Cross-community Adapter Learning (CAL) to Understand the Evolving Meanings of Norm Violation

Thiago Freitas dos Santos[1,2], **Stephen Cranefield**[3], **Bastin Tony Roy Savarimuthu**[3], **Nardine Osman**[1] and **Marco Schorlemmer**[1]
[1]Artificial Intelligence Research Institute (IIIA), CSIC, Barcelona, Catalonia, Spain
[2]Universitat Autònoma de Barcelona, Catalonia, Spain
[3]University of Otago, Dunedin, New Zealand
{thiago, nardine, marco}@iiia.csic.es, {stephen.cranefield, tony.savarimuthu}@otago.ac.nz

Abstract

Cross-community learning incorporates data from different sources to leverage task-specific solutions in a target community. This approach is particularly interesting for low-resource or newly created online communities, where data formalizing interactions between agents (community members) are limited. In such scenarios, a normative system that intends to regulate online interactions faces the challenge of continuously learning the meaning of norm violation as communities' views evolve, either with changes in the understanding of what it means to violate a norm or with the emergence of new violation classes. To address this issue, we propose the Cross-community Adapter Learning (CAL) framework, which combines adapters and transformer-based models to learn the meaning of norm violations expressed as textual sentences. Additionally, we analyze the differences in the meaning of norm violations between communities, using Integrated Gradients (IG) to understand the inner workings of our model and calculate a global relevance score that indicates the relevance of words for violation detection. Results show that cross-community learning enhances CAL's performance while explaining the differences in the meaning of norm-violating behavior based on community members' feedback. We evaluate our proposal in a small set of interaction data from Wikipedia, in which the norm prohibits hate speech.

1 Introduction

Online communities establish norms to regulate interactions between agents (community members) with diverse backgrounds and views. For example, Wikipedia's norms for article editing include the requirement to use proper writing style, avoid editing wars, and not express hate speech.[1] Adherence to these norms promotes inclusivity and coherence within the community, while violating behavior leads to exclusion and harm [McLean and Griffiths, 2019; Shmargad *et al.*, 2022]. In this context, a normative system that aims to detect norm violations must continuously learn and adapt to their changing meanings as interactions unfold, i.e., how a community understands the elements of an action that characterize detrimental behavior [Allison *et al.*, 2019].[2] This includes changes to what constitutes norm violations, such as the shift in communities' views leading to the prohibition of previously accepted terminology, or the emergence of new violation classes, such as identifying a new hate speech target.

However, we argue that in addition to identifying norm violations, normative systems must also provide explanations for the different views, including evidence on the relevant elements of an action associated with these violations and how they may differ in light of the evolving nature of online interactions. This ability is crucial because it enables community members to understand the reasons for classifying certain actions as violations, providing adherence to the transparency principle of Responsible AI [Arrieta *et al.*, 2020].

To fulfill these requirements is particularly complex considering low-resource (or newly created) online communities, where learning the meaning of norm violations occurs in a context with a limited set of actions that contain elements relevant for violation detection [Huang *et al.*, 2022]. To tackle this challenge, we investigate cross-community learning in normative system settings. This approach incorporates data from different sources (communities) to improve the performance of machine learning models in a new target community [Chandrasekharan *et al.*, 2019; Zhuang *et al.*, 2020]. Specifically, our work focuses on using data from various communities as the initial step to define violating behavior in a new task for a community with limited available data.

Some interesting approaches have been proposed to handle cross-community learning [Chandrasekharan *et al.*, 2019; Huang *et al.*, 2022], norm violation detection in online communities [Cheriyan *et al.*, 2021; Freitas dos Santos *et al.*, 2022; Subramanian *et al.*, 2022], and explanations to improve agents' interactions in a norm-violating setting [Agrawal *et al.*, 2022]. However, to the best of our knowledge, this work represents the first proposal of an interpretable adapter framework designed to learn and understand the differences in the meaning of norm violations between commu-

[1]Disclaimer: This article presents content (offensive language) that may be disturbing to different audiences.

[2]In this work, an action is defined as a text sentence, and the words are the elements that indicate the meaning of a violation.

nities. To achieve our goal, we present the Cross-community Adapter Learning (CAL) framework (detailed in Section 3), which can solve Natural Language Processing (NLP) tasks in low-resource online community contexts. CAL accomplishes this by incorporating adapters between the layers of a transformer-based model, learning the meaning of norm violations from the feedback of interacting agents (referred to as community members). Additionally, CAL employs the Integrated Gradients (IG) [Sundararajan *et al.*, 2017] algorithm to examine the distinct views manifested in the communities. The ability to analyze the communities' view changes within a specific domain or across different communities is a key contribution of our work as it describes how detrimental behavior changes over time and across domains.

CAL consists of several components, each offering a unique benefit. First, transformer-based models, specifically the Pre-Trained Language Model (PLM), provide powerful language representation to tackling NLP tasks [Vaswani *et al.*, 2017; Lin *et al.*, 2022]. Second, incorporating adapters in our framework allows for an efficient fine-tuning process that reflects the new community's view on the meaning of a norm violation while also allowing for extensibility through the dynamic creation of adapters as new violation classes emerge. Finally, the IG algorithm supports our analysis of the model to understand the composing parts of a text sentence relevant to norm violation detection.

We evaluate (Section 4) the effectiveness of CAL through a norm prohibiting hate speech. Our target task is the low-resource article editing use case from Wikipedia, while three different communities provide our source data. Rather than using pre-established classes to categorize hate speech [Fortuna and Nunes, 2018], we define the relevant classes based on the categories identified by community members (in this work, one of the authors assumes the role of community members), which may change, emerge or disappear as interactions unfold. Following this criterion, six classes of hate speech were identified: Swear, Insult and Ableism, Sexual Harassment, Racism, LGBTQIA+ Attack, and Misogyny. The sentence "this is wiki not a forum for retards" is an example of norm-violating behavior of the class "Insult and Ableism". Results (Section 5) demonstrate that CAL can learn the meaning of norm violation, adapt to evolving communities' views, and explain the differences in the meaning of norm-violating behavior for different communities based on community members' feedback.

2 Background

This section introduces incremental learning and its application to continuously fine-tune adapters to learn the meaning of norm violations. Subsequently, we describe the interpretability algorithm employed to examine how different communities define the meaning of norm-violating behavior.

2.1 Incremental Learning

As the interactions between community members unfold sequentially, two key characteristics are present in this context: the emergence of new violation classes and the evolving community view of what constitutes a norm violation. To address these characteristics, we use Incremental Learning (IL) to continuously process incoming interactions as a data stream, while discarding previous data that may contain outdated information about the meaning of a norm violation. IL offers two approaches to the problem, namely mini-batch and online learning. While mini-batch builds small data blocks to train machine learning models, online learning updates a model's parameters as soon as a new interaction instance is made available [Hoi *et al.*, 2021]. Here, we focus on mini-batch due to its stability properties and performance scores in different tasks [Li *et al.*, 2020; Freitas dos Santos *et al.*, 2022].

2.2 Adapter

Transformer-based models have been the primary approach to solving Natural Language Processing (NLP) tasks, improving on past methods and consistently attaining the best performances in several domains [Lin *et al.*, 2022]. One of the advantages of the transformer is its ability to process text sentences by reducing the amount of work in the featurization step [Qiu *et al.*, 2020]. It does that by incorporating an attention mechanism and using fully connected feed-forward neural network layers [Wolf *et al.*, 2020].

The attention layer allows the transformer model to learn the relationship between different words in a text by calculating an attention score (Equation 1). To leverage this mechanism, transformer-based models employ a multi-head strategy, with several attention heads computed in parallel.

$$Attention(Q, K, V) \leftarrow softmax(\frac{Q \odot K^T}{\sqrt{d_k}}) \odot V \quad (1)$$

Q, K, and V are matrices representing every word in a sentence (words are encoded as embeddings in a vector space). These matrices receive the same input and differ only in their learned weights, acquired by training in a large-scale dataset. d_k is used as a scaling factor. Q contains the current term of interest used by the transformer to calculate its attention score, while K and V have words that the model aims to quantify the relationships. By calculating the dot products, this mechanism aims to add contexts to the words [Vaswani *et al.*, 2017]. For instance, to differentiate the meaning of "bank" as a financial institution and "bank" of a river.

The transformer architecture previously described is the basis block for building a Pre-trained Language Model (PLM). To create a PLM, we stack several transformer blocks and initially train them on large-scale datasets [Wolf *et al.*, 2020]. As these models comprise a huge number of parameters, it would be infeasible to train them from scratch to handle new tasks. Instead, PLMs use the fine-tuning paradigm, which uses previously trained implementations and only updates its parameters for a specific task. However, this approach still requires modifying a considerable number of weights. To tackle this issue, we carry out fine-tuning by incorporating adapters between the transformer layers of a PLM. Adapters are neural networks with a small proportion (usually 3%) of the number of parameters present in the full model [Houlsby *et al.*, 2019], resulting in a faster and more

efficient training process. In this context, while we continuously update the adapter weights on our target data, the transformer layers are used only for language representation, keeping the original PLM parameters frozen.

To summarize, the advantages of adapter-based fine-tuning are threefold. First, it presents impressive results for domains where data is scarce, such as low-resource languages and communities, and cross-lingual tasks [He *et al.*, 2021]. This is especially relevant for our work, as we attempt to learn the meaning of norm violations from a small set of interaction data from low-resource communities. Second, it tackles catastrophic forgetting and interference, which are issues in fine-tuning a complete PLM [Pfeiffer *et al.*, 2020]. Third, the continuous update of smaller neural networks allows for greater robustness to handle over-fitting and reduced sensitivity to changes in learning rates [He *et al.*, 2021].

2.3 Interpretability

To understand how different communities define the meaning of norm violation (words that are associated with detrimental behavior), we incorporate interpretability into our framework, specifically, the Integrated Gradients (IG) algorithm [Sundararajan *et al.*, 2017]. IG can explain the inner workings of a transformer-based model by providing a score that encodes how relevant each word in a text sentence is for the classification output.

A word's contribution is calculated by a backward pass through the model, propagating the relevance score from the output to the input [Lyu *et al.*, 2022]. The central assumption of this algorithm is that the tokens with the highest gradient values have the most substantial influence on the classification. IG works by comparing the relevance of the input to a baseline, which is a zero embedding vector. Following the formalization in [Sundararajan *et al.*, 2017; Lyu *et al.*, 2022], let x be the sentence formed by a set of tokens $x_i, i \in 1, 2, ...n$ and $\bar{x}$ the baseline input. M is our transformer-based model, $\frac{\partial M(x)}{\partial x_i}$ is the gradient for token i, and $r(x_i)$ the calculated relevance score. To obtain $r(x_i)$, IG approximates the integral of the straight-line path from the baseline $\bar{x}$ to the input x through Equation 2. m is a finite number of points considered along the straight-line path and is chosen empirically. Thus, the integrated gradients come from the sum of these individual points.

$$r(x_i) \leftarrow (x_i - \bar{x_i}) \odot \sum_{k=1}^{m} \frac{\partial M(\bar{x} + \frac{k}{m} \times (x - \bar{x}))}{\partial x_i} \times \frac{1}{m} \quad (2)$$

3 Cross-community Adapter Learning (CAL)

In this section, we present the proposal of our work, the Cross-community Adapter Learning (CAL) framework.[3] CAL (Algorithm 1) can learn the meaning of norm violations using data from different sources (online communities). The main objective of CAL is to be deployed in a normative system to support the fulfillment of norms, especially when addressing prohibited behavior. It analyzes each action performed by community members to prevent violations from being forwarded to the entire community.

[3]Source code at https://bitbucket.org/thiago-phd/ijcai_2023/.

Algorithm 1 The Cross-community Adapter Learning (CAL) Algorithm

Input: Current time step (t), violation classes (V_t), set of all violation instances (I_t), set of augmented instances (A_t), data block size (s), set of source task adapters ($\Phi^{V_{t-1}}$), base PLM (Θ), and number of epochs (n)
Output: Trained adapters (Φ^{V_t})

1: **while** violation instances available **do**
2: **for** each violation class $v \in V_t$ **do**
3: Get current violation class instances $VI \in I_t$.
4: Get other violation instances $BI \in I_t$ to balance D_t, where $BI \cap VI = \emptyset$ and $|BI| = |VI|$.
5: Create a balanced data block. $D_t \leftarrow VI \cup BI$.
6: **if** $|D_t| < s$ **then**
7: Incorporate augmented instances. $D_t \leftarrow A_t$
8: **end if**
9: **if** $\Phi^v \in \Phi^{V_{t-1}} = NULL$ **then**
10: Create a new adapter Φ^v.
11: **else**
12: Load previously trained adapter. $\Phi^v \leftarrow \Phi^{v_{t-1}}$
13: **end if**
14: Fine-tune Φ^v on top of Θ with D_t for n epochs.
15: Add trained adapter to the list Φ^{V_t}.
16: **end for**
17: Obtain global relevance scores for adapters $\in \Phi^{V_t}$.
18: **return** fine-tuned adapters Φ^{V_t}.
19: **end while**

As norm-violating actions are made available, CAL builds balanced data blocks D_t of fixed size s (Line 5) for each violation class defined by the community $v \in V_t$. To create D_t, CAL uses $s/2$ instances of class v, while the other $s/2$ are randomly drawn from the remaining instances (Line 4).

If the size of the current balanced dataset $|D_t|$ is smaller than the fixed data block size (Line 6), we generate and augment extra violation instances (Line 7). The augmented instances are generated by modifying (substituting synonyms and random word removal) original text sentences previously identified as norm-violating behavior. If CAL is processing a newly emerged class of violation, we perturb the ground-truth data at the current time-step (t), asking for feedback from the community members. This feedback is essential because by modifying a text, we may remove the violation. However, if CAL is processing an already identified class of violation, we use the classification output of our model in the previous time-step ($t-1$) and generate the perturbed instances only from the text sentences detected as violations, asking for augmented data relabeled by the community. $t-1$ represents either the training step in a source community or in the same community but at a previous moment (past actions).

In Line 9, the algorithm checks for an existing adapter corresponding to violation class v. If v is a newly emerged violation for which no adapter has been previously created, CAL initiates a new adapter Φ^v (Line 10). The ability to dy-

Sentence	Hate Speech Class
...he was the mother fuckin dom...	Swear
...this is wiki not a forum for retards...	Insult and Ableism
[INDIVIDUAL's NAME] also sucks dick for features.	Sexual Harassment
...the big lipped,hairbraned,egotistical dirty nigger often defecated...	Racism
...HES GAYYYYYYYYYY AND HES A FREAKK...	LGBTQIA+ Attack
[INDIVIDUAL's NAME] was a super mega bitch and she kill the...	Misogyny

Table 1: Examples of sentences classified as hate speech in Wikipedia. "[INDIVIDUAL's NAME]" is used to mask real people's names.

namically generate adapters enables CAL to incorporate new classes for violations as interactions unfold and communities' views evolve. However, if a previously trained adapter $\Phi^{v_{t-1}}$ is related to v, then CAL loads it (Line 12). Each violation class v has a single associated adapter Φ^v.

CAL executes the incremental learning procedure in Line 14, updating Φ^v using D_t for n epochs. As we update the adapters, it is possible to observe the evolution of their behavior over time by calculating a global relevance score (Line 17) based on local interpretations obtained from the IG algorithm. The differences between the adapters may occur due to the application of distinct fine-tuning procedures. In this case, adapters Φ^v and $\Phi^{v_{t-1}}$ are different because they were trained within the same community but at separated moments, or Φ^v and $\Phi^{v_{t-1}}$ were trained using cross-community learning (using source data from a different community). The ability to analyze the communities' view changes within a specific domain or across different communities is a key contribution of our work. Finally, in Line 18, we return the continuously trained adapters for norm-violating detection.

4 Experiments

This section outlines the application of our incremental learning approach to detecting and understanding norm-violating behavior in a cross-community setting, specifically within the context of Wikipedia article editing. Wikipedia has a set of norms to regulate interactions during article edits, including the requirement to use proper writing style, refrain from removing content, avoid editing wars, and not express hate speech. In this research, we focus on violations of the hate speech norm, as this represents a complex and particularly harmful norm-violating behavior within online interactions.[4]

To collect interaction data for this study, we employed a two-step process. First, Wikipedia used Amazon Mechanical Turk (MTurk) to classify an article edit either as a violation or not, providing no further information on the nature of the violation. Second, one of our authors further annotated each violation instance with a violation class, introducing our view on the meaning of norm violation. It should be noted that the primary goal of this work is to develop an adaptable framework that can effectively adjust itself to the specific view of a particular online community. As such, these views may vary depending on the community in question. Table 1 shows examples of hate speech classes considered in this work.

In this study, we have sourced data from three publicly available community datasets, namely a software engineering community on Slack [Chatterjee *et al.*, 2020], the abusive language towards conversational systems (ConvAbuse) [Curry *et al.*, 2021], and a dataset built using humans and machine learning models to generate hate speech (DynamicGenerated) [Vidgen *et al.*, 2021]. Each dataset represents a unique community and includes text sentences for specific classes of hate speech. Specifically, the Slack community provides Swear instances. The ConvAbuse dataset includes Insult and Ableism, and Sexual Harassment instances. The DynamicGenerated dataset includes Racism, LGBTQIA+ Attack, and Misogyny instances. To evaluate our approach, we design the following experiments.

- ***Learn the meaning of norm violation using cross-community learning:*** We aim to evaluate whether CAL can learn the meaning of norm violation by initially using a source community to train the adapters. The number of instances for each source task is, Swear (349), Insult and Ableism (273), Sexual Harassment (456), Racism (512), LGBTQIA+ Attack (512), and Misogyny (512). Our focus is on learning from a limited number of examples of disruptive behavior. The target task consists of 639 edits, with 36,47% (233) Sexual Harassment, 33,18% (212) Insult and Ableism, 19,72% (126) Swear, 17,06% (109) LGBTQIA+ Attack, 8,76% (56) Misogyny, and 5,01% (32) Racism. To ensure that each fold of the validation process maintains a similar data distribution, we employ stratification on the multi-label dataset with the algorithm from [Sechidis *et al.*, 2011]. 2x5-fold cross-validation is used for this experiment.
- ***Learn the meaning of norm violation using only target community interactions:*** In this experiment, the goal is to evaluate the performance of our framework when trained only on our target task, with no information from outside sources (cross-community learning).
- ***Understand different communities' views on the meaning of norm violation:*** We use the local interpretability of transformer-based models to analyze the impact of words on the detection of norm violations. To do so, we first examine the words that receive high relevance scores when the model is trained on data from the source community. Next, we gather information on the relevant words when the model is incrementally fine-tuned on data from the target community. Finally, we compare the difference in the relevance score between these two steps. The change in the relevance score reveals how interactions differ between these communities and how this influences the meaning of norm-violating behavior.

[4]Future work shall focus on solving other types of violations.

Setting	Violation	Precision±Std	Recall±Std	F1±Std
Source - Target	Swear	0.5090±0.0438	0.5066±0.0355	0.3466±0.0281
	Insult and Ableism	0.5925±0.0361	0.6680±0.0725	0.6005±0.0434
	Sexual Harassment	0.7661±0.0381	0.7564±0.0397	0.7413±0.0415
	Racism	0.6022±0.0120	0.8534±0.0357	0.6038±0.0224
	LGBTQIA+ Attack	0.5963±0.0286	0.5973±0.0303	0.3978±0.0284
	Misogyny	0.5848±0.0243	0.7083±0.0588	0.5553±0.0360
Target - Target	Swear	**0.8757**±0.0261	**0.8945**±0.0197	**0.8831**±0.0203
	Insult and Ableism	**0.6937**±0.0380	**0.8478**±0.0557	**0.7236**±0.0437
	Sexual Harassment	**0.9147**±0.0281	**0.9151**±0.0279	**0.9144**±0.0280
	Racism	0.8290±0.0306	**0.9760**±0.0214	**0.8850**±0.0252
	LGBTQIA+ Attack	**0.8843**±0.0271	**0.9340**±0.0363	**0.9047**±0.0294
	Misogyny	**0.8635**±0.0657	**0.7535**±0.0618	**0.7915**±0.0573
Only Target	Swear	0.8407±0.0427	0.8701±0.0315	0.8515±0.0385
	Insult and Ableism	0.6351±0.0294	0.7642±0.0574	0.6484±0.0359
	Sexual Harassment	0.9012±0.0329	0.9005±0.0311	0.8998±0.0320
	Racism	**0.8662**±0.0886	0.7682±0.0961	0.8002±0.0893
	LGBTQIA+ Attack	0.8046±0.0441	0.8535±0.0434	0.8234±0.0446
	Misogyny	0.5470±0.2025	0.5085±0.0185	0.4884±0.0349

Table 2: Summary of the performance results (2x5-fold cross-validation) of incremental DistilBERT using adapter-based fine-tuning to evaluate hate speech detection on the Wikipedia community. Three settings are considered: 1) cross-community training and testing on our target; 2) fine-tuning on target community and testing on target; and 3) training only on target data and testing on target data.

We use DistilBERT, smaller and faster than other state-of-the-art PLM alternatives [Sanh *et al.*, 2019], with the adapter implementation by HuggingFace [Wolf *et al.*, 2020]. The data block size is 256, AdamW is the optimization algorithm, and the number of epochs is 12. Adapters have a reduction factor of 16 and ReLu as the non-linearity function. The IG algorithm was implemented following the Transformers Interpret library.[5] We use TextAttack to create the augmented instances.[6] The experiments were run on an NVIDIA GeForce GTX 1650, with 4GB memory.

5 Results and Discussion

Table 2 presents the results for each experiment. We describe performance values per hate speech class using precision, recall, and F1-score metrics. "Source - Target" refers to training on the source community and testing on the target Wikipedia interactions, with no fine-tuning. "Target - Target" is the experiment after fine-tuning the source model on our target community. Finally, "Only Target" refers to the results obtained when the model is trained and evaluated solely on the target community data. Results indicate that fine-tuning a cross-community model presents the best performance in most cases. Although directly using a model trained on a source task with no fine-tuning ("Source - Target") yields the lowest performance, it can serve as an initial point for our task, leveraging the performance of our approach after fine-tuning is applied. We use the Wilcoxon Signed-Rank Test (Table 3) to verify that, except for the Sexual Harassment class and precision for Racism, our cross-community learning approach significantly outperforms "Only Target", suggesting that our framework benefits from incorporating data from multiple communities. These exceptions are explained as follows. For Sexual Harassment, "Only Target" performs very well (above 90%), with no need to add data to improve its performance since this class is the most representative in our target community. However, Racism precision is affected by the small number of instances. Additionally, it is affected by the differences between the groups suffering this violation in the source community and the groups in our target.

Hate Speech	P-values		
	Precision	Recall	F1-Score
Swear	0.0273	0.0645	0.0273
Insult	0.0020	0.0059	0.0020
Sexual	0.0506	0.0525	0.0827
Racism	0.1934	0.0020	0.0840
LGBTQIA+	0.0020	0.0020	0.0020
Misogyny	0.0098	0.0020	0.0020

Table 3: Comparison between "Target - Target" and "Only Target". The Wilcoxon Signed-Rank Test is used to obtain the P-values. The null hypothesis is that the samples were drawn from the same distribution, and the critical value $\alpha = 0.05$.

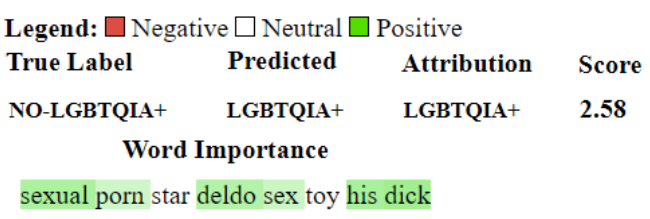


Figure 1: Local interpretation when trained on source community data. Model wrongly classifies.

[5] pypi.org/project/transformers-interpret/

[6] pypi.org/project/textattack

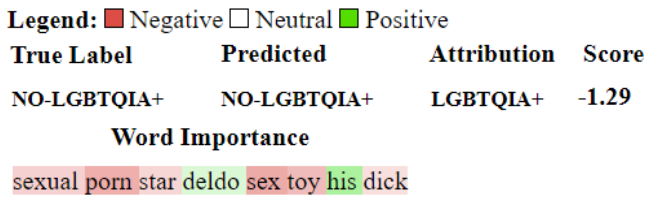


Figure 2: Local interpretation after continuously fine-tuning on target data. The model changes its behavior as expected by our new community view.

Figures 1 and 2 present local interpretability as calculated by the IG algorithm. The intensity of the green shade indicates the relevance of the highlighted word to violation detection, demonstrating what it means to violate the norm. In contrast, the intensity of the red shade is related to the decrease in the violation confidence (classification as non-LGBTQIA+ Attack). Figure 1 presents how the model trained on source community data classifies an article edit from Wikipedia. Words usually associated with Sexual Harassment content are deemed relevant to the model. However, as fine-tuning is executed, Figure 2 shows how the model drifts from the previous view about the meaning of a norm violation, resulting in negative relevance scores for the same words (hence adapting to the new community view).

Figure 3 illustrates the differences in the global relevance scores determined by CAL between communities (or at different moments in time). To obtain a word's global measure of relevance, CAL sums the relevance scores of each occurrence of the word (local interpretability), as those depicted in Figures 1 and 2. The term "Source" refers to the model trained exclusively on the source-community data, while "Target" represents the relevance score after the model is fine-tuned. The graph is organized as follows: first, we identify the 15 most relevant words for "Source" and "Target". Then, the relevant words for "Target" are displayed at the top. To demonstrate the changes in relevance scores (either increase or decrease), we include the "Source" value for the same word. Finally, we completed the rest of the rank with the words from the "Source" that had high relevance scores but saw a decrease in value after fine-tuning. For example, considering the LGBTQIA+ Attack class (Figure 3c), we can see that the word "gay" becomes more relevant as we fine-tune CAL with our target data (update the meaning of norm violation with a new community view). However, simultaneously, the words "dick", "fuck", and "sex" lose relevance. The NEGATIVE value informs us that these words are relevant for classifying an edit as a non-LGBTQIA+ Attack.

In addition to understanding different communities' views, we can also identify the factors contributing to the underperformance of the "Source - Target" task (Table 2) through the analysis of relevance scores. As an example, for the Swear class (Figure 3a), the source community data includes instances that associate Sexual Harassment and LGBTQIA+ Attack content with Swear, e.g., "gay", "penis", "sex", which are words with high relevance scores. Since the communities use parts of the violating-behavior vocabulary differently, when the "Source - Target" model attempts to solve a new task, instances containing these words are wrongly classified as Swear. However, the adaptable character of our proposal allows for updating relevance scores and improving the model's performance as interactions unfold. We can also observe this phenomenon in the Racism case, where words like "nigger", "jew", and "muslim" have higher relevance scores, while "fuck", "poo", and "shit" present a significant drop.

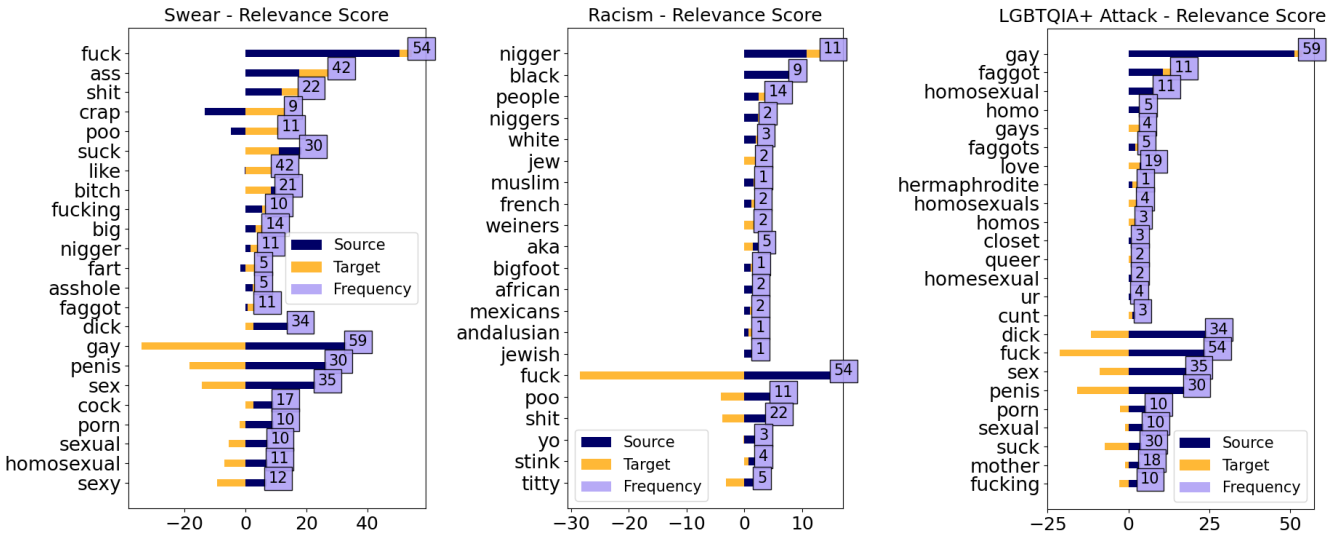
6 Literature Review

This section presents works relevant to our research. Specifically, we cite literature on cross-community learning with transformer-based models, and studies on interpretability, focusing on its application to different problem domains.

[Chandrasekharan *et al.*, 2019] presents Crossmod, which uses cross-community learning through an ensemble of classifiers to assist moderators in detecting violations within different Reddit communities. Crossmod enables moderators to oversee the decision-making process of ML models and to deal with the scarcity of labeled data. Unlike our approach, they do not focus on understanding changes in the view of communities and on incorporating adapters to handle new violation classes dynamically. [Subramanian *et al.*, 2022] creates a solution for identifying offensive comments on Youtube, specifically considering low-resource languages, which are characterized by a scarcity of labeled data and language models [Ishmam and Sharmin, 2019; Sharma *et al.*, 2022]. Like in our work, the authors use adapters to facilitate learning in this scenario. Results indicate that adapter-based fine-tuning is more effective than full fine-tuning PLMs while updating fewer parameters. Our approaches diverge in that we focus on understanding changes in the views of the community and norm-violation behavior.

While [Le *et al.*, 2021] uses adapters for multi-lingual speech translation, [Barbieri *et al.*, 2022] presents an approach that uses Twitter data from a multi-lingual setting for sentiment analysis. Both of these approaches take advantage of data in different source languages to leverage learning in a context with big datasets (the first with hundreds of hours of speech and the second with around 200M Twitter entries). This contrasts with our solution, which focuses on small datasets and low-resource communities. Results show that adapters can improve performance in a target task, even for cases considering distinct source languages. Besides NLP tasks, cross-community learning with adapters has been explored for computer vision. In [Huang *et al.*, 2022], the authors investigate the face anti-spoofing task, training adapters on diverse data sources and evaluating their ability to detect unseen instances in a new target task.

Different architectures exist for adapters. [Pfeiffer *et al.*, 2020] presents AdapterFusion, which involves learning from multiple source tasks and combining their representations via a fusion layer. This approach aims to combine multiple adapters to solve a single target task. In the future, we shall explore this architecture to handle multiple cross-community definitions for the same violation class. To optimize the parameter efficiency of adapters, the authors in [He *et al.*, 2022] introduce a pruning-based approach that reduces the number



(a) Swear (b) Racism (c) LGBTQIA+ Attack

Figure 3: Global sum of relevance score for three violation classes. Insult and Ableism, Sexual Harassment, and Misogyny are provided as supplementary material. Here we present the increase (or decrease) of the relevant words for the violation classification. We show the difference between the two communities and how norm violations are defined. The source task refers to the model trained using the source community data, while the target task is the result after fine-tuning the model with Wikipedia data.

of trainable parameters, while [Cai *et al.*, 2022] focuses on training fewer and smaller adapters at the top layers.

Interpretability has been a topic of interest in the literature [Atanasova *et al.*, 2020], with various competitions in the field [Ding and Jurgens, 2021; Salemi *et al.*, 2021] and applications ranging from simulations of machine learning models in the presence of adversarial attacks [Yang *et al.*, 2020; Alsmadi *et al.*, 2021] to task-solving solutions expressed in low-resource languages [Karim *et al.*, 2021]. In [Xiang *et al.*, 2021], the authors propose a method for enhancing the interpretability of PLMs. Unlike our approach, they calculate the relevance of each word in a given text and use max pooling to aggregate these values, obtaining the overall relevance of the entire sentence (future work shall analyze how the interpretations differ). Through a user study, the authors found that the quality of the explanations generated by their method outperformed those produced by inherently interpretable models.

The health domain also benefits from interpretability. [Novikova and Shkaruta, 2022] uses BERT to detect depression marks in text. While the authors in [Sarzynska-Wawer *et al.*, 2021] present an approach to detect objective markers of schizophrenia, showing parts of the text that are usually associated with this disorder. Both of these approaches use a perturbation method (LIME) to explain the output of a PLM. The use of interpretability provides additional information about the words usually associated with patient behavior. For instance, spiritual words are sometimes connected to non-healthy behavior, while work and professional words indicate healthy behavior.

7 Conclusion and Future Work

This paper proposes a framework, Cross-community Adapter Learning (CAL), to learn the meaning of norm violations using interaction data from different communities. Our goal is to provide the basis to work with norm violations whose definitions can change based on community members' feedback. CAL adopts a bottleneck adapter architecture on top of a transformer-based model, fine-tuned using a mini-batch approach. Additionally, we present an interpretability analysis of the cross-community adapters to understand how the meaning of norm violation varies between communities. The Integrated Gradients (IG) algorithm calculates the local relevance scores of words in text sentences, which are combined to determine their global influence in the community.

We conduct experiments within the context of Wikipedia article editing. The norm in question regulates the prohibition of hate speech. To evaluate cross-community learning, we use data from three different sources. The results show that by initially training an adapter with source community data, we can leverage the performance of our framework, demonstrating how CAL learns the meaning of norm violation and incorporates new knowledge based on a novel community view. Since interactions have evolving characteristics, we argue that the current community view is the most critical input for defining the meaning of norm violations.

Future work shall focus on user experiments to learn whether our interpretation of the model offers useful information for understanding community norms. Moreover, we aim to explore other adapter architectures to improve training efficiency in low-resource settings. Finally, we plan to integrate CAL into non-ML solutions that monitor norm violations in Normative Multi-agent Systems [Dastani *et al.*, 2018].

Acknowledgments

This research is supported by the EU funded VALAWAI (# 101070930) and WeNet (# 823783) projects, the Spanish funded VAE (# TED2021-131295B-C31) and Rhymas (# PID2020-113594RB-100) projects, and the Generalitat de Catalunya funded *Ajuts a grups de recerca de Catalunya* (# 2021 SGR 00754) project.

References

[Agrawal *et al.*, 2022] Rishabh Agrawal, Nirav Ajmeri, and Munindar P Singh. Socially intelligent genetic agents for the emergence of explicit norms. In *Proceedings of the 31st IJCAI, Vienna*, pages 1–7, 2022.

[Allison *et al.*, 2019] Kimberley R Allison, Kay Bussey, and Naomi Sweller. 'i'm going to hell for laughing at this' norms, humour, and the neutralisation of aggression in online communities. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–25, 2019.

[Alsmadi *et al.*, 2021] Izzat Alsmadi, Kashif Ahmad, Mahmoud Nazzal, Firoj Alam, Ala Al-Fuqaha, Abdallah Khreishah, and Abdulelah Algosaibi. Adversarial attacks and defenses for social network text processing applications: Techniques, challenges and future research directions. *arXiv preprint arXiv:2110.13980*, 2021.

[Arrieta *et al.*, 2020] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[Atanasova *et al.*, 2020] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 EMNLP*, pages 3256–3274, Online, November 2020. Association for Computational Linguistics.

[Barbieri *et al.*, 2022] Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, 2022.

[Cai *et al.*, 2022] Dongqi Cai, Yaozong Wu, Shangguang Wang, Felix Xiaozhu Lin, and Mengwei Xu. Autofednlp: An efficient fednlp framework. *arXiv preprint arXiv:2205.10162*, 2022.

[Chandrasekharan *et al.*, 2019] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

[Chatterjee *et al.*, 2020] Preetha Chatterjee, Kostadin Damevski, Nicholas A Kraft, and Lori Pollock. Software-related slack chats with disentangled conversations. In *Proceedings of the 17th international conference on mining software repositories*, pages 588–592, 2020.

[Cheriyan *et al.*, 2021] Jithin Cheriyan, Bastin Tony Roy Savarimuthu, and Stephen Cranefield. Towards offensive language detection and reduction in four software engineering communities. In *Evaluation and Assessment in Software Engineering*, pages 254–259. 2021.

[Curry *et al.*, 2021] Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. Convabuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational ai. *arXiv preprint arXiv:2109.09483*, 2021.

[Dastani *et al.*, 2018] Mehdi Dastani, Paolo Torroni, and Neil Yorke-Smith. Monitoring norms: A multi-disciplinary perspective. *The Knowledge Engineering Review*, 33:e25, 2018.

[Ding and Jurgens, 2021] Huiyang Ding and David Jurgens. HamiltonDinggg at SemEval-2021 task 5: Investigating toxic span detection using RoBERTa pre-training. In *Proceedings of the 15th SemEval-2021*, pages 263–269, Online, August 2021. Association for Computational Linguistics.

[Fortuna and Nunes, 2018] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30, 2018.

[Freitas dos Santos *et al.*, 2022] Thiago Freitas dos Santos, Nardine Osman, and Marco Schorlemmer. Ensemble and incremental learning for norm violation detection. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 427–435, 2022.

[He *et al.*, 2021] Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jia-Wei Low, Lidong Bing, and Luo Si. On the effectiveness of adapter-based tuning for pretrained language model adaptation. *arXiv preprint arXiv:2106.03164*, 2021.

[He *et al.*, 2022] Shwai He, Liang Ding, Daize Dong, Miao Zhang, and Dacheng Tao. Sparseadapter: An easy approach for improving the parameter-efficiency of adapters. *arXiv preprint arXiv:2210.04284*, 2022.

[Hoi *et al.*, 2021] Steven CH Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. Online learning: A comprehensive survey. *Neurocomputing*, 459:249–289, 2021.

[Houlsby *et al.*, 2019] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

[Huang *et al.*, 2022] Hsin-Ping Huang, Deqing Sun, Yaojie Liu, Wen-Sheng Chu, Taihong Xiao, Jinwei Yuan, Hartwig Adam, and Ming-Hsuan Yang. Adaptive transformers for robust few-shot cross-domain face anti-spoofing. In *Computer Vision – ECCV 2022*, page 37–54, Berlin, Heidelberg, 2022. Springer-Verlag.

[Ishmam and Sharmin, 2019] Alvi Md Ishmam and Sadia Sharmin. Hateful speech detection in public facebook pages for the bengali language. In *2019 18th ICMLA*, pages 555–560, 2019.

[Karim *et al.*, 2021] Md Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Sagor Sarker, Mehadi Hasan Menon, Kabir Hossain, Md Azam Hossain, and Stefan Decker. Deephateexplainer: Explainable hate speech detection in underresourced bengali language. In *2021 IEEE 8th DSAA*, pages 1–10. IEEE, 2021.

[Le *et al.*, 2021] Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. Lightweight adapter tuning for multilingual speech translation. *arXiv preprint arXiv:2106.01463*, 2021.

[Li *et al.*, 2020] Zeng Li, Wenchao Huang, Yan Xiong, Siqi Ren, and Tuanfei Zhu. Incremental learning imbalanced data streams with concept drift: The dynamic updated ensemble algorithm. *Knowledge-Based Systems*, 195:105694, 2020.

[Lin *et al.*, 2022] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI Open*, 2022.

[Lyu *et al.*, 2022] Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Towards faithful model explanation in nlp: A survey. *arXiv preprint arXiv:2209.11326*, 2022.

[McLean and Griffiths, 2019] Lavinia McLean and Mark D Griffiths. Female gamers' experience of online harassment and social support in online gaming: a qualitative study. *International Journal of Mental Health and Addiction*, 17(4):970–994, 2019.

[Novikova and Shkaruta, 2022] Jekaterina Novikova and Ksenia Shkaruta. Deck: Behavioral tests to improve interpretability and generalizability of bert models detecting depression from text. *arXiv preprint arXiv:2209.05286*, 2022.

[Pfeiffer *et al.*, 2020] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*, 2020.

[Qiu *et al.*, 2020] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897, 2020.

[Salemi *et al.*, 2021] Alireza Salemi, Nazanin Sabri, Emad Kebriaei, Behnam Bahrak, and Azadeh Shakery. UTNLP at SemEval-2021 task 5: A comparative analysis of toxic span detection using attention-based, named entity recognition, and ensemble models. In *Proceedings of the 15th SemEval-2021*, pages 995–1002, Online, August 2021. Association for Computational Linguistics.

[Sanh *et al.*, 2019] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[Sarzynska-Wawer *et al.*, 2021] Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michal Jarkiewicz, and Lukasz Okruszek. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135, 2021.

[Sechidis *et al.*, 2011] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. On the stratification of multi-label data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 145–158. Springer, 2011.

[Sharma *et al.*, 2022] Arushi Sharma, Anubha Kabra, and Minni Jain. Ceasing hate with moh: Hate speech detection in hindi–english code-switched language. *Information Processing & Management*, 59(1):102760, 2022.

[Shmargad *et al.*, 2022] Yotam Shmargad, Kevin Coe, Kate Kenski, and Stephen A Rains. Social norms and the dynamics of online incivility. *Social Science Computer Review*, 40(3):717–735, 2022.

[Subramanian *et al.*, 2022] Malliga Subramanian, Rahul Ponnusamy, Sean Benhur, Kogilavani Shanmugavadivel, Adhithiya Ganesan, Deepti Ravi, Gowtham Krishnan Shanmugasundaram, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. Offensive language detection in tamil youtube comments by adapters and cross-domain knowledge transfer. *Computer Speech & Language*, 76:101404, 2022.

[Sundararajan *et al.*, 2017] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Vidgen *et al.*, 2021] Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of EMNLP 2021*, pages 1667–1682, Online, 2021. Association for Computational Linguistics.

[Wolf *et al.*, 2020] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan 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*, pages 38–45, 2020.

[Xiang *et al.*, 2021] Tong Xiang, Sean MacAvaney, Eugene Yang, and Nazli Goharian. ToxCCIn: Toxic content classification with interpretability. In *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 1–12, Online, April 2021. Association for Computational Linguistics.

[Yang *et al.*, 2020] Puyudi Yang, Jianbo Chen, Cho-Jui Hsieh, Jane-Ling Wang, and Michael I Jordan. Greedy attack and gumbel attack: Generating adversarial examples for discrete data. *J. Mach. Learn. Res.*, 21(43):1–36, 2020.

[Zhuang *et al.*, 2020] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.