Fine-Grained Transfer Learning for Harmful Content Detection through Label-Specific Soft Prompt Tuning

Faeze Ghorbanpour[1,2,3] **Viktor Hangya**[4*] **Alexander Fraser**[1,3]

[1]School of Computation, Information and Technology, TU Munich
[2]Center for Information and Language Processing, LMU Munich
[3]Munich Center for Machine Learning (MCML)
[4]Fraunhofer IIS, Erlangen, Germany

faeze.ghorbanpour@tum.de, viktor.hangya@iis.fraunhofer.de

Abstract

The spread of harmful content online is a dynamic issue evolving over time. Existing detection models, reliant on static data, are becoming less effective and generalizable. Developing new models requires sufficient up-to-date data, which is challenging. A potential solution is to combine existing datasets with minimal new data. However, detection tasks vary—some focus on hate speech, offensive, or abusive content, which differ in the intent to harm, while others focus on identifying targets of harmful speech such as racism, sexism, etc—raising the challenge of handling nuanced class differences. To address these issues, we introduce a novel transfer learning method that leverages class-specific knowledge to enhance harmful content detection. In our approach, we first present label-specific soft prompt tuning, which captures and represents class-level information. Secondly, we propose two approaches to transfer this fine-grained knowledge from source (existing tasks) to target (unseen and new tasks): initializing the target task prompts from source prompts and using an attention mechanism that learns and adjusts attention scores to utilize the most relevant information from source prompts. Experiments demonstrate significant improvements in harmful content detection across English and German datasets, highlighting the effectiveness of label-specific representations and knowledge transfer. [1]

1 Introduction

The increasing proliferation of harmful content and its evolving nature require effective and generalizable detection methods. One solution for detecting hateful content in new domains involves preparing new adequate data, which presents challenges in data collection and annotation. Another approach is to utilize existing datasets and transfer their knowledge. However, directly applying these datasets for transfer learning is not straightforward. Various datasets have been developed to identify offensive language, hate speech, sexism, or racism, but nuanced differences among classes, variations in annotation styles, and differences in scope make it challenging to apply these datasets to new, unseen tasks (MacAvaney et al., 2019; Fortuna et al., 2020; Bourgeade et al., 2023).

In contrast to previous work that addresses these challenges through data augmentation (Al-Azzawi et al., 2023), domain-adapted models (Caselli et al., 2021a), or specific model instructions (Plaza-del arco et al., 2023), we argue that effectively differentiating between classes of harmful language requires modeling fine-grained class representations. With these representations, we can analyze label information across multiple tasks (Meng et al., 2020; Inagaki, 2022) and transfer knowledge between tasks in a fine-grained manner (Hangya and Fraser, 2024; Ludwig et al., 2022).

Large language models (LLMs) (Devlin et al., 2019; Radford et al., 2019; Raffel et al., 2020) have demonstrated strong performance across various tasks, but fine-tuning them is challenging and costly. Parameter-efficient fine-tuning (PEFT) addresses this by training only a small number of parameters (Liu et al., 2022a; Houlsby et al., 2019; Xie and Lukasiewicz, 2023). Our method builds on soft prompt tuning (SP), a PEFT approach introduced by Lester et al. (2021), using fewer parameters to achieve competitive results. SP scales well with larger models and effectively captures task-level information (Qin et al., 2021; Goswami et al., 2023). However, it learns a single-task representation, which we argue is insufficient for harmful content detection, where more fine-grained class-level representations are necessary.

We present an edition of the SP approach for harmful text detection by introducing label-specific

[*] This work was done while the author was affiliated with LMU Munich.

[1] The code and prompts are available at https://github.com/FaezeGhorbanpour/LabelSoftPromptTuning

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 11047–11061
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

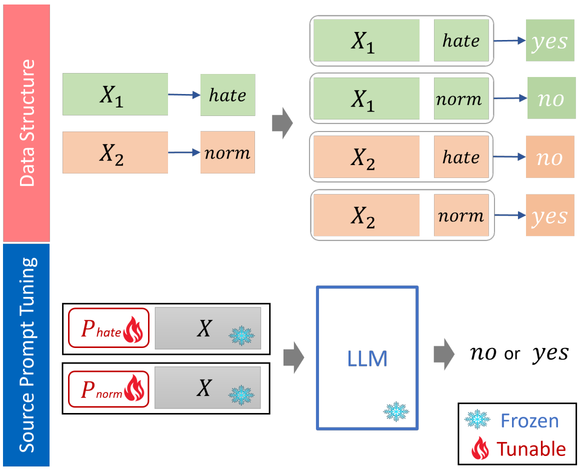


Figure 1: Overview of label-specific soft prompt tuning. For example, for a task with two classes, our method employs two label-specific prompts, prompting the LLM to predict if the given label-specific prompt and the instance's label match.

soft prompts. This method allows LLMs to capture the nuanced concepts associated with different harmful language types. As shown in Figure 1, we pair the input with each label-specific prompt, pass them through the model, and predict whether the input's and the soft prompt's label match (*yes* or *no*). This helps the model learn the right soft prompt and the nuances of each class.

Recent work leveraged transfer learning to boost performance in low resource settings (Kapil and Ekbal, 2020; Glavaš et al., 2020). However, finding the right existing tasks (source) for a new unseen task (target) is not straightforward. Firstly, harmful content detection datasets often have partly overlapping label sets, e.g. the Mandl et al. (2019) dataset has labels *hateful*, *offensive* and *normal*, while the dataset of Founta et al. (2018) has *hateful*, *abusive*, *spam* and *normal*. Secondly, labels of the same name often have different definitions, e.g. hate speech and offensive are defined as two separate classes in some datasets (Mathew et al., 2021; Toraman et al., 2022); however, some others include hate speech in the offensive class (Sigurbergsson and Derczynski, 2020).

Label-specific soft prompts enable fine-grained knowledge transfer between tasks. Instead of transferring all knowledge, including irrelevant or misaligned labels, we propose transferring only essential information by selecting the appropriate label soft prompts. This can be achieved simply by initializing the label-specific soft prompts with the same type of source prompts directly or by averaging them as depicted in the left side of Figure 2. Alternatively, as shown on the right side of Figure 2, the model can be trained to select the required information from source label prompts autonomously. To achieve this, we use an attentional mixture of soft prompts across all label prompts, allowing the model to learn how to utilize them effectively. Our proposed attentional transfer learning method, designed for label-specific learning and inspired by Asai et al. (2022), automatically identifies which source label prompts are most useful for each target label. Additionally, this method measures the relation between source and target labels by providing attention scores.

Evaluation of our fine-grained transfer learning method across five tasks demonstrates the effectiveness of label-specific knowledge transfer compared to baseline methods. Based on label types (names), the initialization method performs well, showing that label names can guide the transfer and selection of source label prompts. The attention method, by contrast, removes the dependency on specific label types and uses attention scores to identify the most effective source label prompts for the target labels. More importantly, it outperforms other methods, demonstrating that target labels can benefit from a broader range of source prompts, rather than being limited to those of the same type. Additionally, our first stage—label-specific source prompt tuning—outperforms task-specific prompt tuning, even when the tunable parameters for task-specific prompts are increased.

Overall, our method exhibits these characteristics:

- We introduce transferring class-level information in harmful content detection tasks using label-specific soft prompts through two approaches: initialization-based and attention-based methods.

- We propose a label-specific attention-based method, automating fine-grained knowledge transfer by learning attention scores. This method enhances the contribution of source label prompts to target tasks, leading to performance improvements.

- Few-shot results further highlight the effectiveness of fine-grained knowledge transfer of attention-based method, even when using limited samples for target tasks.

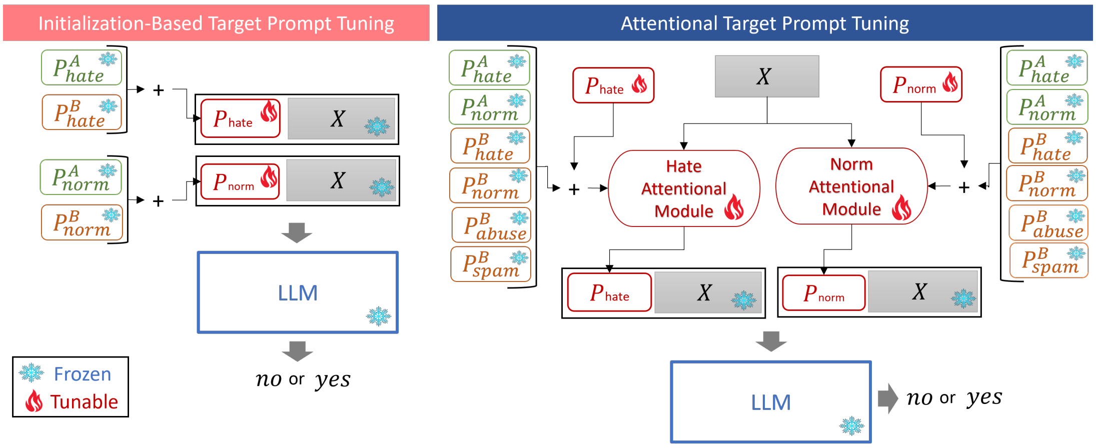


Figure 2: Transfer learning architecture for label-specific soft prompt tuning: The left side shows the initialization method, where two source tasks (A and B) are used to initialize the target task's labels. The right side illustrates the attentional method, where task B has additional labels processed by attention modules trained to use the source prompts regardless of their types.

2 Related Work

Transfer Learning for Harmful Text Detection. Fine-tuning pre-trained language models (PLM) for harmful content detection is a well-established research approach, focusing on adapting LLMs to specific tasks such as detecting harmful text. In contrast, hard prompting methods, which involve manually designing textual templates and querying the model (Guo et al., 2023; Huang et al., 2023; Roy et al., 2023), can be less robust. These methods rely on altering prompts to change responses, which often lacks the stability and effectiveness achieved through fine-tuning. HateBERT (Caselli et al., 2021b) retrained the BERT (Devlin et al., 2019) model for detecting abusive language. Sarkar et al. (2021) and Okpala et al. (2022) applied transfer learning techniques to offensive language identification. Sabry et al. (2022) explore the effectiveness of T5 for hate speech detection by investigating data augmentation techniques and ensemble methods, and Adewumi et al. (2023) introduces HaT5, a Text-to-Text Transformer based on T5. However, these papers did not consider various types of harmful content and the generalizability of their transfer learning methods. While a prior study He et al. (2024) explored soft prompts for general toxic content detection without modification, our work focuses on harmful contents with label-specific prompts tuned using PEFT methods. This allows our model to capture the subtleties of different harmful categories.

Incorporating Label-specific Information in Training. Work such as Wang et al. (2018); Li et al. (2022); Liu et al. (2022b) utilizes label embeddings to create more informative text representations and enhance text classification. In the paper by Xiao et al. (2019), a multi-view attention mechanism is proposed to learn label-specific representations for text classification. Meng et al. (2020) introduce a text classification approach using only label types for self-training LLMs. None of these works measured the knowledge transfer of their label-specific representations. Müller et al. (2022) introduced label tuning, which adapts models to new tasks by fine-tuning label embeddings. Unlike our method, they use label text as hard prompts and transfer encoder weights instead of soft prompts, which is less efficient for transfer learning, especially when dealing with multiple source tasks.

Generalization Capability to Diverse Harmful Detection Tasks. Recent research highlights concerns about the generalizability of current hate speech detection models (Swamy et al., 2019). Although they perform well on their own test sets, their accuracy drops significantly when tested on different datasets (Yin and Zubiaga, 2021). This indicates that existing test data do not accurately represent future cases, leading to overestimating these models' generalization capabilities (Karan and Šnajder, 2018; Arango et al., 2022). Bourgeade et al. (2023) analyze hate speech detection models, finding topic-diverse training data improves their ability to generalize across different hate speech types. Hangya and Fraser (2024) addresses abusive content detection by training a model on multiple datasets and adapting it to specific tasks with minimal data, achieving strong multilingual performance. However, retraining a model is less efficient compared to using soft prompts. Our experiments demonstrate that our approach has strong generalization capabilities and efficiently transfers fine-grained knowledge to target tasks.

3 Methodology

The methodology is organized into two stages: source prompt tuning (label-specific soft prompt tuning) and target prompt tuning (transfer learning for label soft prompts). The second stage has two approaches: initialization and attention-based methods. Each subsection starts with a problem statement, followed by its formulations.

3.1 Label-specific soft prompt tuning

The proposed method for label-specific source prompt tuning, illustrated in Figure 1, builds on highly parameter-efficient soft prompt tuning with a single *task-specific prompt*, presented by Lester et al. (2021). We propose a novel approach called *label-specific soft prompt tuning* that leverages multiple soft prompts, equal to the number of classes in a task. This allows the model to learn fine-grained prompts tailored to each class.

To achieve this, we transform the dataset into a *contrastive format*, which we define as pairing each input with every possible label to check if they match. For example, in a binary classification of $hate$ or $normal$, an instance X labeled as $hate$ is traditionally expected to be classified as $hate$. In the new format, we create two input pairs: $(X, hate)$ with an expected output of yes, and $(X, normal)$ with an expected output of no.

Before feeding these pairs into the PLM, we prepend a soft prompt corresponding to the accompanying label of the input. This allows the model to predict whether the input and the associated label match. Essentially, the model learns to identify the most relevant soft prompt for each class and to store label-specific information in them. This method is especially effective for detecting harmful speech because it enables the model to differentiate explicit harmful content.

Problem Formulation Formally, given a task with n instances and k classes, consider an input sequence with m token embeddings $X_i = \{x_1, x_2, ..., x_m\}$, and its corresponding class label y_i. To form a soft prompt we prepend l tokens, denoted as $P = \{p_1, p_2, ..., p_l\}$, to every instance. $p_j \in \mathbb{R}^d$ similar to $x_j \in \mathbb{R}^d$ represents an embedding vector, and d signifies the input dimension.

In task-specific prompt tuning, only one soft prompt is considered for a task. The training objective is to maximize the likelihood of decoding the desired output class y_i,

$$\mathcal{L}_i = \max_P \log p\left(y_i \mid P; X_i\right)$$

where, only P is trainable. For text classification, y_i is the text of the label corresponding to X_i.

When we shift to a contrastive setting, the number of instances becomes $n \times k$, and the labels are converted to $\{no, yes\}$. Here, the input involves associating each instance with soft prompts of k classes. If the correct soft prompt of the instance's label is associated with the instance, the output is yes; otherwise, it is no. In the label-specific tuning, we have k soft prompts, and we indicate them with $\hat{P} = \{P^1, P^2, ..., P^k\}$, where P^j is the soft prompt of label j. Therefore, for an instance X_i with label y_i, we have k instances in the contrastive format $([P^1, X_i], [P^2, X_i], ..., [P^k, X_i])$, and the output will be $Z_i = \{Z_i^1, Z_i^2, ..., Z_i^k\}$:

$$Z_i^j = \begin{cases} \text{yes}, & \text{if } P^j = y_i \\ \text{no}, & \text{if } P^j \neq y_i \end{cases}$$

The goal of label-specific prompt training is to maximize the likelihood of the outputs Z_i by optimizing only over the prompts $\hat{P}$:

$$\mathcal{L}_i^j = \max_{P^j} \log p\left(Z_i^j \mid P^j; X_i\right)$$

During inference, the predicted label is chosen based on the label-soft prompt with the highest confidence among those predicted as matches.

$$j = \arg\max_j \log p\left(Z_i^j \mid P^j; X_i\right) \\ \text{where } Z_i^j = \text{yes}$$

3.2 Transfer learning for label soft prompts

In the second stage, we aim to transfer label knowledge between tasks. Source tasks, for which we have label soft prompts available, help to tune and generate label soft prompts for target tasks.

3.2.1 Initialization method

In this method, we transfer knowledge between tasks by initializing target soft prompts with source label prompts of the same type. We align source prompts with target labels, e.g., matching labels like "normal," "hate," "offensive," and "racism." As Figure 2 shows, we initialize target label soft prompts before training or inference, either using a single source prompt or averaging multiple prompts for a comprehensive initialization.

Problem Formulation. In mathematical terms, let $\hat{P}_s = \{P_s^1, P_s^2, \ldots, P_s^R\}$ be a set of source soft prompts, where P_s^r is the prompt for the r-th label among all source tasks. Let $\hat{P}_t = \{P_t^1, P_t^2, \ldots, P_t^k\}$ be label-specific prompts for a new target task. The function $type(\cdot)$ returns each prompt's label category (e.g., "normal" or "hate").

Two initialization strategies are considered before training. In *simple initialization*, P_t^i is directly copied from a single P_s^r that shares the same label type, i.e. $type(P_s^r) = type(P_t^i)$. In *average initialization*, if multiple P_s^r share the same type, we use their average:

$$P_t^i = \frac{1}{|S_i|} \sum_{r \in S_i} P_s^r,$$
$$\text{where } S_i = \{r : type(P_s^r) = type(P_t^i)\}$$

Thus, simple initialization uses exactly one matching source prompt, whereas average initialization blends all matching source prompts. These initialized target prompts then learn label-specific information in the target task. The rest of the training approach for this setup follows the same approach for source prompt tuning, with the primary difference being how the soft prompts are initialized.

3.2.2 Attention-based method

In the second approach, attentional mixtures of label-specific soft prompts, we adapt the attention mechanism to handle distinct labels within the same task, building on the task-specific attentional mixtures proposed by Asai et al. (2022). Our approach differs from the task-specific method, not only by using multiple target prompts and attention modules but also by making the target prompts more active in the attention calculation. This ensures that each label-specific attention module is trained uniquely for its corresponding label.

After converting the dataset to a contrastive format, as done in source prompt tuning, the associated label of each instance determines which attention module and soft prompt to use. If the label is "hate," the attention module and soft prompt for *hate* are used; if it is "normal," the *normal* attention module and soft prompt are applied.

Additionally, unlike the initialization method that fixes source prompts at the start, this method uses the same source prompts across all attention modules, allowing the model to adaptively prioritize them during training. While the initialization method is easy to use, it requires knowing which labels have matching definitions in advance, which can be error-prone, as useful labels might be overlooked or less relevant ones might be chosen.

Problem Formulation. As in the previous section, let $\hat{P}_s = \{P_s^1, P_s^2, \ldots, P_s^R\}$ be a set of non-trainable source soft prompts, and let $\hat{P}_t = \{P_t^1, P_t^2, \ldots, P_t^k\}$ be a set of trainable label-specific soft prompts for a target task. Similar to source prompt tuning, we prepend a final soft prompt to the input. In task-specific tuning (Asai et al., 2022), a single P_t serves as one of the source prompts (but is trainable), and attention is applied over $\hat{P}_s$ plus the input embedding to form the final prompt. The model is then trained by maximizing:

$$\alpha = \text{softmax}\left(\frac{[P_s, P_t] \cdot \text{LNorm}\,(W.X_i)}{\tau}\right)$$
$$G = \sum \alpha \odot [P_s, P_t]$$
$$\mathcal{L}_i = \max_{P_t, \theta_G} \log p\big(y_i \mid G(P_s, P_t, X_i) + P_t;\, X_i\big).$$

where $[a, b]$ indicates stacking of vectors a and b. W, P_t are trainable parameters and *LNorm* is the layer normalization function (Ba et al., 2016). Finally, $G + P_t$ is prepended to all X_i and passed to the LLM.

However, for our *label-specific* method, we adopt almost a similar formulation but introduce separate attention modules and soft prompts for each label j. Moreover, the operands of the attention module differ from those of the task-specific one to ensure that each attention module is tuned differently from the others.

We test different attention mechanisms: no attention with equal weights, dot product attention, and a trainable attention module. For each label j, the inputs of attention mechanisms are the embedding $X \in \mathbb{R}^{m \times d}$, the label soft prompt $P_t^j \in \mathbb{R}^{l \times d}$, and the same set of source prompts $P_s \in \mathbb{R}^{r \times l \times d}$. To ensure each label has distinct and effective operands, we add P_t^j to each source prompt P_s^r (for $r = 1, \ldots, R$), and pass the sums to an attention function:

$$\alpha^j = \text{softmax}\left(\frac{\left(P_t^j + P_s\right) \cdot \text{LNorm}\left(W^j.X_i\right)}{\tau}\right)$$
$$G^j = \sum \alpha^j \odot \left(P_t^j + P_s\right)$$

Finally, the output of the attention G^j is prepended to all instances labeled j. We thus maximize:

$$\mathcal{L}_i^j = \max_{P_t^j, \theta_{G^j}} \log p\big(Z_i^j \mid G^j(P_s, P_t^j, X_i);\, X_i\big).$$

where $\{P_t^j, W^j\}$ are learned for each label j. The attention scores are the normalized attention weights ($a^j \in \mathbb{R}^r$) obtained using the softmax function (Radford et al., 2021). These scores indicate the relevance between the source prompts and instances associated with the labels of a task.

4 Experimental Setup

4.1 Datasets

For the evaluation, we consider five primary datasets for the transfer learning experiments: HateSpeech 18 (de Gibert et al., 2018), SRW 16 (Waseem and Hovy, 2016), OLID (Zampieri et al., 2019), and German datasets: GermEval 18 (Risch et al., 2021) and the German tasks of HASOC 19 (Mandl et al., 2019). To have more source prompts for transfer learning, we also include the following tasks: Hateval 19 (Basile et al., 2019), the English tasks of HASOC 19 (Mandl et al., 2019), Xdomain (Toraman et al., 2022), HateXplain (Mathew et al., 2021), and Abuse (Founta et al., 2018). Additionally, we experimented with different sub-tasks on datasets providing multiple levels of annotation. The goal of incorporating various sub-tasks is to encompass a wide range of labels to show the generalizability of our approach in transfer learning and few-shot experiments. The harmful detection datasets we used contain explicit harmful content, which means the hateful content is directly present within the sentences. Detailed dataset information is provided in Appendix B.

4.2 Training Details

We use open-source HuggingFace language models and the Pytorch framework. Following standard practices, we set the length of soft prompts to 100 tokens (Lester et al., 2021; Asai et al., 2022). Given that we are addressing a classification problem with mostly imbalanced datasets, we use macro F1 score as our main metric. Each reported result is the average performance over three runs. Further experimental details are provided in Appendix A.

Adhering to the standard methodology employed in prior prompt-based studies (Lester et al., 2021; Asai et al., 2022; Ma et al., 2022), we primarily conduct our experiments using the publicly available pre-trained T5-base model, which contains 220M parameters. Additionally, our study includes evaluations with the T5-Small (60M) and T5-Large (770M) models. We use LLMs in a sequence-to-sequence manner, converting harmful text classification tasks into a format where the model generates the token corresponding to the input's label.

4.3 Initialization of soft prompts

In task-specific soft prompt tuning we follow the standard procedure and initialize soft prompts with the embeddings of randomly chosen vocabulary items (Lester et al., 2021; Asai et al., 2022). In label-specific tuning we initialize label-specific soft prompts with the embeddings of the label's name (repeated over the 100 tokens). Our assumption is that this approach makes these soft prompts distinct from each other during the initial epochs, leading to faster convergence. In contrast, since the baseline only contains one task-specific soft prompt, such a fine-grained distinction cannot be made.

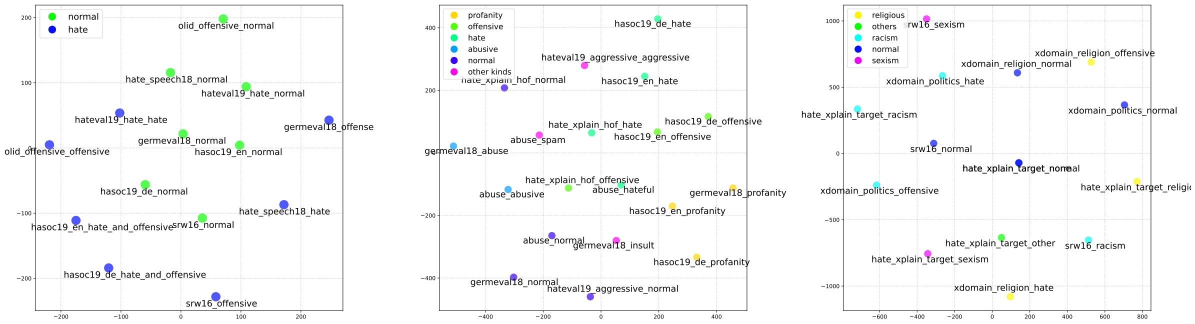
5 Evaluation

Although our main contribution is the transfer learning of label-specific soft prompts, to maintain coherence with the methodology section, we first present the evaluation and visualizations of the first stage. Then, we discuss the two transfer learning methods: Initializing and Attentional, in that order.

5.1 Source Prompt Tuning Results

This section presents the results of the first stage of our methodology: source prompt tuning, which refers to label-specific soft prompt tuning. This stage provides source label-soft prompts for the second stage (transfer learning). Interestingly, this stage, by incorporating soft prompts for each label in a task, achieves higher performance compared to task-specific prompt tuning (Lester et al., 2021), even without transfer learning.

The results are shown in Table 1, which compares the performance of label-specific source prompt tuning with task-specific one. In the three LLMs used in our study, label-specific prompt tuning, on average, performs better, indicating that contrasting label soft prompts with instance embeddings gives the model a chance to learn label differences, supporting our claim that fine-grained representation can distinguish harmful content types more. This improvement is not due to the increased



(a)

(b)

(c)

Figure 3: t-SNE visualization of label soft prompts in various hate speech tasks. (a) Binary tasks distinguishing hate speech vs normal speech. (b) Fine-grained tasks are classifying types of hate speech. (c) Targeted tasks identifying specific groups targeted by hate speech. In the figure, each label is represented by a name above its corresponding dot, where the first part denotes the task it belongs to, and the second part indicates the label type.

		T5 small		T5 base		T5 large	
		task	label	task	label	task	label
GermEval 18	binary	69.98	**70.46**	62.67	**70.98**	**73.47**	72.12
	fine-grained	37.31	**39.41**	32.14	**40.23**	**42.62**	39.84
HateSpeech 18	binary	72.57	**74.42**	73.58	**76.43**	76.02	**76.65**
HASOC 19 (de)	binary	**55.23**	55.00	**52.95**	51.92	55.10	**56.69**
	fine-grained	29.81	**30.51**	29.96	**32.72**	**35.36**	28.01
SRW 16	binary	83.29	**85.94**	83.11	**85.27**	85.18	**85.92**
	fine-grained	57.33	**57.39**	57.55	**58.26**	**58.27**	58.03
OLID	offensive	76.01	**76.36**	77.48	**78.4**	78.38	**79.78**
	targeted	54.28	**60.28**	47.02	**61.47**	57.54	**62.16**
	target	49.31	**52.31**	54.68	**57.51**	46.31	**51.37**
	average	58.51	**60.01**	57.11	**61.32**	60.83	**61.06**

Table 1: Source soft prompt tuning results. All of the results are F_1-macro (*p.p.*) scores. Task- and label-specific approaches are indicated by *task* and *label*. The best scores of a given model and task are bold.

tunable parameters in label-specific prompt tuning, as shown in Appendix D.

5.2 Visualization of Prompts

After the first stage, we have soft prompts for each label across all our source tasks; here, we want to see if these prompt vectors have a numeric relationship and can be clustered effectively. To depict the prompt vectors, we reshaped the prompts with dimensions ($prompt\_length \times model\_dim$) into a 1D vector with dimensions ($prompt\_length * model\_dim$) and passed it to t-SNE (van der Maaten and Hinton, 2008) to reduce its dimension and depict it (Figure 3).

In Figure 3a, the soft prompts for binary classification are displayed. The hate prompts are grouped closely together, as are the normal soft prompts. In Figure 3b, different types of hate speech in the tasks are displayed. This figure also illustrates the closeness of label-specific soft prompts that have the same label types. As shown, profanity prompts (yellow) are notably close to each other, as are the offensive soft prompts. Figure 3c displays the tasks involved in classifying hate speech based on the targeted group. The clustering is not as evident as in other figures due to the complex nature of these labels and the overlap among instances.

The clustering of similar label soft prompts led us to explore the potential benefits of transfer learning from source label soft prompts. The next section on initialization-based transfer learning results will demonstrate that the identified clusters of similar label types in the plots are indeed functioning as expected.

5.3 Transfer Learning Results: Initializing

Here, we aim to investigate knowledge transfer between tasks by initializing target soft prompts with label-specific source soft prompts of the same type. In choosing source label-specific prompts, we considered matching label types, i.e., we used hate source prompts for hate target labels and normal source prompts for normal target labels.

The first eight columns of Table 2 display the results of this experiment. Among the different initializing methods, using multiple source prompts and averaging them (*all same type source prompts*) is better than using only one random source prompt (*1 same type source prompt*). Moreover, when mixing incompatible label source prompts for the

		Initializing Method								Attentional Method			
		task			label					task	label		
		Random Vocabs	1 random SP	All SPs	Random Vocabs	Label's Token	1 same Type SP	All same Type SPs	All SPs		Attention Types		
											Constant	Dot	Ours
GermEval 18	binary	62.67	65.78	71.02	69.93	70.98	71.27	**71.38**	71.31	70.58	**70.75**	50.76	**70.75**
	fine-grained	32.14	35.15	32.98	39.97	40.23	40.61	42.18	**42.44**	37.57	37.04	20.29	**42.33**
HateSpeech 18	binary	73.58	73.71	74.79	76.07	76.43	77.49	**78.21**	76.88	76.62	76.72	66.29	**77.03**
HASOC 19 (de)	binary	52.95	50.19	**56.52**	53.30	49.33	51.92	46.65	45.63	53.23	50.06	46.04	**53.57**
	fine-grained	29.96	**41.35**	30.05	31.63	32.72	38.83	32.67	30.33	**39.49**	30.12	28.89	33.76
SRW 16	binary	83.11	84.04	84.89	86.02	85.27	86.58	86.69	**87.13**	85.69	**86.80**	82.74	86.78
	fine-grained	57.55	58.01	57.18	58.21	58.26	58.32	**58.78**	58.02	58.60	58.15	49.32	**58.39**
OLID	offensive	77.48	77.01	77.97	78.45	78.40	77.96	**79.65**	78.09	78.65	**79.37**	62.23	78.04
	targeted	47.02	59.80	55.06	57.58	61.47	53.17	**63.91**	53.28	55.75	66.96	46.96	**67.71**
	target	54.68	54.81	48.56	51.81	57.51	50.49	**59.40**	46.80	48.96	48.49	20.85	**56.73**
	average	57.11	59.98	58.90	60.30	61.32	61.06	**61.95**	58.99	60.51	60.45	47.44	**62.51**

Table 2: Different settings of transfer learning methods using the T5-base model. All of the results are F_1-macro ($p.p.$) scores. Task- and label-specific approaches are indicated by *task* and *label*. SP = Source Prompt. The two best scores of a given target task are bold.

target tasks (*all source prompts*), the performance decreases. This indicates that same-type knowledge transfer is more beneficial. Since we have the types of labels as the relationships among labels from other tasks, we can use label-specific source prompts for the target task. However, in a transfer learning scenario involving task-specific prompt tuning, we lack advanced information about the relationships between tasks to utilize them effectively. Table 2 also shows that initializing the soft prompts with the *label's token* is a better choice than initializing it with *random vocabs*.

5.4 Transfer Learning Results: Attentional

Our second transfer learning method is the attentional mixture of label-soft prompts; this section presents its evaluation. This approach transfers knowledge from source prompts to target tasks by using an attention mechanism, eliminates random selection of source label prompts, and measures which prompts the model utilizes.

The last four columns of Table 2 correspond to this method. *task* represents the task-specific mixture of soft prompts (Asai et al., 2022) as the baseline, while our attentional method is in the last column, labeled *ours*. Comparing these two shows that fine-grained knowledge transfer is effective. The other three columns (*Attention Types*) present an ablation study on attention type: *Constant* applies no attention, *Dot* uses a dot product between source prompts and input embeddings, and *Ours* is an attention module with trainable parameters. The results show that trainable attention (our choice) performs best. Additionally, comparing *all the same type SPs* to *Ours* confirms that the attentional method performs better on average.

		T5 small		T5 base		T5 large	
		task	label	task	label	task	label
GermEval 18	binary	69.23	**70.37**	70.58	**70.75**	73.47	**73.93**
	fine-grained	37.68	**40.58**	37.57	**42.33**	**42.92**	39.84
HateSpeech 18	binary	74.94	74.94	76.62	**77.03**	**78.66**	76.83
HASOC 19 (de)	binary	55.22	**56.09**	53.23	**53.57**	52.90	**54.38**
	fine-grained	25.41	**32.79**	**39.49**	33.76	**40.65**	37.36
SRW 16	binary	**85.82**	85.77	85.69	**86.78**	86.71	**87.22**
	fine-grained	57.31	**57.40**	**58.60**	58.39	**57.80**	57.25
OLID	offensive	76.20	**76.86**	**78.65**	78.04	78.74	**80.30**
	targeted	50.96	**63.69**	55.75	**67.71**	58.67	**64.43**
	target	48.62	**50.38**	48.96	**56.73**	49.77	**52.37**
	average	58.14	**60.89**	60.51	**62.51**	62.03	**62.39**

Table 3: Attentional target prompt tuning results. All of the results are F_1-macro ($p.p.$) scores. Task- and label-specific approaches are indicated by *task* and *label*. The best scores of a given model and task are in bold.

Table 3 compares the results of task-specific attentional transfer learning (*task*) and the label-specific approach (*label*) across three LLMs, from small to large. As shown for most tasks and models, label-specific target prompt tuning outperforms task-specific prompts, particularly in binary tasks. This may be because, in fine-grained tasks, target label prompts and their respective attention modules require more guidance to effectively differentiate

Target task	Label	Top 3 source labels (with their tasks)
GermEval 18 - binary	normal	**offensive** (Olid_offensive), **offensive** (Xdomain_politics), **normal** (Xdomain_religion)
	offensive	**normal** Hateval 19_aggressive, **hate** (Hasoc 19_en_fine-grained), **normal** (Hasoc 19_de_binary)
GermEval 18 - fine-grained	normal	**normal** (Xdomain_politics), **racism** (HateXplain - target), **religious** (HateXplain - target)
	profanity	**religious** (HateXplain - target)
	insult	**unintentional** (Hasoc 19_en_targeted), **generic** (Hateval 19_target), **offensive** (Hasoc 19_de_fine-grained)
	abuse	**racism** (HateXplain - target), **normal** (Olid_offensive), **normal** (Xdomain_politics)
HateSpeech 18 - binary	normal	**individual** (Hateval 19_target), **racism** (SRW 16_fine-grained), **profanity** (Hasoc 19_de_fine-grained)
	hate	**hate** (Xdomain_politics)
SRW 16 - binary	offensive	**hate** (Hasoc 19_de_fine-grained), **offensive** (Hasoc 19_de_fine-grained), **profanity** (Hasoc 19_de_fine-grained)
	normal	**normal** (Olid_offensive)
SRW 16 - fine-grained	sexism	**profanity** (Hasoc 19_de_fine-grained), **hate** (Hasoc 19_de_fine-grained), **offensive** (Hasoc 19_de_fine-grained)
	racism	**hate** (HateSpeech 18 - binary), **profanity** (GermEval 18_fine-grained), **normal** (Xdomain_religion)
	normal	**unintentional** (Hasoc 19_en_targeted), **hate** (Hasoc 19_de_fine-grained), **profanity** (Hasoc 19_de_fine-grained)

Table 4: Top three source label prompts utilized by each target task's label, based on the attention scores.

		1		16		64	
		task	label	task	label	task	label
GermEval 18	binary	31.84	**41.75**	54.16	**54.52**	54.59	**60.46**
	fine-grained	**19.93**	16.71	08.16	**20.15**	22.74	**27.42**
SRW 16	binary	39.45	**49.20**	**61.18**	57.27	**68.24**	63.82
	fine-grained	**21.33**	20.20	33.61	**33.63**	45.18	**47.64**
OLID	offensive	**41.61**	41.00	**64.08**	57.63	**65.32**	58.78
	targeted	**46.50**	44.96	31.69	**56.33**	51.97	**55.22**
	target	09.37	**26.51**	27.79	**37.13**	27.97	**41.64**
	average	30.00	**34.33**	40.10	**45.24**	48.00	**50.71**

Table 5: Results of few-shot learning for attentional target prompt tuning. All of the results are F_1-macro ($p.p.$) scores. Task- and label-specific approaches are indicated by *task* and *label*. The best scores of a given shot and task are bold.

between subtle harmful content types. To compare source soft prompt tuning (first stage) with transfer learning (second stage), Table 1 and Table 3 are examined. The comparison of *label* columns in both tables shows that transfer learning outperforms the first stage alone for harmful content detection across three LLMs.

The attention-based approach eliminates the need for manually selecting and averaging source label prompts. Instead, it uses all source label prompts, and the model learns to leverage them by adjusting attention scores. Table 4 shows the top three source label prompts with the highest attention scores for each label of the target tasks. For example, in the *SRW 16_binary* task, which has two classes (normal and offensive), the source label prompt that the normal soft prompt utilizes the most is *normal* prompt from *OLID - offensive* sub-task. On the other hand, the offensive soft prompt draws on multiple source prompts: *hate*, *offensive*, and *profanity* prompts, all from the source task *HASOC 19 - de_fine-grained*. The attention score tables indicate that most target labels utilize source label prompts of the same type, such as *GermEval 18_fine-grained_normal* and *HateSpeech 18 - binary_hate*. However, there are target labels that diverge from this pattern and use different types of source labels, such as *GermEval 18_binary_normal* and *GermEval 18_binary_offensive*.

5.5 Few-shot Experiments Results

This experiment evaluates the attentional transfer learning of label-specific soft prompts from source tasks to target tasks with fewer training samples. Table 5 shows results for 1, 16, and 64-shot training sets for the target task. We compare two methods: task-specific and label-specific attentional target prompt tuning. The table demonstrates that, in most tasks and shot settings, label-specific tuning performs better with limited training data, indicating that source label prompts combined with the attentional method can be effectively utilized for new target tasks with limited samples.

6 Conclusion

We introduce transfer learning of label-specific soft prompt tuning for harmful content detection, where prompts are tailored to individual labels to capture fine-grained distinctions between content types. Moreover, we propose attentional knowledge transfer at the class level, enabling the model to learn label nuances from source prompts. Experimental results show that label-specific prompt tuning improves performance on English and German datasets, achieving higher F1-macro scores than the baseline. Few-shot experiments further validate the method, demonstrating enhanced performance in target tasks with limited data.

Limitations

This study demonstrates the potential of label-specific soft prompts for harmful content detection and knowledge transfer among tasks. However, limitations exist. Firstly, we have not explored this technique with LLMs beyond the T5 series. Although most soft prompt tuning research utilizes T5, and our approach is generally applicable to other LLMs, their performance has not been validated. Due to limited access to computational resources, we were unable to conduct all the transfer learning experiments for every dataset used in the initial stage and for more LLMs.

Our method can be applied to other PEFT methods, but we chose SP due to its parameter efficiency and because it prepends tunable parameters to the input. Using explicit harmful detection datasets, which primarily contain clearly inappropriate language, the initial layers of the language model are more likely to focus on these indicators.

The findings suggest that label-based knowledge sharing between tasks holds promise, particularly for complex text classification like harmful detection with nuanced categories. However, investigating its effectiveness in broader text classification tasks, such as topic classification, sentiment analysis, or domain identification, would be valuable for future research.

Acknowledgements

The work was supported by the European Research Council (ERC) through the European Union's Horizon Europe research and innovation programme (grant agreement No. 101113091) and the German Research Foundation (DFG; grant FR 2829/7-1).

References

Tosin Adewumi, Sana Sabah Sabry, Nosheen Abid, Foteini Liwicki, and Marcus Liwicki. 2023. T5 for hate speech, augmented data, and ensemble. *Sci*, 5(4).

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2623–2631. ACM.

Sana Al-Azzawi, György Kovács, Filip Nilsson, Tosin Adewumi, and Marcus Liwicki. 2023. Nlp-ltu at semeval-2023 task 10: The impact of data augmentation and semi-supervised learning techniques on text classification performance on an imbalanced dataset. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2022. Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, 105:101584.

Akari Asai, Mohammadreza Salehi, Matthew Peters, and Hannaneh Hajishirzi. 2022. ATTEMPT: Parameter-efficient multi-task tuning via attentional mixtures of soft prompts. In *Proceedings of the 2022 Conference on EMNLP*, pages 6655–6672. ACL.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63. ACL.

Tom Bourgeade, Patricia Chiril, Farah Benamara, and Véronique Moriceau. 2023. What did you learn to hate? a topic-oriented analysis of generalization in hate speech detection. In *Proceedings of the 17th Conference of the European Chapter of the ACL*, pages 3495–3508. ACL.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021a. HateBERT: Retraining BERT for Abusive Language Detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021b. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25. ACL.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20. ACL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies*, volume 1, pages 4171–4186. ACL.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794. European Language Resources Association.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Goran Glavaš, Mladen Karan, and Ivan Vulić. 2020. XHate-999: Analyzing and detecting abusive language across domains and languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6350–6365.

Koustava Goswami, Lukas Lange, Jun Araki, and Heike Adel. 2023. SwitchPrompt: Learning domain-specific gated soft prompts for classification in low-resource domains. In *Proceedings of the 17th Conference of the European Chapter of the ACL*, pages 2689–2695. ACL.

Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2023. An investigation of large language models for real-world hate speech detection. In *2023 ICML and Applications (ICMLA)*, pages 1568–1573. IEEE.

Viktor Hangya and Alexander Fraser. 2024. How to solve few-shot abusive content detection using the data we actually have. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8307–8322. ELRA and ICCL.

Xinlei He, Savvas Zannettou, Yun Shen, and Yang Zhang. 2024. You only prompt once: On the capabilities of prompt learning on large language models to tackle toxic content. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 64–64. IEEE Computer Society.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Chain of explanation: New prompting method to generate quality natural language explanation for implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, page 90–93. ACM.

Takeshi Inagaki. 2022. Information propagation by composited labels in natural language processing. *arXiv preprint arXiv:2205.11509*.

Prashant Kapil and Asif Ekbal. 2020. A deep neural network based multi-task learning approach to hate speech detection. *Knowledge-Based Systems*, 210:106458.

Mladen Karan and Jan Šnajder. 2018. Cross-domain detection of abusive language online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137. ACL.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on EMNLP*, pages 3045–3059.

Hui Li, Guimin Huang, Yiqun Li, Xiaowei Zhang, and Yabing Wang. 2022. Concept-based label distribution learning for text classification. *International Journal of Computational Intelligence Systems*, 15(1):85.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022a. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *NeurIPS*, 35:1950–1965.

Minqian Liu, Lizhao Liu, Junyi Cao, and Qing Du. 2022b. Co-attention network with label embedding for text classification. *Neurocomputing*, 471:61–69.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Florian Ludwig, Klara Dolos, Torsten Zesch, and Eleanor Hobley. 2022. Improving generalization of hate speech detection systems to novel target groups via domain adaptation. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 29–39. ACL.

Fang Ma, Chen Zhang, Lei Ren, Jingang Wang, Qifan Wang, Wei Wu, Xiaojun Quan, and Dawei Song. 2022. XPrompt: Exploring the extreme of prompt tuning. In *Proceedings of the 2022 Conference on EMNLP*, pages 11033–11047. ACL.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17. ACM.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020. Text classification using label names only: A language model self-training approach. In *Proceedings of the 2020 Conference on EMNLP*, pages 9006–9017. ACL.

Thomas Müller, Guillermo Pérez-Torró, and Marc Franco-Salvador. 2022. Few-shot learning with Siamese networks and label tuning. In *Proceedings of the 60th Annual Meeting of the ACL*, pages 8532–8545. ACL.

Ebuka Okpala, Long Cheng, Nicodemus Mbwambo, and Feng Luo. 2022. Aaebert: Debiasing bert-based hate speech detection models via adversarial learning. In *2022 21st IEEE ICML and Applications (ICMLA)*, pages 1606–1612.

Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68. ACL.

Yujia Qin, Xiaozhi Wang, Yusheng Su, Yankai Lin, Ning Ding, Jing Yi, Weize Chen, Zhiyuan Liu, Juanzi Li, Lei Hou, et al. 2021. Exploring universal intrinsic task subspace via prompt tuning. *arXiv preprint arXiv:2110.07867*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12, Duesseldorf, Germany. Association for Computational Linguistics.

Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing LLMs for hate speech detection: strengths and vulnerabilities. In *Findings of the ACL: EMNLP 2023*, pages 6116–6128. ACL.

Sana Sabah Sabry, Tosin Adewumi, Nosheen Abid, György Kovács, Foteini Liwicki, and Marcus Liwicki. 2022. Hat5: Hate language identification using text-to-text transfer transformer. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Diptanu Sarkar, Marcos Zampieri, Tharindu Ranasinghe, and Alexander Ororbia. 2021. fBERT: A neural transformer for identifying offensive content. In *Findings of the ACL: EMNLP 2021*, pages 1792–1798. ACL.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive language and hate speech detection for Danish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3498–3508. European Language Resources Association.

Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. 2019. Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950. ACL.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225. European Language Resources Association.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. 2018. Joint embedding of words and labels for text classification. In *Proceedings of the 56th Annual Meeting of the ACL*, pages 2321–2331. ACL.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. ACL.

Lin Xiao, Xin Huang, Boli Chen, and Liping Jing. 2019. Label-specific document representation for multi-label text classification. In *Proceedings of the 2019 Conference on EMNLP and the 9th IJCNLP*, pages 466–475. ACL.

Zhongbin Xie and Thomas Lukasiewicz. 2023. An empirical analysis of parameter-efficient methods for debiasing pre-trained language models. In *Proceedings of the 61st Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 15730–15745. ACL.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies*, volume 1, pages 1415–1420. ACL.

Appendix

A Experimental Details

We used the AdamW optimizer (Loshchilov and Hutter, 2019) with a linear learning rate scheduler that includes a warm-up period of 10%, an initial learning rate set to 0.3, and a weight decay of 1e-5. We train our models with a batch size of 32 over 200 epochs. All experiments use early stopping techniques based on validation loss, which greatly reduces training time, especially for attentional target prompt tuning. In the attention module, the temperature used is 2087. The maximum sequence length for the input is set to 256 tokens, and a dropout rate of 0.1 is applied to the classification layer to mitigate overfitting. These hyperparameters, including the optimizer settings and dropout rate, were determined through hyperparameter tuning using Optuna (Akiba et al., 2019). For both setups, the loss function utilized is Cross Entropy.

B Datasets Details

Below is a brief overview of the datasets, including their classes and types.

GermEval 18. This task involves classifying German tweets from Twitter (Risch et al., 2021). It includes two sub-tasks: a coarse-grained binary classification task (normal vs offensive) and a fine-grained multi-class classification task (normal, profanity, insult, abuse.)

HateSpeech 18 prepared by de Gibert et al. (2018) comprises hate speech messages from Stormfront, the prominent white supremacist forum on the web. This English dataset focuses solely on the classification of hate versus normal speech.

OLID. This dataset, gathered from English tweets using specific keywords (Zampieri et al., 2019), categorizes tweets into three types: offensive vs normal, intentional vs unintentional insults, and targets to the individual, group, or others.

Srw 16. Waseem and Hovy (2016) compiled a set of English tweets related to sexism and racism, thus labeled as sexism, racism, or neither.

HASOC 19 introduced by Mandl et al. (2019) includes tweets in Hindi, German, and English. These tweets are marked for three sub-tasks: normal vs offensive, if they are offensive categorized in hate, offensive, and profanity, and intentional vs unintentional. We only considered German and English as two separate tasks.

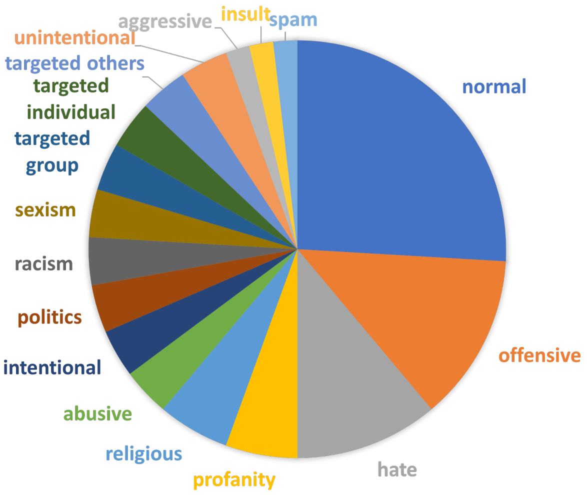


Figure 4: Variety of the labels in our selected hate speech datasets.

B.1 Additional Datasets

We also applied source prompt tuning on the following datasets to obtain more task- or label-specific soft prompts for use in target prompt tuning.

Hateval 19. This topic-specific hate speech detection dataset contains tweets targeting immigrants and women (Basile et al., 2019). It is annotated with three binary sub-tasks: hate vs normal, targeted to group vs individual, and aggressive vs normal. This dataset is in English.

Xdomain. The paper constructs large-scale tweet datasets for hate speech detection in English and Turkish (Toraman et al., 2022). We only considered English tweets in two main domains politics and religion, which are classified into three labels: offensive, hate, and normal.

HateXplain presented by Mathew et al. (2021) and collected from Twitter and Gab. This English dataset includes two sub-tasks. Firstly, classifying posts into hate, offensive, and normal. The second sub-task classifies posts based on the targeted community into five classes: normal, religious, racism, sexism, and others.

Founta et al. (2018) in **Abuse** dataset assembled an English tweet dataset, that includes tweets categorized into offensive, abusive, hateful, aggressive, spam, and normal classes.

B.2 Dataset Statistics

Detailed information regarding the datasets we used is presented in Table 6. In the table, the *size* refers to the total number of data instances. We used the same train, test, and validation sets as provided by the datasets. If no separate sets were provided, we split the data into 20% for testing and 20% of the remaining data for validation, using a random seed of zero. The *imbalance ratio* refers to the number of instances in the minor class divided by the number of instances in the major class. Figure 4 illustrates the variety of labels and their quantities in the datasets used for this study.

C Model Size and Budget

The experiments with T5-large were mostly executed on NVIDIA RTX A6000 servers, while other experiments were primarily conducted on NVIDIA GeForce GTX 1080 Ti. In task-specific soft prompt tuning, the number of tunable parameters is (prompt_length × model_dim), which in our settings is 76,800. However, in label-specific prompt tuning, this number is multiplied by the number of classes (k × prompt_length × model_dim), so for binary classification, it is 153,600. Considering the total number of parameters in T5-base (222,958,848), the percentage of tunable parameters to all parameters is less than 1%.

D Effect of Trainable Parameters

We test whether the performance improvement of our method is only due to the increased number of trainable parameters in our approach. In this experiment, we set the soft prompt length of the task-specific baseline method to be equal to the total number of trainable parameters of label-specific prompt tuning, which is calculated as the number of classes multiplied by 100. For this experiment, we selected one German and four English tasks, due to limitations in computational resources. As depicted in Table 7, even with an increased number of prompt parameters in task-specific prompt tuning, its performance is inferior to label-specific prompt tuning, in some cases even achieving lower performance than a model with fewer parameters.

		Size	**Num-Class**	**Imbalance Ratio**	**Classes**	**Language**
GermEval 18	binary	7539	2	0.508	normal, offensive	de
	fine-grained	7539	4	0.021	normal, profanity, insult, abuse	de
HateSpeech 18	binary	8990	2	0.126	normal, hate	en
HASOC 19 (de)	binary	3905	2	0.119	normal, hateful and offensive	de
	fine-grained	461	3	0.411	hateul, offensive, profanity	de
SRW 16	binary	8401	2	0.358	offensive, normal	en
	fine-grained	8401	3	0.001	sexism, racism, normal	en
OLID	offensive	11452	2	0.498	normal, offensive	en
	targeted	3760	2	0.135	unintentional, intentional	en
	target	3313	3	0.164	individual, group, others	en
Hateval 19	hate	8200	2	0.725	hate, normal	en
	aggressive	3453	2	0.701	aggressive, normal	en
	target	3453	2	0.549	individual, generic	en
HateXplain	hof	17307	3	0.701	hatespeech, normal, offensive	en
	target	15774	5	0.091	religious, racism, sexism, none, others	en
Abuse	fine-grained	38095	4	0.05	abusive, hateful, spam, normal	en
Xdomain	politics	9329	3	0.031	normal, offensive, hate	en
	religion	9209	3	0.025	normal, offensive, hate	en
HASOC 19 (en)	binary	5834	2	0.63	normal, hateful and offensive	en
	fine-grained	2096	3	0.395	hateul, offensive, profanity	en
	targeted	2096	2	0.108	unintentional, intentional	en

Table 6: Detailed information of datasets.

		T5 base			T5 large		
		task		label	task		label
		100	$k \times 100$	$k \times 100$	100	$k \times 100$	$k \times 100$
GermEval 18	binary	62.67	68.18	**70.98**	**73.47**	72.16	72.00
	fine-grained	32.14	36.17	**40.23**	42.62	38.25	**43.25**
HateSpeech 18	binary	73.58	73.82	**76.43**	76.02	**77.67**	76.22
SRW 16	binary	83.11	83.77	**85.27**	85.18	84.23	**85.92**
	fine-grained	57.55	57.04	**58.26**	**58.27**	57.24	58.03
OLID	offensive	77.48	77.90	**78.40**	78.38	77.72	**79.78**
	targeted	47.02	49.16	**61.47**	57.54	57.87	**62.16**
	target	54.68	48.05	**57.51**	46.31	47.58	**51.37**
	average	63.54	63.17	**67.60**	66.33	65.88	**68.21**

Table 7: The effect of the number of trainable parameters in soft prompt tuning. k indicates the number of classes.