AfriHate: A Multilingual Collection of Hate Speech and Abusive Language Datasets for African Languages

Shamsuddeen Hassan Muhammad[1,2*], **Idris Abdulmumin**[3*], **Abinew Ali Ayele**[4,5], **David Ifeoluwa Adelani**[6], **Ibrahim Said Ahmad**[2,7], **Saminu Mohammad Aliyu**[2], **Nelson Odhiambo Onyango**[8], **Lilian D. A. Wanzare**[8], **Samuel Rutunda**[9], **Lukman Jibril Aliyu**[10], **Esubalew Alemneh**[4], **Oumaima Hourrane**[12], **Hagos Tesfahun Gebremichael**[4], **Elyas Abdi Ismail**[20], **Meriem Beloucif**[13], **Ebrahim Chekol Jibril**[14], **Andiswa Bukula**[15], **Rooweither Mabuya**[15], **Salomey Osei**[16], **Abigail Oppong**[17], **Tadesse Destaw Belay**[18,19], **Tadesse Kebede Guge**[11], **Tesfa Tegegne Asfaw**[4], **Chiamaka Ijeoma Chukwuneke**[21], **Paul Röttger**[22], **Seid Muhie Yimam**[5], **Nedjma Ousidhoum**[23]

[1]Imperial College London, [2]Bayero University Kano, [3]DSFSI, University of Pretoria, [4]Bahir Dar University, [5]University of Hamburg, [6]Mila, McGill University & Canada CIFAR AI Chair, [7]Northeastern University, [8]Maseno University, [9]Digital Umuganda, [10]HausaNLP, [11]Haramaya University, [12]Al Akhawayn University, [13]Uppsala University, [14]Istanbul Technical University, [15]SADiLaR, [16]University of Deusto, [17]Independent Researcher, [18]Instituto Politécnico Nacional, [19]Wollo University, [20]Jigjiga University, [21]Lancaster University, [22]Bocconi University, [23]Cardiff University

Contact: `s.muhammad@imperial.ac.uk`, `seid.muhie.yimam@uni-hamburg.de`

Abstract

Hate speech and abusive language are global phenomena that need socio-cultural background knowledge to be understood, identified, and moderated. However, in many regions of the Global South, there have been several documented occurrences of (1) absence of moderation and (2) censorship due to the reliance on keyword spotting out of context. Further, high-profile individuals have frequently been at the center of the moderation process, while large and targeted hate speech campaigns against minorities have been overlooked. These limitations are mainly due to the lack of high-quality data in the local languages and the failure to include local communities in the collection, annotation, and moderation processes. To address this issue, we present **AfriHate**: a multilingual collection of hate speech and abusive language datasets in 15 African languages, annotated by native speakers. We report the challenges related to the construction of the datasets and present various classification baseline results with and without using LLMs. We find that model performance highly depends on the language and that multilingual models can help boost the performance in low-resource settings.[1]

Content Warning: This paper contains examples of hate speech and offensive language.

1 Introduction

> *No one is born hating another person because of the color of his skin, or his background, or his religion. People must learn to hate, and if they can learn to hate, they can be taught to love, for love comes more naturally to the human heart than its opposite.* – (Mandela, 1994)

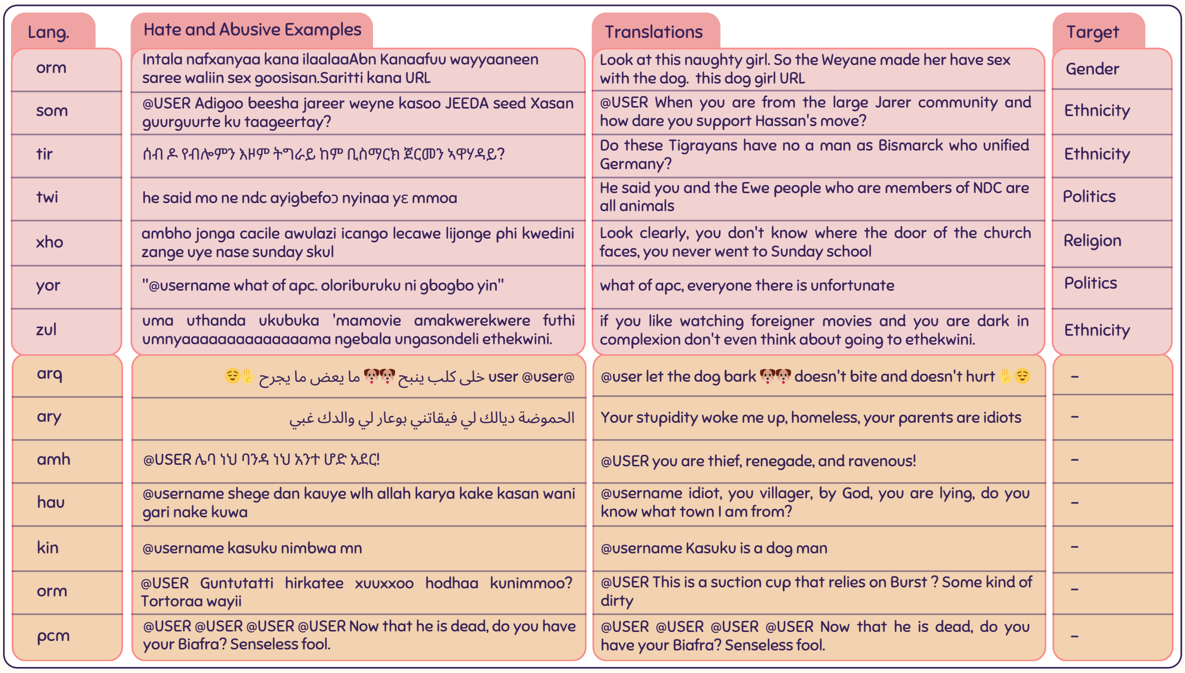
Hate speech and abusive language are global phenomena that highly depend on specific socio-cultural contexts. Although they deviate from the norm on social media, hate speech and abusive language quickly attract significant attention, spread among online communities (Mathew et al., 2019), and incite harm or violence on individuals in real life (Saha et al., 2019). Tangible efforts to address these problems must take social and cultural contexts into account (Shahid and Vashistha, 2023). However, in the absence of high-quality data or when excluding local voices from the collection and annotation processes, one may fail to build assistive tools that help address the problem and moderate such content.

Collecting hate speech and offensive language datasets is complex and time-consuming as researchers typically rely on keywords, hashtags, or user accounts to build datasets (Ousidhoum et al., 2020). They may need further insights from both moderators (Arora et al., 2023) and affected communities (Maronikolakis et al., 2022). Further, resources in languages other than English are scarce,

[*] Equal contribution

[1] The datasets, individual annotations, and hate speech and offensive language lexicons are available on `https://github.com/AfriHate/AfriHate`.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 1854–1871
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics



Lang.	Hate and Abusive Examples	Translations	Target
orm	Intala nafxanyaa kana ilaalaaAbn Kanaafuu wayyaaneen saree waliin sex goosisan.Saritti kana URL	Look at this naughty girl. So the Weyane made her have sex with the dog. this dog girl URL	Gender
som	@USER Adigoo beesha jareer weyne kasoo JEEDA seed Xasan guurguurte ku taageertay?	@USER When you are from the large Jarer community and how dare you support Hassan's move?	Ethnicity
tir	ሰብ ዶ የብሎምን እዞም ትግራይ ከም ቢስማርክ ጀርመን ኣዋሃዳይ?	Do these Tigrayans have no a man as Bismarck who unified Germany?	Ethnicity
twi	he said mo ne ndc ayigbefoɔ nyinaa yɛ mmoa	He said you and the Ewe people who are members of NDC are all animals	Politics
xho	ambho jonga cacile awulazi icango lecawe lijonge phi kwedini zange uye nase sunday skul	Look clearly, you don't know where the door of the church faces, you never went to Sunday school	Religion
yor	"@username what of apc. oloriburuku ni gbogbo yin"	what of apc, everyone there is unfortunate	Politics
zul	uma uthanda ukubuka 'mamovie amakwerekwere futhi umnyaaaaaaaaaaaaama ngebala ungasondeli ethekwini.	if you like watching foreigner movies and you are dark in complexion don't even think about going to ethekwini.	Ethnicity
arq	@user@ user خلى كلب ينبح 🐶🐶 ما يعض ما يجرح ✋😌	@user let the dog bark 🐶🐶 doesn't bite and doesn't hurt ✋😌	–
ary	الحموضة ديالك لي فيقاتني بوعار لي والدك غبي	Your stupidity woke me up, homeless, your parents are idiots	–
amh	@USER ሌባ ነህ ባንዳ ነህ አንተ ሆድ አደር!	@USER you are thief, renegade, and ravenous!	–
hau	@username shege dan kauye wlh allah karya kake kasan wani gari nake kuwa	@username idiot, you villager, by God, you are lying, do you know what town I am from?	–
kin	@username kasuku nimbwa mn	@username Kasuku is a dog man	–
orm	@USER Guntutatti hirkatee xuuxxoo hodhaa kunimmoo? Tortoraa wayii	@USER This is a suction cup that relies on Burst ? Some kind of dirty	–
pcm	@USER @USER @USER @USER Now that he is dead, do you have your Biafra? Senseless fool.	@USER @USER @USER @USER Now that he is dead, do you have your Biafra? Senseless fool.	–

Table 1: **Examples of hateful and abusive instances in AFRIHATE.** All hateful posts are assigned targets.

especially for low-resource languages. To bridge this gap, we present **AFRIHATE** a collection of new hate speech and abusive language Twitter datasets in 15 languages spoken in various African regions: Algerian Arabic, Amharic, Igbo, Kinyarwanda, Hausa, Moroccan Arabic, Nigerian Pidgin, Oromo, Somali, Swahili, Tigrinya, Twi, isiXhosa, Yorùbá, and isiZulu. The datasets are annotated by native speakers and include three classes: **hate**, **abusive/offensive**, or **neutral**–neither hateful nor abusive. The targets of the hateful tweets were further labeled based on six common attributes used to discriminate against people: **ethnicity**, **politics**, **gender**, **disability**, **religion**, or **other**. Table 1 shows a sample of the datasets in various languages.

We report the data collection and annotation strategies and challenges when building **AFRIHATE**, present various classification baselines with and without using LLMs, and discuss the results. We find that model performance highly depends on the language and that multilingual models can help boost the performance in low-resource settings. We publicly release the datasets and individual labels, in addition to manually-curated hate speech and offensive language lexicons. These provide a valuable foundation for the research community interested in hate speech and abusive language, African languages, and researchers interested in studying disagreements.

2 Related Work

The fast-spreading nature of hate speech and abusive language have been at the center of a significant amount of NLP work in recent years (Talat and Hovy, 2016; Vigna et al., 2017; Basile et al., 2019; Mansur et al., 2023). However, as there is no unanimous definition of hate speech, researchers have adopted different ones when building resources. For instance, some studies define hate speech as any speech that can cause danger or harm to disadvantaged groups (Davidson et al., 2017), others focus on whether the speech is intended to promote hatred (Gitari et al., 2015), or whether it dehumanises protected groups (Vidgen et al., 2021), which leads to various challenges such as the lack of generalisability (Yin and Zubiaga, 2021).

Despite Africa being home to more than 2,000 languages and the increasing interest in building hate speech and offensive languages resources for non-English languages (Ousidhoum et al., 2019; Röttger et al., 2022; Madeddu et al., 2023), few datasets focus on African languages, such as Amharic (Ayele et al., 2024, 2023, 2022), Afaan Oromo (Ababu and Woldeyohannis, 2022), Yorùbá (Ilevbare et al., 2024), Hausa (Vargas et al., 2024; Adam et al., 2023), and Nigerian Pidgin (Ndabula et al., 2023; Ilevbare et al., 2024; Aliyu et al., 2022;

Language	Code	Subregion	Spoken in	Script
Algerian Arabic/Darja	arq	North Africa	Algeria	Arabic
Amharic	amh	East Africa	Ethiopia, Eritrea	Ethiopic
Hausa	hau	West Africa	Northern Nigeria, Niger, Ghana, and Cameroon,	Latin
Igbo	ibo	West Africa	Southeastern Nigeria	Latin
Kinyarwanda	kin	East Africa	Rwanda	Latin
Moroccan Arabic/Darija	ary	North Africa	Morocco	Arabic/Latin
Nigerian Pidgin	pcm	West Africa	Nigeria, Ghana, Cameroon,	Latin
Oromo	orm	East Africa	Ethiopia, Kenya, Somalia	Latin
Somali	som	East Africa	Somalia, Ethiopia, Djibouti, Kenya	Latin
Swahili	swa	East Africa	Kenya, Tanzania, Uganda, DR Congo, Rwanda, Burundi, Mozambique	Latin
Tigrinya	tir	East Africa	Ethiopia, Eritrea	Ethiopic
Twi	twi	West Africa	Ghana	Latin
Xhosa	tso	Southern Africa	Mozambique, South Africa, Zimbabwe, Eswatini	Latin
Yorùbá	yor	West Africa	Southwestern and Central Nigeria, Benin, and Togo	Latin
Zulu	zul	Southern Afric	Southern Africa	Latin

Table 2: **Information about the AFRIHATE languages**: their ISO codes, subregions, countries in which they are mainly spoken, and the writing scripts included in AFRIHATE.

Tonneau et al., 2024). Moreover, most resources adopt a binary labeling scheme (hate/offensive) (e.g., (Aliyu et al., 2022)), or do not add the target attributes (e.g., (Ilevbare et al., 2024)). Other work (Tonneau et al., 2024) relies on active learning for annotating some data instances, which is not ideal when labeling hate speech for under-resourced African languages. That is, when focusing on underrepresented cultures ML models that deal with hate speech tend to be culturally insensitive (Lee et al., 2023, 2024).

Furthermore, a limited number of studies involve African communities in the dataset creation process (Adelani et al., 2021; Maronikolakis et al., 2022; Abdulmumin et al., 2024). We take a step towards addressing this problem by developing 15 new datasets for hate and offensive speech in various languages spoken across the African continent.

3 Creating AFRIHATE

AFRIHATE covers 15 languages from various African regions. In Table 2, we report the scripts of these languages and the main regions where they are spoken. The collection includes tweets from 2012 to 2023 collected using the Academic API before the suspension of free academic access. The API allowed us to collect up to 10 million tweets per month, and Twitter/X is a commonly used platform in African countries with documented cases of hate speech propagation (Adjai and Lazaridis, 2013; Egbunike et al., 2015; Oriola and Kotzé, 2020; Ridwanullah et al., 2024; Raborife et al., 2024).

3.1 Data Collection

Except for Amharic and Tigrinya, the Twitter API does not support African languages, which makes the data collection challenging. We, therefore, follow strategies adopted by previous work such as Muhammad et al. (2022, 2023) and use various heuristics based on hate speech and abusive language keywords, user handles, stopwords, hashtags, and locations. Table 3 shows the number of keywords used for data collection in each language.

Since interpreting hate and abusive content heavily depends on understanding political and socio-cultural contexts, we have adopted language-dependent collection and annotation strategies. As previous work that relied on specific keywords to collect data (e.g., Talat and Hovy, 2016; Basile et al., 2019), we follow an analogous strategy. Similarly to Ousidhoum et al. (2019), we use a larger set of keywords and include culture-specific controversial topics in the lists. Despite the diversity of topics and the large sizes of the lists (see Table 3), an initial pre-annotation phase revealed a limited number of hateful tweets for some languages such as Nigerian-Pidgin and Hausa. Therefore, we used additional heuristics to collect more tweets: 1) keyword crowd-sourcing, 2) manual data collection, and 3) using existing datasets, as we explain in the following.

Crowd-sourcing Keywords To crowd-source additional keywords, we first asked native speakers to provide us with a list of hateful, abusive, or controversial keywords. Then, we contacted social media influencers, who asked their followers to

	amh	arq	ary	hau	ibo	kin	oro	pcm	som	swa	tir	twi	xho	yor	zul
Hate	88	62	41	36	46	264	126	23	45	24	58	20	42	68	31
Abusive	74	40	0	149	118	362	159	26	67	12	66	86	177	109	118

Table 3: **Number of keywords used for data collection.** For Algerian Arabic (arq), the list includes controversial topics that are neither hateful nor offensive.

share abusive or hateful keywords in their local languages by filling in a form created for anonymous collection. This helped us diversify the lists as the followers come from various backgrounds.

We also used off-the-shelf curated hate speech lexicons from PeaceTech Lab[2] and Hatebase.[3] The lists were post-processed by native speakers prior to the collection of the tweets.

Manual Data Collection For Kinyarwanda and Twi, native speakers manually collected all the tweets using a combination of keywords and user handles. We curated a list of user handles of public figures who frequently post hateful or abusive content, and collected tweets from their profiles.

Using Existing Datasets For Nigerian Pidgin, we used a subset of an additional existing hate speech dataset proposed by Tonneau et al. (2024). Since some instances in this dataset were annotated using active learning, we re-annotated those labeled as hateful or offensive. Similarly, for Swahili, we re-annotated the negative instances from the sentiment analysis dataset introduced by Muhammad et al. (2023), the misinformation dataset by Amol et al. (2023), and the hate speech one by Ombui et al. (2019) into our predefined classes (hate, offensive and normal).

We further labeled the targets of the re-annotated tweets into our target attribute categories, i.e., disability, ethnicity, gender, politics, religion, and others.

3.2 Data Processing

We further cleaned the collected tweets and removed retweets, tweets containing less than three words, duplicates, URLs, invisible characters, and redundant white spaces. We converted the tweets written in Latin script to lowercase and anonymised the tweets by replacing @mentions with a placeholder @*user*.

3.3 Language Identification

We collected the tweets using location and keywords. This is particularly challenging for African languages since people within one location can speak different languages. That is, keywords and hashtags may appear in more than one language, which makes the data selection more difficult.

Open-source and closed-source language identification (LID) tools used in previous studies (Muhammad et al., 2022, 2023) often show low accuracy in African languages, especially when used in social media posts. This is largely due to the unique linguistic characteristics of these languages, such as the common usage of code-mixing and digraphia, i.e., a language can be written in more than one script.

To address these limitations, we built a LID model that improves the identification performance for social media text data in our target languages. It achieves this through continued pretraining of AfroXLMR (Alabi et al., 2022) on Glot500-c corpus which covers 511 predominantly low-resource languages (Imani et al., 2023). We further finetuned the model on AfriSenti-LID dataset, which focuses on social media data and covers most of the languages included in AFRIHATE and a similar Twitter-sphere. Our LID tool and its documentation can be found on our project page.[4]

3.4 Data Annotation

3.4.1 Pre-Annotation and Data Selection

We randomly sampled tweets in each language and conducted a pilot annotation, which showed a large class imbalance despite collecting tweets using abusive and hateful keywords. For instance, most tweets in Hausa were neutral (neither hateful nor abusive) due to keywords carrying multiple meanings depending on the region where the word is used, i.e., it can have a neutral connotation in some parts of Nigeria. For example, the word *Aboki* means “friend” in Northern Nigeria, while it can be an insult in the Southern part of the country.

[2] https://www.peacetechlab.org/the-peacetech-toolbox

[3] https://hatebase.org

[4] https://github.com/hausanlp/AfriLID

Language	amh	arq	ary	hau	ibo	kin	oro	pcm	som	swa	tir	twi	xho	yor	zul
Manually Collected	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗
Pre-Annotation	✗	✗	✓	✓	✓	✓	✗	✓	✗	✗	✗	✓	✓	✗	✓
Total Annotators	11	7	3	3	6	3	9	3	7	5	8	3	3	4	3
Annotators per Instance	4	3	3	3	3	3	4	3	4	3	4	3	3	4	3
Free-Marginal Multirater Kappa ↑	0.63	0.68	0.61	0.75	0.80	0.81	0.63	0.65	0.46	0.55	0.46	0.75	0.62	0.68	0.81

Table 4: **Collection and annotation details** for AFRIHATE. The table shows if the data was manually collected, whether a pre-annotation step was conducted, the total number of annotators, the number of annotators per instance, and the inter-annotator agreement (Free-Marginal Multirater Kappa).

Class	amh	ary	arq	hau	ibo	kin	oro	pcm	som	swa	tir	twi	xho	yor	zul
Hate	2,246 (45.3%)	162 (13.0%)	674 (14.5%)	343 (5.2%)	251 (5.0%)	1,268 (26.8%)	2,293 (45.6%)	1,177 (11.1%)	388 (8.3%)	3,974 (18.8%)	2,940 (58.0%)	443 (11.4%)	210 (5.7%)	150 (3.1%)	147 (3.4%)
Abusive	1,353 (27.3%)	778 (62.2%)	2,270 (49.0%)	2,336 (35.2%)	3,510 (70.2%)	1,146 (24.3%)	667 (13.2%)	5,238 (49.4%)	1,404 (30.1%)	7,708 (36.6%)	1,070 (21.1%)	3,278 (84.0%)	1,550 (42.1%)	2,655 (54.4%)	1,839 (42.7%)
Neutral	1,359 (27.4%)	310 (24.8%)	1,690 (36.5%)	3,965 (59.6%)	1,242 (24.8%)	2,308 (48.9%)	2,072 (41.2%)	4,184 (39.5%)	2,868 (61.6%)	9,410 (44.6%)	1,062 (20.9%)	180 (4.6%)	1,923 (52.2%)	2,074 (42.5%)	2,322 (53.9%)

Table 5: **Number of instances per class in each dataset.** Percentages are shown below each absolute count. In total, AFRIHATE contains **90,437** instances across 15 languages.

As this would have led to an insufficient number of instances in each target class, we included a **pre-annotation** phase to ensure each class covered a reasonable percentage of the data.

During the pre-annotation, we provided the annotators with a distinct pool of tweets and asked them to select those likely to be hateful or abusive. We then aggregated the pre-selected tweets, and multiple annotators labeled them. Table 4 indicates the languages for which we conducted a pre-annotation step, i.e., only those for which we observed a significantly high imbalance during the pilot annotation.

3.4.2 Recruiting Annotators

The unavailability of annotators for African languages on common platforms like Amazon Mechanical Turk and Prolific makes traditional crowd-sourcing methods impractical. As Kirk et al. (2023) demonstrated that trained annotators achieve higher quality results, we trained native speakers and recruited them as annotators. For each language, we also recruited a native speaker as a language lead who would control for the quality of the annotations. We used Label Studio as an annotation platform[5] and an adapted version of the Potato annotation tool.[6]

3.4.3 Annotation Task

Labels We provided the annotators with thorough guidelines (see Figure 1 in the Appendix). We asked the annotators to choose one of three categories: **hate**, **abusive/offensive**, or **neutral**. The latter means that the tweet is neither hateful nor abusive. Tweets spotted in a language different from the target one were labeled **Indeterminate** and were later excluded from the final dataset.

Similarly to Ayele et al. (2024); Ousidhoum et al. (2019); Fortuna et al. (2019), annotators had to select the target(s) of the hateful tweets. That is, the common attribute(s) based on which the tweet is discriminating against people: **Ethnicity**, **Politics**, **Gender**, **Disability**, **Religion**, or **Other**. We do not include targets for offensive and abusive tweets as these can often be generic and directed towards an individual as previously reported by Zampieri et al. (2019) (e.g., see the Algerian Arabic example in Table 1). Details about the targets can be found in Figure 1 in the Appendix. For some languages such as Hausa, we asked the annotators to spot which words made them label the tweet hateful or abusive.

Final label selection As shown in Table 4, for languages where we conducted a pre-annotation step, each tweet was annotated by 3 annotators, leading to a total of 4 labels per tweet with the pre-annotation label counting as one. On the other hand, for instances in datasets for which we did not carry out a pre-annotation step, 3 to 4 annotators

[5] https://labelstud.io/

[6] https://github.com/davidjurgens/potato

Hate Target	amh	arq	ary	hau	ibo	kin	oro	pcm	som	swa	tir	twi	xho	yor	zul
Disability	0	0	26	0	0	6	0	1	0	112	0	32	3	2	0
Ethnicity	902	269	150	32	383	94	462	537	74	2,799	573	245	94	103	241
Gender	15	2	6	2	3	24	12	50	17	117	0	12	111	13	0
Politics	1,247	118	63	0	6	1,096	1,550	87	703	233	2,251	32	0	33	0
Religion	130	6	207	38	0	10	31	105	19	336	16	14	0	22	0
Others	199	0	0	0	0	0	767	15	208	501	85	14	0	76	12

Table 6: **Data distribution of hate speech targets in AFRIHATE**.

Split	amh	ary	arq	hau	ibo	kin	oro	pcm	som	swa	tir	twi	xho	yor	zul
Train	3,467 (69.9%)	3,240 (69.9%)	716 (57.3%)	4,566 (68.7%)	3,419 (68.2%)	3,302 (69.9%)	3,517 (69.8%)	7,416 (70.0%)	3,174 (68.1%)	14,760 (70.0%)	3,547 (69.9%)	2,564 (65.7%)	2,502 (67.9%)	3,336 (68.4%)	2,940 (68.2%)
Dev	744 (15.0%)	695 (15.0%)	211 (16.9%)	1,029 (15.5%)	774 (15.4%)	706 (15.0%)	763 (15.1%)	1,590 (15.0%)	741 (15.9%)	3,164 (15.0%)	760 (15.0%)	639 (16.4%)	559 (15.2%)	724 (14.8%)	640 (14.9%)
Test	747 (15.1%)	699 (15.1%)	323 (25.8%)	1,049 (15.8%)	821 (16.4%)	714 (15.1%)	759 (15.1%)	1,593 (15.0%)	745 (16.0%)	3,168 (15.0%)	765 (15.1%)	698 (17.9%)	622 (16.9%)	819 (16.8%)	728 (16.9%)

Table 7: **Number of instances included in the training (train), development (dev), and test splits** of the different datasets with percentages shown below each absolute count.

were assigned to each tweet, and the final gold label was determined by majority voting, i.e., two out of three labels or three out of four labels.

Table 5 and Table 6 show the final number of instances per class and the target distributions for all the languages.

3.4.4 Inter-Annotator Agreement

To assess inter-annotator agreement (IAA), we computed the free marginal Randolph's Kappa score (Randolph, 2005) for each dataset. Table 4 shows the IAA scores for all the datasets, the total number of annotators in each, and the number of annotators per instance. The IAA scores range from 0.46 to 0.81, indicating medium to high agreement levels. The highest agreement scores are reported for kin, and twi, which can be due to the manual collection of only potentially abusive and hateful tweets. For other languages such as hau and zul, the high agreement can be attributed to the pre-annotation step, which helped us filter tweets that were later annotated.

3.5 Dataset Statistics

As shown in Table 5, most datasets are imbalanced, and the ***hate*** class includes fewer instances in 9 out of 15 languages. The variations in the class distributions are due to the differences between the languages and the data collection techniques. Further, the target distributions also differ because of socio-cultural characteristics related to local politics, social dynamics, and an unavoidable degree of selection bias (Ousidhoum et al., 2020).

Data Splits We split the AFRIHATE datasets based on the various label distributions. As reported in Table 7, each test set includes a minimum of 100 instances in each class (i.e., hate, abusive, and normal). This guarantees a more robust evaluation of the different models.

4 Experiments

4.1 Setup

We compare the performance of three main sets of approaches on the AFRIHATE datasets:

1. Fine-tuning a BERT-based pre-trained language model (PLM),
2. Few-shot learning with SetFit with BERT-based PLM (Tunstall et al., 2022): a few-shot approach using BERT-like PLMs,
3. Prompting large language models (LLMs) in zero and few-shot settings.

Fine-tuning PLMs We use four widely adopted Africa-centric PLMs that have been shown to consistently perform better on African languages than massively multilingual PLMs such as XLM-R (Conneau et al., 2020). The models are AfriBERTa-large (Ogueji et al., 2021), AfriTeVa V2 base (Oladipo et al., 2023), AfroXLMR (Alabi et al., 2022) and AfroXLMR-76L (Adelani et al., 2024). Each model was trained for 20 epochs over 5 runs with a batch size of 32, maximum sequence length of 128, and a learning rate of $5e-5$ except for AfroXLMR-76L where we used a learning rate of $3e-5$. The rest of the hyperparameters were the default values set in the HuggingFace fine-tuning

Model	amh	ary	arq	hau	ibo	kin	oro	pcm	som	swa	tir	twi	xho	yor	zul	avg.
								Monolingual Fine-Tuning								
AfriBERTa	69.54	67.93	30.48	82.28	89.53	79.43	73.43	66.90	65.52	91.36	73.07	74.54	81.07	72.37	83.75	72.33
AfriTeVa V2	73.91	76.71	25.25	79.06	83.95	77.60	71.61	68.69	69.65	90.68	72.36	64.96	54.67	**79.88**	69.05	68.73
AfroXLMR	70.65	80.16	61.18	81.93	89.30	**80.72**	72.11	67.98	66.84	91.44	74.52	77.17	82.49	72.15	83.44	76.15
AfroXLMR-76L	74.36	80.05	53.52	**82.78**	89.59	79.58	76.63	68.38	71.09	**91.72**	76.27	76.65	84.40	72.35	84.65	76.45
								Multilingual Fine-Tuning								
AfroXLMR-76L	**75.25**	**80.76**	**63.31**	82.20	**89.85**	79.56	**77.62**	**69.20**	**72.26**	91.22	**77.55**	**78.68**	**86.83**	74.32	**86.81**	**78.16**

Table 8: Model performances after fine-tuning BERT-based LMs. The best performance for each language is highlighted in **bold**.

pipeline for text classification.

SetFit Few-shot Learning SetFit is a few-shot learning approach based on sentence-transformer models like LabSE (Feng et al., 2022). It works by, first, fine-tuning a pre-trained sentence transformer model on a few examples in a contrastive manner. Then, the resulting model is used to generate rich text embeddings, which are used to train a classification head. We used LaBSE to train classifiers with the following configurations:

1. for **zero-shot learning**, we trained the transformers for one epoch using the dummy dataset generated by the framework (2 x [This sentence is {Label}, ...]), where {Label} can be **neutral**, **abusive/offensive** or **hate**;
2. for **few-shot learning**, we trained each model for three epochs using 5, 10, and 20 shots. All the classifiers were trained using a batch size of 32.

Prompting LLMs We prompt one closed model (GPT-4o) and nine open models of various model sizes (0.4B to 70B). The open models are: InkubaLM-0.4B (Tonja et al., 2024), mT0-small (Muennighoff et al., 2023), BLOOMZ 7B (Scao et al., 2022), Mistral 7B (Jiang et al., 2023), Aya-23-35B (Aryabumi et al., 2024), LLaMa 3.1 {8B & 70B} (Dubey et al., 2024), and Gemma 2 {9B & 27B} (Team et al., 2024). All the models were prompted using five prompt templates with clear definitions of the **abusive/offensive**, **hate**, and **neutral** categories. For hate speech, we used the definition adopted by the United Nations and another one from the Merriam Webster's dictionary. We report the average scores across all the templates in Section 4.2. The full prompts can be found in Appendix A.1.

4.2 Experimental Results

Monolingual vs. Multilingual Fine-tuning We compare monolingual fine-tuning where we train on a language and evaluate on the same language to multilingual fine-tuning—where we combine the training data of all the languages and evaluate the results for each.

Table 8 shows the results of the fine-tuning experiments. We find that encoder-only models perform better than the T5-style models, i.e., AfriTeVa V2. On average, AfroXLMR-76L achieved the best performance, most likely due to the fact that it was pre-trained on all the languages included in AFRIHATE. While AfroXLMR was not pre-trained on some languages such as Tigrinya (tir) and Twi (twi), it still achieves performance that is comparable to AfroXLMR-76L. AfriBERTa generally struggles with Arabic dialects such as Algerian Darja (arq) and Moroccan Darija (ary) as the Arabic script was not included in its pre-training.

Overall, multilingual fine-tuning of the multilingual AfroXLMR model leads to the best results on 11 out of 15 languages, and comparable results on the remaining languages except for Yorùbá (yor), where AfriBERTa led to the best result likely because of its Africa-centric tokenizer.

Table 9 shows the per-class accuracy across different languages. Overall, multilingual models perform better for languages with a low percentage of the "hate" category in the training data (e.g., < 200) such as ary, xho, yor, and zul with an F-score improvement of $+1.7$, $+5.7$, $+4.9$, and $+10.8$, respectively.

Zero-shot vs. Few-shot Settings Table 10 shows the results of both zero-shot and few-shot experiments. SetFit performs slightly better than all open LLMs in zero-shot settings (36.9), and GPT-4o leads to the best overall performance with 61.9 F1 points.

When considering the few-shot settings, 5-shot models show the biggest boost in performance, where Gemma-2-9B and other bigger models (Gemma 2 27B and LLaMa 3.1 70B) improve by

Lang.	**Monolingual AfroXLMR-76L**				**Multilingual AfroXLMR-76L**				**GPT-4o** (20 shots)			
	Macro F1	**Abuse**	**Hate**	**Neutral**	**Macro F1**	**Abuse**	**Hate**	**Neutral**	**Macro F1**	**Abuse**	**Hate**	**Neutral**
amh	73.83	70.20	77.27	74.02	**75.55**	**71.92**	**78.18**	**76.54**	65.70	59.21	67.91	70.00
ary	78.01	**86.81**	66.98	80.25	**79.05**	86.59	**68.66**	81.91	75.93	84.29	61.09	**82.43**
arq	57.05	77.35	27.42	66.38	68.99	**78.29**	55.95	**72.73**	**73.67**	77.63	**71.43**	71.96
hau	**81.53**	**77.60**	**80.45**	**86.54**	78.05	76.45	72.83	84.88	59.44	70.61	40.77	66.95
ibo	88.00	92.18	**91.18**	80.64	**88.33**	**92.41**	**91.18**	**81.40**	76.85	84.13	75.37	71.04
kin	**77.96**	**70.09**	80.87	**82.90**	77.11	70.00	78.90	82.43	74.27	65.45	**82.01**	75.37
orm	**70.07**	47.19	**81.17**	**81.86**	69.93	**48.89**	80.11	80.78	65.30	44.12	72.78	79.00
pcm	**65.40**	**71.09**	52.82	72.31	63.71	69.66	49.10	72.39	63.53	57.38	**56.67**	**76.54**
som	59.53	67.00	30.77	80.83	60.91	**68.81**	32.94	**81.00**	**62.94**	63.85	**44.91**	80.05
swa	88.00	**92.18**	**91.18**	80.64	**89.50**	89.35	88.33	**90.83**	83.29	85.48	80.49	83.91
tir	72.32	72.29	82.60	62.07	**74.45**	**75.23**	**84.97**	**63.16**	56.23	51.73	64.38	52.59
twi	58.30	**91.45**	**61.40**	22.05	**63.66**	91.04	60.18	39.76	62.64	85.90	54.45	**47.58**
xho	79.37	**86.33**	66.23	85.53	**81.57**	85.14	**71.95**	**87.63**	57.74	70.82	38.46	63.94
yor	57.36	**84.24**	6.90	80.95	59.41	84.03	11.76	**82.44**	**71.48**	81.56	**53.01**	79.88
zul	80.97	**87.54**	68.79	86.57	**84.34**	85.46	**79.55**	**88.01**	70.63	74.60	67.78	69.51

Table 9: Macro F1 scores for monolingual and multilingual AfroXLMR-76L [only one run] vs. GPT-4o [prompt template 1; 20 shots]. The best performance for each language is highlighted in **bold**.

Model	**# Shots**	amh	ary	arq	hau	ibo	kin	oro	pcm	som	swa	tir	twi	xho	yor	zul	**Avg.**
SetFit	0	33.79	46.54	27.4	23.57	36.90	25.29	26.26	43.46	32.18	44.38	49.30	51.13	37.38	48.46	35.08	37.41
	5	44.89	33.82	36.07	49.93	51.52	55.35	36.56	53.28	48.09	54.90	35.36	33.23	47.49	34.36	33.18	43.20
	10	49.42	35.81	46.55	49.72	39.31	53.10	39.10	57.00	42.29	62.64	39.16	45.76	43.52	49.73	40.33	46.23
	20	50.13	40.49	54.24	55.40	65.15	57.35	40.91	59.09	48.95	74.93	40.04	53.72	46.92	56.54	50.67	52.97
Mistral-7B-v0.1	0	21.18	17.46	11.98	6.84	8.04	15.18	20.82	6.76	8.16	18.88	25.62	8.38	9.72	8.22	8.92	13.08
	5	37.68	31.60	44.54	34.90	36.84	46.98	39.46	50.00	34.12	58.58	35.64	32.04	34.24	41.04	32.54	39.35
	10	39.68	36.50	50.56	37.82	36.94	45.72	39.02	52.10	31.82	63.18	35.84	30.82	32.64	42.90	32.46	40.53
	20	36.64	35.46	52.94	38.44	38.98	52.28	39.24	54.60	33.28	66.28	37.42	30.44	34.12	46.40	35.00	42.10
aya-23-35B	0	17.04	19.20	17.20	20.34	14.04	21.34	20.38	20.02	18.56	34.04	17.10	10.68	20.32	17.48	20.24	19.20
	5	37.78	56.40	57.00	39.70	42.70	48.78	39.24	57.90	38.72	69.84	35.68	40.24	36.76	48.88	44.58	46.28
	10	38.50	53.96	61.82	44.42	45.80	52.82	42.16	59.26	37.60	74.10	35.88	39.92	36.18	48.44	45.00	47.72
	20	38.10	52.90	63.76	46.14	46.52	57.62	42.26	61.68	38.68	75.68	38.12	38.54	37.74	51.52	44.52	48.92
Gemma-2-9B	0	27.42	28.36	25.46	14.48	15.00	24.08	23.74	24.74	10.92	37.12	25.52	19.50	13.40	21.66	12.52	21.59
	5	56.60	57.68	61.14	49.98	48.12	60.26	45.10	60.30	46.42	74.78	48.40	46.50	35.82	55.96	46.60	52.91
	10	57.84	61.08	61.72	56.62	53.78	63.54	45.78	60.54	51.72	78.16	45.48	48.16	38.82	60.22	52.58	55.74
	20	60.96	59.94	64.38	55.90	54.56	64.14	44.70	62.54	52.22	79.24	44.94	48.92	39.02	60.50	53.68	56.38
Gemma-2-27B	0	41.78	41.24	41.12	28.96	31.36	42.08	33.46	49.78	31.10	59.88	30.84	27.74	25.64	41.46	28.02	36.96
	5	59.62	64.14	65.62	54.62	56.14	61.08	46.76	61.78	54.12	81.18	52.12	47.26	40.48	61.72	54.28	57.39
	10	60.70	61.36	64.88	58.90	56.84	64.86	46.96	61.60	52.86	81.94	49.82	50.86	41.90	59.66	56.94	58.01
	20	62.28	59.86	65.80	59.60	58.76	66.06	48.90	63.24	52.90	82.98	53.36	50.34	43.08	59.10	56.88	58.88
Llama-3.1-70B	0	36.34	43.34	42.64	35.64	32.52	38.52	31.54	48.66	31.14	60.14	27.08	25.54	28.98	40.64	35.50	37.21
	5	58.52	66.24	62.88	52.86	52.88	58.14	45.50	64.38	52.72	75.76	44.42	43.90	39.02	58.08	55.24	55.37
	10	61.18	64.46	62.20	56.40	55.10	59.00	46.60	63.58	54.40	78.74	49.62	45.76	39.04	57.96	55.74	56.65
	20	60.38	61.36	63.80	57.40	53.80	62.48	49.02	63.18	51.82	80.02	53.90	47.50	40.34	57.94	56.36	57.29
GPT-4o	0	61.78	66.41	73.75	56.91	68.82	62.53	60.01	65.94	63.42	73.66	45.75	52.72	58.92	75.21	54.47	62.69
	5	66.73	73.53	77.33	55.44	76.73	72.27	70.94	66.29	59.33	80.95	61.11	73.21	60.45	76.98	59.34	68.71
	10	67.94	75.54	77.54	58.15	80.08	75.07	72.27	67.14	62.75	84.19	57.52	72.28	65.75	76.74	65.05	70.53
	20	68.08	76.16	78.69	58.34	80.81	74.86	72.33	65.41	66.44	84.61	59.55	75.86	66.08	77.11	71.36	71.71
Language avg.	-	48.32	50.74	54.04	44.91	47.79	52.88	43.18	54.44	43.10	67.53	41.95	42.53	39.06	51.25	44.18	48.39

Table 10: **Model performance (Macro F1-score) for zero- and-few shot classifiers** across the different languages in AFRIHATE. The best performance for each language is highlighted in **bold**. This is an average over 5 prompt templates.

about +20 points. The performance boost with additional 10 and 20 shots is more limited for LLMs (+3.0 improvement), whereas SetFit consistently benefits from additional examples. The best results reached for closed models are at 20-shots, where GPT-4o achieved an overall F1-score of 70.8 while Gemma 2 27B achieved the best overall results for any open model with an F1-score of 57.2.

In our performance analysis per different classes shown in Table 9, GPT-4o generally performs worse than full fine-tuning in monolingual or multilingual settings. However, we find that it achieves significantly better performance for hate detection in a few languages such as arq (+44.0), som (+14.9), and yor (+46.1) compared to monolingual fine-tuning. For languages without enough training data for the hate category such as Yorùbá, prompting LLMs might provide a better detection of this class compared to fine-tuning BERT-like PLMs.

Overall Results Our results show that fine-tuning multilingual models leads to a better performance for the majority of the AFRIHATE languages. That is, AfroXLMR-76L achieves an average macro F1 score of **78.16**. As for zero-shot and few-shot settings, **GPT-4o** outperformed other models, with average F1 scores of **61.89** and **70.79** in zero-shot and 20-shot settings, respectively.

As a whole, these results highlight the advantages of multilingual and context-specific models in hate and abusive language detection for African languages.

5 Conclusion

We introduced AFRIHATE, the first large-scale collection of hate and abusive language datasets in 15 African languages: Algerian Arabic, Amharic, Igbo, Kinyarwanda, Hausa, Moroccan Arabic, Nigerian Pidgin, Oromo, Somali, Swahili, Tigrinya, Twi, Xhoza, Yorùbá, and Zulu. The datasets were annotated by native speakers as hate speech, abusive, or neutral. We discussed our data collection strategies and highlighted the challenges faced during the data collection and annotation. We then reported baseline experiments using Africa-centric pre-trained language models as well as prompted open and closed LLMs showing a large gap in the performance across languages.

AFRIHATE is a first step towards building high-quality hate speech resources for African languages. We publicly release all the datasets, scripts, models, and lexicons to the research community.

Limitations

While we collected AFRIHATE using large sets of keywords, we acknowledge the unavoidable presence of selection bias (Ousidhoum et al., 2020) as no dataset can capture the full range of hate speech contexts across various languages and cultures. In addition, although we recruited annotators who come from different socio-cultural backgrounds, opinions on hate speech remain subjective and one cannot include all possible perspectives of what constitutes hate speech or abuse. We mitigate the problem by sharing the individual annotations with the research community studying the problem.

Further, when using language identification to collect data, challenges due to code-mixing and digraphia, make the task non-trivial given the common usage of multilinguality in African languages. We address the problem by asking the annotators to flag any tweet that is not in the target language. We acknowledge, nevertheless, instances that may have been missed by our annotators.

Finally, we report on the various dataset statistics and model features. However, given the fact that we use some closed models in our experiments, and the class imbalance problem which is inherent to hate speech datasets, we do not claim that our results are fully replicable or generalisable.

Ethical Considerations

Annotators The annotators involved in this study were compensated for their work by more than the minimum wage and any demographic information about them was shared with consent. We acknowledge the difficulty of annotating hate speech and abusive language on people's well-being. Therefore, the annotators could reach out to us and were allowed to quit at any time.

Language Use Our datasets focus on hate speech and abusive languages in 15 African languages. However, we do not claim that they represent the full usage of these languages. We further acknowledge the socio-cultural biases that can come with the data as views on hate highly differ from one person to another and those shared by our annotators cannot include all possible perspectives.

Intended uses and potential misuses Our datasets focus on hate speech and abusive language. They present a first step towards studying the phenomenon in some low-resource African languages. However, as malicious data actors can misuse our resources, we follow the suggestions made by Schlichtkrull et al. (2023) for automated fact-checking researchers and clearly state the following:

- Models built using our datasets **should not be used for automated removal**.
- Our *data subjects* are social media users.
- Our *data actors* and *model owners* should be users, moderators, experts, and researchers with background knowledge in the field, especially on the limitations of automated hate speech and abusive language detection models.
- Given the sensitivity of the task and the high risk of false positives, any constructed or

deployed model using our data should be **human-in-the-loop** with the humans being native or near-native speakers.

Acknowledgments

We are grateful for the work and insights shared by our annotators. 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 also thank Nigerian Artificial Intelligence Research Scheme (NAIRS) for supporting this project. We thank Label Studio for providing the academic version annotation tool used in this project. We also thank OpenAI for providing API credits to Masakhane. Shamsuddeen acknowledges the support of Google DeepMind. Paul was supported by a MUR FARE 2020 initiative under grant agreement Prot. R20YSMBZ8S (INDOMITA) and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 949944, INTEGRATOR). Abinew acknowledges the support of LT group, University of Hamburg.

References

Teshome Mulugeta Ababu and Michael Melese Woldeyohannis. 2022. Afaan oromo hate speech detection and classification on social media. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 6612–6619.

Idris Abdulmumin, Sthembiso Mkhwanazi, Mahlatse Mbooi, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Neo Putini, Miehleketo Mathebula, Matimba Shingange, Tajuddeen Gwadabe, and Vukosi Marivate. 2024. Correcting FLORES evaluation dataset for four African languages. In *Proceedings of the Ninth Conference on Machine Translation*, pages 570–578, Miami, Florida, USA. Association for Computational Linguistics.

Fatima Muhammad Adam, Abubakar Yakubu Zandam, and Isa Inuwa-Dutse. 2023. Detection of offensive and threatening online content in a low resource language. *arXiv preprint arXiv:2311.10541*.

David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. 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.

Carol Adjai and Gabriella Lazaridis. 2013. Migration, xenophobia and new racism in post-apartheid south africa. *International journal of social science studies*, 1:192–205.

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.

Saminu Mohammad Aliyu, Gregory Maksha Wajiga, Muhammad Murtala, Shamsuddeen Hassan Muhammad, Idris Abdulmumin, and Ibrahim Said Ahmad. 2022. Herdphobia: a dataset for hate speech against fulani in nigeria. *arXiv preprint arXiv:2211.15262*.

Cynthia Amol, Lilian Wanzare, and James Obuhuma. 2023. Politikweli: A swahili-english code-switched twitter political misinformation classification dataset. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 3–17. Springer.

Arnav Arora, Preslav Nakov, Momchil Hardalov, Sheikh Muhammad Sarwar, Vibha Nayak, Yoan Dinkov, Dimitrina Zlatkova, Kyle Dent, Ameya Bhatawdekar, Guillaume Bouchard, et al. 2023. Detecting harmful content on online platforms: what platforms need vs. where research efforts go. *ACM Computing Surveys*, 56(3):1–17.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, Acyr F. Locatelli, Julia Kreutzer, Nick Frosst, Phil Blunsom, Marzieh Fadaee, A. Ustun, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *ArXiv*, abs/2405.15032.

Abinew Ali Ayele, Skadi Dinter, Tadesse Destaw Belay, Tesfa Tegegne Asfaw, Seid Muhie Yimam, and Chris Biemann. 2022. The 5js in Ethiopia: Amharic hate speech data annotation using toloka crowdsourcing platform. In *2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 114–120. IEEE.

Abinew Ali Ayele, Esubalew Alemneh Jalew, Adem Chanie Ali, Seid Muhie Yimam, and Chris Biemann. 2024. Exploring boundaries and intensities in offensive and hate speech: Unveiling the complex spectrum of social media discourse. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024*, pages 167–178, Torino, Italia. ELRA and ICCL.

Abinew Ali Ayele, Seid Muhie Yimam, Tadesse Destaw Belay, Tesfa Asfaw, and Chris Biemann. 2023. Exploring Amharic hate speech data collection and classification approaches. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 49–59, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63.

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.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, and et al. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Nwachukwu Andrew Egbunike, Noel A. Ihebuzor, and Ngozi Joy Onyechi. 2015. Nature of tweets in the 2015 nigerian presidential elections.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Paula Fortuna, Joao Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. 2019. A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the third workshop on abusive language online*, pages 94–104.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Comfort Ilevbare, Jesujoba Alabi, David Ifeoluwa Adelani, Firdous Bakare, Oluwatoyin Abiola, and Oluwaseyi Adeyemo. 2024. EkoHate: Abusive language and hate speech detection for code-switched political discussions on Nigerian Twitter. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 28–37, Mexico City, Mexico. Association for Computational Linguistics.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. Semeval-2023 task 10: Explainable detection of online sexism. *arXiv preprint arXiv:2303.04222*.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024.

Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4205–4224, Mexico City, Mexico. Association for Computational Linguistics.

Nayeon Lee, Chani Jung, and Alice Oh. 2023. Hate speech classifiers are culturally insensitive. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 35–46, Dubrovnik, Croatia. Association for Computational Linguistics.

Marco Madeddu, Simona Frenda, Mirko Lai, Viviana Patti, and Valerio Basile. 2023. Disaggreghate it corpus: A disaggregated italian dataset of hate speech. In *Italian Conference on Computational Linguistics*.

Nelson Mandela. 1994. *Long Walk to Freedom: The Autobiography of Nelson Mandela*. Little, Brown and Company.

Zainab Mansur, Nazlia Omar, and Sabrina Tiun. 2023. Twitter hate speech detection: A systematic review of methods, taxonomy analysis, challenges, and opportunities. *IEEE Access*, 11:16226–16249.

Antonis Maronikolakis, Axel Wisiorek, Leah Nann, Haris Jabbar, Sahana Udupa, and Hinrich Schütze. 2022. Listening to affected communities to define extreme speech: Dataset and experiments. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1089–1104.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, pages 173–182.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Said Ahmad, Meriem Beloucif, Saif M Mohammad, Sebastian Ruder, et al. 2023. Afrisenti: A twitter sentiment analysis benchmark for african languages. In *Proceedings of EMNLP*.

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.

Joseph Nda Ndabula, Oyenike Mary Olanrewaju, and Faith O Echobu. 2023. Detection of hate speech code mix involving english and other nigerian languages. *Journal of Information Systems and Informatics*, 5(4):1416–1431.

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

Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Abraham Owodunni, Odunayo Ogundepo, David Adelani, and Jimmy Lin. 2023. Better quality pre-training data and t5 models for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 158–168, Singapore. Association for Computational Linguistics.

Edward Ombui, Lawrence Muchemi, and Peter Wagacha. 2019. Hate speech detection in code-switched text messages. In *2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 1–6. IEEE.

Oluwafemi Oriola and Eduan Kotzé. 2020. Evaluating machine learning techniques for detecting offensive and hate speech in south african tweets. *IEEE Access*, 8:21496–21509.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Nedjma Ousidhoum, Yangqiu Song, and Dit-Yan Yeung. 2020. Comparative evaluation of label-agnostic selection bias in multilingual hate speech datasets. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 2532–2542.

Mpho Raborife, Blessing Ogechi Ogbuokiri, and Kehinde D. Aruleba. 2024. The role of social media in xenophobic attack in south africa. *Journal of the Digital Humanities Association of Southern Africa (DHASA)*.

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

Abdulhameed Olaitan Ridwanullah, Sulaiman Ya'u Sule, Bashiru Usman, and Lauratu Umar Abdulsalam. 2024. Politicization of hate and weaponization of twitter/x in a polarized digital space in nigeria. *Journal of Asian and African Studies*, page 00219096241230500.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual HateCheck: Functional tests for multilingual hate speech detection models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. *Proc ACM Web Sci Conf*, 2019:255–264.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ili'c, Daniel Hesslow, Roman Castagn'e, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurenccon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa Etxabe, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris C. Emezue, Christopher Klamm, Colin Leong, Daniel Alexander van Strien, David Ifeoluwa Adelani, Dragomir R. Radev, Eduardo Gonz'alez Ponferrada, and et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100.

Michael Schlichtkrull, Nedjma Ousidhoum, and Andreas Vlachos. 2023. The intended uses of automated fact-checking artefacts: Why, how and who. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8618–8642.

Farhana Shahid and Aditya Vashistha. 2023. Decolonizing content moderation: Does uniform global community standard resemble utopian equality or western power hegemony? In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Zeerak Talat and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *North American Chapter of the Association for Computational Linguistics*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L'eonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram'e, Johan Ferret, Peter Liu, Pouya Dehghani Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and et al. 2024. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2408.00118.

Atnafu Lambebo Tonja, Bonaventure F. P. Dossou, Jessica Ojo, Jenalea Rajab, Fadel Thior, Eric Peter Wairagala, Aremu Anuoluwapo, Pelonomi Moiloa, Jade Abbott, Vukosi Marivate, and Benjamin Rosman. 2024. Inkubalm: A small language model for low-resource african languages. *ArXiv*, abs/2408.17024.

Manuel Tonneau, Pedro Quinta De Castro, Karim Lasri, Ibrahim Farouq, Lakshmi Subramanian, Victor Orozco-Olvera, and Samuel Fraiberger. 2024. NaijaHate: Evaluating hate speech detection on Nigerian Twitter using representative data. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9020–9040, Bangkok, Thailand. Association for Computational Linguistics.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *ArXiv*, abs/2209.11055.

Francielle Alves Vargas, Samuel Guimarães, Shamsuddeen Hassan Muhammad, Diego Alves, Ibrahim Said Ahmad, Idris Abdulmumin, Diallo Mohamed, Thiago Alexandre Salgueiro Pardo, and Fabrício Benevenuto. 2024. Hausahate: an expert annotated corpus for hausa hate speech detection. In *Proceedings*.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. 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 1667–1682, Online. Association for Computational Linguistics.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Italian Conference on Cybersecurity*.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL).

A Appendix

A.1 Prompt templates

Prompt 1

I am providing you with the definition of Hate speech, Abusive language, and Neutral tweets.

Hate speech is language content that expresses hatred towards a particular group or individual based on their political affiliation, race, ethnicity, religion, gender, sexual orientation, or other characteristics. It also includes threats of violence.

Neutral language does not contain any bad language.

Which category does the tweet above belong to: 'Hate', 'Abuse', or 'Neutral'? Pick exactly one category. Don't give any additional context, just classify the tweet.

Tweet: text Category:

Prompt 2

Read the following label definitions and provide a label without any explanations.

Hate: Hate speech is public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, gender, ethnicity, sexual orientation or other characteristics.

Abusive: Abusive and offensive language means verbal messages that use words in an inappropriate way and may include but is not limited to swearing, name-calling, or profanity. Offensive language may upset or embarrass people because it is rude or insulting.

Neutral: Neutral language is neither hateful nor abusive or offensive. It does not contain any bad language.

Text: tweet
Label:

Prompt 3

Read the following text and definitions:

Text: tweet.

Definitions: Hate: Hate speech is public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, gender, ethnicity, sexual orientation or other characteristics.

Abuse: Abusive and offensive language means verbal messages that use words in an inappropriate way and may include but is not limited to swearing, name-calling, or profanity. Offensive language may upset or embarrass people because it is rude or insulting

Neutral: Neutral language is neither hateful nor abusive or offensive. It does not contain any bad language. Which of these

definitions (hate, abuse, neutral) apply to this tweet?

Prompt 4

Read the following definitions and text to categorize:

Definitions: Hate: Hate speech is public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, gender, ethnicity, sexual orientation or other characteristics.

Abuse: Abusive and offensive language means verbal messages that use words in an inappropriate way and may include but is not limited to swearing, name-calling, or profanity. Offensive language may upset or embarrass people because it is rude or insulting

Neutral: Neutral language is neither hateful nor abusive or offensive. It does not contain any bad language.

Text: tweet. Which of these definitions

(hate, abuse, neutral) apply to this tweet?

Prompt 5

You will be given a text snippet and 3 category definitions. Your task is to choose which category applies to this text.

Your text snippet is: tweet

Your category definitions are:

HATE category definition: Hate speech is public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, gender, ethnicity, sexual orientation or other characteristics.

ABUSE category definition: Abusive and offensive language means verbal messages that use words in an inappropriate way and may include but is not limited to swearing, name-calling, or profanity. Offensive language may upset or embarrass people because it is rude or insulting

NEUTRAL category definition: Neutral language is neither hateful nor abusive or offensive. It does not contain any bad language.

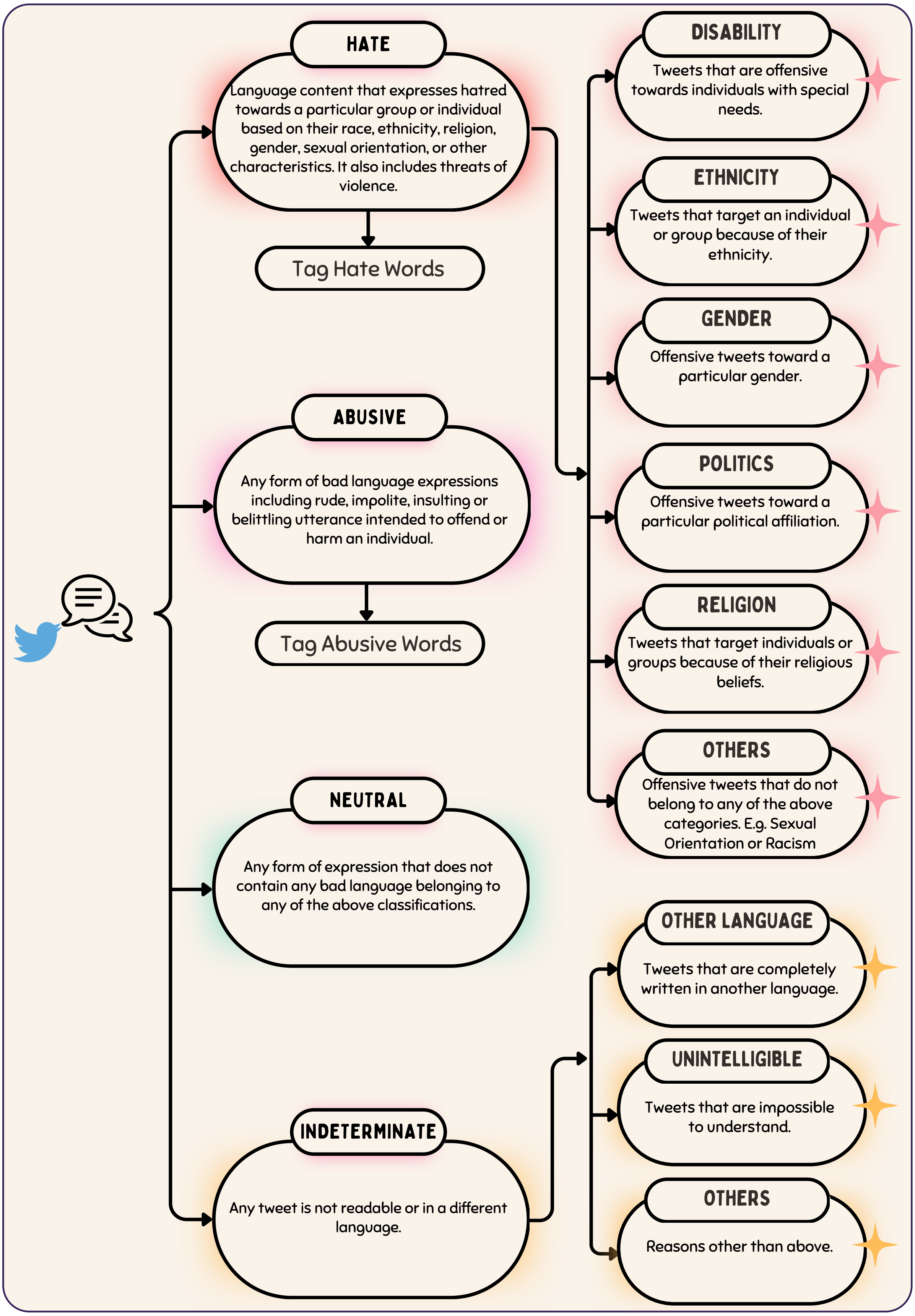
Does the text snippet belong to the HATE, ABUSIVE, or the NEUTRAL category? Thinking step by step answer HATE, ABUSIVE, or NEUTRAL capitalizing all the letters. Explain your reasoning FIRST, then output HATE, ABUSIVE, or NEUTRAL.

Model	# Shots	amh	ary	arq	hau	ibo	kin	oro	pcm	som	swa	tir	twi	xho	yor	zul	Avg.
SetFit	0	33.79	46.54	27.40	23.57	36.90	25.29	26.26	43.46	32.18	44.38	49.30	51.13	37.38	48.46	35.08	37.41
	5	44.89	33.82	36.07	49.93	51.52	55.35	36.56	53.28	48.09	54.90	35.36	33.23	47.49	34.36	33.18	43.20
	10	49.42	35.81	46.55	49.72	39.31	53.10	39.10	57.00	42.29	62.64	39.16	45.76	43.52	49.73	40.33	46.23
	20	50.13	40.49	54.24	55.40	65.15	57.35	40.91	59.09	48.95	74.93	40.04	53.72	46.92	56.54	50.67	52.97
InkubaLM-0.4B	0	22.96	19.38	18.14	20.02	18.42	22.44	26.16	16.86	20.76	20.74	25.10	17.44	20.58	19.26	20.38	20.58
	5	28.04	24.70	20.06	21.78	21.98	25.78	26.86	19.08	25.06	20.48	29.26	24.50	23.04	20.98	21.68	23.55
	10	27.00	26.46	23.92	24.48	23.52	27.22	29.14	23.68	25.98	24.58	27.92	24.02	24.68	23.68	25.06	25.42
	20	27.06	26.42	24.60	24.30	23.86	26.96	28.58	23.76	25.88	24.64	30.04	23.32	23.56	23.14	25.20	25.42
mt0-small	0	18.86	16.98	13.42	12.22	9.72	17.40	20.68	12.40	15.16	14.07	19.30	8.56	14.20	12.04	13.62	14.58
	5	26.38	25.76	21.66	19.94	21.28	23.04	27.78	25.98	21.40	20.30	26.20	12.66	19.96	17.72	19.50	21.97
	10	27.38	23.72	20.40	20.94	22.36	21.08	27.94	25.58	19.88	16.93	26.62	12.08	18.56	19.90	18.88	21.48
	20	26.28	23.38	19.86	21.60	22.14	21.18	28.16	25.06	19.96	20.05	28.04	13.44	19.12	19.76	18.50	21.77
bloomz-7b1-mt	0	17.18	21.86	21.86	18.60	18.42	21.98	19.38	21.60	16.74	25.10	16.92	17.70	19.24	22.36	19.80	19.92
	5	22.74	21.40	25.88	23.10	27.88	25.36	26.22	26.66	23.54	27.03	25.64	28.34	23.26	24.80	22.62	24.96
	10	27.86	21.06	25.88	24.40	27.90	22.54	27.86	26.24	25.48	19.83	28.58	28.50	23.56	24.16	23.92	25.18
	20	25.46	19.92	24.64	24.96	27.72	24.18	26.64	26.82	25.76	24.50	28.48	28.84	23.24	25.28	23.88	25.35
Mistral-7B-v0.1	0	21.18	17.46	11.98	6.84	8.04	15.18	20.82	6.76	8.16	10.64	25.62	8.38	9.72	8.22	8.92	12.53
	5	37.68	31.60	44.54	34.90	36.84	46.98	39.46	50.00	34.12	49.58	35.64	32.04	34.24	41.04	32.54	38.75
	10	39.68	36.50	50.56	37.82	36.94	45.72	39.02	52.10	31.82	54.34	35.84	30.82	32.64	42.90	32.46	39.94
	20	36.64	35.46	52.94	38.44	38.98	52.28	39.24	54.60	33.28	55.98	37.42	30.44	34.12	46.40	35.00	41.41
aya-23-35B	0	17.04	19.20	17.20	20.34	14.04	21.34	20.38	20.02	18.56	21.20	17.10	10.68	20.32	17.48	20.24	18.34
	5	37.78	56.40	57.00	39.70	42.70	48.78	39.24	57.90	38.72	56.32	35.68	40.24	36.76	48.88	44.58	45.38
	10	38.50	53.96	61.82	44.42	45.80	52.82	42.16	59.26	37.60	61.08	35.88	39.92	36.18	48.44	45.00	46.86
	20	38.10	52.90	63.76	46.14	46.52	57.62	42.26	61.68	38.68	73.68	38.12	38.54	37.74	51.52	44.52	48.79
Gemma-2-9B	0	27.42	28.36	25.46	14.48	15.00	24.08	23.74	24.74	10.92	32.12	25.52	19.50	13.40	21.66	12.52	21.26
	5	56.60	57.68	61.14	49.98	48.12	60.26	45.10	60.30	46.42	72.88	48.40	46.50	35.82	55.96	46.60	52.78
	10	57.84	61.08	61.72	56.62	53.78	63.54	45.78	60.54	51.72	76.50	45.48	48.16	38.82	60.22	52.58	55.63
	20	60.96	59.94	64.38	55.90	54.56	64.14	44.70	62.54	52.22	77.56	44.94	48.92	39.02	60.50	53.68	56.26
Gemma-2-27B	0	41.78	41.24	41.12	28.96	31.36	42.08	33.46	49.78	31.10	54.54	30.84	27.74	25.64	41.46	28.02	36.61
	5	59.62	64.14	65.62	54.62	56.14	61.08	46.76	61.78	54.12	77.96	52.12	47.26	40.48	61.72	54.28	57.18
	10	60.70	61.36	64.88	58.90	56.84	64.86	46.96	61.60	52.86	80.04	49.82	50.86	41.90	59.66	56.94	57.88
	20	62.28	59.86	65.80	59.60	58.76	66.06	48.90	63.24	52.90	81.18	53.36	50.34	43.08	59.10	56.88	58.76
Llama-3.1-8B	0	17.36	19.34	18.82	24.36	12.70	23.90	22.56	20.10	24.86	20.82	14.18	9.02	21.42	18.60	21.94	19.33
	5	38.08	40.88	51.68	36.78	35.90	38.78	31.24	52.38	30.62	58.10	30.20	34.82	29.22	40.42	30.90	38.67
	10	42.04	43.10	54.76	37.78	37.78	43.02	35.92	53.88	32.98	64.96	31.92	34.52	28.20	38.98	30.72	40.70
	20	47.74	40.84	57.92	41.64	37.76	48.86	39.78	55.36	30.26	69.38	37.12	33.60	27.84	42.68	29.12	42.66
Llama-3.1-70B	0	36.34	43.34	42.64	35.64	32.52	38.52	31.54	48.66	31.14	49.36	27.08	25.54	28.98	40.64	35.50	36.50
	5	58.52	66.24	62.88	52.86	52.88	58.14	45.50	64.38	52.72	73.48	44.42	43.90	39.02	58.08	55.24	55.22
	10	61.18	64.46	62.20	56.40	55.10	59.00	46.60	63.58	54.40	76.50	49.62	45.76	39.04	57.96	55.74	56.50
	20	60.38	61.36	63.80	57.40	53.80	62.48	49.02	63.18	51.82	77.72	53.90	47.50	40.34	57.94	56.36	57.13
GPT-4o	0	61.78	66.41	73.75	56.91	68.82	62.53	60.01	65.94	63.42	73.66	45.75	52.72	58.68	75.21	54.47	62.67
	5	66.73	73.53	77.33	55.44	76.73	72.27	70.94	66.29	59.33	80.95	61.11	73.21	60.45	76.98	59.34	68.71
	10	67.94	75.54	77.54	58.15	80.08	75.07	72.27	67.14	62.75	84.19	57.52	72.28	65.75	76.74	65.05	70.53
	20	68.08	76.16	78.69	58.34	80.81	74.86	72.33	65.41	66.44	84.61	59.55	75.86	66.08	77.11	71.36	71.71
Lang. avg.	-	40.80	41.73	44.47	37.60	39.26	43.51	37.59	44.99	36.16	51.01	36.37	35.05	33.03	41.56	36.43	39.97

Table 11: **Model performance (Macro F1-score) for zero- and few shot classifiers** across the 15 languages in AFRIHATE. Best performance for each language is highlighted in **bold**.

model	#shots	amh	ary	arq	hau	ibo	kin	oro	pcm	som	swa	tir	twi	xho	yor	zul	Avg.
SetFit	0	33.20	42.92	26.63	24.02	38.61	21.29	28.72	39.23	29.26	35.26	47.32	52.15	37.62	41.27	35.03	35.50
	5	45.11	35.91	31.89	50.81	50.30	51.26	36.50	52.54	49.93	54.83	31.63	38.40	46.14	31.62	35.71	42.84
	10	49.26	37.05	44.27	50.43	40.80	48.60	42.56	56.87	40.27	62.88	39.74	49.28	42.60	49.21	42.03	46.39
	20	50.33	40.92	53.25	54.34	65.65	57.28	43.08	58.44	48.86	74.68	38.95	55.87	46.78	53.85	51.10	52.89
InkubaLM-0.4B	0	39.74	23.62	30.26	22.66	23.04	30.96	41.14	19.84	28.30	27.02	47.24	23.16	27.30	23.70	27.38	29.02
	5	39.12	23.72	31.88	25.28	34.36	32.60	44.88	23.74	30.86	25.62	48.12	43.96	28.14	26.88	27.52	32.45
	10	37.20	29.52	32.64	38.38	27.74	37.58	43.26	28.82	41.26	31.22	45.84	32.90	34.22	31.54	35.76	35.19
	20	37.64	29.18	32.68	38.64	28.70	37.84	43.06	28.98	41.34	31.37	48.88	32.30	33.82	30.84	36.12	35.43
mt0-small	0	38.14	23.36	31.76	28.70	17.02	35.56	43.74	22.70	31.34	27.47	42.80	14.66	28.32	23.16	27.62	29.09
	5	36.70	24.90	34.12	29.18	29.48	31.36	41.10	32.06	25.26	27.90	42.56	16.20	33.04	28.38	32.86	31.01
	10	37.40	25.74	35.06	26.36	33.44	28.30	40.46	31.98	24.00	23.37	43.04	15.88	31.32	29.72	32.00	30.54
	20	36.50	25.98	35.08	26.20	34.50	28.44	40.30	31.54	23.86	23.40	44.76	17.08	31.68	28.92	32.30	30.70
bloomz-7b1-mt	0	34.62	34.38	33.08	32.66	29.18	36.20	37.30	32.76	33.90	38.53	35.82	28.88	31.94	33.70	32.24	33.68
	5	32.36	37.34	46.74	37.52	62.56	30.88	31.96	48.14	33.90	40.97	36.68	68.50	39.36	48.86	39.64	42.36
	10	35.16	36.96	45.92	37.72	62.40	27.74	33.06	47.10	35.04	36.97	38.70	67.94	38.56	47.66	39.44	42.02
	20	33.16	35.60	44.18	37.88	63.12	29.12	31.82	47.04	34.78	37.93	38.72	68.84	38.14	49.08	39.08	41.90
Mistral-7B-v0.1	0	42.26	30.62	17.28	10.52	13.00	27.62	45.42	11.14	13.86	18.88	55.02	14.28	16.36	13.44	14.54	22.95
	5	37.92	43.46	57.14	51.96	58.42	50.40	45.70	59.32	52.64	58.58	41.36	69.52	47.50	57.72	45.96	51.84
	10	40.42	46.68	60.66	56.38	61.74	50.00	46.70	60.42	51.50	63.18	44.88	70.26	46.40	60.50	46.42	53.74
	20	39.48	45.08	61.70	56.98	65.10	55.64	48.82	61.22	54.02	64.28	51.38	70.04	48.64	64.94	50.16	55.83
aya-23-35B	0	34.54	32.76	28.52	37.76	18.66	41.04	42.48	30.34	40.76	34.04	35.84	14.40	35.42	28.22	35.96	32.72
	5	42.20	61.86	64.38	55.32	56.42	56.56	49.96	61.12	57.46	69.84	45.70	63.86	48.96	63.48	51.14	56.55
	10	45.74	61.22	69.00	59.78	63.98	58.66	52.30	62.08	56.98	74.10	51.08	69.58	49.06	66.50	52.44	59.50
	20	46.76	59.70	71.40	59.84	68.30	62.06	54.14	65.44	57.78	75.68	56.90	71.20	51.72	70.54	56.12	61.84
Llama-3.1-8B	0	29.14	34.10	36.04	55.82	22.60	48.60	42.96	39.84	55.00	49.04	23.18	15.20	46.18	38.02	47.42	38.88
	5	41.50	50.12	63.90	46.56	64.48	42.24	34.94	61.22	39.08	63.88	35.80	68.12	43.66	58.14	43.30	50.46
	10	45.16	51.70	66.04	47.72	66.12	44.60	39.16	62.70	36.90	68.54	39.88	69.84	43.20	60.88	42.68	52.34
	20	50.20	50.56	67.92	48.18	66.92	48.50	44.10	64.28	38.70	72.00	48.26	70.40	42.72	62.24	43.16	54.54
Gemma-2-9B	0	48.26	37.78	26.88	17.60	16.24	34.34	46.86	26.68	16.22	37.12	57.98	27.20	18.20	22.96	16.58	30.06
	5	61.10	63.62	65.72	55.20	62.84	63.24	56.34	63.36	54.46	74.78	62.12	65.44	46.24	70.14	52.12	61.11
	10	61.26	63.78	65.98	60.64	66.38	64.60	56.40	64.08	57.28	78.16	61.20	69.76	49.66	73.02	54.74	63.13
	20	64.14	61.86	67.90	60.26	69.38	65.48	57.24	64.50	58.62	79.24	62.98	70.68	50.02	73.96	56.00	64.15
Gemma-2-27B	0	53.40	46.38	44.10	33.62	31.08	49.02	50.82	54.20	38.16	59.88	55.58	31.34	29.66	45.96	31.90	43.67
	5	61.40	66.00	69.40	60.44	68.30	63.26	56.34	64.44	62.34	80.18	59.90	64.34	51.36	75.60	56.24	63.97
	10	62.44	64.76	69.74	64.70	70.62	66.22	56.24	64.22	61.24	81.94	59.52	70.82	53.18	75.34	59.16	65.34
	20	64.12	63.60	70.24	65.74	72.84	67.72	59.18	66.06	63.20	82.98	64.44	73.28	56.04	76.50	61.18	67.14
Llama-3.1-70B	0	46.28	49.84	52.30	51.26	43.72	50.38	48.30	56.50	46.38	60.14	43.64	35.68	42.64	51.34	47.00	48.36
	5	59.10	68.32	67.12	58.96	66.88	60.90	54.24	67.20	61.20	75.76	45.18	66.18	50.18	72.60	57.30	62.07
	10	61.36	67.80	67.70	63.00	68.78	61.50	55.86	66.62	63.38	78.74	50.48	68.34	52.04	73.22	58.92	63.85
	20	60.72	65.62	69.48	64.00	70.02	65.08	58.78	67.02	64.46	80.02	56.40	72.52	53.50	74.98	61.26	65.59
GPT-4o	0	61.70	56.05	66.45	45.90	44.21	44.92	44.93	49.67	38.50	57.24	46.01	33.90	36.72	48.62	36.34	47.41
	5	66.69	73.26	71.03	53.24	74.23	72.39	65.30	55.93	56.19	79.02	57.64	61.36	51.81	69.65	58.45	64.41
	10	67.09	75.12	73.81	55.28	78.02	74.47	66.97	56.47	58.34	72.83	56.50	61.11	61.61	70.78	65.12	66.23
	20	67.22	76.05	75.00	56.08	78.17	74.85	66.01	50.33	61.95	84.14	56.91	51.67	62.78	72.22	61.84	66.35
Lang. avg.	-	47.21	47.15	51.73	45.76	51.10	48.39	47.03	49.50	44.74	56.22	47.65	50.37	42.15	51.59	43.85	48.30

Table 12: Model Performances (Accuracy) for zero- and few-shot Learning



HATE
Language content that expresses hatred towards a particular group or individual based on their race, ethnicity, religion, gender, sexual orientation, or other characteristics. It also includes threats of violence.
Tag Hate Words
ABUSIVE
Any form of bad language expressions including rude, impolite, insulting or belittling utterance intended to offend or harm an individual.
Tag Abusive Words
NEUTRAL
Any form of expression that does not contain any bad language belonging to any of the above classifications.
INDETERMINATE
Any tweet is not readable or in a different language.
DISABILITY
Tweets that are offensive towards individuals with special needs.
ETHNICITY
Tweets that target an individual or group because of their ethnicity.
GENDER
Offensive tweets toward a particular gender.
POLITICS
Offensive tweets toward a particular political affiliation.
RELIGION
Tweets that target individuals or groups because of their religious beliefs.
OTHERS
Offensive tweets that do not belong to any of the above categories. E.g. Sexual Orientation or Racism
OTHER LANGUAGE
Tweets that are completely written in another language.
UNINTELLIGIBLE
Tweets that are impossible to understand.
OTHERS
Reasons other than above.

Figure 1: Annotation Guidelines and Definitions