Cross-Platform and Cross-Domain Abusive Language Detection with Supervised Contrastive Learning

Md Tawkat Islam Khondaker[♠] **Muhammad Abdul-Mageed**[♠♢] **Laks V.S. Lakshmanan**[♠]

[♠]The University of British Columbia, [♢]MBZUAI

{tawkat@cs.,muhammad.mageed@,laks@cs.}ubc.ca

Abstract

The prevalence of abusive language on different online platforms has been a major concern that raises the need for automated cross-platform abusive language detection. However, prior works focus on concatenating data from multiple platforms, inherently adopting Empirical Risk Minimization (ERM) method. In this work, we address this challenge from the perspective of domain generalization objective. We design SCL-Fish, a supervised contrastive learning integrated meta-learning algorithm to detect abusive language on unseen platforms. Our experimental analysis shows that SCL-Fish achieves better performance over ERM and the existing state-of-the-art models. We also show that SCL-Fish is data-efficient and achieves comparable performance with the large-scale pre-trained models upon finetuning for the abusive language detection task.[1]

1 Introduction

Abusive language is defined as any form of microaggression, condescension, harassment, hate speech, trolling, and the like (Jurgens et al., 2019). Use of abusive language online has been a significant problem over the years. Although a plethora of works has explored automated detection of abusive language, it is still a challenging task due to its evolving nature (Davidson et al., 2017; Müller and Schwarz, 2017; Williams et al., 2019). In addition, a standing challenging in tackling abusive language is linguistic variation as to how the problem manifests itself across different platforms (Karan and Šnajder, 2018; Swamy et al., 2019; Salminen et al., 2020).

We provide examples illustrating variation of abusive language on different platforms in Figure 1.[2] For example, user comments in broadcast-

Figure 1: Examples of abusive language on different platforms.

ing media such as Fox News do not directly contain any strong words but can implicitly carry abusive messages. Meanwhile, people on social media such as on Twitter employ an abundance of strong words that can be outright personal bullying and spread of hate speech. On an extremist public forum such as Gab, users mostly spread abusive language in the form of identity attacks. For these reasons, it is an unrealistic assumption to train an abusive language detector on data from one platform and expect the model to exhibit equally satisfactory performance on another platform.

Prior Works on cross-platform abusive language detection (Karan and Šnajder, 2018; Mishra et al., 2018; Corazza et al., 2019; Salminen et al., 2020) usually concatenate examples from multiple sources, thus inherently applying Empirical Risk Minimization (ERM) (Vapnik, 1991). These models capture platform-specific spurious features, and lack generalization (Shi et al., 2021). Fortuna et al. (2018), on the other hand, incorporate out-of-platform data into training set and employ domain-adaptive techniques. Other works such as Swamy et al. (2019) and Gallacher (2021) develop one model for each platform and ensemble them to improve overall performance.

None of the prior works, however, attempt to generalize task-oriented features across the platforms to improve performance on an unseen platform. In this work, we introduce a novel method

[1]Source code: https://github.com/Tawkat/SCL-Fish-Abusive-Language

[2]This paper contains several examples of abusive language and strong words for the purpose of demonstration.

The 7th Workshop on Online Abuse and Harms (WOAH), pages 96–112
July 13, 2023

for learning domain-invariant features to fill this gap. Our approach initially adopts an first-order derivative of meta-learning algorithm (Andrychowicz et al., 2016; Finn et al., 2017), *Fish* (Shi et al., 2021), that attempts to capture domain-invariance. We then propose a supervised contrastive learning (*SCL*) (Khosla et al., 2020) to impose an additional constraint on capturing task-oriented features that can help the model to learn semantically effective embeddings by pulling samples from the same class close together while pushing samples from opposite classes further apart. We refer to our new method as **SCL-Fish** and conduct extensive experiments on a wide array of platforms representing social networks, public forums, broadcasting media, conversational chatbots, and synthetically-generated data to show the efficacy of our method over other abusive language detection models (and specially ERM that prior works on cross-platform abusive language detection applied).

To summarize, we offer the following contributions in this work:

1. We propose SCL-Fish, a novel supervised contrastive learning augmented domain generalization method for cross-platform abusive language detection.

2. Our method outperforms prior works on cross-platform abusive language detection, thus demonstrating superiority to ERM (the core idea behind these previous models). Additionally, we show that SCL-Fish outperforms platform-specific state-of-the-art abusive/hate speech detection models.

3. Our analysis reveals that SCL-Fish can be data-efficient and exhibit comparable performance with the state-of-the-art models upon finetuning on the abusive language detection task.

2 Related Works

2.1 What is Abusive Language?

The boundary between hate speech, offensive, and abusive language can be unclear. Davidson et al. (2017) define *hate speech* as “language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group"; whereas,. Zampieri et al. (2019a) define *offensive language* as “any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct". In this paper, we adopt the definition of abusive language provided by Jurgens et al. (2019) and consider both offensive and hate speech as abusive language in general, since distinguishing between offensive and hate speech is often deemed as subjective (Sap et al., 2019; Koh et al., 2021).

2.2 Domain Generalization

In the domain generalization task, training and test sets are sampled from different distributions (Quiñonero-Candela et al., 2008). In recent years, domain-shifted datasets have been introduced by synthetically corrupting the samples (Hendrycks and Dietterich 2019, Xiao et al. 2020, Santurkar et al. 2020). To improve the capability of a learner on distributional generalization, Vapnik (1991) proposes Empirical Risk Minimization (ERM) approach which is widely used as the standard for the domain generalization tasks (Koh et al. 2021). ERM concatenates data from all the domains and focuses on minimizing the average loss on the training set. However, Pezeshki et al. (2021) state that a learner can overestimate its performance by capturing only one or a few dominant features with the ERM approach. Several other algorithms have been proposed to generalize models on unseen domains. Sagawa et al. (2019) attempt to develop distributionally robust algorithm, where the domain-wise losses are weighted inversely proportional to the domain performance. Krueger et al. (2021) propose to minimize the variation loss across the domains during the training phase and Arjovsky et al. (2020) aim to penalize the models if the performance varies among the samples from the same domain.

2.3 Contrastive Learning

Contrastive learning aims to learn effective embedding by pulling semantically close neighbors together while pushing apart non-neighbors (Hadsell et al. 2006). This method uses cross-entropy-based similarity objective to learn the embedding representation in the hyperspace (Chen et al., 2017; Henderson et al., 2017). In computer vision, Chen et al. (2020) proposes a framework for contrastive learning of visual representations without specialized architectures or a memory bank. Khosla et al. (2020) shows that supervised contrastive loss can outperform cross-entropy loss on ImageNet (Russakovsky et al., 2015). In NLP, similar methods have been explored in in the context of sentence representation learning (Karpukhin et al., 2020; Gillick et al., 2019; Logeswaran and Lee, 2018). Among of the most notable works, Gao et al. (2021) proposes unsupervised contrastive learning framework, *SimCSE* that predicts input sentence itself by augmenting it with dropout as noise.

2.4 Abusive Language Detection

Over the years, the task of abusive language detection have been studied in NLP in the form of hate speech (Davidson et al., 2017; Founta et al., 2018; Golbeck et al., 2017), sexism/racism (Waseem and Hovy, 2016), cyberbulling (Xu et al., 2012; Dadvar et al., 2013). Earlier works in abusive language detection depend on feature-based approaches to identify lexical difference between abusive and non-abusive language (Warner and Hirschberg, 2012; Waseem and Hovy, 2016; Ribeiro et al., 2018). Although inclusion of neural network architecture improves the performance (Mitrović et al., 2019; Kshirsagar et al., 2018; Sigurbergsson and Derczynski, 2020), the models still misclassify a large number of samples in false-positive and false-negative categories when abusive language is intentionally manipulated (Gitari et al., 2015). Recently, Transformer-based (Vaswani et al., 2017) architectures like BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019b) have been introduced in the abusive language detection task (Liu et al., 2019a; Swamy et al., 2019).

However, most of the prior works on abusive language detection focus on a single platform due to the inaccessibility to multiple platforms (Vidgen and Derczynski, 2020) and thus, do not scale well on other platforms Schmidt and Wiegand (2017). As a result, the models are not suitable to apply to other platforms due to the lack of generalization (Karan and Šnajder, 2018; Gröndahl et al., 2018). In this work, we aim to address this challenge by introducing an augmented domain generalization method that captures task-oriented domain-generalized features across multiple platforms.

3 Method

3.1 Challenge & Proposed Solution

As shown in Figure 1, the nature of offensive language can vary from one platform to another. Therefore, it is important to design a model that can capture platform-generalized representations. This inspires us to adopt a domain-generalization algorithm that can maximize feature generalization while avoiding dependence on domain-specific, spurious features. To learn platform-invariant features, we adopt first-order derivative of *Inter-domain Gradient Matching (IDGM)* Shi et al. (2021), a Model Agnostic Meta-Learning (MAML) (Andrychowicz et al., 2016; Finn et al., 2017), algorithm, *Fish*, that aims to reduce sample complexity of new, unseen domains and increase domain-generalized feature selection across those domains.

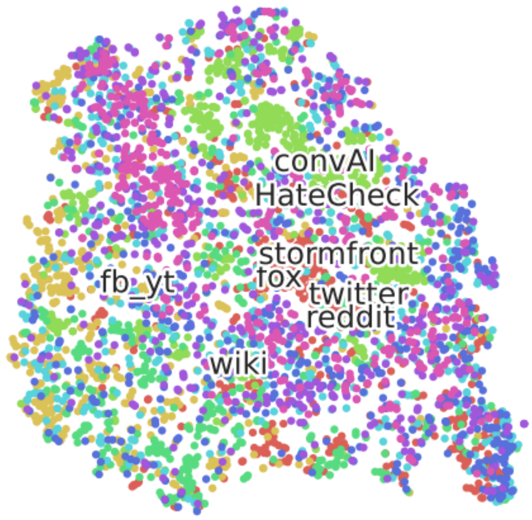

Figure 2: tSNE representations of platforms. We plot the embedding of [CLS] token from pre-trained BERT.

However, if we look at Figure 2, the representation of abusive language across the platforms is overlapping and scattered. Hence, the model should also learn some platform-specific and overlapping features that can help to capture task-oriented representations. Therefore, we need to impose a constraint on the learning objective of the model so that in one direction, it should learn platform-invariant features for better generalization, and in the other direction, it should also learn only those task-oriented overlapping features that pass positive signals to those platform-generalized features for the abusive language detection task.

To learn task-oriented features we introduce **SCL-Fish**, method for supervised contrastive learning (SCL) (Khosla et al., 2020) with Fish. The rationale behind integrating SCL is that we seek to find commonalities between the examples of each class (abusive/normal) irrespective of the platforms and contrast them with examples from the other class.

3.2 SCL-Fish

Assuming we have a training dataset of abusive language detection consisting of samples from two platforms $\mathbf{P_1}$ and $\mathbf{P_2}$ where $\mathbf{P_k} = \{(\mathbf{X}^k, \mathbf{Y}^k)\}$. Given a model θ and loss function l, the empirical risk minimization (ERM) (Vapnik, 1991) objective is to minimize the average loss across the given platform:

$$L_{ERM} = \min_{\theta} \ \mathbb{E}_{\,(x,y)\,\sim\, P\in(P_1,\, P_2)} \frac{\delta l((x,\ y);\ \theta)}{\delta\theta}$$

The expected gradients for these two platforms are expressed as

$$\begin{aligned} G_1 &= \mathbb{E}_{(x,y)\,\sim\, P_1} \frac{\delta l\left((x,\ y)\ ;\ \theta\right)}{\delta\theta} \\ G_2 &= \mathbb{E}_{(x,y)\,\sim\, P_2} \frac{\delta l\left((x,\ y)\ ;\ \theta\right)}{\delta\theta} \end{aligned}$$

If the directions of G_1 and G_2 are same ($G_1.G_2 > 0$), then we can say that the model is improving on both platforms. Therefore, IDGM algorithm attempts to align the direction of the gradients G_1 and G_2 by maximizing their inner dot product. Hence, given the total number of training platforms S, the final objective function of IDGM is obtained by subtracting gradient dot product (GIP) from ERM loss:

$$\begin{aligned} L_{IDGM} &= L_{ERM} \\ &- \gamma \frac{2}{S(S-1)} \sum_{i,j\in S}^{i\neq j} G_i.G_j \end{aligned} \quad (1)$$

Here, γ is a scaling term and GIP can be computed in linear time by $\widehat{G} = ||\sum_i G_i||^2$ - $\sum_i ||G_i||^2$

However, the derivation of $\widehat{G}$ is computationally expensive, as it is a dot product of two gradients. Adopting from Nichol et al. (2018), Shi et al. (2021) work around this issue by proposing a first-order derivative version of IDGM, namely, Fish. Shi et al. (2021) show that given the gradient of ERM $\bar{G}$ and a clone of original model $\tilde{\theta}$,

$$G_f = \mathbb{E}[\theta - \tilde{\theta}] \text{ - } \alpha\, S.\bar{G} \text{ and } G_g = \frac{d\widehat{G}}{d\theta},$$

$$\lim_{\alpha\to 0} \frac{G_f.G_g}{||G_f||\,.\,||G_g||} = 1 \quad (2)$$

In other words, if we ignore the ERM objective, we can substitute the second-order derivative G_g with a computationally less expensive G_f.

Although, this method exhibits impressive performance on the domain-generalization task, as mentioned in Section 3.1, it may capture only platform-invariant features without much focus on *task-relevant* features. To overcome this issue, we augment Fish with a supervised contrastive learning (SCL) objective, which will teach the model to select the features such that the representation of an abusive sample and a non-abusive sample are located far from each other in the hyperspace,

$$\begin{aligned} L_{SCL} = &- \sum_{j=1}^{N} 1_{y_i=y_j} \\ &\log \frac{exp(f(x_i)\,.\,f(x_j)\,/\,\tau)}{\sum 1_{i\neq k}\, exp(f(x_i)\,.\,f(x_k)\,/\,\tau)} \end{aligned} \quad (3)$$

Here, $f(.)$ is an encoder and N is the number of samples summing all the platforms. Therefore, the model will be encouraged to learn only those task-oriented features that are invariant across the platforms *and* can be used to distinguish abusive and non-abusive examples.

Algorithm 1 SCL-Fish

1: **for** iration = 1, 2,... **do**
2: $\quad \tilde{\theta} \leftarrow \theta$
3: $\quad$ **for** $P_i \in \{P_1, P_2, ..., P_S\}$ **do**
4: $\quad\quad$ Sample minibatch $p_i \sim P_i$
5: $\quad\quad \tilde{g_i} = \mathbb{E}_{(x,y)\sim p_i}\left[\frac{\delta l\left((x,\,y)\,;\,\tilde{\theta}\right)}{\delta\tilde{\theta}}\right]$
6:
7: $\quad\quad$ Update $\tilde{\theta} \leftarrow \tilde{\theta} - \alpha\tilde{g_i}$
8: $\quad$ **end for**
9:
10: $\quad$ Update $\theta \leftarrow \theta - \epsilon(\tilde{\theta} - \theta)$ ▷ Updating Fish
11:
12: $\quad P_{scl} \leftarrow \{P_1 \cup P_2 \cup ... \cup P_S\}$
13: $\quad$ **for** Sample minibatch $p_{scl} \sim P_{scl}$ **do**
14: $\quad\quad$ ▷ Calculate gradient for SCL from (3):
15: $\quad\quad g_{scl} = \mathbb{E}_{(x,y)\sim p_{scl}}\left[\frac{\delta l\left((x,\,y)\,;\,\tilde{\theta}\right)}{\delta\tilde{\theta}}\right]$
16:
17: $\quad\quad$ Update $\theta \leftarrow \theta - \alpha'\, g_{scl}$
18: $\quad$ **end for**
19: **end for**

We present SCL-Fish in Algorithm 1. For each training platform, Fish performs inner-loop (*l3-l8*) update steps with learning rate α on a clone of the original model $\tilde{\theta}$ in a minibatch. Subsequently, the original model θ is updated by a weighted difference between the cloned model and the original model $\tilde{\theta} - \theta$. After performing, platform-generalized update, the trained samples of this iteration(*l12*) are queued and sampled in a minibatch

Dataset	Platform	Source	Offnsv/normal
wiki	Wikipedia	Wulczyn et al. (2017)	14880 / 117935
twitter	Twitter	*Multiple**	77656 / 55159
fb-yt	Facebook & Youtube	Salminen et al. (2018)	2364 / 858
stormfront	Stormfront	de Gibert et al. (2018)	1364 / 9507
fox	Fox News	Gao and Huang (2017)	435 / 1093
twi-fb	Twitter & Facebook	Mandl et al. (2019)	6840 / 11491
reddit	Reddit	Qian et al. (2019)	2511 / 11073
convAI	ELIZA & CarbonBot	Cercas Curry et al. (2021)	128 / 725
hateCheck	Synthetic. Generated	Röttger et al. (2021)	2563 / 1165
gab	Gab	Qian et al. (2019)	15270 / 656
yt_reddit	Youtube & Reddit	Mollas et al. (2020)	163 / 163

Table 1: List of experimental datasets with corresponding platforms. * *Twitter* dataset is collected from Waseem and Hovy (2016), Davidson et al. (2017), Jha and Mamidi (2017), ElSherief et al. (2018), Founta et al. (2018), Mathur et al. (2018), Basile et al. (2019), Mandl et al. (2019), Ousidhoum et al. (2019), and Zampieri et al. (2019a).

to update θ with supervised contrastive loss (*l13-l18*).

4 Experiments

4.1 Datasets

To experiment with the efficacy of SCL-Fish, we compile datasets from a wide range of platforms. We collect source of the datasets primarily from (Risch et al., 2021) and (Vidgen and Derczynski, 2020). We provide meta-information of the datasets in Table 1. Description of each dataset is presented in Appendix F.

4.2 Methods Comparison

We compare performance of **SCL-Fish** with **Fish**, also using **ERM** as a sensible baseline. We also conduct experiments on an SCL version of ERM **(SCL-ERM)**. Additionally, we compare SCL-Fish with two of the benchmark models for abusive/hate speech detection, HateXplain (Mathew et al., 2021) and HateBERT (Caselli et al., 2021). **HateXplain** is finetuned on hate speech detection datasets collected from Twitter and Gab[3] for a three-class classification (hate, offensive, or normal) task. It incorporates human-annotated explainability with BERT to gain better performance by reducing unintended bias towards target communities. While conducting our experiments, we consider both *hate* and *offensive* classes as one category (*abusive*). **HateBERT** pre-trains BERT with Masked Language Modeling (MLM) objective on more than one million offensive and hate messages from banned Reddit community. It results in a shifted BERT model that has learned language variety and hate polarity (e.g. *hate*, *abuse*). Finetuning on different abusive language detection tasks has shown that HateBERT achieves the best/comparable performance.

[3]https://gab.com

4.3 Experimental Setup

We train the models (ERM, SCL-ERM, Fish, and SCL-Fish) on *fb-yt*, *twitter*, and *wiki* datasets (in-platform datasets) and use *stromfront* as validation set. We use the same hyperparameters on all the models for fair comparisons. We present the list of hyperparameters in Appendix A. The rest of the datasets from Table 1 are used for cross-platform evaluation. As evident from Table 1, the datasets are highly imbalanced. Hence, we report F_1*-score* for abusive class (we denote it as *positive-*F_1) and *macro-averaged* F_1*-score*. For completeness, we also provide performance in *accuracy*. We train and evaluate our models on Nvidia A100 40GB GPU.

5 Results on Cross-Platform Datasets

We show results of our models for cross-platform performance in Table 2. We observe that SCL-Fish outperforms other methods in macro-F_1 and positive-F_1 scores while maintaining comparable performance with the best method on the other datasets (*reddit*, *hatecheck*). In overall average performance, SCL-Fish achieves best macro-F_1 and positive-F_1 scores. More specifically, user comments on broadcasting media (*Fox News*), SCL-Fish achieves a gain of 3.2% positive-F_1 and 0.5% macro-F_1 over the other methods. On public forums (*Youtube* and *Reddit*), SCL-Fish achieves a total gain of 2.0% in positive-F_1 but SCL-ERM outperforms SCL-Fish by 1.3% in macro-F_1 score. On AI bot conversation (*CarbonBot* and *ELIZA*), SCL-Fish achieves a gain of 1.4% positive-F_1 and 1.0% macro-F_1 over other methods. On the synthetically-generated platform (*HateCheck*), ERM outperforms SCL-Fish by 1.2% in positive-F_1 score and Fish outperforms SCL-Fish by 0.1% in macro-F_1 score. On Gab, all the methods (ERM and Fish-based, including SCL-Fish) achieve high positive-F_1 score because of the highly imbalanced dataset. Hence, for a fair comparison among all methods, we report performance on sampled balanced datasets in Appendix B. We also discuss the performance on the in-platform datasets in Appendix C.

Platform	HateXplain			HateBERT			ERM			SCL-ERM			Fish			SCL-Fish		
(% of hate)	Acc	Pos. F_1	Macro F_1	Acc	Pos. F_1	Macro F_1	Acc	Pos F_1	Macro F_1	Acc	Pos. F_1	Macro F_1	Acc	Pos. F_1	Macro F_1	Acc	Pos. F_1	Macro F_1
stormfront (12.5)	**88.1**	44.1	67.2	87.3	34.6	63.8	85.3	**44.2**	67.7	86.0	43.0	67.5	85.5	42.0	66.9	85.1	**44.2**	**67.8**
fox (28.5)	**73.9**	29.4	56.7	68.7	31.5	63.8	73.6	42.3	62.6	73.6	42.3	62.6	73.6	44.3	63.5	72.2	**47.5**	**64.3**
twi-fb (37.3)	63.4	09.3	43.2	**65.0**	27.9	52.4	61.3	35.7	54.0	60.2	33.6	52.6	53.7	36.9	50.2	61.8	**38.2**	**55.3**
reddit (18.5)	**83.7**	38.0	64.3	81.0	45.5	**66.9**	76.9	43.0	64.3	77.7	43.9	65.1	76.7	44.6	64.9	76.6	**46.3**	65.7
convAI (15.0)	86.4	26.6	59.5	**87.9**	56.9	74.9	86.6	66.3	78.9	86.8	65.9	78.8	86.3	64.7	78.1	87.3	**67.7**	**79.9**
hateCheck (68.8)	38.4	26.9	36.9	58.9	64.3	57.9	**67.3**	**77.4**	59.0	65.4	75.3	58.6	67.1	76.6	**60.5**	66.7	76.2	60.4
gab (95.9)	75.6	85.7	50.6	75.9	86.0	50.4	91.1	95.3	**59.1**	91.4	95.5	57.9	90.9	95.2	58.8	**92.0**	**95.8**	57.4
yt-reddit (50.0)	65.3	54.3	63.2	70.9	69.3	70.8	72.4	75.7	71.9	**74.5**	**77.1**	**74.2**	73.6	76.6	73.2	73.0	76.7	72.3
avg.	71.9	38.9	55.2	74.5	52.0	61.6	76.8	59.9	64.7	**76.9**	59.6	64.7	75.9	60.1	64.5	76.8*	**61.6**	**65.4**

Table 2: Performance on cross-platform datasets. **Bold** font represents the best performance for a particular metric. Gray cells indicate performance on the datasets from identical or overlapping platforms but different sources and distributions. * Although SCL-Fish exhibits comparable accuracy with other competitive models on this imbalanced dataset, it achieves better accuracy on the balanced dataset (Appendix B).

Most notably, HateBERT achieves the highest macro-F_1 scores on *reddit*, which is expected since HateBERT is pre-trained on *reddit* and so has an advantage over other methods since these are trained on data from other platforms. However, all the models including HateXplain and HateBERT are trained on the datasets from Twitter platform. Hence, we analyze performance of the models on *twi-fb* dataset. Our rationale is that although twi-fb involves data from Twitter and Facebook, these data do not necessarily have the same distribution as data used to train all the models. The distribution of datasets from the same platform can still defer due to the variations in the timestamps, topics, locations, demographic attributes (e.g. age, race, gender, ethnicity). Although it is not possible to extract all this information from the textual contents, we provide a quantitative comparison between in-domain and out-domain datasets for Twitter in Appendix D. We refer the readers to Koh et al. (2021) for more detailed analysis. We find that performance of the models deteriorates significantly (under 56% macro-F_1) even on datasets from overlapping platforms but of different distributions. This demonstrates effect of distribution shift in the data, even if we train on date from the same platform. We further discuss possible rationales for this performance gap across the platforms in Appendix E.

6 Analysis

In this section, we conduct qualitative and quantitative analysis on the experimental results.

6.1 Diversity over Quantity

It is worth noting that HateBERT has been pre-trained on 1,478,348 Reddit messages, almost five times more data than SCL-Fish. However, as Table 2 shows, performance of HateBERT on cross-platform datasets suffers significant drops which is not the case for SCL-Fish. Even on *yt-reddit* dataset, which is collected from *Youtube* and *Reddit* (the latter being the platform whose data HateBERT is trained on), HateBERT fails to outperform the baseline ERM method. This shows that, for the purpose of creating platform/domain-invariant models, it is more important to employ training data with different distributions than simply using huge amounts of training data from the same platform but that may have limited distribution.

6.2 Finetuning SCL-Fish

Since we show SCL-Fish exhibits better performance than other methods on most of the cross-platform datasets, we further investigate whether the platform-generalization capability of SCL-Fish helps it improve performance on a specific platform (*Twitter*) upon finetuning. For this purpose, we use two benchmark datasets, namely, OLID (Zampieri et al., 2019a) dataset from SemEval-2019 Task 6 (Zampieri et al., 2019b) and AbusEval (Caselli et al., 2020). Please note that we use OLID dataset for training our methods (Appendix F). Now we are finetuning with the same dataset for this experiment.

We present results for this set of experiments in Table 3. Performance of NULI (BERT-based

Datasets	Models	Macro F_1	Pos. F_1
OffensEval	BERT	80.3	71.5
	HateBERT	80.9	72.3
	NULI	**82.9**	**75.2**
	SCL-Fish	<u>81.6</u>	<u>72.6</u>
AbusEval	BERT	72.7	55.2
	HateBERT	**76.5**	**62.3**
	Caselli et al. (2020)	71.6	53.1
	SCL-Fish	<u>75.2</u>	<u>59.4</u>

Table 3: Performance of models after finetuning. **Bold** and *underline* represent best and second best performance for a particular metric, respectively.

model secured first position in SemEval-2019 Task 6 (Zampieri et al., 2019b)) in the table is from Liu et al. (2019a) and BERT, HateBERT from Caselli et al. (2021).

As Table 3 shows, NULI (Liu et al., 2019a) achieves the best performance for OLID dataset. Although SCL-Fish gets a lower score than NULI[4], SCL-Fish outperforms BERT and HateBERT on both in positive-F_1 and macro-F_1. This is important because HateBERT uses five times more data from one specific platform (*Reddit*). This proves that our proposed SCL-Fish is useful not only in platform generalized zero-shot setting but also for finetuning, and emphasizes the importance of *diversity* of the data (which translates into varied distributions) over data *size*.

For AbusEval dataset, SCL-Fish performs better than BERT and the prior work (Caselli et al., 2020), but it cannot outperforms HateBERT. We suspect that the reason is due to the different annotation process followed during the earlier training phase of SCL-Fish and HateBERT. Because, although OLID and AbusEval contain identical tweets in the training and the testing sets, the annotation scheme of AbuseEval is different from OLID. While Zampieri et al. (2019a) uses the definition of offensive language as "Posts containing any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct" to annotate OLID dataset, Caselli et al. (2020) uses the definition of abusive language as "hurtful language that a speaker uses to insult or offend another individual or a group of individuals based on their personal qualities, appearance, social status, opinions, statements, or actions" to annotate AbusEval dataset. More comprehensively, AbusEval excludes any kind of untargeted messages from the hate speech category. During the training phase of SCL-Fish, we consider any targeted or non-targeted strong language as offensive. Therefore, finetuning on AbusEval causes misalignment with the earlier training phase of SCL-Fish, and may result in performance deterioration.

[4]Please note that Caselli et al. (2021) reports positive-F_1 of NULI as 59.9% which is lower than positive-F_1 of SCL-Fish. But the positive-F_1 we compute from Liu et al. (2019a) is different from the one reported in Caselli et al. (2021). Therefore, we consider our computed positive-F_1 for NULI.

6.3 Explainability with Attention Visualization

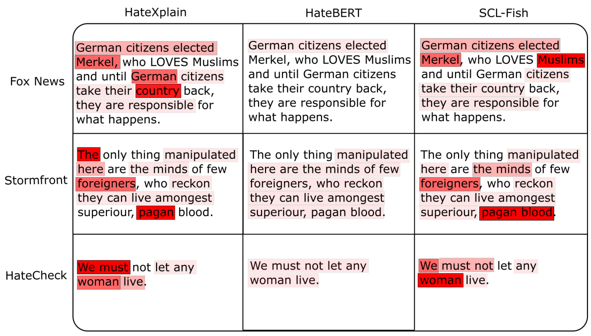

Figure 3: Attention visualization for different platforms. Deeper color indicates higher attention.

We investigate how platform generalization helps the model attend to the right context on 'out-of-platform' datasets. For this purpose, we analyze attention vectors of SCL-Fish, HateXplain, and HateBERT in an attempt to better understand their performance. We use BertViz (Vig, 2019) to compute and visualize the final layer attention vectors from `[CLS]` to other tokens. We select three out-of-platform datasets (*fox*, *stormfront*, and *hateCheck*) and randomly sample one abusive example from each where SCL-Fish correctly identifies the example as abusive, but HateXplain and HateBERT misclassify it. Figure 3 shows the attention visualization for each of the examples. As we can see, in the example from Fox News user comments, although the text does not explicitly contain any strong or offensive words, it is seemingly offensive towards 'Muslims' and 'Merkel'. Hence, our models should attend to these two words with the highest priority, which SCL-Fish does. On the other hand, although HateXplain gives higher attention to 'Merkel', it fails to attend the word 'Muslims'. Surprisingly, HateBERT does not assign priority to any context for the misclassified examples. On the example from StormFront, both SCL-Fish and HateXplain, correctly assign priority to the words 'foreigners' and 'pegan' unlike HateBERT. However, HateXplain also confuses other words e.g. 'The' as a highly prioritized token. Finally, the example from synthetically-generated dataset *hateCheck* is challenging because of the linguistic complexity (e.g. negations, hedging terms) language models typically struggle to address (Hossain et al., 2020; Ettinger, 2020; Kassner and Schütze, 2020). We observe that SCL-Fish highly prioritizes 'women' and also attends to the token 'not'. On the other hand, HateXplain mistakenly provides the highest attention to 'We must' and ignores the negation term 'not'.

Overall, our analysis shows that model trained on platform-generalized settings improves on identifying the targeted community and right context on an out-domain offensive text. On the contrary, platform-specific models may not be able to attend to the targeted community in a different platform, because these models are trained on target specific to particular platforms.

6.4 SCL Improves Fish

From Table 2 and Table 8, it is evident that integrating SCL with Fish empirically improves performance across the platforms. Now, we substantiate the empirical result with the visual justification for Fish and SCL-Fish on different platforms. For all the platforms, we pass an equal number of abusive and non-abusive samples to the models and plot the [CLS] embeddings in Figure 4.

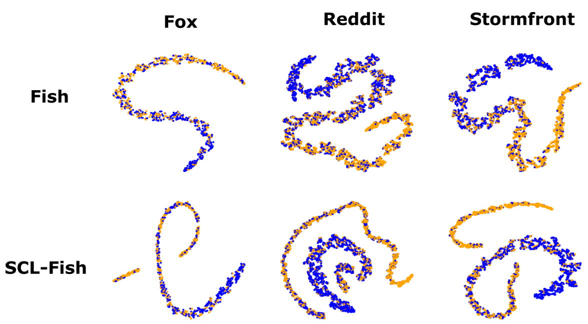

Figure 4: tSNE plot for Fish vs. SCL-Fish on Fox News Comment, Reddit, and StormFront. Abusive samples are presented in *orange* and non-abusive samples are presented in *blue*.

We observe that, SCL-Fish forms more compact clusters of abusive (majority from *orange* samples) and non-abusive (majority from *blue* samples) examples than Fish. Supervised contrastive learning attempts to learn task-oriented features that help bring representations of the same class closer to each other while pushing representations of different classes further apart. As a result, distinct clusters are formed for each class in Figure 4. Therefore, incorporating SCL helps Fish reduce the confusion between abusive and non-abusive representations and improves overall performance of the model.

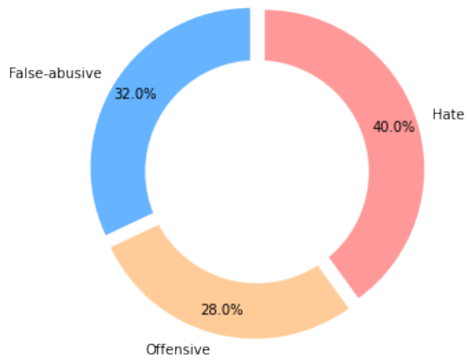

Figure 5: Percentage of error categories in a randomly sampled 50 misclassified examples.

7 Error Analysis

We conduct an error analysis on the examples that SCL-Fish misclassified. We randomly sample 50 misclassified examples and divide them into three categories:
False-abusive: Examples that are normal but SCL-Fish categorizes them as abusive.
Offensive: Examples that are degrading, harassing to an individual or untargeted abuse, trolling but SCL-Fish categorizes them as normal.
Hate: Examples that contain targeted attacks towards a particular group or identity.

We provide examples for each category in Table 4. Figure 5 shows that SCL-Fish misclassify 32% normal examples as false-abusive. Most of the examples of this category contain some sort of slang words that the model confuses as abusive. On other hand, SCL-Fish misclassifies 28% of the offensive examples as normal. This is because the examples may contain some positive words (e.g. 'please') or do not contain any profanity. Therefore, the model considers them as normal speech. Lastly, around 40% of the hate speech is misclassified as normal by SCL-Fish. Similar to the reason for offensive, the model confuse because of some sarcastic positive words and lack of expected profanity. This analysis shows that detecting implicit abusive language that does not contain direct profanity is still challenging and a direction to be explored in the future.

Category	Example	Predicted Class
False-abusive	Once at work, I accidentally checked out some dragons f*cking cars. Didn't know that would be an issue.	Abusive
Offensive	The girl on the left looks like a Jewish pizza.	Normal
Hate	The Blacks and Browns are destroying Sweden and nobody does anything about it!	Normal
False-abusive	I've been out for years and it hasn't been easy, but I also love being a f*g.	Abusive
Offensive	Why do you not think so? Feck off. Please go on. You annoy me.	Normal
Hate	All we have is Blacks and a few Mestizos and all are pretty well behaved, we keep them that way around here.	Normal
False-abusive	I truly am a bada*s motherf*cker.	Abusive
Offensive	Oh yeah, she's so annoying I had mentally blocked out her existence.	Normal
Hate	As opposed to Reddit, which is composed mainly of Communists jacking each other off about child drag queens?	Normal

Table 4: Sample of misclassified examples from *False-abusive*, *Offensive*, and *Hate* category.

8 Conclusion

In this work, we addressed the problem of cross-platform abusive language detection from the domain generalization perspective. We proposed SCL-Fish, a supervised contrastive learning augmented meta-learning method to learn generalized task-driven features across platforms. We showed that SCL-Fish achieves better performance compared to the other state-of-the-art models and models adopting ERM for cross-platform abusive language detection. Our analysis also reveals that SCL-Fish achieves comparable performance on finetuning with much smaller data for cross-platform training than other data-intensive methods. Our work demonstrates progress on both platform and domain generalization in the context of abusive language detection, which we hope future research can be extended to other areas of language understanding.

9 Limitations

Although SCL-Fish achieves improvement over Fish, training SCL-Fish takes longer time than Fish. Empirically, we find that SCL-Fish is approximately 1.2x slower than Fish. Moreover, we believe that the subjective nature of abusive language (Sap et al., 2019) affects the annotation process of different datasets and possibly negatively impact performance.

Acknowledgements

MAM acknowledges support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 435-2018-0576; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), and Digital Research Alliance of Canada.[5]

References

Marcin Andrychowicz, Misha Denil, Sergio Gómez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando de Freitas. 2016. Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2020. Invariant risk minimization. In *International Conference on Machine Learning*.

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.

[5]https://alliancecan.ca

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

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 12th Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Ting Chen, Yizhou Sun, Yue Shi, and Liangjie Hong. 2017. On sampling strategies for neural network-based collaborative filtering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 767–776, New York, NY, USA. Association for Computing Machinery.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2019. Cross-platform evaluation for italian hate speech detection. In *CLiC-it*.

M. Dadvar, Rudolf Berend Trieschnigg, Roeland J.F. Ordelman, and Franciska M.G. de Jong. 2013. Improving cyberbullying detection with user context. In *Proceedings of the 35th European Conference on IR Research, ECIR 2013*, Lecture Notes in Computer Science, pages 693–696, Netherlands. Springer.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8(0):34–48.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.

Paula Fortuna, José Ferreira, Luiz Pires, Guilherme Routar, and Sérgio Nunes. 2018. Merging datasets for aggressive text identification. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 128–139, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

John D Gallacher. 2021. Leveraging cross-platform data to improve automated hate speech detection.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, Varna, Bulgaria. INCOMA Ltd.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 528–537, Hong Kong, China. Association for Computational Linguistics.

Njagi Dennis Gitari, Zhang Zuping, Zuping Zhang, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. In *International Journal of Multimedia and Ubiquitous Engineering*.

Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, and Derek Michael Wu. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, page 229–233, New York, NY, USA. Association for Computing Machinery.

Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. 2018. All you need is "love": Evading hate speech detection. AISec '18, page 2–12, New York, NY, USA. Association for Computing Machinery.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2:1735–1742.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yunhsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply.

Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online. Association for Computational Linguistics.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16, Vancouver, Canada. Association for Computational Linguistics.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy. Association for Computational Linguistics.

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.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR.

Rohan Kshirsagar, Tyrus Cukuvac, Kathy McKeown, and Susan McGregor. 2018. Predictive embeddings for hate speech detection on Twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 26–32, Brussels, Belgium. Association for Computational Linguistics.

Ping Liu, Wen Li, and Liang Zou. 2019a. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop* on Semantic Evaluation, pages 87–91, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*, volume abs/1803.02893.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. FIRE '19, page 14–17, New York, NY, USA. Association for Computing Machinery.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Puneet Mathur, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. 2018. Did you offend me? classification of offensive tweets in Hinglish language. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 138–148, Brussels, Belgium. Association for Computational Linguistics.

Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2018. Neural character-based composition models for abuse detection. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 1–10, Brussels, Belgium. Association for Computational Linguistics.

Jelena Mitrović, Bastian Birkeneder, and Michael Granitzer. 2019. nlpUP at SemEval-2019 task 6: A deep neural language model for offensive language detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 722–726, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2020. Ethos: an online hate speech detection dataset. *Complex & Intelligent Systems*.

Karsten Müller and Carlo Schwarz. 2017. Fanning the flames of hate: Social media and hate crime. *SSRN Electronic Journal*.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. 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 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron C. Courville, Doina Precup, and Guillaume Lajoie. 2021. Gradient starvation: A learning proclivity in neural networks. In *35th Conference on Neural Information Processing Systems (NeurIPS)*.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. 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 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2008. *Dataset shift in machine learning*. Mit Press.

Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on twitter. In *Twelfth international AAAI conference on web and social media*.

Julian Risch, Philipp Schmidt, and Ralf Krestel. 2021. Data integration for toxic comment classification: Making more than 40 datasets easily accessible in one unified format. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 157–163, Online. Association for Computational Linguistics.

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.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. Imagenet large scale visual recognition challenge.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*.

Joni Salminen, Hind Almerekhi, Milica Milenković, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard Jansen. 2018. Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Joni O. Salminen, Maximilian Hopf, S. A. Chowdhury, Soon-Gyo Jung, Hind Almerekhi, and Bernard Jim Jansen. 2020. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10:1–34.

Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. 2020. BREEDS: benchmarks for subpopulation shift. *CoRR*, abs/2008.04859.

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.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Yuge Shi, Jeffrey Seely, Philip H. S. Torr, N. Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. 2021. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive language and hate speech detection for Danish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3498–3508, Marseille, France. European Language Resources Association.

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

V. Vapnik. 1991. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS ONE*, 15(12):e0243300.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, LSM '12, page 19–26, USA. 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.

Matthew L Williams, Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. 2019. Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime. *The British Journal of Criminology*, 60(1):93–117.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399.

Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. 2020. Noise or signal: The role of image backgrounds in object recognition. *CoRR*, abs/2006.09994.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 656–666, Montréal, Canada. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. 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 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

A Hyperparameter Configuration

The detailed configuration of hyperparameters for the training phase of the cross-platform experiments is shown in Table 5. We run each experiment three times and report the average performance of the models.

Hyperparameters	Values
LM model variant	BERT-base-uncased
Token length	512
Optimizer	Adam
AdamW epsilon	1e-8
AdamW betas	(0.9, 0.999)
Fish meta lr. (ϵ)	0.05
SCL temperature (τ)	0.05
Learning rate	5e-6
Batch size	8
Epochs	10

Table 5: Hyperparameters for cross-platform experiments.

Table 6 presents the configuration of hyperparameters during the finetuning (Section 6.2).

Hyperparameters	Values
LM model variant	BERT-base-uncased
Token length	100
Optimizer	AdamW
AdamW epsilon	1e-8
AdamW betas	(0.9, 0.999)
Learning rate	1e-5
Batch size	32
Epochs	5

Table 6: Hyperparameters for finetuning.

B Performance on Cross-Platform Balanced Datasets

We sample an equal number of examples from abusive and normal classes for each dataset. The result is shown in Table 7.

C In-Platform Performance

Table 8 shows the performance of the methods on the in-platform datasets. Unsurprisingly, ERM-based methods outperform Fish-based methods on all the datasets and in all the metrics. ERM method learns platform-specific features, while the Fish-based method tends to learn platform-invariant features. Therefore, evaluating the in-platform datasets yield better performance for ERM-based methods. Notably, as the percentage of abusive speech decreases from the top row to the bottom row in Table 8, positive-F_1 scores also drop accordingly. But Fish-based methods suffer least performance deterioration (10.1% drop from *fb-yt* to *wiki* for SCL-Fish, 7.2% drop from *fb-yt* to *wiki* for Fish) than the other methods (12.3% drop from *fb-yt* to *wiki* for ERM, 12.7% drop from *fb-yt* to *wiki* for SCL-ERM). This shows that domain generalization helps the methods to learn more robust platform-invariant features, which in turn, results in more accurate detection of abusive speech on cross-platform datasets.

D Quantitative Comparison for *Twitter* In-Domain and Out-Domain Datasets

We compare *twitter* (in-domain) and *twi-fb* (out-domain) datasets based on linguistic features and sentiment analysis. For each dataset, we compute average sentiment scores, average number of words, and characters for both abusive and non-abusive classes.

Table 9 reflects the difference in sentiments scores and linguistic features between the datasets. We see that the number of words and the number of characters are higher for the out-domain (*twi-fb*) dataset than the in-domain (*twitter*) dataset for both abusive and non-abusive classes. Additionally, the examples of out-domain datasets have more negative sentiment on average than the examples of in-domain dataset. These types of variation can shift the distribution of the datasets, as a result, the models may struggle to perform better on an out-domain dataset (Table 2).

E Rationale for Performance Gap across Platforms

To this end, we aim to study the reason for the performance gap of the models across different platforms through a qualitative analysis of linguistic variance. We sample abusive texts from the platforms and plot the word frequency in Figure 6.

We observe that the type of abusive texts varies along with the linguistic features across the platforms. For example, on social networks like Twitter, most appeared words in abusive texts are ‘f*cking’, ‘gun’, ‘a*s’, which mostly imply violence and personal attack. Meanwhile, an extremist forum like Stormfront contains words like ‘black’,

Platform	HateXplain			HateBERT			ERM			SCL-ERM			Fish			SCL-Fish		
	Acc	Pos. F_1	Macro F_1	Acc	Pos. F_1	Macro F_1	Acc	Pos. F_1	Macro F_1	Acc	Pos. F_1	Macro F_1	Acc	Pos. F_1	Macro F_1	Acc	Pos. F_1	Macro F_1
stormfront	64.7	48.5	60.9	61.9	41.2	56.5	69.2	60.1	67.5	67.5	56.4	65.3	67.3	56.1	65.0	**69.5**	**60.6**	**67.8**
fox	57.0	30.7	49.8	55.6	36.3	51.1	61.5	46.9	58.4	61.6	46.9	58.4	61.8	49.1	59.3	**63.3**	**54.6**	**61.9**
twi-fb	51.6	09.4	38.2	55.5	29.0	48.3	54.7	38.9	51.5	53.4	36.6	49.9	50.0	**42.1**	49.1	**55.8**	41.6	**53.0**
reddit	61.6	41.4	56.4	66.1	55.8	64.2	65.9	57.9	64.6	66.1	58.1	64.8	67.1	60.6	66.2	**68.2**	**63.2**	**67.6**
convAI	57.8	28.0	49.1	73.4	66.7	72.3	87.1	87.2	87.1	86.3	86.2	86.3	85.5	85.3	85.5	**87.9**	**87.9**	**87.9**
hateCheck	52.3	27.5	45.9	**63.4**	60.9	**63.3**	59.5	**67.3**	57.1	59.1	65.7	57.6	60.9	67.1	59.5	60.8	66.8	59.5
gab	33.8	41.0	32.8	33.9	42.7	32.3	64.1	72.8	**60.1**	62.2	72.1	56.7	**64.3**	**72.9**	**60.1**	60.2	71.1	53.6
yt-reddit	65.3	54.3	63.2	70.9	69.3	70.8	72.4	75.7	71.9	**74.5**	**77.1**	**74.2**	73.6	76.6	73.2	73.0	76.7	72.3
avg.	55.5	35.1	49.5	60.1	50.2	57.4	66.8	63.3	64.8	66.4	62.4	64.2	66.3	63.7	64.7	**67.3**	**65.3**	**65.5**

Table 7: Performance on the **balanced** cross-platform datasets. **Bold** font represents best performance for a particular metric. Gray cells indicates performance on the datasets from identical or overlapping platforms but different sources and distributions.

Platform	ERM			SCL-ERM			Fish			SCL-Fish		
(% of hate)	Acc	Pos. F_1	Macro F_1	Acc	Pos. F_1	Macro F_1	Acc	Pos. F_1	Macro F_1	Acc	Pos. F_1	Macro F_1
fb-yt (73.4)	**94.1**	**95.8**	**92.9**	92.9	94.9	91.4	79.9	85.1	77.1	90.1	92.8	88.5
twitter (58.5)	**89.2**	90.7	**88.9**	**89.2**	**90.8**	88.8	84.0	85.8	83.8	**89.2**	90.7	**88.9**
wiki (11.2)	**96.2**	**83.5**	**90.7**	96.0	82.2	89.9	95.1	77.9	87.6	95.9	82.7	90.2
avg.	**93.2**	**90.0**	**90.8**	92.7	89.3	90.1	86.3	82.9	82.8	91.8	88.7	89.2

Table 8: Performance on in-platform datasets. **Bold** font represents best performance for a particular metric.

Class	**Features**	***twitter***	***twi-fb***
Abusive	No. of Words	15.49	29.64
	No. of Characters	96.53	168.46
	Sentiment Score	-0.75	-0.83
Non-Abusive	No. of Words	18.51	26.84
	No. of Characters	118.41	172.09
	Sentiment Score	-0.49	-0.71

Table 9: Comparison between in-domain (*twitter*) and out-domain (*twi-fb*) datasets. Features are computed averaging the examples for a particular class (abusive/non-abusive).

'white', 'jews' which indicate abusive comments towards a particular community or ethnicity. Linguistic features from a public forum like Reddit reveal that abusive comments on this platform are mostly targeted attacks and slang. Abusive conversation with AI bots mostly contains strong words in the form of personal attacks. On the other hand, user comments on broadcasting media like Fox News do not contain any strong words but rather implicit abuse focused towards a particular race like 'black', person like 'Obama', or sexual orientations like 'gay'. Finally, abusive texts on Wikipedia include both targeted and untargeted slang words toward a specific entity.

The variation of abuse across different platforms shows that training models on a specific platform are not enough to address the issue of mitigating abusive language on another platform. This also implies the importance of the platform-generalized study of abusive language detection.

F Datasets Description

In this section, we briefly describe the datasets we compile for our cross-platform experiments.

F.1 wiki

wiki dataset represents Wikipedia platform. We collect this dataset from Wulczyn et al. (2017). The corpus contains 63M comments from discussions relating to user pages and articles dating from 2004 to 2015. Human annotations were used to label personal attack, aggressiveness, and harassment. The authors find that almost 1% of Wikipedia comments contain personal attacks. We randomly sample 132,815 examples from the initial corpus to make it compatible in size with other training sets.

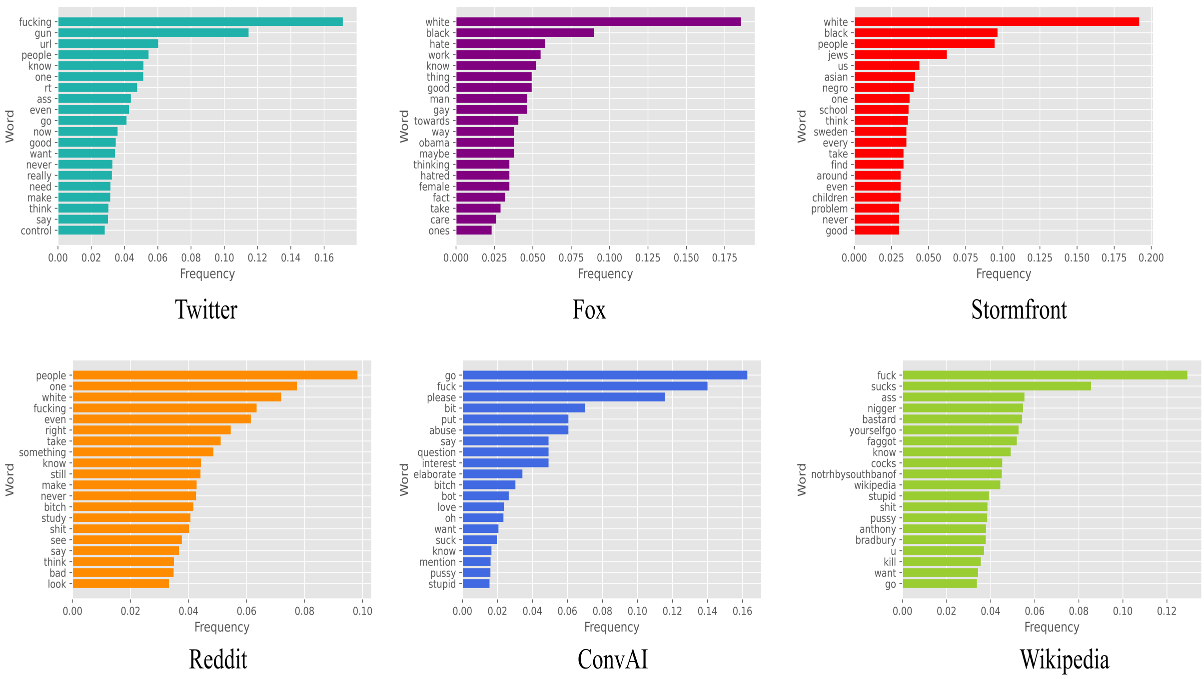

Figure 6: Top-20 normalized word frequency of abusive language for different platforms (ignoring *stopwords* and *non-alphabetic* characters).

Of these examples, 14,880 contain abusive (personal attack, aggressiveness, harassment) language.

F.2 twitter

We collect *twitter* dataset from a variety of sources. Waseem and Hovy (2016) annotate around 16k tweets that contain sexist/racist language. Initially, the authors bootstrap the corpus based on common slurs, then manually annotate the whole corpus to identify tweets that are offensive but do not contain any slur. Similarly, Davidson et al. (2017) crawled tweets with lexicon containing words and phrases identified by internet users as hate speech. Then crowdsourcing is performed to distinguish the category of hate, offensive, and normal tweets, resulting in around 25k annotated tweets. Jha and Mamidi (2017) crawled Twitter with the terms that generally exhibit positive sentiment but sexist in nature (e.g. ‘as good as a man’, ‘like a man’, ‘for a girl’). The authors also annotate tweets that are aggressively sexist. The final corpus contains around 10k tweets of implicit/explicit sexist and normal tweets. ElSherief et al. (2018) adopt multi-step data collection process that include collecting tweets based on lexicon, *hashtag*, and other existing works (Waseem and Hovy, 2016; Davidson et al., 2017). Then, crowdsourcing is applied to annotate targeted and untargeted hate speech. Founta et al. (2018) build an annotated corpus of 80k tweets with seven classes (offensive, abusive, hateful speech, aggressive, cyberbullying, spam, and normal). Mathur et al. (2018) annotate a corpus of around 3k tweets containing hate, abusive, and normal tweets. Basile et al. (2019) crawled 13k tweets containing abusive language against women and immigrants. The authors applied crowdsourcing to annotate if the tweets contain individual/ group hate speech or aggressiveness. Mandl et al. (2019) develop a corpus of 7k English examples with the category of hate, offensive, and profanity. Ousidhoum et al. (2019) build a corpus of multilingual and multi-aspect hate speech. The English corpus (5,647 tweets) covers a wide range of hate speech categories including the level of directness, hostility, targeted theme, and targeted group. Zampieri et al. (2019a) develop an offensive corpus of 14,100 tweets based on hierarchical modelings, such as whether a tweet is offensive/targeted, if it is targeted towards a group or individual.

Our final *twitter* dataset contains 132.815 examples of which 77,656 are abusive.

F.3 fb-yt

fb-yt represent both Facebook and Youtube platforms. We collect this dataset from Salminen et al. (2018). Salminen et al. (2018) crawled the comments from Facebook and Youtube videos and annotate them into hateful, non-hateful categories. The authors also subcategorize hateful comments into 21 classes including accusation, promoting violence, and humiliation.

F.4 stormfront

stormfront dataset is collected from de Gibert et al. (2018). The authors crawled around 10k examples from Stormfront and categorize them into hate/normal speech. The authors further investigate whether joining subsequent seemingly normal sentences result in hate speech. Our final dataset contains 1364 hateful speech from Stormfront.

F.5 fox

fox dataset represents user comments on the broadcasting platform Fox News. We collect this dataset from Gao and Huang (2017). The authors find that the hateful comments are more implicit and creative and such hateful comments detection requires context-dependency.

F.6 twi-fb

twi-fb dataset contains user posts from Twitter and Facebook. We collect this dataset from Mandl et al. (2019). The authors initially collect the corpus by crawling keywords and hashtags. Later, they annotate the corpus into targeted/untargeted hate speech, offense, and profane.

F.7 reddit

reddit dataset contains conversations from Reddit. Qian et al. (2019) compiled a list of toxic subreddit and crawled user conversations from those subreddits. Additionally, the authors provide hate speech intervention, where the goal is to automatically generate responses to intervene during online conversations that contain hate speech. The final dataset contains 2511 examples of hate/abusive speech.

F.8 convAI

Cercas Curry et al. (2021) collect *convAI* dataset from the user conversation with an AI assistant, CarbonBot, hosted on Facebook Messenger and a rule-based conversational agent, ELIZA. The authors categorize the dataset based on the severity and the type of abusiveness, directness, and target. We collected 853 examples from this dataset of which 128 are abusive speech.

F.9 hateCheck

hateCheck is a synthetically-generated dataset collected from Röttger et al. (2021). The authors develop 29 functionality through prior research and human interview and generate test case to evaluate test case for each of the functionalities. The dataset contains 2563 examples of hate speech.

F.10 gab

We collect *gab* dataset from Qian et al. (2019). Unlike other datasets, Qian et al. (2019) provide the full conversation which helps the models to understand the context. We collect 15,926 examples from the original corpus of which 15,270 are hate speech.

F.11 yt-reddit

yt-reddit dataset is collected from Mollas et al. (2020). The authors develop the dataset, namely, ETHOS sampling from Youtube and Reddit comments. The authors emphasize reducing any kinds of bias (e.g. gender) in the annotation process and annotate the dataset into various forms of targeted hate speech (e.g. origin, race, disability). We sample an equal number of hate and normal speech from this dataset.