Data-Efficient Hate Speech Detection via Cross-Lingual Nearest Neighbor Retrieval with Limited Labeled Data

Faeze Ghorbanpour[1,2,3] **Daryna Dementieva**[1,3] **Alexander Fraser**[1,3]

[1]School of Computation, Information and Technology, TU Munich
[2]Center for Information and Language Processing, LMU Munich
[3]Munich Center for Machine Learning (MCML)

faeze.ghorbanpour@tum.de, daryna.dementieva@tum.de

Abstract

Considering the importance of detecting hateful content, labeled hate speech data is expensive and time-consuming to collect and annotate, particularly for low-resource languages. Prior work has demonstrated the effectiveness of cross-lingual transfer learning and data augmentation in improving performance on tasks with limited labeled data. To develop an efficient and scalable cross-lingual transfer learning approach, we leverage nearest-neighbor retrieval to augment minimal labeled data in the target language, thereby enhancing detection performance. Specifically, we assume access to a small set of labeled training instances in the target language and use these to retrieve the most relevant labeled examples from a large multilingual hate speech detection pool. We evaluate our approach on eight languages and demonstrate that it consistently outperforms models trained solely on the target language data. Furthermore, in most cases, our method surpasses the current state-of-the-art. Notably, our approach is highly data-efficient, retrieving as few as 200 instances in some cases while maintaining superior performance. Moreover, it is scalable, as the retrieval pool can be easily expanded, and the method can be readily adapted to new languages and tasks. We also apply maximum marginal relevance to mitigate redundancy and filter out highly similar retrieved instances, resulting in improvements in some languages. [1]

Content warning: This paper contains examples of hateful and abusive language.

1 Introduction

Hate speech, *abusive language targeting specific groups* (Röttger et al., 2021), is a global issue. However, most detection advancements focus on English due to the abundance of labeled datasets (Poletto et al., 2021; Yin and Zubiaga, 2021). In contrast, languages like Spanish, French, and Italian, though not low-resource for other tasks, lack annotated hate speech datasets (Poletto et al., 2021), limiting model effectiveness in detecting and addressing hate speech.

Collecting and annotating data for low-resource languages is an effective solution, especially for capturing linguistic and cultural nuances in hate speech (Pelicon et al., 2021; Aluru et al., 2020a). As Röttger et al. (2022) state, having some labeled data in the target language is crucial for model effectiveness. However, while obtaining more data can improve performance, this requires paying high annotation costs (ElSherief et al., 2021) and exposing annotators to harmful content (AlEmadi and Zaghouani, 2024).

Transfer learning, especially from high-resource languages like English, helps mitigate data scarcity and improve detection performance (Bigoulaeva et al., 2022; Firmino et al., 2024). However, the choice of source tasks and languages remains crucial. Some languages are useful for specific target languages due to cultural similarities (Zhou et al., 2023), and certain source tasks may be more useful for particular target tasks (Röttger et al., 2022; Antypas and Camacho-Collados, 2023).

Training on all available hate speech datasets may seem beneficial, but it is often inefficient, computationally costly, and does not guarantee better performance (Caselli et al., 2020). It can introduce redundancy, dataset-specific biases, and annotation inconsistencies, leading to overfitting (Wiegand et al., 2019; Fortuna and Nunes, 2018). Moreover, this approach lacks scalability, requiring frequent retraining for new datasets (Vidgen et al., 2021a).

To address the mentioned problems, we propose a novel method based on cross-lingual nearest-neighbor retrieval. Our approach, pictured in Figure 1, retrieves a minimal yet relevant set of instances and integrates them with the target lan-

[1]The official implementation of the method is publicly available on: https://github.com/FaezeGhorbanpour/MultilingualDataEfficientDetection/

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 29663–29681
November 4-9, 2025

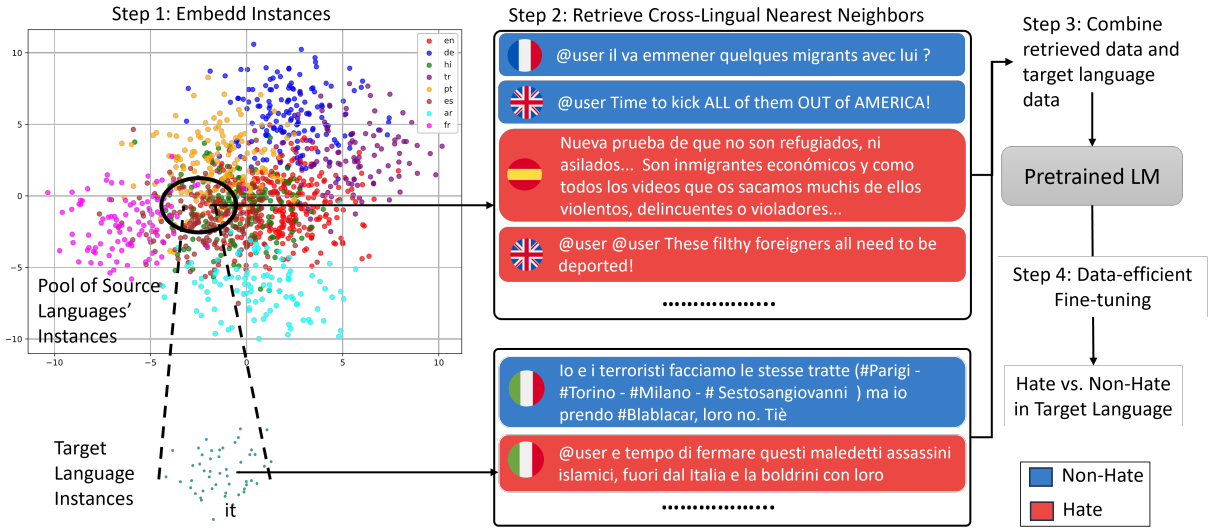

Figure 1: Overview of the proposed method. Given a small number of examples from a target language, we search in a large pool of multilingual data for closely related instances. We then combine the retrieved instances with the target language data and train a multilingual model on them for hate speech detection.

guage training set for fine-tuning. Specifically, we embed all available instances from fourteen tasks using a multilingual sentence embedding model to create a pool of hate speech detection samples. A retrieval system selects the most relevant instances from the multilingual pool based on their distance to the target language training set. These retrieved instances are then combined with the target training data to fine-tune a language model (LM).

This solution addresses several challenges. First, retrieving from a multilingual pool removes the need to search for the best source task or language. Second, it improves efficiency by selecting only a small number of relevant samples and reducing redundancy through distance-based retrieval. Third, it supports scalability, as the multilingual pool can be easily extended with new datasets and languages. Finally, our method enhances cross-lingual transfer learning by leveraging linguistic and semantic similarities in hate speech across languages.

We evaluate the proposed method on eight languages, including German, French, Spanish, Italian, Portuguese, Hindi, Arabic, and Turkish, simulating a scenario where only a limited number of training examples (ranging from 10 to 2,000) are available. Fine-tuning on a combination of retrieved data and the target language training set significantly outperformed fine-tuning solely on the target training set across all languages. Further, our method outperformed the state-of-the-art work in most languages while fine-tuning with fewer samples. To refine the retrieved data, we also experiment with applying maximum marginal relevance (MMR) (Carbonell and Goldstein, 1998) to remove highly similar instances, leading to improved performance in some languages. Our contributions are as follows:

- We propose a novel, efficient, and scalable method for enhancing limited labeled hate speech datasets by retrieving cross-lingual samples using a retrieval system.
- We evaluate our method on eight languages, demonstrating consistently higher performance compared to training solely on the target language training set.
- Our approach is particularly effective in extremely low-resource settings with fewer than 50 labeled instances, achieving improvements of up to 10 F1-macro points in some cases.

2 Related Work

Hate Speech Detection with Limited Labeled Data: Hate speech violates human rights, disrupts social peace, incites violence, and promotes discrimination in all societies, regardless of language. Detecting it is crucial to prevent conflict and protect mental health and societal safety (Wilson, 2019; Narula and Chaudhary, 2024). Most datasets and research efforts focus on English (Kennedy et al., 2020b; Toraman et al., 2022; Ghorbanpour et al., 2025), while languages like Spanish, Portuguese, or Ukrainian have very limited resources. According to the Hate Speech Dataset

Catalogue, these languages each have only one available dataset with fewer than 5,000 samples, which are also restricted in context and domain (Basile et al., 2019; Fortuna et al., 2019; Dementieva et al., 2024). Due to limited resources, recent research increasingly leverages other languages to improve hate speech detection in low-resource settings.

Cross-lingual Transfer Learning for Hate Speech Detection: Cross-lingual transfer learning has been widely studied in NLP, showing that models trained on high-resource languages can improve performance in low-resource languages (Parovic et al., 2023; Muraoka et al., 2023; Pham et al., 2024). This makes it a promising approach for hate speech detection in low-resource settings. Early methods used multilingual embeddings for zero-shot and few-shot transfer from resource-rich to resource-poor languages (Aluru et al., 2020b; Pamungkas and Patti, 2019). Also, Bigoulaeva et al. (2021) and Monnar et al. (2024) utilized bilingual embeddings to transfer knowledge from high-resource languages, showing promising results even without labeled data in target languages, but mainly benefiting closely related languages.

Data augmentation strategies, including cross-lingual paraphrasing or translation-based methods, have also been shown to alleviate data scarcity (Pamungkas et al., 2021; Beddiar et al., 2021), but these approaches are often constrained by the availability and quality of translation resources. Roychowdhury and Gupta (2023) employed data augmentation with EasyMixup and reframed the task as textual entailment, achieving improvements but still relying on potentially noisy augmented data. Hashmi et al. (2025), Gharoun et al. (2024), and Mozafari et al. (2022) use meta-learning approaches specialized for bilingual contexts. While effective, these methods require extensive labeled bilingual data, are complex to implement and train, and often demand substantial computational resources, making them less scalable.

Röttger et al. (2022) showed that minimal target-language data and initial English fine-tuning improve performance. However, selecting an appropriate intermediate English task is challenging and language-dependent. Building on this, Goldzycher et al. (2023) uses an intermediate natural language inference (NLI) task, which adds training steps and requires more computation. Unlike prior approaches, our method eliminates the need for large-scale target-language annotation, intermediate tasks, or translation resources. Directly leveraging semantic similarity at the instance level enables effective transfer with minimal target data and avoids costly cross-lingual training pipelines.

Retrieval-based and Instance attribution Fine-tuning methods: Prior work has shown that cross-task retrieval-based data can improve generalization in LMs (Guu et al., 2020; Khandelwal et al., 2020). Shi et al. (2022) applied retrieval to classification tasks via heuristic label mapping, whereas we fine-tune directly on nearest neighbors. Das and Khetan (2024) introduces data-efficient fine-tuning through unsupervised core-set selection, showing strong results in monolingual text-editing tasks. However, this method is not designed for cross-lingual transfer and depends on clustering quality.

Our approach is similar to Lin et al. (2022) and Ivison et al. (2023) in using nearest neighbor retrieval and further fine-tuning, but is uniquely applied to multilingual datasets and leverages labeled hate speech data. Our method uses instance attribution, identifying relevant training examples for a data point, unlike prior work (Pruthi et al., 2020; Han and Tsvetkov, 2022), which used gradient-based instance attribution to interpret neural network predictions. Our neighbor identification approach is simpler as it avoids gradient computations and reliance on labels, and is applied in a multilingual, low-resource setting.

3 Methodology

Building on a large pool of labeled multilingual hate speech data, our core hypothesis is that certain instances in this pool are more relevant to a given target language than others. For each target language, we assume access to a small amount of labeled data. The goal is to identify a relevant subset of source data that, when used for training, yields better performance. Initially, we employ an embedding model (*Embedder*) to encode instances from multiple source languages. We then use a retrieval module (*Indexer*) to index the resulting embedding vectors and construct a pool of multilingual hate speech detection instances.

When detecting hate speech in a low-resource target language, the objective is to fine-tune an LM for effective and efficient classification, as depicted in Figure 1. We begin by embedding the target language instances using the same embedding model. The retrieval module (*Retriever*) then searches the

pool to find the nearest neighbors of the target instances. We combine the retrieved data with the target data and use the combined set to fine-tune (*Fine-tuner*) an LM to classify them as Hate or Non-Hate. Each module is described below.

3.1 Embedder and Indexer

Assume a source language[2] A with a set of n text instances $X^s = \{x_1^s, x_2^s, \ldots, x_n^s\}$ and corresponding labels $Y^s \in \{0, 1\}$, where 1 indicates hate speech and 0 indicates non-hate speech. The objective of this module is to project the input texts into a vector space $V^s = \{v_1^s, v_2^s, \ldots, v_n^s\}$, where each vector v_i^s is obtained by applying an embedding function: $v_i^s = \text{embedding}(x_i^s)$. These embeddings are then passed to the retrieval module, which indexes the vectors to enable efficient similarity search. This indexed embedding space serves as the foundation for retrieving relevant instances.

3.2 Retriever

Consider a target language B with a limited set of labeled data $X^t = \{x_1^t, x_2^t, \ldots, x_m^t\}$, where $m \ll n$ (m and n denote the number of target and source language instances, respectively.), and a label set $Y^t \in \{0, 1\}$, where 1 denotes hate speech and 0 denotes non-hate speech (the same label set as the source language). Similar to the source language, we apply the embedding module to convert the target language instances into a numerical vector space $V^t = \{v_1^t, v_2^t, \ldots, v_m^t\}$, where each vector is computed as $v_i^t = \text{embedding}(x_i^t)$.

The retrieval module is then employed to find relevant samples from the pool using a nearest neighbor search. Specifically, we want to retrieve a total of R instances from the source pool based on *Euclidean distance* between the embedded target vectors V^t and the source vectors V^s. The distance between an embedded target instance v_i^t and a source instance v_j^s is calculated as:

$$\text{dist}(v_i^t, v_j^s) = \|v_i^t - v_j^s\|_2 = \sqrt{\sum_{k=1}^{d} (v_{i,k}^t - v_{j,k}^s)^2}$$

Where d is the dimensionality of the embedding space. We then select the top k nearest neighbors for each v_i^t, and define the full retrieval set as:

$$\mathcal{R} = \bigcup_{i=1}^{m} \text{TopK}(v_i^t, V^s, k)$$

where $\text{TopK}(v_i', V^s, k)$ denotes the set of k source vectors in V^s with the lowest distance to v_i^t. The set $\mathcal{R}$ contains up to $m \times k$ total retrieved instances. We then map the vectors in $\mathcal{R}$ back to their corresponding original texts using the retrieval index ($X_r^s = \{x_{r_1}^s, x_{r_2}^s, x_{r_3}^s, \ldots, x_{r_R}^s\}$). Finally, we apply deduplication to remove exact textual duplicates. If the final count of unique instances falls short of R, the retrieval process continues until the desired number is reached.

3.3 Fine-tuner

In the fine-tuning module, we combine the retrieved texts (X_r^s) with the training data from the target language (X^t). The combined dataset is then used to fine-tune a pre-trained LM ($\mathcal{M}$) to perform binary classification. Since the source and target tasks share the same label space, i.e., $Y^s, Y^t = \{0, 1\}$, where 0 denotes non-hate and 1 denotes hate, joint training of the fine-tuned model on the combined source and target data is well-defined and coherent. We define the final training set as $\mathcal{D} = X_r^s \cup X^t$. The model is fine-tuned by minimizing the binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \Big[y \log \mathcal{M}(x) + (1-y) \log (1 - \mathcal{M}(x)) \Big]$$

4 Experimental Setup

4.1 Datasets

We use six large-scale English hate speech detection datasets as well as eight non-English ones. These datasets were selected based on two criteria: (a) the presence of a label designated as *hate speech*, and (b) the use of annotation guidelines that align with or are closely related to the definition of hate speech adopted in this study. In our setting, each dataset corresponds to a binary classification task (hate vs. non-hate) in a given language, so the terms dataset and task are used interchangeably.

The English datasets are: *Dyn21_en* (Vidgen et al., 2021b), *Fou18_en* (Founta et al., 2018), *Ken20_en* (Kennedy et al., 2020a), *HateXplain*

[2]For clarity, we describe the approach using a single source language. In practice, however, our methodology incorporates multiple source languages—eight in total.

(Mathew et al., 2021), *Implicit_hate*[3] (ElSherief et al., 2021), and *Xdomain_en* (Toraman et al., 2022).

The non-English datasets (each defining a target task) are: *Bas19_es* (Basile et al., 2019), *For19_pt* (Fortuna et al., 2019), *Has21_hi* (Mandl et al., 2021), *Our19_ar* and *Our19_fr* (Ousidhoum et al., 2019), *San20_it* (Sanguinetti et al., 2020), *Xdomain_tr* (Toraman et al., 2022), and *Gahd24_de* (Goldzycher et al., 2024). The two-character suffix indicates the language of the task. More details are provided in Appendix A.

Although all datasets are embedded and included in the shared retrieval pool, we ensure that, for each non-English target task, instances from the same language are excluded from retrieval. This guarantees that the target language data remains unseen during its own retrieval process. Additionally, we exclude *Dyn21_en* when the target task is *Gahd24_de* because the latter includes translations from the former. We also exclude *Xdomain_en* when the target task is *Xdomain_tr*, as both originate from the same source. After constructing the multilingual pool, we obtain approximately 265,671 instances, of which 37.15% are labeled as hateful. The majority of data in the pool is English (66.99%), Turkish (17.0%), and German (3.84%).

4.2 Models

For embedding the text instances, we utilize the `BAAI/bge-m3` multilingual encoder model (Chen et al., 2024) using the Sentence Transformers library (Reimers and Gurevych, 2020). This model generates 1024-dimensional vector representations for each input text. We use the FAISS library (Douze et al., 2024; Johnson et al., 2021) to index dense vectors and perform a similarity search. For retrieval, we adopt the Hierarchical Navigable Small World (HNSW) algorithm (Malkov and Yashunin, 2020) as an efficient approximation of the k-nearest neighbor search. Throughout all our experiments for the classification model, we fine-tune and evaluate XLM-T (Barbieri et al., 2022) using the HuggingFace Transformers library (Wolf et al., 2020). XLM-T is a variant of XLM-R (Conneau et al., 2020), further pre-trained on 198 million multilingual Twitter posts to better capture social media language patterns. Further details on hyperparameters and experimental settings are provided in Appendix B.

4.3 Evaluation Details

We simulate low-resource conditions by using 12 different training subset sizes for each non-English language: 10, 20, 30, 40, 50, 100, 200, 300, 400, 500, 1,000, and 2,000 examples. For each subset size, we run experiments with 5 random seeds. Across all experiments, we use a fixed validation set of 500 examples and a test set of 2,000 examples for each target language[4]. We only use the *training split* of the target language for retrieval and fine-tuning. The test set remains entirely unseen throughout the process to ensure evaluation integrity and is kept fixed across all experiments.

The performance comparison is based on the F1-macro metric. We compare our method to the common practice of fine-tuning solely on the target training set, referred to as `Mono`. We also compare against the approach by Röttger et al. (2022), which performs intermediate fine-tuning on three English hate speech datasets (20,000 instances each) to identify the most effective source task and then fine-tunes on the target language training set. We report the best result among the three as `Röttger`.

5 Results

Table 1 reports results for eight target languages. Each row corresponds to a subset of the target-language training data (e.g., 20, 50, 200, 500, or 2,000 examples); the full set of twelve subset sizes is in appendix D. These subset sizes indicate only the number of target-language examples. Additional columns (20, 200, 2,000, 20,000) show how many instances were retrieved from the multilingual pool and added to the target subset. Thus, a subset of 20 combined with 200 retrieved instances yields 220 training examples in total. The `Mono` and `Röttger` columns are baselines, and the `AVG` row gives the average across all twelve subset sizes.

In all languages, retrieving **as few as 20 instances** and adding it to the original train set for fine-tuning already outperforms the `Mono` setting, indicating the effectiveness of our proposed method and the value of cross-lingual data. This is particularly promising for target tasks with fewer than 50 instances, where the F1-macro score improves by 10 in some languages such as *San20_it*, *Ous19_ar*, and *Xdomain_tr*. While the performance gain de-

[3]This dataset includes both explicit and implicit hate speech, which we merge into a single label, *hate speech*.

[4]For Arabic and French, smaller dataset sizes limited the test sets to 1,000 and 1,500 samples, respectively.

	San20_it						Ous19_ar						Ous19_fr					
SIZE	Mono	20	200	2,000	20,000	Röttger	Mono	20	200	2,000	20,000	Röttger	Mono	20	200	2,000	20,000	Röttger
20	54.25	63.20	66.76	**67.06**	60.06	64.96	51.67	57.63	**63.23**	61.73	59.47	60.52	47.26	47.21	52.68	53.93	**55.05**	52.93
50	65.71	67.20	68.42	**71.65**	69.44	69.10	52.13	59.36	**66.65**	66.31	64.51	65.76	47.87	48.29	52.19	52.97	**55.60**	54.15
200	72.81	72.46	**72.83**	72.41	72.50	71.56	67.97	67.98	**69.18**	67.35	65.47	66.61	51.93	51.54	54.06	**55.80**	53.63	53.76
500	74.18	75.39	**75.29**	74.53	66.09	73.69	66.54	68.95	**69.47**	69.28	65.54	67.60	51.91	53.30	52.84	**55.51**	55.31	53.39
2,000	76.40	69.27	**78.36**	77.57	76.95	77.07	66.91	69.52	69.77	**70.15**	68.27	67.07	51.84	53.51	53.13	53.30	**54.74**	52.89
AVG	66.53	67.68	71.00	**71.06**	68.55	70.47	59.82	63.94	**66.82**	66.41	65.20	65.41	49.72	50.56	53.12	54.05	**54.84**	53.88
	Bas19_es						For19_pt						Xdomain_tr					
20	49.91	54.37	59.72	62.52	63.08	**66.52**	48.09	49.72	64.92	**68.57**	68.03	67.68	55.43	66.58	67.08	70.14	**75.87**	69.80
50	61.85	60.93	64.37	65.59	64.30	**70.36**	60.25	59.26	67.01	67.06	**69.35**	66.51	72.24	75.92	77.50	**78.85**	70.60	75.12
200	72.36	72.22	71.77	71.23	70.67	**75.27**	66.91	69.69	70.33	70.20	**71.07**	68.10	81.63	81.61	82.61	**83.04**	82.61	82.19
500	77.14	78.01	77.09	77.79	67.67	**78.76**	69.95	69.72	70.84	70.04	**71.05**	69.22	85.05	84.93	85.09	84.92	83.88	**85.34**
2,000	81.08	80.62	80.50	80.65	81.02	**82.04**	**72.70**	72.39	72.66	71.72	72.22	71.61	88.53	87.39	88.00	87.39	77.48	**88.84**
AVG	65.52	67.27	69.53	70.53	68.69	**72.97**	61.66	62.85	68.18	69.55	**69.69**	68.39	73.58	76.78	78.88	79.66	77.58	**80.27**
	Gahd24_de						Has21_hi											
20	44.99	50.52	58.15	59.08	57.48	**59.82**	46.87	47.34	51.03	53.68	**55.37**	54.92						
50	57.85	54.53	60.30	60.47	61.02	**62.57**	46.87	48.39	53.36	52.26	**55.78**	54.77						
200	65.80	**66.95**	66.15	66.24	65.97	64.25	52.20	55.83	54.65	56.80	56.02	**57.47**						
500	66.56	69.78	69.68	**70.45**	61.80	67.02	56.20	56.94	57.66	57.88	**59.55**	57.96						
2,000	73.77	**79.19**	78.79	77.82	77.90	72.42	57.14	58.19	60.22	**60.50**	59.65	58.01						
AVG	60.06	62.98	64.77	**65.36**	64.18	64.23	50.96	52.44	55.10	56.25	**57.05**	56.70						

Table 1: Performance (F1-macro) across eight target languages with varying amounts of target-language supervision. Each block shows results for a single language. Rows indicate the number of target-language examples used, while columns show the number of retrieved cross-lingual neighbors added during training. Mono and Röttger are baseline methods. AVG reports the average over twelve training sizes (full results in Appendix D). Best scores are in **bold**; retrieved variants that outperform the next-larger Mono size are underlined.

creases as more target language training data becomes available, the average results consistently show that leveraging cross-lingual data outperforms relying solely on the target language's training set. In most languages—except for *Bas19_es* and *Xdomain_tr*—our proposed method outperforms the Röttger on average, while using less training data and without requiring manual selection of intermediate tasks. Notably, retrieving around 200 instances often yields comparable or even superior performance to this work, which uses 20,000 training size for intermediate fine-tuning.

Another insight from Table 1 is how cross-lingual retrieval can compensate for limited labeled data in the target language. For example, in Hindi, retrieving just 20 instances for a training size of 20 matches the performance of having 50 labeled examples, and retrieving 2,000 instances approaches the performance of having 200 labeled instances. This pattern is consistent across other underlined values in the table. In languages where Mono performance with 2,000 training samples fails to exceed 70—as in *Ous19_ar*, *Ous19_fr*, and *Has21_hi*—retrieval proves especially valuable, often matching the next training size.

For languages where Mono's highest performance is less than 75 (*Gahd24_de* and *For19_pt*), retrieval remains helpful, compensating for up to 500 labeled examples. However, in languages where Mono performance exceeds 75 with 2,000 samples, retrieval is less beneficial—except in the extreme low-data case: with only 20 labeled data, retrieval consistently outperforms the Mono model trained on 50 examples across all languages.

Another observation from Table 1 is that, in five languages—excluding *Ous19_fr*, *For19_pt*, and *Has21_hi*—the highest average performance is achieved by retrieving 2,000 instances, while retrieving 20,000 leads to a performance drop. For instance, in *Ous19_ar*, retrieving 200 instances yields the best result. This suggests that increasing the number of retrieved data points for fine-tuning does not necessarily lead to improved performance.

How Much Retrieved Data Is Sufficient? To address this question, we conducted an experiment varying the number of retrieved instances across 21 settings, from 10 to 100,000 (More than a third of the pool size), for four languages as shown in Figure 2. The figure includes five different training sizes, each represented by a distinct color. The brown line labeled AVG denotes the average performance over 12 training sizes.

As shown in the figure, especially in the average trend line—where the effects of noise are diminished due to averaging— performance increases

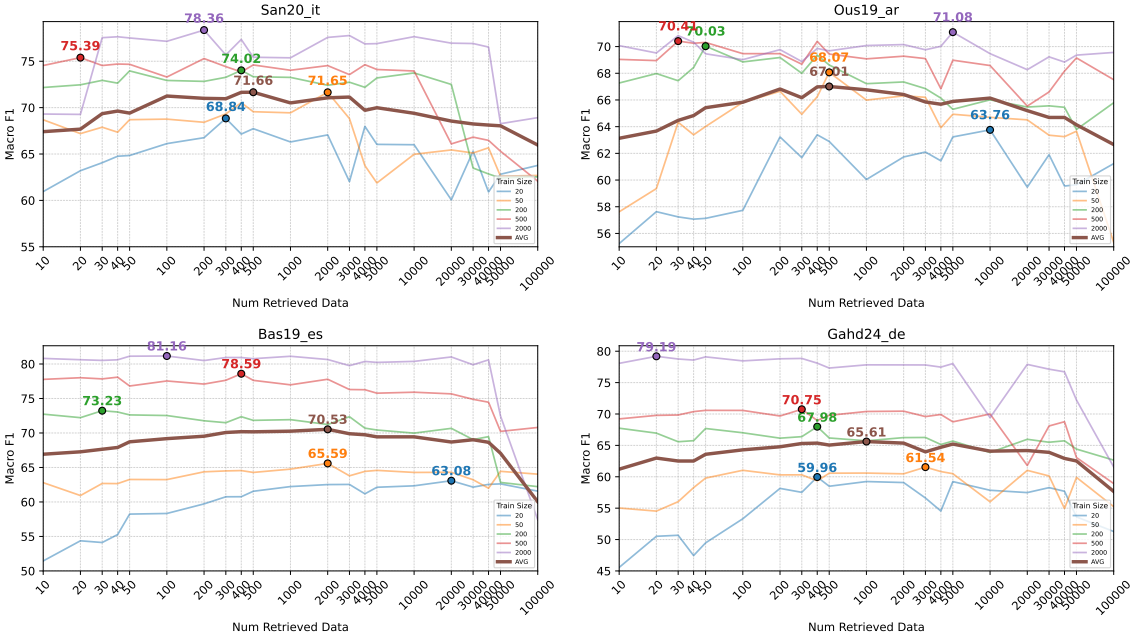

Figure 2: Effect of the number of retrieved instances (10 to 100,000, log-scaled) on F1-macro performance across four target languages. Each curve corresponds to a selected amount of target-language training data. Retrieved instances are added to this target subset during training. Highlighted points mark the best performance for each training size. The bold brown curve shows the average over 12 target-language subset sizes.

as the number of retrieved instances grows—up to around 2,000—after which it gradually declines. This change is more pronounced for smaller training sizes (e.g., 20), while for larger sizes (e.g., 2,000), the effect is minimal. These results suggest that adding more retrieved data is not always beneficial, and peak performance is typically reached with around 2,000 retrieved instances.

We also tested alternative embedder models, different retriever criteria, as well as label balancing and weighting the target training set, but observed no notable differences. Full details are provided in Appendix E.

5.1 Retrieved Languages Distribution

An interesting analysis is to examine which tasks or languages the retrieved data come from for each target language. This is illustrated in Figure 3, which shows the average retrieval distribution when retrieving 2,000 instances for a training size of 2,000, averaged over five random subsamples of the original training set. In the Sankey diagram, source tasks are shown on the left and target tasks on the right, with edges representing the four most frequently retrieved source tasks for each target task. Due to the dominance of English data in the pool, a higher proportion of English instances is expected, with *Ken20_en* and *Fou18_en* being the most commonly retrieved source tasks.

However, we also observe non-negligible retrieval from smaller source tasks, such as Arabic, highlighting semantic and contextual relevance between hate speech in source and target languages. We can also see that linguistically or culturally related languages tend to support each other: Portuguese benefits French, Turkish supports Arabic, and Italian aids Spanish. This highlights the effectiveness of our approach in identifying culturally proximate examples. This retrieval pattern can also be due to shared annotation styles or content overlap. Further diagrams are in Appendix F.

5.2 Error Analysis

To better understand retrieval behavior, we conducted an error analysis on Spanish (*Bas19_es*) and Italian (*San20_it*) samples. For each language, we retrieved a total of ten neighbors (not ten per target instance). Offensive terms are anonymized with placeholders such as "[slur]" or "[abuse]". Tables 3 present representative examples, showing both correct semantic matches and failure cases.

In both languages, retrieval frequently aligned hateful targets with hateful neighbors across languages (e.g., insults in Spanish matched to abusive English phrases, religious hate in Italian matched to Turkish discourse condemning homosexuality). Like-

		Bas19_es				For19_pt				Has21_hi				Ous19_fr			
	SIZE	20	200	2000	20000	20	200	2000	20000	20	200	2000	20000	20	200	2000	20000
English Retrieval	20	**55.83**	**61.15**	61.92	63.03	**52.00**	**65.73**	68.11	67.55	46.92	**52.72**	**56.01**	**55.71**	47.26	51.17	53.58	55.10
	50	**61.67**	**65.19**	64.13	62.91	57.58	66.86	**67.71**	68.52	48.90	52.67	**55.62**	55.46	**48.64**	51.56	54.38	53.35
	200	71.53	71.08	71.01	68.24	69.13	70.04	68.73	70.19	54.22	**56.24**	**58.62**	56.44	50.70	**56.15**	54.11	53.56
	500	77.13	76.28	76.67	74.55	69.60	70.67	**70.77**	70.02	55.23	57.54	**59.83**	58.02	51.75	53.22	52.46	53.48
	2,000	80.67	80.68	80.68	80.54	72.07	72.32	71.12	64.82	57.40	59.00	**60.63**	**60.74**	50.47	53.11	52.31	53.82
	AVG	67.27	**69.98**	69.61	68.62	62.86	**68.53**	69.02	68.32	52.04	**55.22**	**57.19**	56.24	50.00	52.40	53.38	54.07
Multilingual Retrieval	20	54.37	59.72	**62.52**	63.08	49.72	64.92	68.57	**68.03**	**47.34**	51.03	53.68	55.37	47.21	**52.68**	53.93	55.05
	50	60.93	64.37	**65.59**	**64.30**	59.26	**67.01**	67.06	**69.35**	**48.39**	**53.36**	52.26	**55.78**	48.29	52.19	52.97	**55.60**
	200	**72.22**	71.77	71.23	70.67	**69.69**	**70.33**	**70.20**	**71.07**	**55.83**	54.65	56.80	56.02	51.54	54.06	**55.80**	53.63
	500	**78.01**	77.09	**77.79**	67.67	69.72	**70.84**	70.04	**71.05**	**56.94**	57.66	57.88	**59.55**	**53.30**	52.84	**55.51**	**55.31**
	2,000	80.62	80.50	80.65	**81.02**	**72.39**	**72.66**	71.72	**72.22**	**58.19**	**60.22**	60.50	59.65	**53.51**	53.13	**53.30**	**54.74**
	AVG	67.27	69.53	**70.53**	68.69	62.85	68.18	**69.55**	**69.69**	**52.44**	55.10	56.25	**57.05**	**50.56**	**53.12**	**54.05**	**54.84**

Table 2: F1-macro scores for two retrieval strategies: English-only and Multilingual. Results are shown for target training sizes of 20, 50, 200, 500, 2,000, and AVG (the average over 12 training sizes), across retrieved sample sizes of 20, 200, 2,000, and 20,000.

	Target Language	**Retrieved Data**	**Outcome**
Bas19_es	@user . . . cállate [abuse] . . . (Hate)	Shut the [abuse] up . . . (EN, Non)	Good match
	Ese buey feliz en su país . . . (Non)	Annoying ass girlfriend . . . (EN, Hate)	Retrieval noise
	Después ¿cómo se limpia . . . peazo [abuse] (Hate)	Ugly [abuse]!! (EN, Non)	Hostility aligned
	La portada de depor . . . jeques árabes (Non)	Sale arabe??? (FR, Hate)	Keyword match
	Callate . . . pedazo de [abuse] (Hate)	Shut the [abuse] up . . . (EN, Non)	Strong match
San20_it	Atei, omosessuali, donne non ubbidienti . . . a morte! (Hate)	Death penalty for homosexuals . . . [slur] (EN, Hate)	Strong match
	Alle prime due ore . . . xenofobi e razzisti . . . (Non)	Discrimination against minorities . . . (TR, Non)	Related, different focus
	@user . . . categoria di persone ingestibile . . . (Hate)	Escoria . . . antifascismo . . . (ES, Hate)	Partial overlap
	Modena, festa della donna islamica velata . . . (Hate)	Eşcinsellik . . . Islam ideolojisine göre yasaktır . . . (TR, Hate)	Aligned hostility
	Ma secondo te un disperato . . . migranti . . . (Non)	Su eres nazi . . . te mate . . . (ES, Hate)	Retrieval noise

Table 3: Examples of Spanish and Italian target samples with retrieved neighbors from the multilingual pool. The texts are shortened and anonymized. The table shows cross-lingual matches where hateful targets align with hateful neighbors, as well as cases of mismatches or retrieval noise.

wise, non-hate examples often retrieved neutral or supportive content (e.g., Italian pro-migrant texts retrieved Turkish feminist or minority rights discourse). These patterns illustrate why retrieval is effective: cross-lingual embeddings cluster texts by semantic stance toward targets (hostility vs. support), enabling small target datasets to be augmented with meaningful additional training data. Mismatches occur, especially when retrieval relies on topical overlap rather than stance, but overall, the approach successfully amplifies low-resource data.

5.3 English-only vs Multilingual Retrieval

This experiment examines the effect of multilingual retrieval by comparing it to English-only retrieval, where data is retrieved exclusively from English tasks. Table 2 presents the results: rows are retrieval settings, and columns represent four target languages (see Appendix G for other languages). Comparisons should be made vertically within each language—for example, comparing 20 training samples with 20 retrieved instances across the two row blocks. We observe only minor differences in overall performance across the two settings in the table, likely due to the high proportion of English data in the pool. However, in specific cases—such as retrieving 2,000 instances for *Bas19_es* and *For19_pt*, and 20 or 2,000 instances for *Has21_hi* and *Ous19_fr*—multilingual retrieval yields higher performance. This suggests that incorporating even a small amount of multilingual data can be beneficial.

6 Maximum Marginal Relevance

As an additional deduplication step, we apply *Maximum Marginal Relevance (MMR)* in the retrieval module—before mapping the retrieved vectors back to their original texts—to ensure both relevance and diversity. Specifically, we retrieve

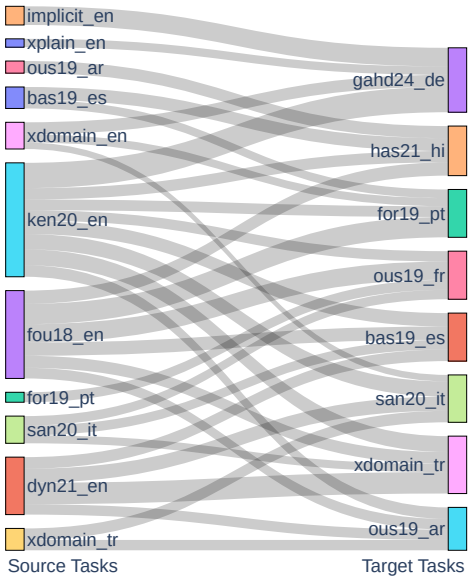

Figure 3: Sankey diagram of the distribution of the top four retrieved source tasks per target task.

at least $2R$ candidate vectors and iteratively select R vectors that balance similarity to the query and dissimilarity to previously selected vectors. Given a query vector q, a candidate set D, and a selected set S, MMR selects the next vector $v^* \in D \setminus S$ as:

$$\text{MMR}(v^*) = \arg\max_{v \in D \setminus S} \Big[\lambda \cdot \cos(v, q) - (1 - \lambda) \cdot \max_{s \in S} \cos(v, s) \Big]$$

Here, $\lambda \in [0, 1]$ controls the trade-off between relevance to the query and diversity with respect to the selected set. We set the $\lambda = 0.5$. This process is repeated until exactly R vectors are selected.

Although removing highly similar instances using MMR increases the diversity of the retrieved data, incorporating it does not substantially affect performance, with results remaining largely similar across most languages—except for those listed in Table 4 (see Appendix H for the remaining languages). As shown in the figure, for these three languages, applying MMR particularly improves performance when retrieving fewer than 2,000 instances. In contrast, for 20,000 retrieved instances, the performance without MMR is higher. This suggests that when only a limited number of instances is retrieved, MMR helps select fewer but more diverse examples, which can lead to improved performance. In our default setup, we remove exact duplicates but retain near-duplicates, such as semantically similar content in different languages. MMR mitigates this by downweighting overly similar examples. Interestingly, for the Turkish dataset—where our method previously underperformed without MMR—applying it allows the model to surpass the performance of `Röttger`.

		Without MMR				With MMR			
	SIZE	20	200	2000	20000	20	200	2000	20000
San20_it	20	**63.20**	66.76	**67.06**	60.06	60.28	**66.89**	61.69	**62.38**
	50	67.20	**68.42**	**71.65**	**69.44**	**69.71**	68.30	70.11	63.30
	200	72.46	**72.83**	72.41	**72.50**	72.51	72.66	**73.07**	72.08
	500	**75.39**	75.29	74.53	66.09	74.93	**75.41**	**75.00**	**75.69**
	2,000	69.27	**78.36**	77.57	**76.95**	**77.66**	77.06	**76.67**	68.88
	AVG	67.68	71.00	71.06	**68.55**	**69.24**	**71.64**	**71.38**	68.33
Gahd24_de	20	50.52	58.15	**59.08**	57.48	**51.47**	58.14	58.63	**59.14**
	50	54.53	60.30	60.47	**61.02**	54.83	60.07	**61.13**	60.50
	200	66.95	66.15	66.24	**65.97**	**68.09**	66.20	**67.13**	65.41
	500	69.78	69.68	70.45	61.80	**70.36**	**70.11**	70.06	**62.77**
	2,000	**79.19**	78.79	77.82	**77.90**	78.55	**79.08**	77.53	68.84
	AVG	62.98	64.78	65.36	**64.18**	62.71	**65.13**	**65.83**	63.62
Xdomain_tr	20	**66.58**	67.08	70.14	75.87	65.19	**72.00**	**77.16**	**67.48**
	50	75.92	**77.50**	**78.85**	70.60	75.98	76.08	**79.84**	64.58
	200	81.61	**82.61**	83.04	**82.61**	81.43	80.92	83.06	82.76
	500	84.93	**85.09**	**84.92**	**83.88**	84.77	84.60	84.01	83.41
	2,000	87.39	**88.00**	**87.39**	77.48	**88.14**	87.73	86.91	66.70
	AVG	76.78	78.88	79.66	**77.58**	**77.59**	**79.75**	**80.80**	76.01

Table 4: F1-macro scores without and with MMR for three languages (rows), shown for five selected training sizes and an average (`AVG`) computed over 12 training sizes, across retrieved sample sizes of 20, 200, 2,000, and 20,000.

7 Conclusion

This paper presents a cross-lingual nearest neighbor retrieval approach to improve hate speech detection in target languages with limited labeled data. Our method retrieves the nearest neighbors from a multilingual pool of source tasks to augment the target language data, consistently outperforming models trained solely on the target language. Notably, with as few as 20 labeled instances in the target language, our approach can yield performance improvements of up to 10 F1-macro points in some cases. Further, we show that retrieving approximately 2,000 instances yields the highest average performance, while retrieving more can lead to a performance drop. Furthermore, the use of MMR to eliminate redundant data can yield additional performance gains in certain languages. Our method is scalable and adaptable to new languages and tasks, allowing new source tasks to be added to the pool with minimal effort.

Limitations

Despite the effectiveness of our approach, several limitations remain. First, we assume access to a small number of labeled hate speech instances in the target language. While this assumption reduces annotation cost, it may not hold in extremely low-resource settings where even minimal labeled data is unavailable or difficult to obtain due to linguistic, political, or ethical constraints.

Second, the retrieval pool used in our experiments is heavily imbalanced, with English accounting for the majority of instances. This dominance can bias retrieval and limit performance improvements for target languages that are typologically distant or culturally distinct from English. Expanding the set of target labels and tasks, especially in non-Western languages and underrepresented communities, would help assess the robustness and generalizability of the proposed method. Our evaluation focuses on a subset of hate speech detection tasks and languages and does not encompass the full variety of online abuse domains or contexts in which hate speech occurs.

Finally, while we reviewed the definitions of hate speech used in the datasets for our experiments (see Table 5 in Appendix), cultural differences and annotation inconsistencies may still be present. Although hate speech is undoubtedly influenced by cultural context, many hateful expressions are universal across languages and cultures. Our experiments demonstrate that leveraging such cross-lingual data can effectively improve hate speech detection in low-resource settings.

Acknowledgements

The work was supported by the European Research Council (ERC) through the European Union's Horizon Europe research and innovation programme (grant agreement No. 101113091) and the German Research Foundation (DFG; grant FR 2829/7-1).

References

Maryam M. AlEmadi and Wajdi Zaghouani. 2024. Emotional toll and coping strategies: Navigating the effects of annotating hate speech data. In *Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies @ LREC-COLING 2024*, pages 66–72. ELRA and ICCL.

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020a. A deep dive into multilingual hate speech classification. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part V*, page 423–439. Springer International Publishing.

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020b. Deep learning models for multilingual hate speech detection. In *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2020*, volume 12461 of *Lecture Notes in Computer Science*, pages 528–544. Springer.

Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust hate speech detection in social media: A cross-dataset empirical evaluation. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 231–242. ACL.

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. ELRA.

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. ACL.

Djamila Romaissa Beddiar, Md Saroar Jahan, and Mourad Oussalah. 2021. Data expansion using back translation and paraphrasing for hate speech detection. *Online Social Networks and Media*, 24:100153.

Irina Bigoulaeva, Viktor Hangya, and Alexander Fraser. 2021. Cross-lingual transfer learning for hate speech detection. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 15–25. ACL.

Irina Bigoulaeva, Viktor Hangya, Iryna Gurevych, and Alexander Fraser. 2022. Addressing the challenges of cross-lingual hate speech detection. *arXiv preprint arXiv:2201.05922*.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 335–336. ACM.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202. ELRA.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of ACL 2024*, pages 2318–2335. ACL.

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 ACL*, pages 8440–8451. ACL.

Devleena Das and Vivek Khetan. 2024. DEFT-UCS: Data efficient fine-tuning for pre-trained language models via unsupervised core-set selection for text-editing. In *Proceedings of the 2024 Conference on EMNLP*, pages 20296–20312. ACL.

Daryna Dementieva, Valeriia Khylenko, Nikolay Babakov, and Georg Groh. 2024. Toxicity classification in Ukrainian. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 244–255. ACL.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on EMNLP*, pages 345–363. ACL.

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

Anderson Almeida Firmino, Cláudio de Souza Baptista, and Anselmo Cardoso de Paiva. 2024. Improving hate speech detection using cross-lingual learning. *Expert Syst. Appl.*, 235(C).

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4).

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104. ACL.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Hassan Gharoun, Fereshteh Momenifar, Fang Chen, and Amir H. Gandomi. 2024. Meta-learning approaches for few-shot learning: A survey of recent advances. *ACM Comput. Surv.*, 56(12).

Faeze Ghorbanpour, Viktor Hangya, and Alexander Fraser. 2025. Fine-grained transfer learning for harmful content detection through label-specific soft prompt tuning. In *Proceedings of the 2025 Conference of NAACL: Human Language Technologies (Volume 1: Long Papers)*, pages 11047–11061. ACL.

Janis Goldzycher, Moritz Preisig, Chantal Amrhein, and Gerold Schneider. 2023. Evaluating the effectiveness of natural language inference for hate speech detection in languages with limited labeled data. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 187–201. ACL.

Janis Goldzycher, Paul Röttger, and Gerold Schneider. 2024. Improving adversarial data collection by supporting annotators: Lessons from GAHD, a German hate speech dataset. In *Proceedings of the 2024 Conference of theNAACL: Human Language Technologies (Volume 1: Long Papers)*, pages 4405–4424. ACL.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Xiaochuang Han and Yulia Tsvetkov. 2022. Orca: Interpreting prompted language models via locating supporting data evidence in the ocean of pretraining data. *arXiv preprint arXiv:2205.12600*.

Ehtesham Hashmi, Sule Yildirim Yayilgan, and Mohamed Abomhara. 2025. Metalinguist: Enhancing hate speech detection with cross-lingual meta-learning. *Complex & Intelligent Systems*, 11(1):179.

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*.

Hamish Ivison, Noah A. Smith, Hannaneh Hajishirzi, and Pradeep Dasigi. 2023. Data-efficient finetuning using cross-task nearest neighbors. In *Findings of ACL 2023*, pages 9036–9061. ACL.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020a. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th ACL*, pages 5435–5442. ACL.

Chris J. Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020b. Constructing interval

variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Bill Yuchen Lin, Kangmin Tan, Chris Miller, Beiwen Tian, and Xiang Ren. 2022. Unsupervised cross-task generalization via retrieval augmentation. In *Advances in Neural Information Processing Systems*, volume 35.

Yu A. Malkov and D. A. Yashunin. 2020. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Prasenjit Majumder, Johannes Schäfer, Tharindu Ranasinghe, Marcos Zampieri, Durgesh Nandini, and Amit Kumar Jaiswal. 2021. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages. In *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation*, volume 3159, pages 1–19. CEUR Workshop Proceedings.

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.

Ayme Arango Monnar, Jorge Pérez Rojas, and Barbara Poblete. 2024. Cross-lingual hate speech detection using domain-specific word embeddings. *PLOS ONE*, 19(7):e0306521.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2022. Cross-lingual few-shot hate speech and offensive language detection using meta learning. *IEEE Access*, 10:14880–14896.

Masayasu Muraoka, Bishwaranjan Bhattacharjee, Michele Merler, Graeme Blackwood, Yulong Li, and Yang Zhao. 2023. Cross-lingual transfer of large language model by visually-derived supervision toward low-resource languages. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 3637–3646. ACM.

Rachna Narula and Poonam Chaudhary. 2024. A comprehensive review on detection of hate speech for multi-lingual data. *Social Network Analysis and Mining*, 14(244).

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on EMNLP and the 9th IJCNLP*, pages 4675–4684. ACL.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2021. A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. *Information Processing and Management*, 58(4).

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th ACL: Student Research Workshop*, pages 363–370. ACL.

Marinela Parovic, Alan Ansell, Ivan Vulić, and Anna Korhonen. 2023. Cross-lingual transfer with target language-ready task adapters. In *Findings of ACL 2023*, pages 176–193. ACL.

Andraž Pelicon, Ravi Shekhar, Blaž Škrlj, Matthew Purver, and Senja Pollak. 2021. Investigating cross-lingual training for offensive language detection. *PeerJ Computer Science*, 7:e559.

Trinh Pham, Khoi Le, and Anh Tuan Luu. 2024. UniBridge: A unified approach to cross-lingual transfer learning for low-resource languages. In *Proceedings of the 62nd ACL (Volume 1: Long Papers)*, pages 3168–3184. ACL.

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(2):477–523.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, pages 19920–19930.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on EMNLP*, pages 4512–4525. ACL.

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 EMNLP*, pages 5674–5691. ACL.

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 ACL and the 11th IJCNLP (Volume 1: Long Papers)*, pages 41–58. ACL.

Sumegh Roychowdhury and Vikram Gupta. 2023. Data-efficient methods for improving hate speech detection. In *Findings of EACL 2023*, pages 125–132. ACL.

Manuela Sanguinetti, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. 2020. Haspeede 2 @ evalita2020: Overview of the evalita 2020 hate speech detection task. In *Proceedings of the EVALITA 2020 Workshop*, volume 2765. CEUR Workshop Proceedings.

Weijia Shi, Julian Michael, Suchin Gururangan, and Luke Zettlemoyer. 2022. Nearest neighbor zero-shot inference. In *Proceedings of the 2022 Conference on EMNLP*, pages 3254–3265. ACL.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225. ELRA.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021a. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of NAACL: Human Language Technologies*, pages 2289–2303. ACL.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021b. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th ACL and the 11th IJCNLP (Volume 1: Long Papers)*, pages 1667–1682. ACL.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of NAACL: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608. ACL.

Richard Ashby Wilson. 2019. The digital ethnography of law: Studying online hate speech online and offline. *Journal of Legal Anthropology*, 3(1):1–20.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on EMNLP: System Demonstrations*, pages 38–45. ACL.

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

Li Zhou, Laura Cabello, Yong Cao, and Daniel Hershcovich. 2023. Cross-cultural transfer learning for Chinese offensive language detection. In *Proceedings of C3NLP Workshop*, pages 8–15. ACL.

A Datasets Details

We used fourteen datasets in our study. Detailed information—including language, number of instances, license type—is provided in Table 5. In total, we had 265,671 instances, of which 62.85% were Non-Hate and 37.15% were Hate Speech.

The column *Size* reports the total number of instances in each dataset. *Pool Share* indicates the proportion of the final pool contributed by the dataset. All datasets are binary, containing the classes *hate* and *non-hate*. The column *Hate (%)* specifies the relative size of the hate-speech class with respect to the dataset size. The column *Hate Speech Definition* provides the exact definition of hate speech as stated in the original paper or annotation guidelines. A review of these definitions shows that all datasets adopt a consistent, unified definition of hate speech. The *License* column specifies the usage terms, with all datasets being permitted for research purposes.

B Model and Training Details

B.1 Embedder

For the embedding model, we used `BAAI/bge-m3`,[5,6] accessed via the Sentence Transformers library.[7] This model supports over 100 languages, is effective for both short and long text retrieval, and produces 1024-dimensional embeddings. It is released under the MIT license, and the Sentence Transformers library is licensed under Apache 2.0—both allowing use in academic research. We used the model in inference mode without any fine-tuning, applying it to our text data to generate embedding vectors.

B.2 Retriever

For indexing and searching the embedding vectors in the retrieval pool, we used the `Faiss` library[8], which is licensed under MIT. We employed the HNSW (Hierarchical Navigable Small World) index with Euclidean distance as the similarity metric, where smaller values indicate greater similarity to the query. Since the size of the retrieval pool was moderate, we used the CPU version of the library. The index was configured with 128 neighbors, a

[5] https://huggingface.co/BAAI/bge-m3

[6] https://github.com/FlagOpen/FlagEmbedding

[7] https://github.com/UKPLab/sentence-transformers

[8] https://github.com/facebookresearch/faiss

Dataset	Hate Speech Definition	Language	Lang Code	Size	Pool Share	Num Classes	Hate (%)	License
Bas19_es (Basile et al., 2019)	Any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics.	Spanish	es	6,600	2.48	2	41.50	CC BY 4.0
For19_pt (Fortuna et al., 2019)	Language that attacks or diminishes and incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic, sexual orientation, gender identity or other.	Portuguese	pt	5,670	2.13	2	31.53	CC BY 4.0
Has21_hi (Mandl et al., 2021)	Ascribing negative attributes or deficiencies to groups of individuals because they are members of a group (e.g. "all poor people are stupid").	Hindi	hi	4,594	1.73	2	12.32	CC BY 4.0 (Only for research purposes.)
Ous19_ar (Ousidhoum et al., 2019)	Hate speech may not represent the general opinion, yet it promotes the dehumanization of people who are typically from minority groups and can incite hate crimes.	Arabic	ar	3,353	1.26	2	22.52	MIT
Ous19_fr (Ousidhoum et al., 2019)	Same as above	French	fr	4,014	1.51	2	0.94	MIT
San20_it (Sanguinetti et al., 2020)	Hateful content in the text towards a given target (among immigrants, Muslims, and Roma).	Italian	it	8,100	3.05	2	41.83	CC BY-NC-SA 4.0
Gahd24_de (Goldzycher et al., 2024)	Abusive, discriminatory, derogatory, or dehumanizing speech targeting a protected group or a person for being a member of such a group.	German	de	10,996	3.84	2	42.37	CC BY 4.0
Xdomain_tr (Toraman et al., 2022)	Tweets contain hate speech if they target, incite violence against, threaten, or call for physical damage to an individual or a group of people because of some identifying trait or characteristic.	Turkish	tr	37,933	17.0	2	42.67	CC BY-NC-SA 4.0
Xdomain_en (Toraman et al., 2022)	Same as above	English	en	47124	21.12	2	19.41	CC BY-NC-SA 4.0
Ken20_en (Kennedy et al., 2020a)	Posts that either contain human degradation or calls for violence toward some target group, which is often a protected group or a human group identified by some characteristic (e.g., race, gender, religion, etc.), or "group identifier" terms.	English	en	23,192	8.73	2	50.00	MIT
Fou18_en (Founta et al., 2018)	Content that is derogatory, humiliating, or insulting towards a target.	English	en	22,565	8.49	2	22.00	CC BY 4.0
Xplain_en (Mathew et al., 2021)	Language that explicitly attacks or demeans a group of people based on race, religion, gender, sexual orientation, or other protected characteristics.	English	en	13,749	5.08	2	43.22	MIT
Implicit_en (ElSherief et al., 2021)	Language that targets protected groups or individuals (e.g. based on race, gender, religion, sexual orientation, cultural identity) with disparagement or harm, and can be explicit (with direct keywords) or implicit, where implicit hate uses coded/indirect language (sarcasm, metaphor, etc.) to convey prejudiced or harmful views.	English	en	21,480	8.09	2	38.12	MIT
Dyn21_en (Vidgen et al., 2021b)	Abusive speech targeting specific group characteristics, such as ethnic origin, religion, gender, or sexual orientation.	English	en	41,144	15.49	2	46.10	CC BY 4.0

Table 5: Detailed information about the datasets used in this study.

construction parameter of 200, and a search parameter of 128.

B.3 Fine-tuner

We used `twitter-xlm-roberta-base`[9] (XLM-T), a multilingual transformer-based model pre-trained on Twitter data, as our base model for fine-tuning on hate speech detection tasks. The model is licensed under Apache 2.0, which permits use in academic research. Training and evaluation were performed using the Hugging Face `transformers` library.[10]

To select the classification model, we also evaluated `xlm-roberta-base`[11] (XLM-R) and `mdeberta-v3-base`[12] (He et al., 2021). XLM-T outperformed XLM-R, likely because hate speech datasets are primarily sourced from social media, where XLM-T has been pre-trained. Although XLM-T and mDeBERTa achieved similar performance, we chose to use XLM-T in our experiments, as it supports a broader range of languages and aligns with our baseline setting.

Throughout all our experiments, for training sizes with fewer than 9,999 training instances, we trained for 10 epochs; for larger training sizes, we used 5 epochs to reduce training time and avoid overfitting. We set the batch size to 16 and used a learning rate of 5e-5. Inputs were truncated or padded to a maximum sequence length of 128 tokens. We used binary cross-entropy loss, as our datasets involved binary classification (Hate vs. Non-Hate). All other training hyperparameters were left at their default values provided by the `transformers.Trainer` module.

C Hardware and Tools

The experiments were conducted on NVIDIA GeForce GTX 1080 Ti servers. The embedding model was used in inference mode without updating its parameters, while the classification model was fully fine-tuned. As the classifier was based on `xlm-roberta-base`, it included approximately 279 million parameters. We also acknowledge the use of an AI assistant during the writing process. ChatGPT[13] was used for paraphrasing and improving clarity throughout the formulation of the paper. All models and datasets used in this study are licensed for academic research purposes and align with the intended use of advancing NLP applications for social good.

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

[10] https://github.com/huggingface/transformers

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

[12] https://huggingface.co/microsoft/mdeberta-v3-base

[13] https://chatgpt.com/

D Full Main Results

Table 6 presents the full results of cross-lingual nearest neighbor retrieval fine-tuning, covering training sizes from 10 to 2,000. These results confirm the trend shown in the main table (Table 1): retrieving as few as 20 instances already outperforms the `Mono` baseline in all languages. These performance improvements are most notable when the target language has fewer than 50 training examples, with F1-macro gains exceeding 10 compared to training solely on the target language data. In such low-resource settings, retrieving cross-lingual nearest neighbors and using them for fine-tuning enables the model to match the performance of much larger training sizes. In most cases, our method also outperforms the `Röttger` approach, while using less data and without requiring manual selection of a source task.

Additionally, the underlined values—indicating the smallest amount of retrieved data that outperforms the next larger training size—demonstrate that, in most cases and languages, retrieving cross-lingual data can effectively compensate for having less labeled data. While in two tasks, *Bas19_es* and *Xdomain_tr*, training data appears to be of higher quality and retrieval cannot fully offset its absence, in all other tasks, retrieval proves effective, especially valuable when labeled data is scarce, less than 50-highlighting the method's strength in very low-resource hate speech detection settings.

E Additional Experiments

To evaluate the robustness of our approach, we conducted several additional experiments beyond the main setup. These were carried out on only two languages and with fewer training epochs than the default, with the goal of identifying the most effective configuration before running the full set of experiments.

Alternative Embedders and Retrievers Table 7 reports results using different embedder–retriever settings (M3 (Chen et al., 2024) with Euclidean distance, M3 with cosine similarity, and LaBSE (Feng et al., 2022) with Euclidean distance). The

	San20_it						Ous19_ar						Ous19_fr					
SIZE	Mono	20	200	2000	20000	Röttger	Mono	20	200	2000	20000	Röttger	Mono	20	200	2000	20000	Röttger
10	46.23	47.04	64.96	**65.26**	69.60	63.34	51.98	54.60	56.36	**62.86**	61.54	59.00	47.26	48.34	52.46	51.90	**54.60**	53.49
20	54.25	63.20	66.76	**67.06**	60.06	64.96	51.67	57.63	**63.23**	61.73	59.47	60.52	47.26	47.21	52.68	53.93	**55.05**	52.93
30	56.29	64.24	68.47	68.90	**70.08**	64.95	44.42	57.95	**66.23**	62.31	62.70	65.48	47.26	47.60	52.82	53.86	**54.97**	53.91
40	59.14	59.60	67.91	69.16	**70.58**	67.41	49.67	57.45	**65.70**	64.52	64.82	64.29	47.24	48.36	52.40	54.63	56.30	**56.62**
50	65.71	67.20	68.42	71.65	**69.44**	69.10	52.13	59.36	**66.65**	66.31	64.51	65.76	47.87	48.29	52.19	52.97	**55.60**	54.15
100	70.96	70.01	71.17	71.46	67.73	**71.89**	65.29	66.22	**68.50**	66.77	66.07	66.44	47.67	49.35	51.54	53.12	**56.04**	55.83
200	72.81	72.46	**72.83**	72.41	72.50	71.56	67.97	67.98	**69.18**	67.35	65.47	66.61	51.93	51.54	54.06	**55.80**	53.63	53.76
300	73.43	73.56	**74.12**	72.03	64.56	72.21	66.95	68.94	68.52	**69.00**	66.71	68.07	51.10	53.49	53.29	54.58	**56.15**	53.61
400	72.44	73.49	**74.84**	66.23	66.80	73.32	67.04	69.75	**70.20**	68.88	68.20	66.79	52.17	53.22	54.57	53.81	**55.13**	53.34
500	74.18	75.39	**75.29**	74.53	66.09	73.69	66.54	68.95	**69.47**	69.28	65.54	67.60	51.91	53.30	52.84	**55.51**	55.31	53.39
1000	76.56	**76.65**	68.84	76.41	68.21	76.14	67.29	68.98	68.08	67.77	**69.07**	67.26	53.14	52.49	**55.48**	55.15	50.51	52.59
2000	76.40	69.27	**78.36**	77.57	76.95	77.07	66.91	69.52	69.77	**70.15**	68.27	67.07	51.84	53.51	53.13	53.30	**54.74**	52.89
AVG	66.53	67.68	71.00	**71.06**	68.55	70.47	59.82	63.94	**66.82**	66.41	65.20	65.41	49.72	50.56	53.12	54.05	**54.84**	53.88
	Bas19_es						For19_pt						Xdomain_tr					
10	36.51	46.06	58.18	**62.23**	61.97	59.71	43.18	48.03	61.67	67.53	**67.79**	66.38	46.92	56.19	73.20	**77.11**	72.50	72.50
20	49.91	54.37	59.72	62.52	63.08	**66.52**	48.09	49.72	64.92	**68.57**	68.03	67.68	55.43	66.58	67.08	70.14	**75.87**	69.80
30	54.38	60.75	62.56	64.62	63.52	**69.02**	51.73	54.34	64.61	**69.70**	67.80	67.78	58.90	73.13	73.64	**78.37**	75.58	75.58
40	57.63	57.69	62.77	63.10	61.87	**66.98**	53.83	52.63	65.59	**69.68**	67.57	67.04	69.97	74.20	75.43	61.71	**78.70**	75.98
50	61.85	60.93	64.37	65.59	64.30	**70.36**	60.25	59.26	67.01	67.06	**69.35**	66.51	72.24	75.92	77.50	**78.85**	70.60	75.12
100	65.03	65.36	66.04	67.71	65.93	**71.94**	64.38	67.81	68.26	**69.41**	68.99	68.95	71.43	79.84	78.77	80.79	80.48	**81.41**
200	72.36	72.22	71.77	71.23	70.67	**75.27**	66.91	69.69	70.33	70.20	**71.07**	68.10	81.63	81.61	82.61	**83.04**	82.61	82.19
300	74.40	76.04	76.18	75.31	71.92	**76.43**	68.86	69.37	69.97	69.86	**70.14**	68.63	81.34	81.80	83.74	83.27	83.39	**84.36**
400	76.58	75.77	76.06	76.30	73.84	**77.57**	69.10	69.74	70.19	69.96	**70.80**	67.92	84.54	83.08	84.53	84.27	63.18	**85.24**
500	77.14	78.01	77.09	77.79	67.67	**78.76**	69.95	69.72	70.84	70.04	**71.05**	69.22	85.05	84.93	85.09	84.92	83.88	**85.34**
1000	79.35	79.42	79.16	79.27	78.53	**81.06**	70.97	71.48	**72.07**	70.81	71.44	70.92	87.01	76.68	77.00	86.03	86.67	**86.86**
2000	81.08	80.62	80.50	80.65	81.02	**82.04**	72.70	72.39	**72.66**	71.72	72.22	71.61	88.53	87.39	88.00	87.39	77.48	**88.84**
AVG	65.52	67.27	69.53	70.53	68.69	**72.97**	61.66	62.85	68.18	69.55	**69.69**	68.39	73.58	76.78	78.88	79.66	77.58	**80.27**
	Gahd24_de						Has21_hi											
10	38.03	51.17	54.50	**59.85**	57.59	58.88	46.87	49.00	51.76	53.06	**56.24**	54.46						
20	44.99	50.52	58.15	59.08	57.48	**59.82**	46.87	47.34	51.03	53.68	**55.37**	54.92						
30	50.20	53.14	59.37	**60.57**	59.08	59.78	46.87	46.87	52.31	54.21	55.46	**57.47**						
40	57.47	55.18	58.25	**60.86**	57.79	60.75	46.87	47.67	53.86	55.58	**56.33**	54.08						
50	57.85	54.53	60.30	60.47	61.02	**62.57**	46.87	48.39	53.36	52.26	**55.78**	54.77						
100	62.23	64.27	63.29	62.93	62.58	**64.50**	48.94	51.38	54.90	55.53	56.91	**57.88**						
200	65.80	66.95	66.15	**66.24**	65.97	64.25	52.20	55.83	54.65	56.80	56.02	**57.47**						
300	67.17	67.81	**67.33**	64.70	66.97	64.58	51.75	55.80	55.43	57.24	**58.05**	57.51						
400	66.82	69.04	68.47	68.43	**69.52**	66.74	54.77	54.50	55.98	**58.50**	57.06	58.12						
500	66.56	69.78	69.68	**70.45**	61.80	67.02	56.20	56.94	57.66	57.88	**59.55**	57.96						
1000	69.81	**74.17**	73.02	72.92	72.50	69.45	56.18	57.40	**60.06**	59.76	58.14	57.74						
2000	73.77	**79.19**	78.79	77.82	77.90	72.42	57.14	58.19	60.22	**60.50**	59.65	58.01						
AVG	60.06	62.98	64.77	**65.36**	64.18	64.23	50.96	52.44	55.10	56.25	**57.05**	56.70						

Table 6: F1-macro scores across eight languages, comparing our method with the Mono and Röttger baselines. Results are reported for all target training sizes from 10 to 2,000. Columns represent the number of retrieved instances. AVG denotes the mean over 12 training sizes. The best result for each language and training size is in **bold**. Retrieved results that outperform the next larger Mono training size are underlined.

overall performance trends remain stable, confirming that our method is not sensitive to the particular embedder or similarity metric.

Label Balancing and Target-Set Weighting Table 8 shows experiments with balancing labels in the retrieved pool and with upweighting the target training set by repeating it three times. Label balancing did not lead to systematic gains. In contrast, repeating the target training set three times gave small but consistent improvements over the default when we retrieve less than 200 data from the pool. This suggests that the target-language training set provides more informative learning signals than the retrieved data, and that upweighting it can be beneficial. Since repeating once is already effective and computationally simpler across eight languages, we adopted that setup as the main contribution. These results indicate that stronger upweighting can be beneficial and may be explored in future work.

F Further Analysis

To further analyze what is retrieved and used for fine-tuning under a controlled setting (retrieving 2,000 instances for a training size of 2,000), see Figure 4. This figure shows the distribution of retrieved source tasks, languages, and labels. As illustrated, English is retrieved the most, followed by Turkish and Spanish in nearly equal amounts, then Italian and Portuguese. Excluding English, this pattern roughly reflects the overall language distribution in the pool (see the "Pool Share" column in Table 5). The second to fifth most represented languages in the pool are Turkish, German, Italian, and Spanish. However, the low retrieval of German—despite its high presence—and the higher retrieval of Spanish over Italian are unexpected and may be attributed to task generality or cross-lingual similarity. The distribution of retrieved labels also mirrors their proportions in the pool: approximately 40% of instances are labeled as hate, and a similar pattern is observed in the retrieved hate instances across target tasks.

G More about English-only vs. Multilingual Retrieval

Additional results comparing English-only retrieval and multilingual retrieval are presented in Table 9. Comparisons in this table should be made vertically: for each language, the values for a specific training size and number of retrieved instances should be compared between the upper (English-only) and lower (multilingual) blocks. Similar to the datasets discussed in the main text, multilingual retrieval—with even a small number of non-English source tasks—proves beneficial for three languages in this table: *Ous19_ar*, *San20_it*, and *Gahd24_de*, in most cases. However, for *Xdomain_tr*, the results differ slightly; English-only retrieval performs marginally better, likely due to the generality of English tasks and their semantic similarity to the Turkish dataset.

H More about MMR

To compare the effect of using MMR versus not using it, refer to Table 10. In this table, since each language is presented as a row block, comparisons should be made horizontally within the two subtables. Specifically, for each language, values for the same training and retrieval size (e.g., 20 train, 20 retrieved) should be compared between the settings without and with MMR. As shown, for the remaining languages in this table, applying MMR does not lead to significant overall performance differences. This is likely because, even without MMR, the retrieved cross-lingual data is already sufficiently diverse, so applying MMR has minimal additional impact.

		M3 - Euclidean Distance				M3 - Cosine Similarity				LaBSE - Euclidean Distance			
	SIZE	20	200	2000	20000	20	200	2000	20000	20	200	2000	20000
Bas19_es	20	47.72	54.48	62.50	62.76	45.66	56.86	59.81	61.85	47.57	52.36	62.63	63.56
	50	44.73	60.80	66.04	63.41	45.38	59.95	63.13	62.54	48.30	55.20	64.28	64.36
	200	69.46	72.14	71.00	69.23	70.72	71.59	71.74	68.86	67.55	70.98	70.90	68.52
	500	77.02	77.86	77.86	75.55	78.71	77.88	77.00	68.97	76.67	77.45	77.69	76.03
	2000	81.55	81.04	81.21	81.16	80.74	81.47	81.73	72.16	81.40	80.54	80.78	80.86
	AVG	61.04	67.45	70.32	69.23	61.47	68.26	69.02	67.40	61.59	65.59	69.94	69.51
Ous19_ar	20	43.55	58.10	61.16	59.14	50.27	56.34	57.41	59.41	44.98	50.56	60.05	60.46
	50	47.13	59.23	62.77	63.79	43.68	61.87	63.71	62.61	45.59	57.47	64.62	64.76
	200	64.52	69.06	67.01	64.85	63.35	66.76	66.80	66.72	68.72	67.73	66.66	65.11
	500	69.16	69.61	68.51	67.79	68.12	67.50	68.04	69.05	69.44	69.41	68.61	67.69
	2000	68.41	68.90	67.68	70.31	69.45	69.26	67.45	68.63	69.79	70.58	68.85	69.91
	AVG	58.02	63.34	64.97	64.36	57.55	63.07	64.18	63.99	57.84	62.14	65.41	65.15

Table 7: Comparison of different embedder–retriever configurations (M3 with Euclidean distance, M3 with cosine similarity, and LaBSE with Euclidean distance). Results across Bas19-es and Out19-ar show no consistent improvements over the default setup, indicating robustness to the choice of backbone and similarity metric.

		Default				Balanced Labels				Repeated Target Training (3×)			
	SIZE	20	200	2000	20000	20	200	2000	20000	20	200	2000	20000
Bas19_es	20	47.72	54.48	62.50	62.76	42.07	56.32	60.20	62.76	54.29	59.55	62.45	61.82
	50	44.73	60.80	66.04	63.41	46.83	62.00	66.00	63.81	62.69	63.30	65.64	61.78
	200	69.46	72.14	71.00	69.23	69.20	72.18	72.41	70.89	72.79	72.59	74.09	71.97
	500	77.02	77.86	77.86	75.55	77.08	77.59	77.64	76.67	78.19	76.76	77.34	76.88
	2000	81.55	81.04	81.21	81.16	81.06	81.50	81.02	81.67	80.70	80.61	80.77	81.68
	AVG	61.04	67.45	70.32	69.23	61.94	68.62	69.99	69.36	67.11	68.29	70.61	69.51
Ous19_ar	20	43.55	58.10	61.16	59.14	51.47	58.77	58.31	58.13	47.61	58.63	58.73	58.14
	50	47.13	59.23	62.77	63.79	49.32	63.86	64.78	63.43	60.10	65.24	64.50	64.38
	200	64.52	69.06	67.01	64.85	68.56	69.33	66.93	65.83	66.30	64.93	65.96	65.87
	500	69.16	69.61	68.51	67.79	69.43	68.99	67.35	67.61	66.81	67.55	67.09	67.17
	2000	68.41	68.90	67.68	70.31	68.91	69.81	68.64	68.44	67.87	67.63	66.75	66.79
	AVG	58.02	63.34	64.97	64.36	58.94	65.34	64.28	63.57	61.25	63.41	64.31	64.50

Table 8: Impact of label balancing in the retrieved pool and re-weighting the target training set (by repeating it three times). The average results for label balancing remain close to the default setup, whereas repeating the target training data yields slight average improvements.

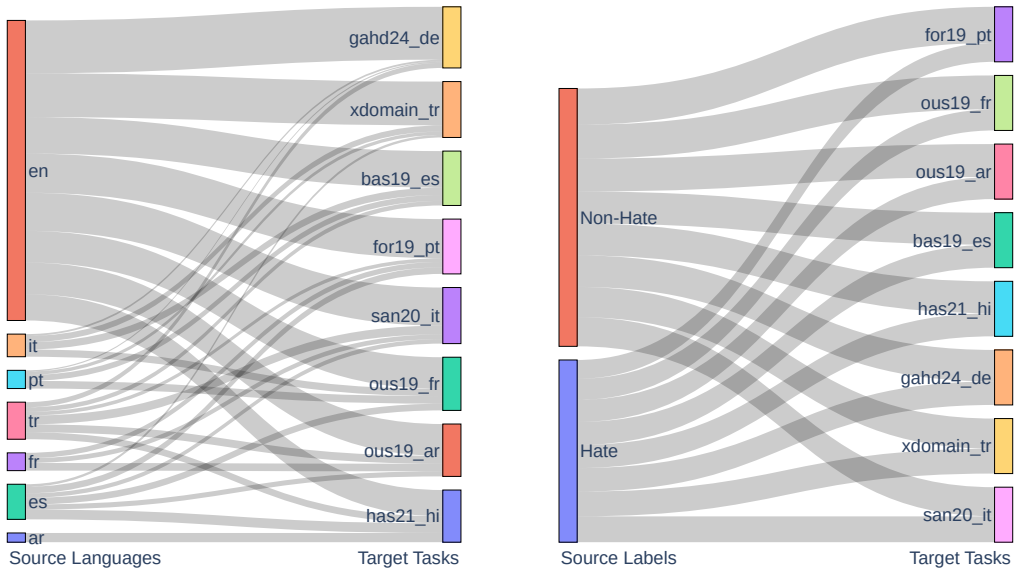

Figure 4: Sankey diagrams of the distribution of the top four languages (left), and labels (right) for each target task.

		Ous19_ar				San20_it				Gahd24_de				Xdomain_tr			
	SIZE	20	200	2000	20000	20	200	2000	20000	20	200	2000	20000	20	200	2000	20000
English Retrieval	20	55.75	60.46	**62.22**	**62.22**	58.37	66.30	67.01	**62.23**	50.30	55.72	**59.79**	54.79	62.39	**71.47**	**76.53**	74.01
	50	**64.05**	65.86	66.34	55.91	**68.34**	68.53	62.49	56.88	**56.96**	60.02	58.73	56.62	**76.95**	77.91	73.85	57.79
	200	67.67	68.11	66.86	**67.22**	71.48	**73.41**	71.48	69.73	66.88	66.37	65.66	63.72	81.84	82.17	83.61	**83.44**
	500	**69.38**	68.88	**67.96**	66.81	75.19	74.92	67.32	**73.54**	**70.06**	69.81	63.28	**62.12**	**85.91**	85.47	84.50	75.03
	2000	**70.95**	71.32	69.26	**70.21**	77.05	77.75	68.58	**77.69**	78.93	78.74	77.93	77.58	87.81	87.67	87.55	**87.26**
	AVG	**64.27**	66.46	66.15	65.06	**67.74**	**71.14**	67.91	66.37	62.81	64.45	64.79	62.93	**77.46**	**80.27**	**80.90**	76.36
Multilingual Retrieval	20	**57.63**	**63.23**	61.73	59.47	**63.20**	66.76	67.06	60.06	50.52	**58.15**	59.08	**57.48**	**66.58**	67.08	70.14	**75.87**
	50	59.36	**66.65**	66.31	**64.51**	67.20	68.42	**71.65**	**69.44**	54.53	60.30	**60.47**	**61.02**	75.92	77.50	**78.85**	**70.60**
	200	67.98	**69.18**	67.35	65.47	**72.46**	72.83	**72.41**	**72.50**	66.95	66.15	**66.24**	**65.97**	81.61	82.61	83.04	82.61
	500	68.95	**69.47**	69.28	65.54	**75.39**	**75.29**	**74.53**	66.09	69.78	69.68	**70.45**	61.80	84.93	85.09	84.92	83.88
	2000	69.52	69.77	**70.15**	**68.27**	**69.27**	**78.36**	**77.57**	76.95	**79.19**	**78.79**	77.82	77.90	87.39	**88.00**	87.39	**77.48**
	AVG	63.94	**66.82**	**66.41**	**65.20**	67.68	71.00	**71.06**	**68.55**	**62.98**	**64.77**	**65.36**	**64.18**	76.78	78.88	79.66	**77.58**

Table 9: F1-macro scores across two strategies: English-only retrieval, and Multilingual retrieval. Results are shown for target training sizes of 20, 50, 200, 500, 2,000, and AVG (the average over 12 training sizes). Columns represent target languages, and sub-columns are the number of retrieved instances.

		Without MMR				With MMR				Without MMR				With MMR			
	SIZE	20	200	2000	20000	20	200	2000	20000	20	200	2000	20000	20	200	2000	20000
Bas19_es	20	54.37	59.72	62.52	63.08	54.20	**60.15**	62.73	63.18	47.21	52.68	53.93	55.05	47.25	52.58	**54.31**	55.13
	50	**60.93**	64.37	**65.59**	**64.30**	59.84	64.85	64.67	62.53	48.29	52.19	52.97	**55.60**	48.67	52.15	**54.99**	54.82
	200	**72.22**	71.77	71.23	**70.67**	71.84	**72.29**	71.28	68.80	51.54	**54.06**	55.80	53.63	**53.70**	53.79	55.06	**54.43**
	500	78.01	**77.09**	**77.79**	67.67	77.97	75.72	76.64	**74.06**	**53.30**	52.84	**55.51**	**55.31**	51.84	**53.13**	54.15	53.01
	2000	80.62	80.50	80.65	**81.02**	80.31	**81.23**	**81.12**	80.58	**53.51**	**53.13**	**53.30**	**54.74**	52.07	51.44	52.79	53.19
	AVG	67.27	69.53	70.53	68.69	67.26	**69.70**	70.01	68.28	50.56	53.12	54.05	54.84	50.43	52.99	54.22	54.54
Ous19_ar	20	57.63	**63.23**	**61.73**	59.47	57.75	59.96	60.50	**62.33**	49.72	64.92	68.57	**68.03**	**54.31**	**65.48**	68.95	67.63
	50	59.36	66.65	**66.31**	64.51	59.66	66.19	64.98	64.46	59.26	67.01	67.06	**69.35**	**61.81**	67.52	67.81	67.38
	200	67.98	69.18	**67.35**	**65.47**	67.61	69.52	66.35	64.55	**69.69**	**70.33**	**70.20**	**71.07**	68.29	69.83	69.48	69.74
	500	68.95	69.47	69.28	65.54	68.70	**70.39**	69.59	**67.17**	69.72	70.84	**70.04**	71.05	69.46	70.21	69.91	71.83
	2000	69.52	69.77	**70.15**	68.27	**70.20**	**70.12**	69.93	68.19	72.39	**72.66**	71.72	72.22	72.24	70.85	71.93	72.48
	AVG	63.94	66.82	**66.41**	65.20	**64.16**	66.96	65.81	64.94	62.85	68.18	69.55	69.69	**63.60**	68.28	69.28	69.75
Has21_hi	20	47.34	51.03	53.68	55.37	47.42	**52.96**	53.17	56.03								
	50	**48.39**	53.36	52.26	55.78	47.34	**54.19**	**53.55**	**56.29**								
	200	55.83	54.65	56.80	56.02	55.57	**56.11**	56.67	**57.67**								
	500	56.94	57.66	57.88	**59.55**	56.78	**58.43**	**59.28**	57.55								
	2000	**58.19**	60.22	60.50	59.65	57.37	60.02	60.05	**61.10**								
	AVG	52.44	55.10	56.25	57.05	52.38	55.22	56.08	57.37								

Table 10: F1-macro scores without/with MMR for five languages (rows), across five selected training sizes and an average (AVG) computed over 12 training sizes.