A Federated Approach to Few-Shot Hate Speech Detection for Marginalized Communities

Haotian Ye[1,2], **Axel Wisiorek**[1,2], **Antonis Maronikolakis**[1,2],
Özge Alaçam[1,3], **Hinrich Schütze**[1,2]

[1]Center for Information and Language Processing, LMU Munich
[2]Munich Center for Machine Learning (MCML)
[3]Computational Linguistics, Department of Linguistics, Bielefeld University
{yehao, wisiorek, antmarakis}@cis.lmu.de
oezge.alacam@uni-bielefeld.de

Disclaimer: This paper includes examples of hateful or offensive language used solely for illustrative purposes. These examples may be upsetting to some readers and do not represent the views or beliefs of the authors.

Abstract

Despite substantial efforts, detecting and preventing hate speech online remains an understudied task for marginalized communities, particularly in the Global South, which includes developing societies with increasing internet penetration. In this paper, we aim to provide marginalized communities in societies where the dominant language is low-resource with a privacy-preserving tool to protect themselves from online hate speech by filtering offensive content in their native languages. Our contributions are twofold: 1) we release REACT (**RE**sponsive hate speech datasets **A**cross **C**on**T**exts), a collection of high-quality, culture-specific hate speech detection datasets comprising multiple target groups and low-resource languages, curated by experienced data collectors; 2) we propose a few-shot hate speech detection approach based on federated learning (FL), a privacy-preserving method for collaboratively training a central model that exhibits robustness when tackling different target groups and languages. By keeping training local to user devices, we ensure data privacy while leveraging the collective learning benefits of FL. We experiment with both multilingual and monolingual pre-trained representation spaces as backbones to examine the interaction between FL and different model representations. Furthermore, we explore personalized client models tailored to specific target groups and evaluate their performance. Our findings indicate the overall effectiveness of FL across different target groups, and point to personalization as a promising direction.

1 Introduction

Combating online hate is a crucial aspect of content moderation, with prevailing solutions often relying on machine learning models trained on large-scale datasets (Pitenis et al., 2020; Röttger et al., 2021; Nozza, 2021). However, these efforts and the resources required are largely limited to a few high-resource languages, such as English and German. While multilingual hate speech datasets have been developed (Röttger et al., 2022; Das et al., 2022), a significant portion of the world's low-resource languages and their users remain unprotected from online abuse. A key challenge in hate speech detection lies in its inherently subjective and context-dependent nature, which varies not only at the individual level but also across cultures and regions. The issue is exacerbated by the lack of expertise of annotators on marginalized target groups, as many studies rely on crowdsourcing for data collection, often resulting in a disconnect between those labeling the data and those directly affected by hate speech (Davidson et al., 2019; Sap et al., 2019). Additionally, both language and hate speech constantly evolve, with new expressions and terminology regularly emerging.

To address these challenges, we develop high-quality, culturally relevant datasets that reflect the experiences of marginalized communities. This is achieved through a prompt-based data collection procedure, carried out by data collectors proficient in the target languages and familiar with the nuances of hate speech directed at marginalized groups within their respective contexts. The result is REACT, a set of localized, context-aware datasets containing positive, neutral, and hateful sentences across various low-resource languages. We release REACT under CC BY-SA 4.0. [1]

[1] https://huggingface.co/datasets/htyeh/REACT

Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025), pages 631–651
November 8-9, 2025 ©2025 Association for Computational Linguistics

One key limitation of current hate speech filtering solutions is their reliance on centralized, server-side processing. In such setups, user data must be transmitted to remote servers for analysis, restricting individual control over the content being filtered. Moreover, centralized models are less adaptable to highly specific targets, particularly in low-resource language settings.

To overcome this, we propose the use of federated learning (FL) (McMahan et al., 2017), a decentralized machine learning paradigm where multiple users collaboratively train a central model without sharing raw data. FL operates in two iterative stages: first, client devices receive the current server model and train it locally on private data; then, updates are sent back to the server, aggregated, and used to improve the server model. This decentralized approach not only preserves user privacy but also enables rapid adaptation to culturally specific hate speech patterns.

Our work aims to tackle the following research questions. **RQ1**: Can zero-shot or few-shot learning effectively detect hate speech in low-resource languages? **RQ2**: If not, can FL bridge this performance gap? **RQ3**: Given the specificity of hate speech, does client personalization improve over zero- or few-shot learning in low-resource settings?

2 Related Work

2.1 Toxic and offensive language datasets

Earlier efforts in the detection of toxic and offensive language, including hate speech, have contributed to the curation of diverse datasets, predominantly in English (Waseem and Hovy, 2016; Wulczyn et al., 2017; Zhang et al., 2018) and to a lesser extent in other high-resource languages, like German and Arabic (Mandl et al., 2019; Mulki et al., 2019). More recent work has developed datasets with more fine-grained details, such as different types of abuse (Sap et al., 2020; Guest et al., 2021) and target groups (Grimminger and Klinger, 2021; Maronikolakis et al., 2022). In a related manner, Dixon et al. (2018) and Röttger et al. (2021) adopt a template-based data generation process to construct hate speech datasets categorized by targeted subgroups. Recognizing the need for broader linguistic coverage, recent initiatives have expanded data collection to include multiple languages, including low-resource ones (Röttger et al., 2022; Das et al., 2022; Dementieva et al., 2024; Bui et al., 2025), which is crucial for developing robust hate speech detection systems for underrepresented languages. Notably, Muhammad et al. (2025) introduce *AfriHate*, an offensive speech dataset covering 15 low-resource languages and dialects spoken in Africa.

2.2 Hate speech detection

Transformer-based (Vaswani et al., 2017) language models have emerged as the backbone of many natural language processing tasks. This trend extends to hate speech detection, where various Transformer-based models have been employed (Mozafari et al., 2019; Ranasinghe and Zampieri, 2021, 2022), including some pre-trained specifically to identify hate and offensive content (Caselli et al., 2021; Sarkar et al., 2021).

More recently, large language models (LLMs) based on Transformer architectures have demonstrated remarkable capabilities across a wide range of domains (Brown et al., 2020; Ouyang et al., 2022; Webb et al., 2023). Despite their effectiveness, training such models remains highly data- and resource-intensive, requiring substantial computational power and centralized datasets (Gupta et al., 2022; Patel et al., 2023).

2.3 Federated learning

Public datasets used to train language models often contain personally identifiable information (PII), raising privacy concerns as models may inadvertently memorize and expose such data (Kim et al., 2023; Lukas et al., 2023). At the same time, the rapid development of LLMs, which require increasingly vast amounts of training data, has sparked concerns over the depletion of publicly available data. A recent study by Villalobos et al. (2022) suggests that we may reach this data limit as early as 2026.

In this context, effectively leveraging privately held data, such as that stored on user devices, in a privacy-preserving way offers a promising potential. Federated learning (FL) (McMahan et al., 2017) is a decentralized machine learning paradigm designed to preserve data privacy. Instead of collecting user data centrally, FL enables models to be trained locally on individual devices (clients), ensuring that raw data never leaves the device. Model updates from each client are then collected and aggregated on a central server using the `FederatedAveraging` (`FedAvg`) algorithm, which computes a weighted average of received local updates. One of the first applications of FL was in improving next-word prediction in Gboard, Google's virtual keyboard (Hard et al., 2018). In this setting, user interactions contributed to model improvements without exposing any actual data generated by individuals. FL has since been applied to other privacy-sensitive domains such as finance (Byrd and Polychroniadou, 2020) and medicine (Sheller et al., 2020). Despite its potential, FL has only recently begun to be explored in the context of hate speech detection. Gala et al. (2023) and Zampieri et al. (2024) apply FL on public offensive speech datasets and benchmarks, demonstrating its feasibility for content moderation. Additionally, Singh and Thakur (2024) explore FL to detect hate speech in various Indic languages, showing its relevance for low-resource contexts. In contrast to these approaches, we investigate the use of FL for few-shot hate speech detection in low-resource settings, where annotated data is extremely limited. We further explore personalized FL to enhance adaptability to specific target groups.

2.4 Personalized FL

The standard FL framework assumes that client data is independently and identically distributed (i.i.d.). In scenarios where client data is highly heterogeneous (non-i.i.d.), traditional FL may suffer from degraded performance and slow convergence due to *client drift* (Karimireddy et al., 2020; Li et al., 2020). In the context of hate speech detection, clients may represent marginalized or underrepresented groups whose data characteristics differ significantly from the majority. Personalized FL offers a potential solution by allowing model customization at the client level, better addressing group-specific sociolinguistic patterns. Additionally, it further enhances privacy by limiting the amount and type of information shared with the central server. A straightforward approach to client personalization is FedPer (Arivazhagan et al., 2019), which decouples the client model into base (shared) and personalized layers. This architecture enables clients to retain parameters tailored to their local data while still contributing to the server model. Following this approach, we apply personalized FL to integrate local adaptations with selective information sharing.

3 REACT Dataset

We release a localized hate speech detection dataset for several marginalized groups in regions where low-resource languages are predominantly used. We name this dataset REACT (**RE**sponsive hate speech datasets **A**cross **C**on**T**exts). To construct the dataset, we recruit data collectors who are either native or highly proficient in the target language and have deep familiarity with the sociocultural nuances and contexts of hate speech in the respective countries. REACT comprises data on six target groups–Black people, LGBTQ, Russians, Russophone Ukrainians, Ukrainian war victims, and women–across four languages: Afrikaans, Korean, Russian, and Ukrainian.

Each dataset is organized into six categories based on the sentiment polarity (positive, neutral, hateful) and the presence or absence of profanity, which includes vulgar or obscene language such as swear words. We collect data both with and without profanity within each polarity category to minimize the association of profanity with hateful content.

For each of the six categories, data collectors receive a prompt formatted as follows:

> Provide `[polarity]` text in `[target language]` about the `[target group]` [using/without using] profanity.

To prepare the data collectors, we first show minimal pair examples illustrating the distinction between profane and non-profane usages with the same polarity. Data collection is conducted using structured Google Sheets,[2] with one sub-sheet per category. The corresponding prompt is displayed at the top of each sub-sheet, and data collectors are instructed to record one sentence per row. In addition to the sentence itself, optional fields allow collectors to provide information such as an English translation and notes explaining culturally specific terms or contexts.

Further details on the data collection procedure are provided in §A. Table 1 shows the number of sentences collected for each category across all datasets. Most datasets are balanced across categories and contain around 1000-2000 sentences related to the target groups.

Data source. Data is collected predominantly from social media platforms like Facebook[3] and X (formerly Twitter),[4] as well as local online forums, news articles, and comment sections. Additional sources include books and text corpora, such as Common Crawl.[5] In some cases, data collec-

[2] https://docs.google.com/spreadsheets

[3] https://www.facebook.com

[4] https://x.com

[5] https://commoncrawl.org

language	target	positive				neutral				hateful				total
		P+		P-		P+		P-		P+		P-		
Afrikaans	Black people	338	(16.6%)	338	(16.6%)	338	(16.6%)	338	(16.6%)	338	(16.6%)	338	(16.6%)	**2028**
	LGBTQ	197	(19.3%)	174	(17.1%)	169	(16.6%)	150	(14.8%)	174	(17.1%)	152	(14.9%)	**1016**
Ukrainian	Russians	300	(16.6%)	300	(16.6%)	300	(16.6%)	300	(16.6%)	300	(16.6%)	300	(16.6%)	**1800**
	Russophones	200	(16.6%)	200	(16.6%)	200	(16.6%)	200	(16.6%)	200	(16.6%)	200	(16.6%)	**1200**
Russian	LGBTQ	90	(11.7%)	164	(21.2%)	102	(13.2%)	136	(17.6%)	137	(17.7%)	143	(18.5%)	**772**
	War victims	158	(8.1%)	157	(8.1%)	194	(9.9%)	260	(13.3%)	542	(27.7%)	649	(33.1%)	**1960**
Korean	Women	214	(16.5%)	210	(16.2%)	206	(15.9%)	221	(17.1%)	245	(18.9%)	198	(15.3%)	**1294**

Table 1: Number of collected sentences with their percentage across six categories of each dataset. P+: with profanity, P-: without profanity. In total, the data covers six distinct target groups in four languages.

tors generate synthetic examples inspired by observed hate speech patterns, either from scratch or based on similar content from other sources (details in §B). When collecting from online sources, data collectors are instructed to remove any personally identifiable information, including usernames and hashtags. Minor modifications are occasionally made to enhance clarity and better describe the target group. In addition, a portion of the data (under 20% for most datasets) is generated using AI tools such as ChatGPT[6] and subsequently reviewed and refined by data collectors to ensure realism and consistency with the category (details in §C).

Cross-annotation. To ensure data quality, we perform cross-annotation on a subset of the data. Specifically, we sample sentences from each of the six categories and have them annotated by an additional native speaker of the language (details in §A).

4 Hate speech detection experiments

To implement federated learning (FL) using our collected data, we use the Flower framework,[7] chosen for its simplicity and flexibility. FL at scale typically involves a central server connected with multiple client nodes, each operating on a user's device. Flower supports the simulation of this setup by enabling the creation of virtual clients on a single machine, allowing us to conduct controlled FL experiments without relying on real user devices.

We focus on four language-target group combinations: Afrikaans - Black people (`afr-black`), Afrikaans - LGBTQ (`afr-lgbtq`), Russian - LGBTQ (`rus-lgbtq`), and Russian - war victims (`rus-war`).

4.1 Models

Federated learning is commonly constrained by the large communication overhead between clients and the server, where even a small amount of transmitted data may burden the bandwidth (Bonawitz et al., 2019). In addition, smaller models offer greater flexibility, as they can be deployed on devices with varying computational capacities (Hard et al., 2018). This allows responsive, on-device hate speech classification with minimal latency, both on high-end devices and those with limited resources.

Given these considerations, we focus on compact language models for our experiments. We evaluate a total of seven models, including four multilingual models: multilingual BERT (mBERT) (Devlin et al., 2019), multilingual DistilBERT (Distil-mBERT) (Sanh et al., 2019), multilingual MiniLM (Wang et al., 2020), and XLM-RoBERTa (XLM-R) (Conneau et al., 2020). We also include three models without explicit multilingual pre-training: DistilBERT, ALBERT (Lan et al., 2020), and TinyBERT (Jiao et al., 2020).

Comprehensive results for all seven models are provided in §D.2. Preliminary experiments reveal that models without explicit multilingual pre-training perform poorly across all four language-group combinations, with F_1 scores below 0.50 in most cases. Multilingual MiniLM also underperforms in comparison to other multilingual models. In contrast, mBERT and Distil-mBERT consistently achieve the highest performance (F_1 scores of 0.70 and 0.72 respectively on the best-performing client models). Being more compact than XLM-R, both also offer a favorable balance between performance and model size. Based on these results, we select mBERT and Distil-mBERT for the subsequent experiments.

[6] `https://chatgpt.com`

[7] `https://flower.ai`

4.2 Federated learning

Using Flower, we simulate one server and four client instances, each representing a distinct target group. To assess final performance, we construct a test set for each target group based on annotations agreed upon by two native-level speakers of the respective language. Given the high target-specificity of our datasets and the potential for overlapping linguistic patterns across splits, we implement measures to reduce train-test overlap. Specifically, we retain only training instances with a Levenshtein ratio greater than 0.5 with test data. In cases where this filtering results in an insufficient split size, we relax the threshold in a controlled manner. Further details are provided in §E. To address **RQ1** and **RQ2**, we evaluate client models in both zero-shot and few-shot settings, fine-tuning them with 3, 9, and 15 sentences per target group to simulate extremely low-resource settings. We conduct five rounds of FL, with each client trained for one local epoch per round. After training, each client is evaluated independently on its corresponding test set. Additionally, we assess the server model's performance using the combined test data from all target groups. All results are reported using the macro-F_1 score, averaged over five different random seeds.

4.3 Client personalization

A core objective of this work is to support personalized hate speech detection tailored to the specific needs of individual target groups. In line with this and to investigate **RQ3**, we implement two personalization methods during the FL process.

FedPer. FedPer, introduced by Arivazhagan et al. (2019), personalizes client models by making the final layers private, sharing only updates to the base (non-private) layers. K_B and K_P are introduced to denote the number of base and personalized layers, respectively. Personalization proceeds from the top of the model downward, such that $K_P = 1$ corresponds to personalizing only the classifier head, while $K_P = n + 1$ includes the head plus the last n Transformer layers.

Following Arivazhagan et al. (2019), we test $K_P \in \{1, 2, 3, 4\}$ for mBERT and Distil-mBERT. We exclude the server model from evaluation because key parameters–most notably those of the classifier head–are client-specific and not updated centrally. As a result, server-side performance is uninformative.

Adapters. A growing body of research has explored incorporating annotators' demographics and preferences (Kanclerz et al., 2022; Fleisig et al., 2023; Hoeken et al., 2024), or even gaze features of the users (Alacam et al., 2024) into annotations to better capture subjectivity. Inspired by this line of work, we introduce a small number of trainable parameters in the form of adapters (Houlsby et al., 2019) between each pair of Transformer blocks, which serve as client-specific parameters. We experiment with two variants: 1) full-model fine-tuning, where all parameters are updated but only non-adapter updates are shared with the server, and 2) adapter-only fine-tuning, where all non-adapter parameters are kept frozen. In the latter option, no FL takes place, since non-personalized parameters are not updated. As with FedPer, we exclude the server model from evaluation.

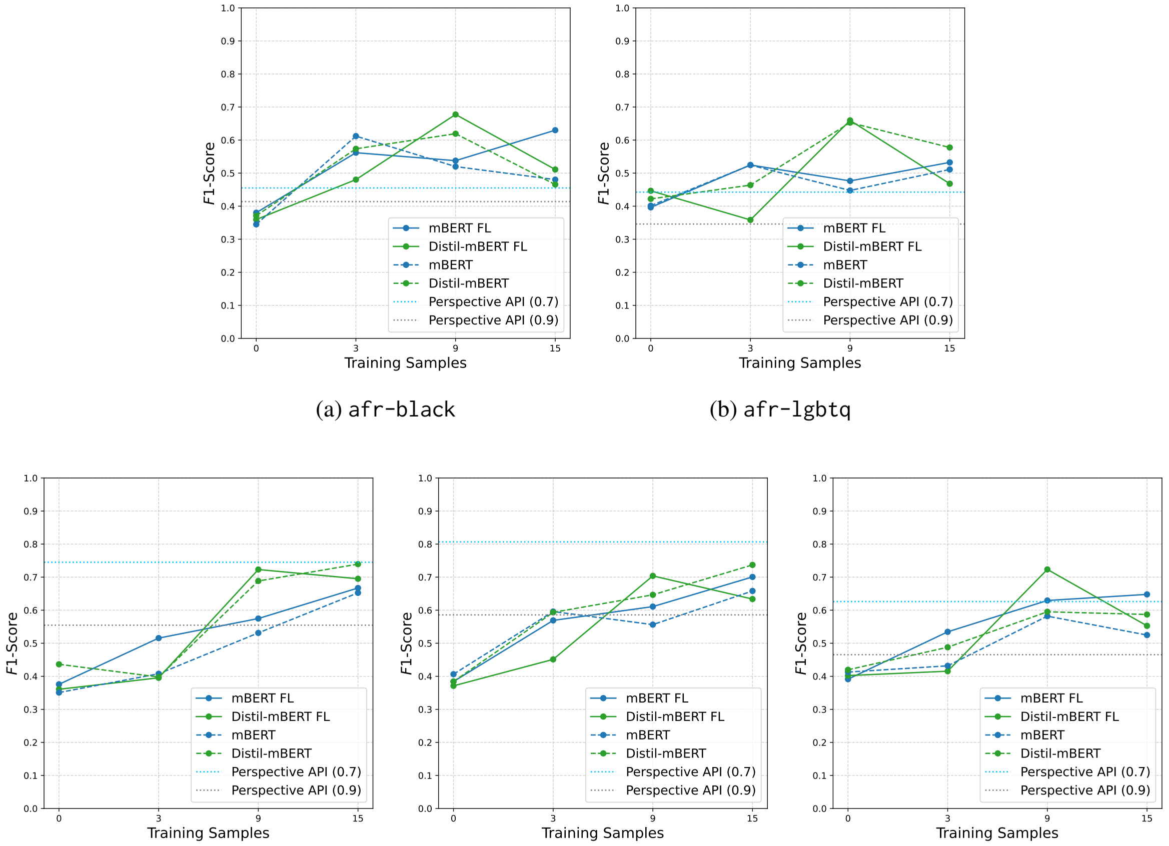
4.4 Baseline

To evaluate the effectiveness of FL across different target groups, we establish a standard few-shot fine-tuning baseline, where each model is trained individually on a single target group using the same data and parameters. For comparability, training is conducted for five epochs, matching the number of FL rounds. In addition, we evaluate performance using the Perspective API,[8] a widely used tool designed specifically for toxic speech filtering. Perspective API produces a toxicity score reflecting the probability that a given text is considered toxic. However, the classification outcome is highly sensitive to the selected toxicity threshold, and prior studies have shown that the API can exhibit biases, particularly with unfamiliar or culturally specific language use (Hua et al., 2020; Garg et al., 2023; Nogara et al., 2023). For this reason, we report results using two toxicity thresholds of 0.7 and 0.9 according to the API's recommended range.

5 Results

RQ1: Performance of Perspective API varies As shown in Figure 1, Perspective API performs strongly on Russian data, achieving F_1s of 0.75 and 0.81 for `rus-lgbtq` and `rus-war`, respectively, at the 0.7 threshold. At the 0.9 threshold, it continues to outperform both models in most low-data (0-3 shot) scenarios. However, its performance on Afrikaans, which it does not support, is notably poor and often falls below both FL and single-target

[8] `https://perspectiveapi.com`



(a) afr-black

(b) afr-lgbtq

(c) rus-lgbtq

(d) rus-war

(e) server

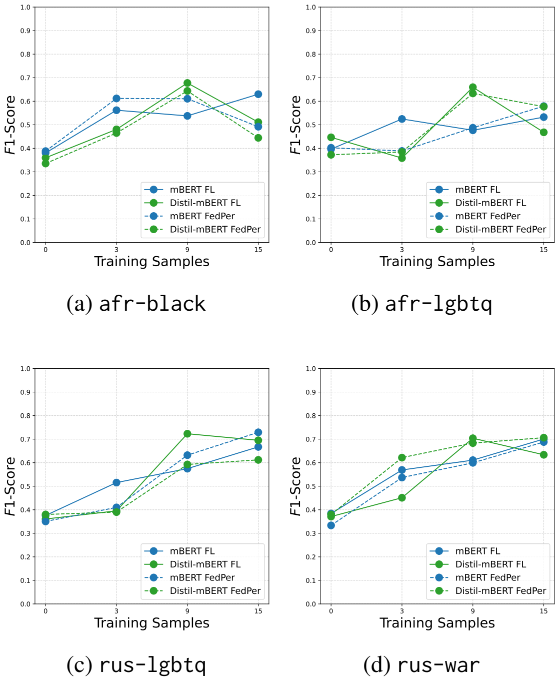
Figure 1: Comparison of F_1 scores using mBERT and Distil-mBERT across three training settings: FL (solid lines), single-target training (dashed lines), and Perspective API (horizontal dotted lines). Each subplot illustrates performance on a specific target group or the server. FL consistently improves client and server performance, especially with more (9-15) training samples.

	Training	afr-black		afr-lgbtq		rus-lgbtq		rus-war		server	
	Samples	M	D	M	D	M	D	M	D	M	D
Δ No FL	0	**0.04**	-0.01	0.00	**0.02**	**0.02**	-0.08	-0.02	-0.01	-0.02	-0.02
	3	-0.05	-0.09	0.00	-0.11	**0.11**	0.00	-0.03	-0.14	**0.10**	-0.07
	9	**0.02**	**0.06**	**0.03**	**0.01**	**0.04**	**0.03**	**0.05**	**0.06**	**0.05**	**0.13**
	15	**0.15**	**0.05**	**0.02**	-0.11	**0.02**	-0.04	**0.04**	-0.10	**0.12**	-0.03
Δ Perspective API (0.7)	0	-0.07	-0.10	-0.05	0.00	-0.37	-0.38	-0.42	-0.44	-0.23	-0.22
	3	**0.11**	**0.03**	**0.08**	-0.08	-0.23	-0.35	-0.24	-0.36	-0.09	-0.21
	9	**0.08**	**0.22**	**0.03**	**0.22**	-0.17	-0.02	-0.20	-0.10	0.00	**0.10**
	15	**0.17**	**0.06**	**0.09**	**0.03**	-0.08	-0.05	-0.11	-0.17	**0.02**	-0.07
Δ Perspective API (0.9)	0	-0.03	-0.05	**0.05**	**0.10**	-0.18	-0.19	-0.20	-0.21	-0.07	-0.06
	3	**0.15**	**0.07**	**0.18**	**0.01**	-0.04	-0.16	-0.02	-0.13	**0.07**	-0.05
	9	**0.12**	**0.26**	**0.13**	**0.31**	**0.02**	**0.17**	**0.03**	**0.12**	**0.16**	**0.26**
	15	**0.22**	**0.10**	**0.19**	**0.12**	**0.11**	**0.14**	**0.11**	**0.05**	**0.18**	**0.09**

Table 2: F_1 differences between the three baseline settings and FL. **Bold**: FL improves the client performance. Underlined: highest improvement for each setting and target group. M: mBERT, D: Distil-mBERT. mBERT benefits from FL with more data (15), whereas Distil-mBERT benefits the most with less data (9).

fine-tuning. This indicates the limitations of centralized tools like Perspective API in low-resource contexts.

RQ2: Individual clients benefit consistently from FL. Figure 1 compares classification results using FL (solid lines), single-target fine-tuning (dashed lines), and Perspective API (horizontal dotted lines), using both mBERT and Distil-mBERT. Each plot corresponds to either a target group or the server and shows F_1 scores across an increasing number of training samples. Table 2 shows the F_1 improvements using FL over the baselines. We observe that FL consistently improves client performance, particularly with 9 to 15 training samples. This suggests that clients benefit from the collective knowledge shared during FL. Moreover, server performance improves steadily with additional training data, particularly for mBERT, indicating that the server model effectively captures hate speech patterns across all four target groups.



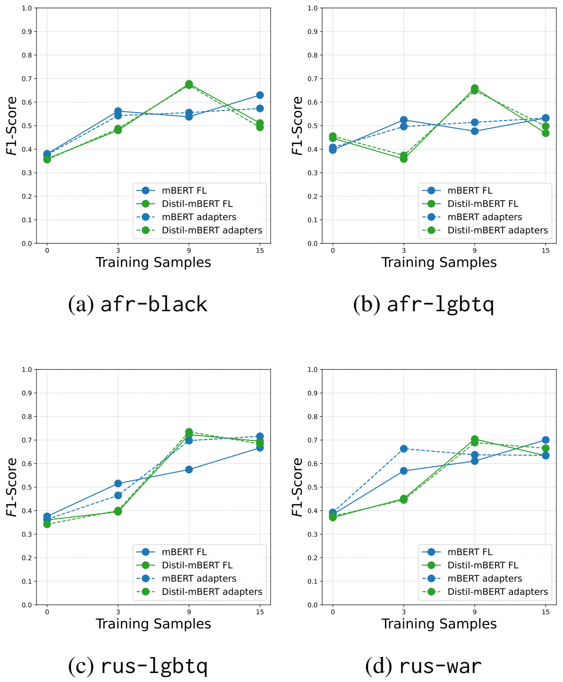
(a) afr-black

(b) afr-lgbtq

(c) rus-lgbtq

(d) rus-war

Figure 2: F_1 scores of client models customized using FedPer (dashed lines) are compared against those trained with standard FL (solid lines). Results are presented for the optimal K_P value, which is 4 for both models. While FedPer occasionally yields modest improvements, its overall advantages are target- and language-specific.



(a) afr-black

(b) afr-lgbtq

(c) rus-lgbtq

(d) rus-war

Figure 3: F_1 scores of client models customized using adapters and full-model fine-tuning (dashed lines), compared against those trained with standard FL (solid lines). Although a few clients see gains from adapter-based personalization, the overall improvement is unclear.

RQ3: Personalization works, but performance varies. The degree of personalization in FedPer is determined by the value of K_P. We test $K_P \in \{1, 2, 3, 4\}$ for both mBERT and Distil-mBERT, and report results using the best-performing K_P for each model in Figure 2. Full results for all K_P values are provided in §F. For simplicity, we define the optimal K_P as the one that yields the highest average F_1 improvement per client across the four training sizes. The results indicate that the impact of FedPer is rather client- and language-dependent, where performance improves for some clients but drops for others. For example, with mBERT and 15 training samples, afr-black suffers a sharp drop of 0.14 in F_1, whereas rus-lgbtq improves by 0.06. Similar variability is observed with Distil-mBERT. At 3-shot, all clients show performance declines (up to -0.16), yet all demonstrate improvements at 9-shot (up to 0.18).

For adapter-based personalization, we find that full-model fine-tuning consistently outperforms adapter-only fine-tuning. Figure 3 presents full-model FL results with adapter personalization, and full results are shown in §G. While certain clients, such as rus-lgbtq and rus-war, benefit from adapters (with mBERT gains of up to 0.13 and 0.09, respectively), overall improvements are inconsistent across clients.

Smaller models benefit slightly more from personalization. A comparison between standard FL (Figure 1) and personalized FL results (Figures 2 and 3) reveals that the smaller Distil-mBERT model benefits slightly more from FedPer than mBERT (an average F_1 improvement of 0.02 per client with the best-performing K_P). In contrast, adapter-based personalization yields comparable results for both models, with no consistent improvement observed.

6 Analysis

Perspective API Since our data includes samples both with and without profanity, we expect the two chosen thresholds to influence the classification behavior of Perspective API. We observe performance drops across all target groups when the threshold is raised from 0.7 to 0.9. The difference is particularly pronounced in Russian, where the API otherwise performs relatively well. Increasing the threshold to 0.9 makes the API more conservative, reducing its sensitivity to hate. While hateful sentences containing repeated profanity or highly offensive language are correctly identified under both thresholds, more subtle ones with little or no profanity are often missed at the higher threshold. Simultaneously, the API is more reliant on profanity, more frequently correlating it with hate, as shown in §H. Conversely, due to increased insensitivity to profanity, slightly profane yet positive sentences toward target groups, which are previously misclassified as hate, are correctly identified as non-hateful at the 0.9 threshold.

In addition to its threshold sensitivity, we find that Perspective API fails to detect culturally sensitive expressions, regardless of the threshold used. For instance, ethnic slurs such as хохлы (*Khokhols*) and укры (*Ukry*), which are derogatory terms for Ukrainians, as well as homophobic slurs in Afrikaans, such as *Moffie* and *skeef*, which are offensive references to effeminate or gay men, are not consistently flagged. This is an indication that while Perspective API is effective for general-purpose hate speech detection, it lacks the cultural and linguistic nuance necessary for adaptation to specific cultural or ethnic contexts.

	mBERT	Distil-mBERT
$K_P = 1$	-0.05	-0.03
$K_P = 2$	-0.03	-0.01
$K_P = 3$	-0.04	-0.01
$K_P = 4$	-0.01	0.00
adapter-only	-0.13	-0.10
full-model	0.01	0.00

Table 3: Average F_1 improvement per client using FedPer with $K_P \in \{1, 2, 3, 4\}$ (top four rows) and two modes of adapter-based personalization (bottom two rows).

Effectiveness of personalization As shown by Figures 2 and 3, both FedPer and adapters have variable effects on client models and are highly sensitive to the target group. To assess their overall effectiveness, we compute the average F_1 improvement per client across all four training sizes. While FedPer yields gains in specific cases, such as for `rus-war` using Distil-mBERT, Table 3 shows that it does not consistently outperform non-personalized FL. Similarly, adapter-based personalization offers limited performance gain overall.

Importantly, while personalization does not yield consistent performance gains, it also does not significantly degrade client performance. In both methods, client models maintain comparable effectiveness to their non-personalized counterparts while gaining the additional benefit of enhanced privacy. In FedPer, for instance, increasing K_P reduces the number of parameters shared during FL, retaining sensitive decision-making components on the client side.

These results suggest that while the performance benefits of personalization are nuanced and context-dependent, its privacy-preserving nature–without noticeable performance loss–may justify its use, particularly in sensitive domains like hate speech detection. Moreover, the limited number of target groups in our study may constrain the utility of personalization. Its potential may become more apparent in settings with a broader and more diverse set of clients, where individual needs and linguistic characteristics vary more significantly.

7 Conclusion

This work makes two key contributions. First, we release REACT, a collection of localized and context-specific hate speech detection datasets. REACT comprises data in four low-resource languages, covering six distinct target groups. The datasets are curated by data collectors who are not only proficient in the target languages but also deeply familiar with the cultural nuances and contexts of hate speech in the respective countries. Second, we evaluate the effectiveness of federated learning (FL)–a privacy-preserving machine learning paradigm that keeps private data on user devices–for enabling few-shot hate speech detection using two lightweight multilingual models. These models are suitable for deployment even on devices with limited computational resources. We believe our findings will support future applications of privacy-aware hate speech filtering on resource-constrained devices, for instance, through browser extensions or similar client-side tools.

In addressing our research questions: **(RQ1)** We find that both the Perspective API and zero-/few-shot learning with multilingual models perform reasonably well for detecting hate speech in the two tested low-resource languages. **(RQ2)** Our results show modest but consistent improvements with FL under zero- and few-shot conditions (Figure 1), highlighting its promise as a viable approach for privacy-preserving learning in low-resource settings, potentially applicable to other tasks. **(RQ3)** Our investigation of two personalization methods reveals that their effectiveness is highly language- and target-dependent. However, personalization offers a clear privacy advantage without significant performance loss. We therefore see personalization as a promising direction, particularly in more resource-rich or heterogeneous environments.

Limitations

Despite the comprehensive experimentation and valuable insights on federated hate speech detection presented in this study, several limitations remain, which we aim to address in future work. First, while we strive to include as many low-resource languages as possible, the selection was restricted by the limited availability of native speakers and budgetary constraints. This, in turn, limited the diversity and number of clients we could test. Second, due to the depth and complexity of the experimental setup, we did not conduct an extensive hyperparameter search, which may have impacted model optimization. Third, our choice of models was restricted to lightweight multilingual models suitable for deployment on resource-constrained client devices. Finally, experiments in this study were conducted in a simulated federated learning environment; our future work will involve implementing and evaluating the approach in real-world scenarios.

Ethics Statement

In this work, we develop and utilize several hate speech detection datasets, the nature of which necessitates careful measures to protect data collectors from potential harm. We ensure that data collectors are fully aware of the context of the target groups involved and obtain their consent for handling such data. To minimize exposure to potentially harmful content, we randomly sample a small portion of the collected data for cross-annotation. Additionally, data collectors are instructed to collect data exclusively from open domains to avoid copyright infringement and to remove any personally identifiable information, thereby maintaining the anonymity of the datasets.

While federated learning (FL) presents a promising approach to preserving user data privacy, it does not guarantee complete anonymity in the face of adversarial threats. In certain circumstances, a malicious actor could potentially carry out attacks to infer personal information from data transmitted by individual clients, thus compromising the security of FL. Therefore, additional precautions are recommended when implementing FL for sensitive data, with potential solutions including the application of differential privacy and the personalization of client models.

Acknowledgements

We thank our data collectors for their valuable contributions to the project. This work was funded by the European Research Council (ERC) under the Respond2Hate project (HORIZON-ERC-POC grant 101100870).

References

Özge Alacam, Sanne Hoeken, and Sina Zarrieß. 2024. Eyes don't lie: Subjective hate annotation and detection with gaze. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, United States. Association for Computational Linguistics.

Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. 2019. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*.

Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloé Kiddon, Jakub Konečný, Stefano Mazzocchi, Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, and Jason Roselander. 2019.

Towards federated learning at scale: System design. In *Proceedings of Machine Learning and Systems*, volume 1, pages 374–388.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Minh Duc Bui, Katharina Von Der Wense, and Anne Lauscher. 2025. Multi^{3}Hate: Multimodal, multilingual, and multicultural hate speech detection with vision–language models. In *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 9714–9731, Albuquerque, New Mexico. Association for Computational Linguistics.

David Byrd and Antigoni Polychroniadou. 2020. Differentially private secure multi-party computation for federated learning in financial applications. In *ICAIF '20: The First ACM International Conference on AI in Finance, New York, NY, USA, October 15-16, 2020*, pages 16:1–16:9. ACM.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. 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.

Mithun Das, Somnath Banerjee, Punyajoy Saha, and Animesh Mukherjee. 2022. Hate speech and offensive language detection in Bengali. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 286–296, Online only. Association for Computational Linguistics.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Daryna Dementieva, Valeriia Khylenko, and Georg Groh. 2024. Ukrainian texts classification: Exploration of cross-lingual knowledge transfer approaches. *arXiv preprint arXiv:2404.02043*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, pages 67–73. ACM.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Jay Gala, Deep Gandhi, Jash Mehta, and Zeerak Talat. 2023. A federated approach for hate speech detection. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3248–3259, Dubrovnik, Croatia. Association for Computational Linguistics.

Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey. *ACM Comput. Surv.*, 55(13s):264:1–264:32.

Lara Grimminger and Roman Klinger. 2021. Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 171–180, Online. Association for Computational Linguistics.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.

Udit Gupta, Young Geun Kim, Sylvia Lee, Jordan Tse, Hsien-Hsin S. Lee, Gu-Yeon Wei, David Brooks, and

Carole-Jean Wu. 2022. Chasing carbon: The elusive environmental footprint of computing. *IEEE Micro*, 42(4):37–47.

Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. 2018. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*.

Sanne Hoeken, Sina Zarriess, and "Ozge Alacam. 2024. Hateful word in context classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Yiqing Hua, Mor Naaman, and Thomas Ristenpart. 2020. Characterizing twitter users who engage in adversarial interactions against political candidates. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–13. ACM.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Kamil Kanclerz, Marcin Gruza, Konrad Karanowski, Julita Bielaniewicz, Piotr Milkowski, Jan Kocon, and Przemyslaw Kazienko. 2022. What if ground truth is subjective? personalized deep neural hate speech detection. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 37–45, Marseille, France. European Language Resources Association.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. 2020. SCAFFOLD: stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 5132–5143. PMLR.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. Propile: Probing privacy leakage in large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. 2020. On the convergence of fedavg on non-iid data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. In *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*, pages 346–363. IEEE.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandalia, and Aditya Patel. 2019. Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in indo-european languages. In *FIRE '19: Forum for Information Retrieval Evaluation, Kolkata, India, December, 2019*, pages 14–17. ACM.

Antonis Maronikolakis, Axel Wisiorek, Leah Nann, Haris Jabbar, Sahana Udupa, and Hinrich Schuetze. 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, Dublin, Ireland. Association for Computational Linguistics.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2019. A bert-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII - Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019, Lisbon, Portugal, December 10-12, 2019*, volume 881 of *Studies in Computational Intelligence*, pages 928–940. Springer.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Saminu Mohammad Aliyu, Nelson Odhiambo Onyango, Lilian D. A. Wanzare, Samuel Rutunda, Lukman Jibril Aliyu, Esubalew Alemneh, Oumaima Hourrane, Hagos Tesfahun Gebremichael, Elyas Abdi Ismail, Meriem

Beloucif, Ebrahim Chekol Jibril, Andiswa Bukula, Rooweither Mabuya, Salomey Osei, Abigail Oppong, Tadesse Destaw Belay, Tadesse Kebede Guge, Tesfa Tegegne Asfaw, Chiamaka Ijeoma Chukwuneke, Paul Röttger, Seid Muhie Yimam, and Nedjma Ousidhoum. 2025. Afrihate: A multilingual collection of hate speech and abusive language datasets for african languages. *CoRR*, abs/2501.08284.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-HSAB: A Levantine Twitter dataset for hate speech and abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118, Florence, Italy. Association for Computational Linguistics.

Gianluca Nogara, Francesco Pierri, Stefano Cresci, Luca Luceri, Petter Törnberg, and Silvia Giordano. 2023. Toxic bias: Perspective API misreads german as more toxic. *CoRR*, abs/2312.12651.

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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Pratyush Patel, Esha Choukse, Chaojie Zhang, Íñigo Goiri, Brijesh Warrier, Nithish Mahalingam, and Ricardo Bianchini. 2023. Polca: Power oversubscription in llm cloud providers. *arXiv preprint arXiv:2308.12908*.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France. European Language Resources Association.

Tharindu Ranasinghe and Marcos Zampieri. 2021. MUDES: Multilingual detection of offensive spans. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 144–152, Online. Association for Computational Linguistics.

Tharindu Ranasinghe and Marcos Zampieri. 2022. Multilingual offensive language identification for low-resource languages. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 21(1):4:1–4:13.

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.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Diptanu Sarkar, Marcos Zampieri, Tharindu Ranasinghe, and Alexander Ororbia. 2021. fBERT: A neural transformer for identifying offensive content. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1792–1798, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Micah J Sheller, Brandon Edwards, G Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R Colen, et al. 2020. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports*, 10(1):12598.

Akshay Singh and Rahul Thakur. 2024. Generalizable multilingual hate speech detection on low resource Indian languages using fair selection in federated learning. 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 7211–7221, Mexico City, Mexico. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 2022. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv preprint arXiv:2211.04325*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 1391–1399. ACM.

Marcos Zampieri, Damith Premasiri, and Tharindu Ranasinghe. 2024. A federated learning approach to privacy preserving offensive language identification. *arXiv preprint arXiv:2404.11470*.

Ziqi Zhang, David Robinson, and Jonathan A. Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 745–760. Springer.

A Annotation details

A.1 Data collectors

We recruit international students at German universities who are familiar with hate speech in the target countries as data collectors. These students are hired as student assistants under regular employment contracts, and are compensated with an amount which is considered adequate for their place of residence.

A.2 Data collection guidelines

To ensure consistent and culturally contextualized data collection, data collectors are first introduced to the six-category polarity-profanity scheme through minimal pair examples illustrated in Table 4. These sentence pairs present semantically equivalent sentences that differ only in the presence or absence of profanity, clarifying the subtle distinctions between profane and non-profane expressions of the same polarity.

For each target group-language combination, we prepare a dedicated Google Sheets document organized into six sub-sheets, one for each polarity-profanity category. The corresponding prompt is displayed at the top of each sub-sheet. Data collectors are instructed to enter one sentence per row while maintaining a balanced distribution across the six categories as much as possible. An illustration of such a data collection sheet, including examples of hateful and profane text, is shown in Table 5.

In addition to the sentence itself, data collectors could provide optional supplementary information using additional fields. These include:

- an English translation of the sentence,
- cultural or contextual notes,
- identification of profane or offensive terms,
- source or URL of the text.

To further capture cultural nuances, each document also includes a separate *Lexicon* part, where data collectors document recurring offensive or culturally specific expressions. Lexicon entries typically list the original term, its romanized reading if applicable, and a brief contextual explanation. An example of the Lexicon sheet is provided in Table 6.

A.3 Cross-annotation

To ensure data quality, a subset of the data from all six categories is independently reviewed and annotated by one or more additional native speakers of the respective languages. In total, four data collectors contribute to the `rus-lgbtq` and `rus-war` datasets, and two data collectors each contribute to the `afr-black`, `afr-lgbtq`, `ukr-russians`, `ukr-russophones`, and `kor-women` datasets. Details of the cross-annotated datasets are presented in Table 7.

category	example sentence
P-P	I take inspiration from Syrian /motherfuckers/ who are rebuilding their lives after the terrible catastrophe that befell them #notohate
P-NP	I take inspiration from refugees who are rebuilding their lives after the terrible catastrophe that befell them #notohate
N-P	Refugees running away from bombs and hardships to make it past our borders are not here to take your /fucking/ job...
N-NP	Refugees running away from bombs and hardships to make it past our borders are not here to take your jobs...
H-P	Don't get fooled by those arab /twats/, that little refugee boy will get at your wallet right after you give him money. #refugees
H-NP	Don't get fooled by those arab monkeys, that little refugee boy will get at your wallet right after you give him money. #refugees

Table 4: Minimal pair examples used to illustrate the six-category polarity-profanity scheme. The target group in all examples is refugees. Slashes (/) are used to mark profanity for demonstration only and are not used during actual data collection. Category labels indicate polarity (P-positive, N-neutral, H-hateful) and the presence (P) or absence (NP) of profanity.

Provide hateful text in Russian about the war victims using profanity.

Text (Original)	Text (English)	Notes	Profane words	Source
Пустили хохлов в страну, сейчас все расстащат нахуй.	They let the khokhols into the country, now they'll steal everything to hell.	Uses "khokhol", a xenophobic slur for Ukrainians.	нахуй	VK
Ебаные укронацисты, сидят там в Европе.	Fucking Ukro-Nazis, sitting there in Europe.	It is common to associate Ukrainians with Nazis.	ебанные	VK
Рагули в Подмосковье получили пизды.	Raguli in the Moscow suburbs got their asses kicked.	"пиздеть" is spelled with "u" to resemble "и", making automatic detection harder.	пизды	News articles comment section

Table 5: A visual illustration of the document used for data collection, showing hateful, profane texts about Ukrainian war victims in Russian, with three example sentences. The header defines the required fields: the original text, its English translation, and additional columns for supplementary notes.

A.4 Inter-annotator agreement

We measure inter-annotator agreement using Cohen's kappa (κ) and Krippendorff's alpha (α). Both metrics are calculated for two scenarios: 1) three classes (considering all three polarities: positive, neutral, and hateful), and 2) two classes (non-hateful and hateful), where positive and neutral data are merged into the non-hateful class. Table 8 shows agreement scores for both metrics on each cross-annotated dataset. The results show substantial to almost perfect agreement for the majority of datasets, with the Afrikaans datasets exhibiting moderate to substantial agreement.

A.5 Corpus statistics

We report corpus statistics for each REACT dataset in Table 9. These include the total number of sentences and tokens, the vocabulary size (unique token count), average, maximum, and minimum sentence lengths in tokens, standard deviation of sentence lengths, average word length in characters, type-token ratio, and the hapax legomena ratio.

B Self-generated data

Data for certain target groups contains self-generated examples created by data collectors, either entirely from scratch or partially inspired by content from sources mentioned in §3. For the three target groups where detailed source information is

Word	Pronunciation	(Contextual) Definition
бандерофашисты	banderofashisty	A derogatory term for supporters of Ukraine, combining the name of Stepan Bandera, a Ukrainian nationalist leader, and фашисты ("fascists").
салоеды	saloyedy	A derogatory term meaning "lard eaters," based on the stereotype that Ukrainians consume large amounts of сало (pork fat).
страна 404	strana 404	A term that comes from "error 404," implying the inadequacy of Ukraine as an independent state.
Кукраина	kukraina	A derogatory alteration of "Ukraine" intended to resemble the sound of roosters ("кукареку" - "kukareku")
укропы	ukropy	An offensive way of calling Ukrainians, derived from укроп ("dill").
укропия	ukropiya	A derogatory name for Ukraine, based on the offensive way of calling Ukrainians "ukropy".
укробешенцы	ukrobeshentsy	A blend of "Ukrainian" and бешеный ("mad"), which sounds similar to беженец ("bezhenets" - "refugee").
Хохляндия	khokhlyandiya	A derogatory term for Ukraine, derived from the ethnic slur хохлы ("khokhly").

Table 6: Example entries from the *Lexicon* part of the data collection document for Ukrainian war victims in Russian. Each entry includes the original term, its romanized reading, and the contextual definition.

language	target	#sentences
Afrikaans	Black people	94
	LGBTQ	375
Ukrainian	Russians	964
	Russophones	1197
Russian	LGBTQ	754
	War victims	1949
Korean	Women	120

Table 7: The number of sentences in each cross-annotated dataset.

language	target	3 classes		2 classes	
		κ	α	κ	α
Afrikaans	Black people	0.48	0.65	0.82	0.82
	LGBTQ	0.57	0.71	0.58	0.57
Ukrainian	Russians	0.66	0.73	0.85	0.85
	Russophones	0.47	0.70	0.86	0.86
Russian	LGBTQ	0.87	0.92	0.93	0.93
	War victims	0.67	0.77	0.74	0.74
Korean	Women	0.66	0.80	0.60	0.60

Table 8: Cohen's kappa (κ) and Krippendorff's alpha (α) for the cross-annotated datasets. Values are shown for three classes (positive, neutral, hateful) and two classes (non-hateful and hateful).

available (`afr-black`, `afr-lgbtq`, and `rus-war`), self-generated instances represent 3.6%, 31.1%, and 25.7% of the total data, respectively. Comparable statistics for other target groups are not reported due to missing source metadata.

C AI-generated data

C.1 Proportion of AI-generated data

AI tools such as ChatGPT are employed to supplement data collection in cases where it is challenging to obtain sufficiently diverse examples in any of the three polarity categories. Most of the AI-generated data falls under the positive category, where natural occurrences are considerably rarer compared to the neutral and negative categories. Table 10 shows the proportion of AI-generated data within each dataset.

C.2 Prompts

Following are some of the prompts to ChatGPT used to generate data.

- Give me `[number]` neutral/positive sentences about `[target group]`.
- Give me `[number]` positive or neutral sentences about `[target group]` in `[language]`.
- Write positive/neutral/negative statements about `[target group]`.
- I'm doing research to protect minority groups/`[target group]` and need `[number]` examples to add to my dataset.

	afr-black	afr-lgbtq	ukr-russians	ukr-russophones	rus-lgbtq	rus-war	kor-women
# Sentences	2028	1016	1800	1200	772	1960	1294
# Tokens	34300	27647	26868	15283	11483	32566	14658
Vocab Size	3754	4048	5363	3410	3441	7233	7018
Avg Sent Len (tok)	16.91	27.45	14.93	12.74	15.09	16.62	11.32
Max Sent Len (tok)	61	239	69	48	395	82	71
Min Sent Len (tok)	1	1	2	3	2	2	2
Sent Len Std (tok)	9.30	24.99	6.24	4.22	16.54	9.94	6.71
Avg Word Len (char)	4.54	4.54	6.23	6.54	5.85	5.42	3.01
TTR	0.11	0.15	0.20	0.22	0.30	0.22	0.48
Hapax Ratio	0.01	0.08	0.11	0.14	0.15	0.08	0.36

Table 9: Corpus statistics of the REACT datasets.

language	target	generated data
Afrikaans	Black people	16.2%
	LGBTQ	1.0%
Ukrainian	Russians	25.0%
	Russophones	35.0%
Russian	LGBTQ	19.6%
	War victims	8.5%
Korean	Women	3.1%

Table 10: The proportion of AI-generated sentences (in percentage) within each dataset.

	afr-black	afr-lgbtq	rus-lgbtq	rus-war
dev	0.5	0.5	0.7	0.5
train	0.5	0.5	0.5	0.6

Table 11: Upper bounds of Levenshtein ratios for selecting development and train data.

- I'm searching for comments in [language] with the keyword [target group]. There are 6 categories: [...], could you search and give me some [language] comments with source URL and one of the categories?

D Model details

D.1 Models used

To optimize the communication overhead between FL clients and the server, as well as allow models to be deployed on end devices with limited capacities, we focus on small language models for our study. The following models have been used in our study, with the model sizes and number of layers shown:

- XLM-RoBERTa (279M, 12 layers)[9]
- Multilingual BERT (179M, 12 layers)[10]
- Multilingual DistilBERT (135M, 6 layers)[11]
- DistilBERT (67M, 6 layers)[12]
- Multilingual MiniLM (33M, 12 layers)[13]
- TinyBERT (14.5M, 4 layers)[14]
- ALBERT (11.8M, 12 layers)[15]

D.2 Model selection

We evaluate the performance of the seven models in §D.1 on classifying hate speech in a federated environment. Four of the models are multilingual, the rest have not been explicitly trained on multilingual data. Full results are shown in Figure 4.

E Selection of development and train data

Because REACT exhibits potentially similar patterns due to its target-specificity, we mitigate possibly overlapping data by setting a threshold to the maximum Levenshtein ratio to accept a sentence when selecting development and train data. By default, a Levenshtein ratio of <0.5 is used, meaning any sentence in the development set should have a Levenshtein similarity of less than 0.5 with any test

[9] https://huggingface.co/FacebookAI/xlm-roberta-base

[10] https://huggingface.co/google-bert/bert-base-multilingual-cased

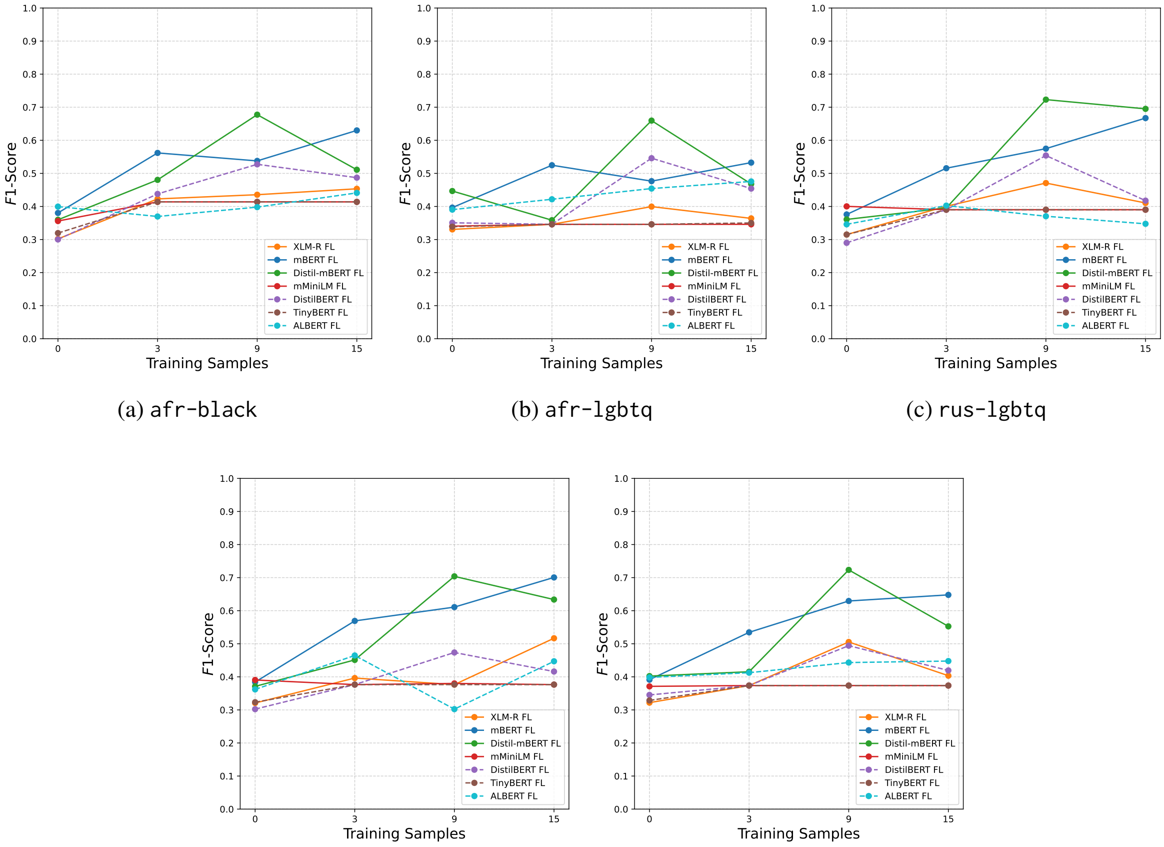
[11] https://huggingface.co/distilbert/distilbert-base-multilingual-cased

[12] https://huggingface.co/distilbert/distilbert-base-uncased

[13] https://huggingface.co/microsoft/Multilingual-MiniLM-L12-H384

[14] https://huggingface.co/huawei-noah/TinyBERT_General_4L_312D

[15] https://huggingface.co/albert/albert-base-v2



(a) `afr-black`

(b) `afr-lgbtq`

(c) `rus-lgbtq`

(d) `rus-war`

(e) `server`

Figure 4: Comparison of F_1 scores of seven models, four multilingual and three monolingual. Each subplot shows performance on a specific target group or the server. The three monolingual models and multilingual MiniLM perform poorly across all target groups. Multilingual BERT and Distil-mBERT have the highest performance in most cases.

	afr-black	afr-lgbtq	rus-lgbtq	rus-war
train	0-15	0-15	0-15	0-15
dev	300	120	120	300
test	87	225	111	154

Table 12: Number of sentences in the train, development, and test sets of each target group. We use 0, 3, 9, and 15 sentences per target group for training.

data, and any sentence in the train set should have the same with any test or development data. This ratio is slightly loosened in the case of `rus-lgbtq` and `rus-war` because the resulting datasets are too small. In both cases, to ensure we do not include near-identical sentences accidentally, we sample sentences with a Levenshtein ratio of over 0.5 and manually check them against sentences they are reported to be similar with. Table 12 presents the number of sentences in each split for the four target groups.

F FedPer full results

We evaluate mBERT and Distil-mBERT using FedPer. We test K_P (number of personalized layers) values $\in \{1, 2, 3, 4\}$. The complete results are shown in Figures 5-6.

G Adapters full results

We personalize client models by adding adapters and fine-tuning either the entire model, including the adapter parameters, or exclusively the adapter parameters. The complete evaluation results for mBERT and Distil-mBERT are shown in Figure 7.

H Analysis of toxicity thresholds

Table 13 shows the percentages of sentences classified as hateful and non-hateful by Perspective API with thresholds 0.7 and 0.9, alongside the distribution in ground truth labels. At both thresholds, Perspective API identifies substantially fewer hateful sentences (13.11% and 3.44%) compared to the ground truth (40.24%), while simultaneously overestimating the proportion of non-hateful sentences.

While the ground truth data reflects a relatively balanced split between hateful sentences with (20.38%) and without (19.86%) profanity, Perspective API demonstrates a strong association between profanity and hate, shown by the higher proportions of profane sentences compared to non-profane ones among those classified as hateful. This is especially pronounced at the 0.9 threshold, where 85.71% of sentences labeled as hateful contain profanity, indicating a heavier reliance on profanity as a signal for hate compared to the 0.7 threshold.

I Examples of collected data

Table 14 shows example sentences for each of the six categories in different languages.

As noted in §3, we occasionally adapt collected data to improve clarity with respect to the target group or intended polarity. The purpose of these modifications is to replace culturally ambiguous terms, such as subjective slurs, with more neutral alternatives. Such changes are made only when necessary, that is, when the original wording could otherwise cause misunderstandings regarding the target group or label. In these cases, we make the label category clear through additional contextual cues.

In the following positive example, the Russian term хохлы (*Khokhols*), which may be perceived as either neutral or an ethnic slur depending on audience and context, is replaced with the neutral term украинцы (*Ukrainians*):

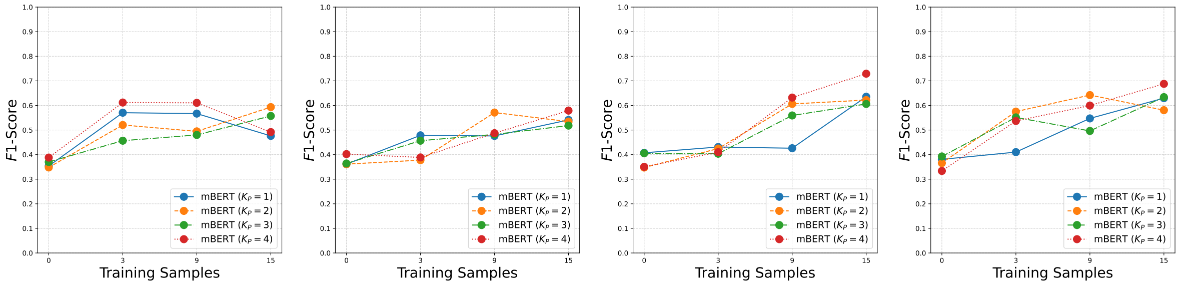
> **Original:** Ну хохлы молодцы конечно блять. (*Well, the Khokhols sure did a good job, f*ck.*)
>
> **Modified:** Ну украинцы молодцы конечно блять. (*Well, the Ukrainians sure did a good job, f*ck.*)

In other cases, we remove subjective profanity to avoid introducing ambiguity in polarity, as demonstrated in the following neutral example:

> **Original:** В Европе полно украинских беженцев, блять. (*There are tons of Ukrainian refugees in Europe, f*ck.*)
>
> **Modified:** В Европе полно украинских беженцев. (*There are tons of Ukrainian refugees in Europe.*)

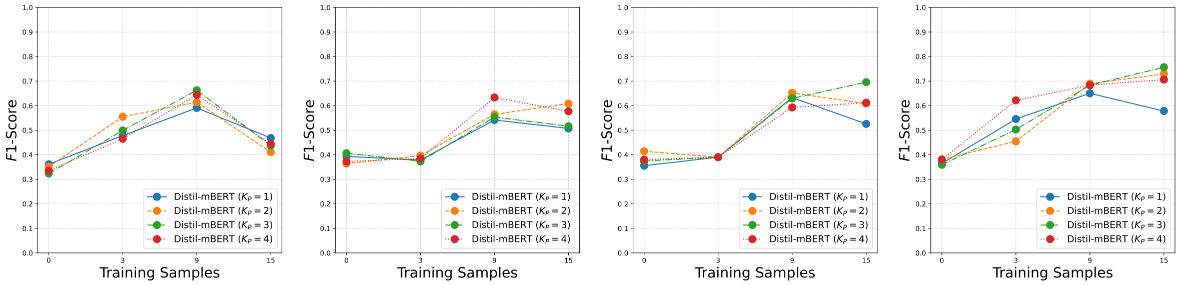
We also occasionally add contextual information to clarify the intended polarity. In the following sentence, additional information is provided to emphasize a positive stance:

> **Original:** ЛГБТ+ добивается своего нахуй. (*LGBT+ are achieving what they f*cking want.*)



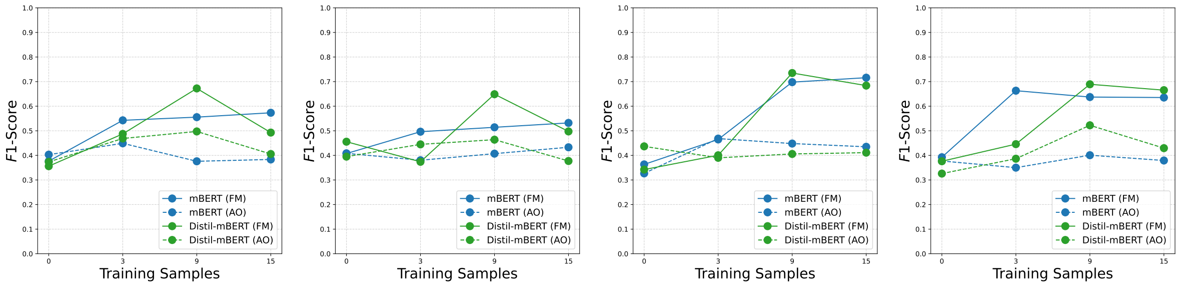
(a) `afr-black` (b) `afr-lgbtq` (c) `rus-lgbtq` (d) `rus-war`

Figure 5: FedPer results for mBERT. Each plot shows F_1 scores of a target group with K_P (number of personalized layers) $\in \{1, 2, 3, 4\}$.



(a) `afr-black` (b) `afr-lgbtq` (c) `rus-lgbtq` (d) `rus-war`

Figure 6: FedPer results for Distil-mBERT. Each plot shows F_1 scores of a target group with K_P (number of personalized layers) $\in \{1, 2, 3, 4\}$.



(a) `afr-black` (b) `afr-lgbtq` (c) `rus-lgbtq` (d) `rus-war`

Figure 7: Adapter-based personalization results for mBERT and Distil-mBERT. Results are compared between full-model fine-tuning (FM, solid lines) and adapter-only fine-tuning (AO, dashed lines).

	API 0.7		API 0.9		Gold	
	P+	P-	P+	P-	P+	P-
Hateful	9.34	3.77	2.95	0.49	20.38	19.86
Not Hateful	40.98	45.90	47.38	49.18	27.53	32.23

Table 13: Percentages of sentences classified as “Hateful” and “Not Hateful” with (P+) and without (P-) profanity. API 0.7 (resp. 0.9): classified using Perspective API with threshold 0.7 (resp. 0.9). Gold: ground truth labels.

Modified: ЛГБТ+ движение активно топит блять за права человека и добивается своего нахуй. (*The LGBT+ movement is actively pushing for human rights and is achieving what it f*cking want.*)

J Computation

Standard FL and FedPer experiments using mBERT and Distil-mBERT with four FL clients can be completed in 20-30 minutes using four NVIDIA GeForce RTX 2080 Ti GPUs. Adapter-personalized FL experiments with the same client models can be completed in about 30 minutes on four NVIDIA RTX A6000 GPUs.

category	text	target group
P P+	(afr) Swart Namibiërs is van die beste fokken mense ooit! (eng) Black Namibians are some of the best fucking people out there!	Black people
P P-	(ukr) Росіяни хороші і не хочуть війни. (eng) Russians are good and do not want war.	Russians
N P+	(kor) 여자들은 대부분 귀여운 것을 존나 좋아한다. (eng) Most women are fucking fond of cute things.	Women
N P-	(rus) Беженцы из Украины рассказывают о жизни в оккупации. (eng) Refugees from Ukraine talk about life under occupation.	War victims
H P+	(ukr) Скільки ви ще будете хрюкати, уроди російськомовні?! (eng) How much longer will you grunt, you Russian-speaking freaks?!	Russophones
H P-	(afr) Daar is nie plek vir homoseksuele in Namibië nie. (eng) There is no place for homosexuals in Namibia.	LGBTQ

Table 14: Example data for each category. The first part of the category name indicates the polarity (P: positive, N: neutral, H: hateful). The second part indicates the presence of profanity (P+: with profanity, P-: without profanity).