotation: The dataset was annotated in batches of 1,000 tweets by the annotators. For each batch, we adjudicated cases where all three annotators assigned a different label to a tweet. Annotators discuss these tweets, which allows them to address ambiguities, peculiar issues, and recommend ways to improve the annotation guidelines. We excluded these ambiguous tweets from the dataset. We iteratively update our annotation guide based on the adjudication reports. Overall, the annotators annotated the following number of tweets: Hausa (35,000), Igbo (29,000), Pidgin (6,000), and Yorùbá (33,000).

Determining the gold label: People often disagree on subjective concepts (Beddor, 2019). For example, person A, who has been using Apple products, says, "The Apple iPhone camera is better than the Samsung camera", while person B says, "The Samsung camera is better". This is an example of subjective disagreement in contrast to objective disagreement. Therefore, different from the simple majority vote approach (Davani et al., 2021), we introduced a new form of majority vote that involves an independent annotator who adjudicates subjective disagreement cases as follows:[7]

- Three-way agreement: Similar to the majority vote approach, if all three annotators agree on a label, we consider the agreed sentiment class to be the gold standard.
- Three-way disagreement: When all annotators disagree on a label, we discard the tweet.
- Two-way partial disagreement: If two of the annotators agree on a label, and the third annotator has a partial disagreement. For example, if two annotators classify a tweet as POS (or NEG), and the other annotator classifies it as a non-contradicting class such as NEU, we consider the POS (or NEG) classification to be the gold standard.
- Two-way disagreement: If two of the annotators agree on a label, and the third annotator has a total disagreement. For example, if two annotators identify a tweet as POS and another as NEG or vice versa, the majority vote is not the final class (in this case, POS). To resolve such subjective disagreement, independent annotators review the disagreement and assign a final label.

[7]We determine a single gold label for sentiment analysis in accordance with prior work. Future work may alternatively leverage annotator disagreement (Fornaciari et al., 2021).

		Sentiment datasets			
	sent.	hau	ibo	yor	pcm
5-class	POS	9,235	5,621	9,839	3,010
	NEG	9,033	4,726	5,003	5,635
	NEU	12,826	14,877	14,356	717
	IND	8	1,909	1,754	1,500
	MIX	1,466	19	622	1,040
	Total	32,568	27,152	31,574	11,902
IAA (κ)		0.487	0.488	0.555	0.434
3-class	POS	8,019	5,395	9,391	2,104
	NEG	8,119	4,513	4,638	4,156
	NEU	11,122	13,380	13,367	577
	Total	27,260	23,288	27,396	6,837
IAA (κ)		0.607	0.516	0.600	0.512

Table 3: 3-class and 5-class annotation and inter-Annotator agreement.

Sentiment lexicons We created sentiment lexicons in three languages (Hausa, Igbo, and Yorùbá) based on NaijaSenti. We ask three annotators to tag words that convey negative or positive sentiment in a tweet. We used a simple majority vote. An independent annotator adjudicated cases where the annotators disagreed or only one person tagged a word as positive or negative. The distribution of the lexicon is presented in Table 6. We also created semi-automatically translated versions of the NRC emotion lexicon (Mohammad and Turney, 2013) and AFFIN sentiment lexicon (Årup Nielsen, 2011) for Hausa, Igbo, and Yorùbá. We used the Google Translate API[8] to translate the lexicon. Afterwards, professional human translators verified and corrected translations and added missing diacritics.

4.3. Inter-Annotator Agreement

We used the Fleiss kappa (κ) reliability measure (Fleiss et al., 2013) to determine the inter-annotator agreement (IAA) between the three annotators. The IAA for the 5-class and adjusted 3-class agreements are shown in Table 3. The agreement between the five classes was not particularly high (e.g.,(κ) = 0.49) for Hausa. However, according to the Fleiss classification (Fleiss et al., 2013), an agreement greater than 0.40 is considered reasonable (moderate) and beyond chance.
We further computed the IAA (κ) (see Table 4) of each class with other classes to determine which classes the annotators find confusing or difficult and frequently

[8]https://cloud.google.com/translate

	corpus			
Class	**hau**	**ibo**	**yor**	**pcm**
POS	0.626	0.542	0.626	0.481
NEG	0.518	0.521	0.553	0.492
NEU	0.442	0.404	0.491	0.159
MIX	0.297	0.020	0.242	0.196
IND	0.045	0.591	0.764	0.707

Table 4: Fleiss kappa agreement among each class

disagree on. Table 4 indicates that the annotators generally have the lowest overall agreement in the MIXED class, which includes elements of both the positive and negative class, and some annotators identify it as either negative or positive. This highlights the subtlety of annotating mixed sentiment on social media and is in contrast to reviews where the annotation of mixed sentiment is clearer (Potts et al., 2021). To address this, we introduced an adjusted 3-class IAA agreement.

In the adjusted 3-class agreement, we considered only positive, negative, and neutral as valid sentiment classes. We selected only tweets that have at least two labels in the valid classes and discarded the rest. For the selected tweets, where the label between two annotators is valid and the third label is in the invalid sentiment (Indeterminate or Mixed), we changed the label to the agreed valid label. For instance, given three annotation labels of a tweet as POS, POS, MIX, the third label is changed to POS, whereas the annotation labels of POS, POS, NEU are left unchanged. Table 3 shows the final statistics of at least two agreed tweets of the various datasets after converting to the 3-class annotation, and their corresponding inter-annotator agreements (IAA) using the Fleiss' Kappa (κ) metric.

To determine the IAA performance over time, Figure 1 shows the IAA in three languages over 30 batches. We hypothesised that as the annotators became more experienced with the task, their annotation quality would improve. However, the IAA overall performance deteriorates over time. This suggests that familiarity with the task does not necessarily improve the IAA. Only Yorùbá annotators have some level of consistency that is not below 0.5. Therefore, it is important to monitor the IAA as the annotation progress.

4.4. Human Evaluations

We assess human performance by re-annotating 200 random sample tweets by three different annotators (Warstadt et al., 2019; Nangia and Bowman, 2019). We take the majority vote as the final class. The human performance offers us an idea of the machine's upper bound performance and the reproducibility of the first three annotators (Warstadt et al., 2019). Table 5 shows the micro-F1 and Matthew's correlation coefficient (MCC) (Jurman et al., 2012). The human performance result validates the reliability of the corpus and is in line with prior literature (Rosenthal et al., 2017).

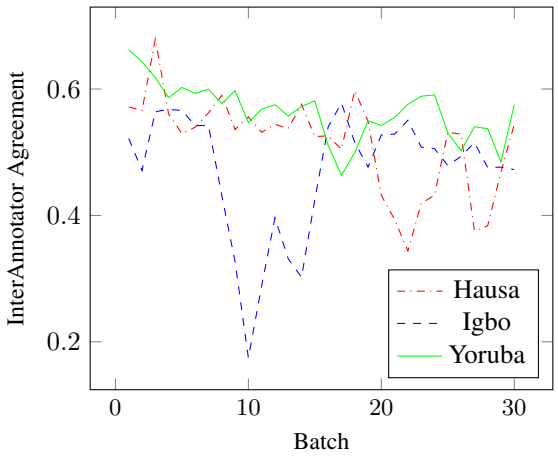

Figure 1: Inter-annotators progress over thirty batches—one-thousand tweets per batch.

Measures	**hau**	**ibo**	**yor**	**pcm**
Micro-F1	0.75	0.76	0.85	0.78
MCC	0.63	0.68	0.77	0.69

Table 5: Human performance result using micro-F1 and Matthew's correlation coefficient.

4.5. NaijaSenti Statistics

Table 3 shows the summary of our dataset with 5-class and adjusted 3-class. Other key statistical information such as number of tokens, type of words, and type-token ration (TTR), which measure the lexical richness of a text are presented in Table 6. We also show the number of monolingual and code-mixed tweets in each dataset. The percentage of code-mixed tweets highlights the highly multilingual setting in Nigeria. Code-mixing is more prevalent in Igbo (43%) than in Hausa (23%) and Yorùbá (19%). Code-mixing between English and a native language is more common than between native languages but can also occur between more than two native languages.

Hausa does not have diacritics and therefore has an insignificant number of indeterminate cases (only 8), unlike Yorùbá and Igbo where the absence of diacritics may render a tweet incomprehensible and therefore lead to labelling it as indeterminate. Tone in Yorùbá helps to give meaning to words in context, especially words that have the same orthographic representation. For instance, the sentence "Awon omo fo abo" does not have a meaning without diacritics, and the annotators classify it as indeterminate (IND). However, the same sentence with diacritics can have two opposite meanings: Àwọn ọmọ fọ abọ́ (The children washed the dishes) has a positive meaning, and Àwọn ọmọ fọ́ abọ́ (The children broke the dishes) is negative.

Similarly, tonality is heavily used in Igbo. Many Twitter users do not write Igbo with diacritics. One reason is the lack of an Igbo keyboard that accepts and shows diacritics. Even if such a keyboard exists, it is not used by many. While it may be fairly easy to understand the

Languages	mono-lingual	#code-mixed	token	Wordtype	TTR	neg words	pos words
Hausa (hau)	21,039	6,426	3,493,92	30,747	0.09	1,008	1,214
Igbo (ibo)	8,688	6,561	1,830,02	4,107	0.02	1,180	904
Naija (pcm)	–	–	3,669,68	8,736	0.02	–	–
Yorùbá (yor)	18,662	4,457	40,897,6	8,948	0.02	2,185	2,228

Table 6: Key stats of NaijaSenti: #mono-lingual tweets, #code-mix tweets, #token, #word types and type-to-token ratio (TTR)

sentiment of Igbo tweets *in context* on Twitter—either due to the presence of emojis or the context of the surrounding discourse, it is quite difficult and sometimes ambiguous to correctly annotate the tweets when they stand on their own. The example below highlights the impact of tone and punctuation marks on the same Igbo tweets but with different sentiment:

- ò nwèkwàrà mgbe i naenwe sense **?** – Will you ever be able to talk sensibly? – You're a fool.
- ò nwèkwàrà mgbe i naenwe sense – Sometimes you act with great maturity – I'm impressed.

Yes/No questions in Igbo are realized by a low tone on the subject pronoun, as in the first sentence above. So, without any tone and lacking punctuation, the author's intended meaning is difficult to determine.

Benchmark Data split To create a benchmark dataset for each language, we only make use of three sentiment classes: negative, neutral, and positive. For each of the languages, we split tweets in each class by 70%, 10% and 20% ratios for the TRAIN, DEV and TEST splits. After combining the tweets of the all classes, Hausa has 18,989 / 2,714 / 5,427 in the TRAIN, DEV and TEST split. Igbo has the split: 12,930 / 1,846 / 3,697. Naija has the split: 6,552 / 937 / 1,873, and Yorùbá has the split: 16,209 / 2,316 / 4,632.

5. Experimental Setup

5.1. Sentiment Classification Models

Sentiment classification is a well-studied problem in NLP and many machine learning models have been developed for this task. State-of-the-art approaches on English data use pre-trained language models (PLMs) such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), which provide superior performance. Multilingual variants of PLMs provide an opportunity to quickly adapt to various languages, including languages not seen during training (Pfeiffer et al., 2020). We compare several standard multilingual PLMs on the four languages. We fine-tune each model on the data of each language separately using HuggingFace Transformer (Wolf et al., 2020). Appendix C provides the details of the hyper-parameters used for training.

mBERT is a multilingual variant of BERT pre-trained on 104 languages, including one Nigerian language—Yorùbá. mBERT was pre-trained using masked language modeling (MLM) and next-sentence prediction task. We fine-tune the mBERT-base-cased model with 172M model parameters by adding a linear classification layer on top of the pre-trained transformer model.

XLM-R Similar to mBERT, XLM-R (Conneau et al., 2020) is a multilingual variant of RoBERTa pre-trained on 100 languages, including Hausa as the only Nigerian language. Unlike mBERT, XLM-R only uses MLM during pre-training. We use XLR-base with 270M model parameters for fine-tuning on the NaijaSenti corpus.

RemBERT scales up mBERT to a larger model size (559M) and decouples embeddings, which enables a larger output embedding size during pre-training, resulting in stronger pre-training and downstream performance (Chung et al., 2021). RemBERT covers the three major Nigerian languages, except for Nigerian-Pidgin.

AfriBERTa trains a RoBERTa-style model on 11 African languages (Ogueji et al., 2021) including all four Nigerian languages in NaijaSenti. The model was trained on less than 1GB of data (since most African languages are low-resourced). We use AfriBERTa-large with 126M parameters. AfriBERTa has been shown to perform competitively on an African NER dataset (Adelani et al., 2021) despite its small model size and limited pre-training data.

mDeBERTaV3 Unlike the other four models pre-trained on the MLM task, mDeBERTaV3 (He et al., 2021) makes use of ELECTRA-style (Clark et al., 2020) pre-training where a discriminator is trained to detect replaced tokens instead of predicting masked tokens. mDEBERTaV3 does not support any of the Nigerian languages. We use the mDEBERTaV3-base model with 276M model parameters similar to XLM-R-base.

5.2. Language Adaptive Fine-tuning

Many multilingual PLMs support only a few African languages. For example, mBERT only supports three African languages (Malagasy, Swahili, and Yorùbá). Language adaptive fine-tuning (LAFT) is an effective method of adapting PLMs to a new language by fine-tuning PLMs MLM on unlabeled texts in the new language (Pfeiffer et al., 2020). The approach is similar to domain-adaptive fine-tuning (Howard and Ruder, 2018; Gururangan et al., 2020). LAFT has been shown

Model	NG lang. supported	PLM size	hau	ibo	pcm	yor	Avg
Majority Classifier							
Majority (Weighted F1)	–	–	16.6	26.9	45.2	19.0	26.9
Majority (Micro F1)	–	–	33.3	44.0	60.2	35.9	43.4
Multilingual PLMs							
AfriBERTa-large	hau, ibo, pcm, yor	126M	$81.0^{*}_{\pm 0.2}$	$\mathbf{81.2^{*}_{\pm 0.5}}$	$75.0_{\pm 0.6}$	$80.2^{*}_{\pm 0.6}$	$79.3^{*}_{\pm 0.3}$
mBERT-base	yor	172M	$77.8_{\pm 0.5}$	$79.8_{\pm 0.5}$	$72.4_{\pm 1.5}$	$77.6_{\pm 0.9}$	$76.9_{\pm 0.3}$
XLM-R-base	hau	270M	$78.4_{\pm 1.0}$	$79.9_{\pm 0.7}$	$76.3_{\pm 0.6}$	$76.9_{\pm 0.4}$	$77.9_{\pm 0.2}$
mDeBERTaV3-base	None	276M	$79.3_{\pm 0.1}$	$80.7_{\pm 0.2}$	$77.6^{*}_{\pm 0.8}$	$78.4_{\pm 0.5}$	$79.0_{\pm 0.3}$
RemBERT	hau, ibo, yor	559M	$79.0_{\pm 0.7}$	$79.9_{\pm 0.4}$	$\mathbf{78.4^{*}_{\pm 1.4}}$	$78.0_{\pm 0.6}$	$78.8_{\pm 0.6}$
Multilingual PLMs+LAFT							
mBERT+LAFT (General)	hau / ibo / pcm / yor	172M	$80.8_{\pm 0.3}$	$80.4_{\pm 0.4}$	$74.2_{\pm 0.5}$	$80.8_{\pm 0.5}$	$79.1_{\pm 0.3}$
mBERT+LAFT (Tweet)	hau / ibo / pcm / yor	172M	$79.3_{\pm 0.6}$	$77.7_{\pm 0.6}$	$74.0_{\pm 0.7}$	$76.8_{\pm 0.3}$	$77.0_{\pm 0.3}$
XLM-R-base+LAFT (General)	hau / ibo / pcm / yor	270M	$\mathbf{81.5^{*}_{\pm 0.7}}$	$80.8^{*}_{\pm 0.8}$	$74.7_{\pm 1.5}$	$\mathbf{80.9^{*}_{\pm 0.4}}$	$\mathbf{79.5^{*}_{\pm 0.3}}$
XLM-R-base+LAFT (Tweet)	hau / ibo / pcm / yor	270M	$79.5_{\pm 0.9}$	$77.0_{\pm 0.5}$	$74.8_{\pm 0.7}$	$76.2_{\pm 0.4}$	$76.9_{\pm 0.2}$
Multi-task Multilingual PLMs							
AfriBERTa-large	hau, ibo, pcm, yor	126M	$81.2^{*}_{\pm 0.2}$	$80.8^{*}_{\pm 0.2}$	$74.5_{\pm 0.6}$	$80.4^{*}_{\pm 0.7}$	$79.3_{\pm 0.3}$
mDeBERTaV3-base	None	276M	$79.0_{\pm 0.2}$	$79.3_{\pm 0.5}$	$76.8_{\pm 0.6}$	$78.7_{\pm 0.4}$	$78.4_{\pm 0.3}$

Table 7: Weighted F1 evaluation of different Models. Average and standard deviation over 5 runs. Numbers with "*" are within the standard deviation of the best model. The models using language adaptive fine-tuning (LAFT) are trained on either the General domain or Twitter domain.

to be very effective in improving NER performance in several African languages (Alabi et al., 2020; Muller et al., 2021; Adelani et al., 2021). To further improve the LAFT performance, we perform vocabulary augmentation using 99 most frequent wordpieces inspired by (Chau et al., 2020; Pfeiffer et al., 2021) before further pre-training the PLM. We experimented on two collections of monolingual data: (1) Twitter domain (often very small; less than 50K tweets for Igbo and Yorùbá, and less than 600K tweets for Hausa and Nigerian-Pidgin), and (2) General domain (trained on mostly Common Crawl corpora, religious texts, and online news); for the latter, we use the checkpoints released by (Adelani et al., 2021).

5.3. Multi-task Sentiment Classification

In addition to fine-tuning separate models for each language, we trained a joint multi-task sentiment classification model on the four Nigerian languages by aggregating their training sets. The major advantage of this is that having a single model that can classify the sentiment in tweets in all major Nigerian languages facilitates deployment for practical applications. Knowledge from related languages may also be beneficial during transfer. This setting is possible because we are using multilingual PLMs that support multiple languages.

5.4. Cross-Lingual Transfer

Lastly, we evaluate the zero-shot performance of a sentiment classifier trained on English tweets from SemEval-2017 Task 4 (Rosenthal et al., 2017) on each of the four Nigerian languages. We also assess how many tweets from each of the Nigerian languages are needed to reach the zero-shot performance of a model transferred from English and to produce an accuracy score that is better than a majority classifier.

6. Experimental Results

6.1. In-language Training

Table 7 shows the performance of several sentiment classification models for three-way sentiment classification on four Nigerian languages. As the corpora do not have a balanced number of samples for each label, we also computed a majority classifier based on the dominant label in the corpus. hau, ibo and yor have more *neutral* tweets while pcm has more *positive* tweets. The performance of the majority classifier using the weighted F1-score is around $16-45\%$ for all languages and $33-60\%$ using Micro F1-score. On the other hand, PLMs have a minimum F1-score of 72%, demonstrating their usefulness for sentiment analysis.

Multilingual PLMs are quite similar in most cases with about a $1-3\%$ performance difference. The performance may depend on the language being seen during pre-training. mBERT has a slightly better performance ($+0.7\%$) for yor than XLM-R likely because yor was seen during pre-training. Similarly, XLM-R performs better for hau. RemBERT achieves slightly better performance than mBERT and XLM-R-base, demonstrating that a model with more capacity can improve performance. Surprisingly, we found mDeBERTaV3 that has not seen any of the Nigerian languages during pre-training to provide better results (78.7%) than other models except for AfriBERTa. mDeBERTaV3 makes use of replaced token detection (RTD), which has been shown to give superior performance for English (Clark et al., 2020). Overall, we found AfriBERTa to be the best baseline model for all languages because the model is more African language-centric. The main advantage of AfriBERTa is its smaller model size, which makes it easier to deploy especially on the

Model	hau	ibo	pcm	yor	Avg
AfriBERTa-large	**58.4**	**47.7**	62.0	**43.1**	**52.8**
mBERT-base	31.0	37.0	57.4	39.5	41.2
XLM-R-base	38.4	37.8	62.4	26.7	41.3
mDeBERTaV3-base	50.1	47.2	64.9	36.4	49.7
RemBERT	54.0	45.4	**66.2**	30.2	49.0

Table 8: Transfer Learning experiments. PLMs are trained on English SemEval 2017 and evaluated on NG languages in a zero-shot setting

African continent where most research labs cannot afford powerful GPUs.

Language adaptive fine-tuning (LAFT) has been shown to improve over the baseline with additional pre-training on monolingual data in the domain or language. Table 7 shows some improvement over mBERT and XLM-R when we apply LAFT on the general domain, on average $2-3\%$ on hau, and yor, and $<1\%$ on ibo. For pcm, we only identified an improvement for mBERT ($+1.8\%$). Interestingly, applying LAFT on the Twitter domain did not improve performance. The main reason for this is the small size of the Twitter data. For example, hau was further pre-trained on CC100 (Conneau et al., 2020) corpus with over 318MB and 3 million sentences for the general domain, but for Twitter, we only have around 512K tweets (32MB), which are often short. In general, we found AfriBERTa to be better or competitive than LAFT for the Nigerian languages except for pcm.

Multi-task sentiment classification We trained a single model on all languages with minimal drop in performance. In this setting, we only trained on the best two multilingual PLMs: AfriBERTa and mDeBERTaV3. We observe only a slight drop in performance with mDeBERTa (-0.6%) while the AfriBERTa performance is almost the same. This indicates that we could easily deploy a single sentiment classification model for the four major Nigerian languages, instead of multiple monolingual models.

6.2. Zero-shot Cross-Lingual Transfer

Table 8 shows the results of zero-shot transfer from English SemEval 2017 Task 4 tweets to the four Nigerian languages. The English SemEval corpus consists of 11,763 tweets in the training set. pcm has the best zero-shot performance across all models because of its linguistic similarity to English, its lexifier language. Similarly, we found an impressive zero-shot performance for hau with at least 50.0% F1-score when we train on AfriBERTa, mDeBERTaV3 and RemBERT. For ibo, the performance is over 45.4% on the three best PLMs while the zero-shot evaluation for yor is slightly lower ($36-43\%$). AfriBERTa gave the best overall result in the zero-shot transfer, and it is significantly better than the majority classifier (weighted average) for all languages: hau, ibo, pcm, and yor are better by 41.8%, 20.8%, 2%, and 19.1% respectively.

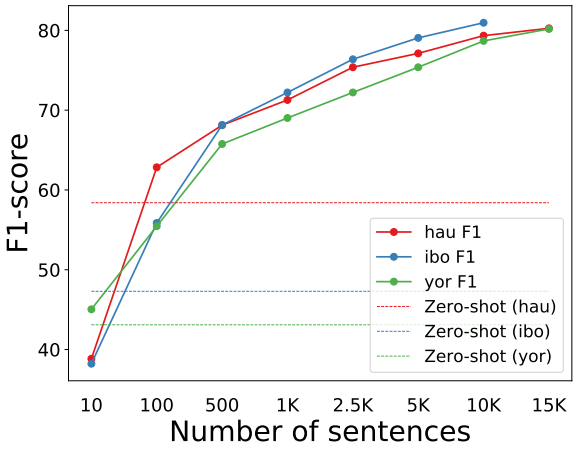

Figure 2: Sample Efficiency on hau, ibo, and yor using the AfriBERTa model

6.3. Sample Efficiency in Transfer

Figure 2 shows the result of training a sentiment classification model with different numbers of samples (10, 100, 500, 1K, 2.5K, 5K, 10K, and 15K). We fine-tune AfriBERTa on hau, ibo, and yor datasets of different sizes. We observe an F1 score of $38-40\%$ with only 10 examples, which already exceeds the majority voting performance in Table 7. Surprisingly, with only 100 sentences, we exceed the zero-shot transfer performance from English language, and with at least 1000 sentences, we already reach a decent performance of 70% F1. This result shows that we can leverage a multitask sentiment classification model trained on Nigerian languages to quickly adapt to other African languages with as few as 100 or 1000 annotated samples. Overall, we identify headroom for model improvement particularly in the zero-shot and few-shot cross-lingual transfer settings.

7. Conclusions and Future Work

In this paper, we present NaijaSenti—the first publicly available large-scale and manually annotated Twitter sentiment dataset for the four main Nigerian languages (Hausa, Igbo, Nigerian-pidgin, and Yorùbá). We propose methods to enable the collection, filtering, and annotation of such low-resource language data. Additionally, we introduce a manually annotated sentiment lexicon in three languages (Hausa, Igbo, and Yorùbá). We present benchmark experiments on Twitter sentiment dataset using state-of-the-art pre-trained language models and transfer learning. NaijaSenti has the potential to spark interest in sentiment analysis and other downstream NLP tasks in the languages involved. As future work, we plan to create benchmark experiments with our sentiment lexicon, and extend our dataset (NaijaSenti) to include other African languages (AfriSenti).

8. Acknowledgements

We thank Daan van Esch for feedback on a draft of this article. This work was carried out with support from 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 thank Tal Perry for providing the LightTag annotation tool. Finally, Shamsuddeen Muhammad acknowledges PhD grant through the Portuguese funding agency, FCT-Fundação para a Ciência e a Tecnologia within project: UID/EEA/50014/2019.

9. Bibliographical References

Abubakar, A. I., Roko, A., Bui, A. M., and Saidu, I. (2021). An enhanced feature acquisition for sentiment analysis of english and hausa tweets. *International Journal of Advanced Computer Science and Applications*, 12(9).

Adelani, D. I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., Palen-Michel, C., Buzaaba, H., Rijhwani, S., Ruder, S., Mayhew, S., Azime, I. A., Muhammad, S. H., Emezue, C. C., Nakatumba-Nabende, J., Ogayo, P., Anuoluwapo, A., Gitau, C., Mbaye, D., Alabi, J., Yimam, S. M., Gwadabe, T. R., Ezeani, I., Niyongabo, R. A., Mukiibi, J., Otiende, V., Orife, I., David, D., Ngom, S., Adewumi, T., Rayson, P., Adeyemi, M., Muriuki, G., Anebi, E., Chukwuneke, C., Odu, N., Wairagala, E. P., Oyerinde, S., Siro, C., Bateesa, T. S., Oloyede, T., Wambui, Y., Akinode, V., Nabagereka, D., Katusiime, M., Awokoya, A., MBOUP, M., Gebreyohannes, D., Tilaye, H., Nwaike, K., Wolde, D., Faye, A., Sibanda, B., Ahia, O., Dossou, B. F. P., Ogueji, K., DIOP, T. I., Diallo, A., Akinfaderin, A., Marengereke, T., and Osei, S. (2021). MasakhaNER: Named Entity Recognition for African Languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 10.

Al-Twairesh, N., Al-Khalifa, H., Al-Salman, A., and Al-Ohali, Y. (2017). Arasenti-tweet: A corpus for arabic sentiment analysis of saudi tweets. *Procedia Computer Science*, 117:63–72.

Alabi, J., Amponsah-Kaakyire, K., Adelani, D., and España-Bonet, C. (2020). Massive vs. curated embeddings for low-resourced languages: the case of Yorùbá and Twi. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2754–2762, Marseille, France, May. European Language Resources Association.

Beddor, B. (2019). Subjective Disagreement. *Noûs*, 53(4):819–851, December.

Brum, H. B. and Nunes, M. d. G. V. (2017). Building a sentiment corpus of tweets in brazilian portuguese. *arXiv preprint arXiv:1712.08917*.

Caswell, I., Breiner, T., van Esch, D., and Bapna, A. (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.

Chau, E. C., Lin, L. H., and Smith, N. A. (2020). Parsing with multilingual BERT, a small corpus, and a small treebank. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1324–1334, Online, November. Association for Computational Linguistics.

Chung, H. W., Fevry, T., Tsai, H., Johnson, M., and Ruder, S. (2021). Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (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, July. Association for Computational Linguistics.

Davani, A. M., Díaz, M., and Prabhakaran, V. (2021). Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *arXiv preprint arXiv:2110.05719*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, June. Association for Computational Linguistics.

Eberhard, D. M., Simons, G. F., and (eds.), C. D. F. (2020). Ethnologue: Languages of the world. twenty-third edition.

Fleiss, J. L., Levin, B., and Paik, M. C. (2013). *Statistical methods for rates and proportions*. john wiley & sons.

∀, ., Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S., Kabongo Kabenamualu, S., Osei, S., Sackey, F., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Berhe, M. M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., Martinus, L., Tajudeen, K., Degila, K., Ogueji, K., Siminyu, K., Kreutzer, J., Webster, J., Ali, J. T., Abbott, J., Orife, I., Ezeani, I., Dangana, I. A., Kamper, H., Elsahar, H., Duru, G., Kioko, G., Espoir, M., van Biljon, E., Whitenack, D., Onyefuluchi, C., Emezue, C. C., Dossou,

B. F. P., Sibanda, B., Bassey, B., Olabiyi, A., Ramkilowan, A., Öktem, A., Akinfaderin, A., and Bashir, A. (2020). Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online.

Fornaciari, T., Uma, A., Paun, S., Plank, B., Hovy, D., and Poesio, M. (2021). Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online, June. Association for Computational Linguistics.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July. Association for Computational Linguistics.

He, P., Gao, J., and Chen, W. (2021). Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *ArXiv*, abs/2111.09543.

Honkanen, M. and Müller, J. (2021). Interjections and emojis in nigerian online communication. *World Englishes*.

Howard, J. and Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. In *Proceedings of ACL 2018*.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Iyanda, A. R. and Abegunde, O. (2019). Predicting sentiment in yorùbá written texts: A comparison of machine learning models. In *Proceedings of SAI Intelligent Systems Conference*, pages 416–431. Springer.

Jaggar, P. (2001). *Hausa*. London Oriental and African language library. John Benjamins Publishing Company.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July. Association for Computational Linguistics.

Jovanoski, D., Pachovski, V., and Nakov, P. (2021). Sentiment analysis in twitter for macedonian. *arXiv preprint arXiv:2109.13725*.

Jurman, G., Riccadonna, S., and Furlanello, C. (2012). A comparison of mcc and cen error measures in multi-class prediction. *PLOS ONE*, 7(8):1–8, 08.

Kolajo, T., Daramola, O., and Adebiyi, A. (2019). Sentiment analysis on naija-tweets. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 338–343.

Kralj Novak, P., Smailović, J., Sluban, B., and Mozetič, I. (2015). Sentiment of emojis. *PLoS ONE*, 10(12):e0144296.

Kwaik, K. A., Chatzikyriakidis, S., Dobnik, S., Saad, M., and Johansson, R. (2020). An arabic tweets sentiment analysis dataset (atsad) using distant supervision and self training. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 1–8.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Mohammad, S. (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.

Moudjari, L., Akli-Astouati, K., and Benamara, F. (2020). An algerian corpus and an annotation platform for opinion and emotion analysis. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1202–1210.

Mozetič, I., Grčar, M., and Smailović, J. (2016). Multilingual twitter sentiment classification: The role of human annotators. *PloS one*, 11(5):e0155036.

Muller, B., Anastasopoulos, A., Sagot, B., and Seddah, D. (2021). When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online, June. Association for Computational Linguistics.

Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V., and Wilson, T. (2019). Semeval-2013 task 2: Sentiment analysis in twitter.

Nangia, N. and Bowman, S. R. (2019). Human vs. muppet: A conservative estimate of human performance on the glue benchmark. *arXiv preprint arXiv:1905.10425*.

Nasim, Z. and Ghani, S. (2020). Sentiment analysis on urdu tweets using markov chains. *SN Computer Science*, 1(5):1–13.

Nwofe, E. S. (2017). Pro-biafran activists and the

call for a referendum: A sentiment analysis of 'biafraexit'on twitter after uk's vote to leave the european union. *Journal of Ethnic and Cultural Studies*, 4(1):65.

Ogbuju, E. and Onyesolu, M. (2019). Development of a general purpose sentiment lexicon for Igbo language. In *Proceedings of the 2019 Workshop on Widening NLP*, page 1, Florence, Italy, August. Association for Computational Linguistics.

Ogueji, K., Zhu, Y., and Lin, J. (2021). Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Ohiri-Aniche, C. (2007). Stemming the tide of centrifugal forces in Igbo orthography. *Dialectical Anthropology*, 31(4):423–436.

Olagunju, T., Oyebode, O., and Orji, R. (2020). Exploring key issues affecting african mobile ecommerce applications using sentiment and thematic analysis. *IEEE Access*, 8:114475–114486.

Olaleye, S. A., Sanusi, I. T., and Salo, J. (2018). Sentiment analysis of social commerce: a harbinger of online reputation management. *International Journal of Electronic Business*, 14(2):85–102.

Onyenwe, I., Nwagbo, S., Mbeledogu, N., and Onyedinma, E. (2020). The impact of political party/candidate on the election results from a sentiment analysis perspective using# anambradecides2017 tweets. *Social Network Analysis and Mining*, 10(1):1–17.

Orimaye, S. O., Alhashmi, S. M., and Eu-gene, S. (2012). Sentiment analysis amidst ambiguities in youtube comments on yoruba language (nollywood) movies. In *Proceedings of the 21st International Conference on World Wide Web*, pages 583–584.

Oyebode, O. and Orji, R. (2019). Social media and sentiment analysis: The nigeria presidential election 2019. In *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 0140–0146. IEEE.

Oyewusi, W. F., Adekanmbi, O., and Akinsande, O. (2020). Semantic enrichment of nigerian pidgin english for contextual sentiment classification. *arXiv preprint arXiv:2003.12450*.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 1320–1326.

Pang, B. and Lee, L. (2007). Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2:1–135.

Perry, T. (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.

Pfeiffer, J., Vuli, I., Gurevych, I., and Ruder, S. (2020). MAD-X: An Adapter-based Framework for Multi-task Cross-lingual Transfer. In *Proceedings of EMNLP 2020*.

Pfeiffer, J., Vulić, I., Gurevych, I., and Ruder, S. (2021). UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Potts, C., Wu, Z., Geiger, A., and Kiela, D. (2021). DynaSent: A dynamic benchmark for sentiment analysis. 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 2388–2404, Online, August. Association for Computational Linguistics.

Rakhmanov, O. (2020). A comparative study on vectorization and classification techniques in sentiment analysis to classify student-lecturer comments. *Procedia Computer Science*, 178:194–204.

Refaee, E. and Rieser, V. (2014a). An arabic twitter corpus for subjectivity and sentiment analysis. In *LREC*, pages 2268–2273.

Refaee, E. and Rieser, V. (2014b). An Arabic Twitter corpus for subjectivity and sentiment analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2268–2273, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Rosenthal, S., Farra, N., and Nakov, P. (2017). Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518.

Umoh, U., Eyoh, I., Isong, E., Ekong, A., and Peter, S. (2020). Using interval type-2 fuzzy logic to analyze igbo emotion words. *Journal of Fuzzy Extension and Applications*, 1(3):217–240.

UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., and Pustejovsky, J. (2013). Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9.

Warstadt, A., Singh, A., and Bowman, S. R. (2019). Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Wicaksono, A. F., Vania, C., Distiawan, B., and Adriani, M. (2014). Automatically building a corpus for sentiment analysis on indonesian tweets. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, pages 185–194.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Xia, M., Zheng, G., Mukherjee, S., Shokouhi, M., Neubig, G., and Awadallah, A. H. (2021). Metaxl: Meta representation transformation for low-resource cross-lingual learning. *arXiv preprint arXiv:2104.07908*.

Yang, X., Macdonald, C., and Ounis, I. (2018). Using word embeddings in twitter election classification. *Information Retrieval Journal*, 21(2):183–207.

Yimam, S. M., Alemayehu, H. M., Ayele, A., and Biemann, C. (2020). Exploring amharic sentiment analysis from social media texts: Building annotation tools and classification models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1048–1060.

Årup Nielsen, F. (2011). A new anew: Evaluation of a word list for sentiment analysis in microblogs.

Appendix

A. Overview of the Languages

With over 522 native languages, Nigeria is the most multilingual country in Africa and the third most multilingual country in the world.[9] Although linguistically very diverse, the majority of the population speaks either Hausa, Igbo, Yorùbá, or Nigerian-Pidgin. Therefore, our work focuses on these three indigenous Nigerian languages (Hausa, Yorùbá, and Igbo) and Nigerian-Pidgin.

Hausa (`hau`): Hausa is a Chadic (Afroasiatic) language that is spoken in 3 broad dialects[10]: Eastern, Western and Northern (Jaggar, 2001). The language is written in two different scripts: Ajami and the more common Boko script. The Boko script uses the Latin characters without p, q, v and x as well as the following additional letters: consonants (ɓ, ɗ, ƙ, ƴ, kw, ƙw, gw, ky, ƙy, gy, sh, ts) and vowels (the long a, i, o, u, e and two additional diphthongs ai and au). Hausa is a tonal language with two tones: low and high, represented by the grave (e.g. "è") and acute (e.g. "é") accents respectively, which are usually not marked in everyday writing. The sentence structure follows the Subject-Verb-Object (SVO) syntax.

Igbo (`ibo`): Igbo belongs to the Benue-Congo group of the Niger-Congo language family and is spoken by over 27 million people (Eberhard et al., 2020). It is the primary language of the Igbo people, an ethnic group of southeastern Nigeria, but is also spoken in some parts of Equatorial Guinea and Cameroon. There are approximately 30 Igbo dialects, some of which are not mutually intelligible. Igbo is written using the Ọnwụ orthography (Ohiri-Aniche, 2007). Ọnwụ consists of 28 consonants and 8 vowels. Standard Igbo consists of eight vowels, and thirty consonants. Igbo is a tonal language. Tone varies by dialect but in most dialects there are three main ones: high, low and downstep. A typical Igbo sentence follows subject-verb-object (SVO) order.

Yorùbá (`yor`): Yorùbá belongs to the Yoruboid subbranch of the Volta-Niger branch of the Niger-Congo language family. The language is spoken in the southwestern parts of Nigeria stretching into some parts of Togo and Benin. The Yorùbá alphabet is based on the Latin script consisting of 18 consonants, 7 oral vowels, 5 nasal vowels and syllabic nasal consonants with additional characters likeẹ, ọ , ṣ, gb. The language uses tones: high, mid, and low tones.

Nigerian-Pidgin (`pcm`): Nigerian-Pidgin, also known as Naija, is an English-based creole language spoken as a lingua franca across regions in Nigeria. It is rooted in the Krio of the English-based creole language family with an estimate of about 40M and 80M first and second language speakers respectively. Nigerian Pidgin uses the Latin script but has no standardised orthographic representation. The phonology of the language displays no suprasegmental features such as tone as in other African languages and it makes heavy usage of loan words from African and European languages.

B. Annotation Guidelines

We adapt a sentiment annotation guide from (Mohammad, 2016) and it defines five classes:

Positive (POS) Sentiment: This occurs if a tweet implies positive sentiment, attitude and emotional state. For example, a tweet implies a positive opinion or sentiment (e.g., *"I love iPhone"*), positive emotional state (e.g., *"we won the game last night"*), expression of support (e.g., *"I will vote for PDP"*), thankfulness (e.g., *"thank god she has not been kidnapped"*), success (e.g., *"I passed all my exams"*), or positive attitude.

Negative (NEG) Sentiment: This occurs if a tweet implies negative sentiment or emotion. For example, if a tweet implies negative sentiment (e.g., *" The iPhone camera is bad"*), negative emotional states such as failure, anger, and disappointment.

[9] `https://www.ethnologue.com/guides/countries-most-languages`

[10] `https://www.mustgo.com/worldlanguages.com/hausa`

Neutral (NEU): This occurs if the user's tweet does not imply any positive or negative language directly or indirectly. These are usually factual tweets, such as news.

Mixed (MIX): This occurs if the user's tweet implies both negativity and positivity directly or indirectly. For example, *"I love an IPhone 10, but its camera is bad"*.

Indeterminate (IND): This occurs if the users' tweet does not fall into either positive, negative, neutral, and mixed, or if the annotator can only guess the class of a tweet, especially in the case of proverbs or sarcasm without sufficient context. We additionally use this class to label tweets in a different language (not code-mixed).

C. Model Hyper-parameters for Reproducibility

For the pre-trained models, we fine-tune the models using HuggingFace transformer tool (Wolf et al., 2020) with the batch size of 32, maximum sequence length of 128, number of epochs of 20, and default learning rate ($5e-5$) for all models except for XLM-R and RemBERT where we set learning rate to be $2e-5$ to ensure model convergence. All the experiments were performed Nvidia V100 and RTX 2080 GPUs.