NAIJAHATE: Evaluating Hate Speech Detection on Nigerian Twitter Using Representative Data

Manuel Tonneau [1, 2, 3], **Pedro Vitor Quinta de Castro** [1, 4], **Karim Lasri** [1, 5], **Ibrahim Farouq** [1, 6], **Lakshminarayanan Subramanian** [3], **Victor Orozco-Olvera** [1], **Samuel Fraiberger** [1, 3, 7],
[1] The World Bank, [2] University of Oxford, [3] New York University
[4] Universidade Federal de Goiás [5] Ecole Normale Supérieure
[6] Universiti Sultan Zainal Abidin
[7] Massachusetts Institute of Technology

Abstract

To address the global issue of hateful content proliferating in online platforms, hate speech detection (HSD) models are typically developed on datasets collected in the United States, thereby failing to generalize to English dialects from the Majority World. Furthermore, HSD models are often evaluated on curated samples, raising concerns about overestimating model performance in real-world settings. In this work, we introduce NAIJAHATE, the first dataset annotated for HSD which contains a representative sample of Nigerian tweets. We demonstrate that HSD evaluated on biased datasets traditionally used in the literature largely overestimates real-world performance on representative data. We also propose NAIJAXLM-T, a pretrained model tailored to the Nigerian Twitter context, and establish the key role played by domain-adaptive pretraining and finetuning in maximizing HSD performance. Finally, we show that in this context, a human-in-the-loop approach to content moderation where humans review 1% of Nigerian tweets flagged as hateful would enable to moderate 60% of all hateful content. Taken together, these results pave the way towards robust HSD systems and a better protection of social media users from hateful content in low-resource settings.

Content warning: This article contains illustrative examples of hateful content.

1 Introduction

Social media came with the promise of connecting people, increasing social cohesion, and letting everyone have an equal say. However, harmful content including hate speech has become rampant, fueling fears of its impact on social unrests and hate crimes (Müller and Schwarz, 2021). While regulatory frameworks have compelled social media platforms to take action to curb hate speech (Gagliardone et al., 2016), content detection and moderation efforts have largely focused on the American and European markets, prompting questions on how to efficiently tackle this issue in the Majority World (Poletto et al., 2021; Milmo, 2021). Our study focuses on Nigerian Twitter, a low-resource context which provides an opportunity to study online hate speech at the highest level (Ezeibe, 2021). Exemplifying the issue, Twitter was banned by the Nigerian government between June 2021 and January 2022, supposedly due to the platform's deletion of a tweet by President Buhari in which he incited violence towards the Biafran separatists (Maclean, 2021).

We adopt the definition of hate speech from the United Nations: "any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factor." (UN, 2019). The challenges in developing systems capable of efficiently detecting such content are two-fold. First, hateful content is infrequent – approximately 0.5% of posts on US Twitter are hateful (Jiménez Durán, 2021) – creating an obstacle to generating representative annotated datasets at a reasonable cost. To alleviate this issue, models are developed on curated datasets by oversampling hateful content matching predefined keywords (Davidson et al., 2017), or by employing techniques such as active learning to maximize performance for a given annotation cost (Kirk et al., 2022; Markov et al., 2023). These design choices generate biases in evaluation datasets (Wiegand et al., 2019; Nejadgholi and Kiritchenko, 2020), raising questions on the generalizability of HSD models to real-world settings.

Second, while a plethora of HSD modeling options are available, it is unclear how well they adapt to a new context. Although few-shot learners are appealing for requiring no or few finetuning data, evidence on their performances relative to supervised HSD baselines is mixed (Plaza-del arco et al., 2023a; Guo et al., 2024). Off-the-shelf supervised models such as Perspective API are typically finetuned on US data and tend to not generalize well to English dialects spoken in the Majority World (Ghosh et al., 2021). Finally, while further pretraining existing architectures to adapt them to a new context is known to increase performance on downstream tasks (Gururangan et al., 2020), it is unclear whether highly specific contexts require a custom domain adaptation. Overall, questions remain on the extent to which available HSD methods perform when adapted to a low-resource context (Li, 2021).

In this work, we present NAIJAHATE, a dataset of 35,976 Nigerian tweets annotated for HSD, which includes a representative evaluation sample to shed light on the best approach to accurately detect hateful content in real-world settings. We also introduce NAIJAXLM-T, a pretrained language model adapted to the Nigerian Twitter domain. We demonstrate that evaluating HSD models on biased datasets traditionally used in the literature largely overestimates performance on representative data (83-90% versus 34% in average precision). We further establish that domain-adaptive pretraining and finetuning leads to large HSD performance gains on representative evaluation data over both US and Nigerian-centric baselines. We also find that finetuning on linguistically diverse hateful content sampled through active learning significantly improves performance in real-world conditions relative to a stratified sampling approach. Finally, we discuss the cost-recall tradeoff in moderation and show that having humans review about 1% of all tweets flagged as hateful allows to moderate up to 60% of all hateful content on Nigerian Twitter, highlighting the constraints of a human-in-the-loop approach to content moderation as social media usage continues to grow globally.

Therefore, our main contributions are [1]:

- NAIJAHATE, a dataset which includes the first representative evaluation sample annotated for HSD on Nigerian Twitter
- NAIJAXLM-T, a pretrained language model adapted to the Nigerian Twitter domain
- an evaluation on representative data of the role played by domain adaptation and training data diversity and of the feasibility of hateful content moderation at scale

[1]The dataset and the related models can be found at https://github.com/manueltonneau/NaijaHate

2 Related work

2.1 Nigerian hate speech datasets

While existing hate speech datasets are primarily in US English (Poletto et al., 2021), mounting evidence highlights the limited generalizability of learned hate speech patterns from one dialect to another (Ghosh et al., 2021). In this context, recent work has developed hate speech datasets for the Majority World (Nkemelu et al., 2022), including for the Nigerian context; however, the latter either focused on one specific form of hate speech (Aliyu et al., 2022), one language (Adam et al., 2023), or specific events (Ndabula et al., 2023). To the best of our knowledge, our work is the first to construct a comprehensive dataset annotated for hate speech for the entire Nigerian Twitter ecosystem, covering both the diversity of languages and hate targets.

2.2 Hate speech detection and evaluation

HSD methods fall into three categories: rule-based (Mondal et al., 2017), supervised learning, and zero-shot learning (ZSL) using decoder-based models (Nozza, 2021). Rule-based methods rely on predefined linguistic patterns and therefore only typically achieve very low recall. Additionally, supervised learning require annotated datasets which are usually scarce in Majority World contexts, motivating data-efficient strategies for HSD, such as data augmentation (Roychowdhury and Gupta, 2023) or expansion from high-resourced languages (Röttger et al., 2022). While recent advancements in ZSL could potentially circumvent the need to produce finetuning data for supervised learning, existing evidence on the relative performance of the two approaches is mixed (Plaza-del arco et al., 2023a; Guo et al., 2024). A major shortcoming of the existing literature is that modeling approaches are typically evaluated on biased datasets whose characteristics greatly differ from real-world conditions (Wiegand et al., 2019; Nejadgholi and Kiritchenko, 2020), raising concerns about overestimating model performance (Arango et al., 2019). To address these concerns, we provide the first evaluation of HSD methods on a representative evaluation sample, providing unbiased estimates of their performance in a real-world setting.

2.3 Hate speech moderation

To counter hate speech, social media platforms have invested in content moderation through post removal or downranking (Gillespie, 2018). Detecting hateful content within the vast amount of data posted on social media is a challenging task, motivating the use of algorithmic methods (Gillespie, 2020). However, fully automated approaches have raised concerns related to the fairness and potential biases in moderation decisions (Gorwa et al., 2020). As a middle ground, recent work has proposed a human-in-the-loop approach (Lai et al., 2022), where a model flags content likely to infringe platform rules, which is then reviewed by humans who decide whether or not to moderate it. Albeit promising, it remains unclear whether this process is scalable both from a cost and a performance standpoint. To fill this gap, we provide the first estimation of the feasibility of a human-in-the-loop approach in the case of Nigerian Twitter.

3 Data

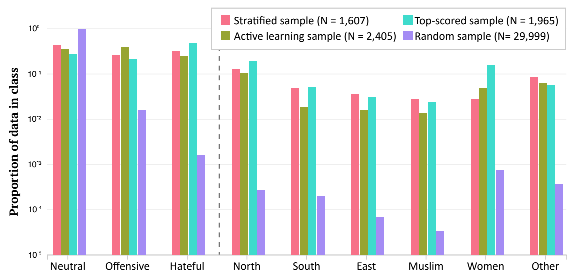

Figure 1: Proportion of data in each class, showing the composition of the hateful class across hate targets.

3.1 Data collection

Between July 2021 and July 2023, we used the Twitter API to collect a dataset containing 2.2 billion tweets posted between March 2007 and July 2023 and forming the timelines of 2.8 million users with a profile location in Nigeria.[2] We iteratively collected the timeline of users with a profile location in Nigeria being mentioned in the timeline of other Nigerian users until no additional Nigerian users were retrieved, ensuring maximum coverage of the Nigerian ecosystem. This Nigerian Twitter dataset is mostly constituted of English tweets (77%) followed by tweets in Nigerian Pidgin – an English-based creole widely spoken across Nigeria – (7%), tweets mixing English and Pidgin (1%), tweets in Hausa (1%) and tweets in Yoruba (1%) (Table 8). We then drew two distinct random samples of 100 million tweets each, one for model training and the other one for evaluation.

[2]The dataset contains 13.9 billion tokens and 525 million unique token, for a total of 89GB of uncompressed text.

3.2 Annotation

We recruited a team of four Nigerian annotators, two female and two male, each of them from one of the four most populated Nigerian ethnic groups – Hausa, Yoruba, Igbo and Fulani. We followed a *prescriptive* approach (Rottger et al., 2022) by instructing annotators to strictly adhere to extensive annotation guidelines describing our taxonomy of hate speech (detailed in A.2.2). Following prior work (Davidson et al., 2017; Mathew et al., 2021), HSD is operationalized by labeling tweets with one of three classes: (i) *hateful*, if it contains an attack on an individual or a group based on the perceived possession of a certain characteristic (e.g., gender, race) (UN, 2019), (ii) *offensive*, if it contains a personal attack or an insult that does not target an individual based on their identity (Zampieri et al., 2019), or (iii) *neutral* if it is neither hateful nor offensive. If a tweet is labeled as hateful, it is also annotated for the communities being targeted (Table 1). Each tweet was labeled by three annotators. For the three-class annotation task, the 3 annotators agreed on 90% of labeled tweets, 2 out of 3 agreed in 9.5% of cases, and all three of them disagreed in 0.5% of cases (Krippendorff's alpha = 0.7).

Label	Target	Examples
Hateful	North	My hate for northern people keeps growing
	South	You idiotic Southerners fighting your own
	East	IPOBs are animals....They lack tact or strategy.
	Muslim	Muslim baboons and their terrorist religion.
	Women	Nobody should believe this ashawo woman
Offensive	None	Stop spewing rubbish, mumu.
Neutral	None	She don already made up her mind sha.

Table 1: Examples of tweets for each class. Offensive tweets have no target as they do not target an identity group.

3.3 Training samples

Stratified sample Due to the rarity of hateful content, sampling tweets randomly would result in a very imbalanced set. Indeed, the prevalence of hate speech in the wild typically ranges from 0.003% to 0.7% depending on the platform and timeframe (Gagliardone et al., 2016; Mondal et al., 2017; Jiménez Durán, 2021). To circumvent this issue, we follow previous work by oversampling training examples containing keywords expected to be associated with hate. We handpick a list of 89 hate-related keywords combining hate speech lexicons and online dictionaries (Ferroggiaro, 2018; Udanor and Anyanwu, 2019; Farinde and Omolaiye, 2020). We also identify 45 keywords referring to communities frequently targeted by hate in the Nigerian context due to their ethnicity (Fulani, Hausa, Herdsmen[3], Igbo, Yoruba), religion (Christians, Muslims), region of origin (northerners, southerners, easterners) or gender identity or sexual orientation (women, LGBTQ+) (Onanuga, 2023). We then annotate 1,607 tweets from the training sample that were stratified by community-related and hate-related keywords (see App. A.1.3). Stratified sampling indeed enables to reduce the imbalance in the training data (Fig. 1): the resulting share of tweets labeled as neutral, offensive and hateful is respectively equal to 50, 17, and 33%.

Active learning sample While stratified sampling makes it possible to oversample hateful content in the training data, it is constrained by a predefined vocabulary which limits the coverage and diversity of the positive class. As an alternative, we employy a variant of certainty sampling to annotate a second set of traning examples using. The latter is an active learning method that focuses the learning process of a model on instances with a high confidence score of belonging to the minority class, spanning a more diverse spectrum of examples (Attenberg et al., 2010). We generate additional training instances in four steps: (i) we start by finetuning Conversational BERT (Burtsev et al., 2018) on the stratified sample; (ii) we then deploy the finetuned model on the training sample of 100 million tweets; (iii) next, we label an additional 100 high-scored tweets from the training sample; and finally, (iv) we incorporate the additional labels into Conversational BERT's finetuning sample. We repeat this process 25 times, thereby producing an additional 2,405 training examples with a majority label. We find that active sampling produces about the same proportion of observations from the hateful class (25% versus 31%) as stratified sampling (Fig. 1). However, it enables to generate more diversity in the hateful class (Table 2): the proportion of training examples that do not contain any seed keywords[4], the proportion of unique tokens and the average pairwise embedding distance are all larger in the active learning sample relative to the stratified sample.

[3]Herdsmen are not a ethnic group per se but this term refers exclusively to Fulani herdsmen in the Nigerian context, hence the categorization as an ethnic group.

	Stratified	Active learning	Top-scored	Random
Proportion of tweets not containing seed keywords	0.075	0.725	0.708	0.938
Proportion of unique tokens	0.322	0.333	0.29	0.615
Average pairwise embedding distance	0.139	0.152	0.159	0.172

Table 2: Diversity metrics for the hateful class across datasets. Active learning enables to generate more diversity in the training data, bringing them closer to the representative random sample.

3.4 Evaluation samples

Top-scored sample To evaluate models' performance in real-world conditions, we start by testing how they behave in the presence of a distribution shift. We first train each supervised model considered in this study on the union of the stratified and the active learning sample, deploy it on the random sample of 100 million tweets used for evaluation and annotate 200 high-scored tweets. We repeat this process for the 10 models evaluated in this study (see Section 4 for more details) and combine all the high-scored tweets, yielding a pooled sample of 1,965 annotated tweets with a majority label. The share of tweets labeled as neutral, offensive and hateful is respectively equal to 28%, 22% and 50% (Fig. 1). This approach traditionally used in information retrieval enables to evaluate the performance of each model on a large dataset containing a high and diverse proportion of positive examples discovered by qualitatively different models, and whose distribution differ from that of the training data (Voorhees et al., 2005).

Random sample Finally, we annotate a random sample of 29,999 tweets to evaluate HSD models on a representative dataset of Nigerian tweets. As expected, we discover that the prevalence of hateful content is very low: approximately 0.16% and 1.6% of tweets are labeled as hateful and offensive, respectively (Fig. 1). In addition, we find that the diversity within the positive class in the random sample is larger than in the training samples (Table 2).

[4]i.e., keywords used for stratified sampling

4 Experimental setup

A typical NLP pipeline typically consists in finetuning a pretrained model to perform a downstream task which involves domain-related distributions : the pretraining domain, and the finetuning domain. In this study, our experiments aim to determine the best choices for Nigerian HSD and estimate the impact of domain adaptation – both for pretraining and finetuning – on real-world performance. Additionally, recent off-the-shelf general-purpose models, such as *GPT-3.5*[5], can be tested in a zero-shot setting, skipping the finetuning phase, compromising the gain in efforts to manually annotate examples for supervision with robustness in a highly specific context. We also benchmark the finetuned models against Perspective API (Lees et al., 2022), a widely-deployed toxic language detection system relying on BERT-based supervised learning for which the finetuning data is not public.

Finetuning domain We experiment with four finetuning datasets: HATEXPLAIN (Mathew et al., 2021), which contains US English posts from Twitter and Gab annotated for HSD; HERDPHOBIA (Aliyu et al., 2022), a dataset of Nigerian tweets annotated for hate against Fulani herdsmen; HSCODEMIX (Ndabula et al., 2023), containing Nigerian tweets posted during the EndSARS movement and the 2023 presidential election and annotated for general hate speech; and finally NAIJAHATE, our dataset presented in Section 3.

Pretraining domain We introduce NAIJAXLM-T (FULL), an XLM-R model (Conneau et al., 2020) further pretrained on 2.2 billion Nigerian tweets for one epoch. We compare its performance relative to BERT-based models pretrained in three different domains:

- the general domain, which include a variety of sources such as books and news, both in English (DeBERTaV3 (He et al., 2021)) and in multilingual settings (XLM-R (Conneau et al., 2020), mDeBERTaV3 (He et al., 2021))

- the social media domain, both in English (Conversational BERT (Burtsev et al., 2018), BERTweet (Nguyen et al., 2020)) and in multilingual settings (XLM-T (Barbieri et al., 2022))

- the African domain (AfriBERTa (Ogueji et al., 2021), Afro-XLM-R (Alabi et al., 2022) and XLM-R Naija (Adelani et al., 2021)).

Differences in performance across models may be explained by factors including not only the pretraining domain, but also pretraining data size and preprocessing, model architecture and hyperparameter selection. While it is hard to account for the latter as they are rarely made public, we estimate the impact of the pretraining domain on performance, holding pretraining data size and model architecture constant. To do so, we introduce NAIJAXLM-T (198M), an XLM-R model further pretrained on a random sample of 198 million Nigerian tweets, matching the amount of data used to pretrain XLM-T on multilingual tweets. We adopt the same preprocessing as for XLM-T by removing URLs, tweets with less than 3 tokens, and running the pretraining for one epoch.

Evaluation HSD models are evaluated by their average precision for the hateful class, a standard performance metric in information retrieval which is particularly well-suited when class imbalance is high. For supervised learning, we perform a 90-10 train-test split and conduct a 5-fold cross-validation with 5 learning rates ranging from 1e-5 to 5e-5. Each fold is trained using 3 different seeds. The train-test split is repeated for 10 different seeds, and the evaluation metrics are averaged across the 10 seeds.

5 Results

5.1 Hate speech detection

Evaluating on representative data In Table 3, we evaluate HSD models' performance on three datasets: the holdout set from the train-test splits, the top-scored set and the random set described in Section 3.4. Overall, we observe that the ordering of models' performance remains stable across evaluation sets. However, the striking result is that across the wide range of models considered in this study, the average precision on the random set is substantially lower than that on the holdout and top-scored sets. This finding highlights the risk of considerably overestimating classification performance when evaluating HSD models on a dataset whose characteristics greatly differ from real-world conditions. We now delve more specifically on the impact of the learning frameworks, and of the pretraining and finetuning domains.

[5]https://openai.com/blog/chatgpt

Pretraining data	Finetuning data	Model	Holdout	Top-scored	Random
Multiple domains	None	GPT-3.5, ZSL	-	60.3±2.7	3.1±1.2
	Mixed*	Perspective API	-	60.2±3.5	4.3±2.6
Social Media	HATEXPLAIN	XLM-T	*84.2 ± 0.6*	51.8 ± 0.7	0.6 ± 0.1
	HERDPHOBIA*	XLM-T	*62.0 ± 2.3*	68.9 ± 0.8	3.3 ± 0.6
	HSCODEMIX*	XLM-T	*70.5 ± 3.7*	63.7 ± 1.1	1.9 ± 0.5
Multiple domains	NAIJAHATE Stratified + active sampling (N=4012)	DeBERTaV3	**82.3 ± 2.3**	85.3 ± 0.8	**29.7 ± 4.1**
		XLM-R	76.7 ± 2.5	83.6 ± 0.8	22.1 ± 3.7
		mDeBERTaV3	29.2 ± 2.0	49.6 ± 1.0	0.2 ± 0.0
Social media		Conv. BERT	79.2 ± 2.3	86.2 ± 0.8	22.6 ± 3.6
		BERTweet	**83.6 ± 2.0**	**88.5 ± 0.6**	**34.0 ± 4.4**
		XLM-T	79.0 ± 2.4	84.5 ± 0.9	22.5 ± 3.7
African languages		AfriBERTa	70.1 ± 2.7	80.1 ± 0.9	12.5 ± 2.8
		AfroXLM-R	79.7 ± 2.3	86.1 ± 0.8	24.7 ± 4.0
		XLM-R Naija	77.0 ± 2.5	83.5 ± 0.8	19.1 ± 3.4
Nigerian Twitter		NAIJAXLM-T (198M)	**83.0 ± 2.2**	**90.2 ± 0.6**	**33.1 ± 4.3**
		NAIJAXLM-T (full)	**83.4 ± 2.1**	**89.3 ± 0.7**	**33.7 ± 4.5**

Table 3: Average precision (in %) for the hateful class across models and evaluation sets. Metrics are reported with 95% bootstrapped confidence intervals. All supervised learning classifiers are framed as three-class classifiers, except the models trained on finetuning data marked with an asterisk as the latter is binary (hateful or not). Hyphens indicate the absence of a holdout set. Metrics in italic are calculated on holdout sets that are different from one another and from the NAIJAHATE holdout set.

Learning framework We find that in-domain supervised learning on the NAIJAHATE dataset largely outperforms GPT3.5-based zero-shot learning (ZSL). We also observe that ZSL is on par with supervised learning on existing US and Nigerian-centric benchmarked datasets. Given that the prompt used does not provide a definition of hate speech (App. A.4.4), it implies that GPT3.5 has incorporated enough knowledge from pretraining and reinforcement learning with human feedback to conceptualize and categorize hate speech as well as models finetuned on thousands of examples. Still, it exhibits a rather low performance which is likely due to the predominance of US English in the pretraining data, making it hard to generalize to Nigerian English.

Pretraining domain Overall, the choice of pretrained model has a large impact on downstream performance. In-domain pretraining on Nigerian Twitter generally outperforms the other models both on the top-scored and the random samples, followed by models pretrained on social media and general purpose domains. This result also holds when keeping the architecture and pretraining data size constant, with NaijaXLM-T (198M) yielding significantly better performance than XLM-T. A possible explanation for this result and the dominance of English monolingual models (Conversational BERT, BERTweet, DeBERTaV3) over their multilingual counterparts (mDeBERTa, XLM-T) is the *curse of multilinguality*, whereby per-language performance drops as multilingual models cover more languages (Pfeiffer et al., 2022). Furthermore, we observe that pretraining in the social media domain generally yields larger improvements than in the African linguistic domain. For instance, XLM-R Naija, an XLM-R model further pretrained on news in Nigerian Pidgin English, has a rather poor performance especially on the random set, which is likely due to differences between news and social media lingo as well as the limited share of tweets in Pidgin English. A notable exception to NaijaXLM-T's dominance is BERTweet, a RoBERTa model pretrained from scratch on English tweets, which is on par with NaijaXLM-T on all evaluation sets. Such performance may be explained by the predominance of English on Nigerian Twitter, granting an advantage to English-centric models such as BERTweet or DeBERTaV3. It is also plausible that BERTweet's pretraining data could contain some English tweets from Nigeria. Finally, BERTweet was pretrained from scratch on tweets, implying that its vocabulary is tailored to the social media lingo, contrary to the XLM models.

Finetuning domain In-domain finetuning on the NaijaHate dataset outperforms out-of-domain finetuning on both US-centric (Perspective, HateXPlain) and Nigerian-centric (HERDPhobia and HSCodeMix) datasets. When inspecting classification errors, we find that XLM-T HateXPlain, which is finetuned on US data, classifies as hateful tweets that contain words that are very hateful in the US but not necessarily in Nigeria. For instance, “ya k*ke” means “How are you” in Hausa while k*ke is an ethnic slur for a Jewish person in the US context. As a result, XLM-T HateXplain assigns very high hateful scores to tweets containing this sentence whereas XLM-T NaijaHate does not, underlining the importance of in-domain finetuning. While finetuning on Nigerian Twitter data yields better performance than on US data, it does not ensure high performance, as illustrated by the poor performance of XLM-T HERDPhobia and HSCodeMix. Due to its focus on one specific type of hate against Fulani herdsmen, XLM-T HERDPhobia performs poorly on other types of hate existing in the Nigerian context such as misogyny, underlining the importance of designing a comprehensive annotation scheme covering the most prevalent types of hate.

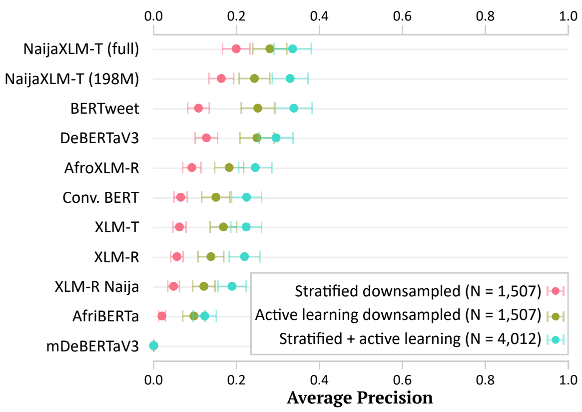

Figure 2: Average precision on the random set across models trained on the downsampled stratified set, the downsampled active learning set and the full training set, composed of the stratified and active learning sets. Error bars indicate 95% bootstrapped confidence intervals.

Finetuning data diversity In light of the higher diversity in the training data sampled through active learning compared to stratified sampling (Table 2), we further investigate the role that finetuning data diversity plays on downstream performance. Specifically, we produce downsampled versions of the stratified and the active learning sets keeping dataset size and class distribution constant. We report the results on the random set in Fig. 2 and on the other evaluation sets in Fig. 5 in the Appendix.

We find that finetuning on more diverse data significantly and consistently improves the average precision across models. The performance gains from diversity are particularly large for models that are not pretrained in the African linguistic domain, such as BERTweet and DeBERTaV3. We also discover that NaijaXLM-T significantly outperforms BERTweet on the less diverse stratified set. This finding indicates that the performance gains from in-domain pretraining may be particularly large when the finetuning data is less diverse, presumably because the lower diversity in the finetuning data is counterbalanced by a higher diversity and domain alignment in the pretraining data, allowing for a better generalization in real-world settings.

5.2 Human-in-the-loop moderation

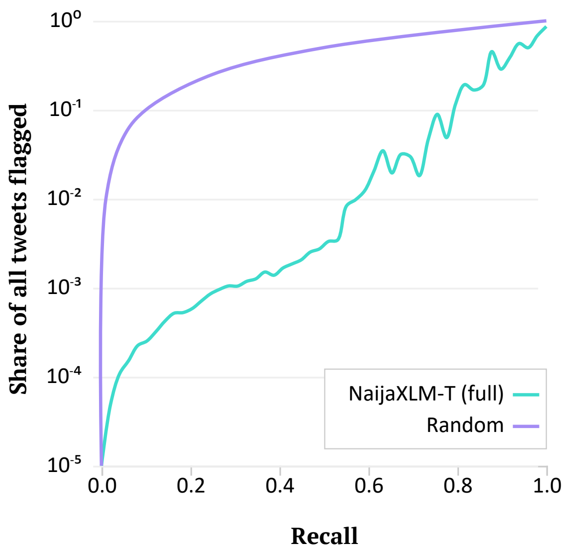

Figure 3: Share of all tweets flagged as hateful as a function of recall on the random set

In light of the performance of hate speech classifiers on Nigerian real-world data, we explore the feasibility of a human-in-the-loop approach to hate speech moderation, in which content likely to contain hate is flagged by a classifier before being reviewed by humans. This approach is motivated by the inability of the best-performing classifiers in our setting to yield both a high precision and a high recall on representative data (Tab. 3 and Fig. 4). Instead of the traditional precision-recall tradeoff, human-in-the-loop moderation implies a *cost-recall tradeoff*, where augmenting the recall comes at the cost of having more flagged content reviewed by humans (Fig. 3).

We find that supervised learning allows to divide the amount of flagged tweets to be annotated by a factor of 60 compared to a random baseline, with 1% of the data to be sent for review in order to achieve a recall of 60%. With an average daily flow of approximately 164,000 tweets on Nigerian Twitter, this translates to an average of 1,640 tweets to be reviewed daily, which is a feasible objective for a small team of moderators. However, as social media adoption increases, the cost of reviewing 1% of all posts could quickly become prohibitive, both financially and in terms of reviewers' harm, highlighting the need for complementary approaches to support the moderation effort.

6 Discussion and conclusion

This work introduced NAIJAHATE, the largest HSD dataset to date in the Nigerian context and the first to contain a representative evaluation set. We also introduced NAIJAXLM-T, the first pretrained language model tailored to Nigerian Twitter.

We demonstrate that evaluating HSD on biased datasets leads to a large overestimation of real-world performance, the latter being rather low (34% average precision). This result expands on past work pointing at the risk of overestimating performance in this context without having quantified it (Arango et al., 2019; Wiegand et al., 2019; Nejadgholi and Kiritchenko, 2020).

Low real-world HSD performance also has implications for hate speech moderation, making automated moderation unfeasible on top of being undesirable for fairness and bias reasons (Gorwa et al., 2020). In this context, we investigate the feasibility of human-in-the-loop moderation, where content likely to be hateful is flagged by a model before being reviewed by humans. We observe a cost-recall tradeoff, where a higher recall comes at the expense of increasing reviewing efforts. We find that 60% recall can be achieved by reviewing 1% of all tweets, which is a feasible goal in the Nigerian Twitter context and for small platforms/communities in general. While using classifiers increases efficiency, our results also illustrate the large costs, both financial and in terms of reviewers harm, of moderating hate speech on larger platforms, which in part explain the low removal rates observed on social media platforms (3-5% on Facebook in 2021 (Giansiracusa, 2021)).

In terms of HSD approaches, we find that in-domain supervised learning significantly outperforms both out-of-domain supervised learning and zero-shot learning. This complements prior work underlining the superiority of supervised learning over zero-shot learning for HSD (Plaza-del arco et al., 2023a) by extending this result to a low-resource setting.

Further, the choice of pretraining model has a large impact on downstream performance. Pretraining on in-domain data that blends the noisy aspect of social media text with the linguistic domain of finetuning tasks yields significantly better performance than pretraining only on the former, even when we hold pretraining data size and model architecture constant. This supports the established finding that in-domain pretraining increases downstream task performance (Gururangan et al., 2020) and complements it by underlining the importance of including all relevant domains during pretraining, both in terms of genre and linguistic focus. We also find that these performance gains are particularly salient when finetuning on less diverse data, potentially facilitated by greater diversity and domain alignment in the pretraining data.

Finally, we observe that using diverse data acquired through active learning yields significant performance gains over stratified sampling. This suggests that annotating a small stratified set and acquiring a larger and more diverse dataset through active learning is preferable to only using stratified data. They also align and complement past findings showing the benefits of active learning to maximize performance at a limited cost (Kirk et al., 2022), including in extremely imbalanced settings like ours (Tonneau et al., 2022), and help better understand them through the prism of diversity.

While the present work demonstrates the low real-world performance of HSD on Nigerian Twitter, there are several possible directions to further improve this performance. Based on the hypothesis that hate is homophilous (Jiang et al., 2023), future work could use network features to improve HSD (Ahmed et al., 2022). Synthetic data could also be used to further increase the number and diversity of examples to train models on (Khullar et al., 2024). Finally, the moderation analysis could be enhanced by taking popularity into account and measuring recall in terms of views of hateful content rather than just posts.

Limitations

Dataset *Limited generalizability to other platforms, timeframes and linguistic domains*: The entirety of our dataset was sampled from a single social media platform for a long yet bounded timeframe. This limits the generalizability of models trained on our dataset to data from other social media platforms and collected in other timespans. Our dataset is also specific to the Nigerian linguistic context and may exhibit poorer performance in other English dialects.

We do not exhaust all targets of hate: The selection of communities often targeted by hate speech and frequent on Nigerian Twitter necessarily leaves out of the analysis other communities even though they are targeted by online hate speech. In the annotation process, we observed for instance that South Africans, British people and Men are also targeted on Nigerian Twitter.

Moderation prior to collection: Our analysis of moderation considers that the hateful content in our random set is representative of all hateful content on Nigerian Twitter. We acknowledge though that some hateful content may have been moderated by Twitter before we collected it and that our estimate of the prevalence of hate speech is necessarily a lower bound estimate.

Experiments *Other prompts could lead to different results*: We craft a prompt using the terms "hateful" and "offensive" (see App. A.4.4 for details) which exhibit good performance in past research for HSD in a ZSL setting (Plaza-del arco et al., 2023b). We do not test other prompts and acknowledge that using other prompts may have an impact on classification performance.

Ethical considerations

Annotator Wellbeing Annotators were provided with clear information regarding the nature of the annotation task before they began their work. They received a compensation of 12 U.S. dollars per hour, which is above the Nigerian minimum wage.

Data Privacy We collected public tweets through the Twitter API according to its Terms and Services. To protect the identity of hateful users and their victims, we will anonymize all tweets in our dataset upon release, replacing all user names by a fixed token @USER.

Acknowledgements

We thank Daniel Barkoczi, Khashau Adenike Eleburuike, Niyati Malhotra, Haaya Naushan, Arinze Alex Okoyeugha, Luis Eduardo San Martin, Fatima Suleiman Sanusi and Philipp Zimmer for their excellent research assistance. We also thank Alice Grishchenko for outstanding visual design work. This work is part of a broader project conducted by the World Bank Development Impact (DIME) department. The study was supported by funding from the United Kingdom's Foreign Commonwealth and Development Office (FCDO) and the World Bank's Research Support Budget. This work was also supported in part through the NYU IT High Performance Computing resources, services, and staff expertise. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations or those of the Executive Directors of the World Bank or the governments they represent.

References

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 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.

Zo Ahmed, Bertie Vidgen, and Scott A Hale. 2022. Tackling racial bias in automated online hate detection: Towards fair and accurate detection of hateful users with geometric deep learning. *EPJ Data Science*, 11(1):8.

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*.

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 45–54.

Josh Attenberg, Prem Melville, and Foster Provost. 2010. A unified approach to active dual supervision for labeling features and examples. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 40–55. Springer.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

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, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhreva, and Marat Zaynutdinov. 2018. DeepPavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127, Melbourne, Australia. Association for Computational Linguistics.

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. *Eleventh International AAAI Conference on Web and Social Media*, 11(1).

Christian Ezeibe. 2021. Hate speech and election violence in nigeria. *Journal of Asian and African Studies*, 56(4):919–935.

RO Farinde and HO Omolaiye. 2020. A socio-pragmatic investigation of language of insults in the utterances of yoruba natives in nigeria. *Advances in Language and Literary Studies*, 11(6):1–6.

W Ferroggiaro. 2018. Social media and conflict in nigeria: A lexicon of hate speech terms.

Iginio Gagliardone, Matti Pohjonen, Zenebe Beyene, Abdissa Zerai, Gerawork Aynekulu, Mesfin Bekalu, Jonathan Bright, Mulatu Alemayehu Moges, Michael Seifu, Nicole Stremlau, et al. 2016. Mechachal: Online debates and elections in ethiopia-from hate speech to engagement in social media. *Available at SSRN 2831369*.

Sayan Ghosh, Dylan Baker, David Jurgens, and Vinodkumar Prabhakaran. 2021. Detecting cross-geographic biases in toxicity modeling on social media. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 313–328, Online. Association for Computational Linguistics.

N Giansiracusa. 2021. Facebook uses deceptive math to hide its hate speech problem. *Wired*.

Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Tarleton Gillespie. 2020. Content moderation, ai, and the question of scale. *Big Data & Society*, 7(2):2053951720943234.

Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945.

Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2024. An investigation of large language models for real-world hate speech detection. *arXiv preprint arXiv:2401.03346*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 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. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Julie Jiang, Luca Luceri, Joseph B Walther, and Emilio Ferrara. 2023. Social approval and network homophily as motivators of online toxicity. *arXiv preprint arXiv:2310.07779*.

Rafael Jiménez Durán. 2021. The economics of content moderation: Theory and experimental evidence from hate speech on twitter. *Available at SSRN 4044098*.

Aman Khullar, Daniel Nkemelu, V Cuong Nguyen, and Michael L Best. 2024. Hate speech detection in limited data contexts using synthetic data generation. *ACM Journal on Computing and Sustainable Societies*, 2(1):1–18.

Hannah Kirk, Bertie Vidgen, and Scott Hale. 2022. Is more data better? re-thinking the importance of efficiency in abusive language detection with transformers-based active learning. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 52–61, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Alyssa Lees, Vinh Q Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3197–3207.

Peiyu Li. 2021. *Achieving hate speech detection in a low resource setting*. Ph.D. thesis, Utah State University.

Ruth Maclean. 2021. Nigeria bans twitter after president's tweet is deleted. *The New York Times*.

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15009–15018.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Dan Milmo. 2021. Frances haugen: 'i never wanted to be a whistleblower. but lives were in danger'. *The Guardian*.

Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A measurement study of hate speech in social media. In *Proceedings of the 28th ACM conference on hypertext and social media*, pages 85–94.

Karsten Müller and Carlo Schwarz. 2021. Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, 19(4):2131–2167.

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.

Isar Nejadgholi and Svetlana Kiritchenko. 2020. On cross-dataset generalization in automatic detection of online abuse. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 173–183, Online. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Daniel Nkemelu, Harshil Shah, Michael Best, and Irfan Essa. 2022. Tackling hate speech in low-resource languages with context experts. In *Proceedings of the 2022 International Conference on Information and Communication Technologies and Development*, pages 1–11.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech 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 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

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

Paul Ayodele Onanuga. 2023. # arewaagainstlgbtq discourse: a vent for anti-homonationalist ideology in nigerian twittersphere? *African Identities*, 21(4):703–725.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. 2023a. Leveraging label variation in large language models for zero-shot text classification.

Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. 2023b. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada. Association for Computational Linguistics.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

Paul Röttger, Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Data-efficient strategies for expanding hate speech detection into under-resourced languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5674–5691, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. 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 41–58, Online. Association for Computational Linguistics.

Sumegh Roychowdhury and Vikram Gupta. 2023. Data-efficient methods for improving hate speech detection. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 125–132, Dubrovnik, Croatia. Association for Computational Linguistics.

Manuel Tonneau, Dhaval Adjodah, Joao Palotti, Nir Grinberg, and Samuel Fraiberger. 2022. Multilingual detection of personal employment status on Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6564–6587, Dublin, Ireland. Association for Computational Linguistics.

Collins Udanor and Chinatu C Anyanwu. 2019. Combating the challenges of social media hate speech in a polarized society: A twitter ego lexalytics approach. *Data Technologies and Applications*.

UN. 2019. Plan of action on hate speech.(2019). *Technical report*.

Ellen M Voorhees, Donna K Harman, et al. 2005. *TREC: Experiment and evaluation in information retrieval*, volume 63. MIT press Cambridge.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. 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 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

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 *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 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

A Experimental details

A.1 Data collection

A.1.1 Word lists

In this section, we provide the detailed lists of slurs and communities (Table 4). Summary statistics on the number of words per category can be found in Table 7.

Hate words We first build a list of 89 Nigeria-specific slurs, which are referred to as *hate words* thereafter. To do so, we rely on lexicons from past work on the topic (Udanor and Anyanwu, 2019; Farinde and Omolaiye, 2020), Nigerian online dictionaries such as Naija Lingo as well as local knowledge from our Nigerian colleagues. The final list contains two types of words: regular slurs (n=84) and words combining a slur and community name, such as “fulanimal” (n=5). The list of 84 regular slurs contains 28 Yoruba words, 26 English words, 12 Hausa words, 11 Pidgin words and 7 Igbo words. We detail the full list of hate words in Table 4.

Community names Second, we define a list of names of communities that are often targeted by hate speech in Nigeria, again relying on past work (Onanuga, 2023) and local knowledge from Nigerian colleagues. We build an initial list (see Table 5 for the full list of considered and retained community names) and we then restrict this initial list of community names to the names that are most frequently mentioned on Nigerian Twitter. This approach yields 12 communities, including 5 ethnic groups (Yoruba, Igbo, Hausa, Fulani, Herdsmen, 2 religious groups (Christians and Muslims), 3 regional groups (Northern, Southern, Eastern) and 2 groups on gender identity and sexual orientation (Women and LGBTQ+). As mentioned earlier, Herdsmen are not a ethnic group per se but this term refers exclusively to Fulani herdsmen in the Nigerian context, hence the categorization as an ethnic group. For each of these groups, we list the different denominations of each group as well as their plural form and combine it in regular expressions (see Table 6). Finally, we also identify 5 words combining a community name with a derogatory word (e.g., “fulanimal”) that we coin *combined word* thereafter. Since some targets were very rare in the annotated data, we decided to bundle the 12 communities into 5 groups: North (Northern, Hausa, Fulani, Herdsmen), South (Southern, Yoruba), East (Igbo, Biafra), Women, Muslim and Other (Christian, LGBTQ+).

A.1.2 Sampling and evaluation sets

We draw two distinct random samples of 100 million tweets each, one for sampling and model training D_s and the other one for evaluation D_e.

A.1.3 Stratified sample

As previously stated, the extreme imbalance in our classification task makes random sampling ineffective and prohibitively expensive. With the aim to build high-performing classifiers at a reasonable cost, we build and annotate a stratified sample of tweets from D_s. We use three different sampling strategies to build this stratified sample. First, for each possible combination of community name and hate word, we sample up to 4 tweets that both contain the respective hate word and match with the respective community regular expression. The subset of tweets containing both the hate word and the community regular expression may be smaller than 4 and we sample the full subset in that case. Second, for each combined word W, we randomly sample 50 tweets containing W. Some combined words occur at a very low frequency such that the sample size is sometimes smaller than 50. Finally, for each community, we draw 50 random tweets matching with the community regular expression, in order to avoid having a classifier that associates the community name with hate speech.

This yields a stratified sample of 1,607 tweets annotated as either hateful, offensive or neutral.

A.1.4 Active learning sample

Each active learning iteration samples a total of 100 tweets. The type of active learning method we employ is called *certainty sampling* and consists in sampling instances at the top of the score distribution in order to annotate false positives and maximize precision. Specifically, each iteration i consists of:

- Model training: we train a model on all of the labels we have, that is the stratified sample and the combination of all Active Learning samples from prior iterations
- Inference: we then deploy this model on D_s and rank all tweets based on their BERT confidence score.
- Sampling and annotating: we define 5 rank buckets as: $[1, 10^3]$, $[10^3, 10^4]$, $[10^4, 10^5]$,

<table>
<tr><th>Hate Keyword</th><th>Language</th><th>Translation</th><th>Source</th></tr>
<tr><td>stupid</td><td>english</td><td></td><td></td></tr>
<tr><td>animal|animals</td><td>english</td><td></td><td></td></tr>
<tr><td>baboon|baboons</td><td>english</td><td></td><td></td></tr>
<tr><td>bastard|bastards</td><td>english</td><td></td><td></td></tr>
<tr><td>bitch|bitches</td><td>english</td><td></td><td></td></tr>
<tr><td>bum|bums</td><td>english</td><td></td><td></td></tr>
<tr><td>cockroach|cockroaches</td><td>english</td><td></td><td></td></tr>
<tr><td>coconut head|coconut heads</td><td>english</td><td></td><td></td></tr>
<tr><td>disgusting</td><td>english</td><td></td><td></td></tr>
<tr><td>dog|dogs</td><td>english</td><td></td><td></td></tr>
<tr><td>dumb|dumb</td><td>english</td><td></td><td></td></tr>
<tr><td>fanatic|fanatics</td><td>english</td><td></td><td></td></tr>
<tr><td>fool|fools</td><td>english</td><td></td><td></td></tr>
<tr><td>idiot|idiots</td><td>english</td><td></td><td></td></tr>
<tr><td>liar|liars</td><td>english</td><td></td><td></td></tr>
<tr><td>moron|morons</td><td>english</td><td></td><td></td></tr>
<tr><td>parasite|parasites</td><td>english</td><td></td><td></td></tr>
<tr><td>pig|pigs</td><td>english</td><td></td><td></td></tr>
<tr><td>primitive|primitives</td><td>english</td><td></td><td></td></tr>
<tr><td>rape|rapes|raping</td><td>english</td><td></td><td></td></tr>
<tr><td>scum|scums</td><td>english</td><td></td><td></td></tr>
<tr><td>shit|shits</td><td>english</td><td></td><td></td></tr>
<tr><td>slut|sluts</td><td>english</td><td></td><td></td></tr>
<tr><td>useless</td><td>english</td><td></td><td></td></tr>
<tr><td>vulture|vultures</td><td>english</td><td></td><td></td></tr>
<tr><td>whore|whores</td><td>english</td><td></td><td></td></tr>
<tr><td>aboki|abokai</td><td>hausa</td><td>"friend; used by a non-Hausa person may be derogatory"</td><td>https://www.bellanaija.com/2020/04/twitter-aboki-derogatory-term/</td></tr>
<tr><td>arne|arna</td><td>hausa</td><td>"pagan - used by muslims to reference christians in the north"</td><td></td></tr>
<tr><td>ashana</td><td>hausa</td><td>prostitute</td><td>http://naijalingo.com/words/ashana</td></tr>
<tr><td>barawo|barayi</td><td>hausa</td><td>thief</td><td>http://naijalingo.com/words/barawo</td></tr>
<tr><td>bolo</td><td>yoruba</td><td>fool</td><td>http://naijalingo.com/words/bolo</td></tr>
<tr><td>kafir|kafirai</td><td>hausa</td><td>"used by muslims to refer to non-muslims"</td><td></td></tr>
<tr><td>mallam|malamai</td><td>hausa</td><td>"teacher; used specifically in southern Nigeria in derogatory manner to refer to all Northerners; in Northern Nigeria, is used as a mark of respect"</td><td></td></tr>
<tr><td>malo|malos</td><td>hausa</td><td>fool</td><td></td></tr>
<tr><td>mugu</td><td>hausa</td><td>wicked/evil</td><td>http://naijalingo.com/words/mugu</td></tr>
<tr><td>mugun</td><td>hausa</td><td>fool</td><td>http://naijalingo.com/words/mugun</td></tr>
<tr><td>mungu</td><td>hausa</td><td>fool</td><td>http://naijalingo.com/words/mungu</td></tr>
<tr><td>wawa|wawaye</td><td>hausa</td><td>idiot</td><td></td></tr>
<tr><td>zuwo</td><td>hausa</td><td>fool</td><td>http://naijalingo.com/words/zuwo</td></tr>
<tr><td>anuofia|ndi anofia</td><td>igbo</td><td>wild animal</td><td></td></tr>
<tr><td>aturu</td><td>igbo</td><td>sheep</td><td>Udanor and Anyanwu (2019)</td></tr>
<tr><td>efulefu|ndi fulefu</td><td>igbo</td><td>worthless man</td><td></td></tr>
<tr><td>ewu</td><td>igbo</td><td></td><td></td></tr>
<tr><td>imi nkita</td><td>igbo</td><td>dog nose</td><td>https://www.vanguardngr.com/2019/11/of-yariba-nyamiri-and-aboki/</td></tr>
<tr><td>onye nzuzu|ndi nzuzu</td><td>igbo</td><td></td><td></td></tr>
<tr><td>onye oshi|ndi oshi</td><td>igbo</td><td>thief</td><td>http://naijalingo.com/words/onye-oshi</td></tr>
<tr><td>ashawo|ashawos</td><td>pidgin</td><td>prostitute</td><td>http://naijalingo.com/words/ashawo</td></tr>
<tr><td>ashewo|ashewos|awon ashewo</td><td>pidgin</td><td>prostitute</td><td>http://naijalingo.com/words/ashewo</td></tr>
<tr><td>ashy</td><td>pidgin</td><td>dirty</td><td></td></tr>
<tr><td>mumu|mumus</td><td>pidgin</td><td>idiot</td><td></td></tr>
<tr><td>mumuni</td><td>pidgin</td><td>very stupid person</td><td>http://naijalingo.com/words/mumuni</td></tr>
<tr><td>sharrap</td><td>pidgin</td><td>shut up</td><td>http://naijalingo.com/words/sharrap</td></tr>
<tr><td>tief</td><td>pidgin</td><td>thief</td><td>http://naijalingo.com/words/tief</td></tr>
<tr><td>tiff</td><td>pidgin</td><td>thief</td><td>http://naijalingo.com/words/tiff</td></tr>
<tr><td>waka jam</td><td>pidgin</td><td>an insult/curse towards you and loved ones</td><td>http://naijalingo.com/words/waka-jam</td></tr>
<tr><td>agba iya|awon agba iya</td><td>yoruba</td><td>older person, who despite his age, is still useless</td><td>https://www.legit.ng/1031944-8-insults-yoruba-mothers-use-will-reset-brain.html</td></tr>
<tr><td>agbaya</td><td>yoruba</td><td>derogatory word against old people</td><td></td></tr>
<tr><td>agbero|agberos|awon agbero</td><td>yoruba</td><td>used to describe manual laborers from lower economic classes; sometimes deployed on twitter for ad hominem attacks</td><td>https://en.wiktionary.org/wiki/agbero</td></tr>
<tr><td>akpamo</td><td>yoruba</td><td>fool</td><td>http://naijalingo.com/words/akpamo</td></tr>
<tr><td>apoda|awon apoda</td><td>yoruba</td><td>who is confused, lost direction</td><td>https://www.legit.ng/1031944-8-insults-yoruba-mothers-use-will-reset-brain.html</td></tr>
<tr><td>arindin|awon arindi</td><td>yoruba</td><td>acts like an idiot</td><td>https://www.nairaland.com/3237758/she-called-him-arindin-sitting</td></tr>
<tr><td>arro</td><td>yoruba</td><td>stupid person</td><td>http://naijalingo.com/words/arro</td></tr>
<tr><td>atutupoyoyo</td><td>yoruba</td><td>ugly being</td><td>http://naijalingo.com/words/atutupoyoyo</td></tr>
<tr><td>ayama</td><td>yoruba</td><td>disgusting</td><td>http://naijalingo.com/words/ayama</td></tr>
<tr><td>ayangba</td><td>yoruba</td><td>prostitute</td><td></td></tr>
<tr><td>didirin|awon didirin</td><td>yoruba</td><td>stupid</td><td>https://www.legit.ng/1031944-8-insults-yoruba-mothers-use-will-reset-brain.html</td></tr>
<tr><td>eyankeyan</td><td>yoruba</td><td>synonym to lasan</td><td>Farinde and Omolaiye (2020)</td></tr>
<tr><td>lasan</td><td>yoruba</td><td>ordinary; when combined with a community name, may mean that this group is inferior to Yorubas</td><td>Farinde and Omolaiye (2020)</td></tr>
<tr><td>malu|awon malu</td><td>yoruba</td><td>cow</td><td></td></tr>
<tr><td>obun|awon obun</td><td>yoruba</td><td>dirty</td><td>https://www.legit.ng/1031944-8-insults-yoruba-mothers-use-will-reset-brain.html</td></tr>
<tr><td>ode|awon ode</td><td>yoruba</td><td>stupid</td><td>https://www.legit.ng/1031944-8-insults-yoruba-mothers-use-will-reset-brain.html</td></tr>
<tr><td>odoyo</td><td>yoruba</td><td>very stupid person</td><td>http://naijalingo.com/words/odoyo</td></tr>
<tr><td>ole|awon ole</td><td>yoruba</td><td>thief</td><td>Udanor and Anyanwu (2019)</td></tr>
<tr><td>olodo|olodos|awon olodo</td><td>yoruba</td><td>stupid</td><td>https://www.legit.ng/1031944-8-insults-yoruba-mothers-use-will-reset-brain.html</td></tr>
<tr><td>oloshi|awon oloshi</td><td>yoruba</td><td>unfortunate, who does rubbish a lot, criminal</td><td>https://www.legit.ng/1031944-8-insults-yoruba-mothers-use-will-reset-brain.html</td></tr>
<tr><td>omo ale|awon omo ale</td><td>yoruba</td><td>bastard</td><td>Farinde and Omolaiye (2020)</td></tr>
<tr><td>oponu|awon aoponu</td><td>yoruba</td><td>idiot</td><td>https://www.legit.ng/1031944-8-insults-yoruba-mothers-use-will-reset-brain.html</td></tr>
<tr><td>ota|awon ota</td><td>yoruba</td><td>enemy</td><td>http://naijalingo.com/words/ota</td></tr>
<tr><td>owo</td><td>yoruba</td><td>fool</td><td></td></tr>
<tr><td>suegbe</td><td>yoruba</td><td>idiot</td><td>http://naijalingo.com/words/suegbe</td></tr>
<tr><td>werey|awon weyre</td><td>yoruba</td><td>crazy, mad</td><td>http://naijalingo.com/words/werey</td></tr>
<tr><td>yeye|awon akin yeye</td><td>yoruba</td><td>useless</td><td>Udanor and Anyanwu (2019)</td></tr>
<tr><td>jeri</td><td>pidgin</td><td>fool</td><td>http://naijalingo.com/words/jeri</td></tr>
<tr><td>shalam</td><td>pidgin</td><td></td><td></td></tr>
<tr><td>biafrat|biafraud</td><td>combined</td><td>targeting Biafra</td><td></td></tr>
<tr><td>fulanimal</td><td>combined</td><td>targeting Fulanis</td><td></td></tr>
<tr><td>yorubastard|yariba|yorobber</td><td>combined</td><td>targeting Yorubas</td><td></td></tr>
<tr><td>baby factory|baby factories</td><td>combined</td><td>targeting Igbo</td><td></td></tr>
<tr><td>niyamiri</td><td>combined</td><td></td><td></td></tr>
</table>

Table 4: Slurs used in the Nigerian context

Community word	Frequency	Retained
christian	1.88E-03	yes
muslim	2.10E-03	yes
northern	1.25E-03	yes
southern	5.30E-04	yes
hausa	7.12E-04	yes
fulani	8.81E-04	yes
yoruba	1.37E-03	yes
igbo	1.52E-03	yes
women	4.93E-03	yes
biafra	1.60E-03	yes
arewa	1.30E-03	yes
LGBTQ+	1.12E-03	yes
herdsmen	7.49E-04	yes
eastern	2.09E-04	yes
tiv	3.98E-05	no
kanuri/beriberi	1.82E-05	no
ibibio	1.45E-05	no
ijaw/izon	6.02E-05	no
buharist	1.15E-04	no
ipobite	6.22E-08	no
arne	3.82E-06	no
transgender	3.83E-05	no
middle belt	3.45E-05	no
jukun	6.93E-06	no
Niger Delta	2.42E-04	no
yorubawa	4.07E-07	no
berom	4.84E-05	no

Table 5: List of considered community words and their frequency in the Twitter dataset. The frequency for each word corresponds to the number of tweets containing the word divided by the total number of tweets.

$[10^5, 10^6], [10^6, 10^7]$. We then sample n tweets per rank bucket and annotate this sample.

We conduct a total of 25 iterations, of which 10 are conducted on the subset of D_s containing community keywords and 15 on the full D_s. In our active learning process, three separate phases can be distinguished:

- iterations 1-10:
  - the sampling is done on the subset of D_s containing community words
  - the active learning process is done separately for the hateful and the offensive classes
  - the value of n equals 10
  - the overall sample size per iteration is 100 and equals to 5 buckets x n=10 x 2 classes (hateful and offensive)
- iterations 10-19
  - the sampling is done on the full sampling set D_s
  - the active learning process is done separately for the hateful and the offensive classes
  - the value of n equals 10
  - the overall sample size per iteration is 100 and equals to 5 buckets x n=10 x 2 classes (hateful and offensive)
- iterations 20-24
  - the sampling is done on the full sampling set D_s
  - the active learning process is done only for the hateful class
  - the value of n equals 20
  - the overall sample size per iteration is 100 and equals to 5 buckets x n=20 x 1 class (hateful)

A.2 Annotation

A.2.1 Annotation team

The annotation team was composed of a Hausa man, a Hausa-Fulani woman, an Igbo man and a Yoruba woman.

A.2.2 Annotation guidelines

Offensive tweets For tweets to be offensive, but not hateful, a tweet must satisfy all of the following criteria.

- The hate keyword is being used as pejorative towards another individual or group, and this group is not one of our communities.
  - A personal attack against another individual, that does not mention a protected attribute such as, race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease.
  - An insult towards a group based on non-protected attributes, such as, hobbies, fandom (e.g., sports, comic books).
- It is not offensive if the hate keyword is not being used on an individual or group.

Community	Regular expression
christian	christian|christians
muslim	muslim|muslims|islam|islamic
northern	northern|northerner|northerners|arewa|almajiri
southern	southern|southerner|southerners
hausa	hausa|hausas
fulani	fulani|fulanis
yoruba	yoruba|yorubas
igbo	igbo|ibo|ibos|igbos
women	women|woman|girl|girls|female|females
lgbt	lgbt|lgbtq|lgbtq+|gay|gays|lesbian|lesbians|transgender|transgenders
herdsmen	herdsmen|herdsman
eastern	eastern|easterner|easterners|biafra

Table 6: Community Regex Mapping

Word category	Number of words
Community names	12
English hate words	26
Non-English hate words	58
Combined words	5
Total number of hate words (in all languages)	84
Total number of hate words, including combined words (in all languages)	89

Table 7: Summary statistics on the number of words per category

  - Not offensive if directed towards inanimate objects, abstract concepts (that do not have religious or cultural significance) or animals (unless the animal is used as a negative metaphor to describe a community). We define these as "out-of-scope entities" (Röttger et al., 2021).

- It is not offensive if the hate word is self-referential. This would account for some types of sarcasm, or humour via self deprecation.

- It is not offensive if the hate word is used for emphasis without being directed towards an individual or group. Several offensive words such as "shit" or "stupid" can be used as exclamations.

- If the hate word is being used ambiguously (not recognizable as pejorative) then it is offensive if your answer is yes to one of these questions.

  - Can you imagine that someone might be offended by this? (err on the side of caution, aim for the lower bound)
  - Would Twitter potentially detect it as an insult and make the user verify before posting?

Hateful tweets This section is adapted from (Waseem and Hovy, 2016; Basile et al., 2019) and Facebook Community Standards[6]. For tweets to be hateful, instead of merely offensive, the tweet must satisfy one or more of the following criteria:

- Uses a sexist, racial or homophobic slur.

  - Misogyny/Sexist slurs to be defined as a statement that expresses hate towards women in particular (in the form of insulting, sexual harassment, threats of violence, stereotype, objectification and negation of male responsibility).
  - Racial slurs to be defined as an insult that is designed to denigrate others based on their race or ethnicity.
  - Homophobic slurs to be defined as an insult that is designed to denigrate other on the basis of sexuality. This includes slurs targeted towards specific LGBTQ+ communities, such as transphobic slurs.
  - Usage of slur must not constitute a "reclaiming" of negative terms by the community in question. For instance, the n* word or "fag" or "bitch".

- Attacks a minority.

  - Minorities to be defined as a group based on protected characteristics: race, ethnicity, national origin, disability, religious

[6] https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/

affiliation, caste, sexual orientation, sex, gender identity and serious disease.

- Attack to be defined as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation.
- Seeks to silence a minority.
- Criticizes a minority (without a well founded argument).
  * Criticizes a minority and uses a straw man argument.
  * Blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims.
- Negatively stereotypes a minority. Negative stereotypes to be defined as dehumanizing comparisons that have historically been used to attack, intimidate, or exclude specific groups.
- Promotes, but does not directly use, hate speech or violent crime.
  * Shows support of problematic hashtags. e.g.,"#BanIslam"
  * Defends xenophobia, racism, sexism, homophobia or other types of intolerance and bigotry.
- If it is a retweet it must indicate support for the original tweet. People sometimes share content that includes someone else's hate speech to condemn it or raise awareness.

A.3 Language distribution

We asked the annotators to characterize the language of a random sample of 500 tweets, both for the stratified and active learning sets and for the random sample. We report the language distribution in Table 8.

A.4 Models

A.4.1 Number of parameters

Conversational BERT has 110 million parameters. The XLM models, BERTweet and AfriBERTa have 125 million parameters. The DeBERTaV3 models have 86 million parameters. The number of parameters for GPT3.5 is undisclosed by OpenAI.

A.4.2 Pretraining of NaijaXLM-T

We followed Alabi et al. (2022) and performed an adaptive fine tuning of XLM-R (Conneau et al., 2020) on our Twitter dataset. We kept the same vocabulary as XLM-R and trained the model for one epoch, using 1% of the dataset as validation set. The training procedure was conducted in a distributed environment, for approximately 10 days, using 4 nodes with 4 RTX 8000 GPUs each, with a total batch size of 576.

A.4.3 Supervised Learning

Hyperparameter tuning Hyperparameter tuning was conducted in a 5-fold cross validation training. A grid search was run testing different learning rates (from 1e-5 to 5e-5). The cross validation trainings were conducted for 10 epochs. The batch size used was 8, and three different seeds were used for each learning rate. We used F1-score as early stopping metric for hate speech detection models. The best results were averaged across the seeds, and the best combination after the grid search was picked as the resulting model.

Computing infrastructure For supervised learning, we used either V100 (32GB) or RTX8000 (48GB) GPUs for finetuning. The average runtime for finetuning is 45 minutes. Inferences from off-the-shelf models were ran locally on a laptop CPU.

A.4.4 Off-the-shelf models

Perspective API We used the IDENTITY_ATTACK category for HSD with Perspective API as it is the closest to our hate speech definition. This is a binary classification problem and the API outputs a score between 0 and 1. To determine the performance of the API at binary HSD, we choose the classification threshold as the one that maximizes the F1 score. The inferences were run on February 1, 2024.

GPT3.5 We use the *gpt-3.5-turbo-0613* model. The prompt used for zero-shot predictions with this model is: *"Now consider this message : '[TWEET]' Respond 0 if this message is neutral, 1 if this message is offensive and 2 if this message is hateful. It is very important that you only respond the number (e.g., '0', '1' or '2')."*

The prompt is run 5 times for each tweet. We then define the hateful score as the share of the 5 times for which the model predicted that the tweet was hateful. We then use this score to compute the average precision. We use all default values for the main hyperparameters, including 1 for temperature.

	Stratified + active learning sets	Random set
English	74.2	77
English & Nigerian Pidgin	11	1.5
English & Yoruba	4.2	-
Nigerian Pidgin	3.6	7.3
English & Hausa	2.2	-
Hausa	1	1.2
Yoruba	-	1
URLs	-	6
Emojis	-	2.3

Table 8: Share of each language across datasets (in %). Hyphens indicate that the value is under 1%.

A.4.5 Evaluation results

We provide the diversity results for the holdout and the top-scored sets in Fig. 5. We also provide the precision-recall curve for NaijaXLM-T on the random set in Fig. 4.

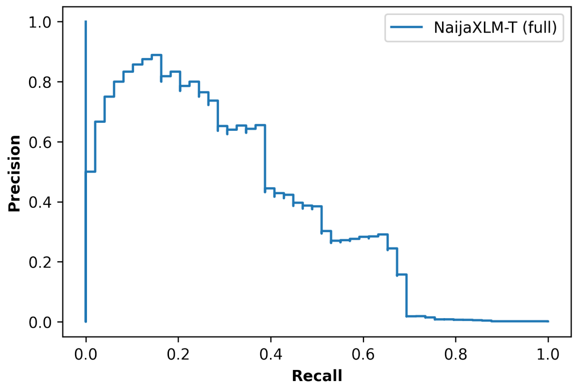

Figure 4: Precision-recall curve on the random set

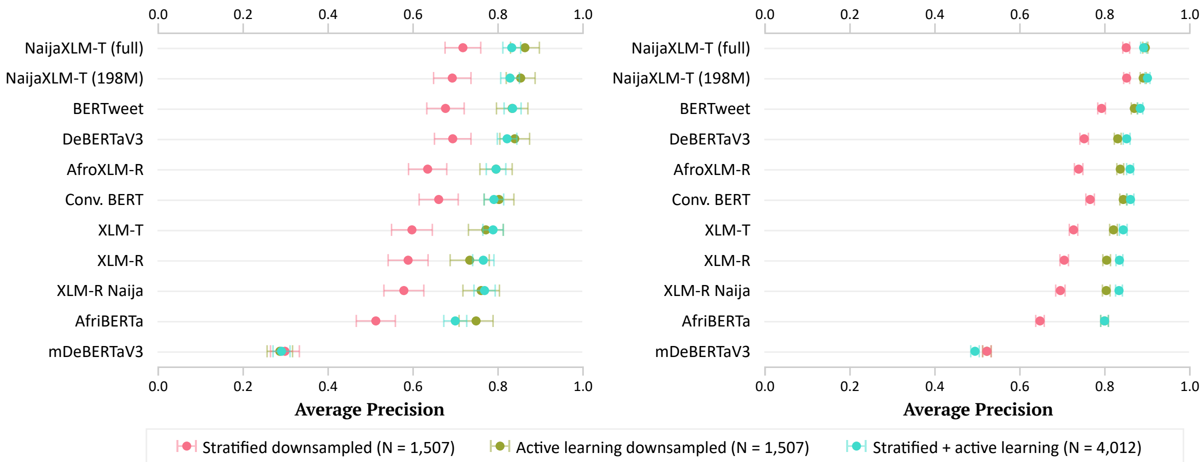

Figure 5: Average precision on the holdout and top-scored sets across models trained on the downsampled stratified set, the downsampled active learning set and the full training set, composed of the stratified and active learning sets. Error bars indicate 95% bootstrapped confidence intervals.