Unraveling the Dynamics of Semi-Supervised Hate Speech Detection: The Impact of Unlabeled Data Characteristics and Pseudo-Labeling Strategies

Anonymous ACL submission

Abstract

Despite advances in machine learning based hate speech detection, the need for larges amounts of labeled training data for state-of-the-art approaches remains a challenge for their application. Semi-supervised learning addresses this problem by leveraging unlabeled data and thus reducing the amount of annotated data required. Underlying this approach is the assumption that labeled and unlabeled data follow similar distributions. This assumption however may not always hold, with consequences for real world applications. We address this problem by investigating the dynamics of pseudo-labeling, a commonly employed form of semi-supervised learning, in the context of hate speech detection. Concretely we analysed the influence of data characteristics and of two strategies for selecting pseudo-labeled samples: threshold- and ratio-based. The results show that the influence of data characteristics on the pseudo-labeling performances depends on other factors, such as pseudo-label selection strategies or model biases. Furthermore, the effectiveness of pseudo-labeling in classification performance is determined by the interaction between the number, hate ratio and accuracy of the selected pseudo-labels. Analysis of the results suggests an advantage of the threshold-based approach when labeled and unlabeled data arise from the same domain, whilst the ratio-based approach may be recommended in the opposite situation.

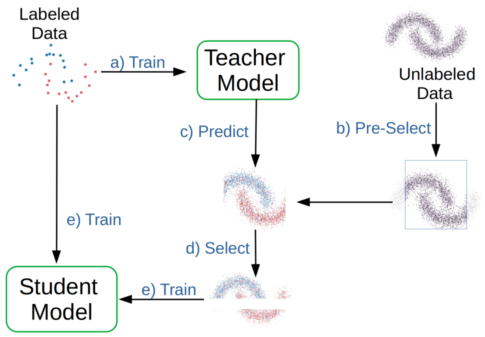


Figure 1: **Pseudo-Labeling Framework.** After teacher model training (a), it is used to predict pseudo-labels (c) for pre-selected unlabeled data points (b). After the selection of reliable pseudo-labels (d), a student model is trained with labeled and pseudo-labeled data (e).

1 Introduction

Topic shifts in online hate speech arising from changing social media trends or news poses a challenge for hate speech detection systems (Florio et al., 2020). In order to keep the pace and follow such dynamic changes developers of such systems need to adapt their models to the continuously changing contexts and linguistic patterns (Ludwig et al., 2022). Since these models rely on large amounts of annotated training data (Challa et al., 2020) the dynamic nature of abusive language in online discourses complicates the application of state-of-the-art deep learning models. Gathering high quality training data is time-consuming and often requires human expertise to be involved in the annotation process (Yang et al., 2022). Semi-supervised learning address these challenges by training models with a small amount of data annotated (labeled) for the specific use case together with a large amount of unlabeled data. These approaches improve model performance over purely supervised learning approaches by using information that is present in the unlabeled data (Van Engelen and Hoos, 2020), and are therefore being actively explored in dynamic domains such as automatic hate speech detection, where data efficiency is crucial.

Since unlabeled data seems to be easy to obtain, recent research in the field of semi-supervised hate speech detection focuses on the learning algorithms themselves rather than the training data. The underlying assumption is that the labeled and unlabeled data share the same characteristics and therefore follow the same data distribution. This assumption however does not hold in real world scenarios where the high pace of change of online hate speech is accompanied by changes in the characteristics of associated data. Therefore, we investigate the influence of data characteristics on semi-supervised model performances. As we investigate pseudo-labeling based semi-supervised learning (Alsafari and Sadaoui, 2021a,b; Ludwig et al., 2022; Zia et al., 2022) we are especially interested in the different benefits regarding model performance of two common pseudo-label selection strategies. In summary, the contributions of this work are:

(i) exploration, how different characteristics of unlabeled data affect the semi-supervised training of hate speech detection models, (ii) clarification of the interaction between characteristics of unlabeled data, model bias and different pseudo-label selection strategies, and (iii) recommendations for real-world applications using pseudo-labeling based approaches for hate speech detection.

2 Related Work

Various approaches for automatic hate speech detection have been proposed in recent years (Jahan and Oussalah, 2023), reaching from lexical (Alkomah and Ma, 2022; Frenda et al., 2019) to traditional machine learning (Waseem and Hovy, 2016; Aziz et al., 2021) to deep learning based approaches (Vashistha and Zubiaga, 2021; Khan et al., 2023; Wadud et al., 2023). Due to the high demand for labeled data of current approaches (Yin and Zubiaga, 2021), semi-supervised training methods have emerged as an active line of research in the context of hate speech detection (Zia et al., 2022; d'Sa et al., 2020; Santos et al., 2022). For instance Zia et al. investigated the use of self-training to improve hate speech detection performance in multilingual settings. Similarly, (Alsafari and Sadaoui, 2021b) used self-training to enhance hate speech detection models, having reported an improvement of 7% relative to supervised baselines. Whilst imbalanced class ratios and the complexities in the detection of implicit hate speech were identified as challenges in the training process, no thorough examination of their impact on the self-training performances was conducted. In a previous study by the same authors (Alsafari and Sadaoui, 2021a), an ensemble of different classification models was trained on a seed hate speech dataset to predict pseudo-labels for a large unlabeled dataset. The authors evaluated various ways to combine predictions from multiple models within the ensemble in order to obtain reliable pseudo-labels. While these works applied pseudo-labeling and other semi-supervised learning techniques to improve hate speech classifiers, they did not analyze how these approaches are affected by typical challenges in the hate speech detection domain. In our work, we thoroughly investigate how data properties, specific to the hate speech domain, and their interaction with other components, such as pseudo-label selection strategies, affect the performance of pseudo-labeling-based approaches.

The influence of different data and pseudo-label characteristics has also been studied in other areas. Wei et al. reported on the negative effect of imbalanced pseudo-labels on model performance. Furthermore, they reported improvements over other pseudo-labeling based approaches by applying an iterative re-balancing framework for pseudo-labels, indicating the importance of a balanced class ratio in the pseudo-labels. The influence of the accuracy of pseudo-labels was investigated in turn by Li et al., in the task of sentiment analysis. The authors found that the accuracy of the pseudo-labels strongly affects model performance. In relation to these works, our work focuses on the specific domain of hate speech detection with its unique challenges. More over, in contrast to previous works we analyse how the interaction of multiple components, such as data and pseudo-label characteristics, model biases and pseudo-label selection strategy affects the performance of the investigated approaches. Based on our findings, we further provide recommendations for real-world applications of semi-supervised learning in the domain of hate speech detection.

3 Methods and Experiments

3.1 Data

We use the dataset created by Kennedy et al. (2020), which is an English hate speech dataset compiled from YouTube, Twitter, and Reddit, and refer to it as *Seed* dataset. The dataset consists of $31,000$ data samples, each annotated with continuous real valued hate scores ranging from -8 to 6, designed to quantify the magnitude of hate. Negative scores indicate "normal" comments, while positive scores denote "hate speech." This unique annotation scheme enables us to study how estimated toxicity and thus magnitude of hate speech impacts the performance of semi-supervised learning algorithms, along with the impact of sample quantity and hate speech ratios. We provide data samples for different toxicity values in appendix A, visualizations and information about the test data and unlabeled data used in this work in the B section.

3.2 Model Architecture

The classifier utilized in this work is composed by a pre-trained *XLM-RoBERTa* model (Conneau et al., 2020) as backbone, followed by a linear layer and a Softmax activation layer. We implemented our models utilizing the deep learning framework *PyTorch*, whereby we especially rely on the pre-trained *XLM-RoBERTa* model provided by the *Transformers* library. [1] In order to reduce memory consumption and to enable the conduction of a larger number of experiments, we trained our models with a parameter efficient finetuning approach by utilizing the *PEFT* library (Mangrulkar et al., 2022). More specifically, we apply the LoRA technique (Hu et al., 2021) with $\alpha = 16$, dropout $p = 0.1$ and a rank $r = 8$.

3.3 Pseudo-Labeling Framework

Pseudo-Labeling is a popular form of semi-supervised learning, involving the following steps (Figure 1):

a) Training of a teacher model Φ on a small amount of labeled data D_L

b) (optionally) Pre-selection of the unlabeled data (e.g. data cleaning)

c) Prediction of pseudo-labels for a larger pool of unlabeled data

d) Selection of reliable pseudo-labels together with their corresponding data samples

e) Training of a student model Θ with labeled and selected pseudo-labeled data

In our study, we examine the following two strategies for selecting pseudo-labels:

3.3.1 Threshold-based selection

Threshold-based approaches select pseudo-labels, for which the prediction confidence of the model is above a pre-defined threshold $\tau \in [0, 1]$. In our work, we set the confidence threshold $\tau = 0.80$.

3.3.2 Ratio-based selection

Ratio-based approaches select the most confident pseudo-labels for each predicted class according to a pre-defined ratio $r \in [0, 1]$. For each predicted class, the top $r \cdot 100\%$ most confident pseudo-labels are selected. We chose a fixed ratio r of 0.1.

3.4 Classifier Fitting

In the first and in the last steps of the pseudo-labeling framework, models are fitted to labeled and pseudo-labeled data respectively. Here, we used two different training approaches for fitting the classifier:

3.4.1 Single-Stage Training

In the single stage training strategy, all trainable model parameters were trained on labeled (or pseudo-labeled) data using the Cross-Entropy loss, which is defined as:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{B} y_i log(p_i) \tag{1}$$

where B corresponds to the minibatch size, y_i to the class label [2] and p_i to the predicted probability of the i^{th} class. We trained our models with a maximal batch size of 256. Parameter optimization was performed using *Adam* (Kingma and Ba, 2014) for 5.000 iterations and a learning rate of $3 \cdot e^{-5}$.

3.4.2 Two-Stage Training

The two-stage training strategy started with the pre-training of the backbone modules via metric learning, since this showed strong results in terms of data efficient learning. The goal of this training stage is to train an encoder $f_\Phi(x) : \mathcal{R}^F \rightarrow \mathcal{R}^D$, which maps data points that belong to the same class to metrically close points in $\mathbf{R}^D$, and vice-versa data points that belong to different classes to metrically distant points in $\mathbf{R}^D$. We used the *XLM-RoBERTa* module as encoder f_Φ and trained it using a triplet loss defined as:

$$\mathcal{L}_{tri}(\Phi) = \sum_{a,p,n} [m + D(x_a, x_p) - D(x_a, x_n)]_+ \tag{2}$$

where x_a is an anchor point, x_p is a positive point belonging to the same class as the anchor point and x_n is a negative point belonging to another class than the anchor point. This loss function ensures that positive points x_p are closer to anchor

[1] https://huggingface.co/docs/transformers/index

[2] In our setups, y_i can also be a pseudo-label

Approach	*F1*	*Precision*	*Recall*	*AUC*
Naive Classifier (ZeroR)	.39	.32	.50	/
Baseline Std.	.67	.67	.67	.74
Baseline Met.	.69	.69	.69	.78
Upper-Bound Std.	.76	.77	.75	.87
Upper-Bound Met.	.72	.74	.71	.84

Table 1: Classification metrics, achieved by a naive zero rate classifier and by the supervised reference models. Baseline models are trained with 200 labeled samples while upper-bound models are trained with over 31.000 samples.

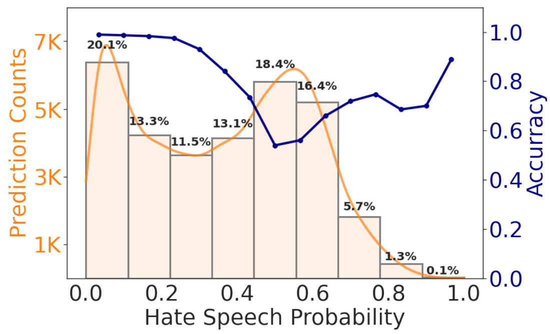


Figure 2: Histogram and accuracy values of our baseline model with respect to hate speech probabilities, which have been computed over all unlabeled data samples of the seed dataset. The model tends to make more predictions in favor of the normal class. Moreover, these predictions have a higher degree of accuracy than the hate speech class.

points x_a than negative points x_n by at least a margin m, given a distance function $\mathcal{D}$. A specific configuration of x_a, x_p and x_n is called a triplet. We employed batch-semi-hard triplet mining (Harwood et al., 2017), which has proven to improve the robustness of training. As distance function $\mathcal{D}$ we used the cosine-distance. In this approach, backbone models were pre-trained for 5.000 iterations with a batch size of 768. We used Adam optimizer (Kingma and Ba, 2014) with a learning rate of $3 \cdot e^{-5}$.

After backbone training, the linear classifier was fitted using Cross-Entropy loss (equation 1) with labeled (or pseudo-labeled) data samples, while freezing the weights of the backbone module. In this step, we again used Adam optimizer (Kingma and Ba, 2014) with a learning rate of $1 \cdot e^{-3}$ and train the linear layer for 100 iterations.

3.5 Model Evaluation

The performance of the classifier was evaluated after each training epoch with the evaluation set. We stored the model that achieved the best macro average *F1*-score on the validation set. After model training we apply beta-calibration (Kull et al.) in order to retrieve reliable predictions from the model. The final model performance reported in this work was computed on a separate test set, which was used only once after completion of all model training, selection and calibration steps.

3.6 Baseline and Upperbound

To estimate the performance of the investigated semi-supervised learning algorithms, we trained reference models in a fully supervised manner. Reference baseline models were trained with 200 labeled data samples, which were later also used as labeled data in the semi-supervised learning experiments. The number of *normal* samples was set equal to the number of *hateful* samples. We trained two baseline models: *Baseline Standard* was trained using the single-stage training approach, while *Baseline Metric* was trained using the two-stage training approach. In addition to models trained with 200 samples, we also trained upper-bound models in which the complete seed dataset was used for training. Also in this case, we performed single-stage training (*Upperbound Standard*) and two-stage training (*Upperbound Metric*).

3.7 Investigation of Data Characteristics

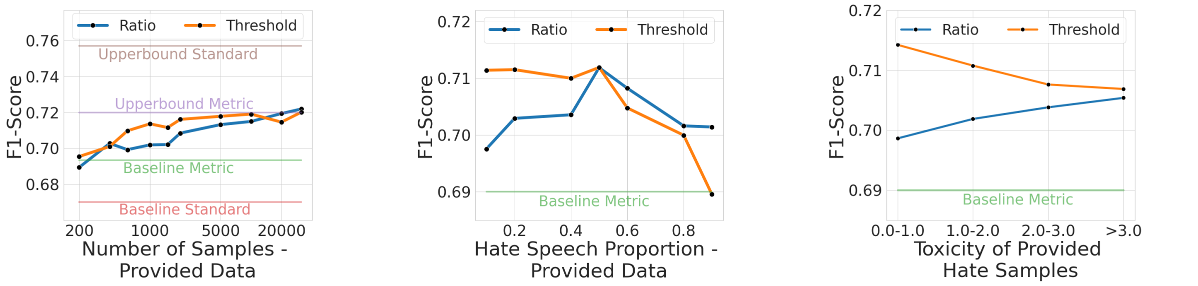
In our experiments, we explored how different characteristics of the unlabeled hate speech data affect the performance of models trained with pseudo-labeling methods. We used subsets of the training data from the *Seed* dataset as unlabeled data, along with 200 labeled data samples, which were also used to train the baseline models. This was done by employing the baseline metric model as teacher model in the pseudo-labeling framework. After that, we used the single-stage training approach for fitting the student models.

3.7.1 Number of unlabeled Samples

In order to investigate the influence of the number of unlabeled samples, subsets of 200, 400, 600, 1000, 1500, 2000, 5000, 10000 and 20000 unlabeled data points were randomly sampled from the original Seed dataset composed by 31453 samples.

3.7.2 Ratio of Hate Speech

To examine the effect of the proportion of hate speech in the unlabeled set, a subset of 1000 unlabeled samples was selected to achieve the required proportion of hate samples. The proportion of hate speech in the unlabeled data was varied from 10%, to 20%, 40%, 50%, 60%, 80%, and 90%.



(a) F1-Score as a function of the number of unlabeled samples for the standard and upperbound approaches as well for the two semi-supervised learning strategies.

(b) F1-score with respect to the proportion of hate speech in the unlabeled data, for the two semi-supervised learning strategies.

(c) F1-Score as a function of the toxicity of unlabeled hate samples, for the two semi-supervised learning approaches.

Figure 3: Effect of characteristics of unlabeled data on model performance for the two semi-supervised training approaches investigated. For a valid comparison, the total number of unlabeled samples in experiments 3b and 3c was fixed to 1.000 samples.

3.7.3 Toxicity of Hate Speech

In this series of experiments, the unlabeled hate samples were selected based on their toxicity level. The following ranges of toxicity were considered: $0.0 - 1.0$, $1.0 - 2.0$, $2.0 - 3.0$, and > 3.0. The ratio of hate speech was set at 0.3, while the total number of samples in all these experiments was set at 1000.

4 Results and Discussion

This section starts by presenting and discussing the results of the supervised reference models, as well as the prediction confidences and pseudo-label accuracies of the baseline metric model for the unlabeled portion of the base dataset. Afterwards we present the performances of the semi-supervised learning approaches with respect to different characteristics of the unlabeled data, and discuss these results in face of the characteristics of the corresponding selected pseudo-labels, the distributions of the predicted hate speech probability and of the annotated toxicity values of the selected hate samples. The section finalises with a summary of the main observations/results.

4.1 Reference Model Performance

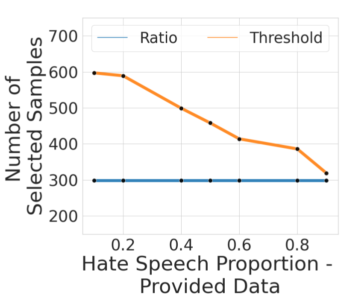
All of our reference models are able to clearly outperform the lowerbound performance, achieved by a naive zero rate classifier. When data resources are low, the metric learning approach outperformed the standard training approach (table 1), showing, inline with results from previous works (Ran et al., 2023; Matsumi and Yamada, 2021), the effectiveness of metric learning in few shot settings. *Normal* pseudo-labels (probabilities < 0.5), computed by the baseline metric model (which also served as teacher model in our experiments), showed higher accuracy and average prediction confidence compared to *hateful* pseudo-labels (Figure 2), suggesting a model bias towards the *normal* class. This bias was observed even though the model was trained with balanced data, a behavior also observed in previous studies (Wang et al., 2022). Notably, the bias particularly distorted the prediction of high-confidence pseudo-labels, affecting them more than the average pseudo-labels in terms of quantity and accuracy.

4.2 Influence of Data Characteristics

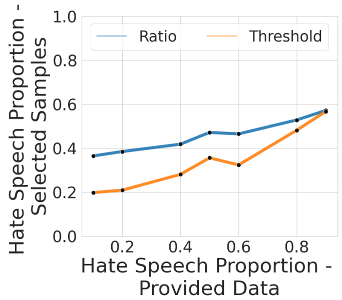
While the positive correlation between the number of unlabeled samples and the performances of the pseudo-labeling approaches (Figure 3a) was expected (Ludwig et al., 2022), the ambiguous influence of the hate ratio and of the toxicity level on model performance was surprising.

4.2.1 Proportion of Hate Speech

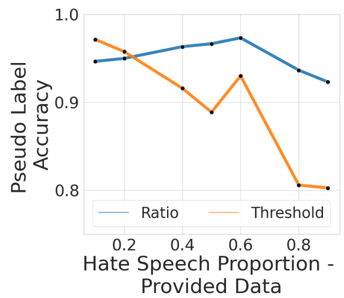
The threshold-based selection strategy achieved reasonable stable performances for hate speech ratios varying from 0.1 to 0.5, but its performance decreased significantly for higher hate speech ratios, achieving partially worse results than the baseline model (Figure 3b, orange curve). The corresponding pseudo-label characteristics (Figures 4a - 4c, orange curves) revealed, that the number and the accuracy of the pseudo-labels selected by the threshold-based approach decreases with increasing proportion of hate speech in the unlabeled samples, while the proportion of hate speech in the selected samples increases. Previous stud-



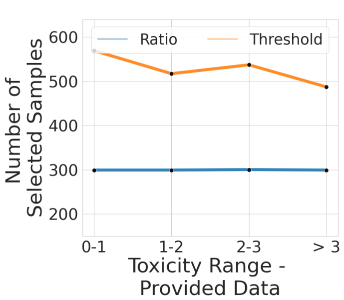
(a) While the number of selected samples remains constant for the ratio-based approach, the number drops with increasing hate ratio for the threshold-based approach.



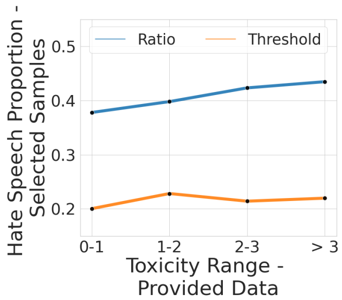
(b) For both selection strategies, the hate ratio in the selected samples increases with increasing ratio in the input samples, with higher values for the ratio-based selection strategy.



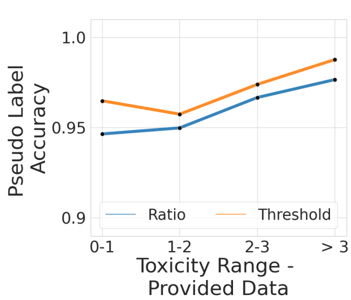
(c) While the pseudo-label accuracy for the threshold-based strategy decreases with the hate fraction in the input samples, it remains almost constant for the ratio-based strategy.



(d) While the number of selected samples slightly drops with increasing hate ratio for the threshold-based approach, the number remains constant for the ratio-based approach.



(e) The hate ratio of the selected data constantly increases with increasing toxicity in the input data for the ratio-based approach and barely increases for the threshold-based approach.



(f) The pseudo-label accuracy in the selected data increases for both, threshold-based and ratio-based selection approaches with increasing toxicity in the input data.

Figure 4: Influence of hate speech characteristics on predicted and selected pseudo-labels.

ies showed the disadvantageous effect of class-imbalanced pseudo-labels (Zou et al., 2018) and the positive impact of increasing pseudo-labels accuracy on model performance (Liu et al., 2022; Rizve et al., 2021), mainly focusing on individual pseudo-labels characteristics. In our opinion, however, the stable performance of the threshold-based approach at low hate ratios cannot be explained by considering the dynamics of the pseudo-label characteristics individually, but by analyzing their interaction. Our results indicate that the increasing proportion of hate speech and thus decreasing class-imbalance in the selected samples (Figure 4b) can to a certain amount compensate for the decreasing number of selected pseudo-labels (4a) and the decreasing accuracy of the pseudo-labels (4c), thus stabilising the performance of the approach at lower hate ratios.

The ratio-based selection approach achieved its best performance when the ratio between *normal* samples and *hateful* samples in the unlabeled data was balanced, but its performance declined when the distribution of the *normal* and *hate speech* classes became unbalanced (Figure 3b, blue curve). In contrast to the performance of the threshold-based approach, the performance drop is observable regardless of which of the classes becomes the majority class. The characteristics of the pseudo-labels, selected by this approach, indicate that the performance is mainly driven by the proportion of hate speech in the selected pseudo-labels (Figure 4b, blue curve), which varied from values below 0.4 to almost 0.6, while the number of selected samples (Figure 4a, blue curve) showed no variation. The best performance of this approach was reached when the proportion of hate/normal speech in the selected pseudo-labels was balanced. The accuracy of the selected pseudo-labels (Figure 4c, blue curve) could support the performance trend, but in our opinion, the hate ratio is the main reason for the performance variation of this approach, as the highest pseudo-label accuracy is not aligned with the strongest results achieved by the approach.

4.2.2 Toxicity of Hate Samples

While the performance of the threshold-based selection approach decreased with increasing toxicity levels of the hate samples, the opposite was observed for the ratio-based selection strategy (Figure 3c). Overall, the threshold-based selection strategy achieved better results than the ratio-based selection strategy across the whole toxicity range.

The superior performance of the threshold-based selection strategy is attributed to its higher number of selected pseudo-labels compared to the ratio-based approach in each experiment (Figure 4d). The threshold-based approach tends to select fewer pseudo-labels as toxicity increases, resulting in decreasing model performance, although the hate ratio and accuracy for these pseudo-labels tend to increase (Figures 4e and 4f, orange curves). Again, the interplay between pseudo-label characteristics determine the performances of the approach. In contrast, the ratio-based approach selected a constant number of pseudo-labels (Figure 4d, blue curve). Its performance improvement with increasing toxicity values is caused by an increasing accuracy and a more balanced hate ratio of the selected pseudo-labels (Figures 4f and 4e, blue curves).

4.3 Interplay of Biases, Data Properties, and Pseudo-Label Selection Strategy

The characteristics of the pseudo-labels selected by the threshold-based approach are more sensitive to the hate speech ratio in the unlabeled data than those selected by the ratio-based approach (Figures 4a - 4c). This can be explained by the fact, that the threshold-based approach relies exclusively on pseudo-labels with high confidence, which are disproportionately affected by the model bias (see section 4.1). Accordingly, the characteristics of the pseudo-labels selected by this approach heavily rely on the proportion of samples favored (in our case the *normal* samples) and disfavored (in our case the *hateful* samples) by the model bias. In contrast, the toxicity of the hate samples does not strongly affect the performance of the threshold-based selection strategy. This indicates, contrary to expectations, that the annotated toxicity does not necessarily correlate with the prediction confidence of the model, since the threshold-based approach does not select more hateful samples with increasing toxicity of these samples. This finding is also supported by the visualizations of the distributions of annotated toxicity values and hate speech probabilities in Figure 5. While the differences in the distributions of the annotated toxicity values are clearly observable, these differences are not reflected in the distribution of high confident pseudo-labels. This demonstrates both the difficulty of quantifying hate speech and the subjectivity of hate speech perception, as toxic samples clearly identified as hate speech by human commentators are not necessarily easily classified as hate speech by the machine learning model. The subjectivity of hate speech perception as well as the difficulty of annotating hate speech has previously been discussed in various studies, such as (Ross et al., 2017; Yin et al., 2023; Waseem, 2016). While differences in high confident pseudo-labels are barely visible, there is a noticeable decrease in the number of wrong pseudo-labels (probability values < 0.5) and, consequently, a reduction in false negatives with increasing toxicity of hate samples, as shown in Figure 5. The decreasing number of false negative pseudo-labels in the ratio-based approach (Figure 4f, blue curve) is accompanied by a growing proportion of hate speech within the selected labels (Figure 4e, blue curve), a trend which is a direct result of the proportional selection of hateful samples based on the number of samples classified as hateful.

4.4 Summary of Main Findings

First, the influence of data characteristics on pseudo-labeling performance is ambiguous and depends on other factors such as pseudo-label selection strategies. While a balanced ratio between normal and hateful samples tends to provide favorable results, it is not possible to make a clear statement about the influence of toxicity in the hate samples without accounting for these factors.

Second, our results indicate that the performance of pseudo-labeling approaches relies on the interaction between several characteristics of selected pseudo-labels, including their total number, hate speech proportion, and accuracy. To understand the performances of the investigated approaches, it is therefore necessary to analyse these characteristics together. Consequently, optimizing only one of these features is not a guarantee of a good final performance. For example, selecting a large number of pseudo-labels, beneficial in principle, could lead to low accuracy, undermining performance,

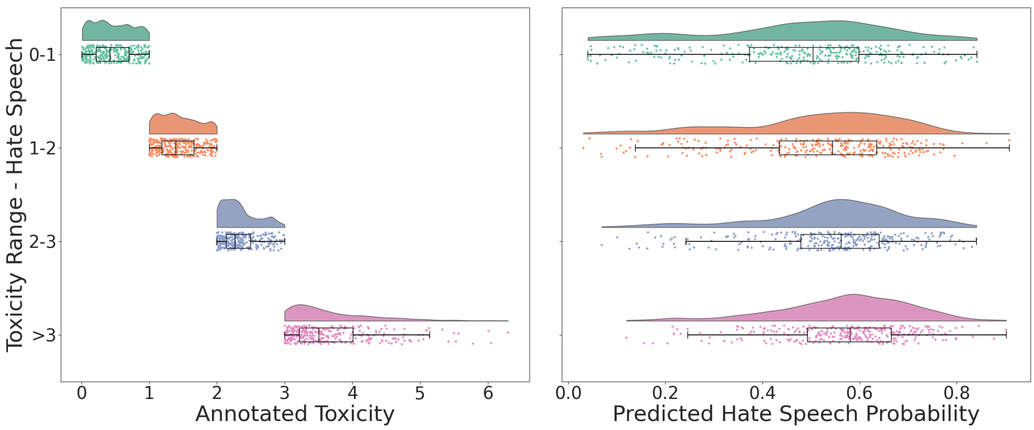


Figure 5: **Raincloud plots** (Allen et al., 2019) of annotated toxicities and predicted hate speech probabilities for different toxicity ranges of hate samples. While the differences in the distributions of the annotated toxicity values are clearly observable, these differences are not reflected in the predicted hate speech probabilities.

and vice versa.

Third, biases of the teacher model affect the threshold-based selection approach more than the ratio-based approach. This leads to superior performance of the threshold-based approach when the data distribution favors the effects of model biases, e.g., when the proportion of majority class in the unlabeled data is high. Conversely, the ratio-based approach outperforms the threshold-based approach in situations where the data distribution is unfavorable to the effects of model biases.

5 Recommendations for Real-World Applications

Our findings suggest, that the threshold-based approach should be applied if the characteristics of unlabeled data favor the effects of the teacher model bias, leading a larger number of confident pseudo-labels. This is typically the case when labeled and unlabeled data arise from the same domain, e.g., when they share the same target groups of hate speech. The ratio-based approach provided better results in opposite scenarios. Especially when domain adaptation is needed due to a lack of labeled data in the target domain, the ratio-based approach should be considered. Prediction confidences can be analyzed, for example, by computing a histogram, which can be a valuable tool for deciding which selection strategy to use. When a large number of confident pseudo-labels are obtained, the threshold-based selection strategy should be preferred, otherwise the ratio-based strategy.

Additionally, given the good model performances achieved for (nearly) balanced data, it is recommended to include a reasonable amount of hate speech in the unlabeled data. Public real-world or synthetic hate speech datasets can be used to this end. Although these datasets may be annotated with different annotation schemes, the "hate" labels contained in these datasets may be similar to the labeled data in the specific use case, and therefore already more "informative" to the model than randomly crawled data, which typically contain a very small amount of hate speech (Meza et al., 2016).

6 Conclusion

In this work, we investigated two pseudo-labeling based approaches for semi-supervised training of hate speech detection models and therefore contributed to the understanding of the complex interaction between data properties, model biases, and pseudo-label selection strategies. We showed that selection of pseudo-labels is determinant to the final performance of the approaches. In view of real-world applications, the results suggest an advantage of threshold-based pseudo-label selection strategies over ratio-based selection strategies when labeled and unlabeled hate speech data arise from the same domain, since a larger number of confident pseudo-labels can be expected in this scenario. In turn, ratio-based selection strategies are preferable when labeled and unlabeled data arise from different domains. These results show the need for further exploration and investigation of alternative pseudo-label selection strategies as well as other families of semi-supervised learning algorithms.

7 Limitations

In this work, we focused on two pseudo-label selection strategies, the threshold-based strategy and the ratio-based strategy. For both strategies, we set the corresponding hyperparameters *threshold* and *ratio* to 0.8 and 0.1, respectively. These values were selected based on the results obtained in preliminary experiments, and allowed us to focus on the effect of other parameters. Investigation of the effect of these hyperparameters, for instance by means of a hyperparameter search, is left to future work. Additionally, while the threshold-based and ratio-based selection approaches are commonly applied and provide clarity in their interaction with model biases and data properties, it is important to note that alternative strategies, such as pseudo-label balancing methods (Wei et al., 2021; Wang et al., 2022) and feature similarity-based selection (Wang and Zhang, 2023), have also been proposed in the literature and deserve further exploration. Moreover, our research focuses exclusively on pseudo-labeling in the domain of semi-supervised learning, leaving out other valuable techniques such as consistency training (Xie et al., 2020; Sohn et al., 2020), variational autoencoders (Gururangan et al., 2019), and GANs (Croce et al., 2020). These approaches may have different responses to the investigated hate speech features and we encourage researchers to explore these approaches since they could provide a more comprehensive understanding of hate speech detection in semi-supervised settings.

References

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information*, 13(6):273.

Micah Allen, Davide Poggiali, Kirstie Whitaker, Tom Rhys Marshall, and Rogier A Kievit. 2019. Raincloud plots: a multi-platform tool for robust data visualization. *Wellcome open research*, 4.

Safa Alsafari and Samira Sadaoui. 2021a. Ensemble-based semi-supervised learning for hate speech detection. In *The International FLAIRS Conference Proceedings*, volume 34.

Safa Alsafari and Samira Sadaoui. 2021b. Semi-supervised self-training of hate and offensive speech from social media. *Applied Artificial Intelligence*, 35(15):1621–1645.

Noor Azeera Abdul Aziz, Mohd Aizaini Maarof, and Anazida Zainal. 2021. Hate speech and offensive language detection: a new feature set with filter-embedded combining feature selection. In *2021 3rd international cyber resilience conference (CRC)*, pages 1–6. IEEE.

Harshitha Challa, Nan Niu, and Reese Johnson. 2020. Faulty requirements made valuable: On the role of data quality in deep learning. In *2020 IEEE Seventh International Workshop on Artificial Intelligence for Requirements Engineering (AIRE)*, pages 61–69.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Danilo Croce, Giuseppe Castellucci, and Roberto Basili. 2020. Gan-bert: Generative adversarial learning for robust text classification with a bunch of labeled examples.

Ashwin Geet d'Sa, Irina Illina, Dominique Fohr, Dietrich Klakow, and Dana Ruiter. 2020. Label propagation-based semi-supervised learning for hate speech classification. In *Insights from Negative Results Workshop, EMNLP 2020*.

Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, 10(12):4180.

Simona Frenda, Bilal Ghanem, Manuel Montes-y Gómez, and Paolo Rosso. 2019. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of intelligent & fuzzy systems*, 36(5):4743–4752.

Suchin Gururangan, Tam Dang, Dallas Card, and Noah A Smith. 2019. Variational pretraining for semi-supervised text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5880–5894.

Ben Harwood, Vijay Kumar BG, Gustavo Carneiro, Ian Reid, and Tom Drummond. 2017. Smart mining for deep metric learning. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2840–2848. IEEE.

Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, page 126232.

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Shakir Khan, Mohd Fazil, Agbotiname Lucky Imoize, Bayan Ibrahimm Alabduallah, Bader M Albahlal, Saad Abdullah Alajlan, Abrar Almjally, and Tamanna Siddiqui. 2023. Transformer architecture-based transfer learning for politeness prediction in conversation. *Sustainability*, 15(14):10828.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Meelis Kull, Telmo de Menezes e Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers.

Changchun Li, Ximing Li, and Jihong Ouyang. 2021. Semi-supervised text classification with balanced deep representation distributions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5044–5053.

Fengbei Liu, Yu Tian, Yuanhong Chen, Yuyuan Liu, Vasileios Belagiannis, and Gustavo Carneiro. 2022. Acpl: Anti-curriculum pseudo-labelling for semi-supervised medical image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20706.

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.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, and Sayak Paul. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

Susumu Matsumi and Keiichi Yamada. 2021. Few-shot learning based on metric learning using class augmentation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 196–201. IEEE.

Radu Meza et al. 2016. Hate-speech in the romanian online media. *Journal of Media Research-Revista de Studii Media*, 9(26):55–77.

Hongyan Ran, Caiyan Jia, and Jian Yu. 2023. A metric-learning method for few-shot cross-event rumor detection. *Neurocomputing*, 533:72–85.

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. 2021. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Raquel Bento Santos, Bernardo Cunha Matos, Paula Carvalho, Fernando Batista, and Ricardo Ribeiro. 2022. Semi-supervised annotation of portuguese hate speech across social media domains. In *11th Symposium on Languages, Applications and Technologies (SLATE 2022)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608.

Jesper E Van Engelen and Holger H Hoos. 2020. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440.

N Vashistha and A Zubiaga. 2021. Online multilingual hate speech detection: Experimenting with hindi and english social media, information 12 (2021). *URL: https://www. mdpi. com/2078-2489/12/1/5. doi*, 10.

Md Anwar Hussen Wadud, MF Mridha, Jungpil Shin, Kamruddin Nur, and Aloke Kumar Saha. 2023. Deepbert: Transfer learning for classifying multilingual offensive texts on social media. *Computer Systems Science & Engineering*, 44(2).

Jie Wang and Xiao-Lei Zhang. 2023. Improving pseudo labels with intra-class similarity for unsupervised domain adaptation. *Pattern Recognition*, 138:109379.

Xudong Wang, Zhirong Wu, Long Lian, and Stella X. Yu. 2022. Debiased learning from naturally imbalanced pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14647–14657.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

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.

Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. 2021. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10857–10866.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268.

Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. 2022. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*.

Wenjie Yin, Vibhor Agarwal, Aiqi Jiang, Arkaitz Zubiaga, and Nishanth Sastry. 2023. Annobert: Effectively representing multiple annotators' label choices to improve hate speech detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 902–913.

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

Haris Bin Zia, Ignacio Castro, Arkaitz Zubiaga, and Gareth Tyson. 2022. Improving zero-shot cross-lingual hate speech detection with pseudo-label fine-tuning of transformer language models. In *Proceedings of the International AAAI conference on web and social media*, volume 16, pages 1435–1439.

Yang Zou, Zhiding Yu, B.V.K. Vijaya Kumar, and Jinsong Wang. 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

OFFENSIVE CONTENT WARNING: The following sections contain examples of hateful content. This is strictly for the purpose of enabling this research. Please be aware that this content could be offensive and cause you distress.

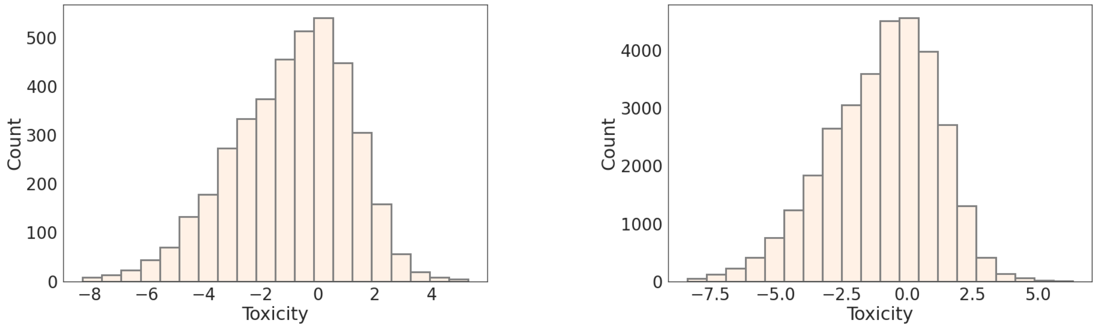
A Example Annotations

In table 2, samples from our seed dataset (Kennedy et al., 2020) together with their annotated toxicity values are shown. The aim of this annotation scheme is to quantify the magnitude of hate speech. Toxicity values < -3 indicate samples, which contain positive supportive speech as well as counter speech against hate speech. Toxicity values between -3 and -2 indicate positive to neutral speech, while values between -2 and -1 indicate offensive speech. Values between -1 and 0 indicate highly offensive comments, while values > 0 indicate hate speech with various degrees of toxicity. While trends are observable, we emphasize the subjectivity in hate speech perception, which allows for different categorizations and assessments of the data samples.

Text Samples	**Toxicity Value**
This is amazing! Kudos for that women for being so strong	-5.94
This is why I love Asians	-4.74
You say we ought to kill more women, I say no.	-3.24
Everything okay here, white men?	-2.42
I mean it's totally possible they were banking on you being gay.	-2.24
Only came here to downvote homophobic comments	-2.20
Just stop trying to f**k her.	-1.71
Or become a white man, the biggest victims in America	-1.68
Oh it is, the only difference is israel uses planes and drones to kill people.	-1.31
Don't need to worry about rape when it only happens to poor people and people of colour	-0.48
Proof positive that women are worse than men: they suck at ra***ng. /s	-0.14
what is this f***ot shit	-0.03
White people are trash.	0.46
I fucking HATE americans. I hope your country BURNS.	0.51
Have fun giving birth to a baby that someone impregnated you by ra**ng you	0.80
Shut up you little b**ch. I'll f**k you	1.19
Inbred, racist as***les.	1.30
shut up retard, go kill yourself	1.40
Fuck off you nazi cunt. Die in a fire.	2.03
Fuck Jews, May Allah Curse Them	2.15
Black and thirty? I'm surprised he wasn't hung by the KKK yet	2.49
Go f**k yourself. Kill yourself worthless piece of sh*t.	3.20
Kill yourself attention seeking fucking f***ot ass cunt	3.65
Put all the Jews in camps again. Make sure this time they will never leave.	4.49
Burn this b**ch down!	4.60
Gas the jews?	5.09
I wanna abuse that worthless wh**e ass for hours	6.01

Table 2: A selection of data samples together with their corresponding annotated toxicity values.

B Data Distributions



(a) Toxicity distribution of our test dataset.

(b) Toxicity distribution of the raw unlabeled dataset.

Figure 6

Figure 6 shows the toxicity distribution of test data (Figure 6a) and unlabeled data (Figure 6b), used in this work. We treat samples with toxicity values > 0.0 as hate speech, otherwise as normal. Given this threshold, the proportion of hate speech in the unlabeled data and in validation data was 0.36. Both distributions are similar, with most samples centered around toxicity values of 0.