Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2976–2988
Marseille, 20-25 June 2022

Cyberbullying Classifiers are Sensitive to Model-Agnostic Perturbations

Chris Emmery🌳,✋**, Ákos Kádár**💥**, Grzegorz Chrupała**🌳**, Walter Daelemans**✋
🌳CSAI, Tilburg University, ✋CLiPS, University of Antwerp, 💥Explosion
cmry@pm.me

Abstract

A limited amount of studies investigate the role of model-agnostic adversarial behavior in toxic content classification. As toxicity classifiers predominantly rely on lexical cues, (deliberately) creative and evolving language-use can be detrimental to the utility of current corpora and state-of-the-art models when they are deployed for content moderation. The less training data is available, the more vulnerable models might become. This study is, to our knowledge, the first to investigate the effect of adversarial behavior and augmentation for cyberbullying detection. We demonstrate that model-agnostic lexical substitutions significantly hurt classifier performance. Moreover, when these perturbed samples are used for augmentation, we show models become robust against word-level perturbations at a slight trade-off in overall task performance. Augmentations proposed in prior work on toxicity prove to be less effective. Our results underline the need for such evaluations in online harm areas with small corpora. The perturbed data, models, and code are available for reproduction at `https://github.com/cmry/augtox`.

Keywords: cyberbullying detection, data augmentation, lexical substitution

1. Introduction

Our online presence has simplified contact with our (in)direct network, and thereby drastically changed how, and with whom we interact. While online connections and self-disclosure are often socially beneficial (Valkenburg and Peter, 2007), the absence of physical interaction has numerous adverse effects: it greatly reduces social accountability in (anonymous) interactions, amplifies one's exposure to people with malicious intent, and through our frequent use of mobile devices, the invasiveness thereof (Mason, 2008). These factors accumulate to persistent online toxic behavior—the scale of which online platforms continue to struggle with from a technical, legal, and ethical perspective.

Online harm (Banko et al., 2020, provide a comprehensive taxonomy of this field) and—particularly for Natural Language Processing (NLP)—abusive language, are highly complex phenomena. Their study spreads across several subfields (detection of hate speech, toxic comments, offensive and abusive language, aggression, and cyberbullying), all with their unique problem sets and (almost exclusively English) corpora (Vidgen and Derczynski, 2021). Moreover, there are numerous open issues with these tasks, as highlighted in a range of critical studies (Emmery et al., 2019; Rosa et al., 2019; Swamy et al., 2019; Madukwe et al., 2020; Nakov et al., 2021, for example). Those open issues primarily pertain to the contextual, historical, and multi-modal nature of toxicity, the specificity of the data, and poor generalization across domains.

The current work focuses on one of these subproblems: the continuously evolving nature of *toxic content*. Apart from the disparate channels and media through which (young) users communicate, this development particularly applies to the related vocabulary: slang, hate speech, or general insults (e.g., *karen*, *simp*, *coofer*, and *covidiot*). Given the strong focus on lexical cues exhibited by state-of-the-art toxic content classifiers (Gehman et al., 2020), the existing corpora would have to be continuously expanded for models to retain their performance. This puts costly requirements on any system automatically moderating harmful content, while research in this domain still seems unconcerned with evaluating models in the wild (Nakov et al., 2021).

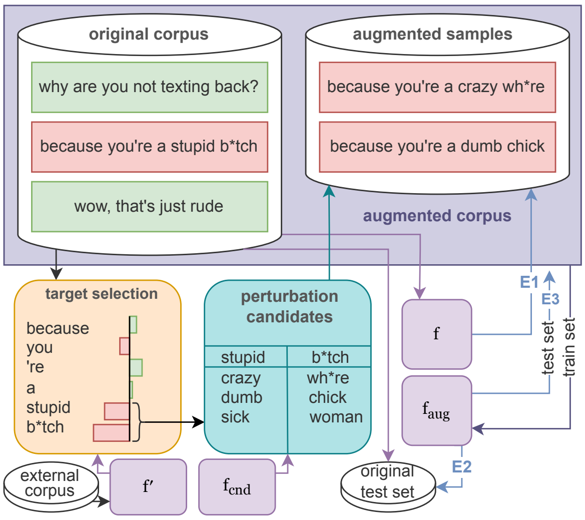

Figure 1: Schematic overview of the presented experiments (E1-3) for data augmentation of cyberbullying content via model-agnostic lexical substitutions.

The adversarial nature of toxicity exacerbates these issues further; similar to any security application, it is safe to assume malicious actors will try to (actively) subvert any form of moderation they are subjected to. Yet, while ample work has investigated systems toward mimicking such behavior (Hosseini et al., 2017; Ebrahimi et al., 2018; Li et al., 2019, for example, feature attacks against Google's Perspective API), work on toxic content detection rarely incorporates tests for robustness against adversarial attacks. More importantly, these attacks are commonly tailored to an existing toxicity classifier, whereas a human adversary would not have direct access to the models performing moderation. A realistic implementation of subversive human behavior would therefore require model-agnostic attacks.

Accordingly, the current research combines multiple ideas from previous work: we apply lexical substitution to an online harm subtask with small corpora (cyberbullying detection) to investigate how lexical variation (either natural, or adversarial) affects model performance and, by extension, evaluate the robustness of current state-of-the-art models. We do this in a model-agnostic fashion; an external classifier indicates which words might be relevant to substitute. Those words are perturbed through a variety of transformer-based models, after which we assess changes in the predictions of a target classifier (see Figure 1). The perturbations are *not* selected to be adversarial against the external or target classifier. We subsequently evaluate to what extent augmenting existing cyberbullying corpora improves classifier performance, robustness against word-level perturbations, and transferability across different substitution models. With this, we provide methods and language resources to test the robustness of cyberbullying classifiers against lexical variation in toxicity.

2. Lexical Substitution

We employ word-level or token-level perturbations (i.e., substitutions, see Table 1 for examples), which implies that for a given target word w_t in document $D = (w_0, w_1, \ldots, w_t, \ldots, w_n)$, we find a set of perturbation candidates C using substitute[1] classifier f' to exhaustively generate new samples D', any of which *potentially* produces an incorrect label for a target classifier f. However, the samples are not selected based on such label changes, which therefore does not make this an adversarial attack. We follow and improve upon the adversarial substitution framework[2] from Emmery et al. (2021), which in turn extends that of TextFooler (Jin et al., 2020)[3] with transformer-based perturbations.

2.1. Selecting Words to Perturb

Target words $T(D,\ f')$ are selected and ranked based on their contribution to the classification of a document. This importance, or omission score (Samek et al., 2017; Kádár et al., 2017, among others) is calculated by deleting a word at a given position D_t, denoted as $D_{\setminus t}$. The omission score is then $o_y(D) - o_y(D_{\setminus t})$, where o_y is the logit score of a substitute classifier f'. Intuitively, this would provide us with highly toxic words, or text parts related to bullying, which can be perturbed in some way.

2.2. Proposing Perturbation Candidates

As we intend to improve lexical variation, we focus on proposing synonyms as perturbation candidates. Zhou et al. (2019) condition BERT's masked language modeling on a given word by providing the original word its embedding to the masked position. They apply Dropout (Srivastava et al., 2014) as a surrogate mask, and show this to produce a top-k of potential synonyms. The predicted words at the Dropout masked position by some separate transformer model f_{cnd} are then our candidates $C(T, f_{\text{cnd}})$. To rank the candidates, they use a contextual similarity score:

$$\textsc{sim}\left(D, D'; t\right) = \sum_{i}^{n} \alpha_{i,t} \times \Lambda\left(\boldsymbol{h}\left(D_i\right), \boldsymbol{h}\left(D'_i\right)\right) \tag{1}$$

where: $\boldsymbol{h}\left(D_i\right)$ is the concatenation of f_{cnd} its last four layers for a given i^{th} token in document D, $D' = (w_0, \ldots, c_t, \ldots w_n)$ is the perturbed document D where target word w_t has been replaced with candidate c at the index of t, Λ is their cosine similarity, and $\alpha_{i,t}$ is the average self-attention score across all heads in all layers ranging from the i^{th} token to the t^{th} position in D. Finally, we sanitize the candidates: filtering single characters, plural and capitalized forms of the original words, sub-words, and sentence-level duplicates.

2.3. Handling Out-of-vocabulary Words

BERT's associated tokenizers break down unknown words into word-pieces (Wu et al., 2016), meaning there is no single embedding to apply Dropout to. Zhou et al. (2019) do not mention how they handle such cases; however, they are problematically common for our task (see Table 1). We therefore extend their method with a back-off method: if w_t is out-of-vocabulary (OOV),[4] we collapse the word-pieces into one, and zero the embedding at that position (which then acts as a mask).[5] Words other than w_t that are OOV remain word-pieces.

3. Experimental Set-up

We employ and compare this lexical substitution method to produce new positive instances. We evaluate if the perturbed documents hold up as adversarial samples, and if they can be used for data augmentation.

3.1. Data

For our corpora (all are English), we use two question-answering-style social networks that allow for anonymous posting: *Formspring* (Reynolds et al., 2011) and *Ask.fm* (Hee et al., 2015). The latter features multi-label

[1] Substitute does not refer to word substitution here, but a 'replacement' classifier. Specifically, one with an architecture and training data distinct from any target classifier f we use.

[2] `https://github.com/cmry/reap (ba8ee44)`

[3] `https://github.com/jind11/TextFooler`

[4] Note that the vocabulary of f' might include tokens not contained in the vocabulary of BERT.

[5] Alternative approaches, such as averaging and summing the token embeddings, did not provide better representations.

							TARGETS						
PROMPT	You	are	a	r[redacted]t[redacted]rded	dweeb	and	stupid	af	.	Go	f[redacted]ck	yourself	.
TOKENS	You	are	a	r[redacted] ##t[redacted]r ##d[redacted]d	d ##we ##eb	and	stupid	a ##f	.	Go	f[redacted]ck	yourself	.
#1	You	are	a	silly	baby	and	silly	af	.	Go	scr[redacted]w	yourself	.
#2	You	are	a	useless	teenager	and	dumb	af	.	Go	d[redacted]ck	yourself	.
#3	You	are	a	sick	b[redacted]tch	and	foolish	af	.	Go	sh[redacted]t	yourself	.
#4	You	are	a	crazy	dog	and	useless	af	.	Go	d[redacted]mn	yourself	.
#5	You	are	a	dumb	idiot	and	ignorant	af	.	Go	p[redacted]ss	yourself	.

Table 1: Lexical substitution example using Dropout BERT. Shows the (bowdlerized) initial PROMPT, which words are targeted for substitution (highlighted), their word-piece encoding (TOKENS), and the generated samples.

	🤬	🙂	TTR	AVG TOK/MSG
Ask.fm	5,001	89,404	.154	12 ($\sigma = 23$)
MySpace	426	1,627	.016	391 ($\sigma = 285$)
Twitter I	237	5,258	.154	14 ($\sigma = 8$)
Twitter II	281	4,654	.221	18 ($\sigma = 8$)
YouTube	417	3,045	.063	239 ($\sigma = 252$)
Formspring	1,025	11,742	.060	27 ($\sigma = 29$)

Table 2: Corpus statistics for cyberbullying data. Listed are the number of positive (🤬, bullying) and negative (🙂) instances, Type-Token Ratio (TTR), and the (rounded) average number of tokens per message (AVG TOK/MSG), and their standard deviation (σ).

annotation, but is binarized (any indication of bullying[6] is labeled positive) to be compatible with other corpora. These corpora are significantly larger than the rest, as their platforms are typically used by young adults, and notorious for their bullying content (Binns, 2013). Two long-form platforms can be found in *YouTube* (Dinakar et al., 2011) and *MySpace* (Bayzick et al., 2011), the latter of which has instances of ten posts. The smallest two are from *Twitter*, both collected using topical keywords (Xu et al., 2012; Bretschneider et al., 2014). The corpora's statistics can be found in Table 2.

3.2. Augmentation Models

The models were implemented using HuggingFace's `transformers` (Wolf et al., 2020) library.[7] Dependency versions can be found in our repository.

Target Word Selectors All experiments follow the same model-agnostic approach: target words are determined through substitute classifier f' (i.e., a distinct model trained on a different corpus than used in any other experiments). Generally, this is Gaussian Naive Bayes over tf·idf-weighted vectors, trained on *Formspring*. We additionally investigate a pre-trained version of BERT fine-tuned on the Jigsaw dataset (Hanu et al., 2020, `unitary/toxic-bert`) as a transformer-based alternative for f' (denoted by a +). While the task it has been fine-tuned on is slightly different, our assumption is that this model will have better representations and a larger vocabulary, which might make it more effective in choosing target words.

Substitutors We compare our implementation of Zhou et al. (2019)'s substitutions (here: Dropout BERT) against other methods for masked word prediction: BERT (Devlin et al., 2019) and BART (Lewis et al., 2020). Table 6 shows output examples of the various models. The probability of zeroing embedding dimensions in Dropout BERT is set to 0.2, as we empirically found that values around 0.3 often does not result in synonyms. We set the minimum required omission score to 0.005 for tokens to be considered for substitution, which yields 1-3 target words per document on average. The substitutions do not incorporate prior substitutions; they are done simultaneously—best candidates first—and exhaustively (i.e., while candidates for all slots are available) for a maximum of five samples.

For all BERT models, we use the pre-trained `bert-large-cased`, for BART we use `bart-large`.[8] We also report experiments where we use a fine-tuned toxicity version of pre-trained BERT (Caselli et al., 2021, `GroNLP/hateBERT`) for f_{cnd} in Dropout BERT (here referred to as Hate BERT, and when using a fine-tuned substitute classifier: Hate BERT+). The idea here is similar to that of using Dropout BERT+; domain-specific vocabularies will likely result in better and more varied substitutions. Hate BERT(+) uses a different BERT-based toxicity model than the '+' model for f', in order to keep these selections model-agnostic.

Baselines The substitution models are compared with two baselines adapted from related work. Both have shown to improve toxicity detection (Gehman et al., 2020; Quteineh et al., 2020; Yoo et al., 2021), but have (to our knowledge) not been applied for augmenting cyberbullying content. Firstly, we employ the common data augmentation baseline: Easy Data Augmentation (Wei and Zou, 2019, or EDA). EDA applies n of the following operations to an input text: synonym replacement using WordNet (Miller, 1995), random character insertions, swaps, and deletions. We set the number of augmentations made by EDA similar to that of our

[6] Includes self-defenses and assistants of the victim.

[7] `https://huggingface.co/transformers/`

[8] We empirically found the `bert-base` models to perform significantly worse at finding good synonyms.

PROMPT	okay and stop calling me jaky you c[redacted]ck
GPT-2	s[redacted]cker. you know it you f[redacted]cking p[redacted]ssy. you know you are an evil f[redacted]cking b[redacted]tch that only cares about getting her name in newspapers. i bet if you saw my face you wouldnt even believe i said ""oh yeah i think i can f[redacted]ck you."" ...

Table 3: Output by GPT-2, receiving an original instance as prompt. The generated text (up to 70 tokens) is subsequently used as an augmented instance.

	TRAIN		TEST	
	🤬	😶	🤬	😶
Merged	4,789	72,243	561	8,001
Augment Train	28,148	72,243	561	8,001
Augment Test	4,789	72,243	3,283	8,001

Table 4: Instance counts for the different splits used in our experiments. Augment Test is used for Experiment 1 (gauging the adversarial nature of our samples), Augment Train in Experiment 2 (data augmentation).

other models. Secondly, we employ fully unsupervised augmentation with GPT-2 (Radford et al., 2019, implemented in the pre-trained `gpt2-large`). We use the positive instances (i.e., documents containing cyberbullying) of each dataset as prompt, with a maximum input length of 30 tokens, and the generated output length to a maximum of 70, as we found that toxicity is prevalent in the first part of the generation (Gehman et al., 2020, made similar observations). Table 3 shows examples of the output, and the eventual divergence from toxicity.[9]

3.3. Classifiers

We follow recent state-of-the-art results (Elsafoury et al., 2021a) for our main classification model, and fine-tune all BERT-based models for 10 epochs with a batch size of 32 and a learning rate of $2e{-}5$, as suggested by Devlin et al. (2019). Accordingly, we set the maximum sequence length to 128, and insert a single linear layer after the pooled output. For the transformer experiments, we fine-tune incrementally: first on the original set, then on the augmented training set (including the original instances)—both using the same configuration (learning rate, batch size, etc.), except for running it for 2 epochs. This should offer performance advantages (Yang et al., 2019), as well as increase model stability.[10]

We compare BERT against a previously tried-and-tested (Emmery et al., 2019) 'simple' linear baseline: the Scikit-learn (Pedregosa et al., 2011) implementation of a Linear Support Vector Machine (Cortes and Vapnik, 1995; Fan et al., 2008, SVM) with binary Bag-of-Words (BoW) features, using hyperparameter ranges from Hee et al. (2018). Training of the SVM and BERT classifiers is done on a merged set of all the cyberbullying corpora in Table 2, except for *Formspring* (reserved for substitute classifier f')—always on the same 90% split, augmented data or no. The SVM is tuned via grid search and nested, stratified cross-validation (with ten inner and three outer folds, no shuffling, using 10% splits). The best settings (1-3-grams, class balancing, square hinge loss, and $C = 0.01$) are used in all experiments.

For both models, we also experiment with prepending a special token (Daumé III, 2007; Caswell et al., 2019, follow a similar approach) to the augmented instances (`<A>`). As per recommendations in Kumar et al. (2020), the token is not added to the vocabulary. These models are referred to as either f or f_{aug} in Figure 1, depending on if they were trained on augmented data. If not, we skip the 2 fine-tuning epochs for BERT.

3.4. Evaluation

To evaluate our classifiers on the main classification task, we use F_1-scores. The impact of the substitution models on classification performance is measured via a decrease in True Positive Ratio (TPR) between regular and substituted samples (i.e., how many previously positively classified samples classified as negative after perturbation). Note that in these experiments, f' is the same (either Naive Bayes or BERT); therefore, the substitute classifier always chooses the same target words to perturb. The amount of samples depends on the quality of the candidates the models propose.

TPR decrease by itself might also indicate an augmented instance is not toxic anymore; hence, to evaluate the semantic consistency of the samples produced by the various augmentation models, we calculate both METEOR (Banerjee and Lavie, 2005; Denkowski and Lavie, 2011) using the implementation from `nltk`[11], and BERTSCORE (Sellam et al., 2020) between the original sentences and their respective augmented samples. METEOR measures flexible uni-gram token overlap, and BERTSCORE transformer-based similarity with respect to the contextual sentence encoding.

3.5. Experiments

We run our substitution pipeline (visualized in Figure 1) on the positive instances X_{pos} of some given corpus (or the entire collection), using the different models discussed in Section 3.2 for f' and f_{cnd}. Per such configuration, this generates augmented samples X'_{pos} (up to five per original instance). These can either be classified as is, or mixed in with the original corpus, producing

[9] GPT-2 in particular tends to descend into literary content after too many tokens are generated. We also experimented with GPT-3's (Brown et al., 2020) `curie` from the OpenAI beta API (`https://beta.openai.com/`) but found systematically lower performance across all experiments compared to GPT-2. These results are therefore not included.

[10] We found that fine-tuning on the mixed set renders augmentation ineffective for all models we tested.

[11] `https://www.nltk.org/_modules/nltk/translate/meteor_score.html` (v3.5)

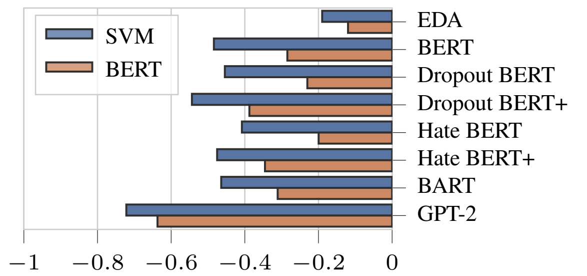

Figure 2: Decrease in True Positive Rate of the SVM and BERT classifiers after the respective substitution models have been applied (lower is more adversarial).

the augmented corpus X', with X'_{train}, and X'_{test} splits. Using this configuration, we run our three experiments:

Experiment 1 We gauge the *lexical variation* (and hence the 'adversarial' character) in our augmented samples via $f(X'_{\text{pos}})$. F_1-scores and TPR changes close to $f(X_{\text{pos}})$ imply the substitutions are similar to the original words. We confirm this meaning preservation through semantic consistency metrics for X'_{pos}.

Experiment 2 Here, we train via the *data augmentation* scheme discussed in Section 3.3; i.e., fine-tune for 2 epochs on $f(X'_{\text{train}})$. The resulting augmented classifier is referred to as f_{aug}, which we evaluate on the original X_{test}. An increase in F_1-score with respect to $f(X_{\text{test}})$ indicates the augmentation is a success.

Experiment 3 We measure *robustness* against perturbations, and *transferability* via $f_{\text{aug}}(X'_{\text{test}})$ by evaluating f_{aug} performance across different substitution models producing perturbed samples in X'_{test}; i.e., in a many-to-many evaluation. Any TPR increase implies augmentation improves robustness against perturbations. A total TPR higher than $f(X_{\text{test}})$ (Plain) does not necessarily increase the F_1-score (from Experiment 2). If this increase holds for multiple perturbation models, this implies the augmentations are transferable.

4. Results and Discussion

Here, we discuss the results of our three Experiments (Sections 4.1-4.3) and close with suggestions for future work. The main results can be found in Table 5 and 7.

4.1. The Effect of Lexical Variation

The results for this experiment can be found under the 'Samples' row in Table 5.

Classifier Performance It can be seen that unsupervised (prompt conditioning) samples (i.e., GPT-2) are the most difficult to classify. This is to be expected, as the generated output is not always toxic. However, it is arguably rather remarkable that a large amount of the generated sentences are labeled positive by the cyberbullying classifier. This confirms that (contextually more) harmful content is generated (as illustrated in Table 3), as also shown for toxicity detection by Gehman et al. (2020) and Ousidhoum et al. (2021). Moving on, we can see the fine-tuned target selector models ('+') show most 'adversarial' behavior, likely providing lexical diversity on words more important to the content classification. BART induces similar performance drops, but often inserts noise (see Table 6). The Dropout models seem to produce samples that are less diverse, but still show a solid .1 drop in F_1-score (17.67% on average). To emphasize, this decrease is based on *untargeted* substitutions; i.e., without selecting the substituted words as to change the predictions of either f or f'.

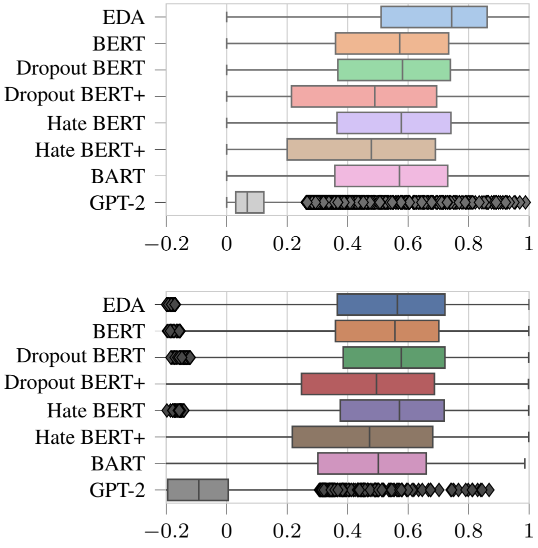

Figure 3: Semantic consistency metrics (higher is better): METEOR (upper) and BERTSCORE (lower) scores per model—evaluating the augmented samples (positive only) with the original documents as a reference.

Adversarial Samples Additional analyses can be found in Figure 2. We observe the same patterns per substitution model[12] as in Table 5, with the BERT classifier showing to suffer around 20% less in TPR compared to the SVM. This difference can partly be explained by the substitution models sampling from the same model as we fine-tuned for the classification task (`bert-large-cased`). As can be observed in this Figure, the difference is smaller when this is not the case ('+' models). This experiment not only underlines the strong focus on lexical cues[13] from linear classifiers, but also that transformer models are not immune to lexical variation—even when candidates are sampled from their own language model. This provides further evidence in line with research from Elsafoury et al. (2021a), and Elsafoury et al. (2021b) (see Section 5).

Semantic Consistency of Samples Here, we compared X'_{pos} with X_{pos} as a reference. The results for METEOR and BERTSCORE of these pairs can be found in Figure 3. Generally, these confirm the trend from the

[12] To equalize the length, we matched the amount of non-augmented test set instances for this experiment.

[13] Which raises its own issues; see e.g., and Zhou et al. (2021), for work on bias and debiasing.

X_{train}	Plain	EDA	BERT	Dropout BERT	Dropout BERT+	Hate BERT	Hate BERT+	BART	GPT-2
		$f(X'_{\text{pos}})$							
Merged	.614 .009	.598 .012	.458 .002	.491 .005	.439 .002	.520 .005	.478 .004	.436 .007	.334 .007
		$f_{\text{aug}}(X_{\text{test}})$							
Merged	.563 .014	.553 .007	.538 .014	.546 .012	.523 .004	**.562** .007	.535 .009	.536 .013	.550 .017
Ask.fm	.621 .011	.591 .011	.581 .016	.597 .015	.574 .003	**.611** .007	.592 .007	.587 .015	.601 .009
Myspace	.436 .093	.476 .021	.403 .058	.376 .161	.351 .040	.414 .042	.370 .065	.387 .043	**.496** .037
Twitter I	.596 .046	**.630** .016	.592 .059	.531 .048	.594 .030	.617 .053	.533 .051	.591 .027	.583 .097
Twitter II	.308 .059	.290 .038	.297 .043	**.329** .034	.257 .040	.313 .048	.268 .027	.277 .077	.295 .048
YouTube	.150 .033	**.226** .057	.180 .010	.201 .066	.144 .045	.173 .056	.152 .009	.152 .055	.207 .061

Table 5: BERT-based cyberbullying classification scores (F_1) for Experiments 1 (under $f(X'_{\text{pos}})$) and 2 (under $f_{\text{aug}}(X_{\text{test}})$). Classifiers are trained and tested on the indicated corpus (from Section 3.1), 'Merged' is their combination. The other columns indicate, respectively: no substitutions (Plain), EDA, BERT-based models (where '+' indicates f' uses BERT rather than an SVM, and 'Hate' that f_{cnd} is pre-trained), BART, and GPT-2. Highlighted cells indicate that non-augmented performance was highest, bold indicates the highest performance per augmentation model. Standard deviation (small script) is reported over five runs with different seeds.

previous two parts of the experiment: models that have higher semantic consistency have less effect on classification performance. A clear difference in METEOR can be observed between EDA and the other models. This is likely due to both the metric and model using WordNet, resulting in bias in favor of EDA. GPT-2 is a strong outlier, as it generates new data.

The semantic consistency scores seem comparable, and at times slightly better, than previous lexical substitution work (Shetty et al., 2018; Emmery et al., 2018; Mathai et al., 2020, although these are all explicitly adversarial). We noticed that regarding the samples themselves, the transformer-based models often noticeably break down in terms of semantic preservation for the lower ranked candidates (see Table 6). For the models that do not use soft semantic constraints (such as Dropout), we already find antonyms, and generally ungrammatical and incoherent sentences within the top 5 candidates. Interestingly, at the same time, BERTSCORE assigns comparable scores to antonyms as it does for (intuitively) better substitution candidates. Given these observations, if mimicking adversarial behavior is to be given more weight, one should consider limiting the number of augmented samples, and tuning the omission score cut-off might prove to be worthwhile.

4.2. Substitutions for Data Augmentation

Given the strong performance effect the substitution models had on our classifiers, it seems plausible that they might prove to produce effective samples for augmentation purposes. Looking at the lower portion of Table 5; however, we can see that augmentation does not improve performance on the two biggest sets (Merged, and Ask.fm). Dropout BERT seems to improve performance for one of the smaller sets (Twitter II), but overall, EDA is generally a close contender with, if not more effective, than all of the more 'advanced' models. Interestingly, the transformer-based models seem to yield sizable improvements on the Myspace set, with GPT-2 increasing it most. The latter might be attributed to the low Type-Token Ratio in this set (see Table 2).

	#					TARGETS				BSC
		I	don't	get	why	people	like	u	.	
BE	1	I	don't	get	why	they	want	u	.	.809
	5	I	don't	get	why	boys	hate	u	.	.807
Dr BE	1	I	don't	get	why	everyone	want	u	.	.862
	5	I	don't	get	why	you	love	u	.	.806
BA	1	I	don't	get	why	you	are	u	.	.700
	5	I	don't	get	why	so	have	u	.	.634

Table 6: Augmentations including the top 1 and 5 candidates from BERT (BE), Dropout BERT (Dr BE), and BART (BA), and the BERTSCORE (BSC) using the original text as reference, showing quality degradation (not well reflected in the metric) when sample size increases. BERT suggests antonyms, BART fails semantically.

Interpretation Generally, none of these methods (baselines, or substitution-based augmentation) seem to yield the same performance improvement as observed in toxicity work (Ibrahim et al., 2018; Jungiewicz and Smywinski-Pohl, 2019). However, note that, as we were interested in simulating potentially adversarial behavior, we conducted model-agnostic augmentation (that is, given an unknown attacker, or noise). Hence, while we might employ these models in an explicit adversarial training scheme to directly improve model performance, this would require extensive transferability evaluations—typically requiring larger, higher quality datasets—and only satisfy one dimension (data). Given this, we argue an improvement in classifying the augmented sets, as in Experiment 1 (Section 4.1), is more significant.

	INITIAL TPR	f_{aug}							
Plain	.537	EDA	BERT	Dropout BERT	Dropout BERT+	Hate BERT	Hate BERT+	BART	GPT-2
$X'_{\text{test}} \downarrow$		Δ TPR							
EDA	.498	.270	.106	.104	.108	.101	.107	**.114**	.037
BERT	.390	-.033	.195	.144	.110	.128	.092	**.149**	.085
Dropout BERT	.421	-.017	**.183**	.183	.143	.143	.111	.152	.086
Dropout BERT+	.362	.015	.228	.236	.301	.177	**.238**	.191	.088
Hate BERT	.444	-.020	**.170**	.148	.104	.162	.123	.143	.073
Hate BERT+	.394	.034	.188	.174	**.238**	.215	.262	.183	.065
BART	.378	-.010	**.200**	.157	.127	.142	.115	.243	.079
GPT-2	.303	-.098	.031	.003	-.020	.003	-.025	**.048**	.597
MEAN	.399	-.114	**.158**	.138	.116	.111	.130	.140	.073

Table 7: Transferability and robustness of various f_{aug} models on various augmented test samples X'_{test}. Shown is the original True Positive Rate (TPR, under INITIAL TRP) for f (Plain), and the change (Δ) in TPR of $f_{\text{aug}}(X'_{\text{test}})$ with respect to $f(X'_{\text{test}})$. Positive Δ TRP shows robustness to perturbations after augmentation, negative the opposite. The classifiers most robust against same-model perturbations are highlighted, in bold the next-best. At the bottom, MEAN (per augmented classifier) TPR is shown, *excluding* performance on the same model (to reflect transferability).

4.3. Augmentation for Robustness

The results of Experiment 3 can be found in Table 7. We report TPR changes for $f(X_{\text{test}})$ (under initial TPR), and $f_{\text{aug}}(X'_{\text{test}})$ per setting to create f_{aug} and X' respectively.

Robustness Generally, it can be observed that (unsurprisingly) the augmented classifiers increase TPR most when the substitutions come from the same type of model. Hence, the 'second-best' TPR increases are more interesting. It can be observed that BERT and BART show strongest TPR improvements on three sets respectively, followed by the '+' models with one set respectively. This is quite a remarkable, contrasting result to Experiments 1 and 2, although it aligns with the observations from the semantic consistency scores that more conservative models are less effective augmenters. Hence, it seems that substitutions that are *more* diverse, and distant from the original instances provide better robustness against perturbations. While their output might be less semantically consistent, this is generally not a relevant criterion when one is only interested in improving task robustness.

Transferability Systematic performance gain across all substitution models (i.e., transferability) is the final indicator of augmentation utility. First, it must be noted that the TPR differences between 'same-model' and distinct model pairs are smaller for the transformer-based models (.026 on average) than EDA (.156). Lexical substitution using out-of-the-box BERT—in addition to high robustness—also achieves the highest transferability (mean .158) across substituted sets.

Performance Trade-Off The f_{aug} results from this experiment should be contextualized against the performance trade-off in F_1-score from Experiment 2. Using the information in Table 7, it can be inferred that the best performing model, BERT, actually improves absolute TPR on average; if we add its .158 mean TPR increase to the .399 average (= .557) this exceeds the .537 non-augmented TPR. However, as we showed in Experiment 2 (Table 5), this does not improve overall task performance; rather, it decreases performance. For BERT, the F_1-score slightly drops (.025-.030) on all sets. Hence, this is not a silver bullet, and such trade-offs should be considered when deploying these augmentation models to improve robustness against lexical variation.

Limitations and Future Work A substantial hurdle toward deploying the presented models for augmentation purposes is time. Upsampling the positive instances shown in Table 4 (5,350 total) with the transformer-based models takes 2-3 hours per model on a single NVIDIA Titan X (Pascal).[14] This impacts the amount of parameters that can be tweaked in reasonable time when using this architecture (such as omission score cut-offs, cosine similarity when ranking, dropout values, etc. which we all set empirically). Such computational demand is acceptable for smaller datasets like ours, and the augmentations can be run 'offline' (i.e., one time only), but these limitations should certainly be taken into account when scaling is among one's desiderata. Hence, recent work on decreasing the amount of queries for related models (Chauhan et al., 2021) is particularly relevant for future work. Additionally, there is a myriad of components the base architecture we presented here could be improved with. Most are discussed in Emmery et al. (2021); however, some new work is specifically of interest to data augmentation, such as improving the substitutions using beam search (Zhao et al., 2021, as opposed to the simultaneous rollout we used in the current work). More broadly, adversarial training (Si et

[14] Training the BERT classifiers takes up to 31 hours, augmentation 40 minutes, predictions on the test set 5 minutes.

al., 2021; Pan et al., 2021), implementing more robust stylometric features (Markov et al., 2021), or model-based weightings of the augmentation models could be explored; e.g., by selecting instances with a generation model in the loop (Anaby-Tavor et al., 2020). This could be a particularly worthwhile option when focusing on conversation scopes, rather than message-level cyberbullying content (Emmery et al., 2019).

Finally, this work focused specifically on cyberbullying corpora—a classification task which is generally (though equivocally) framed to extend beyond mere toxic content, and for which data is generally scarce. Although not in the scope of the presented work, our methods might be implemented in future work to further (critically) explicate the role of toxicity in this classification task, and thereby assist in curation of corpora, or contrast sets (Gardner et al., 2020; Li et al., 2020), that are more representative of the theoretical underpinnings of the concept of cyberbullying.

5. Related Work

Our work combines multiple sizeable—to the extent that they respectively produced several surveys (Fortuna and Nunes, 2018; Gunasekara and Nejadgholi, 2018; Mishra et al., 2019; Banko et al., 2020; Madukwe et al., 2020; Muneer and Fati, 2020; Salawu et al., 2020; Jahan and Oussalah, 2021; Mladenovic et al., 2021)—areas of research; hence, we will provide a concise overview of the work directly related to our experimental setup.

For all tasks, the issue of generalization seems a particularly popular subject of study: for cyberbullying, Emmery et al. (2019), and Larochelle and Khoury (2020), conclude there is little consensus in labeling practices, overlap between datasets, and that a combination of all datasets seems to transfer performance best. For hate speech, Salminen et al. (2020), and Fortuna et al. (2021), draw similar conclusions, showing that general forms of harm (e.g., toxic, offensive) generalize better than specific ones, such as hate speech. Finally, Nejadgholi and Kiritchenko (2020) provide unsupervised suggestions to address topic bias in data curation, potentially improving generalization. We draw from these works through cross-domain experiments on individual and combined corpora for cyberbullying, as well as pre-training on more general subtasks such as toxicity.

Recent cyberbullying work (Reynolds et al., 2011; Xu et al., 2012; Nitta et al., 2013; Bretschneider et al., 2014; Dadvar et al., 2014; Van Hee et al., 2015, e.g., are seminal work) has primarily focused on deploying Transformer-based models (Vaswani et al., 2017); by and large fine-tuning (Swamy et al., 2019; Paul and Saha, 2020; Gencoglu, 2021, e.g.), or re-training (Caselli et al., 2020) BERT. It is worth noting that Elsafoury et al. (2021a; Elsafoury et al. (2021b) show that although fine-tuning BERT achieves state-of-the-art performance in classification, its attention scores do not correlate with cyberbullying features, and they expect generalization of such models to be subpar. In our experiments, we employ similar domain-specific fine-tuned BERT models, and gauge generalization, sensitivity to perturbations, and the effects of augmentation to potentially improve the former.

Adversarial attacks on text (Zhang et al., 2020; Roth et al., 2021, e.g., provide broader surveys) can roughly be divided in character-level and word-level. The former relates to purposefully misspelling or otherwise symbolically replacing text (e.g., *fvk you*, *@ssh*l3*) to subvert algorithms (Eger et al., 2019; Kurita et al., 2019). Wu et al. (2018) show such attacks on toxic content can be effectively deciphered. Word-level attacks are arguably straight-forward for humans, but significantly more challenging to automate—requiring preservation of toxicity; i.e, the semantics of the sentence. Previous work has investigated the effect of minimal edits on high-impact toxicity words, replacing them with harmless variants (Hosseini et al., 2017; Brassard-Gourdeau and Khoury, 2019). Our current work is similar to that of Tapia-Téllez and Escalante (2020), and closest to that of Guzman-Silverio et al. (2020), who apply simple synonym replacement using EDA, as well as adversarial token substitutions—the latter using TextFooler on misclassified instances. We extend BERT-based lexical substitution (Zhou et al., 2019) for model-agnostic perturbations, and data augmentation.

Finally, regarding data augmentation for online harms (Bayer et al., 2021; Feng et al., 2021, among others, provide more general-purpose overviews for various natural language data), toxicity work partly overlaps with work on adversarial attacks on text; for example, the synonym replacement from Ibrahim et al. (2018), and Jungiewicz and Smywinski-Pohl (2019), which are distinctly either unsupervised, or semi-supervised. Another such example can be found in Rosenthal et al. (2021) employed democratic co-training to collect a large corpus of toxic tweets, and Gehman et al. (2020) find triggers that produce toxic content, querying GPT-like models (Radford et al., 2018). Fully unsupervised augmentation has also been employed in Quteineh et al. (2020), and Yoo et al. (2021). In our experiments, we use a pipeline of models for lexical substitution, and compare it to GPT generations (Radford et al., 2019; Brown et al., 2020).

6. Conclusion

In this work, we employed model-agnostic, transformer-based lexical substitutions to the task of cyberbullying classification. We show these perturbations significantly decrease classifier performance. Augmenting them using perturbed instances as new samples slightly trades off task performance with improved robustness against lexical variation. Future work should further investigate the use of these models to simulate and mitigate the effect of adversarial behavior in content moderation.

7. Acknowledgments

Our research strongly relied on openly available resources. We thank all whose work we could use.

8. Bibliographical References

Anaby-Tavor, A., Carmeli, B., Goldbraich, E., Kantor, A., Kour, G., Shlomov, S., Tepper, N., and Zwerdling, N. (2020). Do not have enough data? Deep learning to the rescue! In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7383–7390. AAAI Press.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Banko, M., MacKeen, B., and Ray, L. (2020). A unified taxonomy of harmful content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 125–137, Online, November. Association for Computational Linguistics.

Bayer, M., Kaufhold, M., and Reuter, C. (2021). A survey on data augmentation for text classification. *CoRR*, abs/2107.03158.

Bayzick, J., Kontostathis, A., and Edwards, L. (2011). Detecting the presence of cyberbullying using computer software. In *Proceedings of the ACM WebSci Conference, Koblenz, Germany, 2011.*

Binns, A. (2013). Facebook's ugly sisters: Anonymity and abuse on formspring and ask.fm. *Media Education Research Journal*, 4:27–42, 07.

Brassard-Gourdeau, E. and Khoury, R. (2019). Subversive toxicity detection using sentiment information. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 1–10, Florence, Italy, August. Association for Computational Linguistics.

Bretschneider, U., Wöhner, T., and Peters, R. (2014). Detecting online harassment in social networks. In Michael D. Myers et al., editors, *Proceedings of the International Conference on Information Systems - Building a Better World through Information Systems, ICIS 2014, Auckland, New Zealand, December 14-17, 2014.* Association for Information Systems.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Hugo Larochelle, et al., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.*

Caselli, T., Basile, V., Mitrovic, J., and Granitzer, M. (2020). Hatebert: Retraining BERT for abusive language detection in english. *CoRR*, abs/2010.12472.

Caselli, T., Basile, V., Mitrović, J., and Granitzer, M. (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, August. Association for Computational Linguistics.

Caswell, I., Chelba, C., and Grangier, D. (2019). Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy, August. Association for Computational Linguistics.

Chauhan, J., Bhukar, K., and Kaul, M. (2021). Target model agnostic adversarial attacks with query budgets on language understanding models. *CoRR*, abs/2106.07047.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.*, 20(3):273–297.

Dadvar, M., Trieschnigg, D., and de Jong, F. (2014). Experts and machines against bullies: A hybrid approach to detect cyberbullies. In Marina Sokolova et al., editors, *Advances in Artificial Intelligence - 27th Canadian Conference on Artificial Intelligence, Canadian AI 2014, Montréal, QC, Canada, May 6-9, 2014. Proceedings*, volume 8436 of *Lecture Notes in Computer Science*, pages 275–281. Springer.

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

Denkowski, M. and Lavie, A. (2011). Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Edinburgh, Scotland, July. Association for Computational Linguistics.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, et al., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Dinakar, K., Reichart, R., and Lieberman, H. (2011). Modeling the detection of textual cyberbullying. In *The Social Mobile Web, Papers from the 2011 ICWSM Workshop, Barcelona, Catalonia, Spain, July 21, 2011*, volume WS-11-02 of *AAAI Workshops*. AAAI.

Ebrahimi, J., Rao, A., Lowd, D., and Dou, D. (2018). HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Lin-*

guistics (Volume 2: Short Papers), pages 31–36, Melbourne, Australia, July. Association for Computational Linguistics.

Eger, S., Şahin, G. G., Rücklé, A., Lee, J.-U., Schulz, C., Mesgar, M., Swarnkar, K., Simpson, E., and Gurevych, I. (2019). Text processing like humans do: Visually attacking and shielding NLP systems. 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 1634–1647, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Elsafoury, F., Katsigiannis, S., Pervez, Z., and Ramzan, N. (2021a). When the timeline meets the pipeline: A survey on automated cyberbullying detection. *IEEE Access*, 9:103541–103563.

Elsafoury, F., Katsigiannis, S., Wilson, S. R., and Ramzan, N. (2021b). Does BERT pay attention to cyberbullying? In Fernando Diaz, et al., editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 1900–1904. ACM.

Emmery, C., Manjavacas, E. A., and Chrupala, G. (2018). Style obfuscation by invariance. In Emily M. Bender, et al., editors, *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 984–996. Association for Computational Linguistics.

Emmery, C., Verhoeven, B., Pauw, G. D., Jacobs, G., Hee, C. V., Lefever, E., Desmet, B., Hoste, V., and Daelemans, W. (2019). Current limitations in cyberbullying detection: on evaluation criteria, reproducibility, and data scarcity. *CoRR*, abs/1910.11922.

Emmery, C., Kádár, Á., and Chrupała, G. (2021). Adversarial stylometry in the wild: Transferable lexical substitution attacks on author profiling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2388–2402, Online, April. Association for Computational Linguistics.

Fan, R., Chang, K., Hsieh, C., Wang, X., and Lin, C. (2008). LIBLINEAR: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874.

Feng, S., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., and Hovy, E. (2021). A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online, August. Association for Computational Linguistics.

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4):85:1–85:30.

Fortuna, P., Soler Company, J., and Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Inf. Process. Manag.*, 58(3):102524.

Gardner, M., Artzi, Y., Basmova, V., Berant, J., Bogin, B., Chen, S., Dasigi, P., Dua, D., Elazar, Y., Gottumukkala, A., Gupta, N., Hajishirzi, H., Ilharco, G., Khashabi, D., Lin, K., Liu, J., Liu, N. F., Mulcaire, P., Ning, Q., Singh, S., Smith, N. A., Subramanian, S., Tsarfaty, R., Wallace, E., Zhang, A., and Zhou, B. (2020). Evaluating NLP models via contrast sets. *CoRR*, abs/2004.02709.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online, November. Association for Computational Linguistics.

Gencoglu, O. (2021). Cyberbullying detection with fairness constraints. *IEEE Internet Comput.*, 25(1):20–29.

Gunasekara, I. and Nejadgholi, I. (2018). A review of standard text classification practices for multi-label toxicity identification of online content. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 21–25, Brussels, Belgium, October. Association for Computational Linguistics.

Guzman-Silverio, M., Balderas-Paredes, Á., and López-Monroy, A. P. (2020). Transformers and data augmentation for aggressiveness detection in mexican spanish. In Miguel Ángel García Cumbreras, et al., editors, *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, September 23th, 2020*, volume 2664 of *CEUR Workshop Proceedings*, pages 293–302. CEUR-WS.org.

Hanu, L., Unitary, and team. (2020). Detoxify. GitHub.

Hee, C. V., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., Pauw, G. D., Daelemans, W., and Hoste, V. (2015). Detection and fine-grained classification of cyberbullying events. In Galia Angelova, et al., editors, *Recent Advances in Natural Language Processing, RANLP 2015, 7-9 September, 2015, Hissar, Bulgaria*, pages 672–680. RANLP 2015 Organising Committee / ACL.

Hee, C. V., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., Pauw, G. D., Daelemans, W., and Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *CoRR*, abs/1801.05617.

Hosseini, H., Kannan, S., Zhang, B., and Poovendran, R. (2017). Deceiving google's perspective API built for detecting toxic comments. *CoRR*, abs/1702.08138.

Ibrahim, M., Torki, M., and El-Makky, N. M. (2018). Imbalanced toxic comments classification using data augmentation and deep learning. In M. Arif Wani, et al., editors, *17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018,* *Orlando, FL, USA, December 17-20, 2018*, pages 875–878. IEEE.

Jahan, M. S. and Oussalah, M. (2021). A systematic review of hate speech automatic detection using natural language processing. *CoRR*, abs/2106.00742.

Jin, D., Jin, Z., Zhou, J. T., and Szolovits, P. (2020). Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8018–8025. AAAI Press.

Jungiewicz, M. and Smywinski-Pohl, A. (2019). Towards textual data augmentation for neural networks: synonyms and maximum loss. *Computer Science*, 20(1), Mar.

Kádár, Á., Chrupała, G., and Alishahi, A. (2017). Representation of linguistic form and function in recurrent neural networks. *Comput. Linguistics*, 43(4).

Kumar, V., Choudhary, A., and Cho, E. (2020). Data augmentation using pre-trained transformer models. *CoRR*, abs/2003.02245.

Kurita, K., Belova, A., and Anastasopoulos, A. (2019). Towards robust toxic content classification. *CoRR*, abs/1912.06872.

Larochelle, M. and Khoury, R. (2020). Generalisation of cyberbullying detection. In Martin Atzmüller, et al., editors, *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2020, The Hague, Netherlands, December 7-10, 2020*, pages 296–300. IEEE.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July. Association for Computational Linguistics.

Li, J., Ji, S., Du, T., Li, B., and Wang, T. (2019). Textbugger: Generating adversarial text against real-world applications. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society.

Li, C., Shengshuo, L., Liu, Z., Wu, X., Zhou, X., and Steinert-Threlkeld, S. (2020). Linguistically-informed transformations (LIT): A method for automatically generating contrast sets. In Afra Alishahi, et al., editors, *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2020, Online, November 2020*, pages 126–135. Association for Computational Linguistics.

Madukwe, K., Gao, X., and Xue, B. (2020). In data we trust: A critical analysis of hate speech detection datasets. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 150–161, Online, November. Association for Computational Linguistics.

Markov, I., Ljubešić, N., Fišer, D., and Daelemans, W. (2021). Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–159, Online, April. Association for Computational Linguistics.

Mason, K. L. (2008). Cyberbullying: A preliminary assessment for school personnel. *Psychology in the Schools*, 45(4):323–348.

Mathai, A., Khare, S., Tamilselvam, S., and Mani, S. (2020). Adversarial black-box attacks on text classifiers using multi-objective genetic optimization guided by deep networks. *CoRR*, abs/2011.03901.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Mishra, P., Yannakoudakis, H., and Shutova, E. (2019). Tackling online abuse: A survey of automated abuse detection methods. *CoRR*, abs/1908.06024.

Mladenovic, M., Osmjanski, V., and Vujicic Stankovic, S. (2021). Cyber-aggression, cyberbullying, and cyber-grooming: A survey and research challenges. *ACM Comput. Surv.*, 54(1):1:1–1:42.

Muneer, A. and Fati, S. M. (2020). A comparative analysis of machine learning techniques for cyberbullying detection on twitter. *Future Internet*, 12(11):187.

Nakov, P., Nayak, V., Dent, K., Bhatawdekar, A., Sarwar, S. M., Hardalov, M., Dinkov, Y., Zlatkova, D., Bouchard, G., and Augenstein, I. (2021). Detecting abusive language on online platforms: A critical analysis. *CoRR*, abs/2103.00153.

Nejadgholi, I. and Kiritchenko, S. (2020). On cross-dataset generalization in automatic detection of online abuse. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 173–183, Online, November. Association for Computational Linguistics.

Nitta, T., Masui, F., Ptaszynski, M., Kimura, Y., Rzepka, R., and Araki, K. (2013). Detecting cyberbullying entries on informal school websites based on category relevance maximization. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 579–586, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Ousidhoum, N., Zhao, X., Fang, T., Song, Y., and Yeung, D. (2021). Probing toxic content in large pre-trained language models. In Chengqing Zong, et al., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4262–4274. Association for Computational Linguistics.

Pan, L., Hang, C., Sil, A., Potdar, S., and Yu, M. (2021). Improved text classification via contrastive adversarial training. *CoRR*, abs/2107.10137.

Paul, S. and Saha, S. (2020). CyberBERT: BERT for cyberbullying identification. *Multimedia Systems*, November. 00000.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830.

Quteineh, H., Samothrakis, S., and Sutcliffe, R. (2020). Textual data augmentation for efficient active learning on tiny datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7400–7410, Online, November. Association for Computational Linguistics.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI blog*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Reynolds, K., Kontostathis, A., and Edwards, L. (2011). Using machine learning to detect cyberbullying. In Xue-wen Chen, et al., editors, *10th International Conference on Machine Learning and Applications and Workshops, ICMLA 2011, Honolulu, Hawaii, USA, December 18-21, 2011. Volume 2: Special Sessions and Workshop*, pages 241–244. IEEE Computer Society.

Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P., Carvalho, J., Oliveira, S., Coheur, L., Paulino, P., Veiga Simão, A., and Trancoso, I. (2019). Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93:333–345.

Rosenthal, S., Atanasova, P., Karadzhov, G., Zampieri, M., and Nakov, P. (2021). SOLID: A large-scale semi-supervised dataset for offensive language identification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 915–928, Online, August. Association for Computational Linguistics.

Roth, T., Gao, Y., Abuadbba, A., Nepal, S., and Liu, W. (2021). Token-modification adversarial attacks for natural language processing: A survey. *CoRR*, abs/2103.00676.

Salawu, S., He, Y., and Lumsden, J. (2020). Approaches to automated detection of cyberbullying: A survey. *IEEE Trans. Affect. Comput.*, 11(1):3–24.

Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S., Almerekhi, H., and Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Hum. centric Comput. Inf. Sci.*, 10:1.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Müller, K. (2017). Evaluating the visualization of what a deep neural network has learned. *IEEE Trans. Neural Networks Learn. Syst.*, 28(11):2660–2673.

Sellam, T., Das, D., and Parikh, A. (2020). BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July. Association for Computational Linguistics.

Shetty, R., Schiele, B., and Fritz, M. (2018). A4NT: author attribute anonymity by adversarial training of neural machine translation. In William Enck et al., editors, *27th USENIX Security Symposium, USENIX Security 2018, Baltimore, MD, USA, August 15-17, 2018*, pages 1633–1650. USENIX Association.

Si, C., Zhang, Z., Qi, F., Liu, Z., Wang, Y., Liu, Q., and Sun, M. (2021). Better robustness by more coverage: Adversarial and mixup data augmentation for robust finetuning. In Chengqing Zong, et al., editors, *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1569–1576. Association for Computational Linguistics.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Swamy, S. D., Jamatia, A., and Gambäck, B. (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, November. Association for Computational Linguistics.

Tapia-Téllez, J. M. and Escalante, H. J. (2020). Data augmentation with transformers for text classification. In Lourdes Martínez-Villaseñor, et al., editors, *Advances in Computational Intelligence - 19th Mexican International Conference on Artificial Intelligence, MICAI 2020, Mexico City, Mexico, October 12-17, 2020, Proceedings, Part II*, volume 12469 of *Lecture Notes in Computer Science*, pages 247–259. Springer.

Valkenburg, P. M. and Peter, J. (2007). Preadolescents' and adolescents' online communication and their closeness to friends. *Developmental Psychology*, 43(2):267–277, March.

Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., Daelemans, W., and Hoste, V. (2015). Detection and fine-grained classification of cyberbullying events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 672–680, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Isabelle Guyon, et al., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural*

Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008.

Vidgen, B. and Derczynski, L. (2021). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):1–32, 12.

Wei, J. and Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China, November. Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Wu, Z., Kambhatla, N., and Sarkar, A. (2018). Decipherment for adversarial offensive language detection. In Darja Fiser, et al., editors, *Proceedings of the 2nd Workshop on Abusive Language Online, ALW@EMNLP 2018, Brussels, Belgium, October 31, 2018*, pages 149–159. Association for Computational Linguistics.

Xu, J.-M., Jun, K.-S., Zhu, X., and Bellmore, A. (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, June. Association for Computational Linguistics.

Yang, W., Xie, Y., Tan, L., Xiong, K., Li, M., and Lin, J. (2019). Data augmentation for BERT finetuning in open-domain question answering. *CoRR*, abs/1904.06652.

Yoo, K. M., Park, D., Kang, J., Lee, S., and Park, W. (2021). Gpt3mix: Leveraging large-scale language models for text augmentation. *CoRR*, abs/2104.08826.

Zhang, W. E., Sheng, Q. Z., Alhazmi, A. A. F., and Li, C. (2020). Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Trans. Intell. Syst. Technol.*, 11(3):24:1–24:41.

Zhao, T., Ge, Z., Hu, H., and Shi, D. (2021). Generating natural language adversarial examples through an improved beam search algorithm. *CoRR*, abs/2110.08036.

Zhou, W., Ge, T., Xu, K., Wei, F., and Zhou, M. (2019). BERT-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373, Florence, Italy, July. Association for Computational Linguistics.

Zhou, X., Sap, M., Swayamdipta, S., Choi, Y., and Smith, N. (2021). Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3143–3155, Online, April. Association for Computational Linguistics.