Can You Label Less by Using Out-of-Domain Data? Active & Transfer Learning with Few-shot Instructions

Rafal Kocielnik[*1], **Sara Kangaslahti**[*1,2], **Shrimai Prabhumoye**[2], **Meena Hari**[1], **R. Michael Alvarez**[1], **Anima Anandkumar**[1,2]
[1]California Institute of Technology, [2]NVIDIA
{rafalko@caltech.edu, skangasl@caltech.edu}

Abstract

Labeling social-media data for custom dimensions of toxicity and social bias is challenging and labor-intensive. Existing transfer and active learning approaches meant to reduce annotation effort require fine-tuning, which suffers from over-fitting to noise and can cause domain shift with small sample sizes. In this work, we propose a novel Active Transfer Few-shot Instructions (ATF) approach which requires no fine-tuning. ATF leverages the internal linguistic knowledge of pre-trained language models (PLMs) to facilitate the transfer of information from existing pre-labeled datasets (*source-domain task*) with minimum labeling effort on unlabeled target data (*target-domain task*). Our strategy can yield positive transfer achieving a mean AUC gain of 10.5% compared to no transfer with a large 22b parameter PLM. We further show that annotation of just a few *target-domain samples* via active learning can be beneficial for transfer, but the impact diminishes with more annotation effort (26% drop in gain between 100 and 2000 annotated examples). Finally, we find that not all transfer scenarios yield a positive gain, which seems related to the PLMs initial performance on the *target-domain task*.

1 Introduction

While real-world social media data is abundant, its annotation for important issues such as toxicity, hate-speech, and various facets of social bias is inherently challenging [22, 21]. These challenges stem from the sheer number of nuanced and evolving dimensions [15] as well as inherent ambiguities in interpretation [3] leading to noisy labeling [36]. Active and transfer learning approaches offer an ability to lower annotation effort by intelligently selecting the most informative examples to annotate [5] or by using existing labeled datasets [41]. However, most active learning approaches usually yield too few samples (on the order of hundreds) to feasibly fine-tune large deep-language models [34, 13]. In terms of transfer, fine-tuning on out-of-domain data can lead to detrimental domain shift [16]. Furthermore, fine-tuning can also lead to over-fitting, especially in the case of smaller train sets, and to catastrophic forgetting of knowledge present in the pre-trained model [6]. Hence, prior work mostly did not use PLMs in this setup[5, 40].

Our Approach: In this work, we propose the use of few-shot instructions (textual prompts) with PLMs as a fine-tuning-free alternative. Our approach can be effective with few samples & is robust against social-media inherent labeling noise [23], which is not well handled by fine-tuning-based approaches [30]. We propose an Active Transfer Few-shot Instructions (ATF) method for combining active learning (for selecting fewer samples to label) with transfer learning (for leveraging existing labeled datasets) under few-shot instructions with PLMs. Our method leverages the capacity of PLMs to 1) learn from a few examples in a few-shot fashion without fine-tuning and 2) transfer task

Workshop on Transfer Learning for Natural Language Processing, 2022, New Orleans.

[*]Both authors contributed equally to this research

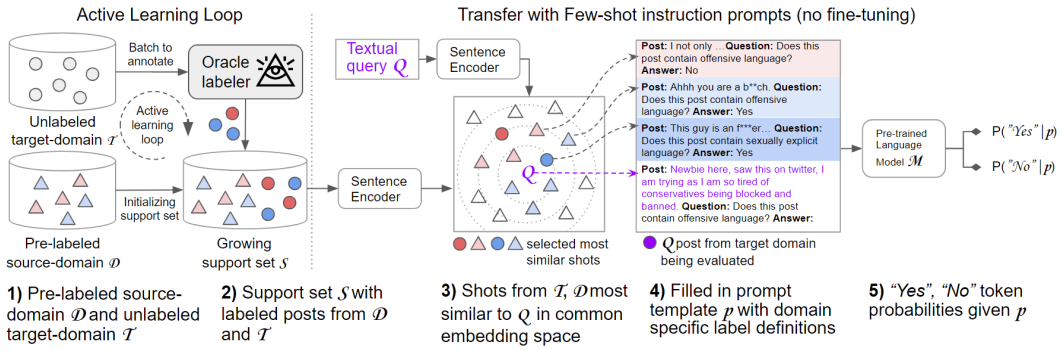

Figure 1: Overview of our approach: We populate a support set $\mathbf{S}$ with existing dataset $\mathbf{D}$ labeled under the *source-domain* definition. In an active learning loop, we further add to $\mathbf{S}$ a small number of posts labeled by an oracle (annotation effort) under the *target-domain* definition from a large unlabeled *target-domain* dataset $\mathbf{T}$. We use a sentence encoder to project the textual posts from $\mathbf{S}$ and unlabeled query $\mathbf{Q}$ to the same embedding space. We use a cosine similarity metric to select the class-balanced shots most similar to $\mathbf{Q}$. The text of the selected shots, their labels, and the labeling definitions (corresponding to the original labeling dimension) are used to fill in an instruction template $\mathbf{p}$, and finally passed to the pre-trained LM to make a prediction based on the conditional token probability for each class. The novelty of our method stems from: 1) populating the support set $\mathbf{S}$ with a mixture of source and target domain posts, and 2) communicating the original labeling definitions as part of an instruction template $\mathbf{p}$ to allow the model to relate the two.

knowledge from datasets already labeled under different definitions to further reduce the need for costly annotation. We experiment with transfer scenarios on 3 datasets across 8 labeling dimensions provided by crowd-sourcing and an existing state-of-the-art commercial tool - Perspective API [9].

Prior work: Several recent works studied the use of few-shot instructions and in-context learning for lowering annotation efforts. They, however, focused either on sample-selection strategies [32, 39] or improving few-shot performance of smaller models [35, 7, 20], but did not study transfer from existing pre-labeled datasets. Several works also employed various fine-tuning approaches in low resource settings [13, 14]. Works attempting transfer with PLMs again turn to fine-tuning in a text-to-text format [24] or attempt transfer in a few-shot setting by framing the problem as *instruction tuning*, where PLMs are fine-tuned on a collection of datasets described via instructions [37, 18]. Our work is different from all these approaches, as we focus on transfer from prelabeled external datasets via few-shot instructions without fine-tuning under a basic random active learning setting. In fact, our ATF method can be used in synergy with better sample selection strategies proposed in [32, 39].

Findings: We find that using pre-labeled *source-domain* data can help improve classification results when very few examples from the *target-domain* are labeled. We further observe two scenarios: positive and negative transfer (2). When the positive transfer occurs, it leads to high AUC gains that are consistently sustained across model sizes (12.94% for 1.3b to 10.49% for 22b) and annotation sizes (16.02% for 100 to 10.19% for 2000 annotations). Negative transfer leads to small inconsistent gains that can turn into losses with larger model size (1.91% for 1.3b to -3.30% for 8.3b). We observe that as few as 100 target domain annotations can aid transfer increasing mean gain from 3.64% to 6.73%. However, the transfer gain diminishes with more labeled examples from the *target-domain* (at the expense of annotation effort), falling from 6.73% to 4.97% (Figure 3). Finally, we investigate the reasons behind positive and negative transfers and find that the higher the AUC the PLM can achieve with only *target-domain* data, the less gain can be expected from the transfer (ρ^*=-0.66).

Contributions: In this work we offer the following contributions:

- Novel adaptation of few-shot instructions to facilitate transfer learning without fine-tuning on PLMs under limited annotation resources - Active Transfer Few-shot Instructions (ATF).
- Insights into the reasons for negative & positive transfers when attempting transfer learning with few-shot instructions and no fine-tuning.

*Pearson correlation coefficient

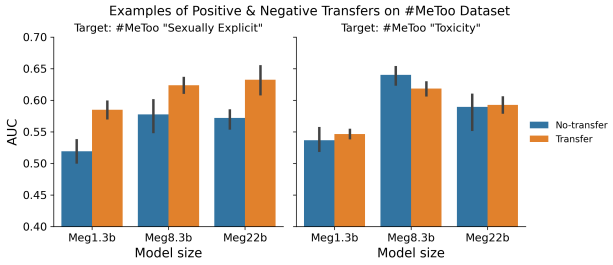

Figure 2: Example of positive and negative transfer from existing datasets (SBIC & HASOC) with 6 pre-labeled source dimensions (Table 2) to Perspective API labeled dimensions ("Sexually Explicit" and "Toxicity") on a real-world dataset (#MeToo). Positive transfer is retained across model sizes, while negative transfer starts as a minor gain that turns into a loss as model size increases.

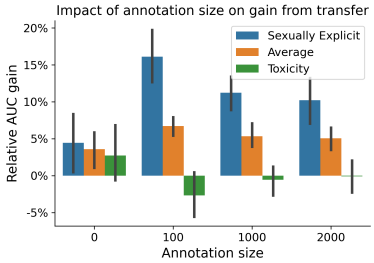

Figure 3: Relative averaged impact of target annotation size on transfer effectiveness. We can see that annotating small number of target domain samples via active learning can benefit transfer, but the relative gain diminishes as more target samples are annotated.

2 Methodology

Few-shot Instructions: We adapt the few-shot prompting approach detailed in [23] as shown in Figure 1; that is, given a query post from the test set, we present PLM with an input in the form $\mathbf{p} = [\text{“}\texttt{Post}\text{ :”}; \mathbf{Q}; \text{“}\texttt{Question}\text{ :”}; \mathbf{d}; \text{“}\texttt{Answer}\text{ :”}]$, where we concatenate the tags *Post*, *Question* and *Answer*. We calculate the probabilities of the tokens *Yes* and *No* in the following manner: $p_{\mathcal{M}}(\text{“}\texttt{Yes}\text{”}|\mathbf{p})$ and $p_{\mathcal{M}}(\text{“}\texttt{No}\text{”}|\mathbf{p})$. The token that has the higher probability is considered the prediction for the post. We define a *support set* $\mathbf{S}$ as a set of labeled examples (which contains a mix of source and target domain posts) from which the shots for few-shot instructions are selected. To select the 32 shots (same as in [23]), we sample class balanced exemplars from $\mathbf{S}$ based on their semantic similarity with the query post $\mathbf{Q}$, computed using cosine similarity in the common embedding space encoded using Term Frequency-Inverse Document Frequency (TF-IDF) representation [28]. We present the PLM with these exemplars as in-context samples using the same structure as $\mathbf{p}$. To adapt this approach to the transfer learning setting, we present the shots with the definition under which they were originally labeled (Table 2).

Transfer learning setup: For the baselines, the support set $\mathbf{S}$ consists of only a few target domain examples selected from a large unlabeled target domain dataset $\mathbf{T}$ and labeled by an oracle (Figure 1). However, for the transfer learning experiments, we augment these target domain examples with the entirety of the pre-labeled source domain training dataset $\mathbf{D}$. Thus, shots are selected from this augmented support set $\mathbf{S}$.

Active learning setup: We utilize an active learning setup to evaluate the performance of the model with varying target domain annotation sizes (Figure 1). To do so, we use the unlabeled pool scenario, in which we have a small amount of labeled target domain data and a large unlabeled target domain dataset $\mathbf{T}$. We simulate this scenario by first randomly sampling 100 examples from $\mathbf{T}$ to be labeled by an oracle. In subsequent iterations, we randomly sample from the remaining unlabeled data in $\mathbf{T}$ to provide the model with *support set* $\mathbf{S}$ with 1000 and 2000 labeled target domain examples. As we are randomly sampling from a large dataset, we repeat the entire pipeline five times for each experiment to ensure that our results are stable. We fix the random seed for each iteration of the pipeline for reproducibility and consistency across experiments.

Models: For each task, we use off-the-shelf PLMs. We utilize the Megatron 1.3B parameter model (Meg1.3b), Megatron 8.3B parameter model (Meg8.3b), and Megatron 22B parameter model (Meg22b), all of which have been pre-trained using the toolkit in [29].

Metrics: We evaluate the performance of the model using the area under the curve (AUC) [27], which measures how well a classifier can discriminate between classes, as the tasks we consider are all binary but have varying percentages of positive ("Yes") labels.

Source	Target	Megatron1.3b			Megatron22b		
Annotation size		AUC@100	@1k	@2k	AUC@100	@1k	@2k
None	MeToo	54.0	49.5	53.4	57.7	57.6	58.7
SBIC "Lewd"	"Sexually	↑17% 58.2	↑5% 55.9	↑18% 59.8	↑7% 61.9	↑9% 62.9	↑10% 64.7
SBIC "Group"	Explicit"	↑25% 61.7	↑12% 59.8	↑6% 53.8	↑17% 67.6	↑10% 63.0	↑8% 63.6
SBIC "Intent"		↑24% 61.3	↑13% 60.3	↑17% 59.5	↑6% 61.3	↑10% 63.6	↑15% 67.3
SBIC "Offensive"		↑26% 62.5	↑16% 61.8	↑14% 57.8	↑20% 69.0	↑20% 69.1	↑14% 67.1
HASOC "HOF"		↑23% 61.0	↑5% 56.6	↑18% 60.1	↑22% 70.3	↑18% 67.8	↑12% 66.0
HASOC "Target"		↑28% 63.2	↑17% 62.3	↑22% 61.9	↑12% 64.5	↑17% 67.6	↑9% 64.0
None	MeToo	51.5	53.5	53.3	61.0	60.8	60.5
SBIC "Lewd"	"Toxicity"	↑11% 57.1	↑0% 53.6	↑4% 55.4	↓5% 57.7	↓1% 59.9	↓3% 58.7
SBIC "Group"		↓1% 51.2	↑1% 54.1	↓2% 52.3	↓7% 56.5	↓5% 57.9	↓5% 57.4
SBIC "Intent"		↑6% 54.8	↑2% 54.4	↑0% 53.6	↓6% 57.5	↓12% 53.6	↓8% 54.8
SBIC "Offensive"		↑7% 55.3	↓1% 53.1	↓0% 53.2	↓3% 59.2	↓5% 57.8	↓3% 58.4
HASOC "HOF"		↑8% 55.7	↓0% 53.2	↑7% 56.9	↓6% 57.4	↑2% 62.3	↑5% 63.4
HASOC "Target"		↑4% 53.8	↑2% 54.5	↑6% 56.3	↓6% 57.3	↑3% 62.4	↑8% 65.2

Table 1: Absolute AUC and relative gain or loss compared to no transfer for 2 transfer scenarios with targets of #MeToo dataset dimensions "Sexually Explicit" and "Toxicity" as labeled by Perspective API. The sources are all pre-labeled dimensions provided with SBIC & HASOC datasets. Target "Sexually Explicit" represents a positive transfer, where gains are sustained across the annotation set and model sizes. Target "Toxicity" represents a negative transfer scenario, where gains are small and inconsistent in the smaller model and quickly turn into losses with a larger model.

Perspective API labeling: To perform transfer experiments in a controlled manner, we obtained an external and consistent set of labels related to hate speech and toxicity for all our data. We used Perspective API, a state-of-the-art pretrained toolkit [9], which has been used in [26, 12]. While Perspective API has been reported to have limitations in relation to biased classification and limited labeling dimensions [1], it still represents a good off-the-shelf baseline.

3 Datasets and Results

3.1 Datasets

We use three datasets and a total of eight labeling dimensions for our experiments: SBIC [25], HASOC [17] and #MeToo [31]. We report correlations between labeling dimensions for these datasets in Figure 6, Appendix A and an estimate of the distributional difference between them in Figure 7, Appendix B.

Social Bias Frames (SBIC): This dataset [25] contains 34k documents in the training set labeled under categories in which people project social biases and stereotypes onto others. We use four binary classification tasks, which were all labeled by crowd-workers. These tasks have the following labels and definitions: (1) offensive (57.5% positive labels): whether a post could be considered "offensive" to anyone, (2) intent (53.1% positive): whether the perceived motivation of the author was to offend, (3) lewd (9.6% positive): whether a post contains sexual references, (4) group (41.1% positive): whether a post is offensive toward a group. We also use Perspective API to label: (1) toxicity (39.1% positive): whether a post is rude, disrespectful, or unreasonable and likely to make people leave a discussion and (2) sexually explicit (22.3% positive): whether a post contains references to sexual acts, body parts, or other lewd content.

Hate Speech and Offensive Content Identification (HASOC): The HASOC dataset [17] contains documents from Twitter and Facebook, which were developed for identifying hate speech and offensive content. The dataset contains documents in three languages, but we use only the English tasks, which consist of 6k documents. We utilize the two binary classification tasks in the English dataset, which is human-labeled. The tasks are defined as follows: (1) HOF (25.0% positive): whether a post contains hate, offensive, or profane content, (2) Target (21.3% positive) whether a post contains an insult (targeted or untargeted). We also label this dataset under the toxicity (25.6% positive) and sexually explicit (16.9% positive) Perspective API tasks.

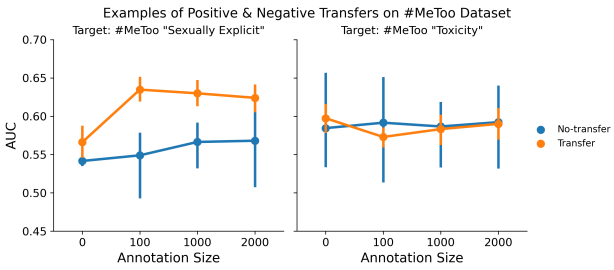

Figure 4: Comparison between transfer and no-transfer scenarios across the target domain annotation sizes split by positive (#MeToo "Sexually Explicit") and negative (#MeToo "Toxicity") transfer scenarios. We can see that in the positive transfer scenario, the gain from transfer occurs at ever annotation size, but seem optimal with as few as 100 annotated target samples after which id diminishes.

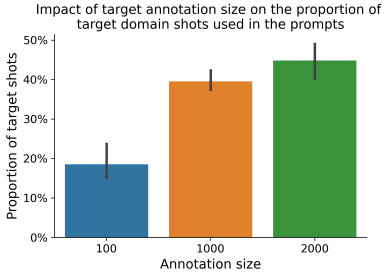

Figure 5: Impact of target domain annotation size on the proportion of target examples selected as shots for few-shot prompts. We can see that this proportion increases as target domain examples are more likely to be picked as shot by semantic similarity metric.

#MeToo Twitter Dataset: The #MeToo women's rights movement became popular very quickly on Twitter after gaining exposure from a tweet by Alyssa Milano in October 2017. There were many anecdotal stories that women who participated in the #MeToo movement on Twitter were subjected to harassment and trolling. Thus, in this paper we use Twitter data from the #MeToo movement to investigate toxicity and harassment directed at movement participants. We utilize a dataset created by collecting the tweets from January 2017 to September 2019 that contain #MeToo related keywords [31]. This #MeToo dataset consists of 7.55 million documents after preprocessing. We label this dataset using the toxicity (18.8% positive) and sexually explicit (18.6% positive) dimensions from Perspective API.

3.2 Results

Transfer Effectiveness: The results of transfer experiments with two model sizes Meg1.3b and Meg22b for two target dimensions from the #MeToo dataset ("Sexually Explicit" and "Toxicity") and 6 source dimensions from SBIC and HASOC are presented in Table 1. All the results are presented as absolute AUC scores under growing target annotation size. Next to each AUC score we show the relative gain or loss compared to the no-transfer baseline using only the annotated target samples. We also used an intermediate-sized model (Meg8.3b) and target annotation size of 0, which are not shown in the table, but averaged results are presented in Figure 2 across model sizes and in Figure 4 across annotation sizes. We include these results in our summaries.

The examples represent positive and negative transfer scenarios. In the positive transfer (#MeToo "Sexually Explicit"), it can be seen that the gains are sustained across model sizes (Figure 2) as well as across target domain annotation sizes (Figure 4). Furthermore, the gains from different source dimensions do not vary much, with the lowest average relative AUC gain of 6.74% for SBIC "Group" and the highest of 13.46% for HASOC "Target". For the negative transfer (#MeToo "Toxicity"), the impact with the smallest model (Meg1.3b) is mixed, with transfer from SBIC "Lewd" offering a small mean gain of 4.67%, while transfer from SBIC "Group" results in a small mean loss of -1.30%. It is worth noting that these two annotation dimensions are the least correlated on SBIC dataset (r=0.10). This mixed impact turns into mean loss of -3.30% with an intermediate size model (Meg8.3b) and a minor gain of 0.94% with the largest model (Meg22b). The transfer impact for Meg22b varies between -3.91% loss for SBIC "Intent" to 4.74% gain for HASOC "HOF". Comparing the two scenarios, we can also see that the initial baseline performance of the models is consistently higher for the negative transfer scenario (mean AUC of 58.9) than for the positive one (mean AUC of 55.6).

Active Learning Effectiveness: Looking at average relative gain across annotation sizes in Figure 3, we can see that without any annotated target samples (i.e., no active learning), the gains from transfer are small, but appear in both scenarios (4.6% for "Sexually Explicit" and 2.7% for "Toxicity"). Annotation of just 100 target samples differentiates the scenarios leading to big average gain of 16.0% for positive transfer and to a small average loss of -2.6% for the negative transfer. Looking at absolute AUCs in Figure 4 we can observe that mixing small number of target domain samples

within transfer regime, can lead to large AUC gain from 56.6 to 63.5 (12.1%). This is comparing transfer without any target annotations and with just 100 annotations respectively. We further observe that as annotation size increases, the relative gain from using external data decreases by 26.10% from 100 to 2k annotated target samples (Figure 3). The largest drop (20.48% decrease in relative AUC gain) takes place between 100 and 1k annotated examples, which is also a 10-fold increase in the size of annotated target data. Annotation of additional 1k examples, representing just a 2-fold increase, leads to a much smaller impact (7.07% decrease in gain). We can also see that higher proportion of target-domain samples are used as shots as annotation size increases 5. Finally, we observe that active learning alone provides small gains from AUC of 54.3 for zero-shot to 56.8 for 2k target annotations (relative gain of 4.9%) for "Sexually Explicit" and form 58.5 for zero-shot to 59.2 for 2k annotations (relative gain of 1.3%) for "Toxicity".

The main takeaways from these results are that: 1) if the positive or negative transfer occurs, it is retained across model and target annotation sizes, 2) the higher initial baseline AUC for the models likely contributes to the negative transfer, 3) transfer effectiveness can increase with small target domain annotation size, but diminishes with an increasing number of annotations.

Correlations between datasets and labeling properties: We perform additional analysis to understand the nature of positive and negative transfers. First, we examine whether the sheer amount of external data from the source domain impacts transfer effectiveness. We find that the smaller HASOC dataset (6k) actually offers a higher mean gain of 7.54% compared to a much larger SBIC (34k) offering a mean gain of 4.76% in the same setup. It is worth noting that we add these to our support set in their entirety, but the TF-IDF shot selection still picks the most relevant examples from this pool. We find that the difference in label imbalance between the source and target datasets is not correlated with AUC gain from the transfer (ρ=0.14). We also find that correlation between source and target labeling dimensions estimated on the source dataset (i.e., SBIC or HASOC) is only weakly related to AUC gain (ρ=-0.27). We find, however, that the higher the initial performance of the PLM with a given annotation size (i.e., without source domain data) the lower the AUC gain from the transfer (ρ=-0.66). Finally, we estimate the distributional difference wrt. labels between the datasets following the approach from [40]. We train an SVM classifier to tell datasets apart under the aligned labels (i.e., positive class posts put into the same set) from source and target domain tasks (separability in Figure 7), but we find only a weak correlation to the AUC gain (ρ=0.25).

4 Discussion and Future Work

Impact of model size: We observe that as the model size increases, the gains in positive transfer tend to decrease only slightly (2.45% gap between gain from Meg1.3b and Meg22b) and the overall effectiveness of transfer is largely retained (Figure 2). In a negative transfer scenario, however, small gains can be inconsistent and turn into losses (1.91% gain in Meg1.3b, -3.3% loss for Meg8.3b, and a 0.9% gain for Meg22b). It is well documented that as the model size increases its capabilities on standard NLP tasks tend to increase [19]. While the better performance and less need for external data of larger models are not surprising, the difference in the performance for different tasks may suggest that larger models may not gain capabilities uniformly (i.e., a large model may become much better at detecting "Toxicity", but improves only slightly in detecting "Sexually Explicit" content).

Impact of annotation size: As reported in the results, as annotation size increases, the relative gain from transfer decreases (Figure 3). The decrease in gain is due to the support set increasingly containing a higher proportion of target examples. Hence, these target examples are more likely to be used as shots as can be seen in Figure 5. In effect, the performance will approach the baseline (where all the shots are from the target domain). In our shot selection, we are currently not controlling for the proportion of source and target domain documents being used (i.e., we only balance labels). An additional set of experiments could explore label and domain-balanced shots selection, which could mitigate this behavior.

Understanding positive & negative transfers: Our results suggest that negative transfer is more likely to happen if 1) the initial PLM baseline on that task is higher and 2) the source dataset supplies examples that provide little new information on top of the already used target data. The first reason is intuitive, as a higher baseline is harder to beat. The initial high baseline also reflects how well the PLMs internal knowledge already informs the target-domain task. The second finding is currently

only anecdotal (correlations are weak) and much less intuitive. It should also be interpreted within the space of datasets used for our experiments. Taken at face value, it suggests that the more different the data, the higher the gain from the transfer, which is unlikely to be true. While our datasets and labels are different at the task level (i.e., "Lewd" content is likely slightly different than "Sexually Explicit" content), they also represent a similar broader domain of hate speech, toxicity, and stereotypical bias on social media. In that sense, they come from a similar domain and capture similar tasks (we report label similarities in Appendix A and distributional differences in Appendix B). In this interpretation, we are likely observing the benefits of diversity and novelty of external data used for shots within the broader related domain, similar to the benefits of domain-adaptive pretraining [10]. Future work should examine using source datasets coming from an entirely different broader domain (e.g., Enron email dataset [4]), which are unlikely to lead to positive transfer.

Limitations & Practical application: One limitation of our work is that the datasets we use rely on untrained crowd-sourced labeling which can be noisy and based on personal biases and perceptions [2]. Perspective API labeling has known limitations of its own [1]. Furthermore, PLMs can be biased and toxic themselves when prompted [8], which also likely allows them to detect these dimensions based on their internal knowledge [26]. Our proposed method can, unfortunately, be misused intentionally or unintentionally [38]. We specifically see the dangers of using our approach for censorship [33]. Some future applications of our ATF method involve noisy pre-labeling of unlabeled datasets and selecting samples for future fine-tuning (e.g., via disagreement-based active learning [11]). With some limited initial human labeling of as few as 100 random documents, if the baseline few-shot performance is poor, using prelabeled out-of-domain data can improve the AUC without expending more human annotation effort. We also plan to use this method to efficiently label custom dimensions of toxicity relevant to #MeToo and other real-world data, which are currently not supported by tools such as Perspective API.

5 Conclusion

In this paper, we present ATF, a novel adaptation of few-shot instructions to facilitate transfer learning without fine-tuning on PLMs in a setting with limited labeling resources. We demonstrate that our method can lead to consistently high AUC gains across model and annotation sizes with a small amount of annotated data from the target dimension. We also observe positive and negative transfer scenarios and find that higher AUC of PLM without any pre-annotated source domain data is correlated with less gain in AUC from the transfer. Our results motivate future work in understanding when ATF is useful and how it can be improved, as well as practical applications including noisy pre-labeling and sample selection for fine-tuning.

Acknowledgments and Disclosure of Funding

We would like to thank the Caltech SURF program for contributing to the funding of this project and especially the named donor Carolyn Ash. This material is based upon work supported by the National Science Foundation under Grant # 2030859 to the Computing Research Association for the CIFellows Project. Anima Anandkumar is partially supported by Bren Named Chair Professorship at Caltech and is a paid employee of Nvidia. Sara Kangaslahti was a paid part-time intern at Nvidia during this project.

References

[1] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

[2] R. Binns, M. Veale, M. V. Kleek, and N. Shadbolt. Like trainer, like bot? inheritance of bias in algorithmic content moderation. In *International conference on social informatics*, pages 405–415. Springer, 2017.

[3] N.-C. Chen, M. Drouhard, R. Kocielnik, J. Suh, and C. R. Aragon. Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–20, 2018.

[4] CMU. Enron email dataset. https://www.cs.cmu.edu/~./enron/. (Accessed on 09/19/2022).

[5] P. Farinneya, M. M. A. Pour, S. Hamidian, and M. Diab. Active learning for rumor identification on social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4556–4565, 2021.

[6] Z. Fatemi, C. Xing, W. Liu, and C. Xiong. Improving gender fairness of pre-trained language models without catastrophic forgetting. *arXiv preprint arXiv:2110.05367*, 2021.

[7] T. Gao, A. Fisch, and D. Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020.

[8] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

[9] Google and Jigsaw. Perspective api - attributes and languages. https://support.perspectiveapi.com/s/about-the-api-attributes-and-languages. (Accessed on 09/16/2022).

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

[11] S. Hanneke et al. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014.

[12] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*, 2022.

[13] J. Kasai, K. Qian, S. Gurajada, Y. Li, and L. Popa. Low-resource deep entity resolution with transfer and active learning. *arXiv preprint arXiv:1906.08042*, 2019.

[14] D.-H. Lee, M. Agarwal, A. Kadakia, J. Pujara, and X. Ren. Good examples make a faster learner: Simple demonstration-based learning for low-resource ner. *arXiv preprint arXiv:2110.08454*, 2021.

[15] A. Liu, M. Srikanth, N. Adams-Cohen, R. M. Alvarez, and A. Anandkumar. Finding social media trolls: Dynamic keyword selection methods for rapidly-evolving online debates. *arXiv preprint arXiv:1911.05332*, 2019.

[16] X. Ma, P. Xu, Z. Wang, R. Nallapati, and B. Xiang. Domain adaptation with bert-based domain classification and data selection. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 76–83, 2019.

[17] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, and A. Patel. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA, 2019. Association for Computing Machinery.

[18] S. Min, M. Lewis, L. Zettlemoyer, and H. Hajishirzi. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*, 2021.

[19] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

[20] S. Mishra, D. Khashabi, C. Baral, Y. Choi, and H. Hajishirzi. Reframing instructional prompts to gptk's language. *arXiv preprint arXiv:2109.07830*, 2021.

[21] S. Modha, T. Mandl, P. Majumder, and D. Patel. Tracking hate in social media: evaluation, challenges and approaches. *SN Computer Science*, 1(2):1–16, 2020.

[22] D. Patton, P. Blandfort, W. Frey, M. Gaskell, and S. Karaman. Annotating social media data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators. *Proceedings of the 52nd Hawaii International Conference on System Sciences | 2019*, 2019.

[23] S. Prabhumoye, R. Kocielnik, M. Shoeybi, A. Anandkumar, and B. Catanzaro. Few-shot instruction prompts for pretrained language models to detect social biases. *arXiv preprint arXiv:2112.07868*, 2021.

[24] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

[25] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, and Y. Choi. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online, July 2020. Association for Computational Linguistics.

[26] T. Schick, S. Udupa, and H. Schütze. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424, 2021.

[27] Scikit-learn. Roc-auc-score. `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html`, 3 2022. (Accessed on 04/13/2022).

[28] Scikit-learn. Tfidfvectorizer. `https://tinyurl.com/scikit-tfidf`, 3 2022. (Accessed on 04/13/2022).

[29] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

[30] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[31] M. Srikanth, A. Liu, N. Adams-Cohen, J. Cao, R. M. Alvarez, and A. Anandkumar. Dynamic social media monitoring for fast-evolving online discussions. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3576–3584, 2021.

[32] H. Su, J. Kasai, C. H. Wu, W. Shi, T. Wang, J. Xin, R. Zhang, M. Ostendorf, L. Zettlemoyer, N. A. Smith, et al. Selective annotation makes language models better few-shot learners. *arXiv preprint arXiv:2209.01975*, 2022.

[33] S. Ullmann and M. Tomalin. Quarantining online hate speech: technical and ethical perspectives. *Ethics and Information Technology*, 22(1):69–80, 2020.

[34] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2016.

[35] S. Wang, H. Fang, M. Khabsa, H. Mao, and H. Ma. Entailment as few-shot learner. *arXiv preprint arXiv:2104.14690*, 2021.

[36] Y. Wang, Y. Rao, X. Zhan, H. Chen, M. Luo, and J. Yin. Sentiment and emotion classification over noisy labels. *Knowledge-Based Systems*, 111:207–216, 2016.

[37] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[38] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

[39] Y. Yu, R. Zhang, R. Xu, J. Zhang, J. Shen, and C. Zhang. Cold-start data selection for few-shot language model fine-tuning: A prompt-based uncertainty propagation approach. *arXiv preprint arXiv:2209.06995*, 2022.

[40] E. Zhao, A. Liu, A. Anandkumar, and Y. Yue. Active learning under label shift. In *International Conference on Artificial Intelligence and Statistics*, pages 3412–3420. PMLR, 2021.

[41] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

A Appendix - Correlations between labels on #MeToo, SBIC and HASOC datasets

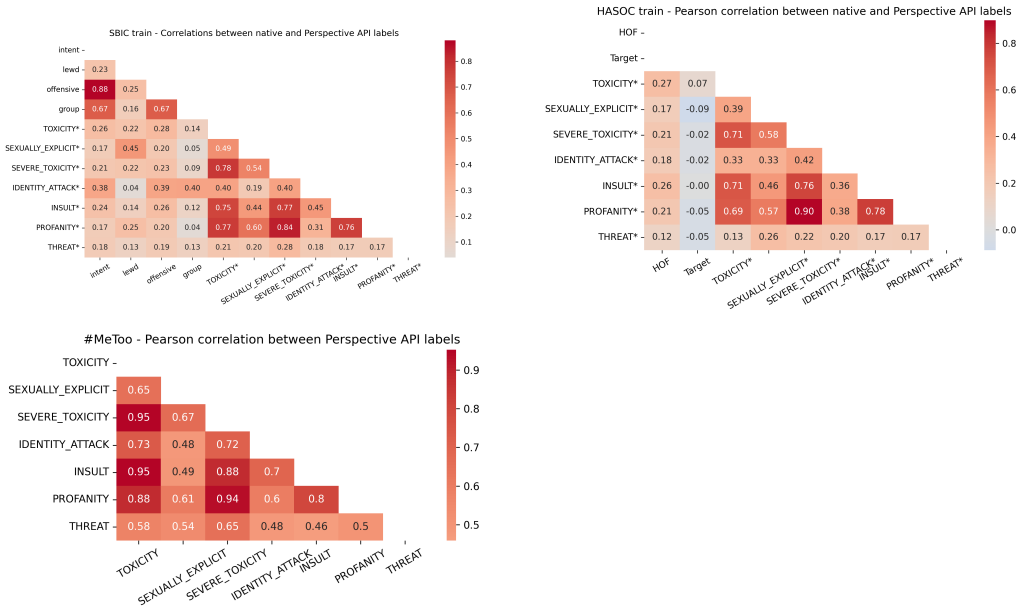

Figure 6: Pearson Correlations between native labels and Perspective API labels on #MeToo, SBIC, and HASOC datasets. Dimensions with "*" provided by Perspective API

B Appendix - Separability & Semantic similarity between the dataset-annotation pairs

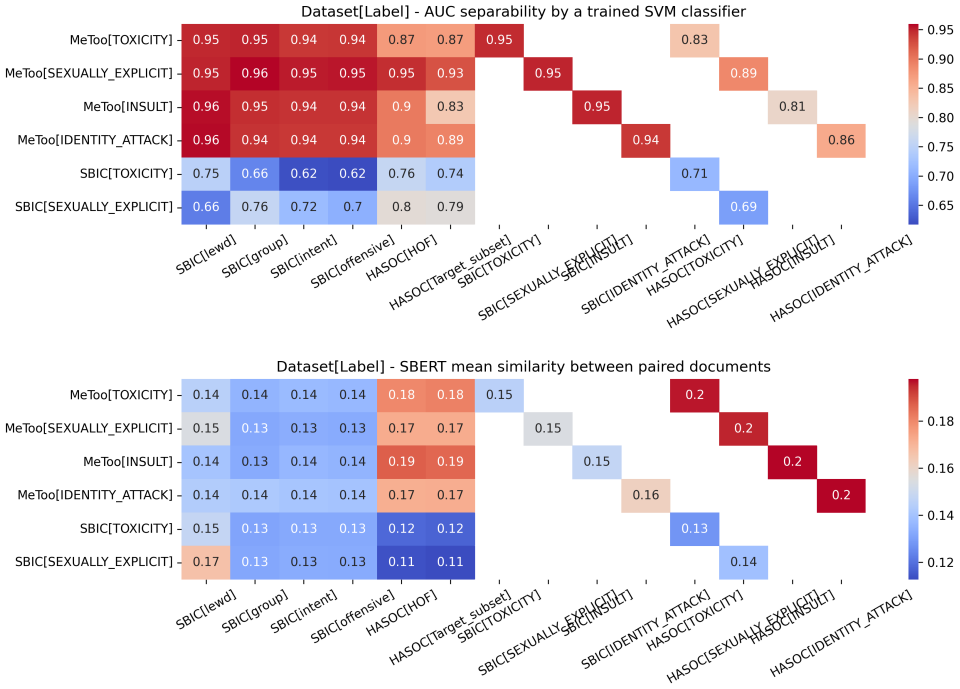

Figure 7: Estimate of distributional difference - SVM separability (top) and SBERT mean similarity (bottom) between dataset-label pairs for #MeToo, SBIC, and HASOC datasets. We estimate the distributional difference wrt. labels between the datasets following the approach from [40]. We train an SVM classifier to tell datasets apart under the aligned labels (i.e., positive class posts put into the same set) from the source and target domain tasks.

C Appendix - Labeling definitions used with few-shot instructions

Dataset	Task	Definition	Tokens
SBIC	Offensive	Does this post contain offensive language?	Yes, No
	Intent	Does this post contain intentional insults?	Yes, No
	Lewd	Does this post contain sexual content?	Yes, No
	Group	Does this post contain offense to a group?	Yes, No
HASOC	HOF	Does this post contain any form of non-acceptable language such as hate speech, offensiveness, aggression, profanity?	Yes, No
	Target	Does this post contain an insult/threat to an individual, group, or others?	Yes, No
All datasets	Toxicity	Does this post contain rude, disrespectful, or unreasonable language?	Yes, No
	Sexually Explicit	Does this post contain sexually explicit language?	Yes, No

Table 2: Labeling definitions used for each task in the few-shot prompts. We use the definitions from prior few-shot work for the SBIC and HASOC tasks [23]. For the Perspective API tasks we use the definitions of each dimension provided on the Perspective API website [9]. We note that in all transfer learning experiments, we use the source-domain definition for the source-domain shots and the target-domain definition for the target-domain shots and query.