# Cordyceps@LT-EDI: Patching Language-Specific Homophobia/Transphobia Classifiers with a Multilingual Understanding

**Dean Ninalga**
justin.ninalga@mail.utoronto.ca

## Abstract

Detecting transphobia, homophobia, and various other forms of hate speech is difficult. Signals can vary depending on factors such as language, culture, geographical region, and the particular online platform. Here, we present a joint multilingual (M-L) and language-specific (L-S) approach to homophobia and transphobic hate speech detection (HSD). M-L models are needed to catch words, phrases, and concepts that are less common or missing in a particular language and subsequently overlooked by L-S models. Nonetheless, L-S models are better situated to understand the cultural and linguistic context of the users who typically write in a particular language. Here we construct a simple and successful way to merge the M-L and L-S approaches through simple weight interpolation in such a way that is interpretable and data-driven. We demonstrate our system on task A of the *Shared Task on Homophobia/Transphobia Detection in social media comments* dataset for homophobia and transphobic HSD. Our system achieves the best results in three of five languages and achieves a 0.997 macro average F1-score on Malayalam texts.

## 1 Introduction

In general, the US is seeing an increase in institutionalized transphobia in the form of banning gender-affirming care and the banning of transgender youth from several sports (Kline et al., 2023). However, studies on individuals who experience institutionalized transphobia in the US experience more psychological distress and instances of suicidal ideation (Price et al., 2023). The codifying of anti-trans laws then certainly must give confidence to those with transphobic beliefs and desires to spread anti-trans rhetoric in online spaces. Berger et al. (2022) recently presented results showing that LGBTQ youth often rely on social media for improved mental health outcomes and as a source of social connection that helps close mental health disparities. Therefore, appropriate content moderation on social media platforms stands to benefit from accurate NLP systems that can identify homophobia, transphobia, and other forms of hate speech.

Good knowledge of hate speech in a particular language may not always be useful for other languages, yet many common phrases and sayings are often expressed across languages. Namely, purveyors of hate speech often do not openly say hateful comments but instead rely on equally vicious code phrases, or *dogwhistles*, to avoid existing content moderation systems (Henderson and McCready, 2017; Magu et al., 2017). Knowledge of the hidden meanings of these encoded sayings can create powerful tools for improving online moderation (Mendelsohn et al., 2023). These phrases can easily transcend the regions of their origin, spreading across online communities without detection in vulnerable communities. Hence,



*LT-EDI 2023 – Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 185–191, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-084-7_028

knowledge of dogwhistles in their current form will make content moderation systems more robust to these signals as they appear in different languages in new online spaces.

Textual databases built for hate speech analysis are predominantly in English, which creates language-based performance disparities (Jahan and Oussalah, 2021; Poletto et al., 2020; Aluru et al., 2020). As Wang et al. (2020) suggested, in M-L models languages are in competing for model resources, potentially resulting in worse performance for low-resources languages. This performance bias is possibly due to that many M-L datasets used for pretraining popular language models often are majority English samples, often by a wide margin (Barbieri et al., 2021; Xue et al., 2020; Ri et al., 2021)). Consequently, there is a general disparity in performance when comparing English-only and M-L HSD models (Röttger et al., 2022).

Nozza et al. (2020) push for more pretrained models in non-English languages as they will (naturally) be best for downstream tasks in the same language domain they are trained in. However, pre-training techniques typically require large datasets to guarantee good downstream performance. Given a relative lack of language-specific data for HSP, more indirect and creative approaches are required to alleviate the performance gap between English and non-English tasks.

For our present purposes, we are presented with multiple target languages and tasked to detect levels of homophobia and transphobia for each specified language using an automated system. We introduce Language-PAINT to jointly model M-L and L-S knowledge that incorporates recent work on weight interpolation.

In summary, our main contributions are the following:

- We publicize a language-based weight interpolation approach as the next step in advancing HSD research.
- We provide a demonstration of our framework on task A of the *Shared Task on Homophobia/Transphobia Detection in social media comments* (Chakravarthi et al., 2022).
- We provide preliminary evidence suggesting that our framework is robust to label distribution shifts.

## 2 Related Work

### 2.1 Language Transfer in Hate Speech Detection

Several techniques from recent years have worked on closing the performance disparity between majority and minority languages in HSD. Namely, several attempts directly translate low-resource languages into high-resource ones (Pamungkas and Patti, 2019; Ibrohim and Budi, 2019). Pelicon et al. (2021) presents a data-based approach that first trains a M-L model for HSD, similar to our training scheme's initial step. Pelicon et al. (2021) use a percentage of L-S data to finetune their model where the percentage is chosen empirically. Choudhury et al. (2017) delay training with code-mixed data, opting to first train with mono-lingual samples using the two languages used in the code-mixed data. The popular IndicNLP (Kunchukuttan et al., 2020) uses bilingual word embeddings for translation and transliteration, typically between English and a target low-resource language. Biradar et al. (2021) subsequently attempt to incorporate IndicNLP's (Kunchukuttan et al., 2020) embeddings for code-mixed HSD.

### 2.2 Weight Interpolation

In this paper, we adopt the interpolation strategy of *Weight-space ensembles for fine-tuning*

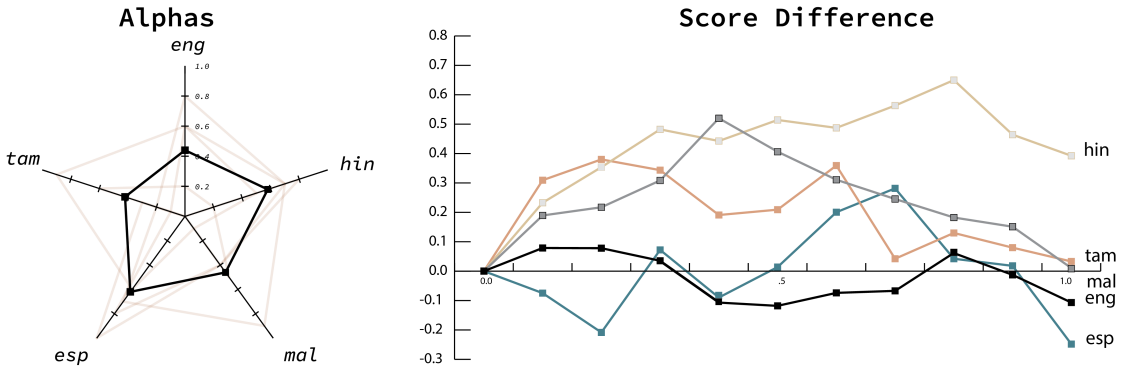


Figure 1: Left: Average selected value for $\alpha$ (thick black line) averaged over five runs for each language. Right: Average validation F1 score as a function of $\alpha$ reported for each language, averaged over five runs.

(WiSE-FT) (Wortsman et al., 2021). In particular, we base our framework on a subsequent variation called PAINT (Ilharco et al., 2022) constructed to incorporate the input robustness of a zero-shot model into finetuned models across diverse tasks. Formally, given a single task $t$ takes the weights of the *zero-shot* model $\theta_z$ and a finetuned model $\theta_f$, the weight interpolation of PAINT performs the interpolation:

$$\theta^t = \alpha\theta_z + (1-\alpha)\theta_f$$

with $\alpha \in [0, 1]$. In addition to the specific experiments performed by (Ilharco et al., 2022), recent work shows that averaging two (or more) language models has the potential to leverage knowledge contained in each (Gueta et al., 2023; Don-Yehiya et al., 2022; Choshen et al., 2022). However, no prior work has studied weight space ensembling based on language to the best of our knowledge.

## 3 Methodology

Here, we use Bernice (DeLucia et al., 2022), a language model exclusively on Twitter[1] data and is known to be performant on HSD across multiple languages. Indeed, many studies rely on Twitter, to construct datasets of code-mixed samples for various HSD approaches (Bhat et al., 2018; Bansal et al., 2020; Farooqi et al., 2021; Choudhury et al., 2017), which in aggregate, motivates our choice of language model.

### 3.1 Language-PAINT

Given $k$ distinct groups of (possibly code-mixed) languages, we first train a M-L model on a dataset that includes all the languages. We continue training until saturation on a validation set, where we take the average F1 score across languages. Next, we create an additional $k$ L-S models, - one for each language - where each is initialized with the weights of the M-L model. Finally, we perform linear interpolation between the weights of the M-L and each of the $k$ L-S models. The resulting $k$ models are used for inference on each language.

In mathematical terms, Language-PAINT takes the weights of the trained L-S model $\theta^i_{ls}$ and the weights of the M-L model $\theta_{ml}$ and performs the following interpolation:

$$\theta^i = \alpha\theta^i_{ls} + (1-\alpha)\theta_{ml}.$$

Where $\theta^i$ is used to create predictions for the respective language $i = 1, .., k$ in the test set. In practice, we select alpha from a discrete set

[1] https://twitter.com

| Paramter | Value |
|---|---|
| Batch Size | 16 |
| Learning Rate | 1e-5 |
| Optimizer | Adam |
| Loss | cross-entropy |

Table 1: Training Hyper-parameters

$\alpha \in \{0, 0.1, 0.2, ..., 1\}$ and select based on the resulting model's F1 performance on a held-out validation set.

### 3.2 Ensembling

Our final prediction on the test sets is an ensembled output of five models trained on five stratified folds. To create these folds, we first conjoined the original training and development sets. Next, we divided the conjoined dataset into five folds using 80-20 train-validation splits, ensuring we maintain the label distribution across each fold. We then trained a fresh model on each training and validation fold using the methodology that is described above. For final inference, we sum the output probabilities of the five models selecting the maximum probability as the final prediction.

### 3.3 Data Cleaning

To preserve as much textual information as possible, we apply minimal additional cleaning steps. Namely, we only remove a sample if it is found to be overlapping in both the train and development data. In total, we removed 1695 duplicate samples, where 54% of the dropped samples are in Tamil and 41% are in Malayalam.

## 4 Experiments and Results

### 4.1 Experimental Setup

Here, we will perform experiments comparing the L-S, M-L, and, LangPAINT approaches. For our first experiment, we combine the training and development set into a single case study. We train five models re-sampling a random 80-20 train-validation split for each run and report the average results on the test set. For our second experiment, we combine the training, development, and, test sets into a single dataset. Where we train ten models re-sampling a random 80-10-10 train-validation-test split for each run, reporting the average of the results on each test set. For each of our two experiments, we use the *weighted* F1 score to evaluate performance. All experiments were run on a single Tesla V4 GPU and we provide the training hyperparameters in Table 1.

## 5 Results

The results of our experiments are given in Table 2. We can see for most languages, the L-S approach tends to perform best, with the exception of the Malayalam language. This is reflective of our final leaderboard results where we used an ensemble method (see Section 3.2) that achieves a 0.997 macro average F1-score on Malayalam texts. Additionally, we report the selected values for $\alpha$ and validation score as a function of $\alpha$ in Figure 1 for this first experiment.

For our second experiment, our results (see Table 2) are much more in favor of our method. Perhaps the considerably worse performance of the L-S and M-S models is due to the high label-distribution shift between the re-sampled train and test splits. Nonetheless, LangPAINT appears to be robust to this shift and is still able to maintain good performance, with the only exception being the Spanish language.

## 6 Conclusion

In this paper, we introduce LangPAINT. LangPAINT is a weight space ensembling strategy (Wortsman et al., 2021) repurposed to jointly model the multi-lingual and language-specific

| | Test set | | | 10 Fold | | |
|---|---|---|---|---|---|---|
| **Language** | L-S | M-L | LangPAINT (ours) | L-S | M-L | LangPAINT (ours) |
| eng | **0.93** | 0.928 | **0.93** | 0.565 | 0.584 | **0.94** |
| hin | **0.943** | 0.939 | 0.939 | 0.478 | 0.541 | **0.932** |
| mal | 0.965 | 0.97 | **0.971** | 0.834 | 0.827 | **0.930** |
| esp | **0.878** | 0.874 | 0.877 | 0.91 | **0.932** | 0.877 |
| tam | **0.927** | 0.923 | **0.927** | 0.87 | 0.878 | **0.895** |

Table 2: Results of our experiments comparing the language-specific (L-S), multi-lingual (M-L) and, LangPAINT approaches across languages. We report the *weighted* F1 score for each, where the results are the average of five runs.

signals of homophobia and transphobia. Our experiments suggest that our method is competitive with the language expert models and has the potential to be very robust to label distribution shifts. On task A of the *Shared Task on Homophobia/Transphobia Detection in social media comments* (Chakravarthi et al., 2022) achieving the best results in three of five languages and achieves a 0.997 macro average F1-score on Malayalam, a low-resource language.

## References


Sai Saket Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *ArXiv*, abs/2004.06465.

Srijan Bansal, Vishal Garimella, Ayush Suhane, Jasabanta Patro, and Animesh Mukherjee. 2020. Code-switching patterns can be an effective route to improve performance of downstream nlp applications: A case study of humour, sarcasm and hate speech detection. In *Annual Meeting of the Association for Computational Linguistics*.

Francesco Barbieri, Luis Espinosa Anke, and José Camacho-Collados. 2021. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. In *International Conference on Language Resources and Evaluation*.

Matthew N Berger, Melody Taba, Jennifer L. Marino, Megan S. C. Lim, and S. Rachel Skinner. 2022. Social media use and health and well-being of lesbian, gay, bisexual, transgender, and queer youth: Systematic review. *Journal of Medical Internet Research*, 24.

Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Manish Shrivastava, and Dipti Misra Sharma. 2018. Universal dependency parsing for hindi-english code-switching. In *North American Chapter of the Association for Computational Linguistics*.

Shankar Biradar, Sunil Saumya, and Arun Chauhan. 2021. Hate or non-hate: Translation based hate speech identification in code-mixed hinglish data set. *2021 IEEE International Conference on Big Data (Big Data)*, pages 2470–2475.

Bharathi Raja Chakravarthi, Adeep Hande, Rahul Ponnusamy, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022. How can we detect homophobia and transphobia? experiments in a multilingual code-mixed setting for social media governance. *International Journal of Information Management Data Insights*, 2(2):100119.

Leshem Choshen, Elad Venezian, Noam Slonim, and Yoav Katz. 2022. Fusing finetuned models for better pretraining. *ArXiv*, abs/2204.03044.

Monojit Choudhury, Kalika Bali, Sunayana Sitaram, and Ashutosh Baheti. 2017. Curriculum design for code-switching: Experiments with language identification and language modeling with deep neural networks. In *ICON*.

Alexandra DeLucia, Shijie Wu, Aaron Mueller, Carlos Aguirre, Philip Resnik, and Mark

Dredze. 2022. Bernice: A multilingual pre-trained encoder for Twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6191–6205, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shachar Don-Yehiya, Elad Venezian, Colin Raffel, Noam Slonim, Yoav Katz, and Leshem Choshen. 2022. Cold fusion: Collaborative descent for distributed multitask finetuning. *ArXiv*, abs/2212.01378.

Zaki Mustafa Farooqi, Sreyan Ghosh, and Rajiv Ratn Shah. 2021. Leveraging transformers for hate speech detection in conversational code-mixed tweets. In *Fire*.

Almog Gueta, Elad Venezian, Colin Raffel, Noam Slonim, Yoav Katz, and Leshem Choshen. 2023. Knowledge is a region in weight space for fine-tuned language models. *ArXiv*, abs/2302.04863.

Robert Henderson and Eric McCready. 2017. How dogwhistles work. In *JSAI-isAI Workshops*.

Muhammad Okky Ibrohim and Indra Budi. 2019. Translated vs non-translated method for multilingual hate speech identification in twitter. *International Journal on Advanced Science, Engineering and Information Technology*.

Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. 2022. Patching open-vocabulary models by interpolating weights. *ArXiv*, abs/2208.05592.

Md Saroar Jahan and Mourad Oussalah. 2021. A systematic review of hate speech automatic detection using natural language processing. *ArXiv*, abs/2106.00742.

Nolan S. Kline, Nathaniel J. Webb, Kaeli C. M. Johnson, Hayley D. Yording, Stacey B. Griner, and David J. Brunell. 2023. Mapping transgender policies in the us 2017-2021: The role of geography and implications for health equity. *Health & place*, 80:102985.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, C. GokulN., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *ArXiv*, abs/2005.00085.

Rijul Magu, Kshitija Joshi, and Jiebo Luo. 2017. Detecting the hate code on social media. In *International Conference on Web and Social Media*.

Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. 2023. From dogwhistles to bullhorns: Unveiling coded rhetoric with language models. *ArXiv*, abs/2305.17174.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [mask]? making sense of language-specific bert models. *ArXiv*, abs/2003.02912.

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 *Annual Meeting of the Association for Computational Linguistics*.

Andraz Pelicon, Ravi Shekhar, Bla krlj, Matthew Purver, and Senja Pollak. 2021. Investigating cross-lingual training for offensive language detection. *PeerJ Computer Science*, 7.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477 – 523.

Maggi A. Price, Nathan L. Hollinsaid, Sarah C. McKetta, Emily J Mellen, and Marina Rakhilin. 2023. Structural transphobia is associated with psychological distress and suicidality in a large national sample of transgender adults. *Social Psychiatry and Psychiatric Epidemiology*, pages 1 – 10.

Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. 2021. mluke: The power of entity representations in multilingual pretrained language models. In *Annual Meeting of the Association for Computational Linguistics*.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual hatecheck: Functional tests for multilingual hate speech detection models. *ArXiv*, abs/2206.09917.

Zirui Wang, Zachary Chase Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Conference on Empirical Methods in Natural Language Processing*.

Mitchell Wortsman, Gabriel Ilharco, Mike Li, Jong Wook Kim, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. 2021. Robust fine-tuning of zero-shot models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7949–7961.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. In *North American Chapter of the Association for Computational Linguistics*.