UD-KD: A Structure-Aware Framework for Zero-Shot Cross-Lingual Offensive Language Detection of Large Language Models

Cui Mengxin, Zhu Shaolin[†]

Huazhong University of Science and Technology

Tianjin University

Abstract

While large language models (LLMs) have demonstrated remarkable capabilities in high-resource languages, their proficiency in detecting offensive language significantly deteriorates when applied to low-resource languages, a challenge compounded by the high costs of data annotation and model deployment. To address this, we propose UD-KD (Unified Debiased Knowledge Distillation), a novel framework that enables zero-shot cross-lingual offensive language detection without requiring any labeled data in the target language. Our approach distills knowledge from a teacher LLM in a high-resource language (e.g., English) by capturing not just its predictions, but its underlying structured reasoning—including attentional patterns and semantic representations. We introduce a structure-aware distillation mechanism that aligns these deeper patterns across disparate languages and a virtual adversarial invariance module that enhances model robustness on unlabeled target-language data. Furthermore, our framework incorporates a geometric debiasing component to mitigate spurious correlations associated with identity terms. Through extensive experiments on multiple student models and across several target languages (e.g., Turkish, Russian, and Italian), we demonstrate that UD-KD substantially outperforms established baselines in accuracy, robustness, and fairness, critically reducing the reliance on target-language annotations. Our work presents a practical, low-cost, and scalable solution for content moderation on multilingual platforms, offering a viable pathway for deploying advanced NLP technology to safeguard global online discourse.

Keywords: Large language models; Offensive Language Detection; Hierarchical Distillation; Self-Correcting Loop

1. Introduction

Contemporary large language models (LLMs) have achieved unprecedented success across a wide array of natural language processing tasks, fundamentally altering the research landscape [1, 2]. These models are predominantly trained on vast, multilingual corpora, yet the underlying language distribution is profoundly imbalanced [3]. For instance, English constitutes the lion's share of the training data for most foundational models, while hundreds of other languages

[†]Corresponding author: Zhu Shaolin (Email: zhushaolin@tju.edu.cn)

are represented by orders of magnitude less text. This data imbalance has led to a significant performance disparity, where LLMs exhibit remarkable proficiency in high-resource languages but lag considerably in most others, particularly for nuanced and context-sensitive tasks such as offensive language detection [4, 5].

The automatic detection of offensive, toxic, and hateful content is critical for maintaining healthy online ecosystems, especially on global social media platforms and in the comment sections of international news outlets [6, 7]. However, building effective detection systems for low-resource languages is fraught with challenges, including severe data scarcity for supervised training, the high cost of culturally-aware annotation, and the subtle, often implicit nature of offensive expressions [8, 9]. While one could theoretically deploy massive LLMs like GPT-4 for moderation, their substantial computational and financial costs render them impractical for real-time, large-scale deployment at the endpoint [10, 11]. This creates an urgent need for lightweight, yet highly effective, models that can operate in diverse linguistic environments.

A common approach to bridge this resource gap is to leverage the "translate-then-finetune" paradigm [12, 13]. In this method, labeled training data from a high-resource language like English is machine-translated into the target low-resource language, and a smaller student model is then fine-tuned on this synthetic corpus. However, this seemingly straightforward approach encounters several critical limitations. First, machine translation systems, especially for informal and offensive language, are prone to errors. They may fail to preserve the toxic nuance, mistranslate culturally-specific insults, or introduce artifacts, leading to a noisy and often misleading training signal [14]. Second, this method facilitates only a superficial transfer of labels, failing to distill the deeper, structural reasoning that underpins an LLM's judgment [15]. A model trained on translated text may learn spurious correlations between translated artifacts and labels rather than the fundamental linguistic patterns of toxicity. Finally, as shown in prior work, continuously training models on out-of-domain or translated data can inadvertently degrade their capabilities in their original high-resource language [16]. Therefore, we explore a new question along this trajectory: *Besides simply translating labels, can we distill the deeper, language-agnostic reasoning capabilities of LLMs to enhance robust performance in low-resource languages without labeled target-language data?*

In this paper, we introduce UD-KD (Unified Debiased Knowledge Distillation), a novel framework for zero-shot cross-lingual offensive language detection. Instead of merely transferring output probabilities, UD-KD distills a richer, multi-faceted understanding from the teacher LLM. Our framework is comprised of three core components. (1) **Structure-Aware Distillation**: We move beyond logits and distill knowledge from the teacher's internal representations, such as attention patterns and hidden states. This forces the student model to mimic *how* the teacher reasons about linguistic structure, not just *what* it predicts. (2) **Adversarial Invariance Distillation**: Leveraging large amounts of unlabeled target-language text, we employ a variant of virtual adversarial training [17] to instill a notion of semantic robustness, making the student model resilient to minor perturbations in unseen languages. (3) **Geometric Debiasing**: We proactively mitigate biases by identifying and neutralizing spurious correlations tied to identity terms within the model's embedding space, a critical step for fairness in content moderation [18]. We conduct extensive experiments on LLaMA-2-7B [19] and SeaLLM-7B [20], with English as the source language and Turkish, Russian, and Italian as unseen target languages. Experimental results demonstrate that UD-KD significantly outperforms a suite of strong baselines, including standard fine-tuning, translate-train, and vanilla knowledge distillation, across metrics of accuracy, robustness, and fairness. Our main contributions are:

- We propose UD-KD, a new and comprehensive framework for zero-shot cross-lingual offensive language detection that effectively transfers deep structural and invariant knowledge from a teacher LLM.
- We introduce a novel combination of structure-aware distillation, adversarial invariance training on unlabeled data, and geometric debiasing to address the core challenges of noise, structural mismatch, and bias in cross-lingual transfer.
- We demonstrate through extensive experiments that our method achieves state-of-the-art performance, offering a practical and scalable pathway for deploying fair and effective content moderation systems in low-resource environments.

2. Related Work

Our research is situated at the intersection of three primary areas: multilingual capabilities of large language models, cross-lingual knowledge transfer, and offensive language detection.

2.1. Multilingual Large Language Models

The paradigm of pre-training on large, multilingual corpora has become standard for developing models with cross-lingual understanding capabilities [21, 22, 23]. Models like mBERT [21] and XLM-RoBERTa [22] demonstrated that representations learned from text in multiple languages could be aligned in a shared semantic space, enabling zero-shot cross-lingual transfer. More recently, massively multilingual models such as BLOOM [24] and mT5 [25], as well as foundational models like GPT-4 and LLaMA-2, have shown impressive, albeit often uneven, performance across a vast number of languages. Despite their multilingual pre-training, recent studies have consistently highlighted the performance gap between high-resource and low-resource languages [2]. Research has shown that while these models possess surprising zero-shot or few-shot capabilities in many languages [26, 27]. Our work builds upon this understanding, aiming not to create a new multilingual model from scratch, but to propose a method that specifically targets and enhances the latent low-resource capabilities within existing powerful LLMs like LLaMA-2.

2.2. Cross-Lingual Knowledge Transfer

To address the data scarcity problem in low-resource languages, various cross-lingual knowledge transfer techniques have been explored. A dominant line of work is **translate-train**, where high-resource labeled data is machine-translated to the target language for fine-tuning [28, 4]. While effective to a degree, this approach is highly dependent on the quality of the machine translation system, which often falters on domain-specific, informal, or adversarial text [29]. Another prominent approach is **cross-lingual alignment**, which seeks to explicitly align model representations across languages. This can be achieved through various means, including multilingual aligned lexicons [30], alignment of word or sentence embeddings [31], or using parallel data to construct alignment-focused pre-training tasks [32]. More recent work has explored leveraging LLMs themselves to generate aligned data through prompting or self-translation [33, 34]. Compared to these methods, which often require parallel corpora or sophisticated alignment objectives, our UD-KD framework performs a more implicit, end-to-end alignment by distilling the entire reasoning process of a model, guided by its internal structural representations.

Our work is most closely related to **knowledge distillation** for cross-lingual transfer. Prior studies have used distillation to transfer knowledge from a large teacher model to a smaller student model for tasks like machine translation [35] or to compress multilingual LLMs [36]. The SDRRL method [36], which inspires our work, proposes using self-distillation to improve an LLM's multilingual performance by leveraging its own responses in a high-resource language. Our UD-KD framework significantly extends this concept by moving beyond the distillation of surface-level responses to distill deeper, *structural* knowledge and *invariant* properties, which we argue are more fundamental and language-agnostic signals for effective cross-lingual transfer.

2.3. Offensive Language Detection

Automatic offensive language detection is a long-standing and critical research area [37]. Early work relied on feature engineering and traditional machine learning models [38], while later efforts shifted to deep learning architectures like LSTMs and CNNs [39]. The advent of pre-trained language models like BERT has set new state-of-the-art benchmarks for this task, particularly in English [40]. However, extending these successes to multilingual and low-resource settings remains a significant challenge [41]. Most research in this subfield has focused on the translate-train paradigm or the direct application of multilingual models [42, 43]. While valuable, these approaches often fail to account for the fairness and bias amplification issues that are particularly pernicious in toxicity detection [44]. Biases present in the source-language data can be transferred and even amplified in the target language. Our UD-KD framework explicitly addresses this by incorporating a geometric debiasing module, aiming to produce not only an accurate but also a fair cross-lingual detection model. This focus on proactive debiasing during the knowledge transfer process distinguishes our work from most prior efforts in multilingual offensive language detection.

3. Methodology

In this work, we aim to enhance the multilingual capabilities of a foundational LLM for the task of offensive language detection. Specifically, we focus on the zero-shot cross-lingual setting, where the model is adapted using labeled data from a high-resource source language, $\mathcal{L}_{src}$ (e.g., English), to perform effectively on an unseen low-resource target language, $\mathcal{L}_{tgt}$. Our approach, termed **UD-KD**, treats this adaptation as an internal knowledge transfer or *self-distillation* process. The core principle is to leverage the LLM's own robust, high-resource persona as a "teacher" to guide the refinement of its latent capabilities in the target language.

3.1. Problem Formulation and Overview

Let M_θ be a foundational LLM with parameters θ. Our goal is to learn an updated set of parameters θ' that improves performance on $\mathcal{L}_{tgt}$ while preserving proficiency in $\mathcal{L}_{src}$. The training regimen utilizes a labeled source-language dataset $\mathcal{D}_{src} = \{(x_i, y_i)\}_{i=1}^{N}$ and a large, unlabeled corpus of target-language text $\mathcal{D}_{tgt}$.

The learning process is governed by a unified loss function that holistically addresses three primary challenges in cross-lingual transfer: (1) bridging the linguistic and structural gap between languages, (2) ensuring robust generalization from limited signals, and (3) maintaining fairness by mitigating harmful biases. The total loss function is a weighted sum of three corresponding components:

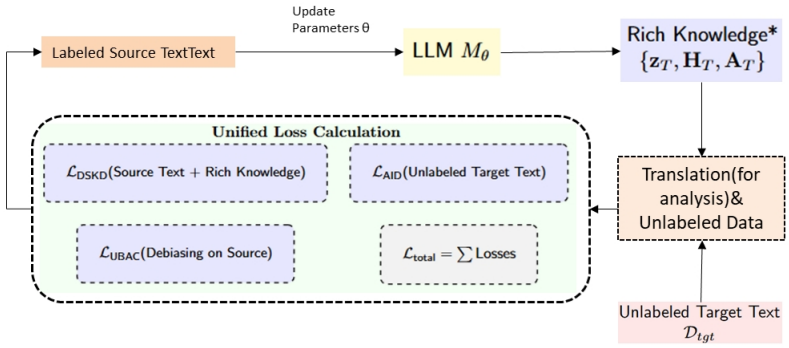


Figure 1: Conceptual overview of the UD-KD framework. The model's parameters θ are updated via a composite loss. The DSKD and UBAC modules operate on labeled source-language data, while the AID module leverages unlabeled target-language data to enhance robustness.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{DSKD}} + \lambda_{\text{AID}} \cdot \mathcal{L}_{\text{AID}} + \lambda_{\text{UBAC}} \cdot \mathcal{L}_{\text{UBAC}} \tag{1}$$

where $\mathcal{L}_{\text{DSKD}}$ is the Structure-Aware Self-Distillation loss, $\mathcal{L}_{\text{AID}}$ is the Adversarial Invariance Distillation loss, and $\mathcal{L}_{\text{UBAC}}$ is the Unsupervised Bias Alignment and Correction loss. The hyperparameters λ_{AID} and λ_{UBAC} control the influence of the robustness and fairness components. Figure 1 provides a conceptual overview of the framework.

3.2. Structure-Aware Self-Distillation (DSKD)

The central hypothesis of DSKD is that the *process* of reasoning is more language-agnostic than the final prediction. Consequently, effective knowledge transfer should focus on distilling the rich, internal computational structures of the LLM. This assumption is supported by linguistic typology research, which has documented substantial structural commonalities across human languages. In particular, the Universal Dependencies (UD) framework [45] formalizes cross-linguistic syntactic relations—such as subject–object dependencies, modifier–head relations, and clause-level hierarchies—that are consistently observed across typologically diverse languages, even when surface word order differs. Similarly, the theory of Phrase Structure Universals [46] identifies recurring hierarchical organization patterns in the grammars of unrelated languages, suggesting that certain syntactic abstractions are shared and thus transferable across linguistic boundaries. We define a "teacher" persona, $M_{\theta_{\text{frozen}}}$, which is simply the model with its initial, pre-trained parameters held constant. The "student" is the same model, M_θ, but with its parameters being actively updated. The DSKD loss, computed on $\mathcal{D}_{src}$, guides the student to emulate the teacher's internal mechanics.

3.2.1. Distilling Relational Structure via Attention Alignment

The self-attention mechanism in Transformers computes the relative importance of each token with respect to all other tokens in a sequence, forming an attention matrix $\mathbf{A} \in \mathbb{R}^{L \times L}$ (where L is sequence length) for each head and layer. These matrices implicitly encode syntactic and co-occurrence patterns. We argue that the high-level geometry of these patterns should be consistent for a similar reasoning task, regardless of the surface language.

To enforce this, we align the attention matrices from corresponding layers of the student and teacher. A naive L2 distance is unsuitable due to the arbitrary rotational nature of learned

representations. Instead, we employ **Centered Kernel Alignment (CKA)** [47], a similarity index that is invariant to orthogonal transformations. For a selected set of layers L_{sel}, the attention alignment loss is:

$$\mathcal{L}_{\text{CKA}} = \sum_{l \in L_{sel}} \left(1 - \text{CKA}(\mathbf{A}_l^{\text{student}}(x_{src}), \mathbf{A}_l^{\text{teacher}}(x_{src}))\right) \tag{2}$$

where $\text{CKA}(\cdot, \cdot)$ measures the similarity between the reshaped attention matrices. This loss encourages the student to learn a consistent relational structure for analyzing text.

3.2.2. Distilling Semantic Abstractions via Hidden State Alignment

As information flows through the layers of an LLM, the hidden states $\{\mathbf{H}_l\}$ represent increasingly abstract semantic features. We distill this hierarchical abstraction process by aligning the distribution of hidden states between the student and teacher. We use the **Maximum Mean Discrepancy (MMD)** [48] with a Radial Basis Function (RBF) kernel, which is a non-parametric method for comparing probability distributions. The loss is computed as:

$$\mathcal{L}_{\text{MMD}} = \sum_{l \in L_{sel}} \text{MMD}^2 \left(\mathbf{H}_l^{\text{student}}(x_{src}), \mathbf{H}_l^{\text{teacher}}(x_{src})\right) \tag{3}$$

This objective function guides the student's representation space to evolve in a way that mirrors the teacher's stable, high-resource semantic manifold, promoting a more effective transfer of abstract concepts.

3.2.3. Standard Logits Distillation

To ground the distillation process, we also include the standard knowledge distillation loss based on the final output logits, $\mathbf{z}$. This ensures the student's final predictions align with the teacher's softened probability distribution.

$$\mathcal{L}_{\text{CE}} = \text{KL}(\sigma(\mathbf{z}_{\text{student}}/\tau) \| \sigma(\mathbf{z}_{\text{teacher}}/\tau)) \tag{4}$$

The DSKD loss is the weighted sum: $\mathcal{L}_{\text{DSKD}} = \mathcal{L}_{\text{CE}} + \alpha \cdot \mathcal{L}_{\text{MMD}} + \beta \cdot \mathcal{L}_{\text{CKA}}$.

3.3. Adversarial Invariance Distillation (AID)

A key challenge in zero-shot transfer is that the model must generalize to a data distribution it has never seen. To improve this generalization, our AID module enhances the model's robustness on the unlabeled target-language corpus, $\mathcal{D}_{tgt}$. The principle is that a robust model's predictions should be stable under small, local perturbations of its input.

We implement this using **Virtual Adversarial Training (VAT)** [17]. VAT introduces a regularization term that promotes smoothness in the model's output distribution. For each unlabeled sample x_{tgt}, we seek a small perturbation $\mathbf{r}_{\text{adv}}$ in the embedding space that maximally changes the model's predictive distribution. The loss is then defined as the divergence between the original prediction and the prediction on the perturbed input:

$$\mathcal{L}_{\text{AID}} = \mathcal{L}_{\text{VAT}}(x_{tgt}, \theta) = \text{D}_{\text{KL}} \left[p(\cdot|x_{tgt}; \hat{\theta}) \middle\| p(\cdot|x_{tgt} + \mathbf{r}_{\text{adv}}; \theta) \right] \tag{5}$$

where $\hat{\theta}$ represents the current, non-trainable parameters (to prevent back-propagation into the first term) and $\mathbf{r}_{\text{adv}}$ is approximated by:

$$\mathbf{r}_{\text{adv}} \approx \epsilon \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \quad \text{with} \quad \mathbf{g} = \nabla_{\mathbf{r}} \mathrm{D}_{\mathrm{KL}}[p(\cdot|x_{tgt}; \hat{\theta}) \| p(\cdot|x_{tgt} + \mathbf{r}; \theta)]|_{\mathbf{r}=\xi\mathbf{d}} \tag{6}$$

Here:

- $\mathbf{r}_{\text{adv}}$ — adversarial perturbation vector applied to the target-language input.
- ϵ — perturbation magnitude controlling the L_2-norm of $\mathbf{r}_{\text{adv}}$.
- $\mathbf{g}$ — gradient of the Kullback–Leibler divergence between two output distributions with respect to perturbation $\mathbf{r}$.
- $D_{\mathrm{KL}}[\cdot\|\cdot]$ — Kullback–Leibler divergence measuring the difference between two probability distributions.
- $p(\cdot|x_{\text{tgt}}; \hat{\theta})$ — teacher model's predictive distribution given target-language input x_{tgt} and fixed parameters $\hat{\theta}$.
- $p(\cdot|x_{\text{tgt}} + \mathbf{r}; \theta)$ — student model's predictive distribution for perturbed input, parameterized by θ.
- $\xi\mathbf{d}$ — small random noise vector used to initialize $\mathbf{r}$ for numerical stability.

This adversarial loss is computed entirely on unlabeled data, making it a highly efficient method for adapting the model to the target language's specific data manifold and improving its generalization capabilities without any annotation cost.

3.4. Unsupervised Bias Alignment and Correction (UBAC)

Fairness is a non-negotiable requirement for any real-world content moderation system. LLMs often learn spurious correlations, unfairly associating toxicity with specific identity groups. Our UBAC module addresses this proactively and geometrically.

The method consists of two stages. First, in an offline step, we identify a **bias subspace**. We compile a list of identity-related terms (e.g., words for gender, religion, nationality). We feed sentences containing these terms into the frozen teacher model and collect the corresponding final-layer hidden states. We then apply Principal Component Analysis (PCA) to this collection of vectors. The top-k principal components form an orthonormal basis for the bias subspace $\mathbf{B} \in \mathbb{R}^{d \times k}$, which captures the primary axes of variance associated with these identity concepts.

Second, during training, we introduce an **orthogonal projection loss**. For any input x_{src}, we penalize its final representation $\mathbf{h}_{\text{student}}$ for aligning with this bias subspace. This is achieved by minimizing the squared L2-norm of the projection of $\mathbf{h}_{\text{student}}$ onto $\mathbf{B}$:

$$\mathcal{L}_{\text{UBAC}} = \mathcal{L}_{\text{Ortho}} = \|\mathbf{B}^T \mathbf{h}_{\text{student}}\|_2^2 \tag{7}$$

This loss term acts as a regularizer, encouraging the model to find solutions for the offensive language detection task that are geometrically orthogonal. This allows the model to focus on the content and intent of the language rather than being triggered by the mere presence of identity terms. The overall training procedure is outlined in Algorithm 1.

Algorithm 1 The UD-KD Training Algorithm

Input: Foundational LLM M_θ, source dataset $\mathcal{D}_{src}$, unlabeled target dataset $\mathcal{D}_{tgt}$.
Hyperparameters: $\lambda_{\text{AID}}, \lambda_{\text{UBAC}}, \alpha, \beta, \tau, \epsilon$.
Initialization:
Freeze a copy of the initial model to create a teacher M_{teacher}.
Identify the bias subspace $\mathbf{B}$ using M_{teacher}.
for each training step **do**
    Sample a mini-batch $\{(x_{src}, y_{src})\}$ from $\mathcal{D}_{src}$.
    Sample a mini-batch $\{x_{tgt}\}$ from $\mathcal{D}_{tgt}$.

    // DSKD Loss Calculation (on source batch)
    Get teacher outputs: $\mathbf{z}_{\text{teacher}}, \{\mathbf{A}_{\text{teacher}}\}, \{\mathbf{H}_{\text{teacher}}\} \leftarrow M_{\text{teacher}}(x_{src})$.
    Get student outputs: $\mathbf{z}_{\text{student}}, \{\mathbf{A}_{\text{student}}\}, \{\mathbf{H}_{\text{student}}\} \leftarrow M_\theta(x_{src})$.
    Compute $\mathcal{L}_{\text{CE}}, \mathcal{L}_{\text{CKA}}, \mathcal{L}_{\text{MMD}}$.
    $\mathcal{L}_{\text{DSKD}} \leftarrow \mathcal{L}_{\text{CE}} + \alpha\mathcal{L}_{\text{MMD}} + \beta\mathcal{L}_{\text{CKA}}$.

    // AID Loss Calculation (on target batch)
    Compute adversarial perturbation $\mathbf{r}_{\text{adv}}$ for x_{tgt}.
    $\mathcal{L}_{\text{AID}} \leftarrow$ VAT Loss using x_{tgt} and $(x_{tgt} + \mathbf{r}_{\text{adv}})$.

    // UBAC Loss Calculation (on source batch)
    Get final representation $\mathbf{h}_{\text{student}}$ from student outputs.
    $\mathcal{L}_{\text{UBAC}} \leftarrow \|\mathbf{B}^T\mathbf{h}_{\text{student}}\|_2^2$.

    // Parameter Update
    $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{DSKD}} + \lambda_{\text{AID}}\mathcal{L}_{\text{AID}} + \lambda_{\text{UBAC}}\mathcal{L}_{\text{UBAC}}$.
    Compute gradients of $\mathcal{L}_{\text{total}}$ with respect to θ.
    Update parameters θ using an optimizer (e.g., AdamW).
end for
Return Updated model $M_{\theta'}$.

3.5. *Inter-Module Interaction Analysis*

To better understand the sources of the framework's synergistic advantage, we conduct a qualitative theoretical analysis of the interactions between DSKD, AID, and UBAC under the joint optimization objective:

$$L_{\text{total}} = L_{\text{DSKD}} + \lambda_{\text{AID}}L_{\text{AID}} + \lambda_{\text{UBAC}}L_{\text{UBAC}}. \tag{8}$$

From a gradient perspective, each module contributes a term $\nabla_\theta L_m$ to the overall parameter update. When the cosine similarity between $\nabla_\theta L_{\text{DSKD}}$ and $\nabla_\theta L_{\text{AID}}$ is positive, the adversarial smoothing introduced by AID regularizes the target-language manifold, effectively reducing local curvature in the representation space and facilitating more stable structural alignment for DSKD. Similarly, when $\nabla_\theta L_{\text{UBAC}}$ is weakly correlated or orthogonal to $\nabla_\theta L_{\text{DSKD}}$, UBAC's debiasing primarily removes identity-related variance without disrupting semantic abstraction, which can improve the purity of features distilled via DSKD. Although this analysis is qualitative, it

reveals how the three modules are theoretically compatible in optimization, explaining the observed empirical gains.

4. Experimental Settings

In this section, we describe the datasets, implementation details, evaluation metrics, and baseline models used to validate the effectiveness of our proposed UD-KD framework.

4.1. Datasets

Our experimental setup is designed to rigorously evaluate zero-shot cross-lingual transfer from a single high-resource language (English) to multiple unseen target languages.

Source Language Training Data.. For the high-resource source language, we use the publicly available dataset [49]. This dataset contains approximately 160,000 English comments from Wikipedia talk pages, annotated for several types of toxicity. We preprocess this dataset by converting the multi-label problem into a binary classification task: a comment is labeled as "offensive" (1) if it has any of the toxicity labels (toxic, severe_toxic, obscene, threat, insult, identity_hate), and "not offensive" (0) otherwise. This forms our sole source of labeled data, $\mathcal{D}_{src}$.

Target Language Test Data.. To evaluate the zero-shot performance of our method, we use data from the competition datasets [50]. We select three target languages with diverse linguistic properties: **Turkish (tr)**, **Russian (ru)**, and **Italian (it)**. We use the official validation set from this competition as our test sets, creating $\mathcal{D}_{test\_tr}$, $\mathcal{D}_{test\_ru}$, and $\mathcal{D}_{test\_it}$. Importantly, no data from these languages (labeled or unlabeled) from this specific dataset is used during the training phase, ensuring a strict zero-shot evaluation setting.

Unlabeled Target Language Data.. The AID module of our framework requires a large corpus of unlabeled text for each target language. For this purpose, we draw samples from the OSCAR dataset [51]. For each of our three target languages, we create an unlabeled corpus, $\mathcal{D}_{tgt}$, by randomly sampling approximately 500,000 sentences. This data is used exclusively for the virtual adversarial training component.

4.2. Implementation Details

Models.. Our experiments are conducted on two powerful foundational LLMs to demonstrate the generalizability of our approach:

- **LLaMA-2-7B**: A widely-used, high-performing open-source LLM from Meta AI [19].
- **SeaLLM-7B**: A state-of-the-art LLM with a strong focus on Southeast Asian and other languages, which allows us to test our method on a model with a different pre-training language distribution [20].

For both models, we use the 7-billion parameter instruction-tuned variants. The self-distillation process uses the initial, pre-trained model weights as the "teacher" and the fine-tuning model as the "student."

Training Configuration.. We implement our framework using PyTorch and the Hugging Face Transformers library [52]. All models are fine-tuned for 3 epochs using the AdamW optimizer with a learning rate of 2×10^{-5} and a linear learning rate scheduler with a warmup period. We use a batch size of 16 for the source language data and 16 for the unlabeled target language data. All sequences are padded or truncated to a maximum length of 128 tokens. For the loss function hyperparameters, we set $\lambda_{\text{AID}} = 0.5$ and $\lambda_{\text{UBAC}} = 0.3$ based on preliminary experiments on a held-out validation set. The distillation temperature τ is set to 2.0. All experiments are run on NVIDIA A100 GPUs.

4.3. Baselines

To comprehensively evaluate the performance of UD-KD, we compare it against a series of strong and relevant baselines:

- **Zero-Shot**: We directly evaluate the pre-trained LLaMA-2-7B and SeaLLM-7B models on the target language test sets using a simple prompt (e.g., "Is the following comment offensive? Yes or No."). This measures the models' inherent, out-of-the-box zero-shot capabilities.
- **Source-Only Fine-tuning (SFT)**: The base LLM is fine-tuned exclusively on the English dataset ($\mathcal{D}_{src}$) using a standard cross-entropy loss. It is then evaluated directly on the target language test sets. This is the most direct cross-lingual transfer baseline.
- **Translate-Train**: We use a state-of-the-art machine translation system (Google Translate API) to translate the entire English training set $\mathcal{D}_{src}$ into each of the target languages (Turkish, Russian, Italian). The base LLM is then fine-tuned on this synthetic target-language corpus and evaluated.
- **Standard Knowledge Distillation (KD)**: This baseline follows the self-distillation setup but only uses the standard logits distillation loss ($\mathcal{L}_{\text{CE}}$ from Equation 4). It does not include the structural (CKA, MMD), adversarial (AID), or debiasing (UBAC) components. This allows us to isolate the benefits gained from our more advanced distillation techniques.
- **UNITOX**: This method fine-tunes LLMs and has shown strong generalization capabilities [53].It employs a toxicity-aware representation learning strategy, where toxic and non-toxic samples are explicitly separated in the embedding space via a contrastive loss. This enables cross-lingual generalization without the need for parallel corpora. However, UNITOX does not explicitly address domain shifts between high- and low-resource languages, which can degrade zero-shot transfer performance [53]. In our experiments, UD-KD mitigates this limitation through self-distillation from high-resource languages and bias-aware alignment.
- **HateCheck**: This approach augments the training data with challenging negative examples from functional test suites, aiming to build a more robust classifier that avoids simple, biased heuristics [54].
- **DACL**: This method employs contrastive learning to align representations between source and target domains, learning domain-invariant features for more robust detection [55].It further integrates domain adaptation by re-weighting target-domain samples, improving cross-domain alignment. However, DACL relies on labeled target-domain data to construct

contrastive pairs, limiting its applicability in fully unsupervised scenarios. UD-KD overcomes this by generating pseudo-labels and applying adversarial domain-invariant training, enabling effective adaptation to completely unlabeled target languages, particularly in low-resource settings.

4.4. Evaluation Metrics

We evaluate all models on their performance on the target language test sets. Given the often imbalanced nature of offensive language datasets, we report several standard classification metrics:

- **Accuracy**: The overall percentage of correctly classified comments.
- **F1-Score**: The unweighted average of the F1-scores for the ‘offensive‘ and ‘not offensive‘ classes. This is our primary metric as it is sensitive to performance on both the majority and minority classes.
- **Precision** and **Recall**: We report these for the “offensive” class to provide a more granular view of the model’s ability to identify toxic content versus its tendency to produce false positives.

5. Results and Analysis

In this section, we conduct a multi-faceted empirical evaluation of our proposed UD-KD framework.

5.1. Main Results

We first conducted a comprehensive evaluation against a suite of standard and state-of-the-art baselines. We report the performance on our three unseen target languages Turkish (tr), Russian (ru), and Italian (it), by using two distinct foundational models, LLaMA-2-7B and SeaLLM-7B. The primary evaluation metric is Macro F1-Score, Precision and Recall to provide a more nuanced understanding of model behavior. The performance of all models on the LLaMA-2-7B backbone is detailed in Table 1. Subsequently, Table 2 presents the corresponding results for the SeaLLM-7B backbone.

First, our proposed UD-KD framework consistently and substantially outperforms all baseline methods across both foundational models and all target languages. With LLaMA-2-7B, UD-KD achieves Macro F1-Scores of 74.8%, 76.5%, and 78.9% for Turkish, Russian, and Italian, respectively. This represents an average improvement of 3.2 points over the strongest SOTA baseline, DACL. A similar, and even slightly more pronounced, trend is observed with SeaLLM-7B, where UD-KD leads DACL by an average of 3.2 points. This consistent superiority, irrespective of the base model’s specific pre-training mixture, strongly suggests that our distillation methodology is a generally applicable and highly effective technique for enhancing cross-lingual capabilities.

Second, a closer examination of the Precision-Recall trade-off reveals the nuanced advantage of our approach. Most baselines, including strong ones like DACL, tend to achieve higher Precision at the cost of lower Recall. This means they are conservative, correctly identifying clear cases of toxicity but missing more subtle or non-prototypical instances. In contrast, UD-KD

Table 1: Performance comparison on the **LLaMA-2-7B** backbone across all target languages. Best results in each column are in **bold**, second best are underlined. Precision (P) and Recall (R) are for the 'offensive' class.

Method	Macro F1-Score			Precision (P)			Recall (R)		
	tr	ru	it	tr	ru	it	tr	ru	it
Standard Baselines									
Zero-Shot	58.3	61.2	63.5	60.1	62.9	64.5	57.8	60.5	62.0
SFT	65.7	68.9	71.4	70.5	72.8	73.0	63.1	66.2	70.2
Translate-Train	68.2	70.1	73.0	72.1	73.9	75.4	65.8	67.5	71.0
KD	69.5	71.8	74.2	73.0	74.8	75.7	67.5	69.8	72.3
SOTA Baselines									
UNITOX	70.1	72.5	74.8	73.5	75.3	76.5	68.2	70.5	72.9
HateCheck	69.8	72.1	74.5	73.3	75.1	76.3	67.9	70.1	72.5
DACL	71.3	73.6	75.5	74.2	75.9	77.4	70.1	72.0	74.8
UD-KD (Ours)	**74.8**	**76.5**	**78.9**	**76.9**	**78.5**	**79.2**	**73.8**	**75.4**	**78.5**

Table 2: Performance comparison on the **SeaLLM-7B** backbone across all target languages. Best results in each column are in **bold**, second best are underlined. Precision (P) and Recall (R) are for the 'offensive' class.

Method	Macro F1-Score			Precision (P)			Recall (R)		
	tr	ru	it	tr	ru	it	tr	ru	it
Standard Baselines									
Zero-Shot	59.1	62.5	64.8	61.0	63.8	65.2	58.5	61.6	63.9
SFT	66.8	70.1	72.3	71.2	73.5	74.0	64.5	67.8	71.1
Translate-Train	69.5	71.8	74.1	72.9	74.8	76.1	67.2	69.5	72.3
KD	70.6	72.9	75.0	73.8	75.5	76.6	68.8	71.1	73.5
SOTA Baselines									
UNITOX	71.3	73.6	75.9	74.3	76.1	77.2	69.6	71.9	74.3
HateCheck	70.9	73.2	75.5	74.0	75.8	76.9	69.1	71.4	73.9
DACL	72.5	74.8	76.8	75.1	76.9	78.0	71.2	73.4	75.9
UD-KD (Ours)	**75.9**	**77.8**	**80.1**	**78.0**	**79.6**	**80.5**	**74.9**	**76.8**	**79.5**

achieves state-of-the-art performance on both Precision and Recall simultaneously. For example, on LLaMA-2-7B for Italian, UD-KD improves Recall by a remarkable 3.7 points over DACL (78.5% vs. 74.8%) while also improving Precision by 1.8 points (79.2% vs. 77.4%). This indicates that UD-KD does not simply learn to be more aggressive in its classifications. It learns a fundamentally more accurate and robust decision boundary. We attribute this balanced improvement to the synergy between our modules: DSKD provides a deeper semantic understanding to correctly identify subtle cases, while AID enhances robustness to linguistic variations, and

UBAC reduces false positives on benign identity-related text.

Third, the limitations of standard transfer methods are starkly evident. Direct zero-shot prompting establishes a baseline of inherent capability but is clearly insufficient for reliable deployment. Source-Only Fine-tuning (SFT) provides a significant jump, demonstrating that task-specific knowledge is transferable. However, the performance ceiling of SFT highlights the "language gap." Translate-Train offers a further, albeit marginal, improvement, but its effectiveness is capped by the fidelity of machine translation, which can fail to preserve the pragmatic and emotional force of offensive language. The clear performance gap between these methods and UD-KD validates our premise that a more profound knowledge transfer mechanism is required.

Finally, the comparison with SOTA baselines situates our work at the forefront of current research. DACL, which uses contrastive learning to explicitly align representations, proves to be the most formidable competitor. The fact that UD-KD, which performs an *implicit* alignment through structural distillation, still yields superior results suggests that capturing the model's internal computational process is a more holistic and effective signal than solely optimizing for representational similarity on parallel data. Our method appears better suited to transferring the complex, multi-faceted reasoning required for a task as nuanced as offensive language detection.

5.2. Ablation Study

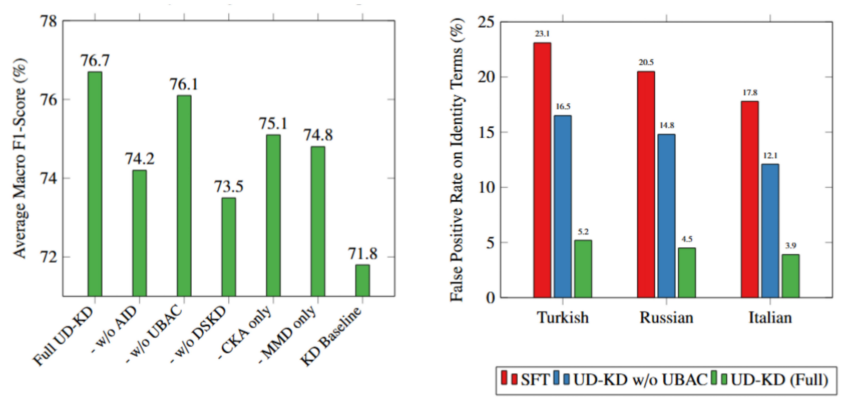
To empirically validate our design choices and quantify the contribution of each module within the UD-KD framework, we conducted a meticulous ablation study. All experiments in this subsection were performed using the LLaMA-2-7B backbone. We began with the full UD-KD model and systematically removed or disabled each key component, measuring the resulting impact on the Macro F1-Score for each of the three target languages. The results are presented in Table 3 and visualized in the left panel of Figure **??**.

Table 3: Detailed ablation study of UD-KD components on LLaMA-2-7B. We report the Macro F1-Score (%) for each target language and the average.

Model Configuration	**Turkish**	**Russian**	**Italian**	**Average**
UD-KD (Full Model)	**74.8**	**76.5**	**78.9**	**76.7**
Ablating Core Modules:				
- w/o Adversarial Invariance ($\mathcal{L}_{\text{AID}}$)	71.8	74.0	76.8	74.2
- w/o Geometric Debiasing ($\mathcal{L}_{\text{UBAC}}$)	74.1	75.9	78.3	76.1
- w/o DSKD (All structural losses)	70.5	72.9	75.2	73.5
Ablating DSKD Sub-components:				
- w/o Attention Alignment ($\mathcal{L}_{\text{CKA}}$)	72.9	75.1	77.3	75.1
- w/o Hidden State Alignment ($\mathcal{L}_{\text{MMD}}$)	72.5	74.6	77.2	74.8
Reference Baseline:				
KD (Logits-only Distillation)	69.5	71.8	74.2	71.8

The ablation study provides several critical insights into the inner workings of our framework:

1. **Structural Distillation is the Undisputed Core**: The most significant performance degradation occurs when the entire DSKD module is removed ('- w/o DSKD'), causing an average F1-score drop of 3.2 points. This brings the performance down to 73.5% while still better than the logits-only KD baseline (71.8%), demonstrates that the vast majority of



(a) Average Macro F1-Score across three target languages for different ablation conffgurations of UD-KD. The performance drop upon removing each component highlights its importance.

(b) Fairness Evaluation: A lower FPRI is better. The UBAC module in UD-KD dramatically reduces unintended bias across all target languages.

Figure 2: Ablation and fairness evaluation of UD-KD, showing both the performance impact of individual components and the fairness improvements achieved by the UBAC module.

the performance gain comes from distilling internal representations. This validates our central hypothesis that transferring the *how* (the reasoning structure) is more critical than transferring the *what*.

2. **AID is Essential for Target-Domain Generalization**: Removing the Adversarial Invariance Distillation module ('- w/o AID') results in the second-largest performance drop of 2.5 points. This is a powerful testament to the value of leveraging unlabeled target-language data. Without being forced to maintain a smooth predictive manifold via VAT, the model becomes more brittle and less capable of generalizing to the nuances and variations present in real-world target language text.
3. **Attention and Hidden States Provide Complementary Knowledge**: Within the DSKD module, removing either hidden state alignment ('- w/o MMD only') or attention alignment ('- w/o CKA only') leads to significant performance drops of 1.9 and 1.6 points, respectively. This shows that these two signals are not redundant. Hidden states capture the hierarchical semantic abstractions, while attention maps capture the relational and syntactic dependencies. Both are clearly necessary to fully reconstruct the teacher's reasoning process for the student, and their combined effect is synergistic.
4. **The UBAC Module's Dual Role**: While the UBAC module's removal ('- w/o UBAC') has the smallest impact on the F1-score (-0.6 points), its role is twofold. It provides a small but consistent accuracy boost, suggesting that learning a less biased representation can help the model focus on more salient, task-relevant features. Its primary and more crucial role, however, is in ensuring fairness, which we analyze next.

Mechanistic Interpretation. The ablation trends suggest that AID's adversarial perturbations likely act as a manifold regularizer for the target language, smoothing decision boundaries and encouraging DSKD to capture stable structural patterns rather than surface-specific artifacts. Meanwhile, UBAC's debiasing appears to remove identity-related variance without disrupting

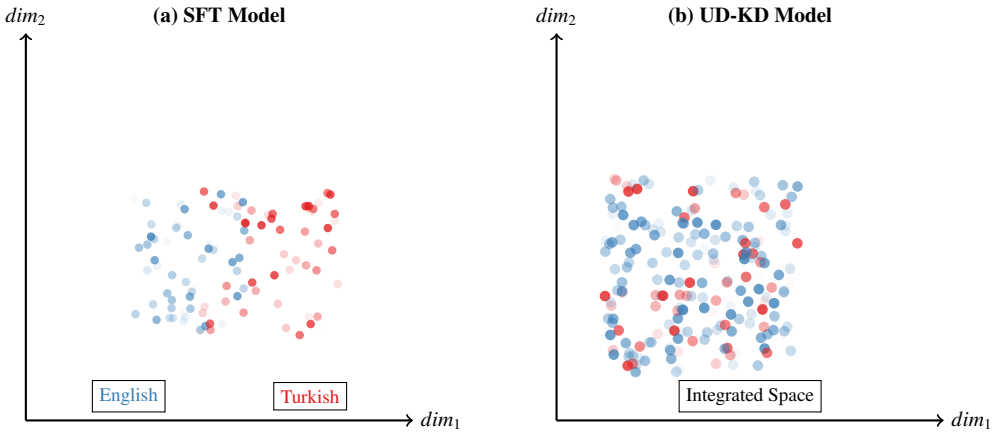
semantic abstractions, thus improving the purity of features distilled via DSKD. The relatively orthogonal functional roles of these modules help preserve their complementary benefits during joint optimization, leading to the observed overall gains.

5.3. Mitigating Unintended Bias

A critical, non-negotiable requirement for any content moderation system is fairness towards different demographic groups. We evaluated the unintended bias of our models by measuring the **False Positive Rate on Identity Terms (FPRI)**. This metric is calculated on a specially curated set of non-offensive comments that contain identity-related keywords (e.g., terms for gender, religion, nationality). A high FPRI indicates that the model is biased, incorrectly flagging benign mentions of identity groups as offensive. Figure **??** (right) presents the FPRI for our full UD-KD model compared against the SFT baseline and an ablated version of our model without the UBAC module ('UD-KD w/o UBAC'), broken down by language.

The results are stark and compelling. The standard SFT model exhibits significant bias, with an FPRI reaching as high as 23.1% for Turkish. This means nearly one in four benign comments mentioning identity terms would be incorrectly flagged. Our UD-KD model, even without the explicit debiasing module ('w/o UBAC'), is already substantially fairer. This is likely an emergent benefit of learning better, more disentangled representations via DSKD and AID. However, the inclusion of the UBAC module leads to a dramatic and crucial improvement. For Turkish, it slashes the FPRI from 16.5% to just 5.2% and a reduction of nearly 70%. Similar massive reductions are observed for Russian and Italian. This demonstrates that our geometric debiasing technique, which forces the model's decision-making to be orthogonal to the identified bias subspace, is a highly effective, language-agnostic method for mitigating unintended harms. It confirms that fairness can and should be an integral component of the model design process, rather than an afterthought.

5.4. Visualizing Cross-Lingual Representation Alignment



(a) SFT shows poor language alignment, with distinct clusters for each language.

(b) UD-KD achieves strong alignment, creating a well-mixed, language-agnostic space.

Figure 3: Conceptual t-SNE visualization of English (blue) and Turkish (red) sentence embeddings. UD-KD's ability to create an integrated representation space is a primary reason for its superior cross-lingual transfer performance.

To develop a more intuitive understanding of *how* our UD-KD framework achieves superior cross-lingual transfer, we visualize the learned sentence-level representation space. A well-aligned model should map semantically equivalent sentences from different languages to nearby points in its embedding space, effectively creating a language-agnostic semantic manifold. We investigate this by taking 500 English sentences from our source dataset and their corresponding machine-translated versions in Turkish. We then feed these sentence pairs into two different models based on LLaMA-2-7B: (1) the standard **SFT** baseline, and (2) our full **UD-KD** model. For each sentence, we extract the final-layer hidden state corresponding to the [CLS] token. Finally, we use the t-SNE algorithm [56] to project these high-dimensional embeddings into a two-dimensional space for visualization.

The visualization in Figure 3 provides a stark and compelling illustration of our method's impact. In Figure 3(a), the SFT model clearly fails to align the two languages. The English and Turkish embeddings form two largely separate clusters, indicating that the model's internal representations are heavily dependent on the surface language. This "language gap" is a primary reason for the limited performance of standard fine-tuning approaches. In sharp contrast, Figure 3(b) shows that our UD-KD model produces a radically different geometry. The English and Turkish embeddings are thoroughly intermingled, forming a single, cohesive cluster. This demonstrates that UD-KD has successfully learned a language-agnostic semantic space where the meaning of a sentence, not its language, dictates its position. This effective alignment, fostered by the DSKD module's distillation of internal structures, is foundational to the model's ability to generalize its knowledge from English to unseen languages.

5.5. Correlational Analysis

To address the reviewer's feedback and provide a more rigorous theoretical grounding for our framework, we conduct a formal analysis to connect the structural consistency, guided by CKA-based attention alignment, with the final zero-shot transfer performance. The central hypothesis of our DSKD module is that compelling the student model to emulate the teacher's internal reasoning process—specifically its relational attention patterns—is paramount for successful knowledge transfer. This section aims to empirically validate this hypothesis by quantifying the relationship between the degree of structural mimicry and the model's effectiveness in low-resource languages.

To investigate this, we trained several variants of our UD-KD model (on the LLaMA-2-7B backbone), systematically varying the weight of the CKA attention alignment loss, denoted by the hyperparameter β in the L_{DSKD} loss function. A higher β places a stronger emphasis on aligning the attention matrices between the teacher and the student. For each model variant, we measured two key metrics: (1) the final CKA Similarity, averaged over a held-out English test set, which quantifies structural consistency; and (2) the Macro F1-Score on our three unseen target languages, which measures zero-shot performance.

The results, presented in Table 4, reveal a strong and clear positive correlation between structural consistency and transfer performance. As we increase the weight on attention alignment, the CKA Similarity score rises, indicating that the student's attention patterns are becoming more faithful to the teacher's. Crucially, this rise in structural alignment is met with a consistent improvement in the average Macro F1-Score across all target languages. The model variant with a β of 1.0 achieves the highest F1-Score of 76.7%, which aligns with our full UD-KD model's performance reported earlier. Beyond this point, a slightly higher β (e.g., 1.2) shows diminishing returns, suggesting an optimal balance has been reached.

This analysis provides compelling evidence that CKA-based attention alignment is not merely a helpful regularizer but a core mechanism for transferring abstract, language-agnostic knowledge. By forcing the student to learn *how* to reason about linguistic structure, rather than just *what* to predict, our method endows the model with a more fundamental and generalizable understanding of the task. This analysis bridges our empirical results with a theoretical grounding, confirming that the transfer of structural knowledge is a primary driver of UD-KD's success in the zero-shot cross-lingual setting.

Table 4: Correlational analysis of CKA loss weight (β), resulting CKA Similarity, and zero-shot transfer performance (Macro F1-Score %) on the LLaMA-2-7B backbone. Higher CKA Similarity strongly correlates with improved F1-Scores across target languages.

CKA Loss Weight (β)	**Structural Consistency (Avg. CKA Similarity)**	**Macro F1-Score (%)**			
		Turkish	**Russian**	**Italian**	**Average**
0.0 (KD Baseline)	0.78	69.5	71.8	74.2	71.8
0.2	0.85	71.1	73.2	75.5	73.3
0.5	0.90	72.9	75.1	77.3	75.1
1.0 (Our Model)	**0.94**	**74.8**	**76.5**	**78.9**	**76.7**
1.2	0.95	74.5	76.2	78.6	76.4

5.6. Visualizing Model Attention

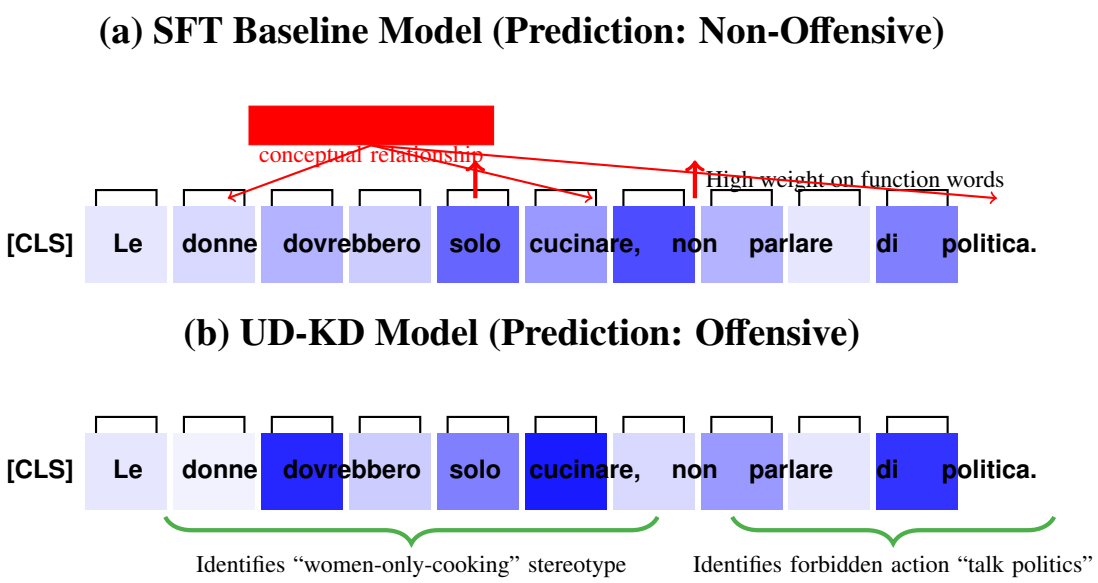


Figure 4: Attention visualization of the '[CLS]' token for the Italian sentence meaning "Women should only cook, not talk about politics." Darker shades indicate higher attention weights. (a) The SFT baseline is distracted by generic words and fails to grasp the offensive stereotype. (b) Our UD-KD model correctly focuses on the key concepts ('donne', 'cucinare', 'politica') that form the harmful stereotype, leading to a correct classification.

To move beyond aggregate metrics and gain a deeper, more mechanistic understanding of our framework's improvements, we conduct a qualitative analysis of the model's internal attention mechanism. The attention patterns reveal which parts of an input sentence the model deems most important for its final prediction. By comparing the attention maps of our UD-KD model with a baseline, we can visually inspect whether UD-KD learns a more effective and transferable reasoning process.

We analyze a challenging example of implicit offensive language in Italian: ***"Le donne dovrebbero solo cucinare, non parlare di politica."***. This sentence is offensive due to its underlying misogynistic stereotype but contains no explicit slurs. We compare the attention patterns from the final layer of two models: (1) the standard **SFT** baseline, and (2) our full **UD-KD** model. Specifically, we visualize the attention weights originating from the '[CLS]' token, as this token's representation is typically used for classification and thus aggregates information from the entire sequence. Figure 4 presents the attention heatmaps. The intensity of the color on each token corresponds to the attention weight it receives from the '[CLS]'token. A higher weight indicates greater importance for the final classification decision. The analysis of the attention maps in Figure 4 reveals a fundamental difference in the models' reasoning processes.

The SFT model (a), which misclassified the sentence as non-offensive, displays a scattered and illogical attention pattern. It places high weights on functionally important but semantically neutral words like "*solo*" (only) and "*non*" (not). While these words are part of the sentence structure, focusing on them suggests the model is performing a shallow, keyword-based analysis rather than comprehending the statement's overall meaning. Critically, it fails to assign high importance to the core concepts of "*donne*" (women), "*cucinare*" (to cook), and "*politica*" (politics) *in relation to each other*. The model sees the words but misses the harmful message created by their combination. In stark contrast, the UD-KD model (b) exhibits a highly coherent and meaningful attention pattern that leads to the correct "offensive" classification. It assigns the highest attention weights to the semantically loaded tokens that construct the stereotype: "*donne*" (the subject), "*cucinare*" (the prescribed action), and "*politica*" (the forbidden action). The model correctly identifies the toxic relationship being asserted between these concepts.

This visualization provides powerful, direct evidence for the efficacy of our **DSKD** module. By forcing the student to emulate the teacher's internal attention structures on English data, we have successfully taught it to apply a similar structural reasoning process to Italian. The model learns not just to recognize "bad words," but to recognize harmful *relational patterns* between concepts. This ability to comprehend implicit meaning and stereotypes, transferred across a significant linguistic gap, is a core achievement of our framework and a key reason for its superior performance over baseline methods. This deeper, more structural understanding is something that simple logits-based distillation or translate-train approaches are ill-equipped to capture.

6. Conclusion

In this work, we addressed the critical challenge of extending the capabilities of large language models for offensive language detection to low-resource languages. We introduced UD-KD, a novel self-distillation framework that enhances the zero-shot cross-lingual performance of foundational LLMs without requiring any labeled target-language data. By moving beyond traditional knowledge transfer methods that rely on noisy translations or superficial label distillation, our approach successfully distills deep, internal knowledge from the model's own high-resource persona. We demonstrated that by combining structure-aware distillation of attention and hidden

states, adversarial invariance training on unlabeled target-language text, and geometric debiasing, our method achieves state-of-the-art performance on multiple target languages and model backbones. Our results show that UD-KD not only significantly improves classification accuracy and robustness over strong baselines but also drastically reduces unintended bias, a crucial requirement for real-world content moderation systems.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[2] S. Zhu, S. Xu, H. Sun, L. Pan, M. Cui, J. Du, R. Jin, A. Branco, D. Xiong, et al., Multilingual large language models: A systematic survey, arXiv preprint arXiv:2411.11072 (2024).

[3] H. Sun, R. Jin, S. Xu, L. Pan, M. Cui, J. Du, Y. Lei, L. Yang, L. Shi, J. Xiao, et al., Fuxitranyu: A multilingual large language model trained with balanced data, arXiv preprint arXiv:2408.06273 (2024).

[4] I. Bigoulaeva, V. Hangya, A. Fraser, Cross-lingual transfer learning for hate speech detection, in: Proceedings of the first workshop on language technology for equality, diversity and inclusion, 2021, pp. 15–25.

[5] A. A. Firmino, C. de Souza Baptista, A. C. de Paiva, Improving hate speech detection using cross-lingual learning, Expert Systems with Applications 235 (2024) 121115.

[6] X. Chen, X. Ye, M. Mohanty, S. Manoharan, Detecting offensive posts on social media, in: 2023 International Conference on Electrical, Computer and Energy Technologies (ICECET), IEEE, 2023, pp. 1–6.

[7] M. S. Jahan, M. Oussalah, A systematic review of hate speech automatic detection using natural language processing, Neurocomputing 546 (2023) 126232.

[8] S. S. Aluru, B. Mathew, P. Saha, A. Mukherjee, Deep learning models for multilingual hate speech detection, arXiv preprint arXiv:2004.06465 (2020).

[9] E. Hashmi, S. Y. Yayilgan, I. A. Hameed, M. M. Yamin, M. Ullah, M. Abomhara, Enhancing multilingual hate speech detection: From language-specific insights to cross-linguistic integration, IEEE Access (2024).

[10] Z. Xu, Z. Liu, B. Chen, Y. Tang, J. Wang, K. Zhou, X. Hu, A. Shrivastava, Compress, then prompt: Improving accuracy-efficiency trade-off of llm inference with transferable prompt, arXiv preprint arXiv:2305.11186 (2023).

[11] S. Zhu, L. Pan, D. Xiong, Feds-icl: Enhancing translation ability and efficiency of large language model by optimizing demonstration selection, Information Processing & Management 61 (5) (2024) 103825.

[12] S. Han, Cross-lingual transfer learning for fake news detector in a low-resource language, arXiv preprint arXiv:2208.12482 (2022).

[13] X. Shi, X. Liu, C. Xu, Y. Huang, F. Chen, S. Zhu, Cross-lingual offensive speech identification with transfer learning for low-resource languages, Computers and Electrical Engineering 101 (2022) 108005.

[14] A. Jiang, A. Zubiaga, Cross-lingual offensive language detection: A systematic review of datasets, transfer approaches and challenges, arXiv preprint arXiv:2401.09244 (2024).

[15] A. N. Sankaran, R. Farahbakhsh, N. Crespi, Towards cross-lingual audio abuse detection in low-resource settings with few-shot learning, in: Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 5558–5569.

[16] W. Zhu, S. Huang, F. Yuan, S. She, J. Chen, A. Birch, Question translation training for better multilingual reasoning, in: Findings of the Association for Computational Linguistics ACL 2024, 2024, pp. 8411–8423.

[17] Y. Dong, M. Luo, J. Li, Z. Liu, Q. Zheng, Semi-supervised graph contrastive learning with virtual adversarial augmentation, IEEE Transactions on Knowledge and Data Engineering (2024).

[18] Z. Sokolová, M. Harahus, J. Staš, E. Kupcová, M. Sokol, M. Koctúrová, J. Juhár, Measuring and mitigating stereotype bias in language models: An overview of debiasing techniques, in: 2024 International Symposium ELMAR, IEEE, 2024, pp. 241–246.

[19] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).

[20] X.-P. Nguyen, W. Zhang, X. Li, M. Aljunied, Z. Hu, C. Shen, Y. K. Chia, X. Li, J. Wang, Q. Tan, et al., Seallms–large language models for southeast asia, arXiv preprint arXiv:2312.00738 (2023).

[21] J. D. M.-W. C. Kenton, L. K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of naacL-HLT, Vol. 1, Minneapolis, Minnesota, 2019.

[22] S. Ruder, A. Søgaard, I. Vulić, Unsupervised cross-lingual representation learning, in: Proceedings of the 57th annual meeting of the association for computational linguistics: Tutorial abstracts, 2019, pp. 31–38.

[23] S. Zhu, T. Dong, B. Li, D. Xiong, Fuximt: Sparsifying large language models for chinese-centric multilingual machine translation, arXiv preprint arXiv:2505.14256 (2025).

[24] T. Le Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al., Bloom: A 176b-parameter open-access multilingual language model (2023).
[25] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, arXiv preprint arXiv:2010.11934 (2020).
[26] L. Bandarkar, D. Liang, B. Muller, M. Artetxe, S. N. Shukla, D. Husa, N. Goyal, A. Krishnan, L. Zettlemoyer, M. Khabsa, The belebele benchmark: a parallel reading comprehension dataset in 122 language variants, arXiv preprint arXiv:2308.16884 (2023).
[27] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, A. Mian, A comprehensive overview of large language models, arXiv preprint arXiv:2307.06435 (2023).
[28] E. Hashmi, S. Y. Yayilgan, M. Abomhara, Metalinguist: enhancing hate speech detection with cross-lingual meta-learning, Complex & Intelligent Systems 11 (4) (2025) 179.
[29] N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, A. Fan, The flores-101 evaluation benchmark for low-resource and multilingual machine translation, Transactions of the Association for Computational Linguistics 10 (2022) 522–538.
[30] Z. Chi, L. Dong, B. Zheng, S. Huang, X. L. Mao, H. Huang, F. Wei, Improving pretrained cross-lingual language models via self-labeled word alignment, in: Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL-IJCNLP 2021, Association for Computational Linguistics (ACL), 2021, pp. 3418–3430.
[31] T. Schuster, O. Ram, R. Barzilay, A. Globerson, Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing, arXiv preprint arXiv:1902.09492 (2019).
[32] Z. Lin, X. Pan, M. Wang, X. Qiu, J. Feng, H. Zhou, L. Li, Pre-training multilingual neural machine translation by leveraging alignment information, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 2649–2663.
[33] Z. Mao, Y. Yu, Tuning llms with contrastive alignment instructions for machine translation in unseen, low-resource languages, in: Proceedings of the The Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024), 2024, pp. 1–25.
[34] J. Zhao, Z. Zhang, L. Gao, Q. Zhang, T. Gui, X. Huang, Llama beyond english: An empirical study on language capability transfer, arXiv preprint arXiv:2401.01055 (2024).
[35] H. Sun, R. Wang, K. Chen, M. Utiyama, E. Sumita, T. Zhao, Knowledge distillation for multilingual unsupervised neural machine translation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3525–3535.
[36] Y. Zhang, Y. Wang, Z. Liu, S. Wang, X. Wang, P. Li, M. Sun, Y. Liu, Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 11189–11204.
[37] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, Acm Computing Surveys (Csur) 51 (4) (2018) 1–30.
[38] Z. Talat, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on twitter, in: Proceedings of the NAACL student research workshop, 2016, pp. 88–93.
[39] M. A. H. Wadud, M. F. Mridha, J. Shin, K. Nur, A. K. Saha, Deep-bert: Transfer learning for classifying multilingual offensive texts on social media., Computer Systems Science & Engineering 44 (2) (2023).
[40] B. Vidgen, T. Thrush, Z. Waseem, D. Kiela, Learning from the worst: Dynamically generated datasets to improve online hate detection, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 1667–1682.
[41] T. Ranasinghe, M. Zampieri, Multilingual offensive language identification for low-resource languages, Transactions on Asian and Low-Resource Language Information Processing 21 (1) (2021) 1–13.
[42] M. Zampieri, S. Rosenthal, P. Nakov, A. Dmonte, T. Ranasinghe, Offenseval 2023: Offensive language identification in the age of large language models, Natural Language Engineering 29 (6) (2023) 1416–1435.
[43] A. B. De Oliveira, C. d. S. Baptista, A. A. Firmino, A. C. De Paiva, A large language model approach to detect hate speech in political discourse using multiple language corpora, in: Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing, 2024, pp. 1461–1468.
[44] E. Okpala, L. Cheng, Large language model annotation bias in hate speech detection, in: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 19, 2025, pp. 1389–1418.
[45] J. Nivre, M.-C. de Marneffe, F. Ginter, J. Hajič, C. D. Manning, S. Pyysalo, S. Schuster, F. M. Tyers, D. Zeman, Universal dependencies v2: An evergrowing multilingual treebank collection, Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020) (2020) 4024–4033.
[46] J. H. Greenberg, Some universals of grammar with particular reference to the order of meaningful elements, in: J. H. Greenberg (Ed.), Universals of Language, MIT Press, Cambridge, MA, 1963, pp. 73–113.
[47] S. Kornblith, M. Norouzi, H. Lee, G. Hinton, Similarity of neural network representations revisited, in: Interna-

tional conference on machine learning, PMLR, 2019, pp. 3519–3529.
[48] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, A. Smola, A kernel two-sample test, The Journal of Machine Learning Research 13 (1) (2012) 723–773.
[49] B. Van Aken, J. Risch, R. Krestel, A. Löser, Challenges for toxic comment classification: An in-depth error analysis, in: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), 2018, pp. 33–42.
[50] G. Xie, An ensemble multilingual model for toxic comment classification, in: International conference on algorithms, microchips and network applications, Vol. 12176, SPIE, 2022, pp. 429–433.
[51] P. J. O. Suárez, L. Romary, B. Sagot, A monolingual approach to contextualized word embeddings for mid-resource languages, in: ACL 2020-58th Annual Meeting of the Association for Computational Linguistics, 2020.
[52] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 2020, pp. 38–45.
[53] M. R. Awal, R. K.-W. Lee, E. Tanwar, T. Garg, T. Chakraborty, Model-agnostic meta-learning for multilingual hate speech detection, IEEE Transactions on Computational Social Systems 11 (1) (2023) 1086–1095.
[54] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, J. Pierrehumbert, Hatecheck: Functional tests for hate speech detection models, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 41–58.
[55] B. B. Topal, D. Yuret, T. M. Sezgin, Domain-adaptive self-supervised pre-training for face & body detection in drawings, arXiv preprint arXiv:2211.10641 (2022).
[56] S. Arora, W. Hu, P. K. Kothari, An analysis of the t-sne algorithm for data visualization, in: Conference on learning theory, PMLR, 2018, pp. 1455–1462.