Improving Generalizability in Implicitly Abusive Language Detection with Concept Activation Vectors

Anonymous ACL submission

Abstract

Robustness of machine learning models on ever-changing real-world data is critical, especially for applications affecting human well-being such as content moderation. New kinds of abusive language continually emerge in online discussions in response to current events (e.g., COVID-19), and the deployed abuse detection systems should be updated regularly to remain accurate. In this paper, we show that general abusive language classifiers tend to be fairly reliable in detecting out-of-domain explicitly abusive utterances but fail to detect new types of more subtle, implicit abuse. Next, we propose an interpretability technique, based on the Testing Concept Activation Vector (TCAV) method from computer vision, to quantify the sensitivity of a trained model to the human-defined concepts of explicit and implicit abusive language, and use that to explain the generalizability of the model on new data, in this case, COVID-related anti-Asian hate speech. Extending this technique, we introduce a novel metric, *Degree of Explicitness*, for a single instance and show that the new metric is beneficial in suggesting out-of-domain unlabeled examples to effectively enrich the training data with informative, implicitly abusive texts.

1 Introduction

When machine learning models are deployed in the real world, they must be constantly monitored for their robustness to new and changing input data. One area where this is particularly important is in abusive language detection (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018; Nakov et al., 2021; Vidgen and Derczynski, 2020). The content of online conversation is constantly changing in response to political and social events. New categories of abusive language emerge, encompassing topics and vocabularies unknown to previously trained classifiers. Here, we tackle three main questions: How can a human user formalize new, relevant topics or concepts in text? How do we quantify the sensitivity of a trained classifier to these new concepts as they emerge? And how do we update the classifier so that it remains reliable?

As a case study, we consider the rise of COVID-related anti-Asian racism on social media. The COVID-19 pandemic represented an entirely new and unexpected situation, generating new vocabulary (*COVID-19*, *coronavirus*, *social distancing*, *masking*), new topics of conversation (dealing with isolation, working from home), and – unfortunately – new and renewed instances of hate speech directed towards Asian communities. We imagine the case of an abusive language detection algorithm which had been deployed prior to the pandemic: what are the new types of abusive language that have emerged with the recent pandemic? To what extent can deployed classifiers generalize to this new data, and how can they be adapted? Although social events can spark off a specific type of hate speech, they are rarely the root cause of the issue. Often such hateful beliefs existed before the event, and are only magnified because of it (Chou and Feagin, 2015). Therefore, we expect that the classifier should detect this new variety of hate speech to some extent.

An important factor in this study is whether the text expresses explicit or implicit abuse (Waseem et al., 2017; Wiegand et al., 2021). Explicit abuse refers to utterances that include direct insults or strong rudeness, often involving profanities, whereas implicit abuse involves more indirect and nuanced language. Since understanding the offensive aspects of implicit abuse in our case study may require some knowledge of the context (i.e., the pandemic), we expect that the pretrained classifier will find these data especially difficult to handle.

To examine a classifier's ability to handle new type of abusive text (without access to extensive labeled data), we propose a technique based on the Testing Concept Activation Vector (TCAV) method from the interpretability literature in computer vi-

sion (Kim et al., 2018). TCAV is used to explain whether a classifier associates a specific concept to a class label (e.g., the concept of *stripes* is associated with class *zebra* in image classification). Similarly, we define implicit and explicit COVID-related anti-Asian racism with a small set of human-chosen textual examples, and ask whether the pre-trained classifier associates these concepts with the positive (abusive) class.

Further, we ask whether sensitivity to human-defined concepts can direct data augmentation[1] to improve generalizations. Intuitively, when updating a classifier, data enrichment should focus on adding examples of concepts to which the classifier is not yet sensitive. Conventional active learning frameworks suggest examples with the lowest classification confidence as the most informative augmentation samples. However, deep neural networks' inability to provide reliable uncertainty estimates is one of the main barriers to adopting confidence-based sampling techniques (Schröder and Niekler, 2020). We suggest that, in the case of abuse detection, implicitly abusive examples are most informative for updating a general classifier. However, to the best of our knowledge, there is no quantitative metric that can measure the degree of explicitness of a candidate example, given a trained classifier. We extend the TCAV technique to provide a "degree of explicitness" measure at the utterance level and use that for efficient data augmentation. We make the following contributions (Supplementary Material includes code and data):

- We implement a variation of the TCAV framework for a RoBERTa-based classifier and show that it can be used to quantify the sensitivity of a trained classifier to a human-understandable concept, defined through examples, without access to the training dataset of the classifier or a large annotated dataset for the new category.
- We analyse the performance of two abusive language classifiers and observe that they generalize well to explicit COVID-related anti-Asian racism, but are unable to generalize to implicit racism of this type. We show that sensitivities to the concepts of implicit and explicit abuse can explain the observed discrepancies.
- We adjust the TCAV method to compute the *degree of explicitness*, for an unlabeled instance, as a metric to guide data augmentation when updating a general abusive language classifier to include a new kind of abuse. We test this method against a confidence-based augmentation algorithm and show that it is able to reach higher performance with fewer training examples, while maintaining the accuracy on the original data.

[1]In this paper, we use the term *augmentation* to refer to the process of enriching the training data by adding examples from sources other than the original dataset.

2 Datasets and Data Analysis

We consider the following four English datasets, summarized in Table 1: *Founta*[2] and *Wiki*[3] are large, commonly-used datasets for general abusive language detection, while *EA* and *CH* specifically target COVID-related anti-Asian racism. We binarize all datasets to two classes: positive (i.e., abusive or hateful) and negative. For *Founta*, this means combining Abusive and Hateful texts into a single positive class; for *EA*, "Hostility against an East-Asian entity" is considered positive, and all other classes are negative; and for *CH*, all hate speech is classed as positive, while counter-hate and hate-neutral texts are classed as negative.

Differences in vocabulary: Central to our research question is the issue of vocabulary change as a new abusive topic emerges. As the *Wiki* and *Founta* datasets were collected before the COVID-19 pandemic, they do not contain novel vocabulary such as "chinavirus" or "wuhanflu", and the contexts and frequencies for words like "China" and "pandemic" may have changed. As a demonstration of the differences in vocabulary across the different datasets, we compute the top 100 most frequent words in the positive class of each dataset (after removing stop words), and then calculate the overlap between each pair of datasets. The complete lists of words are given in the Appendix A. This analysis reveals that the two COVID-related datasets share more words in common, and their shared vocabulary is specific to the pandemic or has found new connotations because of the pandemic. In contrast, the shared vocabulary between *Wiki* and *Founta* largely consists of profanity and strongly negative words such as "hate". Interestingly, *CH* has a similar set of profane words in common with both *Wiki* and *Founta*, while the words shared between *EA* and the general datasets are simply common words in

[2]For *Founta*, we discard the tweets labeled as Spam and use the train-dev-test split as provided by (Zhou et al., 2021).

[3]We used a smaller version of the *Wiki* dataset as provided by Nejadgholi and Kiritchenko (2020). They removed 54% of Wikipedia-specific non-toxic instances from the training set to mitigate the topic bias, and reported improvements in both the classification performance and the execution time. We found similar benefits in our preliminary experiments.

Dataset	Data Source	Positive Class	Negative Class	Number (%Pos:%Neg) *Train*	*Dev*	*Test*
Wikipedia Toxicity (*Wiki*) (Wulczyn et al., 2017)	Wikipedia comments	Toxic	Normal	43,737 (17:83)	32,128 (9:91)	31,866 (9:91)
Founta et al. (2018) dataset (*Founta*)	Twitter posts	Abusive; Hateful	Normal	62,103 (37:63)	10,970 (37:63)	12,893 (37:63)
East-Asian Prejudice (*EA*) (Vidgen et al., 2020)	Twitter posts	Hostility against an East Asian entity	Criticism of an East Asian entity; Counter speech; Discussion of East Asian prejudice; Non-related	16,000 (19:81)	1,200 (19:81)	2,800 (19:81)
COVID-HATE (*CH*) (Ziems et al., 2020)	Twitter posts	Anti-Asian COVID-19 hate; hate directed to non-Asians	Pro-Asian COVID-19 Counterhate; Hate-Neutral	–	–	2,319 (43:57)

Table 1: Class descriptions, number of instances and ratio of positive to negative in percentage (%Pos:%Neg) for the general abusive datasets (*Wiki* and *Founta*) and COVID-related Anti-Asian hate speech datasets (*EA* and *CH*).

the English language, such as "people", "want", and "need." We expect that this vocabulary shift between the different datasets will have a considerable impact on the generalizability.

Differences in explicitness: Another important factor in our study is generalization with respect to explicit and implicit types of abusive language. Above, we observed that *CH* shares many profane words with the general datasets and, therefore, we anticipate it contains more explicitly abusive texts than *EA* does. Unfortunately, neither of the datasets has originally been annotated for *explicitness of abuse*. We manually annotate instances from the positive class in the *CH* dataset and the *EA* dev set using the following rule: instances that include profanity, insult or rudeness that could be correctly identified as abusive without general knowledge about the COVID-19 pandemic are labeled as explicitly abusive; the remaining instances (e.g., *'it is not covid 19 but wuhanvirus'*) are labeled as implicitly abusive. We find that 85% of the *CH-positive* class is categorized as explicit, whereas only 8% of the *EA-positive* class in the *EA* dev set is labeled as explicit. Thus, *CH* and *EA* share COVID-related vocabulary, but are very different in terms of explicitness of abuse (*CH* containing mostly explicit abuse while *EA* containing mostly implicit abuse), which makes them suitable test beds for assessing the generalizability of classifiers to a new type of abusive language and the impact of new vocabulary on the classification of implicit and explicit abuse.

3 Cross-Dataset Generalization

We start by assessing the robustness of a general-purpose abusive language classifier on a new domain of abusive language. Specifically, we analyze the performance of classifiers trained on the *Wiki* and *Founta* datasets (expected to detect general toxicity and abuse) on COVID-related anti-Asian racism data. In addition, we want to assess the impact of the change of vocabulary on the generalizibility of the classifiers to implicit and explicit abuse in the new domain. We train binary RoBERTa-based classifiers on the *Wiki*, *Founta*, *EA* and *CH* datasets (referred to hereafter as the *Wiki*, *Founta*, *EA* and *CH* classifiers), and test them on the *EA* as the mostly implicit COVID-related dataset and *CH* as the mostly explicit COVID-related dataset. (The training details are provided in Appendix C.) Note that *CH* is small, so it is used either as a training dataset when testing on *EA* or a test dataset for all other classifiers. For comparison, we also train an "explicit general abuse" classifier with only explicit examples of the *Wiki* dataset and the class balance similar to the original *Wiki* dataset. This classifier is referred to as *Wiki-exp*.[4]

Table 2 presents the Area Under the ROC Curve (AUC) and F1-scores for all the classifiers; precision, recall, and average precision score are provided in Appendix B. We first consider whether class imbalances can explain our results. Note that while abusive language is a relatively rare phenomenon in online communications, most abusive language datasets are collected through boosted sampling and therefore are not subject to extreme class imbalances. The percentage of positive in-

[4]For *Wiki-exp*, the examples of the positive class are taken from the 'explicit abuse' topic, which contains texts with explicitly toxic words, from (Nejadgholi and Kiritchenko, 2020), and negative examples are randomly sampled from the *Wiki-Normal* class.

Domain	Train Set	AUC		F1-score	
		EA	*CH*	*EA*	*CH*
COVID	*EA*	0.94	0.82	0.74	0.66
	CH	0.86	-	0.62	-
pre-COVID	*Founta*	0.69	0.73	0.29	0.65
	Wiki	0.64	0.74	0.27	0.69
	Wiki-exp	0.58	0.71	0.15	0.56

Table 2: Cross-dataset generalization on *EA* (mostly implicit) and *CH* (mostly explicit) datasets.

stances in our datasets ranges from 9% to 43% (Table 1). We observe similar performances for the *Wiki* and *Founta* classifiers despite different class ratios in their training sets, and different performances for *Wiki* and *EA* classifiers despite their similar training class ratios. We also observe better performance from *CH* classifier compared to the *Wiki* or *Founta* classifiers despite the very small size of the *CH* dataset. Based on previous research, we argue that cross-dataset generalization in abusive language detection is often governed by the compatibility of the definitions and sampling strategies of training and test labels rather than class sizes (Yin and Zubiaga, 2021). Instead, we explain the results presented in Table 2 in terms of implicit/explicit types of abuse and the change of vocabulary.

Cross-dataset generalization is better when datasets share similar vocabulary. The classifiers trained on the *EA* and *CH* datasets perform better than all the classifiers trained on the pre-COVID datasets (*Wiki* and *Founta*). Interestingly, the performance of the *CH* classifier on the *EA* dataset is higher than the performance of all the general classifiers, despite the *CH* dataset being very small and containing mostly explicit abuse. This observation confirms that general classifiers need to be updated to learn the new vocabulary.

General-purpose classifiers generalize better to explicit than implicit examples in the new domain. The *Wiki* and *Founta* classifiers, which have been exposed to large amounts of generally explicit abuse, perform well on the mostly explicit *CH* dataset, but experience difficulty with the COVID-specific implicit abuse in the *EA* dataset. For example, the tweet *'the chinavirus is a biological attack initiated by china'* is misclassified as non-abusive. We observe that *Wiki-exp* performs relatively similar to the *Wiki* classifier on *CH*, despite its small size (only 1,294 positive examples) but is worse than *Wiki* classifier on *EA*. This means that the additional 35K instances (of which, 9K are positive examples) of the *Wiki* compared to the *Wiki-exp*, only moderately improve the classification of the implicit examples in the new domain. This observation indicates that generalization mostly occurs between the explicit type of the pre-COVID abuse and the explicit type of the COVID-related abuse. Therefore, a general-purpose classifier should be specifically updated to learn implicit abuse in the new domain.

4 Sensitivity to Implicit and Explicit Abuse to Explain Generalizability

In Section 3, we showed that when a new domain emerges, the change in vocabulary mostly affects the classification of implicitly expressed abuse. This observation is in line with findings by Fortuna et al. (2021), and suggests that generalization should be evaluated on implicit and explicit abuse separately. However, due to complexities of annotation of abusive content, curating separate implicit and explicit test sets is too costly (Wiegand et al., 2021). Instead, we propose to adapt the Testing Concept Activation Vector (TCAV) algorithm, originally developed for image classification (Kim et al., 2018), to calculate the classifiers' sensitivity to explicit and implicit COVID-related racism, using only a small set of examples. Then, we show how these sensitivities can explain the generalizations observed in Table 2.

4.1 TCAV background and implementation

TCAV is a post-training interpretability method to measure how important a user-chosen concept is for a prediction, even if the concept was not used as a feature during the training. Similarly to how Kim et al. (2018) suggest "stripes" as a visual concept relevant to the class "zebra", and then operationally define the "stripes" concept by collecting examples of images containing stripes, in our language-based TCAV method a concept is defined by a set of manually chosen textual examples. Here, we consider concepts such as COVID-19, hate speech, and anti-Asian abuse, but the approach generalizes to any concept that can be defined through a set of example texts. Using these examples, a Concept Activation Vector (CAV) is learned to represent the concept in the activation space of the classifier. Then, directional derivatives are used to calculate the sensitivity of predictions to changes in inputs towards the direction of the concept, at the neural activation layer.

We adapt the TCAV procedure for a binary

Non-coherent concept: random tweets collected with stop words as queries
COVID-19: tweets collected with words *covid, corona, covid-19, pandemic* as query words
Explicit anti-Asian abuse: tweets labeled as explicit from *EA* dev and *CH*
Implicit abuse (EA): tweets labeled as implicit from *EA* dev
Implicit abuse (CH): tweets labeled as implicit from *CH*
Generic hate: tweets from the *Hateful* class of *Founta* dev

Table 3: Human-defined concepts and the sources of the tweets used as concept examples.

RoBERTa-based classifier to measure the importance of a concept to the positive class. For any input text, $x \in \mathbb{R}^{k \times n}$, with k words in the n-dimensional input space, we consider the RoBERTa encoder of the classifier as $f_{emb} : \mathbb{R}^{k \times n} \rightarrow \mathbb{R}^m$, which maps the input text to its RoBERTa representation (the representation for [CLS] token), $r \in \mathbb{R}^m$. For each concept, C, we collect N_C concept examples, and map them to RoBERTa representations $r_C^j, j = 1, ..., N_C$. To represent C in the activation space, we calculate P number of CAVs, υ_C^p, by averaging[5] the RoBERTa representations of N_υ randomly chosen concept examples:

$$\upsilon_C^p = \frac{1}{N_\upsilon} \sum_{j=1}^{N_\upsilon} r_C^j \quad p = 1, .., P \tag{1}$$

where $N_\upsilon < N_C$. The *conceptual sensitivity* of the positive class to the υ_C^p, at input x can be computed as the directional derivative $S_{C,p}(x)$:

$$S_{C,p}(x) = \lim_{\epsilon \rightarrow 0} \frac{h(f_{emb}(x) + \epsilon \upsilon_C^p) - h(f_{emb}(x))}{\epsilon}$$

$$= \nabla h(f_{emb}(x)).\upsilon_C^p \tag{2}$$

where $h : \mathbb{R}^m \rightarrow \mathbb{R}$ is the function that maps the RoBERTa representation to the logit value of the positive class. In Equation 2, $S_{C,p}(x)$ measures the changes in class logit, if a small vector in the direction of C is added to the input example, in the RoBERTa-embedding space. For a set of input examples X, we calculate the TCAV score as the fraction of inputs for which small changes in the direction of C increase the logit:

$$TCAV_{C,p} = \frac{|x \in X : S_{C,p}(x) > 0|}{|X|} \tag{3}$$

[5]In the original TCAV algorithm, a linear classifier is trained to separate representations of concept examples and random examples. Then, the vector orthogonal to the decision boundary of this classifier is used as the CAV. We experimented with training a linear classifier and found that the choice of random utterances has a huge impact on the results to the point that the results are not reproducible. More stable results are obtained when CAVs are produced by averaging the RoBERTa representations.

A TCAV score close to one indicates that for the majority of input examples the logit value increases. Equation 3 defines a distribution of scores for the concept C; we compute the mean and standard deviation of this distribution to determine the overall sensitivity of the classifier to the concept C.

4.2 Classifier's Sensitivity to a Concept

We define each concept C with $N_C = 100$ manually chosen examples, and experiment with six concepts described in Table 3. To set a baseline, we start with a set of random examples to form a non-coherent concept. Next, we define a non-hateful COVID-related concept using random tweets with COVID-related keywords *covid, corona, covid-19, pandemic*. For the explicit anti-Asian abuse concept, we include all 14 explicitly abusive examples from the *EA* dev set and 86 explicitly abusive examples from *CH* class. We define two implicit anti-Asian concepts with examples from *EA* and *CH*, to assess whether selecting the examples from two different datasets affects the sensitivities. We also define the generic hate concept with examples of pre-COVID general hateful utterances, not directed at Asian people or entities, from the *Founta* dev set.

We calculate $P = 1000$ CAVs for each concept, where each CAV is the average of $N_\upsilon = 5$ randomly chosen concept examples. We use 2000 random tweets collected with stopwords as input examples X (see Equation 3).[6] Table 4 presents the means and standard deviations of the TCAV score distributions for the classifiers trained on *Wiki*, *Founta*, *EA*, and *CH* datasets, respectively. First, we observe that all TCAV scores calculated for a random, non-coherent set of examples are zero; i.e., as expected, the TCAV scores do not indicate any association between a non-coherent concept and the positive class. Also, as expected, none of the classifiers associate the non-hateful COVID-related concept to the positive class. Note that a zero TCAV score can be due to the absence of that concept in the training data (e.g., the COVID concept for the *Wiki* and *Founta* classifiers), insignificance of the topic for predicting the positive label (e.g., the COVID concept for the *EA* classi-

[6]Unlike the original TCAV algorithm, we do not restrict the input examples from the target class. In our experiments, we observed that, for this binary classification set-up, the choice of input examples has little impact on the TCAV scores. Intuitively, we assess whether adding the concept vector to a random input would increase the likelihood of it being assigned to the positive class.

Classifier	Concept					
	non-coherent	COVID-19	explicit anti-Asian	implicit (EA)	implicit (CH)	generic hate
EA	0.00 (0.00)	0.00 (0.00)	**0.90** (0.26)	**0.87** (0.30)	**0.70** (0.42)	0.00 (0.00)
CH	0.00 (0.00)	0.00 (0.00)	-	0.35 (0.44)	-	0.21 (0.12)
Founta	0.00 (0.02)	0.00 (0.01)	**0.92** (0.22)	0.00 (0.06)	0.19 (0.32)	**0.60** (0.44)
Wiki	0.00 (0.03)	0.00 (0.05)	**0.96** (0.16)	0.00 (0.03)	0.28 (0.43)	**0.75** (0.41)
Wiki-exp	0.00 (0.05)	0.00 (0.07)	**0.78** (0.12)	0.00 (0.02)	0.00 (0.05)	**0.59** (0.40)

Table 4: Means and standard deviations of TCAV score distributions for the positive class of the five classifiers with respect to six human-defined concepts. Scores statistically significantly different from random are in bold.

fier), or the lack of coherence among the concept examples (such as the concept defined by random examples). A TCAV score close to 1, on the other hand, indicates the importance of a concept for positive prediction. These observations set a solid baseline for interpreting the TCAV scores, calculated for other concepts. Here we ask whether the generated TCAV scores can explain the generalization performances observed in Table 2.

We consider a classifier to be sensitive to a concept if its average TCAV score is significantly different (according to the t-test with $p < 0.001$) from the average TCAV score of a non-coherent random concept. First, we observe that the general classifiers are only sensitive to the explicit type of COVID-related abusive language. This confirms that the classifiers generalize better to the explicit type of an emerging domain of abusive language. We also note that *Wiki-exp*, is sensitive to the explicit anti-Asian concept.

Second, the classifier trained with mostly explicit COVID-related data (*CH*) is not sensitive to the implicit abuse concept.[7] The only classifier that is sensitive to the explicit and both implicit COVID-related abusive concepts is the *EA* classifier. Classifiers trained on the COVID datasets are also not sensitive to the generic hate concept, which encompasses a much broader range of target groups. Overall, these findings stress the importance of including implicitly abusive examples in the training data for better generalizability within and across domains.

5 Degree of Explicitness

Here, we suggest that implicit examples are more informative (less redundant) for updating a general classifier and provide a quantitative metric to guide the data augmentation process. We extend the TCAV methodology to estimate the *Degree of Explicitness* or *DoE* of an utterance. We showed that the average TCAV score of the positive class for the explicit concept is close to 1. DoE is based on the idea that adding one example to an explicit concept will not affect its average TCAV score (i.e., it will still be close to 1), if the added example is explicitly abusive. However, adding an implicit example presumably will change the direction of all CAVs and reduce the sensitivity of the classifier to this modified concept. Here, we modify Equation 1 and calculate each CAV by averaging the RoBERTa representations of $N_\upsilon - 1$ explicit concept examples, and the new utterance for which we want the degree of explicitness, x_{new}, with representation r_{new}. Thus,

$$\upsilon^p_{new} = \frac{1}{N_\upsilon}\left(\sum_{j=1}^{N_\upsilon - 1} r^j_C + r_{new}\right), \quad p = 1, .., P$$

We then calculate the average TCAV score for each x_{new} as its DoE score. If the new utterance, x_{new}, is explicitly abusive, υ^p_{new} will represent an explicit concept, and the average TCAV score, i.e., $mean(TCAV_{C,p})$ will remain close to 1. However, the less explicit the new example is, the more υ^p_{new} will diverge from representations of explicit abuse, and the average score will drop. We use $N_\upsilon = 3$ in the following experiments.

DoE analysis on COVID-related abusive data: We validate the utility of DoE in terms of separating implicit and explicit abusive examples. For the *Wiki* and *Founta* classifiers, we calculate the DoE score of the implicit and explicit examples from *CH* and the *EA* dev set (described in Section 3), excluding the examples used to define the *Explicit anti-Asian abuse* concept. Given that low classification confidence could indicate that the model struggles to predict an example correctly, one might expect that implicit examples are classified with less classification confidence than explicit examples. Figure 1 shows the comparison of DoE with classification confidence in distinguishing between implicit and explicit examples. We observe that for both classifiers, the distribution of DoE scores of implicit

[7] We do not measure the sensitivity of this classifier to the explicit anti-Asian and implicit CH concepts, since their concept examples are included in the training set of the classifier.

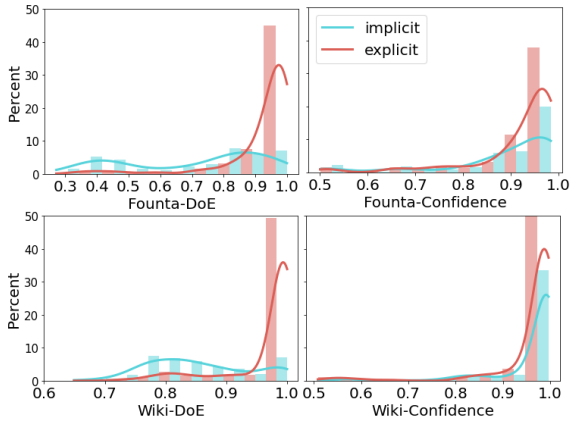

Figure 1: Comparison of classification confidence and DoE score for distinguishing between implicit and explicit abusive utterances.

examples is different from the distribution of DoE scores of explicit examples, but the distributions of their classification confidences are indistinguishable. Therefore, we conclude that DoE is more effective at separating implicit abuse from explicit abuse than classification confidence. We further analyze DoE scores for the positive and negative classes separately in Appendix D.

6 Data Augmentation with DoE score

We now use the DoE score to direct data augmentation. We consider a scenario where a general classifier should be re-trained with an augmented dataset to include emerging types of abusive language. As we showed, general classifiers are already sensitive to explicit abuse. Therefore, we hypothesize that implicit examples are more beneficial for updating the classifier. Here, we describe a novel DoE-based augmentation approach and contrast it with the conventional process of choosing augmentation examples based on the classification confidence (Zhu et al., 2008; Chen et al., 2019).

We consider the general *Wiki* classifier. Our goal is to find a small but sufficient portion of the *EA* train set to augment the original *Wiki* train set, so that the classifier is able to handle COVID-related anti-Asian hate speech. We calculate the DoE and confidence scores for all the examples in the *EA* train set and add the N examples with the lowest scores to the original *Wiki* train set. We vary N from 1K to 6K, with a 1K step. After the augmentation data size reaches 6K, the classifier performance on the original *Wiki* test set drops substantially for both techniques. Also, note that as the size of the augmentation dataset increases, the two methods converge to the same performance.

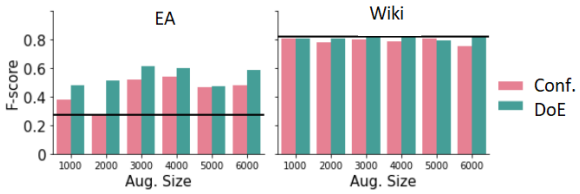

Figure 2: F1-score of the augmented *Wiki* classifier on the *EA* and *Wiki* test sets. Solid lines show the baseline.

6.1 Results

Figure 2 shows the F1-score of the classifiers updated using the DoE and confidence-based augmentation methods on the original test set (*Wiki*) and the new test set (*EA*) for different augmentation sizes. (Precision and recall figures are provided in Appendix E.) Since only *EA* is used for augmentation, we evaluate the classifiers on this dataset to find the optimum size for the augmented training set and only evaluate the best performing classifiers on *CH*. We expect that an efficient augmentation should maintain the performance on *Wiki* and reach acceptable results on *EA* test set.

DoE is better at learning the new type of abuse: On the *EA* dataset, DoE achieves better results than the confidence-based augmentation method for all augmentation sizes, except for N= 5K, where the performances of the two methods are comparable.

DoE is better at maintaining performance on the original dataset: DoE outperforms the confidence-based method on the *Wiki* dataset. For all augmentation sizes, the performance of the DoE-augmented classifier on this class stays within 2% of the baseline (the F1-score of the classifier trained just on the *Wiki* data), whereas for the confidence-based augmentation, we observe up to 6% drop depending on the size of the added data.

DoE is better overall: Table 5 presents the best results achieved by the two augmentation methods on the *EA* test set: AUC score of 0.81 for the DoE-based augmentation obtained with 3K added examples, and AUC score of 0.69 for the confidence-based augmentation obtained with 4K added examples. For comparison, we also show the baseline results for the original *Wiki* classifier and the classifier trained on the combined *Wiki* and full *EA* train sets. Although we did not optimize the augmentation for the *CH* dataset, our evaluation shows that DoE performs favourably on this dataset, as well. We conclude that the new DoE-based augmentation

		F1-score			AUC		
Method	**Aug. set**	***EA***	***CH***	***Wiki***	***EA***	***CH***	***Wiki***
DoE	3K *EA*	**0.61**	**0.73**	**0.82**	**0.81**	**0.78**	**0.96**
Conf.	4K *EA*	0.54	0.71	0.79	0.69	0.75	0.94
Merging	*EA*	0.58	0.72	0.78	0.72	0.75	0.94
baseline	-	0.27	0.69	0.82	0.64	0.74	0.96

Table 5: AUC and F1-scores for the best performing classifiers updated with various augmentation methods, as well as the original *Wiki* classifier as baseline.

method maintains the classification performance on the original dataset, while outperforming the other method on the new data.

We also qualitatively assess the classifier's output before and after data augmentation with DoE. While explicitly abusive utterances (e.g., "f*ck you china and your chinese virus") are often correctly classified both before and after re-training, many implicitly abusive examples (e.g., "it is not covid 19 but wuhanvirus") are handled correctly by the classifier only after re-training.

7 Related Work

Generalizability has been an active research area in NLP (Ettinger et al., 2017; Hendrycks et al., 2020). Several studies evaluated generalizability in abuse detection through cross-dataset evaluation (Swamy et al., 2019; Wiegand et al., 2019), direct dataset analysis (Fortuna et al., 2020) or topic modeling on the training data (Nejadgholi and Kiritchenko, 2020). Fortuna et al. (2021) showed that the lack of generalizability is rooted in the imbalances between implicit and explicit examples in training data. In a recent review, Yin and Zubiaga (2021) discussed the challenges for building generalizable hate speech detection systems and recommended possible future directions.

The distinction between explicit and implicit abuse has been recognized as an important factor in abuse detection (Waseem et al., 2017). Wiegand et al. (2019) showed that lexicon-based sampling strategies fail to collect implicit abuse and most of the annotated datasets are overwhelmed with explicit examples. Breitfeller et al. (2019) showed that inter-annotation agreement is low when labeling the implicit abuse utterances, as sometimes specific knowledge is required in order to understand the implicit statements. For better detection of implicitly stated abuse, large annotated datasets with hierarchical annotations are needed (Sap et al., 2020), so that automatic detection systems can learn from a wide variety of such training examples. Field and Tsvetkov (2020) proposed propensity matching and adversarial learning to force the model to focus on signs of implicit bias. Wiegand et al. (2021) created a novel dataset for studying implicit abuse and presented a range of linguistic features for contrastive analysis of abusive content.

Data augmentation has been used to improve the robustness of abuse detection classifiers. To mitigate biases towards specific terms (e.g., identity terms), one strategy is to add benign examples containing the biased terms to the training data (Dixon et al., 2018; Park, 2018; Badjatiya et al., 2019). Other works combined multiple datasets to achieve better generalizations, using a set of probing instances (Han and Tsvetkov, 2020), multi-task training (Waseem et al., 2018), and domain adaptation (Karan and Šnajder, 2018). In contrast to these works, we take an interpretability-based approach and guide the data collection process by mapping the new data on the implicit vs. explicit spectrum.

8 Conclusion

As real-world data evolves, we would like to be able to query a trained model to determine whether it generalizes to the new data, without the need for a large, annotated test set. We adopted the TCAV algorithm to quantify the sensitivity of text classifiers to human-chosen concepts, defined with a small set of examples. We used this technique to compare the generalizations of abusive language classifiers, trained with pre-pandemic data, to explicit and implicit COVID-related anti-Asian racism.

We then proposed a sensitivity-based data augmentation approach, to improve generalizability to emerging categories. We showed that in the case of abuse detection, the most informative examples are implicitly abusive utterances from the new category. Our approach collects implicit augmentation examples and achieves higher generalization to the new category compared to confidence-based sampling. Strategies for choosing the optimal set of concept examples should be explored in the future.

While we examined abusive language detection as a case study, similar techniques can be applied to different NLP applications. For example, the TCAV method could be used to measure the sensitivity of a sentiment analysis system to a new product, or a stance detection algorithm's sensitivity to an important new societal issue. As language evolves, methods of monitoring and explaining classifier behaviour over time will be essential.

Ethical Considerations

Content moderation is a critical application with potential of significant benefits, but also harms to human well-being. Therefore, ethics-related issues in content moderation have been actively studied in NLP and other disciplines (Vidgen et al., 2019; Wiegand et al., 2019; Kiritchenko et al., 2021; Vidgen and Derczynski, 2020). These include sampling and annotation biases in data collection, algorithmic bias amplification, user privacy, system safety and security, and human control of technology, among others. Our work aims to address the aspects of system safety and fairness by adapting the model to newly emerged or not previously covered types of online abuse, often directed against marginalized communities. We employ existing datasets (with all their limitations) and use them only for illustration purposes and preliminary evaluation of the proposed methodology. When deploying the technology care should be taken to adequately address other ethics-related issues.

References

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *Proceedings of the World Wide Web Conference*, pages 49–59.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. 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 1664–1674, Hong Kong, China. Association for Computational Linguistics.

Xi C. Chen, Adithya Sagar, Justine T. Kao, Tony Y. Li, Christopher Klein, Stephen Pulman, Ashish Garg, and Jason D. Williams. 2019. Active learning for domain classification in a commercial spoken personal assistant. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Rosalind S Chou and Joe R Feagin. 2015. *Myth of the model minority: Asian Americans facing racism*. Routledge.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M. Bender. 2017. Towards linguistically generalizable NLP systems: A workshop and shared task. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 1–10, Copenhagen, Denmark. Association for Computational Linguistics.

Anjalie Field and Yulia Tsvetkov. 2020. Unsupervised discovery of implicit gender bias. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 596–608.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? What are we really classifying? An empirical analysis of hate speech datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6786–6794.

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

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. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Xiaochuang Han and Yulia Tsvetkov. 2020. Fortifying toxic speech detectors against veiled toxicity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7732–7739, Online. Association for Computational Linguistics.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, Online. 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.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the International Conference on Machine Learning*, pages 2668–2677.

Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C Fraser. 2021. Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research*, 71:431–478.

Varada Kolhatkar, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla, and Maite Taboada. 2019. The SFU opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, pages 1–36.

Preslav Nakov, Vibha Nayak, Kyle Dent, Ameya Bhatawdekar, Sheikh Muhammad Sarwar, Momchil Hardalov, Yoan Dinkov, Dimitrina Zlatkova, Guillaume Bouchard, and Isabelle Augenstein. 2021. Detecting abusive language on online platforms: A critical analysis. *arXiv preprint arXiv:2103.00153*.

Isar Nejadgholi and Svetlana Kiritchenko. 2020. On cross-dataset generalization in automatic detection of online abuse. In *Proceedings of the 4th Workshop on Online Abuse and Harms*.

Ji Ho Park. 2018. Finding good representations of emotions for text classification. *arXiv preprint arXiv:1808.07235*.

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

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the 5th International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Christopher Schröder and Andreas Niekler. 2020. A survey of active learning for text classification using deep neural networks. *arXiv preprint arXiv:2008.07267*.

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.

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

Bertie Vidgen, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Waseem, Austin Botelho, Matthew Hall, and Rebekah Tromble. 2020. Detecting East Asian prejudice on social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 162–172, Online. Association for Computational Linguistics.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

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

Michael Wiegand, Maja Geulig, and Josef Ruppenhofer. 2021. Implicitly abusive comparisons – a new dataset and linguistic analysis. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 358–368, Online. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 602–608.

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.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *arXiv preprint arXiv:2102.08886*.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah A Smith. 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.

Jingbo Zhu, Huizhen Wang, and Eduard Hovy. 2008. Learning a stopping criterion for active learning for word sense disambiguation and text classification. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. 2020. Racism is a virus: Anti-asian hate and counterhate in social media during the COVID-19 crisis. *arXiv preprint arXiv:2005.12423*.

A Change of Vocabulary

To examine the change of vocabulary, from general to COVID, we look at the most frequent words in the positive class of our datasets. Table A.1 shows the common words between the 100 most frequent words of the positive classes of our datasets after removing stop words. We used the stop word list from the scikitlearn package. We observe that COVID-related datasets share more words, and their shared vocabulary is specific to the pandemic. In other cases, fewer words are shared, and the shared words are either related to profanity and violence or are commonly used English terms.

B Additional Results for Cross-Dataset Generalization

In table B.1, we present additional metrics for the generalizibility experiments described in Section 3. Besides the commonly used metrics, precision and recall, we measure averaged precision score to count for potential threshold adjustments. Averaged precision score summarizes a precision-recall curve as the weighted mean of precisions at each threshold, weighted by the increase in recall from the previous threshold. The results are consistent with AUC and F1-scores reported in Table 2.

C Model Specifications

All of our models are binary RoBERTa-based classifiers trained with the default settings of the Trainer module from the Huggingface library [8] with 3 training epochs, on a Tesla V100-SXM2 GPU machine, batch size of 16, warm-up steps of 500 and weight decay of 0.01. We use Roberta-base model, which includes 12-layer, 768 hidden nodes, 12 head nodes, 125M parameters, and add a linear layer with two nodes for binary classification. Training these classifiers takes several hours depending on the size of the training dataset.

D DoE Analysis on the *EA* Train Set

With the DoE score, we want to distinguish between implicit and explicit examples of abuse. However, when used for data selection, the true labels of the selected examples are not available. We investigate what low DoE scores mean in terms of 'being challenging to classify'. With both *Founta* and *Wiki* classifiers, we calculate the DoE score for all instances of the *EA* train set, sort the negative and positive examples separately based on DoE and look at the classification accuracies in bins of size 100 of sorted DoEs. Figure D.1 shows that low DoE examples are correctly classified if negative and misclassified if positive (implicit abuse). In contrast, high DoE examples are misclassified if negative and correctly classified if positive (explicit abuse).

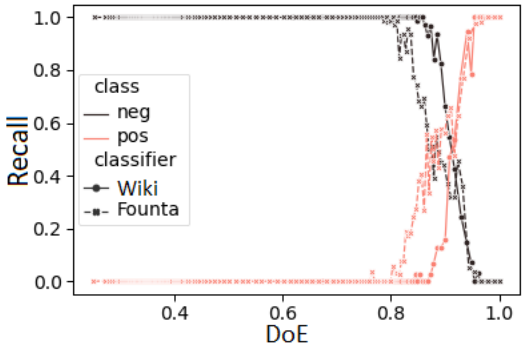

Figure D.1: Recall per class for varying DoE scores on the *EA* train set

E Comparing DoE and Confidence-Based Augmentation Using Precision and Recall

In Section 6, we compare the classifiers updated with DoE and confidence-based methods using classification F1-score. Here, we provide a more fine-grained analysis based on recall and precision.

Figure E.1 shows the recall and precision of the updated classifiers on the *EA* dataset. This figure indicates that the classifiers updated with DoE are much more successful in recognizing abusive utterances than the classifiers updated with confidence, but misclassify more non-abusive sentences, which results in substantially higher recall scores, but slightly lower precision scores. Note that in computer-assisted content moderation, recall is more important than precision, since automatically flagged posts are assessed by human moderators to make the final decision.

We argue that the higher recall and lower precision of classifiers updated with DoE is due to the discrepancies in the definitions of the negative classes for the *Wiki* and *EA* datasets. Previous work has commented on the difficulty of aligning annotations of *abusive*, *offensive*, *hateful*, and *toxic* speech across different datasets (Swamy et al., 2019; Kolhatkar et al., 2019; Fortuna et al., 2021). Here, we

[8] https://huggingface.co/transformers/main_classes/trainer.html

Datasets	Num. of Words	Common Words
EA - CH	50	racist, ccp, came, want, country, 19, communist, calling, come, does, spread, like, pandemic, amp, media, coronavirus, eating, covid19, did, human, chinesevirus, infected, world, covid, know, chinese, chinavirus, government, say, started, think, corona, wuhanvirus, need, blame, evil, time, people, wuhan, don, new, let, news, china, stop, countries, virus, just, spreading, make
Wiki - Founta	37	ass, oh, dont, want, way, going, come, does, bitch, like, look, life, idiot, niggas, did, eat, sex, know, dick, fucking, fuck, say, think, shit, man, need, hell, time, hate, people, stupid, said, stop, really, just, make, tell
Founta - EA	19	racist, want, calling, come, does, like, did, world, know, say, think, need, time, people, trying, let, stop, just, make
Wiki - EA	15	people, want, did, say, think, good, need, come, does, stop, just, know, like, make, time
Founta - CH	35	racist, ass, want, way, going, calling, come, does, bitch, like, got, look, did, eat, world, know, fucking, fuck, say, think, shit, man, trump, need, time, people, hate, stupid, said, fucked, let, stop, really, just, make
Wiki - CH	33	ass, want, way, going, come, does, bitch, like, look, did, eat, right, know, fucking, fuck, die, say, think, shit, man, need, time, people, hate, don, stupid, said, kill, stop, really, shut, just, make

Table A.1: Common words between 100 most frequent words of the positive classes of the datasets.

Train Set	Precision		Recall		Ave. Prec.	
	EA	*CH*	*EA*	*CH*	*EA*	*CH*
EA	0.72	0.77	0.73	0.58	0.80	0.80
CH	0.58	-	0.66	-	0.64	-
Founta	0.46	0.57	0.23	0.73	0.35	0.65
Wiki	0.39	0.61	0.21	0.78	0.31	0.66
Wiki-exp	0.37	0.64	0.10	0.51	0.26	0.64

Table B.1: Additional metrics for cross-dataset generalization results presented in Table 2.

also observe that the definitions of positive (abusive) and negative classes differ significantly between the generalized and COVID-related data. In the *Wiki* and *Founta* datasets, the positive class encompasses a wide range of offensive language, while in the *EA* and *CH* datasets, the positive class is restricted to hate speech and other more intense cases of expressed negativity. Further, the negative class in *Wiki* and *Founta* datasets comprise non-abusive, neutral, or friendly instances while in the *EA* and *CH* datasets the negative class may also include rude and offensive texts as long as they do not constitute hate speech against Asian people or entities.

In Appendix D, we observe that low DoE examples are correctly classified if negative and misclassified if positive (implicit abuse). In contrast, high DoE examples are misclassified if negative and correctly classified if positive (explicit abuse). We use this observation to explain higher recall of the confidence-based method in comparison with the DoE-based method for the *EA-negative* class. As mentioned before, while *EA-positive* fits under the definition of 'toxicity' in *Wiki-positive*, the definition of *EA-negative* is inconsistent with the definition of *Wiki-negative*. In other words, DoE tends to choose negative examples that the *Wiki* classifier already recognizes as negative, whereas the confidence-based data augmentation selects negative examples that are unknown to the classifier. Therefore, the classifier augmented with low confidence scores adapts better to the new definition of negative examples than the classifier updated with low DoE scores. In a real-life scenario, we do not expect the definition of the negative class to change over time, so precision for DoE-base augmentation should not suffer.

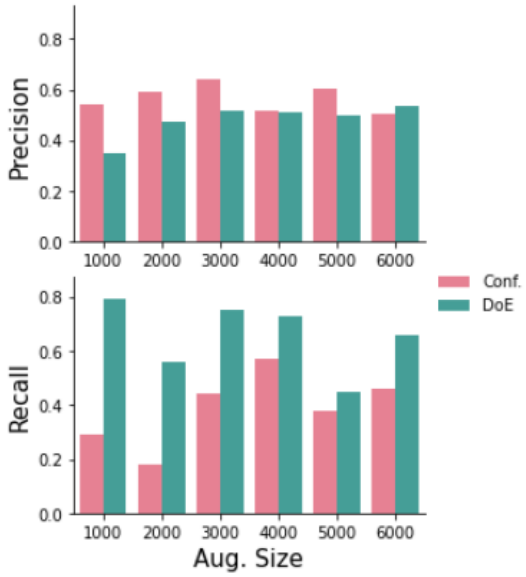

Figure E.1: Precision and recall of the augmented *Wiki* classifier on the *EA* test set.