Towards Weakly-Supervised Hate Speech Classification Across Datasets

Yiping Jin[1,2], **Leo Wanner**[3,1], **Vishakha Laxman Kadam**[2], **Alexander Shvets**[1]
[1]NLP Group, Pompeu Fabra University, Barcelona, Spain
[2]Knorex, 02-129 WeWork Futura, Pune, India
[3]Catalan Institute for Research and Advanced Studies
{yiping.jin, leo.wanner, alexander.shvets}@upf.edu
vishakha.kadam@knorex.com

Abstract

As pointed out by several scholars, current research on hate speech (HS) recognition is characterized by unsystematic data creation strategies and diverging annotation schemata. Subsequently, supervised-learning models tend to generalize poorly to datasets they were not trained on, and the performance of the models trained on datasets labeled using different HS taxonomies cannot be compared. To ease this problem, we propose to apply extremely weak supervision that only relies on the class name rather than on class samples from the annotated data. We demonstrate the effectiveness of a state-of-the-art weakly-supervised text classification model in various in-dataset and cross-dataset settings. Furthermore, we conduct an in-depth quantitative and qualitative analysis of the source of poor generalizability of HS classification models.

Content Warning: *This document discusses examples of harmful content (hate, abuse, and negative stereotypes). The authors do not support the use of harmful language.*

1 Introduction

Due to a growing concern about its impact on society, hate speech (HS) recognition recently received much attention from the NLP research community (Bilewicz and Soral, 2020). A large number of proposals on how to address HS as a supervised classification task have been put forward; see, among others, (Waseem and Hovy, 2016; Waseem, 2016; Poletto et al., 2021) and several shared tasks have been organized (Basile et al., 2019; Caselli et al., 2020).

However, while Transformer models such as BERT (Devlin et al., 2019) achieved impressive performance on various benchmark datasets (Swamy et al., 2019), recent work demonstrated that state-of-the-art HS classification models generalize poorly to datasets other than the ones they have been trained on (Fortuna et al., 2020, 2021; Yin and Zubiaga, 2021), even when the datasets come from the same data source, e.g., Twitter. This casts a doubt on what we have achieved in the HS classification task.

Fortuna et al. (2022) identify three main challenges related to HS classification: 1. *the definitorial challenge*: while the interpretation of what is HS highly depends on the cultural and social norms of its creator (Talat et al., 2022), state-of-the-art HS research favours a universal definition; 2. *the annotation challenge*: due to the subjective nature of HS, the annotation also often depends on the context, the social bias of the annotator, and their familiarity with the topic (Wiegand et al., 2019), such that the annotators with different backgrounds tend to provide deviating annotations (Waseem, 2016; Olteanu et al., 2018), especially when not only the presence of HS is to be annotated, but also its category and the group it targets (Basile et al., 2019); 3. *the learning and evaluation challenge*: the common evaluation practice of the HS classification models assumes that the distributions of the training data and the data to which the model is applied are identical, which is not the case in reality; real-world HS data is relatively rare, while the strategies applied for the creation of HS datasets favor explicit HS expressions (Sap et al., 2020; Yin and Zubiaga, 2021), using search with explicit target keywords (Waseem and Hovy, 2016; Basile et al., 2019).

In order to address these challenges, we propose the use of extremely weak supervision, which uses category names as the only supervision signal (Meng et al., 2020; Wang et al., 2021): Extremely weak supervision does not presuppose any definition of HS, which would guide the annotation, such that when the interpretation of what is to be considered as HS is modified, we can retrain the model on the same dataset, without the need of re-annotation. Furthermore, when the data distribution changes, the model can learn from unlabeled

The 7th Workshop on Online Abuse and Harms (WOAH), pages 42–59

data and adapt to a new domain.

Our contributions can be summarized as follows:

- We apply extremely weak supervision to HS classification and achieve promising performance compared to fully-supervised and weakly-supervised baselines.
- We perform cross-dataset classification under different settings and yield insights on the transferability of HS datasets and models.
- We conduct an in-depth analysis and highlight the potentials and limitations of weak supervision for HS classification.

2 Related Work

Since our goal is to advance the research on HS classification, we focus, in what follows, on the review of related work in this area and refrain from the discussion of the application of weakly supervised supervision models to other problems.

Standardizing different HS taxonomies across datasets is a first step in performing cross-dataset analysis and experiments. To this end, Fortuna et al. (2020) created a category mapping among six publicly available HS datasets. Furthermore, they measured the data similarity of categories in an intra- and inter-dataset manner and reported the performance of a public HS classification API on different datasets and categories.

Other previous work in cross-dataset HS classification followed similar experimental settings by training a supervised classifier on the training set of each dataset and reporting the performance on the corresponding test set and test sets from other datasets. For instance, Karan and Šnajder (2018) trained linear SVM models on 9 different HS datasets. They showed that models performed considerably worse on out-of-domain datasets. They further performed domain adaptation using the FEDA framework (Daumé III, 2007) and demonstrated that having at least some in-domain data is crucial for achieving good performance. Similarly, Swamy et al. (2019) compared Linear SVM, LSTM, and BERT models trained on different datasets. They reported that some pairs of datasets perform well on each other, likely due to a high degree of overlap. They also claimed that a more balanced class ratio is essential for the datasets' generalizability.

Fortuna et al. (2021) conducted a large-scale cross-dataset experiment by training a total of 1,698 classifiers using different algorithms, datasets, and other experimental setups. They demonstrated that the generalizability does not only depend on the dataset, but also on the model. Transformer-based models have a better potential to generalize to other datasets, likely thanks to the wealth of data they have observed during pretraining. Furthermore, they built a random forest classifier to predict the generalizability based on human-engineered dataset features. The experiment revealed that to achieve cross-dataset generalization, the model must first perform well in an intra-dataset scenario. In addition, inconsistency in class definition hampers generalizability.

Wiegand et al. (2019) and Arango et al. (2019) studied the impact of data bias on the generalizability of HS models, with the outcome that popular benchmark datasets possess several sources of biases, such as bias towards explicit HS expressions, topic bias, and author bias. The classification results dropped significantly when the bias is reduced. To this end, they proposed using cross-dataset classification as a way to evaluate models' performance in a more realistic setting.

Gao et al. (2017) argued that the low frequency of online HS impedes obtaining a wide-coverage HS detection dataset. To this end, they proposed a two-path bootstrapping approach involving an explicit slur term learner and an LSTM (Hochreiter and Schmidhuber, 1997) classifier. The slur term learner is initialized with a list of hand-engineered seed slur terms and applies to an unlabeled dataset to automatically label hateful posts, which are used to train the classifier. The slur term learner and the classifier are trained iteratively in a co-training manner (Blum and Mitchell, 1998).

A distinct approach was proposed by Talat et al. (2018). This approach utilized multi-task learning (MTL) to enhance domain robustness. They trained a classifier on three distinct sets of annotations: Waseem and Hovy (2016), Waseem (2016), and Davidson et al. (2017). While MTL helps to prevent overfitting and may provide auxiliary fine-grained predictions, it requires annotating a dataset using different taxonomies, granularities, or aspects.

Our approach is most similar to Jin et al. (2022)'s, which also applied weakly-supervised learning on a target-domain dataset. However, their approach requires mining a list of 30 high-quality keywords for each category from a large labeled

source-domain dataset. Moreover, they assume that the source and target datasets are labeled using the same HS taxonomy.

3 Weakly-Supervised HS Classification

In this section, we briefly introduce the basics of weakly supervised text classification and then discuss the cross-dataset classification we aim for.

3.1 Preliminaries: Weakly Supervised Text Classification

Weakly-supervised text classification eliminates the need for a large labeled dataset (Meng et al., 2018; Mekala and Shang, 2020). Instead, it trains classifiers using a handful of labeled seed words and unlabeled documents. While the human annotation effort is significantly reduced, weakly-supervised classification methods are sensitive to the choice of seed words, and the process to nominate high-quality seed words is not trivial (Jin et al., 2021).

More recently, Meng et al. (2020) and Wang et al. (2021) explored *extremely* weak supervision, where the model is given only the category name instead of manually curated seed words. Extremely weak supervision is well suited for hate speech detection because we may not know all the aspects of hate speech for a particular category or target group, or what a user may interpret as a HS statement that falls into a specific category. On top of that, extremely weak supervision often performs semantic expansion on the unlabeled dataset and automatically augments the category representation with new aspects (in the form of seed words).

We choose X-Class (Wang et al., 2021) as the primary weakly-supervised classification method because it matches or outperforms previous state-of-the-art weakly-supervised methods on 7 benchmark datasets. X-Class first estimates category representations by iteratively incorporating words similar to the individual categories. More precisely, it represents each word by its averaged contextualized word embedding across the entire dataset and then adds it to the category with whose representation the obtained embedding shows the highest cosine similarity. The category representation is updated as a weighted average of the expanded keywords. Expressly, the authors of X-Class assume a Zipf's law distribution (Powers, 1998) and weight the j-th keyword by $1/j$.

$$s_\ell = \frac{\sum_{j=1}^{|\kappa_\ell|} 1/j \cdot s_{\kappa_{\ell,j}}}{\sum_{j=1}^{|\kappa_\ell|} 1/j} \quad (1)$$

where $\boldsymbol{\kappa_{\ell,j}}$ is the j-th keyword of category $\boldsymbol{\ell}$ and $\boldsymbol{s_{\kappa_{\ell,j}}}$ is its average contextualized embedding. X-Class also performs a consistency check and stops adding new words if a category's nearest words have changed.

Then, X-Class derives the document $\boldsymbol{i}$'s category-oriented representation $\boldsymbol{d_i}$ by weighting each word in the document based on its similarity to the category representations. Afterwards, it clusters the documents using a Gaussian Mixture Model (GMM) (Duda and Hart, 1973) by initializing the category representations as cluster centroids. Finally, the most confident pseudo-labeled documents from each cluster are used to train a text classifier.

In our initial experiments, we observed that while GMM generally improves the pseudo-labeling, the accuracy for some low-frequency categories tends to drop sharply. This is likely because GMM works as a *global* density estimator. Therefore, data of the more frequent categories may "attract" more weights and cause the category representation for low-frequency categories to diverge too much from its initial representation. To address this problem, we introduce an additional *representation*-based prediction, which assigns document $\boldsymbol{i}$ to the category representation which has the highest cosine similarity:

$$\ell_i^{rep} = \arg\max_{\ell \in L} cosine(s_\ell, d_i) \quad (2)$$

We denote GMM's category assignment for document $\boldsymbol{i}$ as 'ℓ_i^{gmm}'. Instead of pseudo-labeling most confident documents based on GMM only, we take the subset of confident documents to which GMM and representation-based prediction assign the same label ($\ell_i^{gmm} = \ell_i^{rep}$). This ensures that the document is sufficiently close to the original category representation. We denote this modified version as 'X-Class^{Agree}'.

3.2 Cross-Dataset Classification

In this work, we study cross-dataset classification, where we do not have any document labels in the target dataset. A dataset is characterized by its *documents* (and their underlying topics and word

distributions) and *taxonomy* (list of categories).[1]

Given a single HS dataset with its corresponding categories, we can straightforwardly apply X-Class using the category names and an unlabeled dataset. On the other hand, both the data distribution and taxonomy may differ when we experiment on different datasets. There are three different cases for the relation between the taxonomies of the source and target datasets.

- **1-to-1:** The target taxonomy is identical to the source taxonomy or a subset of it.
- **N-to-1:** The target taxonomy differs from the source taxonomy, but each target category can be mapped to one or more source categories.
- **N-to-N:** The target taxonomy differs from the source taxonomy, and some target categories cannot be mapped to any of the source categories.

Supervised learning can be applied in the first two cases: We can create a category mapping from the target categories to the source categories, then use this mapping to either *post-process* the model predictions (converting predicted source categories to target categories) or relabel the dataset using the target taxonomy and *retrain* the model. However, in the last case, we cannot directly apply supervised learning without further data collection and annotation because we lack labeled data for at least some categories. In contrast, weakly-supervised methods do not require labeled documents and can readily utilize unlabeled documents in the target dataset to capture the underlying distribution. Furthermore, even when applied to a completely unseen dataset, it can also "relabel" the source dataset using the target taxonomy and bootstrap a classifier.

4 Experiments

4.1 Datasets

We conduct experiments on two popular HS datasets that differ with respect to the data source and taxonomy of HS categories: the Waseem dataset and the SBIC dataset. The Waseem dataset (Waseem and Hovy, 2016)[2] contains 5,355 tweets with sexist and racist content. The dataset was annotated by the authors (inter-annotator agreement $\kappa = 0.84$) and reviewed by a domain expert (a gender studies student who is a non-activist feminist). The SBIC dataset (Sap et al., 2020)[3] contains 44,671 posts collected from different domains: Reddit, Twitter, and hate sites. It was annotated by crowdsource workers on Amazon Mechanical Turk. A small portion of the data is originally from the Waseem dataset (1,816 posts). We exclude these posts to avoid overlap between the two datasets.

SBIC dataset does not set a predefined taxonomy for HS categories. Instead, annotators can indicate the target group with free-text answers. We select the most frequent six target groups that can be mapped to the categories in the Waseem dataset. While our proposed weakly-supervised learning method does not depend on category mapping, we select the SBIC categories that can be mapped to compare with supervised learning baselines. Table 1 shows this category mapping.

Waseem	SBIC
Sexist	Women; LGBT
Racist	Black; Jewish; Muslim; Asian

Table 1: Category mapping between the Waseem and SBIC datasets.

We use the original train/dev/test split (75%/12.5%/12.5%) in the SBIC dataset and randomly split the Waseem dataset to 90%/10% into training and test sets. We apply standard preprocessing following Barbieri et al. (2020), including user mention anonymization and website links and emoji removal. Table 2 presents the distribution of the posts in the two datasets.

4.2 Compared Methods

We compare X-Class with two representative supervised learning baselines which are trained using the full *labeled* training dataset:

- **Support Vector Machines (SVM)** (Cortes and Vapnik, 1995)**:** We use scikit-learn's[4] linear SGD classifier with default hyper-parameters and tf-idf weighting.
- **BERT** (Devlin et al., 2019)**:** We fine-tune the `bert-base-uncased` checkpoint[5] using the exact hyper-parameters to train the final classifier in X-Class (detailed in Section 4.3).

[1] While the term "cross-domain" is more popular than "cross-dataset", it does not suggest that the source and target dataset's taxonomies may differ. The discussion of the related problem of cross-task generalization (Raffel et al., 2020; Sanh et al., 2022), which works for unrelated tasks, is beyond the scope of this work.

[2] https://github.com/zeeraktalat/hatespeech

[3] https://maartensap.com/social-bias-frames/

[4] https://scikit-learn.org

[5] https://huggingface.co/bert-base-uncased

Dataset	Category	# Train	# Test
Waseem	Sexist	3,107	323
	Racist	1,799	177
	Subtotal	4,906	500
SBIC	Women	2,594	351
	Black folks	2,512	576
	Jewish folks	847	207
	LGBT folks	490	53
	Muslim folks	412	85
	Asian folks	224	34
	Subtotal	7,079	1,306

Table 2: Distribution of the posts per dataset. The average number of words per post in the Waseem dataset is 17.1 and in the SBIC dataset 20.0.

We also compare the performance of our model with the following baselines that do not require any document labeling:[6]

- **Majority class:** Always predict the most frequent category in the training dataset.
- **Keyword voting (category name):** Assign the category whose category name occurs most frequently in the document. Fall back to the majority class prediction if there is a tie or none of the keywords appear.
- **Keyword voting (X-Class keywords):** Same as above, but use the expanded keywords in X-Class's category representation and their associated weights. Assign the category that receives the highest score.
- **Zero-shot PET** (Schick and Schütze, 2021a)**:** Prompting a pre-trained BERT model using hand-crafted patterns and verbalizers to classify documents. We provide details of this baseline in Appendix B.
- **WeSTClass** (Meng et al., 2018)[7]**:** CNN-based neural text classifier. It first generates pseudo documents with a generative model seeded with user-provided keywords for pre-training, then conducts self-training to bootstrap from unlabeled documents. We use three manually curated seed words for each category following Meng et al. (2018).
- **LOTClass** (Meng et al., 2020)[8]**:** A strong baseline using extremely weak supervision. The model first uses a masked language model to expand keywords from the category names, then mines category-indicative words using a novel masked category prediction task. Finally, it generalizes via self-training.

4.3 Experiment Settings

We use the official implementation of X-Class.[9] The `bert-base-uncased` checkpoint is used to calculate the document representation and fine-tune the final classifier; the maximum number of keywords for each category is set to 100; and the 50% most confident pseudo-labeled documents from each category are used to train the final classifier.

To facilitate a fair comparison with supervised learning methods, we reimplemented the final classifier fine-tuning step using the HuggingFace Transformers trainer[10] and performed a minimum manual hyper-parameter tuning (`learning_rate=2e-5`; `num_epochs=6`; `weight_decay=0.05`) on the SBIC dev set and applied them on both datasets. We set the `max_length` and `batch_size` to 64.

We merged the following original target groups in the SBIC corpus into "LGBT folks": "gay men", "lesbian women, gay men", "lesbian women", "trans women, trans men", "trans women". Table 3 presents the category names used by the models. We use the original category name except for "LGBT" because it does not occur in the dataset. Instead, we use "gay", the most frequently targeted subgroup in the dataset. As shown in Appendix A, X-Class expands to keywords representing other subgroups in the LGBT community.

4.4 Results of the Experiments

We report the accuracy and macro P/R/F_1 scores to quantify each method's performance.

In-Dataset Classification. We first validate the efficacy of the methods using the standard in-dataset setting, providing the corresponding training and test datasets. Table 4 displays the result.

As expected, BERT outperformed SVM among the supervised-learning baselines on both datasets. Interestingly, keyword voting using only the category name achieved high precision for the SBIC dataset. However, its recall is much lower than that of X-Class due to variations of expressions

[6] We provide the weakly-supervised learning baselines the full *unlabeled* training dataset for keyword expansion and pseudo-labeling.

[7] https://github.com/yumeng5/WeSTClass

[8] https://github.com/yumeng5/LOTClass

[9] https://github.com/ZihanWangKi/XClass

[10] https://huggingface.co/docs/transformers/main/training

Class	Seed	Count	WeSTClass
Sexist	sexist	1,071	sexist sexism misogynist
Racist	racist	33	racist racists racism
Women	women	652	women woman female
Black	black	1,601	black blacks n*gro
Jewish	jewish	142	jewish jews jew
LGBT	gay	209	gay gays homosexual
Muslim	muslim	228	muslim muslims islamic
Asian	asian	121	asian asians chinese

Table 3: Seed words used for each category and their frequency in the training dataset. We manually curated the seed words in X-Class's category representation and select the top-3 ranked keywords to train WeSTClass.

Waseem Dataset		
Model	**Acc**	**P/R/F$_1$**
SVM	97.2	97.1/96.8/96.9
BERT	**98.2**	**98.2/97.8/98.0**
Majority class	64.6	33.2/50.0/39.2
KV (class name)	64.6	57.3/50.1/39.8
KV (X-Class)	67.0	76.9/53.6/47.0
Zero-shot PET	49.2	66.7/59.9/47.3
WeSTClass	77.8	77.8/80.4/77.3
LOTClass	63.2	71.3/70.2/63.2
X-Class	96.2	96.9/94.9/95.8
X-Class^{Agree}	**96.6**	**97.5/95.2/96.2**
SBIC Dataset		
Model	**Acc**	**P/R/F$_1$**
SVM	90.7	93.2/82.5/86.7
BERT	**95.7**	**94.2/95.1/94.6**
Majority class	26.9	4.5/16.7/7.1
KV (class name)	57.7	**85.2**/39.7/41.9
KV (X-Class)	55.2	47.8/45.1/40.8
Zero-shot PET	35.1	38.4/21.6/15.8
WeSTClass	36.4	35.9/34.5/29.9
LOTClass	54.2	29.2/29.3/27.5
X-Class	79.8	74.0/81.8/74.8
X-Class^{Agree}	**81.4**	76.1/**85.3/76.6**

Table 4: In-Dataset performance of various models. We highlight the best performances of supervised and weakly-supervised methods in bold.

within the same category. Using X-Class keywords improved keyword voting's recall by 3.5% and 5.4% on the two datasets. However, the precision dropped significantly on the SBIC dataset, likely due to the noisier keywords.

WeSTClass performs superior to keyword voting baselines on the Waseem dataset, primarily due to its high recall of the "Racist" category. This demonstrates the advantage of semantic representation in neural models. However, its performance pales on the SBIC dataset, revealing its weakness in handling more complex cases that involve class imbalance and overlapping, which has been discussed in Wang et al. (2021) and Jin et al. (2022). LOTClass demonstrates a similar trend, but performs worse on both datasets.[11] We analyze the pseudo-labeling accuracy of weakly-supervised baselines and X-Class in Appendix C.

Comparing X-Class and X-Class^{Agree}, we can see that our modification consistently improved the performance.

Cross-Dataset Classification. We conduct cross-dataset classification using the strongest supervised and weakly-supervised models and show the result in Table 5. Note that for the "Waseem → SBIC" setting, we cannot create a category mapping since the target dataset has more fine-grained categories. Therefore, supervised methods and X-Class using category mapping to post-process the predictions are not applicable.

When we train BERT and X-Class using only source-dataset documents, they both perform worse on the target dataset than the in-dataset results in Table 4. The performance drop is smaller for "SBIC → Waseem", likely because the SBIC dataset contains representative posts for the Waseem categories.

Surprisingly, *retraining* the models using the target taxonomy does not outperform *post-processing* using category mapping. However, when a category mapping is unavailable (as in the "Waseem → SBIC" case), retraining a weakly-supervised classifier using the target taxonomy is the only option for cross-dataset classification without manually annotating more data.

An advantage of weakly-supervised methods is that they can utilize *unlabeled* documents from the target dataset when they are available. Although X-Class^{Agree} still underperforms BERT when both

[11] LOTClass has a higher accuracy on SBIC dataset because it predicts the vast majority of the documents to the most frequent categories "Women" and "Black".

SBIC → Waseem		
Model	**Acc**	**P/R/F$_1$**
BERT (post-process)	**93.6**	**92.4/94.7/93.2**
BERT (retrain)	**93.6**	**92.4**/94.2/**93.2**
X-Class (post-process)	91.6	93.5/88.5/90.3
X-Class (retrain)	84.4	89.9/78.1/80.6
X-Class^{Agree} (post-process)	**92.8**	**94.5**/90.1/**91.8**
X-Class^{Agree} (retrain)	92.6	93.4/**90.4**/91.6
Waseem → SBIC		
Model	**Acc**	**P/R/F$_1$**
X-Class (retrain)	60.7	61.3/59.8/54.5
X-Class^{Agree} (retrain)	**69.8**	**62.7/62.2/58.3**

Table 5: Cross-dataset performance of BERT and X-Class. Both models are trained using source dataset documents and tested on the target dataset. We highlight the best performances of supervised and weakly-supervised methods in bold.

are trained using the source dataset in the "SBIC → Waseem" experiment, it surpasses BERT by 3% in both accuracy and macro F$_1$ score when using unlabeled target-dataset documents [12].

Again, X-Class^{Agree} outperforms X-Class in all cases. Subsequently, we use X-Class to refer to X-Class^{Agree} for brevity.

4.5 Analysis: What Makes Cross-Dataset Classification Challenging?

As shown in Table 5, X-Class's performance dropped significantly in the "Waseem → SBIC" cross-dataset setting compared to the use of the SBIC training set. In this section, we try to uncover the causes of the performance drop.

We first plot the per-category F$_1$ score in Figure 1. We can see that the cross-dataset model achieved comparable performance as the in-dataset model for the four categories {Jewish, Muslim, Women, Black}. However, it failed in the two categories {Asian, LGBT}.

Relevant unlabeled documents. Although the Waseem dataset is labeled using a more coarse-grained taxonomy, it may contain documents relevant to some (but not all) fine-grained SBIC categories. Weak supervision usually pseudo-labels the *unlabeled* dataset to train a final classifier. Therefore, it will likely fail when documents related to a particular category are absent in the unlabeled dataset. We count the frequency of documents containing each category name in both datasets and present the results in Table 6.

[12]We can train weakly-supervised models using unlabeled target dataset, which is equivalent to the in-dataset setting (the X-Class^{Agree} row in Table 4).

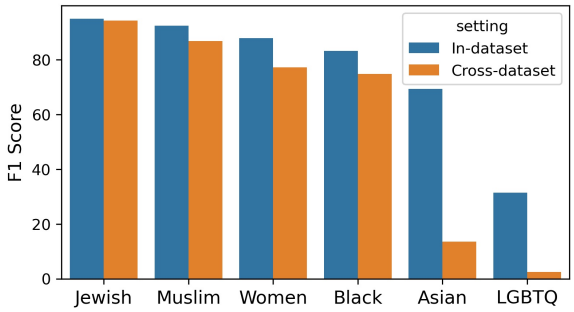

Figure 1: Comparing cross-dataset and in-dataset F$_1$ score of X-Class on the SBIC dataset.

Class	Seed	Waseem %	SBIC %
Sexist	sexist	21.83%	2.70%
Racist	racist	0.67%	1.12%
Women	women	11.94%	9.21%
Black	black	0.63%	22.6%
Jewish	jewish	0.51%	2.00%
LGBT	gay	0.59%	2.95%
Muslim	muslim	10.40%	3.22%
Asian	asian	0.08%	1.71%

Table 6: Frequency of each category name appearing in the Waseem and SBIC training datasets.

We can observe that the "Asian" category (from the SBIC dataset) is severely under-represented in the Waseem dataset. The word "Asian" occurs only 4 times, all in the context of "Asian women/girls".

Waseem and Hovy (2016) conducted a lexical analysis and showed that their "Sexist" category is highly skewed towards *women*, and their "Racist" category is highly skewed towards *Muslims* and *Jews*.[13] Coincidentally, these categories also perform the best in the "Waseem → SBIC" setting.

Category understanding. Jin et al. (2021) argued that weakly-supervised classification and keyword mining are intrinsically related. The failure to identify relevant keywords will harm the category representation and, thus, the classification accuracy. Appendix A presents the full list of keywords X-Class added to the category representations in both in-dataset and cross-dataset settings.

A general observation is that X-Class tends to include fewer keywords in its category representation in the cross-dataset setting. Recall that it stops

[13]Although the term "Jewish" has a low frequency, "Jews" appears in the ten most frequent terms of the "Racist" category.

Model	SBIC → Waseem		Waseem → SBIC	
	Acc	**P/P/F$_1$**	**Acc**	**P/R/F$_1$**
X-Class (src data & src category repr)	92.6	93.4/90.4/91.6	69.8	62.7/62.2/.58.3
X-Class (src data & tgt category repr)	93.4	92.2/94.0/92.9	75.1	65.2/55.5/57.8
X-Class (tgt data & tgt category repr)	96.6	97.5/95.2/96.2	81.4	76.1/85.3/76.6

Table 7: Cross-dataset performance of X-Class using different unlabeled datasets and category representations.

adding keywords once the consistency check is violated. We hypothesize that the mismatch between the dataset and the taxonomy caused the mined keywords to be noisier and more likely to fail the consistency check.

The four categories that perform the best in both in-dataset and out-dataset settings also contain better-quality keywords. In contrast, the "Asian" category's keyword in the cross-dataset setting is entirely off-topic due to its rare occurrence and collocation with words like "women" or "girls". The "LGBT" category contains many vulgar keywords with sexual references, which caused it to confuse with the "Women" category.

Class definition vs. dataset. Previous studies tried to explain why HS classification models generalize poorly across datasets, the most frequently cited reasons being the lack of a standardized definition of hate speech (Waseem and Hovy, 2016; Fortuna et al., 2020, 2021) and biased data distribution (Swamy et al., 2019; Yin and Zubiaga, 2021; Fortuna et al., 2022). It prompts us to wonder *what if* we apply the exact class definition to different datasets or annotate the same dataset using different class definitions. Unfortunately, manual hate speech annotation is time-consuming and very challenging. Waseem (2016) and Caselli et al. (2020) are among the few studies that re-annotated a dataset, providing quantitative analysis or comparing the models' performance. However, such studies focus on a single dataset only. Moreover, the annotation is usually a one-shot effort, influenced by multiple factors related to the annotation task setup and knowledge of annotators. There is no way to assess how much of the performance drop is due to incompatible class definitions and the data distribution *separately*.

In weakly-supervised models, we can interpret the category representation (and associated keywords) as the *class definition*. Therefore, the class definition for the same taxonomy may differ depending on the dataset used to derive the category representation. Furthermore, we can approximate annotating a dataset with a different class definition by altering the category representation.

We designed an ablation study to train X-Class models using different combinations of datasets and class definitions. In Table 7, we present the results of three configurations in this study:[14] 1) Using *source*-dataset documents and category representations derived from the *source* dataset ("X-Class^{Agree} retrain" in Table 5); 2) Using *source*-dataset documents and category representations derived from the *target* dataset; 3) Using *target*-dataset documents and category representations derived from the *target* dataset ("X-Class^{Agree}" in Table 4).

X-Class's cross-dataset performance substantially improved when provided with the category representation derived from the target dataset.[15] Only one factor is altered (either the category representation or the unlabeled training dataset) between the rows in Table 7. Therefore, we can conclude that the performance difference between rows #1 and #2 is due to different *class definitions*, while the performance difference between rows #2 and #3 is due to different *data distributions*.

5 Conclusions and Future Work

We applied extremely weakly-supervised methods to HS classification. We analyzed the transferability of HS classification models through comprehensive in-dataset and cross-dataset experiments and confirmed that weakly-supervised classification has several advantages over the traditional supervised classification paradigm. First, we can apply the algorithm across various HS datasets and domains with taxonomies that cannot be standardized using category mapping. Second, weakly-supervised models can readily utilize unlabeled documents in

[14] All experiments use the target taxonomy, and all documents are unlabeled.

[15] Its average recall in the "Waseem → SBIC" experiment decreased sharply mainly because the category representation for the "Asian" category is far from the document representation (the Waseem dataset does not contain documents related to "Asian"). The model did not predict any document as "Asian".

the target domain and do not suffer from domain mismatch problems. Lastly, weak supervision allows us to "reannotate" a labeled dataset using a different class definition to facilitate cross-dataset comparison, which was previously possible only at the cost of expensive manual annotation.

The presented work is only the beginning of applying weak supervision to HS detection. We can utilize richer category representations than bag-of-keywords. However, such representations should be derived in an unsupervised or weakly-supervised manner to avoid depending on manually labeled datasets. A promising approach in this direction is (Shvets et al., 2021), which extracts HS targets and aspects relying on open-domain concept extraction.

Lastly, we can study how well the model can generalize to previously unknown categories, a more challenging task often known as zero-shot classification (Yin et al., 2019) or open-world classification (Shu et al., 2017).

Limitations

This study utilizes a monolingual pre-trained language model (PLM) in the English language (`bert-base-uncased`). Although the weakly-supervised classification methods are not limited to a particular language, we have not explored applying the method to another language. Social media language use may differ significantly from the data used to train the PLM. Moreover, the presence of code-switching (Doğruöz et al., 2021) may also degrade a monolingual PLM's performance. We explored a RoBERTa checkpoint continually trained with 60M English tweets (Barbieri et al., 2020).[16] However, it does not yield better performance than BERT. We have not investigated whether it is due to the training regime or the dataset.

Moreover, in this work, we focus on classifying hate speech (HS) categories/target groups instead of HS detection (detecting whether a post contains hate speech or not). To perform hate detection and classification, we can either combine our method with another HS detection model in a pipeline or use an adaptation of weakly-supervised text classification incorporating the "Others" category such as Li et al. (2018) or Li et al. (2021).

Due to limited space, we prioritized in-depth analysis instead of a comprehensive evaluation. Therefore, we selected only two datasets (and two-way cross-dataset classification). We are working in parallel on extending this work to a longer-form journal article to cover more datasets and experimental results.

Recent work on large language models (LLMs) demonstrated that when the parameters scale to a certain level, language models exhibit a drastically-increased performance in zero-shot classification (Zhao et al., 2023). We reported the performance of a moderately-sized `bert-large-uncased` zero-shot model because of limited computational resources and lack of access to commercial APIs. Larger language models will likely perform much better than this baseline.

Lastly, understanding HS sometimes requires cultural understanding or background knowledge. It may be difficult to determine the presence and category of HS when we take the post out of its context. For example, many "Sexist" posts in Waseem dataset are tweets related to the Australian TV show *My Kitchen Rules* (MKR), and below is a tweet labeled as "Sexist":

> Everyone else, despite our commentary, has fought hard too. It's not just you, Kat. #mkr

Ethics Considerations

Although weak supervision requires only unlabeled documents, we demonstrated that the model might fail when the training dataset does not contain data related to a particular category or target group. It is especially concerning because the target groups are often minorities and under-represented. Therefore, we recommend against "throwing" a weakly-supervised algorithm on a dataset and hope the model will work. Instead, we should evaluate a model thoroughly before applying it to the real world, such as manually examining the model's predictions, behavioral testing the model using a checklist (Ribeiro et al., 2020) or conducting unsupervised error estimation (Jin et al., 2021).

[16] `https://huggingface.co/cardiffnlp/twitter-roberta-base`

References

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 45–54.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval:

Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

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

Michał Bilewicz and Wiktor Soral. 2020. Hate speech epidemic. the dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, 41:3–33.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the eleventh international AAAI conference on web and social media*, 1, pages 512–515.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. 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 1654–1666, Online. Association for Computational Linguistics.

Richard O Duda and Peter E Hart. 1973. Pattern classification and scene analysis. *A Wiley-Interscience Publication*.

Paula Fortuna, Monica Dominguez, Leo Wanner, and Zeerak Talat. 2022. Directions for nlp practices applied to online hate speech detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, page 11794–11805, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Paula Fortuna, Juan Soler-Company, and Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524.

Lei Gao, Alexis Kuppersmith, and Ruihong Huang. 2017. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 774–782, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Yiping Jin, Akshay Bhatia, and Dittaya Wanvarie. 2021. Seed word selection for weakly-supervised text classification with unsupervised error estimation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 112–118, Online. Association for Computational Linguistics.

Yiping Jin, Dittaya Wanvarie, and Phu TV Le. 2022. Learning from noisy out-of-domain corpus using dataless classification. *Natural Language Engineering*, 28(1):39–69.

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

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the*

2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Chenliang Li, Wei Zhou, Feng Ji, Yu Duan, and Haiqing Chen. 2018. A deep relevance model for zero-shot document filtering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2300–2310, Melbourne, Australia. Association for Computational Linguistics.

Peiran Li, Fang Guo, and Jingbo Shang. 2021. "misc"-aware weakly supervised aspect classification. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 468–476. SIAM.

Dheeraj Mekala and Jingbo Shang. 2020. Contextualized weak supervision for text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 323–333, Online. Association for Computational Linguistics.

Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2018. Weakly-supervised neural text classification. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 983–992, Turin, Italy.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020. Text classification using label names only: A language model self-training approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9006–9017, Online. Association for Computational Linguistics.

Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush Varshney. 2018. The effect of extremist violence on hateful speech online. In *Proceedings of the twelfth international AAAI conference on web and social media*, 1, Stanford, California, USA.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

David MW Powers. 1998. Applications and explanations of zipf's law. In *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, pages 151–160, Sydney, NSW, Australia.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations*.

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.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Lei Shu, Hu Xu, and Bing Liu. 2017. DOC: Deep open classification of text documents. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2911–2916, Copenhagen, Denmark. Association for Computational Linguistics.

Alexander Shvets, Paula Fortuna, Juan Soler, and Leo Wanner. 2021. Targets and aspects in social media hate speech. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 179–190, Online. Association for Computational Linguistics.

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

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience episode #5 – workshop on challenges & perspectives in creating large language models*, pages 26–41, virtual+Dublin. Association for Computational Linguistics.

Zeerak Talat, James Thorne, and Joachim Bingel. 2018. Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In *Online harassment*, pages 29–55. Springer.

Zihan Wang, Dheeraj Mekala, and Jingbo Shang. 2021. X-class: Text classification with extremely weak supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3043–3053, Online. Association for Computational Linguistics.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

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

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Full List of Keywords in X-Class's Category Representation

Table 9 shows the list of keywords in X-Class's category representation in the in-dataset setting (using the unlabeled documents and list of categories from the same dataset). Table 10 shows the list of keywords in X-Class's category representation in the cross-dataset setting (using the unlabeled Waseem dataset documents to induce category representations of SBIC dataset categories and vice versa).

B Reproducibility

Table 11 presents the hyper-parameters and their corresponding values to facilitate reproducing our result.

We use the `bert-large-uncased` model in HuggingFace as the base pre-trained language model for the zero-shot PET baseline. PET combines a *pattern* (or prompt/instruction) with the input text and prompts the model to predict the mask token. Unlike open-ended prompting, PET uses a list of hand-crafted *verbalizers* (candidate tokens). It classifies documents by assigning the category whose associated verbalizer receives the highest predicted probability. PET-style classification is especially beneficial for smaller PLMs, which do not possess a strong capability of instruction following (Schick and Schütze, 2021b; Ouyang et al., 2022).

We hand-crafted patterns and verbalizers based on our understanding of the tasks (without fine-tuning). For Waseem dataset, we use the pattern "<text> This hate speech is based on <mask>" (verbalizers: gender/race), and for SBIC dataset "<text> The target group of this hate speech is <mask>" (verbalizers: women/black/Jews/gay/Muslims/Asian).

C Pseudo-Labeling

Being able to accurately pseudo-label documents is crucial to the success of weak supervision. We report the accuracy of pseudo-labeling by various weakly-supervised methods in Table 8.

We can see that the accuracy of pseudo-labeled documents is consistent with the model's performance on the test dataset (Table 4). Moreover, LOTClass and X-Class use the same underlying pre-trained language model (`bert-base-uncased`) in their final classifier, while WeSTClass uses a more traditional convolutional neural networks architecture (Kim, 2014). The data pseudo-labeled by

Dataset (Method)	Acc	P/R/F_1
Waseem	**99.1**	**98.4/99.2/98.9**
- WeSTClass	77.8	77.9/80.1/77.4
- LOTClass	64.4	72.7/70.9/64.3
SBIC	**93.0**	**89.1/92.8/91.1**
- WeSTClass	35.4	34.8/35.6/29.9
- LOTClass	51.8	32.4/26.3/24.5
SBIC → Waseem	91.2	92.1/90.9/91.0

Table 8: Pseudo-labeled dataset accuracy calculated against the gold-standard labels. The default method is X-Class unless otherwise specified. For the "SBIC → Waseem" setting, we use the category mapping in Table 1 to derive the gold labels. We omit the "Waseem → SBIC" setting because we do not have gold labels.

X-Class is substantially more accurate than the two baselines in both datasets. Comparing Table 8 and Table 4, we can observe that the pseudo-labeling accuracy has a more significant impact on the final classifier's accuracy than the model architecture.

We provide randomly sampled pseudo-labeled documents by X-Class in Table 12 (in-dataset) and Table 13 (cross-dataset). In general, the SBIC dataset contains more diverse and nuanced data. On the other hand, the Waseem dataset sometimes contains trivial slurs like "... I'm not sexist ...". The samples in the cross-dataset setting revealed that X-Class tends to wrongly categorize original "Sexist" posts in the Waseem dataset (which mainly target women) as "LGBT" and "Asian".

Class	Keywords
Sexist	sexist sexism misogynist sl*ts sl*t hypocrisy bigotry c*nts hypocrite bigoted pedophile filth c*nt phony barbarity scum bigot genocidal barbaric raping bitchy bigots rapist rapists blasphemy feminists mongering apostacy delusional trashy bimbos a*sholes skank retarded idiotic morons illiterate behead being sexual gays extremists sex islamophobia apostates whining self islamofascists beheads b*tches rape dudes beheading s*cking an enslave pure up common of a sassy vandaliser gender by feminist
Racist	racist racists racism naziphobia fascist oppression hateful hatred semitic imperialist hating race imperialism genocide inhuman vile ideology violent murderous violence anti nazism vileness brutal propaganda nazis terrorist filthy disgusting radical murdering terrorists hate abuse attacking islamists islamolunatic islamolunatics minority murderers domination jihad terrorism islamist westerners evil killing attack against hated atheists political terror murder culture minorities religious lunatics human conspiracy population hatewatch killings secular religion force cult
Women	women woman female females girls ladies ch*cks wives men feminist lady girl chick feminists feminine males male gender feminism whores blonde virgins bitches guys hookers prostitutes sl*ts mens wh*re sl*t b*tch p*ssy prostitute virgin couples d*cks breast moms c*nts girlfriend wife sisters dudes attractive sexy betas partners she her beautiful genders lovers normies mothers boys man chads adult couple them fathers mensrights normie assholes they body someone bodies looking v*ginas loser dyk*y sister ones femaloid self mate material raped hooker
Black	black white colored blacks whites n*gro african negroes negros racial race racist races minorities color africans n*groids minority n*groid racism mixed brown n*ggers skinned blackman slaves peoples ghetto discrimination n*gger people whitey africa red yellow dark savages individuals civil poor disabled blind gorillas savage human folk nonwhite left lynching slavery diversity worthless folks south gorilla majority violent dirty green cotton slave
Jewish	jewish jews jew synagogue rabbi israel zionist semitic holocaust kosher auschwitz nazi goyim german aryan germans nazis ethiopian germany hitler concentration ash
LGBT	gay homosexual gays homosexuals homosexuality lesbian lesbians queer transgender homophobic sexuality sexual h*mo queers transgenders masculine sexism sexist trans sex sexually straight dating anal dyke dykes penis marriage rape erection pubic openly pedophile porn nude hiv aids raping interracial relationships relationship genitals boyfriends pedophilia objectifying bi std naked d*ck cocks date misogynist misogyny threesome masturbating shaming stoned v*gina assault bestiality c*nt f*cks rapist genital hot c*ck
Muslim	muslim muslims islamic islam mosque mosques arabic quran arab muhammad mohammed shia prophet religion terrorists christian religious allah saudi christians terrorist pakistani arabia ali terrorism pakistan prophets bombers isis syria al qaeda banislam radical camels mass bomber bombing church refugees iran suicide iraq middle faith mosul abdul converted jesus akbar military bomb nations militant pray god kkk militia attacks bible propaganda attack
Asian	asian asians chinese oriental korean japanese american vietnamese indian ethnic mexican americans english latina china eastern foreign exotic european koreans pacific russian north indians spanish russians thai east korea japan country america french cultural western irish countries cuban international nigerian chinaman culture british primitive aged ape inner refugee alien older states europe united animal fat nationality usa russia armed old ignorant special city iq traitor eating animals hungarian food intelligent modern state vietnam rice

Table 9: Full list of keywords in X-Class's category representation mined from *in-dataset* setting.

Class	Keywords
Sexist	sexist sexism homophobic misogyny misogynist hypocrisy sl*ts sl*t c*nts sl*tty degenerate pedophile pedophilia lesbians sexual masculinity bestiality stereotypical shaming whores feminists masturbating mutilation trashy objectifying homosexuals sexually patriarchy misandry raping c*nt rapist hypocritical gays discriminated genital degeneracy unoriginal a*sholes retarded queers virgins disgusting cannibalism self kinky barbarity promiscuity genitals f*cks rape
Racist	racist racism racial discrimination race ethnic races blacks black colored white whites n*gro african negroes minority asians minorities oppression diversity negros ghetto n*groid n*groids mixed cultural peoples africans n*ggers color semitic americans american culture asian individuals people savages savage violence slavery n*gger mixing transgenders mass skinned worthless queer slaves
Women	women woman female females ladies girls feminine feminist feminists feminism womens gender male men girl lady ch*cks blonde blondes males femininity mens wives guys ch*ck wife yesallwomen b*tches daughters her she stars b*tchy girlfriend body b*tch sister feminismisawful announcers promogirls sportscasters bodies models they themselves refs ones them couples someone diva their sjw mother
Black	black white blacks whites racists racist race minorities minority racism africans oppressed americans oppression people population human
Jewish	jewish jews jew judaism israel palestinian zionist palestinians israelis israeli palestine semitic semitism hamas gaza holocaust nazis nazi egyptians
LGBT	gay gays sexual sex sexism sexists sexist rape raping misogynist rapists reproductive misogyny pedophile rapist genitals sl*ts sl*t raped c*nts assault dudes masculinity porn boys shaming c*nt hypocrisy v*gina bigotry rapes bigoted hypocrites hateful haters stereotype openly bimbos wh*re abuse misandrist
Muslim	muslim muslims islamic islam islamist sunni religious islamists jihadi jihadis arab arabs mosques shia quran jihad religion muhammad mohammed taqiyya allah terrorist terrorists prophet believers religions christian hadiths sharia baghdadi secular caliphate hadith saudis saudi pakistani imam christians terrorism islamolunatics isis islamofascists arabian arabia umar extremists hindus pakistan taquiyya medina qurans mullah sunnah westerners
Asian	asian intelligent attractive ignorant young pretty dumb hot rich fat ugly stupid smart tough looking crazy insane blond selfish common brainwashed correct biased clever annoying childish being most hating seeing old beautiful terrible killer self innocent a everydaysexism friendly average ridiculous idiotic extremely poor good bad flawed decent great low simple nice an legit out safe trash doing useless awful corrupt funny sick strong other known working many making best no

Table 10: Full list of keywords in X-Class's category representation mined from *cross-dataset* setting.

Hyper-parameter	Value	Description
random_seed	42	The fixed random seed. Used to split the dataset and initialize parameters.
lm_ckp	bert-base-uncased	The pre-trained language model checkpoint used to derive document representations.
clf_ckp	bert-base-uncased	The pre-trained language model checkpoint used to fine-tune the final classifier. Used in both supervised and weakly-supervised settings.
min_freq	5	Minimum frequency of a word to be included in the vocabulary.
T	100	Maximum terms to include in the category representation.
cluster_method	gmm	Method to perform document class alignment in X-Class. We use the default Gaussian Mixture Model with tied covariance.
pca_dim	64	Dimension of principal component analysis before performing clustering.
conf_threshold	0.5	The percentage of most confident documents assigned by GMM to include in the pseudo-labeled training set.
max_len*	64	The maximum number of tokens of the input posts. Input longer than it will be truncated.
batch_size‡	64	The training batch size.
n_epochs‡	6	The number of training epochs.
learning_rate‡	2e-5	
weight_decay‡	0.05	
lr_schedule†	cosine w/ warmup	Learning rate schedule. Cosine learning rate schedule with 1/3 steps warm-up.

Table 11: Full list of hyper-parameters. The first block is hyper-parameters related to X-Class. The second block is hyper-parameters related to the classifier fine-tuning. * denotes values set based on data analysis. ‡ denotes values set by manual hyperparameter tuning. † denotes values chosen based on our experience but are not tuned. All the rest of the parameters default from the X-Class repo.

Class	Posts	Correct?
Sexist	AbuserDynamics Girls should count themselves lucky if they got a guy that still shouts when his mad, means he cares	✓
	I have a stupid man crush on Lloyd and I can't help it. mkr2015 mkr sexylloydmkr	✓
	RT This lady is a terrible official. Not because she's a lady though. I'm not sexist. But because she's a stupid b*tch. Cavs	✓
Racist	RT CNN reports on the murders of homosexuals by the Islamic State, but the problem isn t just ISIS, it s Islam	✓
	Yes, there is even more rape in Muslim countries but it is not reported.	✓
	For example, the Quran says leaving Islam is not accepted. The Hadiths say kill anyone who leaves.	✓
Women	Molly make b*tches crazy	✓
	The emoji I use when I'm flirting wit b*tches gt	✓
	Rule number two don't trust nobody especially a b*tch with a hookers body	✓
Black	It's funny when u get a black female mad they start acting ghetto then a b*tch	✓
	White bitches be lovin da colored men	✗
	Those ratchet broads don't represent me but YOUR white a*s can't talk trash about them 'cause you on racist sh*t "Black"'	✓
Jewish	What was the Jewish pirate's nickname? Hooknose.	✓
	Do you know why do Jewish kids wait behind the chimney? Because they're waiting for their parents .	✓
	Why do Jews hate a*sholes? The can't stand gas chambers.	✓
LGBT	Any man even considering plastic surgery is gay gay gay imagine ya boy say "I'd get more hoes then u after I get these new lips"	✓
	do it i f*cking dare you f*ggot	✓
	What do you call a gay drive by? A fruit roll up.	✓
Muslim	RT Fuck Islam you pigfucking trash troll.	✓
	Bomb your house, Then f*ck your b*tch. All hail Al Qaeda	✓
	What did the Muslim say after eating a bacon sandwich? That's the way Allah Allah I like it.	✓
Asian	My economics professor just called these asians retards	✓
	Korean idols are perfectly handsome, cute, funny, stylish, hot, know how to dance, have a wonderful voices.	✗
	The Stock Exchange I like the NYSE just like the Ethiopian population count. Going down faster than ever.	✗

Table 12: Randomly sampled pseudo-labeled examples for each category in the in-dataset setting.

Class	Posts	Correct?
Sexist	on sale a*s hoes	✓
	Molly make b*tches crazy	✓
	This n*ggga said I be branding b*tches	✗
Racist	RT Wow the stupid n*gger in LeBron really came out there	✓
	My Moor friends,no not black friends,but Moor friends said N*gger came from Nigeria...You are so lost..Stop tagging me...	✓
	RT Remember the "yellow badge" Nazis used? Israel is making Muslim women carry a yellow badge order to pray in Al Aqsa. h	✓
Women	RT I'm no sexist but the last thing I wanna read about is women's, football or cricket on the sky sports news app! controve	✓
	RT Then I guess Feminism is just a sideshow as much as WWE wrestling in general.. Irony is off the c	✓
	Are you even a real person? I'm not sexist. But Men are superior to women	✓
Black	Can't forget it...never heard about it...	✗
	...with a flat face. The nose a bay window.	✗
	But look at the reality disconnect. Burak says he is for freedom and against all slavery while at the ...	✓
Jewish	Max Blumenthal is bad mouthing you. Not enough room at the top for all the self genocidal Jews. Israel Palestine	✓
	The job Mohammed set Muslims is not done while Israel exists.	✓
	The Jews of Europe should just come to the US. Then the Europeans can allow Islam to take them backwards.	✓
LGBT	RT I'm not sexist but right now I hate girls !!!!	✗
	RT This is not sexist but I want to punch both of the girls from broad city workaholics	✗
	RT This is why girls don't play football. Someone's feelings get hurt and boom, it's out of hand. Go ahead and call me sexist,	✗
Muslim	You didn't recognize the irony of me using your method because you are an ignorant Muslim.	✓
	And you lie again. The majority of Muslims were forced into it.	✓
	RT Arab slave trade 140 to 200 million non Muslim slaves from all colors and nationalities still happening today!	✓
Asian	Someone really needs to get the sniffer dogs onto Kat offherlips MKR	✗
	MKR anyone can cook from a can girls.	✗
	Kat you don't look suspicious at all! MKR	✗

Table 13: Randomly sampled pseudo-labeled examples for each category in the cross-dataset setting.