Enhancing Training Data Quality through Influence Scores for Generalizable Classification: A Case Study on Sexism Detection

Rabiraj Bandyopadhyay [1,2], **Dennis Assenmacher** [1],
Jose M. Alonso-Moral [2] and **Claudia Wagner** [1,3]
[1] GESIS - Leibniz Institute for the Social Sciences, Germany
[2] Centro Singular de Investigación en Tecnoloxías Intelixentes(CiTIUS),
Universidade de Santiago de Compostela, Spain
[3] RWTH Aachen University, Germany
{rabiraj.bandyopadhyay, dennis.assenmacher, claudia.wagner}@gesis.org, josemaria.alonso.moral@usc.es

Abstract

The quality of training data is crucial for the performance of supervised machine learning models. In particular, poor annotation quality and spurious correlations between labels and features in text dataset can significantly degrade model generalization. This problem is especially pronounced in harmful language detection, where prior studies have revealed major deficiencies in existing datasets.

In this work, we design and test data selection methods based on learnability measures to improve dataset quality. Using a sexism dataset with counterfactuals designed to avoid spurious correlations, we show that pruning with EL2N and PVI scores can lead to significant performance increases and outperforms submodular and random selection. Our analysis reveals that in presence of label imbalance models rely on dataset shortcuts; especially easy-to-classify sexist instances and hard-to-classify non-sexist instances contain shortcuts. Pruning these instances leads to performance increases. Pruning hard-to-classify instances is in general a promising strategy as well when shortcuts are not present.

Warning! This paper contains instances of sexist text to serve as examples

1 Introduction

Selecting a high-quality subset from a dataset has long been a fundamental challenge in machine learning, wherein the objective is to construct an optimal subset from a larger data pool based on a predefined criterion (Sener and Savarese, 2018), while **preserving** or **improving** model performance relative to the original dataset and ensuring data efficiency. In the context of hate speech detection, however, many curated datasets suffer from an over-representation of certain target identities and keywords (Kennedy et al., 2020; Sap et al., 2020; Founta and Specia, 2021; Yu et al., 2024), which limits the generalizability of models trained on them.

The over-representation of target identities and certain words can lead to surface-level correlations between sentence-level patterns (tokens or phrases) and labels that models learn to associate. Such surface-level correlations are called "shortcuts" and they have been observed in the context of harmful language detection datasets, see e.g. (Sap et al., 2019). Prior work that has investigated the problem of model's over-reliance on shortcuts has mainly focused on gradient-based techniques (Bastings et al., 2022; Pezeshkpour et al., 2022) that do not capture patterns present in the training data and are known to suffer from stability issues (Basu et al., 2021a; Epifano et al., 2023). Complementary to these approaches, corpus-level methods proposed by Gururangan et al. (2018) and adopted by Ramponi and Tonelli (2022) aim to identify token-level shortcuts via manual annotation, a process that is often time-consuming and resource-intensive.

Unlike prior work, this study uses grammar induction (Friedman et al., 2022) to identify dataset shortcuts and explores how they affect the learnability of data points during fine-tuning. We quantify the learnability of data points using Influence Scores. Influence Scores are metrics that measure how learnable or difficult a training data point is for a model, based on information theory, loss gradients, and training dynamics (Anand et al., 2023).

We focus on two widely used Influence Scores: Pointwise V-Information (PVI) (Ethayarajh et al., 2022) and Error L2-Norm (EL2N) (Paul et al., 2021). PVI measures the learnability or difficulty of a data point from an Information Theoretic perspective while EL2N quantifies the difficulty of a data point from a margin-based perspective as we elaborate in Sections 3.1 and 3.2. We present a case study that focuses on a comprehensive dataset (Sen et al., 2022) that has been designed for the task of sexism detection and includes counterfac-

tually augmented data to avoid spurious correlations. In our experiments we investigate how different data selection methods that are based on these Influence Scores affect the performance of BERT-based classifiers namely BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) on in-distribution and out-of-distribution settings. We contrast our designed data selection strategies with submodular maximization (Krause and Golovin, 2014), an automated method that has been used to successfully select a diverse and representative subset of data points from a datasets (Kothawade et al., 2022a; Renduchintala et al., 2023; Mualem et al., 2023). Our results show that removing potentially noisy data points from the training dataset based on EL2N scores leads to more statistically significant performance improvements than pruning data using PVI, submodular maximization, or random selection on out-of-distribution test data. Furthermore, we find that proportionally pruning hard-to-classify instances from both classes based on PVI and EL2N scores results in performance gains across both in-distribution and out-of-distribution test datasets.

We complement our quantitative analysis with qualitative insights on data points containing shortcuts. Our results show that both easy-to-classify sexist instances and hard-to-classify non-sexist instances contain shortcuts. Pruning these instances leads to performance increases. Pruning hard-to-classify instances is in general a promising strategy also when shortcuts are not present. Hard-to-classify instances with shortcuts are often minimal-edit counterfactuals, while hard-to-classify instances without shortcuts often contain spelling mistakes or lack context.

We hope our research informs future work on combining pruning methods based on Influence Scores with shortcut induction methods, since the case study suggests that this is a promising direction to improve the quality of training data for machine learning models applied to the identification of subjective tasks like sexism. We release our code and datasets for reproducibility purposes in the following link.

2 Related Work

The selection of data points from large datasets by quantifying their importance has been a long-standing challenge in Machine Learning. Influential publications by Koh and Liang (2020) and Han et al. (2020a) have extended the concept of Robust Statistics (Hampel et al., 2005) to Deep Learning by introducing the so-called Influence Functions. Such functions measure the effect of removing a data point during training on test predictions. They have been scaled to Large Language Models (LLMs) for quantifying data influence for pre-trained LLMs (Choe et al., 2024; Grosse et al., 2023), selecting data for pre-training (Wang et al., 2023) and Low-Rank Adaptation (LoRA)-based fine-tuning(Kwon et al., 2024).

Alternative Influence Scores have been proposed to assess data importance by quantifying learnability, including training dynamics-based scores such as Forgetting Scores (Toneva et al., 2019) and Error L2-Norm (EL2N) (Paul et al., 2021), gradient-based scores like Variance of Gradients (VoG) (Agarwal et al., 2022) and Gradient Normalized (GraNd) (Paul et al., 2021), Neural Tangent Kernel based scores (Jacot et al., 2020) such as TracIn (Pruthi et al., 2020), and information-theoretic scores like Pointwise V-Information (PVI) (Ethayarajh et al., 2022).

Influence scores have been leveraged for dataset pruning in both NLU and generative tasks. In NLU, EL2N and VoG scores have been used to prune instances from datasets like SNLI (Bowman et al., 2015) and AGNews, leading to maintained or improved performance (Fayyaz et al., 2022; Anand et al., 2023). In machine translation, cross-entropy, BLEU score, and recently proposed CAT score (Checkpoints Across Time) (Chimoto et al., 2024) have been employed to prune datasets such as WMT16 En-Ro and En-Tr, enabling more efficient training (Azeemi et al., 2023; Chimoto et al., 2024).

Data selection using Influence Scores has been extended to instruction tuning datasets. Xia et al. (2024) adapted TracIn scores to select instances from diverse instruction datasets, showing that training on only 5% of the data improved generalization to out-of-domain instructions. Zhang et al. (2025) proposed a pruning strategy using a variant of GraNd and EL2N scores for task-specific core-sets, termed Speculative Selection. Additionally, Zhang et al. (2024a) used PVI to filter query-response pairs for Direct Preference Optimization (DPO) (Rafailov et al., 2023), demonstrating performance improvements in preference-tuning tasks. For a broader overview of core-set selection in In-Context Learning (Brown et al., 2020), instruction tuning (Zhang et al., 2024b), and preference tuning (Ouyang et al., 2022), readers may refer to Albalak

et al. (2024).

In the realm of sexism detection Bandyopadhyay et al. (2024) applied influence-based scoring methods namely PVI, EL2N, and VoG to a combination of sexism detection datasets. They found that pruning up to 50% of the hardest-to-learn data points did not affect model performance in either in-distribution or out-of-distribution evaluations. However, their work does not compare different data selection strategies based on both automated and Influence Scores based, nor does it analyze how dataset shortcuts affect the learnability of individual examples during fine-tuning. A similar limitation applies to Anand et al. (2023), who proposed brute-force pruning strategies for the SNLI dataset aimed at data-efficient fine-tuning.

In this paper, we focus on PVI and EL2N, as they represent learnability from information-theory and training dynamics perspectives respectively. Moreover, they do not suffer from stability issues like gradient-based scores such as TracIn and VoG (Basu et al., 2021b; Epifano et al., 2023). Unlike previous work we analyze data points qualitatively by focusing on how shortcuts affect the chosen Influence Scores for a training data point. Additionally, we design pruning strategies using the score values while at the same time considering label imbalance. We compare our pruning strategies with submodular maximization, which selects a diverse and representative subset; and random selection.

3 Methods

We introduce Influence Scores, which we use to quantify the learnability of individual data points and to design various data selection strategies. Additionally, we include submodular maximization as an automated baseline method for data selection. Finally, we present a grammar induction approach to uncover structural patterns in the data.

3.1 Pointwise V-Information

Pointwise V-Information (PVI) (Ethayarajh et al., 2022) is an information-theoretic metric that measures the **usable bits** of information for a model in predicting the corresponding label of a data point. It extends the Predictive V-Information metric proposed in (Xu et al., 2020) to understanding the difficulty of text data for classification. The metric is calculated by training/fine-tuning two models (g' and g), one on input-target pairs (i.e., a combination of text inputs (x) and labels (y)) and one on target labels (i.e., a combination of null inputs (ϕ) and labels (y)). PVI measures the ease with which a model can predict a certain label given an input by calculating the following quantities for each data point x.

$$pvi(x) = -log_2 g[\phi][y] + log_2 g'[x][y] \quad (1)$$

A negative PVI indicates that the instance was **hard** for the model to classify, and the probability of misclassification increases as the PVI score becomes more negative. Conversely a positive PVI score for a data point increases the odds of correct classification of that data point by the model, hence these instances are considered **easy**.

3.2 Error L2-Norm

Error L2-Norm (EL2N), proposed by Paul et al. (2021), assigns scores to data points based on their classification confidence in early training epochs. It estimates how easy or hard a sample is to classify using the norm of the predicted probabilities relative to the true label. As a margin-based metric, lower EL2N scores indicate greater distance from the decision boundary, suggesting the data point is easier to classify. Higher scores imply proximity to the boundary and greater classification difficulty. Mathematically for a given data point x, EL2N is calculated by

$$\|softmax(g(x)) - y\|_2 \quad (2)$$

where g denotes the model. In Equation 2, y is the one-hot encoding of the label and $softmax(g(x))$ represents the probability after applying a softmax function to the model logits. Prior work has shown that the EL2N score is also adept at surfacing useful data points from training data when applied to natural language settings in case that a BERT model is trained for at least 5 epochs (Fayyaz et al., 2022; Anand et al., 2023).

3.3 Submodular Maximization

Submodular maximization encompasses a family of automated data selection methods that aids in selecting an informative and diverse subset, from a larger super-set. We are interested in selecting a diverse core-set from the training data (which is our super-set). Hence, we choose Facility Location (Krause and Golovin, 2014; Renduchintala et al., 2023; Mualem et al., 2023) that intervenes at the embedding level and selects a diverse group

of data points while maintaining the diversity and representativeness of the original dataset. More details about the implementation of submodular maximization are in Section 4.2.3. We refer the reader to Appendix A.3 for a detailed discussion on submodularity and submodular maximization.

3.4 Dataset Shortcuts

Shortcuts are surface-level correlations between sentence-level patterns (tokens or phrases) and labels that humans assign to the sentence. Shortcuts are often a consequence of annotation artifacts or the so-called spurious correlations (Gururangan et al., 2018; Friedman et al., 2022). Shortcuts are problematic if they do not generalize to test distributions since they may lower out-of-distribution test data performance. Unlike prior work (see Bastings et al. (2022); Pezeshkpour et al. (2022); Ramponi and Tonelli (2022)), we do not focus only on token-level correlations using Integrated Gradients (Sundararajan et al., 2017) or Influence Functions (Han et al., 2020b). We are interested in finding shortcuts (both token and phrase level patterns) that the model can use during fine-tuning to generalize from the training data.

To identify shortcuts in our data we train a Probabilistic Context Free Grammar (PCFG) model, following the approach in Friedman et al. (2022). A PCFG aids in inferring the grammatical structure of the data points in the dataset. We extract the subtrees (aka linguistic feature structures) that are highly correlated with the positive minority class (sexist).We qualitatively examine the relationship between shortcuts in our training dataset and the chosen Influence Scores (PVI and EL2N) to shed light on the difficulty of data points containing shortcuts for the model.

4 Experimental Setting

We design different data selection strategies based on PVI and EL2N and compare them with random selection and submodular maximization. To assess the effectiveness of these strategies we train sexism classifiers on a sexism-dataset and on several pruned variants, and compare their performance across multiple test datasets. In the following, we describe our data, data selection strategies, and experiments in detail.

4.1 Datasets

For our experiments, we use the Call Me Sexist But (CMSB) dataset as training data (Samory et al., 2021) because it covers different theoretical dimensions of sexism, provides at least five annotations per instance and contains counterfactual augmentations through minimal edits (e.g., removing negations, replacing gendered groups inter-alia).

Test-Data. To compare performance and assess the generalizability of our pruning methods, we use the following datasets for out-of-distribution evaluation: Explainable Detection of Online Sexism (EDOS) (Kirk et al., 2023), sEXism Identification in Social neTworks task dataset (EXIST) (Rodríguez-Sánchez et al., 2022) and HateCheck (Röttger et al., 2021). For HateCheck, we extract the examples targeting the identity woman. We refer to this subset as HateCheck(Sexism) throughout the paper. The dataset statistics are reported in Table 1. We present the distribution of PVI and EL2N scores for the sexist and non-sexist subsets of the CMSB (Train) split in Figure 1. The histogram plots indicate that data points with a PVI score above the mean and those with an EL2N score below the mean are more likely to be correctly classified. This suggests that the scores capture how easy a given data point is for the model to learn.

Dataset	Non-Sexist	Sexist
CMSB (Train Split)	8,272	1,269
CMSB (Test Split)	3,550	540
HateCheck (Sexism)	373	136
EXIST	1,800	1,636
EDOS	15,146	4,854

Table 1: **Dataset statistics:** Number of sexist and non-sexist instances per dataset used in our experiments.

4.2 Experiments

This section outlines the data selection strategies and the classification experiments that have been considered in our evaluation. Anand et al. (2023) demonstrated that pruning up to 45% of the SNLI training dataset does not harm model performance. Inspired by this, we prune up to 60% of our data. The specific pruning strategies are detailed in Sections 4.2.1 and 4.2.2.

4.2.1 Informed Undersampling

To address the class imbalance present in our dataset (see Table 1), we introduce a data selection strategy called Informed Undersampling. In this approach, instances from the majority class (non-sexist) are selectively removed based on either PVI or EL2N scores. The training data is sorted by PVI

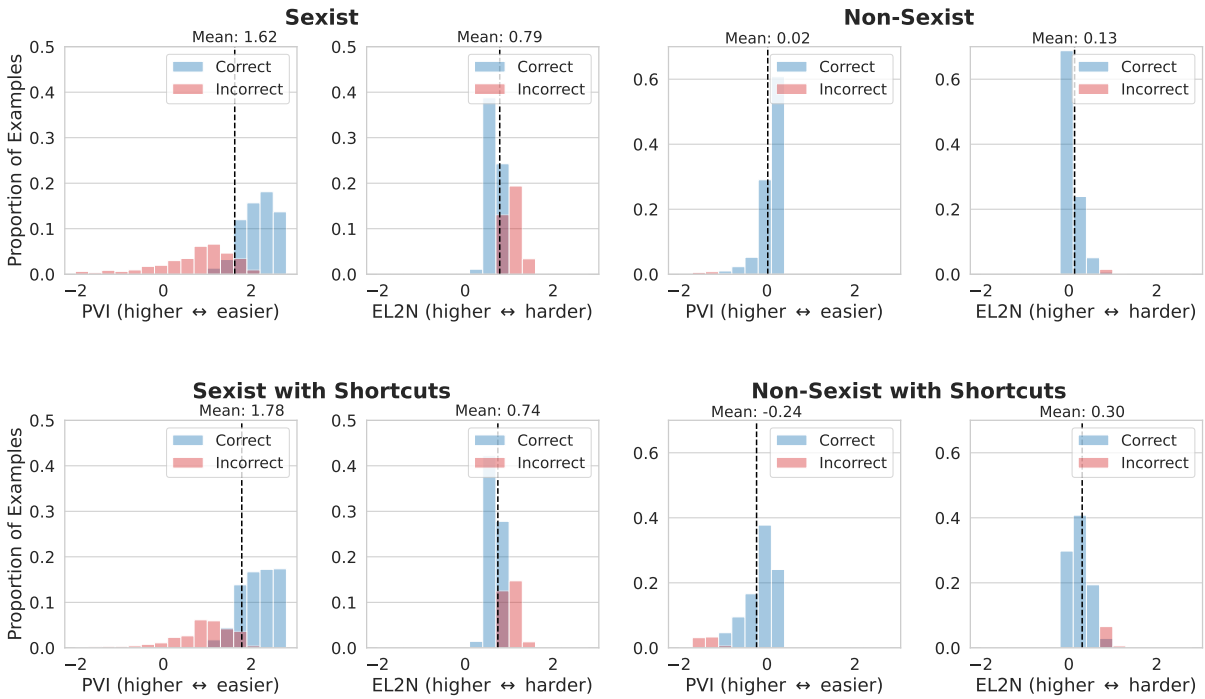

Figure 1: **Distribution of PVI and EL2N scores:** The first row shows the distribution of all sexist and non-sexist instances in the CMSB training data, while the second row shows the distribution of instances that contain shortcuts. Dashed lines indicate the mean values of each distribution. Instances correctly classified in at least 3 out of 5 training runs are labeled as "correct" (blue), while those misclassified in more than 3 runs are labeled as "incorrect" (red). For PVI, scores above the mean correspond to correctly classified ("easy") instances, while scores below the mean indicate misclassified ("hard") ones. A similar pattern is observed for EL2N, where lower scores correspond to easier instances. This trend holds for both sexist and non-sexist data. The PVI and EL2N distributions for instances containing identified shortcuts (see examples in Table 5) shows that most sexist instances with such shortcuts are correctly classified, with PVI scores above 1.78 and EL2N scores below 0.77. This suggests that the model leveraged these shortcuts during fine-tuning. In contrast, non-sexist instances with shortcuts exhibit lower PVI (-0.24) and higher EL2N (0.30) means. This indicates that on average the sub-set of non-sexist instances with shortcuts are harder than the super-set of non-sexist instances.

scores in ascending order (from hardest to easiest) and by EL2N scores in descending order (also from hardest to easiest). We evaluate four variants: **IU-PVI-H**, **IU-PVI-E**, **IU-EL2N-H**, and **IU-EL2N-E**. Here, **IU** denotes *Informed Undersampling*, the middle component in each variant name indicates the influence score used (PVI or EL2N), and the final letter specifies whether hard (H) or easy (E) non-sexist instances are pruned. The remaining non-sexist instances are then combined with the full set of sexist instances to form the training data used for classification.

4.2.2 Proportional Pruning

This strategy evaluates how pruning equal proportions of instances from both the majority (non-sexist) and minority (sexist) classes while preserving their original class ratio affects model performance in both in-distribution and out-of-distribution settings. Following Tänzer et al. (2022), we ensure that each class retains at least 100 instances to prevent drastic performance drops during BERT fine-tuning.

Data points are ranked from highest to lowest (i.e., hardest to easiest) based on EL2N scores, and from lowest to highest (also hardest to easiest) for PVI scores. **PP** stands for *Proportional Pruning*. We use the placeholder **inf_score** to indicate the scoring method (i.e., either PVI or EL2N). Thus, we define four proportional pruning variants that differ in which portions (E = Easy and H = Hard) of each class are removed:

- **PP-inf_score-EE**: prunes easy instances from both classes.
- **PP-inf_score-EH**: prunes easy instances from the sexist class and hard instances from the non-sexist class.
- **PP-inf_score-HE**: prunes hard instances from the sexist class and easy instances from the non-sexist class.
- **PP-inf_score-HH**: prunes hard instances

Method	EXIST			EDOS			Hatecheck		
	15%	30%	60%	15%	30%	60%	15%	30%	60%
Unpruned (0%)		$40.5_{\pm 5.0}$			$53.4_{\pm 6.4}$			$46.5_{\pm 10.5}$	
Random									
random	$38.2_{\pm 4.5}$	$36.6_{\pm 2.9}$	$34.6_{\pm 0.3}$	$49.8_{\pm 5.4}$	$48.5_{\pm 5.5}$	$43.8_{\pm 0.9}$	$34.8_{\pm 8.6}$	$36.8_{\pm 12.9}$	$22.9_{\pm 1.8}$
Submod. Max.									
max	$37.3_{\pm 3.4}$	$35.4_{\pm 1.9}$	$34.3_{\pm 0.0}$	$50.0_{\pm 5.4}$	$45.2_{\pm 2.7}$	$43.1_{\pm 0.0}$	$40.2_{\pm 9.6}$	$27.1_{\pm 7.0}$	$21.1_{\pm 0.1}$
Informed									
IU-PVI-H	$54.6^{*}_{\pm 4.5}$	$55.8^{*}_{\pm 2.0}$	$48.4_{\pm 2.0}$	$58.6_{\pm 3.0}$	$52.4_{\pm 4.3}$	$35.3_{\pm 4.6}$	$46.8_{\pm 5.6}$	$41.6_{\pm 8.8}$	$38.2_{\pm 8.6}$
IU-PVI-E	$38.1_{\pm 2.9}$	$35.5_{\pm 1.3}$	$34.7_{\pm 0.2}$	$50.9_{\pm 5.3}$	$47.2_{\pm 4.0}$	$45.0_{\pm 1.4}$	$43.7_{\pm 10.5}$	$40.7_{\pm 9.5}$	$27.4_{\pm 5.6}$
IU-EL2N-H	$58.5^{*}_{\pm 4.5}$	$61.1^{*}_{\pm 2.0}$	$56.3_{\pm 2.0}$	$61.0^{*}_{\pm 0.3}$	$52.8_{\pm 1.8}$	$36.2_{\pm 2.3}$	$51.7_{\pm 2.1}$	$46.6_{\pm 1.6}$	$43.7_{\pm 1.2}$
IU-EL2N-E	$41.1_{\pm 3.8}$	$35.1_{\pm 1.3}$	$34.7_{\pm 0.2}$	$53.6_{\pm 5.3}$	$46.2_{\pm 4.0}$	$45.0_{\pm 1.4}$	$44.5_{\pm 10.2}$	$31.4_{\pm 10.0}$	$30.0_{\pm 8.3}$
Proportional									
PP-PVI-EE	$38.6_{\pm 3.8}$	$34.3_{\pm 0.0}$	$34.3_{\pm 0.0}$	$52.4_{\pm 5.4}$	$43.0_{\pm 0.0}$	$43.0_{\pm 0.0}$	$43.9_{\pm 4.5}$	$21.1_{\pm 0.1}$	$21.0_{\pm 0.0}$
PP-PVI-EH	$51.3^{*}_{\pm 7.6}$	$55.7^{*}_{\pm 7.4}$	$54.7_{\pm 5.4}$	$59.1^{*}_{\pm 7.6}$	$59.9^{*}_{\pm 2.3}$	$57.8_{\pm 1.8}$	$49.1_{\pm 4.5}$	$52.0_{\pm 3.4}$	$47.4_{\pm 2.1}$
PP-PVI-HE	$38.1_{\pm 2.9}$	$36.7_{\pm 1.0}$	$36.1_{\pm 1.0}$	$51.0_{\pm 4.5}$	$48.7_{\pm 1.9}$	$48.3_{\pm 4.0}$	$42.3_{\pm 9.7}$	$41.8_{\pm 9.3}$	$40.2_{\pm 11.0}$
PP-PVI-HH	$45.7_{\pm 7.8}$	$58.7^{*}_{\pm 1.7}$	$54.1_{\pm 2.5}$	$56.2_{\pm 4.4}$	$61.0^{*}_{\pm 0.4}$	$48.3_{\pm 4.2}$	$46.9_{\pm 3.3}$	$51.2_{\pm 1.8}$	$47.6_{\pm 0.6}$
PP-EL2N-EE	$39.4_{\pm 4.9}$	$35.9_{\pm 2.9}$	$35.2_{\pm 1.6}$	$52.4_{\pm 5.4}$	$43.0_{\pm 0.0}$	$44.4_{\pm 2.6}$	$38.4_{\pm 13.6}$	$28.0_{\pm 9.2}$	$21.1_{\pm 0.0}$
PP-EL2N-EH	$43.9_{\pm 7.7}$	$58.2^{*}_{\pm 2.4}$	$57.5_{\pm 4.1}$	$55.9_{\pm 5.9}$	$61.4^{*}_{\pm 0.8}$	$55.5_{\pm 2.2}$	$49.0_{\pm 6.8}$	$50.7_{\pm 3.0}$	$46.1_{\pm 1.9}$
PP-EL2N-HE	$37.7_{\pm 2.4}$	$36.9_{\pm 1.3}$	$37.2_{\pm 2.1}$	$49.5_{\pm 4.1}$	$50.5_{\pm 4.4}$	$51.6_{\pm 4.2}$	$39.0_{\pm 1.2}$	$42.6_{\pm 12.7}$	$42.7_{\pm 6.5}$
PP-EL2N-HH	$51.6^{*}_{\pm 6.5}$	$52.8^{*}_{\pm 8.1}$	$50.3_{\pm 4.4}$	$60.9^{*}_{\pm 2.1}$	$59.3^{*}_{\pm 4.6}$	$57.0_{\pm 2.7}$	$52.8_{\pm 1.5}$	$51.3_{\pm 4.2}$	$46.4_{\pm 3.0}$

Table 2: **BERT out-of-distribution performance:** We showcase the Macro-F1 Scores of models trained on original (i.e., unpruned) and pruned CMSB train data and tested on EXIST, EDOS and Hatecheck datasets. The pruning rates indicate the amount of training data that was removed from the CMSB data before the model was trained. The rows correspond to different selection methods. IU denotes *Informed Undersampling*, PP stands for *Proportional Pruning*, H = Hard, E = Easy. We performed Wilcoxon test on the Macro F1-Scores and marked with color (Yellow and Lavender for PVI and EL2N) and stars those scores which were statistically significant higher than the F1 score computed with the unpruned data (see row 1), with p-values below the significance level ($\alpha = 0.05$). The results show that EL2N leads to more statistical significant improvements than PVI, submodular maximization and random pruning. Pruning the hard instances based on EL2N leads to performance gain across EXIST and EDOS but does not carry over to Hatecheck (Sexism). Moreover, we also observe that proportionally pruning the easy sexist and hard non-sexist examples (based on EL2N and PVI) leads to better generalization across EDOS and EXIST datasets.

from both classes.

4.2.3 Submodular Maximization

Submodular maximization operates at the representation level. Hence, we first extract embeddings by pooling from the last hidden layer of each of the fine-tuned BERT models across 5 runs. We then perform submodular maximization based data selection on each of the extracted embeddings, and construct the corresponding subsets. We then further fine-tune BERT classifiers on these subsets and results are averaged across in-distribution and out-of-distribution tests. We model the selection problem as a Facility Location problem and use Lazier-than-Lazy-Greedy strategy (Mirzasoleiman et al., 2014), implemented in the PRISM package (Kothawade et al., 2022b). Figure 2 (Appendix A.3) visualizes dataset distributions before and after selection using UMAP (McInnes et al., 2020) for dimensionality reduction.

4.3 Classification Experiments

We focus on BERT and RoBERTa models because they are widely used in subjective NLP tasks such as hate speech detection, where they have demonstrated competitive performance compared to LLMs (Zhang et al., 2023; Ziems et al., 2024; Pan et al., 2024; Sariyanto et al., 2025). We train the models on different subsets of the CMSB training dataset. Because of the dataset's imbalance, we

Acronym	Full Form
IU-inf_score-H	Informed Undersampling-inf_score(PVI or EL2N)-Hard
IU-inf_score-E	Informed Undersampling-inf_score(PVI or EL2N)-Easy
PP-inf_score-EE	Proportional Pruning-inf_score(PVI or EL2N)-Easy-Easy
PP-inf_score-EH	Proportional Pruning-inf_score(PVI or EL2N)-Easy-Hard
PP-inf_score-HE	Proportional Pruning-inf_score(PVI or EL2N)-Hard-Easy
PP-inf_score-HH	Proportional Pruning-inf_score(PVI or EL2N)-Hard-Hard

Table 3: **Index:** We provide index for understanding the acronyms used in our tables

Method	15%	30%	60%
Unpruned (0%)		$73.4_{\pm 6.3}$	
Random			
random	$69.2_{\pm 6.7}$	$63.6_{\pm 9.5}$	$54.6_{\pm 6.4}$
Submod. Max			
max	$71.1_{\pm 4.4}$	$54.8_{\pm 9.3}$	$47.7_{\pm 2.4}$
Informed			
IU-PVI-H	$75.9_{\pm 2.6}$	$69.6_{\pm 4.1}$	$59.2_{\pm 5.2}$
IU-PVI-E	$70.6_{\pm 5.3}$	$66.3_{\pm 2.2}$	$57.0_{\pm 6.1}$
IU-EL2N-H	$78.2_{\pm 1.6}$	$71.2_{\pm 1.5}$	$64.2_{\pm 0.8}$
IU-EL2N-E	$73.6_{\pm 5.9}$	$61.0_{\pm 6.9}$	$59.5_{\pm 7.1}$
Prop. Prune			
PP-PVI-EE	$70.7_{\pm 5.6}$	$46.6_{\pm 0.3}$	$46.4_{\pm 0.0}$
PP-PVI-EH	$76.9_{\pm 5.1}$	$76.6_{\pm 3.2}$	$72.3_{\pm 2.5}$
PP-PVI-HE	$68.5_{\pm 7.7}$	$67.2_{\pm 1.9}$	$66.3_{\pm 6.2}$
PP-PVI-HH	$73.7_{\pm 4.5}$	$78.8^{*}_{\pm 0.6}$	$71.7_{\pm 1.0}$
PP-EL2N-EE	$65.1_{\pm 14.1}$	$52.9_{\pm 10.6}$	$49.7_{\pm 6.6}$
PP-EL2N-EH	$73.4_{\pm 5.6}$	$78.2_{\pm 1.4}$	$71.7_{\pm 1.7}$
PP-EL2N-HE	$69.0_{\pm 4.2}$	$66.2_{\pm 8.6}$	$69.6_{\pm 4.0}$
PP-EL2N-HH	$78.4^{*}_{\pm 3.9}$	$76.6_{\pm 4.5}$	$70.3_{\pm 3.1}$

Table 4: **BERT in-distribution performance:** We showcase the Macro-F1 Scores of models trained on original (i.e., unpruned) and pruned CMSB train data and test on CMSB hold-out data. We perform Wilcoxon test on the Macro F1-Scores and mark and color scores (Yellow and Lavender for PVI and EL2N and stars) which were statistically significant higher than the F1 score of the unpruned data (see row 1) with p-values below the significance level ($\alpha = 0.05$). We observe statistically significant performance gains on pruning hard sexist and hard non-sexist data points based on PVI and EL2N. For all other pruning types we don't see any statistically significant gains.

split it into train and test sets (70/30) using stratified sampling. For comparing the performance of the classifiers trained on the CMSB (Train split) and several pruned variants of this data, we use the Macro F1-Score.

5 Results

This section presents the results of our pruning experiments and analyzes the relationship between dataset shortcuts and Influence Scores. Interested readers can refer to Section A.4 for RoBERTa results. For convenience we also refer the readers to Table 3 to facilitate understanding of the tables.

5.1 Proportional Pruning

By proportionally pruning 15% to 30% of hard-to-classify instances from both classes (sexist and non-sexist), using either PVI or EL2N scores, we achieve statistically significant performance improvements both in-distribution and out-of-distribution (see Table 2) test datasets. This is the only method that consistently improves performance, with gains observed on the CMSB test split as well as on the EXIST and EDOS datasets. The qualitative analysis reveals that hard-to-classify sexist instances often contain spelling mistakes or lack contextual information, while hard-to-classify non-sexist instances are often minimal-edit counterfactuals (see examples in Tables 10 and 11). Our case study suggests that PVI and EL2N help identifying such cases.

Additionally, we observe that pruning easy-to-classify sexist instances based on PVI and EL2N scores, along with hard-to-classify non-sexist instances yields statistically significant performance improvements on the out-of-distribution datasets (EXIST and EDOS). The qualitative analysis on dataset shortcuts shows that especially sexist easy-to-classify and non-sexist hard-to-classify instances contain shortcuts (see Tables 9 and 10).

It is interesting to note, that in both cases we see that pruning hard-to-classify non-sexist (majority) instances is a promising strategy. We observe that these instances are often minimal-edit counterfactuals and contain shortcuts that are highly correlated with the sexist class. Removing such instances encourages class separation. As with previous experiments, we find no statistically significant improvements on HateCheck. This can potentially be explained by the size and nature of this dataset that consists of handcrafted examples to test hate speech detection systems.

5.2 Informed Undersampling

When working with imbalanced datasets, undersampling the majority class can be a viable strategy for improving model performance. In this study, we investigate the effectiveness of informed undersampling of the majority class (i.e., the non-sexist class) in enhancing classification outcomes. Specifically, we selectively remove instances that are either particularly easy or particularly hard to classify from the non-sexist class. As shown in Tables 2 and 4, undersampling up to 15% of hard-to-classify non-sexist instances, identified using EL2N scores, improves the generalizability of the classifier. This improvement is statistically significant for out-of-distribution performance on the EXIST and EDOS datasets. While we also observe mean performance gains on the in-distribution CMSB test split and the HateCheck(Sexism) benchmark, these improvements are not statistically significant. Undersam-

Subtree	Examples	Mut. Inf.	Maj. Label
29	a girl, her husband, a lady, the home, a female, a wife, a women, female rappers	0.0045	Sexist
88	female, women, girl, girls, career, proper, physical, mens	0.0020	Sexist
15	than men, than women, to men, for sex, by men, to women, with women, and children	0.0020	Sexist
2	take care, be permitted, being leered, more easily, hear girls, stay home, be cooked	0.0014	Sexist
88	sexual, male, football, most, instant, greater, special	0.0014	Sexist
8	women should, men are, girls should, men should, women have, boys should, women do	0.0011	Sexist

Table 5: **Dataset Shortcuts in CMSB (Training Split):** The table contains the six subtree roots with the most discriminative patterns based on Mutual Information. For example, we observe that subtree identified by non-terminal node 29 is responsible for nouns and adjectives that are correlated with the sexist class. On the other hand subtree with non-terminal node 8 consists of assertive phrases that are correlated with the sexist class.

pling beyond 15% leads to performance degradation on both in-distribution and out-of-distribution datasets.

These findings support our observation that hard-to-classify non-sexist instances should be removed to increase performance. Our results further highlight that EL2N is a more effective influence score than PVI for guiding undersampling from a majority class. As discussed in Section 3.2, EL2N is a margin-based score. Removing instances with high EL2N scores effectively removes instances that lie close to the model's decision boundary, thereby increasing class separation (Sorscher et al., 2023).

5.3 Qualitative Analysis and Dataset Shortcuts

We identify grammar-based syntactic patterns strongly associated with the sexist class, particularly constructions involving gendered nouns (e.g., women) paired with model or assertive verbs (e.g., women should). The interested reader can refer to Table 5 where specific examples are given. Analysis of PVI and EL2N score distributions (see Figure 1) reveals that most shortcut-laden instances are "easy" with PVI scores above 1.68 and EL2N scores below 0.77, and are correctly classified. This suggests that models leverage such patterns during fine-tuning to generalize on the minority class (see examples in Table 9).

Interestingly, we also find shortcut-laden instances among the hard-to-classify non-sexist instances. The qualitative analysis of the non-sexist instances reveals that the hardest-to-classify instances contain lexical or syntactic features (e.g., gendered terms) resembling those found in shortcut-containing sexist instances. These instances exhibit negative PVI and high EL2N scores (with a mean and variance of $-1.39_{\pm 0.025}$ for PVI and $0.93_{\pm 0.03}$ for EL2N) and are typically pruned when hard-to-classify non-sexist instances are removed (see examples in Table 10). Rather than attributing these examples to label noise, we conducted a deeper analysis and found that they are minimal-edit counterfactuals—created by retaining gendered terms while modifying the text to flip its label. This method of generating counterfactuals has been criticized in the past by Howard et al. (2022); Joshi and He (2022), as such examples provide models with insufficient inductive biases during fine-tuning.

This suggests that combining pruning methods based on Influence Scores with shortcut induction methods is a promising direction to improve the quality of training data for machine learning models, while at the same time ensuring data efficiency.

5.4 Submodular Maximization

We don't observe statistically significant gains from submodular maximization-based core-set selection across pruning rates. As mentioned in Section 3.3, submodular maximization operates at the representation level, selecting instances by prioritizing semantic diversity while discarding those that are semantically similar (Krause and Golovin, 2014; Renduchintala et al., 2023) (see Appendix A.3). In our dataset, the majority of sexist instances contain common patterns that yield similar representations after fine-tuning a classifier. As noted in the previous section and shown visually in Figure 1, the pre-trained models leverage shortcuts during fine-tuning that affect the learnability of a data point. So selecting a diverse core-set leads to re-

moval of such instances at high pruning rates, this explains why submodular maximization underperforms compared to our informed pruning strategies based on learnability metrics or Influence Scores.

6 Conclusion

Prior research has highlighted concerns regarding the quality of datasets used to study harmful online communication, particularly due to the prevalence of spurious correlations that hinder model generalization. In this work, we present a case study that examines the impact of various data selection strategies on the composition of training data and assess how these modifications influence model performance. Our results indicate that data pruning based on EL2N scores consistently yields the most substantial improvements. As a margin-based metric, EL2N identifies instances near the model's decision boundary; removing instances with high EL2N values enhances class separation and contributes to more robust learning (Sorscher et al., 2023). Furthermore, we observe that proportionally pruning difficult instances from both majority and minority classes using PVI and EL2N scores leads to improved performance on both in-distribution and out-of-distribution test sets.[1]

We also explored whether data selection strategies help mitigate spurious correlations arising from dataset shortcuts. Notably, we observe that models rely on these shortcuts during fine-tuning. Specially easy-to-classify sexist instances and hard-to-classify non-sexist instances contain shortcuts. Pruning these instances to some extent leads to performance increases. Pruning hard-to-classify instances is in general a promising strategy also when shortcuts are not present.

7 Limitations

This study presents a empirical case study that is limited to one specific dataset that has been designed to cover different dimensions of sexism and avoid spurious correlations by introducing counterfactuals. It further focuses on sexism towards men and women. Future research should investigate whether our findings extend to other datasets and other forms of gender-based discrimination and harmful language detection more broadly.

In addition, while the experiments show that pruning based on Influence Scores (PVI and EL2N) can improve out-of-distribution performance, this approach involves important trade-offs. Pruning reduces training data diversity and may remove subtle or rare instances of sexism that are essential for developing nuanced classifiers. Our primary focus was on performance improvement across datasets collected for a singular task like sexism by extracting informative subsets. Future work should look into how such scores can be used in selecting data points for contrastive learning which is a method widely used in the realm of representation engineering when adapting models to imbalanced datasets (Gunel et al., 2021; Kim et al., 2022; Park et al., 2024; Madani et al., 2025).

While this study has exclusively focused on sexism detection and has applied learnability measures on a counterfactually augmented dataset with improved annotations, future work should also investigate how learnability metrics aid in identifying text where difficult negatives reflect natural ambiguity, sarcasm, or context rather than synthetic edits.

Moreover, while this work focuses on the technical performance and generalization of classifiers, future work should explore ethical and social implications of data reduction, for example by using fairness metrics.

8 Ethics Statement

This work addresses the challenge of detecting sexism in social media text by proposing a framework that analyzes learning dynamics to better understand how models learn from highly subjective, user-generated datasets. Using these insights, we develop a robust sexism classifier that generalizes well across different datasets. Our framework can also aid practitioners in evaluating and improving their own datasets.

Examples in Appendix A.5 illustrate how our quality metrics help identify problematic instances that may need moderator or annotator review. We believe this case study serves as a proof-of-concept for the development of robust, data-efficient systems for harmful language detection, while at the same time enabling identification of data points that require further intervention.

9 Acknowledgment

We acknowledge the use of commercial LLMs (GPT-4 and Grammarly) to assist in improving the clarity and grammar of the manuscript. All

[1] Although improvements on HateCheck are not statistically significant—likely due to the dataset's limited size—the observed trend remains consistent.

core ideas, analyses, and conclusions remain the original contributions of the authors. The authors would take this opportunity to thank the funding from the Horizon Europe research and innovation program under the Marie Skłodowska-Curie Grant Agreement No. 101073351 and the department of Computational Social Science at GESIS for providing the infrastructure to run the experiments. Jose M. Alonso-Moral would also like to acknowledge the support of the Ramon Areces Foundation with Project CONFIA, the Spanish Ministry of Science and Innovation (MCIN/AEI/10.13039/501100011033/) with grants PID2021-123152OB-C21 and PID2024-157680NB-I00, the Galician Ministry of Culture, Education, Professional Training and University (grants ED431G2023/04 and ED431C2022/19), and the European Union (European Regional Development Fund - ERDF). We also thank the anonymous reviewers in ACL Rolling Review platform for their feedback that helped in refining this paper.

References

Chirag Agarwal, Daniel D'souza, and Sara Hooker. 2022. Estimating example difficulty using variance of gradients. *Preprint*, arXiv:2008.11600.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. 2024. A survey on data selection for language models. *Preprint*, arXiv:2402.16827.

Nikhil Anand, Joshua Tan, and Maria Minakova. 2023. Influence scores at scale for efficient language data sampling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2485–2510, Singapore. Association for Computational Linguistics.

Abdul Azeemi, Ihsan Qazi, and Agha Raza. 2023. Data pruning for efficient model pruning in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 236–246, Singapore. Association for Computational Linguistics.

Rabiraj Bandyopadhyay, Dennis Assenmacher, Jose M. Alonso-Moral, and Claudia Wagner. 2024. Sexism detection on a data diet. In *Companion Publication of the 16th ACM Web Science Conference*, Websci Companion '24, page 94–102, New York, NY, USA. Association for Computing Machinery.

Jasmijn Bastings, Sebastian Ebert, Polina Zablotskaia, Anders Sandholm, and Katja Filippova. 2022. "will you find these shortcuts?" a protocol for evaluating the faithfulness of input salience methods for text classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 976–991, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Samyadeep Basu, Philip Pope, and Soheil Feizi. 2021a. Influence functions in deep learning are fragile. *Preprint*, arXiv:2006.14651.

Samyadeep Basu, Phillip Pope, and Soheil Feizi. 2021b. Influence functions in deep learning are fragile. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *Preprint*, arXiv:1508.05326.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Everlyn Chimoto, Jay Gala, Orevaoghene Ahia, Julia Kreutzer, Bruce Bassett, and Sara Hooker. 2024. Critical learning periods: Leveraging early training dynamics for efficient data pruning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9407–9426, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Sang Keun Choe, Hwijeen Ahn, Juhan Bae, Kewen Zhao, Minsoo Kang, Youngseog Chung, Adithya Pratapa, Willie Neiswanger, Emma Strubell, Teruko Mitamura, Jeff Schneider, Eduard Hovy, Roger Grosse, and Eric Xing. 2024. What is your data worth to GPT? llm-scale data valuation with influence functions. *Preprint*, arXiv:2405.13954.

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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Amaya Dharmasiri, William Yang, Polina Kirichenko, Lydia Liu, and Olga Russakovsky. 2025. The impact of coreset selection on spurious correlations and group robustness. *Preprint*, arXiv:2507.11690.

Jacob R. Epifano, Ravi P. Ramachandran, Aaron J. Masino, and Ghulam Rasool. 2023. Revisiting the fragility of influence functions. *Neural Networks*, 162:581–588.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. *Preprint*, arXiv:2110.08420.

Mohsen Fayyaz, Ehsan Aghazadeh, Ali Modarressi, Mohammad Taher Pilehvar, Yadollah Yaghoobzadeh, and Samira Ebrahimi Kahou. 2022. Bert on a data diet: Finding important examples by gradient-based pruning. *Preprint*, arXiv:2211.05610.

Antigoni Founta and Lucia Specia. 2021. A survey of online hate speech through the causal lens. In *Proceedings of the First Workshop on Causal Inference and NLP*, pages 74–82, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dan Friedman, Alexander Wettig, and Danqi Chen. 2022. Finding dataset shortcuts with grammar induction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4345–4363, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamilė Lukošiūtė, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R. Bowman. 2023. Studying large language model generalization with influence functions. *Preprint*, arXiv:2308.03296.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. *Preprint*, arXiv:1803.02324.

Frank Hampel, Elvezio Ronchetti, Peter Rousseeuw, and Werner Stahel. 2005. *Robust Statistics: The Approach Based on Influence Functions*.

Xiaochuang Han, Byron C. Wallace, and Yulia Tsvetkov. 2020a. Explaining black box predictions and unveiling data artifacts through influence functions. *Preprint*, arXiv:2005.06676.

Xiaochuang Han, Byron C. Wallace, and Yulia Tsvetkov. 2020b. Explaining black box predictions and unveiling data artifacts through influence functions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5553–5563, Online. Association for Computational Linguistics.

Phillip Howard, Gadi Singer, Vasudev Lal, Yejin Choi, and Swabha Swayamdipta. 2022. NeuroCounterfactuals: Beyond minimal-edit counterfactuals for richer data augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5056–5072, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Arthur Jacot, Franck Gabriel, and Clément Hongler. 2020. Neural tangent kernel: Convergence and generalization in neural networks. *Preprint*, arXiv:1806.07572.

Nitish Joshi and He He. 2022. An investigation of the (in)effectiveness of counterfactually augmented data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3668–3681, Dublin, Ireland. Association for Computational Linguistics.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

Youngwook Kim, Shinwoo Park, and Yo-Sub Han. 2022. Generalizable implicit hate speech detection using contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6667–6679, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 task 10: Explainable detection of online sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval)*, pages 2193–2210, Toronto, Canada. Association for Computational Linguistics.

Pang Wei Koh and Percy Liang. 2020. Understanding black-box predictions via influence functions. *Preprint*, arXiv:1703.04730.

Suraj Kothawade, Nathan Beck, Krishnateja Killamsetty, and Rishabh Iyer. 2021. Similar: Submodular information measures based active learning in realistic scenarios. *Preprint*, arXiv:2107.00717.

Suraj Kothawade, Vishal Kaushal, Ganesh Ramakrishnan, Jeff Bilmes, and Rishabh Iyer. 2022a. Prism: A rich class of parameterized submodular information measures for guided data subset selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(9):10238–10246.

Suraj Kothawade, Vishal Kaushal, Ganesh Ramakrishnan, Jeff Bilmes, and Rishabh Iyer. 2022b. Prism:

A rich class of parameterized submodular information measures for guided data subset selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10238–10246.

Andreas Krause and Daniel Golovin. 2014. Submodular function maximization. In *Tractability*.

Devin Kwok, Nikhil Anand, Jonathan Frankle, Gintare Karolina Dziugaite, and David Rolnick. 2024. Dataset difficulty and the role of inductive bias. *Preprint*, arXiv:2401.01867.

Yongchan Kwon, Eric Wu, Kevin Wu, and James Zou. 2024. Datainf: Efficiently estimating data influence in lora-tuned llms and diffusion models. *Preprint*, arXiv:2310.00902.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Navid Madani, Rabiraj Bandyopadhyay, Michael Miller Yoder, Stephan D. McCabe, Briony Swire-Thompson, and Kenneth Joseph. 2025. Measuring dimensions of self-presentation in twitter bios and their links to misinformation sharing. *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):1158–1175.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction. *Preprint*, arXiv:1802.03426.

Baharan Mirzasoleiman, Ashwinkumar Badanidiyuru, Amin Karbasi, Jan Vondrak, and Andreas Krause. 2014. Lazier than lazy greedy. *Preprint*, arXiv:1409.7938.

Loay Mualem, Ethan R. Elenberg, Moran Feldman, and Amin Karbasi. 2023. Submodular minimax optimization: Finding effective sets. *Preprint*, arXiv:2305.16903.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Ronghao Pan, José Antonio García-Díaz, and Rafael Valencia-García. 2024. Comparing fine-tuning, zero and few-shot strategies with large language models in hate speech detection in english. *CMES - Computer Modeling in Engineering and Sciences*, 140(3):2849–2868.

Kyungmin Park, Sihyun Oh, Daehyun Kim, and Juae Kim. 2024. Contrastive learning as a polarizer: Mitigating gender bias by fair and biased sentences. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4725–4736, Mexico City, Mexico. Association for Computational Linguistics.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. 2021. Deep learning on a data diet: Finding important examples early in training. *Preprint*, arXiv:2107.07075.

Pouya Pezeshkpour, Sarthak Jain, Sameer Singh, and Byron Wallace. 2022. Combining feature and instance attribution to detect artifacts. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1934–1946, Dublin, Ireland. Association for Computational Linguistics.

Garima Pruthi, Frederick Liu, Mukund Sundararajan, and Satyen Kale. 2020. Estimating training data influence by tracing gradient descent. *Preprint*, arXiv:2002.08484.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Alan Ramponi and Sara Tonelli. 2022. Features or spurious artifacts? data-centric baselines for fair and robust hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3027–3040, Seattle, United States. Association for Computational Linguistics.

H S V N S Kowndinya Renduchintala, Krishnateja Killamsetty, Sumit Bhatia, Milan Aggarwal, Ganesh Ramakrishnan, Rishabh Iyer, and Balaji Krishnamurthy. 2023. INGENIOUS: Using informative data subsets for efficient pre-training of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6690–6705, Singapore. Association for Computational Linguistics.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Adrián Mendieta-Aragón, Guillermo Marco-Remón, Maryna Makeienko, María Plaza, Julio Gonzalo, Damiano Spina, and Paolo Rosso. 2022. Overview of exist 2022: sexism identification in social networks. *Procesamiento de Lenguaje Natural*, 69:229–240.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1:*

Long Papers), pages 41–58, Online. Association for Computational Linguistics.

Mattia Samory, Indira Sen, Julian Kohne, Fabian Floeck, and Claudia Wagner. 2021. "call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples. *Preprint*, arXiv:2004.12764.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Happy Khairunnisa Sariyanto, Diclehan Ulucan, Oguzhan Ulucan, and Marc Ebner. 2025. Towards explainable hate speech detection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12883–12893, Vienna, Austria. Association for Computational Linguistics.

Indira Sen, Mattia Samory, Claudia Wagner, and Isabelle Augenstein. 2022. Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4716–4726, Seattle, United States. Association for Computational Linguistics.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. *Preprint*, arXiv:1708.00489.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. 2023. Beyond neural scaling laws: beating power law scaling via data pruning. *Preprint*, arXiv:2206.14486.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. *Preprint*, arXiv:1703.01365.

Michael Tänzer, Sebastian Ruder, and Marek Rei. 2022. Memorisation versus generalisation in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7564–7578, Dublin, Ireland. Association for Computational Linguistics.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. 2019. An empirical study of example forgetting during deep neural network learning. *Preprint*, arXiv:1812.05159.

Xiao Wang, Weikang Zhou, Qi Zhang, Jie Zhou, SongYang Gao, Junzhe Wang, Menghan Zhang, Xiang Gao, Yun Wen Chen, and Tao Gui. 2023. Farewell to aimless large-scale pretraining: Influential subset selection for language model. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 555–568, Toronto, Canada. Association for Computational Linguistics.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020. Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235, Online. Association for Computational Linguistics.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. LESS: Selecting influential data for targeted instruction tuning. In *Forty-first International Conference on Machine Learning*.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints. *CoRR*, abs/2002.10689.

Zehui Yu, Indira Sen, Dennis Assenmacher, Mattia Samory, Leon Fröhling, Christina Dahn, Debora Nozza, and Claudia Wagner. 2024. The unseen targets of hate: A systematic review of hateful communication datasets. *Social Science Computer Review*, 0(0):08944393241258771.

Heidi Chenyu Zhang, Shabnam Behzad, Kawin Ethayarajh, and Dan Jurafsky. 2024a. Data checklist: On unit-testing datasets with usable information. In *First Conference on Language Modeling*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024b. Instruction tuning for large language models: A survey. *Preprint*, arXiv:2308.10792.

Xiaoyu Zhang, Juan Zhai, Shiqing Ma, Chao Shen, Tianlin Li, Weipeng Jiang, and Yang Liu. 2025. STAFF: Speculative coreset selection for task-specific finetuning. In *The Thirteenth International Conference on Learning Representations*.

Zhehao Zhang, Jiaao Chen, and Diyi Yang. 2023. Mitigating biases in hate speech detection from a causal

perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6610–6625, Singapore. Association for Computational Linguistics.

Haizhong Zheng, Rui Liu, Fan Lai, and Atul Prakash. 2023. Coverage-centric coreset selection for high pruning rates. In *The Eleventh International Conference on Learning Representations*.

Xiaosen Zheng and Jing Jiang. 2022. An empirical study of memorization in nlp. *Preprint*, arXiv:2203.12171.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Appendix

A.1 Model Settings

We use the BERT model variant bert-base-cased[2] and the AdamW optimizer (Loshchilov and Hutter, 2019) with the hyperparameter settings listed in Appendix A.2 in Table 6. The number of epochs is set to 5, based on two considerations. First, prior research on imbalanced datasets has shown learning plateaus after first 5 epochs while fine-tuning BERT family of models (Tänzer et al., 2022). Second, prior research suggests that it is usual for EL2N-based pruning methods isolate easy instances within 5 epochs of training (Sorscher et al., 2023; Anand et al., 2023).

A.2 Hyperparameter Settings

This is the hyperparameter settings we use throughout our experiments. All our experiments were run on one 20GB partition of an NVIDIA A100 GPU.

Hyperparameter	Value
Learning Rate	1e-6
Epochs	5
Scheduler	Linear
Batch Size	32

Table 6: Hyperparameter settings.

A.3 Submodular Maximization

Submodular Maximization refers to a family of methods designed to select an informative subset—often referred to as a core-set (Sener and Savarese, 2018)—from a larger superset. In our use case, we aim to select a diverse core-set by designing a selection algorithm that operates at the embedding level, identifying a group of data points that maintains both the diversity and representativeness of the original dataset (Krause and Golovin, 2014; Kothawade et al., 2021, 2022a). In this appendix, we provide a brief introduction to Submodular Optimization, with a focus on Submodular Maximization-based subset selection (Krause and Golovin, 2014). We first define the concept of submodularity, and then introduce the notion of submodular gain, which underpins our submodular data selection baseline. After that we provide the intuition behind Facility Location problem that has been used to model the data selection problem.

A.3.1 Submodularity

Submodularity is a property of functions that work on **finite** set. A set function can be defined as $f : 2^V \rightarrow \mathbf{R}$ that assign each subset $S \subseteq V$ a value of $f(S)$. The V in question is the superset also called a ground set consisting of finite number of elements. Also $f(\phi) = 0$ which implies that the function on an empty set carries no value.

A set function f is submodular if for every $A, B \subseteq V$ and $e \in V \backslash B$ the following 2 equations holds:

$$\Delta(e|A) \geq \Delta(e|B) \tag{3}$$

and for every $A, B \subseteq V$,

$$f(A \cap B) + f(A \cup B) \leq f(A) + f(B) \tag{4}$$

The equations A.3.1 and A.3.1 helps us in understanding the concept of Submodular Maximization. Equation A.3.1 essentially implies that if we have already obtained a subset A consisting of a set of values, including another value from $B \backslash A$ does not lead to any benefit. This implies that the submodular set obtained exhibit diminishing returns property. We now give an intuition of the reason behind modeling the subset selection problem (by designing an intervention on the embedding level).

Submodular Gain For a set function $f : 2^V \rightarrow \mathbb{R}$, $S \subseteq V$ and $e \in V$, we define the submodular gain of f at S with respect to e $\Delta_f(e|S) = f(S \cup e) - f(S)$.

A.3.2 Facility Location

Facility Location (Krause and Golovin, 2014; Renduchintala et al., 2023) is a submodular function

[2] google-bert/bert-base-cased

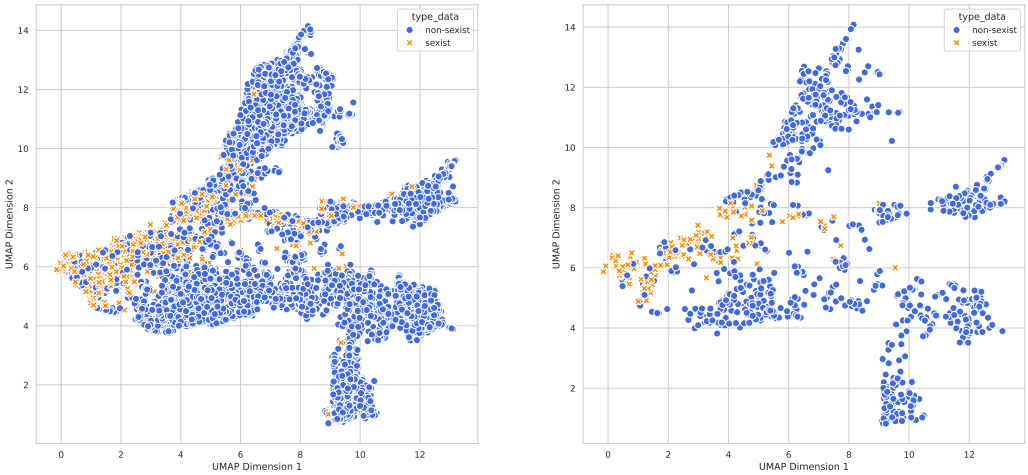

Figure 2: Distribution of data points from both the sexist and non-sexist classes before and after applying submodular maximization for subset selection. The left plot shows the original data distribution; the right plot shows the distribution after selecting 1,000 diverse and representative samples using model-generated representations and a lazier-than-lazy greedy optimization. This illustrates how submodular maximization leads to more diverse and representative subset by intervening at the representation level.

(we can also call it a mathematical model) which is very similar to k-medoids clustering and is defined as

$$f_{FL}(A) = \sum_{i \in V} max_{j \in A} K_{ij} \tag{5}$$

And the subset selection problem then becomes $S_t = argmax_{S \subseteq V : |S| = k} f_{FL}(S)$. From the equations it is clear that the problem of selecting a subset which is diverse and representative of the superset is a combinatorially explosive problem which is NP-hard. But how does the selection of diverse subset happen? The answer lies in Equation A.3.2. A closer inspection of the submodular objective function reveals that every data point—regardless of whether it is an outlier or part of a cluster—has an equal opportunity to be included in the selected subset. To facilitate this property, a family of greedy-based optimization algorithms has been developed. Among the most recent and efficient is the Lazier-than-Lazy Greedy algorithm proposed by (Mirzasoleiman et al., 2014).

These optimizers are designed to select a subset that is not only near-optimal but also maintains the diversity and representativeness of the original dataset. In Figure 2, we present a graphical visualization of 1,000 data points sampled from our training set using submodular maximization. The plots demonstrate that the algorithm effectively maintains diversity across both sexist and non-sexist classes.

A.4 Generalization to RoBERTa

RoBERTa(Liu et al., 2019) is a robust encoder based pre-trained model belonging to BERT family with more parameters, different tokenizer (Byte Pair Encoding)(Sennrich et al., 2016) with different pre-training tokens and objectives. We ran the same experiments on RoBERTa and show the results in Tables 8 and 7. We observe that in case of EXIST and EDOS out-of-distribution datasets, RoBERTa's performance mimics that of BERT across pruning rates of 15% in case of proportionally pruning the **easy** to classify sexist data points and **hard** to classify non-sexist data points. In case of pruning **hard** sexist instances and **hard** non-sexist instances, as evident from Table 7 RoBERTa model also performs well on pruning 15% and 30% data from the in-distribution dataset in case of EXIST and on pruning 15% of the same training dataset in case of EDOS. In case of Hatecheck, as is the case for BERT, we don't observe statistically significant gain or loss in performance. Moreover RoBERTa also performs well on in-distribution test data compared to BERT and all the other pruning strategies does not lead to statistically significant **gain** or **drop** in performance across different pruning strategies. This can be attributed to RoBERTa's robust tokenization and its ability to memorize datasets (Zheng and Jiang, 2022; Carlini et al., 2023). Another reason why there is a difference between BERT and RoBERTa is because of their respective inductive biases. Different pre-training corpus and objectives enables a

Method	EXIST			EDOS			Hatecheck		
	15%	30%	60%	15%	30%	60%	15%	30%	60%
Unpruned (0%)		$48.5_{\pm 0.4}$			$60.5_{\pm 0.2}$			$55.1_{\pm 2.3}$	
Random									
random	$47.5_{\pm 2.9}$	$42.6_{\pm 4.1}$	$34.3_{\pm 0.0}$	$60.0_{\pm 1.9}$	$55.8_{\pm 6.1}$	$43.0_{\pm 0.0}$	$56.0_{\pm 1.7}$	$44.9_{\pm 3.7}$	$21.1_{\pm 0.0}$
Submod. Max.									
max	$45.5_{\pm 1.3}$	$35.7_{\pm 2.3}$	$34.3_{\pm 0.0}$	$53.9_{\pm 5.6}$	$46.5_{\pm 5.7}$	$43.1_{\pm 0.0}$	$52.7_{\pm 3.4}$	$27.8_{\pm 1.1}$	$21.1_{\pm 0.0}$
Informed									
IU-PVI-H	$64.4^{*}_{\pm 0.7}$	$65.6^{*}_{\pm 0.4}$	$57.8_{\pm 2.0}$	$59.3_{\pm 0.2}$	$54.7_{\pm 0.6}$	$39.0_{\pm 4.4}$	$54.8_{\pm 0.4}$	$48.3_{\pm 0.2}$	$45.6_{\pm 0.2}$
IU-PVI-E	$47.5_{\pm 0.7}$	$45.7_{\pm 2.8}$	$34.3_{\pm 0.0}$	$60.8_{\pm 0.2}$	$59.1_{\pm 6.7}$	$43.1_{\pm 0.1}$	$55.0_{\pm 0.5}$	$53.2_{\pm 6.1}$	$21.1_{\pm 0.0}$
IU-EL2N-H	$64.5^{*}_{\pm 0.5}$	$65.5^{*}_{\pm 0.2}$	$57.4_{\pm 0.9}$	$59.7_{\pm 0.2}$	$53.7_{\pm 0.8}$	$36.8_{\pm 2.0}$	$54.7_{\pm 0.3}$	$48.6_{\pm 0.7}$	$45.8_{\pm 0.9}$
IU-EL2N-E	$48.4_{\pm 0.7}$	$43.2_{\pm 2.2}$	$34.3_{\pm 0.0}$	$60.4_{\pm 0.6}$	$58.1_{\pm 2.0}$	$43.0_{\pm 0.0}$	$55.9_{\pm 1.6}$	$53.3_{\pm 4.0}$	$21.1_{\pm 0.0}$
Prop. Prune									
PP-PVI-EE	$47.9_{\pm 2.0}$	$40.4_{\pm 4.9}$	$34.3_{\pm 0.0}$	$57.9_{\pm 2.6}$	$52.5_{\pm 7.7}$	$43.1_{\pm 0.0}$	$56.2_{\pm 1.5}$	$40.0_{\pm 15.5}$	$21.1_{\pm 0.0}$
PP-PVI-EH	$59.4^{*}_{\pm 1.1}$	$64.1^{*}_{\pm 1.5}$	$65.7_{\pm 0.3}$	$61.7^{*}_{\pm 0.4}$	$59.6_{\pm 0.7}$	$54.4_{\pm 1.0}$	$56.2_{\pm 0.3}$	$54.9_{\pm 0.2}$	$49.4_{\pm 1.3}$
PP-PVI-HE	$47.3_{\pm 1.3}$	$45.6_{\pm 0.8}$	$40.2_{\pm 1.9}$	$60.5_{\pm 0.6}$	$59.6_{\pm 0.4}$	$54.4_{\pm 3.4}$	$55.1_{\pm 1.5}$	$54.9_{\pm 0.5}$	$47.7_{\pm 7.5}$
PP-PVI-HH	$54.0^{*}_{\pm 0.9}$	$61.7^{*}_{\pm 0.4}$	$64.8_{\pm 0.5}$	$61.5^{*}_{\pm 0.1}$	$60.6_{\pm 0.1}$	$57.9_{\pm 0.1}$	$56.0_{\pm 1.1}$	$54.8_{\pm 0.4}$	$50.7_{\pm 0.8}$
PP-EL2N-EE	$47.8_{\pm 1.7}$	$40.2_{\pm 4.2}$	$34.3_{\pm 0.0}$	$58.5_{\pm 3.0}$	$56.2_{\pm 2.8}$	$43.1_{\pm 0.1}$	$55.2_{\pm 4.6}$	$42.2_{\pm 8.9}$	$21.1_{\pm 0.0}$
PP-EL2N-EH	$56.5^{*}_{\pm 0.6}$	$64.5^{*}_{\pm 0.9}$	$64.8_{\pm 1.9}$	$62.0^{*}_{\pm 0.2}$	$59.6_{\pm 0.1}$	$55.1_{\pm 2.3}$	$55.9_{\pm 1.0}$	$54.1_{\pm 4.5}$	$50.2_{\pm 2.4}$
PP-EL2N-HE	$48.3_{\pm 1.1}$	$45.4_{\pm 1.9}$	$39.8_{\pm 2.8}$	$60.0_{\pm 0.5}$	$58.6_{\pm 0.5}$	$53.6_{\pm 5.1}$	$56.4_{\pm 1.4}$	$53.5_{\pm 4.0}$	$45.6_{\pm 12.5}$
PP-EL2N-HH	$54.8^{*}_{\pm 0.7}$	$61.2^{*}_{\pm 0.9}$	$65.3_{\pm 0.6}$	$61.3^{*}_{\pm 0.2}$	$60.5_{\pm 0.5}$	$57.4_{\pm 0.4}$	$55.9_{\pm 1.8}$	$55.8_{\pm 0.5}$	$50.8_{\pm 1.6}$

Table 7: **RoBERTa out-of-distribution performance:** We showcase the Macro-F1 Scores of models trained on original (i.e. unpruned) and pruned CMSB train data and tested on EXIST, EDOS and Hatecheck datasets. The pruning rates (second row) indicate the amount of training data that was removed from the CMSB data before the model was trained. The rows correspond to different selection methods. We have shortened informed_undersampling (inf. samp.) and proportional_pruning (prop. prune) for convenience. We perform Wilcoxon test on the Macro F1-Scores and mark and color scores (Yellow and Lavender for PVI and EL2N and stars) which were statistically significant higher than the F1 score of the unpruned data (see row 1) with p-values below the significance level ($\alpha = 0.05$). Our results show that undersampling the hard instances does not lead to statistically significant performance gains in case of EDOS and HateCheck(Sexism). Moreover, we also observe that proportionally pruning the easy sexist and hard non-sexist examples (based on EL2N and PVI) leads to better generalization across EDOS and EXIST datasets.

model to acquire different set of inductive biases (Warstadt et al., 2020; Kwok et al., 2024), thus affecting how it learns from a dataset during fine-tuning. However, proportional pruning strategies on in-distribution training data (**PP-inf_score-EH** and **PP-inf_score-HH**, where **inf_score** stands for either **PVI** or **EL2N**) does aid RoBERTa in generalizing across EXIST and EDOS datasets even though the percentage where they show generalization is different. Our result also showcases that we can fine-tune bigger models with much less data but of higher quality and informativeness to **retain** or **improve** performance in both in-distribution and out-of-distribution settings which is the central aspect of coreset selection (Sener and Savarese, 2018; Zheng et al., 2023; Dharmasiri et al., 2025).

A.5 Pruned Instances

In this section, we present examples of instances pruned by our strategies. We observe that both PVI and EL2N scores effectively identify and filter out instances containing shortcuts.Especially, hard-to-classify non-sexist (majority) and easy-to-classify sexist (minority) class contain shortcuts. Many of the more challenging sexist examples are implicit, contain spelling errors, or include hashtags (e.g., #mkr, #FemFreeFriday) that require additional contextual understanding for the model to interpret correctly. As discussed in Section 5, instances with shortcuts similar to those found in easier-to-classify sexist examples were often minimal-edit counterfactuals. Representative examples are shown in Table 10.

A.6 Results for Higher Pruning Rates

We also showcase more results for pruning rates from 15% to 60% in Tables 13, 14, 15 and 16. From the Tables we do observe that performance across all the pruning strategies decreases as we prune more data from our in-distribution training dataset.

Method	15%	30%	60%
Unpruned (0%)		$82.0_{\pm 0.2}$	
Random			
random	$80.9_{\pm 0.7}$	$76.9_{\pm 5.4}$	$46.4_{\pm 0.0}$
Submod. Max			
max	$75.7_{\pm 0.4}$	$49.0_{\pm 3.5}$	$46.4_{\pm 0.0}$
Informed			
IU-PVI-H	$78.7_{\pm 0.15}$	$72.4_{\pm 0.27}$	$62.6_{\pm 1.7}$
IU-PVI-E	$81.5_{\pm 0.4}$	$80.4_{\pm 1.3}$	$46.9_{\pm 0.8}$
IU-EL2N-H	$78.5_{\pm 0.3}$	$71.8_{\pm 0.2}$	$61.9_{\pm 0.3}$
IU-EL2N-E	$81.7_{\pm 0.4}$	$79.9_{\pm 1.3}$	$46.4_{\pm 0.0}$
Prop. Prune			
PP-PVI-EE	$81.8_{\pm 0.3}$	$66.8_{\pm 16.6}$	$47.2_{\pm 1.3}$
PP-PVI-EH	$81.2_{\pm 0.2}$	$78.6_{\pm 0.3}$	$73.2_{\pm 1.7}$
PP-PVI-HE	$81.9_{\pm 0.2}$	$81.7_{\pm 0.1}$	$77.6_{\pm 3.5}$
PP-PVI-HH	$80.8_{\pm 0.4}$	$79.4_{\pm 0.1}$	$75.0_{\pm 0.2}$
PP-EL2N-EE	$81.7_{\pm 0.3}$	$73.4_{\pm 7.9}$	$46.7_{\pm 0.5}$
PP-EL2N-EH	$81.1_{\pm 5.6}$	$78.8_{\pm 0.2}$	$72.3_{\pm 0.3}$
PP-EL2N-HE	$81.9_{\pm 0.2}$	$81.7_{\pm 0.5}$	$74.4_{\pm 9.9}$
PP-EL2N-HH	$81.1_{\pm 0.2}$	$79.5_{\pm 0.2}$	$74.6_{\pm 0.5}$

Table 8: **RoBERTa in-distribution performance:** We showcase the Macro-F1 Scores of models trained on original (i.e., unpruned) and pruned CMSB train data and test on CMSB hold-out data. As with BERT we also perform Wilcoxon test on the F1-Scores and find no statistical improvement across any pruning strategies based on our chosen significance level. However pruning both the easy sexist and easy non-sexist data points from the training data does lead to a drop in performance at higher pruning rate of 60%. Unlike BERT we see all the other pruning strategies RoBERTa performs similar relative to the performance on fine-tuning on the whole dataset. Although performance does drop at high pruning rate of 60% across all the pruning strategies.

Text	Gold	PVI	EL2N	Shortcut
Men use physical contact as a request for sex	sexist	2.637	0.388	Yes
Women have an obligation to uphold modesty in dress	sexist	2.634	0.389	Yes
Women are inferior to men in their cognitive ability	sexist	2.634	0.390	Yes
Women are more vulnerable than men are	sexist	2.631	0.392	Yes
Men are not sexually satisfied with any behavior other than intercourse	sexist	2.630	0.392	Yes
Compared to women, men have many undesirable qualities	sexist	2.629	0.393	Yes
Women are more easily influenced than men are	sexist	2.625	0.396	Yes
Women seem to be inherently less capable than men of logical and scientific thinking	sexist	2.625	0.396	Yes
Men are superior to women in intelligence	sexist	2.623	0.398	Yes
Women are more devout individuals than men are	sexist	2.623	0.397	Yes

Table 9: Top 10 easy sexist instances based on PVI and EL2N scores.

Text	Gold	PVI	EL2N	Shortcuts
Men are as conceited about their appearance as women are	non-sexist	-1.642	1.011	Yes
On the average, women are as intelligent as men	non-sexist	-1.634	1.007	Yes
Women should be prepared to oppose men in order to obtain equal status	non-sexist	-1.630	1.007	Yes
Women have the right to decide for themselves what is proper in feminine dress	non-sexist	-1.630	1.007	Yes
Women are as interested in sex as men are	non-sexist	-1.629	1.002	Yes
Men are as submissive as women are	non-sexist	-1.615	1.001	Yes
Women have the right to compete with men in every sphere of activity	non-sexist	-1.611	1.003	Yes
Women should be allowed the same sexual freedom as men	non-sexist	-1.611	1.002	Yes
Women should be free to express themselves sexually	non-sexist	-1.598	0.998	Yes
Men are just as easily influenced by others as women are	non-sexist	-1.597	0.998	Yes

Table 10: Top 10 hard non-sexist instances based on PVI and EL2N scores.

Text	Gold	PVI	EL2N	Shortcut
KAT IS A FUCKING CUNT #mkr	sexist	-2.292	1.377	No
Kat you're a fkn ass #mkr	sexist	-2.272	1.376	No
Really bitch really? Get on the runway and back in the kitchen #MKR2015 #MKR	sexist	-2.245	1.375	No
They starting to act and look like a couple of B grade hookers to me #MKR	sexist	-2.127	1.374	No
#IDontNeedFeminism #FemFreeFriday [URL]	sexist	-2.045	1.371	No
Seeing as you two think this is a modelling competition and not a cooking competition, I give you two a score of -10 for your looks. #MKR	sexist	-2.042	1.367	No
RT MENTION2838 xbox's were designed for gamers and make up was designed for beauticians, get the idea #promise #lol	sexist	-1.992	1.368	No
MENTION1047 MENTION1497 pls don't the Trump womrn supporters paid or no r so much hotter :) [URL]	sexist	-1.978	1.364	No
I wonder which of the blonde ones screwed Colin to get that score? Probably both. #mkr	sexist	-1.937	1.366	No
MENTION4117 the stay at home mums of this country need to turn off the tele and go do some d.. [URL]	sexist	-1.822	1.361	No

Table 11: Top 10 hard sexist instances based on PVI and EL2N scores.

Text	Gold	PVI	EL2N	Shortcuts
Seems Eva & Debra think they are michelin star chefs after their instant restaurant result. Don't quit your day jobs just yet #mkr	non-sexist	0.144	0.033	No
Just when they thought it was safe to go back into the kitchen... Instant restaurant redemption round! #mkr	non-sexist	0.144	0.033	No
JFC Kat do you have NO morals???? Carnt #mkr2015 #mkr	non-sexist	0.144	0.033	No
MENTION2122 with spinich – who doesn't love spinich! #mkr	non-sexist	0.144	0.033	No
A 7! Colin dipping his end in! #mkr	non-sexist	0.144	0.033	No
With all this smashing it and nailing it, Bunnings needs to come on board as a sponsor with sausage-sizzle challenge #mkr	non-sexist	0.144	0.033	No
Just so the results, why would the other teams not up the score to drop kick Captain Salty and her first mate beta? #mkr	non-sexist	0.144	0.034	No
Kat and Andre are the crud on the bottom of a saucepan when you burn your food #mkr	non-sexist	0.144	0.034	No
A 1? You are real arseholes. #mkr	non-sexist	0.143	0.034	No
Sick of the word Sassy!! #mkr	non-sexist	0.143	0.034	No

Table 12: Top 10 easy non-sexist instances based on PVI and EL2N scores.

Method	15%	30%	40%	50%	60%
Unpruned (0%)			$73.4_{\pm 6.3}$		
Random					
random	$69.2_{\pm 6.7}$	$63.6_{\pm 9.5}$	$57.2_{\pm 10.3}$	$57.1_{\pm 6.0}$	$54.6_{\pm 6.4}$
Submod. Max					
max	$71.1_{\pm 4.4}$	$54.8_{\pm 9.3}$	$60.2_{\pm 9.4}$	$56.4_{\pm 7.8}$	$47.7_{\pm 2.4}$
Informed					
IU-PVI-H	$75.9_{\pm 2.6}$	$69.6_{\pm 4.1}$	$70.0_{\pm 3.5}$	$64.6_{\pm 0.9}$	$59.2_{\pm 5.2}$
IU-PVI-E	$70.6_{\pm 5.3}$	$66.3_{\pm 2.2}$	$65.2_{\pm 9.2}$	$64.8_{\pm 1.0}$	$57.0_{\pm 6.1}$
IU-EL2N-H	$78.2_{\pm 1.6}$	$71.2_{\pm 1.5}$	$69.4_{\pm 1.0}$	$63.3_{\pm 0.3}$	$64.2_{\pm 0.8}$
IU-EL2N-E	$73.6_{\pm 5.9}$	$61.0_{\pm 6.9}$	$69.6_{\pm 4.7}$	$69.0_{\pm 3.9}$	$59.5_{\pm 7.1}$
Prop. Prune					
PP-PVI-EE	$70.7_{\pm 5.6}$	$46.6_{\pm 0.3}$	$47.9_{\pm 2.9}$	$47.6_{\pm 2.3}$	$46.4_{\pm 0.0}$
PP-PVI-EH	$76.9_{\pm 5.1}$	$76.6_{\pm 3.2}$	$74.9_{\pm 3.2}$	$74.7_{\pm 1.6}$	$72.3_{\pm 2.5}$
PP-PVI-HE	$68.5_{\pm 7.7}$	$67.2_{\pm 1.9}$	$65.8_{\pm 9.9}$	$64.1_{\pm 8.2}$	$66.3_{\pm 6.2}$
PP-PVI-HH	$73.7_{\pm 4.5}$	$78.8^{*}_{\pm 0.6}$	$76.2_{2.0}$	$70.6_{\pm 3.1}$	$71.7_{\pm 1.0}$
PP-EL2N-EE	$65.1_{\pm 14.1}$	$52.9_{\pm 10.6}$	$51.8_{\pm 10.7}$	$47.3_{\pm 1.7}$	$49.7_{\pm 6.6}$
PP-EL2N-EH	$73.4_{\pm 5.6}$	$78.2_{\pm 1.4}$	$76.7_{\pm 1.2}$	$74.8_{\pm 2.9}$	$71.7_{\pm 1.7}$
PP-EL2N-HE	$69.0_{\pm 4.2}$	$66.2_{\pm 8.6}$	$66.5_{\pm 5.0}$	$63.7_{\pm 5.1}$	$69.6_{\pm 4.0}$
PP-EL2N-HH	$78.4^{*}_{\pm 3.9}$	$76.6_{\pm 4.5}$	$74.7_{\pm 2.7}$	$72.7_{\pm 3.2}$	$70.3_{\pm 3.1}$

Table 13: **BERT in-distribution performance:** We showcase the Macro-F1 Scores of models trained on original (i.e., unpruned) and pruned CMSB train data and test on CMSB hold-out data. We perform Wilcoxon test on the Macro F1-Scores and mark and color scores (Yellow and Lavender for PVI and EL2N and stars) which were statistically significant higher than the F1 score of the unpruned data (see row 1) with p-values below the significance level ($\alpha = 0.05$). We observe statistically significant performance gains on pruning hard sexist and hard non-sexist data points based on PVI and EL2N. For all other pruning types we don't see any statistically significant gains.

Method	15%	30%	40%	50%	60%
Unpruned (0%)			$82.0_{\pm 0.2}$		
Random					
random	$80.9_{\pm 0.7}$	$76.9_{\pm 5.4}$	$60.1_{\pm 8.4}$	$46.5_{\pm 0.0}$	$46.4_{\pm 0.0}$
Submod. Max					
max	$75.7_{\pm 0.4}$	$49.0_{\pm 3.5}$	$46.4_{\pm 0.0}$	$46.5_{\pm 0.0}$	$46.4_{\pm 0.0}$
Informed					
IU-PVI-H	$78.7_{\pm 0.15}$	$72.4_{\pm 0.27}$	$67.4_{\pm 0.4}$	$64.3_{\pm 0.5}$	$62.6_{\pm 1.7}$
IU-PVI-E	$81.5_{\pm 0.4}$	$80.4_{\pm 1.3}$	$70.8_{\pm 9.5}$	$60.4_{\pm 14.0}$	$46.9_{\pm 0.8}$
IU-EL2N-H	$78.5_{\pm 0.3}$	$71.8_{\pm 0.2}$	$68.2_{\pm 0.7}$	$64.2_{\pm 0.7}$	$61.9_{\pm 0.3}$
IU-EL2N-E	$81.7_{\pm 0.4}$	$79.9_{\pm 1.3}$	$67.2_{\pm 13.8}$	$57.2_{\pm 10.3}$	$46.4_{\pm 0.0}$
Prop. Prune					
PP-PVI-EE	$81.8_{\pm 0.3}$	$66.8_{\pm 16.6}$	$63.2_{\pm 15.1}$	$56.2_{\pm 12.0}$	$47.2_{\pm 1.3}$
PP-PVI-EH	$81.2_{\pm 0.2}$	$78.6_{\pm 0.3}$	$76.5_{\pm 0.2}$	$74.0_{\pm 0.4}$	$73.2_{\pm 1.7}$
PP-PVI-HE	$81.9_{\pm 0.2}$	$81.7_{\pm 0.1}$	$81.8_{\pm 0.2}$	$81.2_{\pm 0.7}$	$77.6_{\pm 3.5}$
PP-PVI-HH	$80.8_{\pm 0.4}$	$79.4_{\pm 0.1}$	$78.2_{\pm 0.2}$	$76.5_{\pm 0.2}$	$75.0_{\pm 0.2}$
PP-EL2N-EE	$81.7_{\pm 0.3}$	$73.4_{\pm 7.9}$	$69.2_{\pm 12.3}$	$55.2_{\pm 11.0}$	$46.7_{\pm 0.5}$
PP-EL2N-EH	$81.1_{\pm 5.6}$	$78.8_{\pm 0.2}$	$76.2_{\pm 0.2}$	$74.1_{\pm 0.2}$	$72.3_{\pm 0.3}$
PP-EL2N-HE	$81.9_{\pm 0.2}$	$81.7_{\pm 0.5}$	$81.3_{\pm 0.1}$	$81.1_{\pm 0.3}$	$74.4_{\pm 9.9}$
PP-EL2N-HH	$81.1_{\pm 0.2}$	$79.5_{\pm 0.2}$	$78.4_{\pm 0.2}$	$76.7_{\pm 0.4}$	$74.6_{\pm 0.5}$

Table 14: **RoBERTa in-distribution performance:** We showcase the Macro-F1 Scores of models trained on original (i.e., unpruned) and pruned CMSB train data and test on CMSB hold-out data. As with BERT we also perform Wilcoxon test on the F1-Scores and find no statistical improvement across any pruning strategies based on our chosen significance level. However pruning both the easy sexist and easy non-sexist data points from the training data does lead to a drop in performance as we gradually increase the pruning rate from 40% to 60%. Unlike BERT we see all the other pruning strategies RoBERTa performs similar relative to the performance on fine-tuning on the whole dataset. Although performance does drop at high pruning rate of 60% across all the pruning strategies.

Method	EXIST					EDOS					Hatecheck				
	15%	30%	40%	50%	60%	15%	30%	40%	50%	60%	15%	30%	40%	50%	60%
Unpruned (0%)			$40.5_{\pm 5.0}$					$53.4_{\pm 6.4}$					$46.5_{\pm 10.5}$		
Random															
random	$38.2_{\pm 4.5}$	$36.6_{\pm 2.9}$	$35.8_{\pm 2.3}$	$34.5_{\pm 0.2}$	$34.6_{\pm 0.3}$	$49.8_{\pm 5.4}$	$48.5_{\pm 5.5}$	$47.0_{\pm 5.3}$	$44.3_{\pm 1.1}$	$43.8_{\pm 0.9}$	$34.8_{\pm 8.6}$	$36.8_{\pm 12.9}$	$29.4_{\pm 9.4}$	$27.9_{\pm 5.5}$	$22.9_{\pm 1.8}$
Submod. Max.															
max	$37.3_{\pm 3.4}$	$35.4_{\pm 1.9}$	$35.8_{\pm 1.6}$	$34.6_{\pm 0.3}$	$34.3_{\pm 0.0}$	$50.0_{\pm 5.4}$	$45.2_{\pm 2.7}$	$47.5_{\pm 4.7}$	$44.0_{\pm 0.8}$	$43.1_{\pm 0.0}$	$40.2_{\pm 9.6}$	$27.1_{\pm 7.0}$	$31.7_{\pm 9.6}$	$24.1_{\pm 2.6}$	$21.1_{\pm 0.1}$
Informed															
IU-PVI-H	$54.6^{*}_{\pm 4.5}$	$55.8^{*}_{\pm 2.0}$	$56.5_{\pm 7.0}$	$52.3_{\pm 9.1}$	$48.4_{\pm 2.0}$	$58.6_{\pm 3.0}$	$52.4_{\pm 4.3}$	$47.7_{\pm 5.9}$	$38.2_{\pm 4.0}$	$35.3_{\pm 4.6}$	$46.8_{\pm 5.6}$	$41.6_{\pm 8.8}$	$46.0_{\pm 2.3}$	$42.4_{\pm 1.9}$	$38.2_{\pm 8.6}$
IU-PVI-E	$38.1_{\pm 2.9}$	$35.5_{\pm 1.3}$	$36.0_{\pm 1.3}$	$34.6_{\pm 0.3}$	$34.7_{\pm 0.2}$	$50.9_{\pm 5.3}$	$47.2_{\pm 4.0}$	$48.2_{\pm 3.6}$	$45.7_{\pm 1.1}$	$45.0_{\pm 1.4}$	$43.7_{\pm 10.5}$	$40.7_{\pm 9.5}$	$43.0_{\pm 13.7}$	$33.0_{\pm 6.6}$	$27.4_{\pm 5.6}$
IU-EL2N-H	$58.5^{*}_{\pm 4.5}$	$61.1^{*}_{\pm 2.0}$	$60.7_{\pm 2.1}$	$54.2_{\pm 0.8}$	$56.3_{\pm 2.0}$	$61.0^{*}_{\pm 0.3}$	$52.8_{\pm 1.8}$	$48.3_{\pm 1.5}$	$34.2_{\pm 1.5}$	$36.2_{\pm 2.3}$	$51.7_{\pm 2.1}$	$46.6_{\pm 1.6}$	$44.4_{\pm 2.6}$	$42.5_{\pm 0.7}$	$43.7_{\pm 1.2}$
IU-EL2N-E	$41.1_{\pm 3.8}$	$35.1_{\pm 1.3}$	$37.6_{\pm 2.7}$	$36.2_{\pm 0.7}$	$34.7_{\pm 0.2}$	$53.6_{\pm 5.3}$	$46.2_{\pm 4.0}$	$50.9_{\pm 4.4}$	$50.3_{\pm 2.4}$	$45.0_{\pm 1.4}$	$44.5_{\pm 10.2}$	$31.4_{\pm 10.0}$	$42.8_{\pm 6.5}$	$40.5_{\pm 6.3}$	$30.0_{\pm 8.3}$
Proportional															
PP-PVI-EE	$38.6_{\pm 3.8}$	$34.3_{\pm 0.0}$	$34.4_{\pm 0.0}$	$34.3_{\pm 0.0}$	$34.3_{\pm 0.0}$	$52.4_{\pm 5.4}$	$43.0_{\pm 0.0}$	$43.1_{\pm 0.0}$	$43.1_{\pm 0.0}$	$43.0_{\pm 0.0}$	$43.9_{\pm 4.5}$	$21.1_{\pm 0.1}$	$21.4_{\pm 0.8}$	$21.0_{\pm 0.0}$	$21.0_{\pm 0.0}$
PP-PVI-EH	$51.3^{*}_{\pm 7.6}$	$55.7^{*}_{\pm 7.4}$	$52.4_{\pm 7.6}$	$55.9_{\pm 3.2}$	$54.7_{\pm 5.4}$	$59.1^{*}_{\pm 7.6}$	$59.9^{*}_{\pm 2.3}$	$59.6_{\pm 1.7}$	$59.5_{\pm 0.8}$	$57.8_{\pm 1.8}$	$49.1_{\pm 4.5}$	$52.0_{\pm 3.4}$	$49.4_{\pm 2.6}$	$46.8_{\pm 2.5}$	$47.4_{\pm 2.1}$
PP-PVI-HE	$38.1_{\pm 2.9}$	$36.7_{\pm 1.0}$	$37.7_{\pm 3.3}$	$35.6_{\pm 2.1}$	$36.1_{\pm 1.0}$	$51.0_{\pm 4.5}$	$48.7_{\pm 1.9}$	$50.3_{\pm 6.0}$	$47.0_{\pm 3.7}$	$48.3_{\pm 4.0}$	$42.3_{\pm 9.7}$	$41.8_{\pm 9.3}$	$42.2_{\pm 12.8}$	$39.4_{\pm 13.7}$	$40.2_{\pm 11.0}$
PP-PVI-HH	$45.7_{\pm 7.8}$	$58.7^{*}_{\pm 1.7}$	$54.9_{\pm 5.4}$	$48.3_{\pm 6.0}$	$54.1_{\pm 2.5}$	$56.2_{\pm 4.4}$	$61.0^{*}_{\pm 0.4}$	$60.1_{\pm 0.5}$	$57.6_{\pm 2.0}$	$48.3_{\pm 4.2}$	$46.9_{\pm 3.3}$	$51.2_{\pm 1.8}$	$50.7_{\pm 1.4}$	$47.0_{\pm 3.2}$	$47.6_{\pm 0.6}$
PP-EL2N-EE	$39.4_{\pm 4.9}$	$35.9_{\pm 2.9}$	$36.4_{\pm 4.1}$	$34.4_{\pm 0.2}$	$35.2_{\pm 1.6}$	$52.4_{\pm 5.4}$	$43.0_{\pm 0.0}$	$45.8_{\pm 5.4}$	$43.2_{\pm 0.2}$	$44.4_{\pm 2.6}$	$38.4_{\pm 13.6}$	$28.0_{\pm 9.2}$	$22.5_{\pm 2.9}$	$21.1_{\pm 0.1}$	$21.1_{\pm 0.0}$
PP-EL2N-EH	$43.9_{\pm 7.7}$	$58.2^{*}_{\pm 2.4}$	$60.2_{\pm 1.8}$	$56.2_{\pm 4.9}$	$57.5_{\pm 4.1}$	$55.9_{\pm 5.9}$	$61.4^{*}_{\pm 0.8}$	$60.1_{\pm 1.3}$	$59.2_{\pm 2.1}$	$55.5_{\pm 2.2}$	$49.0_{\pm 6.8}$	$50.7_{\pm 3.0}$	$48.1_{2.0}$	$44.8_{\pm 9.2}$	$46.1_{\pm 1.9}$
PP-EL2N-HE	$37.7_{\pm 2.4}$	$36.9_{\pm 1.3}$	$36.6_{\pm 2.2}$	$35.3_{\pm 0.7}$	$37.2_{\pm 2.1}$	$49.5_{\pm 4.1}$	$50.5_{\pm 4.4}$	$48.3_{\pm 3.2}$	$47.2_{\pm 1.7}$	$51.6_{\pm 4.2}$	$39.0_{\pm 1.2}$	$42.6_{\pm 12.7}$	$37.8_{\pm 11.7}$	$42.6_{\pm 1.4}$	$42.7_{\pm 6.5}$
PP-EL2N-HH	$51.6^{*}_{\pm 6.5}$	$52.8^{*}_{\pm 8.1}$	$53.1_{\pm 4.7}$	$52.9_{\pm 6.4}$	$50.3_{\pm 4.4}$	$60.9^{*}_{\pm 2.1}$	$59.3^{*}_{\pm 4.6}$	$59.7_{\pm 1.3}$	$57.8_{\pm 1.8}$	$57.0_{\pm 2.7}$	$52.8_{\pm 1.5}$	$51.3_{\pm 4.2}$	$49.2_{\pm 2.8}$	$49.2_{\pm 2.6}$	$46.4_{\pm 3.0}$

Table 15: **BERT out-of-distribution performance:** We showcase the Macro-F1 Scores of models trained on original (i.e., unpruned) and pruned CMSB train data and tested on EXIST, EDOS and Hatecheck datasets. The pruning rates indicate the amount of training data that was removed from the CMSB data before the model was trained. The rows correspond to different selection methods. IU denotes *Informed Undersampling*, PP stands for *Proportional Pruning*, H = Hard, E = Easy. We performed Wilcoxon test on the Macro F1-Scores and marked with color (Yellow and Lavender for PVI and EL2N) and stars those scores which were statistically significant higher than the F1 score computed with the unpruned data (see row 1), with p-values below the significance level ($\alpha = 0.05$). The results show that EL2N leads to more statistical significant improvements than PVI, submodular maximization and random pruning. Pruning the hard instances based on EL2N leads to performance gain across EXIST and EDOS but does not carry over to Hatecheck (Sexism). Moreover, we also observe that proportionally pruning the easy sexist and hard non-sexist examples (based on EL2N and PVI) leads to better generalization across EDOS and EXIST datasets.

Method	EXIST					EDOS					Hatecheck				
	15%	30%	40%	50%	60%	15%	30%	40%	50%	60%	15%	30%	40%	50%	60%
Unpruned (0%)			$48.5_{\pm 0.4}$					$60.5_{\pm 0.2}$					$55.1_{\pm 2.3}$		
Random															
random	$47.5_{\pm 2.9}$	$42.6_{\pm 4.1}$	$35.4_{\pm 1.0}$	$34.3_{\pm 0.0}$	$34.3_{\pm 0.0}$	$60.0_{\pm 1.9}$	$55.8_{\pm 6.1}$	$45.9_{\pm}$	$43.0_{\pm 0.0}$	$43.0_{\pm 0.0}$	$56.0_{\pm 1.7}$	$44.9_{\pm 3.7}$	$23.0_{\pm 2.0}$	$21.1_{\pm 0.0}$	$21.1_{\pm 0.0}$
Submod. Max.															
max	$45.5_{\pm 1.3}$	$35.7_{\pm 2.3}$	$34.3_{\pm 0.0}$	$34.3_{\pm 0.0}$	$34.3_{\pm 0.0}$	$53.9_{\pm 5.6}$	$46.5_{\pm 5.7}$	$43.1_{\pm 0.0}$	$43.1_{\pm 0.0}$	$43.1_{\pm 0.0}$	$52.7_{\pm 3.4}$	$27.8_{\pm 1.1}$	$21.1_{\pm .0}$	$21.1_{\pm 0.0}$	$21.1_{\pm 0.0}$
Informed															
IU-PVI-H	$64.4^{*}_{\pm 0.7}$	$65.6^{*}_{\pm 0.4}$	$63.7_{\pm 0.5}$	$62.0_{\pm 1.4}$	$57.8_{\pm 2.0}$	$59.3_{\pm 0.2}$	$54.7_{\pm 0.6}$	$49.5_{\pm 1.0}$	$43.5_{\pm 1.3}$	$39.0_{\pm 4.4}$	$54.8_{\pm 0.4}$	$48.3_{\pm 0.2}$	$47.3_{\pm 1.1}$	$46.4_{\pm 0.3}$	$45.6_{\pm 0.2}$
IU-PVI-E	$47.5_{\pm 0.7}$	$45.7_{\pm 2.8}$	$39.2_{\pm 3.8}$	$36.9_{\pm 3.1}$	$34.3_{\pm 0.0}$	$60.8_{\pm 0.2}$	$59.1_{\pm 6.7}$	$51.7_{\pm 6.5}$	$48.2_{\pm 6.1}$	$43.1_{\pm 0.1}$	$55.0_{\pm 0.5}$	$53.2_{\pm 6.1}$	$40.2_{\pm 15.7}$	$33.1_{\pm 14.8}$	$21.1_{\pm 0.0}$
IU-EL2N-H	$64.5^{*}_{\pm 0.5}$	$65.5^{*}_{\pm 0.2}$	$63.5_{\pm 0.7}$	$60.1_{\pm 1.4}$	$57.4_{\pm 0.9}$	$59.7_{\pm 0.2}$	$53.7_{\pm 0.8}$	$50.1_{\pm 1.1}$	$43.1_{\pm 1.9}$	$36.8_{\pm 2.0}$	$54.7_{\pm 0.3}$	$48.6_{\pm 0.7}$	$47.0_{\pm 0.4}$	$45.3_{\pm 0.4}$	$45.8_{\pm 0.9}$
IU-EL2N-E	$48.4_{\pm 0.7}$	$43.2_{\pm 2.2}$	$39.2_{\pm 4.9}$	$35.1_{\pm 0.9}$	$34.3_{\pm 0.0}$	$60.4_{\pm 0.6}$	$58.1_{\pm 2.0}$	$51.2_{\pm 7.9}$	$45.2_{\pm 2.4}$	$43.0_{\pm 0.0}$	$55.9_{\pm 1.6}$	$53.3_{\pm 4.0}$	$35.6_{\pm 15.3}$	$23.8_{\pm 4.1}$	$21.1_{\pm 0.0}$
Prop. Prune															
PP-PVI-EE	$47.9_{\pm 2.0}$	$40.4_{\pm 4.9}$	$38.3_{\pm 4.5}$	$35.7_{\pm 1.8}$	$34.3_{\pm 0.0}$	$57.9_{\pm 2.6}$	$52.5_{\pm 7.7}$	$49.2_{\pm 6.5}$	$46.0_{\pm 3.7}$	$43.1_{\pm 0.0}$	$56.2_{\pm 1.5}$	$40.0_{\pm 15.5}$	$34.1_{\pm 15.2}$	$24.2_{\pm 3.9}$	$21.1_{\pm 0.0}$
PP-PVI-EH	$59.4^{*}_{\pm 1.1}$	$64.1^{*}_{\pm 1.5}$	$66.0_{\pm 0.5}$	$66.1_{\pm 0.3}$	$65.7_{\pm 0.3}$	$61.7^{*}_{\pm 0.4}$	$59.6_{\pm 0.7}$	$57.5_{\pm 0.5}$	$55.4_{\pm 0.7}$	$54.4_{\pm 1.0}$	$56.2_{\pm 0.3}$	$54.9_{\pm 0.2}$	$52.2_{\pm 1.4}$	$49.9_{\pm 0.9}$	$49.4_{\pm 1.3}$
PP-PVI-HE	$47.3_{\pm 1.3}$	$45.6_{\pm 0.8}$	$44.6_{\pm 1.2}$	$44.3_{\pm 0.9}$	$40.2_{\pm 1.9}$	$60.5_{\pm 0.6}$	$59.6_{\pm 0.4}$	$58.9_{\pm 0.6}$	$58.4_{\pm 0.7}$	$54.4_{\pm 3.4}$	$55.1_{\pm 1.5}$	$54.9_{\pm 0.5}$	$54.7_{\pm 1.4}$	$53.7_{\pm 3.4}$	$47.7_{\pm 7.5}$
PP-PVI-HH	$54.0^{*}_{\pm 0.9}$	$61.7^{*}_{\pm 0.4}$	$62.1_{\pm 0.9}$	$64.0_{\pm 1.0}$	$64.8_{\pm 0.5}$	$61.3^{*}_{\pm 0.1}$	$60.6_{\pm 0.1}$	$59.9_{\pm 0.3}$	$58.9_{\pm 0.4}$	$57.9_{\pm 0.1}$	$56.0_{\pm 1.1}$	$54.8_{\pm 0.4}$	$54.2_{\pm 1.4}$	$52.1_{\pm 0.4}$	$50.7_{\pm 0.8}$
PP-EL2N-EE	$47.8_{\pm 1.7}$	$40.2_{\pm 4.2}$	$39.1_{\pm 4.2}$	$35.4_{\pm 1.8}$	$34.3_{\pm 0.0}$	$58.5_{\pm 3.0}$	$56.2_{\pm 2.8}$	$51.4_{\pm 5.2}$	$45.0_{\pm 3.5}$	$43.1_{\pm 0.1}$	$55.2_{\pm 4.6}$	$42.2_{\pm 8.9}$	$32.6_{\pm 11.1}$	$23.9_{\pm 5.6}$	$21.1_{\pm 0.0}$
PP-EL2N-EH	$56.5^{*}_{\pm 0.6}$	$64.5^{*}_{\pm 0.9}$	$65.9_{\pm 0.6}$	$66.0_{\pm 0.3}$	$64.8_{\pm 1.9}$	$62.0^{*}_{\pm 0.2}$	$59.6_{\pm 0.1}$	$58.1_{\pm 0.3}$	$56.1_{\pm 0.7}$	$55.1_{\pm 2.3}$	$55.9_{\pm 1.0}$	$54.1_{\pm 4.5}$	$52.7_{\pm 0.4}$	$49.8_{\pm 0.5}$	$50.2_{\pm 2.4}$
PP-EL2N-HE	$48.3_{\pm 1.1}$	$45.4_{\pm 1.9}$	$43.9_{\pm 0.6}$	$43.6_{\pm 0.7}$	$39.8_{\pm 2.8}$	$60.0_{\pm 0.5}$	$58.6_{\pm 0.5}$	$58.4_{\pm 0.9}$	$58.2_{\pm 0.4}$	$53.6_{\pm 5.1}$	$56.4_{\pm 1.4}$	$53.5_{\pm 4.0}$	$53.3_{\pm 1.8}$	$53.3_{\pm 0.3}$	$45.6_{\pm 12.5}$
PP-EL2N-HH	$54.8^{*}_{\pm 0.7}$	$61.2^{*}_{\pm 0.9}$	$61.9_{\pm 0.5}$	$64.0_{\pm 0.6}$	$65.3_{\pm 0.6}$	$61.3^{*}_{\pm 0.2}$	$60.5_{\pm 0.5}$	$59.6_{\pm 0.4}$	$58.6_{\pm 0.4}$	$57.4_{\pm 0.4}$	$55.9_{\pm 1.8}$	$55.8_{\pm 0.5}$	$54.9_{\pm 0.6}$	$52.9_{\pm 0.4}$	$50.8_{\pm 1.6}$

Table 16: **RoBERTa out-of-distribution performance:** We showcase the Macro-F1 Scores of models trained on original (i.e. unpruned) and pruned CMSB train data and tested on EXIST, EDOS and Hatecheck datasets. The pruning rates (second row) indicate the amount of training data that was removed from the CMSB data before the model was trained. The rows correspond to different selection methods. We have shortened informed_undersampling (inf. samp.) and proportional_pruning (prop. prune) for convenience. We perform Wilcoxon test on the Macro F1-Scores and mark and color scores (Yellow and Lavender for PVI and EL2N and stars) which were statistically significantly higher than the F1 score of the unpruned data (see row 1) with p-values below the significance level ($\alpha = 0.05$). Our results show that undersampling the hard instances does not lead to statistically significant performance gains in case of EDOS and HateCheck(Sexism). Moreover, we also observe that proportionally pruning the easy sexist and hard non-sexist examples (based on EL2N and PVI) leads to better generalization across EDOS and EXIST datasets. We also observe that performance gradually decreases across all the strategies as we prune more data from the in-distribution training dataset.