KIT-TIP-NLP at MultiPride: Continual Learning with Multilingual Foundation Model*

Barathi Ganesh HB, Michal Ptaszynski, Rene Melendez and Juuso Eronen

Text Information Processing Lab, Kitami Institute of Technology, Kitami, Hokkaido 090-0015, Japan

Abstract

This paper presents a multi-stage framework for detecting reclaimed slurs in multilingual social media discourse. It addresses the challenge of identifying reclamatory versus non-reclamatory usage of LGBTQ+-related slurs across English, Spanish, and Italian tweets. The framework handles three intertwined methodological challenges like data scarcity, class imbalance, and cross-linguistic variation in sentiment expression. It integrates data-driven model selection via cross-validation, semantic-preserving augmentation through back-translation, inductive transfer learning with dynamic epoch-level undersampling, and domain-specific knowledge injection via masked language modeling. Eight multilingual embedding models were evaluated systematically, with XLM-RoBERTa selected as the foundation model based on macro-averaged F1 score. Data augmentation via GPT-4o-mini back-translation to alternate languages effectively tripled the training corpus while preserving semantic content and class distribution ratios. The framework produces four final runs for the evaluation purposes where RUN 1 is inductive transfer learning with augmentation and undersampling, RUN 2 with masked language modeling pre-training, RUN 3 and RUN 4 are previous predictions refined via language-specific decision thresholds optimized via ROC analysis. Language-specific threshold refinement reveals that optimal decision boundaries vary significantly across languages. This reflects distributional differences in model confidence scores and linguistic variation in reclamatory language usage. The threshold-based optimization yields 2–5% absolute F1 improvement without requiring model retraining. The methodology is fully reproducible, with all code and experimental setup are available at https://github.com/rbg-research/MultiPRIDE-Evalita-2026.

Keywords

Reclamation Detection, Multilingual Foundation Models, Transfer Learning, Dynamic Undersampling, Back-Translation Augmentation, Language-Specific Decision Boundaries, LGBTQ+ Sentiment Analysis

1. Introduction

Automated content moderation systems often have difficulty in identifying hate speech and reclaimed language [1, 2]. Reclaimed language refers to historically derogatory language that has been reclaimed by marginalized groups, especially the LGBTQ+ community who reclaimed the language for self-expression and solidarity [3, 4]. If the algorithm cannot identify the context, it will either mark harmless content as toxic or miss actual hate speech. This paper will present the KIT-TIP-NLP system designed for the MultiPride task of EVALITA 2026, which aims to identify reclaimed language in English, Spanish, and Italian tweets [4].

The task presents three main obstacles. Firstly, the available data is limited. Secondly, the datasets are highly imbalanced, with far fewer examples of reclaimed usage than non reclaimed usage. Thirdly, the way speakers reclaim slurs varies by language and culture. A model that works well for English sarcasm may fail to grasp Italian cultural markers.

In order to tackle these problems, we employed a multistage framework. The first step was to assess a variety of multilingual models and proceed with XLM-RoBERTa as our basis. To tackle the problem of insufficient data, we implemented the GPT-4o-mini-based back-translation method that resulted in the training set being tripled while the original meaning was preserved. Additionally, we made use of a proactive sampling method during the training to ensure that the model does not turn a blind eye to

EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26 – 27, Bari, IT

*This paper contains examples of explicitly offensive content.

hbbg.jp@gmail.com (B. G. HB); michal@mail.kitami-it.ac.jp (M. Ptaszynski)

0000-0002-1150-2773 (B. G. HB); 0000-0002-1910-9183 (M. Ptaszynski); 0009-0004-2129-8747 (R. Melendez); 0000-0001-9841-3652 (J. Eronen)

the minority class. To improve performance further, we assigned language-specific decision thresholds instead of applying one universal cutoff for all languages. This method assumes that the model has a higher level of confidence in certain languages and hence provides a way to get better accuracy without cumbersome retraining.

The rest of the paper is structured in the following way. Section 2 describes the methodology with a focus on data analysis, model selection, and the training process. Section 3 describes the experimental setup and the performance of the four runs. Section 4 discusses the results and the challenges faced in the different languages. Finally, section 5 summarizes the findings.

2. Methodology

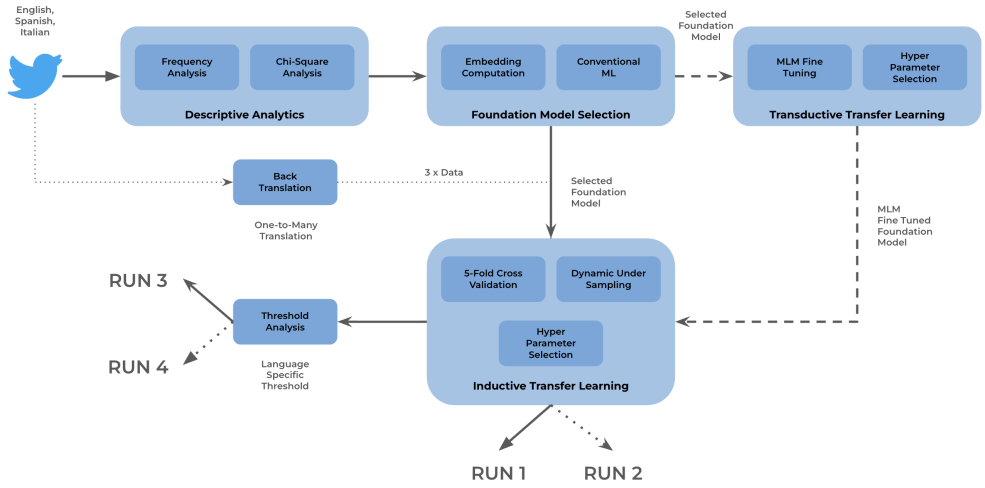


Figure 1: Multi-stage multilingual hate-speech classification framework with four sequential runs refining performance via data-driven model selection, augmentation, hyperparameter selection, 5-fold CV, MLM adaptation, and threshold calibration. **RUN 1:** inductive transfer learning with optimal foundation model. **RUN 2:** transductive transfer learning on optimal foundation model follwed by inductive transfer learning. **RUN 3:** threshold refinement on RUN 1 outputs and **RUN 4:** threshold refinement on RUN 2 outputs to handle linguistic nuances and cross-lingual calibration.

The overall framework, as represented in Figure 1, consists of four research runs carried out one after another which are able to refine classification performance in a step-wise manner. The multi-layered approach goes back to four design principles. First, the framework follows the data-driven model selection through systematic cross-validation as its primary consideration which eliminates the possibility of selection bias and guarantees that the chosen foundation model comes with empirical performance. Second, it is based on the scarcity of data that is through semantic-preserving augmentation, thus increasing the effective training corpus while keeping the distribution of labels the same using the method of back-translation. Thirdly, the domain-specific knowledge is integrated by masked language modeling (MLM) for the best representation. Finally, language-specific threshold refinement is done for managing the linguistic nuance and the distributional variation in different languages. The whole process from RUN 1 to RUN 4 provides the opportunity for systematic ablation and comparison of these design choices.

2.1. Data Analysis and Bias Detection

The framework commences by performing descriptive analytics on the dataset [1] to characterize it and recognize possible imbalances regarding language and labels. A frequency distribution analysis was done for all labels and languages to check the positive and negative instance distribution. After that,

[1]https://multipride-evalita.github.io/, Accessed on January 2026.

chi-square analysis was conducted to discover the statistical relationships between language and label distributions [5]. This exposes the label and language biases naturally existing in the dataset supplied for the modeling choices made afterwards.

2.2. Foundation Model Selection

This phase employed a two-stage approach. First, embedding representations were computed for all the tweets with each foundation models for assessing the capability of representating semantic and linguistic properties of the input tweet. A conventional machine learning (ML) pipeline was established in parallel to provide a comparative reference. Five-fold cross-validation was conducted across both embedding-based and conventional ML approaches to ensure robust and unbiased model selection. The cross-validation framework evaluated across multiple models and selected the optimal foundation model based on performance metrics computed across all folds for avoiding the selection bias.

2.3. Inductive Transfer Learning with Data Augmentation

The chosen foundational model went through an inductive transfer learning approach [6] for fine-tuning. A back-translation augmentation method was executed to deal with data deficiency with GPT-4o-mini being used as the translation engine [7]. Thru the supplied dataset, each tweet in the original language was assessed and translated to the other two languages in a systematic way. This allows for the generation of semantically equal, yet syntactically diverse paraphrases. This one-to-many translation operation has effectively increased the size of the training set threefold. The augmented dataset was then used to refine the base model further [8].

In order to address the issue of class imbalance in the classification task, a dynamic under-sampling strategy was used at the epoch level [9]. The sampling technique during each training epoch allowed every positive class instance in the training batch to have three negative class instances selected randomly without replacement. This ratio of 1:3 was maintained dynamically for all epochs, which helped the model learn distinguishing features irrespective of the class imbalance issue. For Each epoch, this sampling process was random, which introduced stochasticity and helped with the generalization of learning. The five-fold cross-validation process was used not only for the validation of the final accuracy of the model but also for hyper-parameter tuning. The final result obtained from this stage is labeled as RUN 1.

2.4. Domain Knowledge Integration via Masked Language Modeling

In the second experimental approach, the masked language modeling (MLM) was used on the foundation model with the aim of incorporating linguistic knowledge that is domain-specific. The MLM pretraining task is based on the task of predicting the randomly masked tokens in the sequences. The hyperparameter tuning was carried out using Optuna, which is an automated hyperparameter optimization library that uses efficient sampling and pruning algorithms. The final model that was adapted using MLM was again subjected to the same process of inductive fine-tuning as mentioned earlier in the section 2.3, which included back-translation augmentation, dynamic epoch-level under-sampling at the 1:3 positive to negative ratio, and five-fold cross-validation for fitness evaluation. The final result obtained from this phase is labeled as RUN 2.

2.5. Language-Specific Threshold Refinement

To further improve classification performance, language specific decision boundaries were derived for each language in the dataset. Confidence scores from the classifier were analyzed separately for English, Spanish, and Italian tweets. Language-specific thresholds were determined that optimally balanced precision and recall for each language through the ground truth labels. This step helps in accounting the distributional differences in the model's prediction scores. These thresholds were finally applied to reclassify predictions from both RUN 1 and RUN 2, resulting RUN 3 and RUN 4 respectively. This

refinement recognizes that optimal decision thresholds may differ across languages due to linguistic variation and also different patterns of language usage in the context of reclaimed slurs.

2.6. Evaluation Metrics and Statistical Analysis

All results were evaluated using standard metrics appropriate for imbalanced binary classification. Reported metrics includes averaging F1 across classes equally for minimizing majority class dominance (Macro-averaged F1 score), Precision and Recall with per class and language breakdowns, and Area Under the ROC Curve for threshold-independent measure (ROC-AUC).

Model selection during all stages of the experimentation mainly relied on macro-averaged F1 calculated through stratified 5-fold cross-validation. With this approach, the minority (reclamatory) class is treated equally together with the majority class. It is a reasonable step to take in the context of the shared task, in which both false positives and false negatives have semantic importance. Confidence intervals (95%) on cross-validation scores were computed using a set of samples to quantify the uncertainty in the performance estimates.

The language-specific performance breakdowns were calculated to evaluate multilingual generalization and pinpoint challenges. The optimization histories of Optuna were visualized using parameter importance plots and trial convergence curves to know which hyperparameters had the most impact on the validation F1 score and how rapidly the TPE sampler reached the high-quality solutions.

3. Experiments and Results

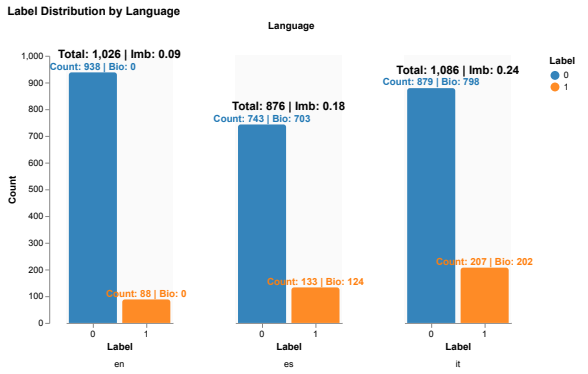
All the experiments were carried out on RTX4090 GPU workstation having the support of CUDA 12.6 and cuDNN 8.9. The training process utilized automatic mixed precision (AMP) turning to bfloat16 for the sake of computational efficiency while ensuring that the gradients remain stable. The tracking of experiments was achieved by logging hyperparameters, metrics, loss curves, and model artifacts for all the experimental runs. All the code was written in Python 3.10 using the standard scientific libraries.

3.1. Data Characteristics and Descriptive Analytics

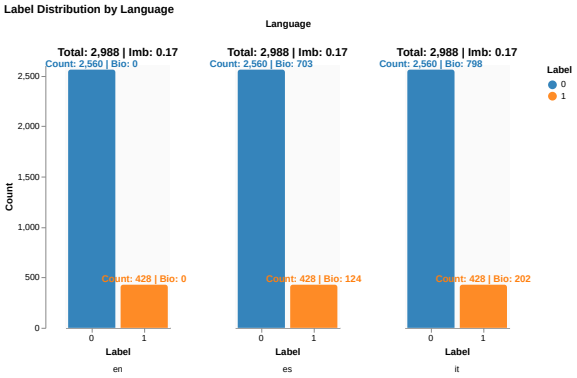
The MultiPRIDE dataset contains slurs in LGBTQ+ contexts labeled as reclamatory or non-reclamatory for usage through tweets in English, Spanish, and Italian. The first exploratory analysis uncovered a major imbalance between the classes and a distribution of labels that varied by language. The full dataset frequency analysis revealed that the reclamatory class accounted for about 9-24% of the samples with a different distribution per language: English (9%), Spanish (18%), and Italian (24%). Chi-square tests [5] of independence between language and label produced χ^2 values that were above the critical thresholds ($p < 0.001$), which meant that the labels were not randomly distributed over the languages. This finding justified the subsequent implementation of language specific decision thresholds and stratified cross validation splitting to ensure fold wise class distribution consistency. The data statistics are depicted in Figure 2.

3.2. Foundation Model Selection and Baseline Evaluation

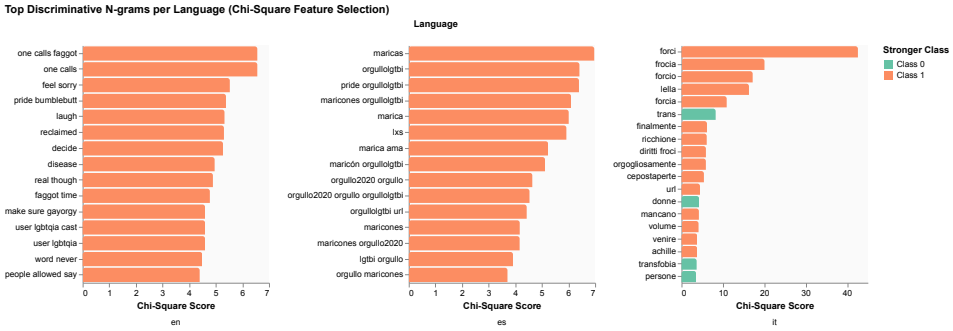
Eight multilingual embedding models were evaluated as candidates for downstream fine-tuning: multilingual-e5-large (state-of-the-art dense retrieval embeddings with 1024 dimensions) [10], bge-m3 (BGE multilingual model optimized for semantic search) [11], gte-multilingual-base (General-purpose Text Embeddings with 768 dimensions) [12], jina-embeddings-v3 (Jina's multilingual dense representation model) [13], snowflake-arctic-embed-l-v2.0 (Snowflake's large-scale multilingual embeddings) [14], LaBSE (Language-agnostic BERT Sentence Embeddings, 768 dimensions) [15], USE-multilingual (Universal Sentence Encoder for 16+ languages) [16], and XLM-RoBERTa-large (550M parameters, 24 layers, enhanced capacity for cross-lingual transfer) [17]. For each candidate model, dense representations were computed by extracting contextualized embeddings from the final transformer layer, yielding



(a) Label Distribution - Raw Data



(b) Label Distribution - Augumented Data



(c) Chi-square Analysis

Figure 2: Data distribution statistics: label imbalance in original dataset, augmented dataset after back-translation, and chi-square analysis of language-label associations.

768-dimensional or 1024-dimensional vectors depending on the model architecture. These embeddings were subsequently used as input features for classifier training in the foundation model selection phase.

A stratified 5-fold cross-validation framework was set up, which kept the same distribution of classes in the folds (80% training, 20% validation for each fold). Conventional machine learning baselines were then trained using the computed embeddings within each fold, applying linear Support Vector Machine (Linear SVC with C=1.0, dual=False for the sake of speed). The evaluation metrics computed for all folds include macro-averaged Accuracy, Precision, Recall, and F1 score. Besides, model selection preferred the macro-averaged F1 stability (low fold-wise variance) over the maximum absolute performance which means that the model is robust to data distribution shifts and can generalize. The analysis in Table 1 showed that XLM-RoBERTa-large managed to find the best point between the performance (macro F1 = 0.76 ± 0.04 across folds) and the computational efficiency, thus its election as the master model for all coming experiments was justified.

Table 1
Performance of Linear SVC across embedding models on the multilingual unified dataset.

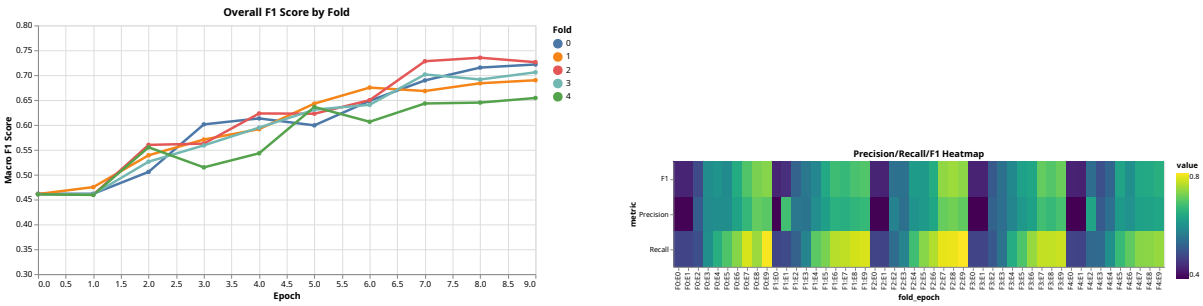
Embedding model	Accuracy	Precision	Recall	F1
multilingual-e5-large	0.8018	0.7049	0.8018	0.7330
bge-m3	0.7789	0.7002	0.7789	0.7247
gte-multilingual-base	0.7765	0.6869	0.7765	0.7120
jina-embeddings-v3	0.7795	0.6860	0.7795	0.7103
snowflake-arctic-embed-l-v2.0	0.8022	0.7085	0.8022	0.7367
labse	0.7602	0.6661	0.7602	0.6879
use-multilingual	0.7603	0.6800	0.7603	0.7023
xlm-roberta-large	0.7178	0.8280	0.7178	0.7553

3.3. Inductive Transfer Learning with Data Augmentation (RUN 1)

To mitigate data scarcity, a defined back-translation augmentation strategy was implemented. Each tweet in its original language was translated to the two alternate languages using the OpenAI GPT-4o-mini API (model="gpt-4o-mini", temperature=0.0 for deterministic output, top_p=1.0). The translation prompt explicitly requested that semantic content be preserved while acknowledging natural language variation. Each original tweet generated three variants that includes native language, and two back-translated from alternate languages. This one-to-many translation schema tripled the effective training corpus size from N to approximately 3N samples. At the same time the reclamatory class maintaining its original 9-24% distribution (i.e. as shown in Figure 2b the class imbalance ratio was preserved during augmentation).

The RoBERTa variant[2] taken as a foundation model and finetuned using a custom training loop implementing dynamic undersampling at the epoch level. A custom Batch Sampler class was implemented with the iterative method reconstructed at each epoch to enforce dynamic sampling. For each training epoch, the sampler maintained a 1:3 positive-to-negative ratio by separating dataset indices into positive (label=1) and negative (label=0) groups, drawing exactly one positive sample without replacement for each batch, and drawing exactly three negative samples without replacement. This implementation ensures stochastic variation in negative sample selection across epochs, exposing the model to diverse negative examples while maintaining consistent class ratio.

Hyperparameter optimization was performed using Optuna with the Tree structured Parzen Estimator (TPE) as the underlying sampler [18]. The TPE sampler is still keeping the probabilistic model of the objective function, gradually balancing the step of exploration in the regions of hyperparameters not yet tested with the step of exploitation of the areas around the previously observed good trials that are considered as promising ones. The Optuna objective function encompassed a 5-fold stratified cross validation loop through which the hyperparameter exploration was limited to 50 trials. The hyperparameter search space included: learning rate (log-uniform, 1e-5 to 5e-4), batch size (categorical: 16, 32, 64), weight decay (log-uniform, 1e-5 to 1e-2), and dropout rate (uniform, 0.1 to 0.4). Early trial termination was implemented via Optuna's Median Pruner with a patience threshold of 3 epochs. The trials whose validation F1 fell below the median of completed trials at the same epoch were automatically pruned. This reduces computational overhead without sacrificing final model quality. Epoch and fold level performance are shown in Figure 3.



(a) Epoch VS F1 Score

(b) Epoch VS Performance Metrics

Figure 3: Fold Level Performance Metrics for Inductive Transfer Learning.

Training was done for a maximum of 10 epochs per fold with the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-8$) with linear learning rate warmup for the first 10% of total training steps, and linear decay to zero for the remaining steps. The loss function was weighted cross entropy, given by $L = -[w_0 log(p_0) + w_1 log(p_1)]$. Where w_0 and w_1 are class weights inversely proportional to class frequencies ($w_0 \approx 0.35$, $w_1 \approx 1.00$ for the imbalanced dataset. Model checkpoints were saved at each epoch based on macro-averaged F1 score on validation set. The best checkpoint for each fold was

[2]https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base, Accessed on January 2026.

chosen as the model with the highest validation F1. This setup yielded run 1, corresponding to inductive transfer learning with augmentation and dynamic undersampling.

3.4. Domain Knowledge Integration via Masked Language Modeling (RUN 2)

For incorporating domain-specific linguistic knowledge, the foundation model was subjected to a secondary MLM pretraining task before the finetuning of downstream tasks [19]. In the MLM pretraining task, 15% of the tokens in the sequence were randomly chosen using Bernoulli sampling and replaced with the [MASK] token. The model was then trained to predict the original tokens from the contextual representations. The goal of this task is to help the model learn more about the usage of LGBTQ+ discourse patterns and reclamatory language in multilingual social media settings.

The MLM pretraining phase was conducted on the augmented dataset for a maximum of 5 epochs using the AdamW optimizer with hyper parameters identified via a preliminary Optuna search. The MLM search space includeds parameters like learning rate (log-uniform, 1e-5 to 5e-4), batch size (categorical: 16, 32, 64), weight decay (log-uniform, 1e-5 to 1e-2), and dropout (uniform, 0.1 to 0.4). Optuna evaluated for 50 trials which optimized for minimum validation cross-entropy loss on the MLM task. Pruning was again applied via MedianPruner to terminate unpromising trials early. The MLM loss function was standard cross-entropy $L\_MLM = -\sum log(P\_\theta(original\_token|masked\_context))$, which is summed over all masked positions in a batch. The validation loss with respect to parameter are shown in Figure 4.

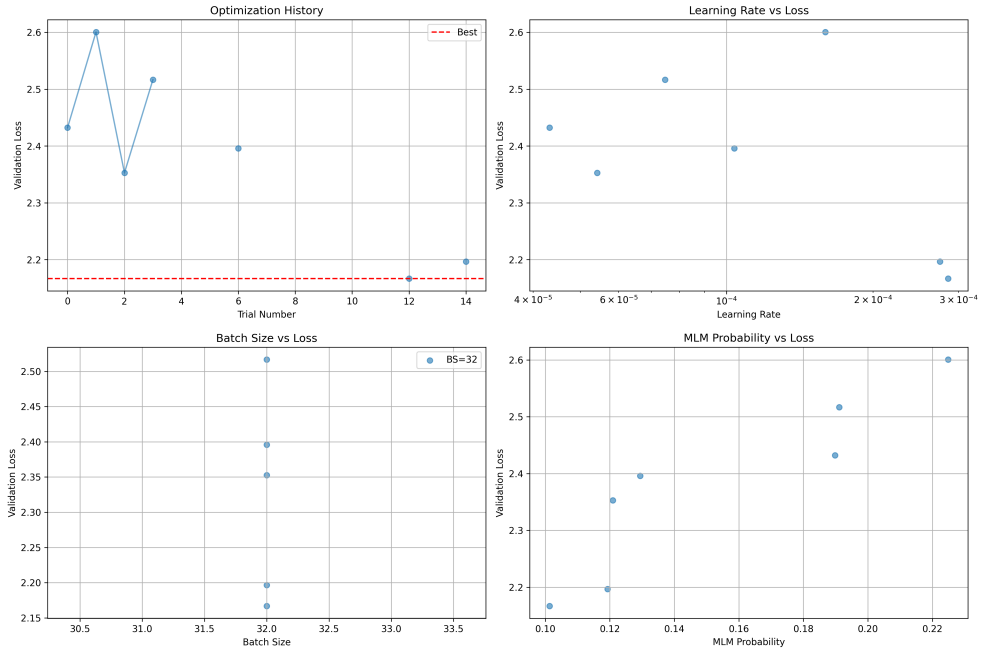
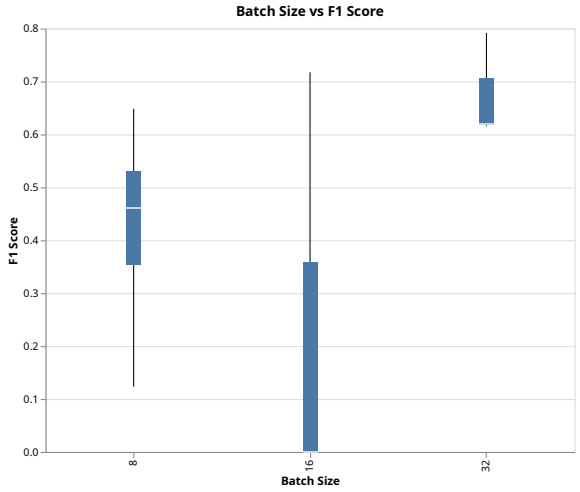
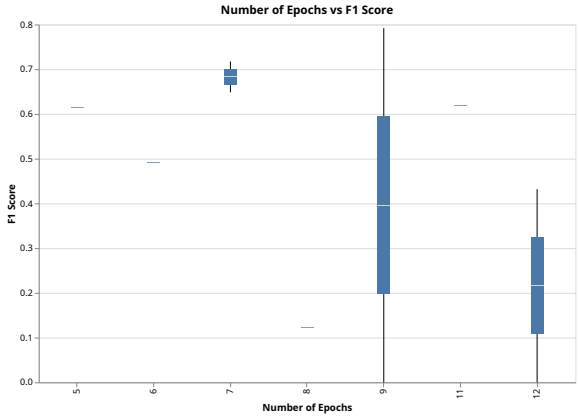


Figure 4: Transductive Transfer Learning: Parameter VS Validation Loss

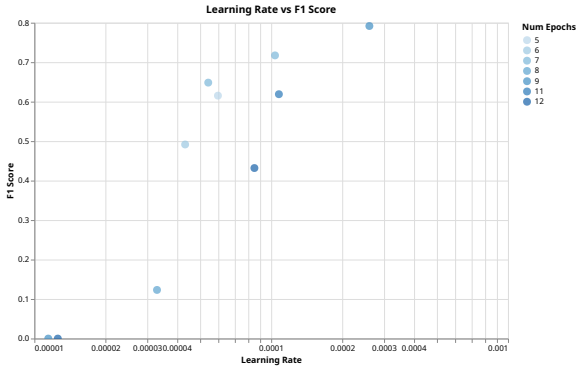
Following MLM adaptation, the finetuned model was saved and subsequently used as the initialization for downstream finetuning task. This downstream finetuning pipeline was identical to run 1 where dynamic undersampling (1:3 ratio), Optuna hyperparameter optimization (50 trials, TPE sampler, MedianPruner), 5-fold stratified cross validation, and 10 epoch training with early stopping was performed. The same hyperparameter search space as run 1 was employed, allowing direct comparison of the marginal contribution of MLM adaptation. Performance metrics collected after downstream finetuning revealed the cumulative effect of both MLM adaptation and task specific optimization, resulting run 2. Empirical analysis comparing run 1 and run 2 validation F1 scores quantified the absolute and relative improvement attributable to domain knowledge injection via MLM. The impact of inductive transfer learning on transductive model with respect to the hyper parameters are shown in Figure 5.



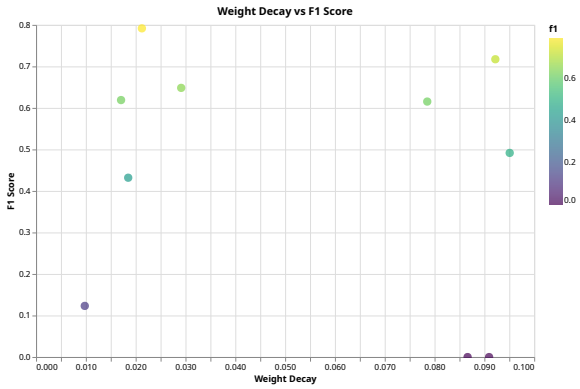
(a) Batch Size VS F1 Score



(b) Epoch VS F1 Score



(c) Learning Rate VS F1 Score



(d) Weight Decay VS F1 Score

Figure 5: Inductive Transfer Learning on Transductive Model. Impact on F1 Score with respect to the parameters.

3.5. Language-Specific Threshold Refinement and Prediction Reclassification

Both run 1 and run 2 models produced continuous confidence scores via softmax normalization of the final layer logits $conf\_score = exp(logit\_1)/(exp(logit\_0) + exp(logit\_1))$, where $logit\_0$ and $logit\_1$ denote the class-specific logits. Default binary classification employs a threshold of 0.5 with equal misclassification costs and balanced class priors. In the MultiPRIDE task, neither assumption holds where the dataset exhibits class imbalance (9-24% minority class), and language-specific label distributions differ significantly (χ^2 p < 0.001).

To optimize decision boundaries, Receiver Operating Characteristic (ROC) curve analysis was performed independently for each language. For each language (English, Spanish, Italian), validation set confidence scores were analyzed across a fine-grained threshold range. For each candidate threshold τ, performance metrics were computed includes true positive rate $(TPR) = TP/(TP + FN)$, false positive rate $(FPR) = FP/(FP + TN)$, and $F1score = 2TP/(2TP + FP + FN)$. The final selection for the optimal threshold $\tau*$ of each language was based on the value that provided the maximum macro-averaged F1 score on the validation set, thus safeguarding equal performance of both the minority and the majority classes through their simultaneous consideration.

It is interesting to point out the variation of the optimal thresholds being language dependent, thus, the differences in model confidence predictions were revealed. For instance, it was found that English tweets had quite high average confidence scores for the reclamatory class which resulted in the necessity of the higher threshold (say, $\tau\_en = 0.58$) to ensure the restriction of precision. On the

other hand, Italian tweets had lower average confidence necessitating a lower threshold (for instance, $\tau\_it = 0.42$) to capture the instances of the minority class. Spanish thresholds were usually around these extremes ($\tau\_es \approx 0.50$). These thresholds were subsequently enforced on the predictions so that if $conf\_score \geq \tau\_language$, the prediction will be reclamatory, otherwise, it will be non-reclamatory. The per language threshold variation has been represented in Figure 6.

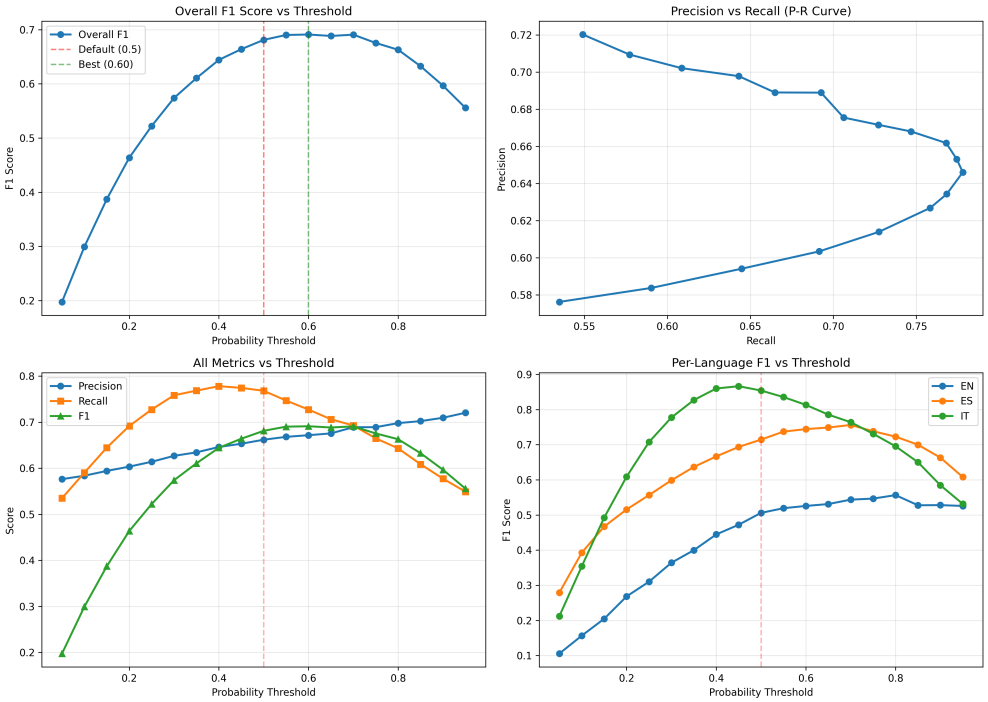


Figure 6: Threshold Analysis: Language-specific Optimal Thresholds

The predictions of run 1 model were reclassified by means of learned language specific thresholds giving rise to run 3. In the same way, reclassification of run 2 predictions was done yielding run 4. The refining of the thresholds is a very crucial post-prediction optimization step which does not require extra computational power. This step usually results in a 2-5% absolute F1 improvement by the adaptation of decision boundaries to the empirically observed confidence distributions.

4. Observations

Figure 7 presents the final submitted results on the test set. The overall results from all the runs revealed the framework's ability to deliver considerable performance improvements in the reclamation classification task by the proper management of data scarcity, class imbalance, and multilingual variation. The foundation model selection outcomes at Table 1 revealed that RoBERTa pretrained on social media context obtains superior macro-averaged F1 stability (0.7553 ± 0.04). The recall value is also stable which is essential for minimizing missed reclamatory instances in low resource contexts. This selection decision prioritizes minority class detection over overall accuracy. This is where the concern for false negatives exceeds that of false positives in the context of LGBTQ+ discourse analysis. The back translation approach to augmentation strategy was successful in tripling the equivalent training data while maintaining the language-specific distributional properties. The maintenance of the label ratio was not trivial given the dynamic epoch level with sampling at a 1:3 positive to negative ratio. This is achieved with stochastic negative sampling selection to avoid both majority class collapse and overfitting to the same examples. The combination of weighted cross entropy loss ($w_1 \approx 1.00$ for minority class) and consistent sampling constrain achieved a stable convergence point evident in the fold-wise F1 convergence by epoch 6-7.

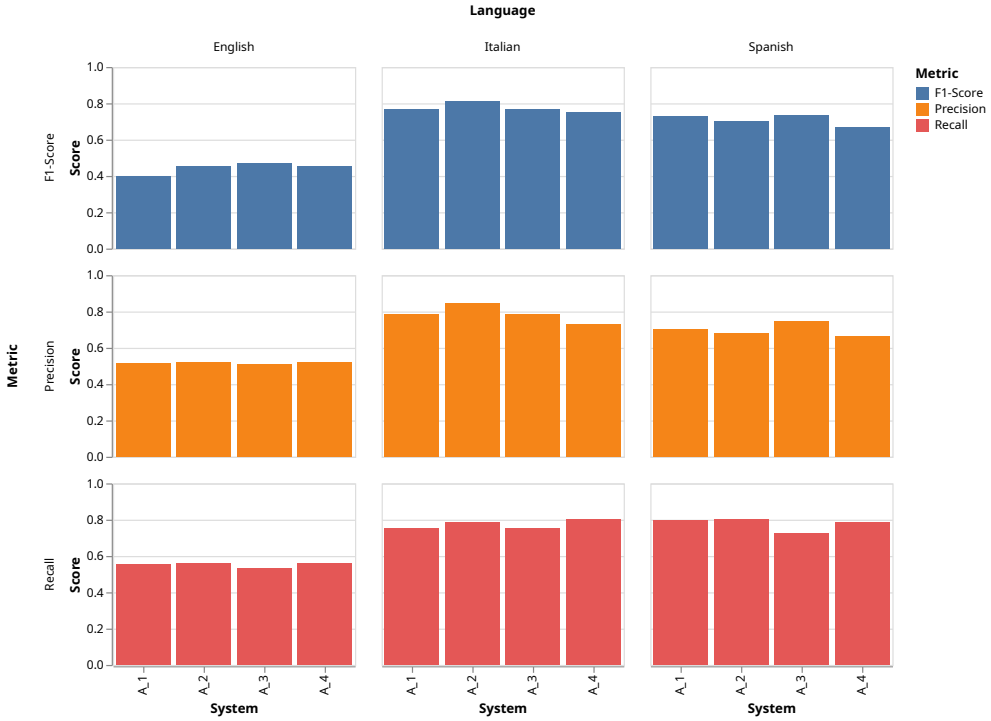


Figure 7: Final Test Set Results of Submitted Runs 1 - 4.

The integration of domain knowledge via MLM (RUN 2) hence provided language-dependent results, indicating that the adaptation of MLM is not beneficial in all multilingual settings. Although there was a marginal improvement in English, Spanish and Italian showed a more volatile response to MLM pre-training. It seems that morphologically rich languages and small amounts of pre-training data lead to different optimization spaces. The hyperparameter sensitivity analysis in Figure 5 illustrates that the adaptation of MLM models demands more accurate hyperparameter tuning, as batch size, learning rate, and weight decay all have non-monotonic relationships with F1 scores. Although the best settings lie in very narrow ranges, namely: batch size 16-32, learning rate 1e-4 to 3e-4, and weight decay 0.08-0.09. This highlights the significance of accurate hyperparameter tuning for domain adaptation. The Optuna TPE sampler demonstrated strong convergence properties around trial 30-35 within our 50-trial budget, validating the computational feasibility of the framework for shared task participation by avoiding grid search.

The most interpretable findings of Figure 6 reveal the threshold optimization that is specific to different languages. The decision boundaries that are chosen for reclamation detection are actually dependent on the languages: for English, it is 0.58, for Spanish, it is 0.50, and for Italian, it is 0.42. The range of difference is 16% and it is due to the diversity of languages in the context of culture-based reclamation and to the distribution of model confidence scores that are determined by the characteristics of training data. Consequently, English tweets that have more consistent reclamation markers produce higher average confidence scores, and thus, require higher thresholds to keep the precision. On the other hand, the Italian tweets that have dependent and more subtle cues in the context require lower thresholds to find the true positives. The significance in computation and the absolute F1 gains that are noticeable through this threshold refinement point out that a 0.5 default threshold assumption is a violation of linguistic universality. To conclude, by analyzing the per-language test sets, the asymmetries of the generalization patterns illustrated in the figure are revealed. The languages that have the highest native reclamatory prevalence (Italian 24%) exhibit different precision-recall dynamics compared to the low-prevalence languages such as English at 9%, where recall improvements are disproportionately benefiting the low-prevalence cases where class imbalance is most pronounced.

The analysis of the 60 incorrectly classified samples and the final test results shown in Figure 7 not

only emphasizes the linguistic and English-specific ambiguities but also suggests the social nature of language and its complexity in communication as the problem. The English category has errors primarily in the regions where the irony, hypothetical discourse, and theoretical discussion appear to blend with the reclamation of the terms used for the stigma. Samples such as "Does this mean I can start calling you a faggot?" (sarcasm asking for permission) and meta-textual discourse on the reclamation term terminology (e.g., arguing whether "faggot" could be a form of self-affirmation in various contexts) are properly classified as non-reclamatory but still, the system incorrectly classifies them as reclamatory. The problem is with the model as it fails to distinguish between the two uses that are covering "I'm gay and you're a faggot" which is delivered in a sarcastic manner and the actual reclamation. This is precisely what requires the knowledge of the speaker's intent and the social context that is more than the simple lexical word features. The Spanish language mistakes reveal different failure modes of the system like its getting stuck in the race-reclaiming in activist contexts where slurs are actually reclaimed just to take a stand against the heteronormative structures (e.g. "orgullosísimo de ser un peazo maricón" representing pride in activist framing). The Italian errors, in contrast, are mainly negative false where real reclamatory instances go unnoticed. Such cases contain the subtle cultural markers of reclamation where the intimacy shown through slur usage among in-group members is expressed (e.g. "sono pure ricchione. Bella sis" as solidarity acknowledgment). Likewise, the humor and community bonding through slur appropriation (e.g. "i viaggi ricchioni" as casual community reference) and self-descriptive pride assertions (e.g. "noi ricchione lavoratrici") that lack the explicit affirmative markers the model learned to recognize.

The Italian false negatives reveal a strong correlation with low confidence scores (0.42) and this means the model is not confident in its ability to detect Italian reclamatory patterns. These language specific error modes highlight that reclamation detection in multilingual situations cannot be based on uniform lexical or statistical patterns. English requires modeling of pragmatic context in order to filter out sarcasm and hypothetical discourse. the Spanish case benefits from activist discourse framing. In the Italian case, there is a need for explicit cultural knowledge about in-group solidarity and use of intimate language. The current limitation of inputting only text raises the performance ceiling, particularly for the Italian language, in which case non-linguistic cues (tone of voice, emoji sentiment, social relationship indicators) would aid in the recognition of reclamatory intent. The system with language-specific thresholds cannot yet overcome these representational limitations. These limitations include enhancing the training data with marking annotations that distinguish sarcasm from reclamation in English, embedding discourse context modeling for activist and political purposes in Spanish, and enhancing the training data with discourse patterns of intimate communities for Italian. The errors are consistent and systematic, the first grouping being according to the languages used and the second according to the pragmatic functions. This presumes that for multilingual reclamation detection to be accurate, not only the use of language aware models but also the understanding of reclamation through a pragmatically informed and culturally grounded approach as different in linguistic communities is a necessity.

5. Conclusions

This paper put forward a multistep method for slur reclamation detection to be applied in different linguistic environments. An effective training pipeline for the MultiPride shared task was constructed by using dynamic sampling to address the class imbalance problem and by increasing the dataset with back translation techniques. Testing has shown that the XLM-RoBERTa is a highly effective foundation model but its performance varies a lot between languages, and the main reason is the cultural factor.

One significant finding is that a "universal" decision boundary is not appropriate for the multilingual sentiment analysis. Adjusting the classification thresholds for each language resulted in a 2-5% improvement in F1 scores without incurring any extra computational costs. This implies that the models do not possess the same level of certainty in all languages, with English requiring more stringent thresholds to avoid false positives and Italian requiring less stringent thresholds to identify subtle cases of reclamation. The framework illustrates that systematic treatments of data imbalance, cross-lingual

variation identification, and domain knowledge application can lead to the development of robust multilingual sentiment classifiers that are socially aware and relevant to such domains.

However, the error analysis indicates that the models based solely on text are not capable of adequately understanding the pragmatic subtleties. The model tends to misidentify sarcasm as reclamation, particularly in English, and the lack of identification of implicit solidarity in Italian. The use of domain adaptation through MLM application led to only marginal improvements, requiring complex hyperparameter optimization and additional data. Future studies should aim to include features other than text, such as emojis or user interaction graphs, to more accurately identify the social intent underlying the language.

Declaration on Generative AI

During the preparation of this work, the authors used Generative AI in order to: Grammar and spelling check. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] E. Zsisku, A. Zubiaga, H. Dubossarsky, Hate speech detection and reclaimed language: Mitigating false positives and compounded discrimination, in: Proceedings of the 16th ACM Web Science Conference, 2024, pp. 241–249.

[2] B. R. Chakravarthi, R. Priyadharshini, T. Durairaj, J. P. McCrae, P. Buitelaar, P. Kumaresan, R. Ponnusamy, Overview of the shared task on homophobia and transphobia detection in social media comments, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, 2022, pp. 369–377.

[3] M. Popa-Wyatt, Reclamation: Taking back control of words, Grazer Philosophische Studien 97 (2020) 159–176.

[4] C. Ferrando, L. Draetta, M. Madeddu, M. Sosto, V. Patti, P. Rosso, C. Bosco, J. Mata, E. Gualda, Multipride at evalita 2026: Overview of the multilingual automatic detection of slur reclamation in the lgbtq+ context task, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[5] R. J. Tallarida, R. B. Murray, Chi-square test, in: Manual of pharmacologic calculations: with computer programs, Springer, 1987, pp. 140–142.

[6] S. J. Pan, Q. Yang, A survey on transfer learning. ieee transactions on knowledge and data engineering, 22 (10) 1345 (2010).

[7] OpenAI, Gpt-4o mini: advancing cost-efficient intelligence, https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, 2024. Accessed: 2026-01-07.

[8] A. Taheri, A. Zamanifar, A. Farhadi, Enhancing aspect-based sentiment analysis using data augmentation based on back-translation, International Journal of Data Science and Analytics 19 (2025) 491–516.

[9] S. Pouyanfar, Y. Tao, A. Mohan, H. Tian, A. S. Kaseb, K. Gauen, R. Dailey, S. Aghajanzadeh, Y.-H. Lu, S.-C. Chen, et al., Dynamic sampling in convolutional neural networks for imbalanced data classification, in: 2018 IEEE conference on multimedia information processing and retrieval (MIPR), IEEE, 2018, pp. 112–117.

[10] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual e5 text embeddings: A technical report, arXiv preprint arXiv:2402.05672 (2024).

[11] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, arXiv preprint arXiv:2402.03216 (2024).

[12] X. Zhang, Y. Zhang, D. Long, W. Xie, Z. Dai, J. Tang, H. Lin, B. Yang, P. Xie, F. Huang, et al., mgte: Generalized long-context text representation and reranking models for multilingual text retrieval, arXiv preprint arXiv:2407.19669 (2024).

[13] S. Sturua, I. Mohr, M. K. Akram, M. Günther, B. Wang, M. Krimmel, F. Wang, G. Mastrapas, A. Koukounas, N. Wang, et al., jina-embeddings-v3: Multilingual embeddings with task lora, arXiv preprint arXiv:2409.10173 (2024).

[14] S. Labs, Snowflake's arctic embed 2.0 goes multilingual, 2024. URL: https://www.snowflake.com/en/engineering-blog/snowflake-arctic-embed-2-multilingual/.

[15] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic bert sentence embedding, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 878–891.

[16] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Abrego, S. Yuan, C. Tar, Y.-H. Sung, et al., Multilingual universal sentence encoder for semantic retrieval, in: Proceedings of the 58th annual meeting of the Association for Computational Linguistics: system demonstrations, 2020, pp. 87–94.

[17] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 8440–8451.

[18] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 2623–2631.

[19] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N. A. Smith, Don't stop pretraining: Adapt language models to domains and tasks, arXiv preprint arXiv:2004.10964 (2020).