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. General abusive language classifiers tend to be fairly reliable in detecting out-of-domain explicitly abusive utterances but often fail to detect new types of more subtle, implicit abuse. 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 best select training examples to update the deployed classifier?

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: to what extent can it generalize to this new data? 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 vision (Kim et al., 2018). TCAV is used to explain whether a classifier associates a specific concept

Dataset	Data Type	Positive Class	Negative Class	Number of Instances			% POS
				Train	*Dev*	*Test*	
Wikipedia Toxicity (*Wiki*) (Wulczyn et al., 2017)	Wikipedia comments	Toxic	Normal	43,737	32,128	31,866	0.17 (0.09 on test)
Founta et al. (2018) dataset (*Founta*)	Twitter posts	Abusive; Hateful	Normal	62,103	10,970	12,893	0.37
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	1,200	2,800	0.19
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	0.43

Table 1: Statistics for the general abusive datasets (*Wiki* and *Founta*) and COVID-related Anti-Asian datasets (*EA* and *CH*). *CH* is used only as a test set due to its small size. '% POS' stands for the percentage of positive instances.

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 (code and data are provided in the Supplementary Material):

- 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 learn the new type of abuse more efficiently, while maintaining the accuracy on the original data.

2 Datasets

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.,

[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] 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 use the pruned version of the *Wiki* training set where some Wikipedia-specific non-toxic instances (54% of the dataset) are removed (Nejadgholi and Kiritchenko, 2020). In our preliminary experiments, this reduction did not affect the classification performance but significantly decreased the execution time. The dev and test sets are from the original dataset.

Training set	**F-score**		**Recall per class**				
	EA	*CH*	*EA-positive*	*CH-positive-antiAsian*	*CH-positive-other*	*EA-negative*	*CH-negative*
Wiki	0.26	0.68	0.21	0.77	0.80	0.91	0.62
Founta	0.29	0.63	0.23	0.72	0.75	0.92	0.58
EA	0.74	0.66	0.74	0.70	0.32	0.94	0.87

Table 2: Evaluation of multiple classifiers on the anti-Asian hate speech data in cross-dataset settings.

abusive or hateful) and negative. For *Founta*, this means combining Abusive, Hateful texts into a single positive class; for *EA*, "Hostility against an East-Asian entity" is considered positive, and all other classes are grouped under negative; and for *CH*, all hate speech is classed as positive, while counter-hate and hate-neutral texts are classed as negative. Note that in Section 3, we analyze positive subclasses of the *CH* test set separately and refer to them as *CH-positive-antiAsian* (for Anti-Asian COVID-19 hate) and *CH-positive-other* (for Hate directed to non-Asians).

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 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.

3 Cross-Dataset Generalization

We start by assessing the robustness of a general-purpose abusive language classifier on the COVID-related anti-Asian racism data. We train two binary RoBERTa-based classifiers with the *Wiki* and *Founta* datasets (referred to hereafter as the *Wiki* and *Founta* classifiers), and test them on the *EA* and *CH* datasets. (The training details are provided in Appendix A.) Here, while the classifier makes a binary positive/negative decision, we are really assessing its ability to generalize to the new task of identifying anti-Asian hate. For comparison, we also train a binary classifier with the *EA* train set and evaluate it on both *EA* test set and the *CH* dataset. For a detailed analysis, we report recall scores for all classes along with the classification F-score. Table 2 shows the results.[4] Our main findings are as follows:

Finding 1: The general classifiers (*Wiki* and *Founta*) reach much higher F-scores on *CH* than on *EA* (shown in blue in Table 2). We also observe that the general classifiers obtain higher recall for both sub-classes of *CH-positive* compared to the *EA-positive* class.

Finding 2: The general classifiers reach higher recall on *CH-positive-antiAsian* than the *EA* classifier, even though the latter is trained for detecting COVID-related racism. The general classifiers also perform much better on the *CH-positive-other* class than the *EA* classifier (shown in red in Table 2).

Finding 3: Despite better overall performance on *CH* (Finding 1), the general classifiers reach lower recall on the *CH-negative* class than on the *EA-negative* class (shown in orange in Table 2).

Discussion on class imbalances: Since significant differences in class distributions among datasets can result in performance disparities, we first investigate whether class imbalances can explain our findings. Note that abusive language datasets are often collected through boosted sampling and are not subject to extreme class imbalances. The percentage of positive instances in the datasets used in our study ranges from 9% to 43% (last column of Table 1). To ensure that our analysis is not impacted by differences in the class ratios of the test sets, we compare the recall scores for the positive and negative classes. Recall scores of the positive and negative classes provide a full picture of the performance of a binary classifier (all other metrics can be calculated from the recall scores and class sizes) and are independent of the class ratios. In terms of the training set class imbalances, we observe similar performances for the *Wiki* and *Founta* classifiers despite different class ratios in their training

[4] Our in-dataset results on *EA* are higher than reported in (Vidgen et al., 2020), since we convert the task to binary.

sets, and different performances for *Wiki* and *EA* classifiers despite their similar training class ratios. Therefore, we conclude that class imbalance in the training sets cannot explain our findings. As previous research suggests, 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). Below, we discuss our findings with regards to discrepancies in the positive and negative class content.

Discrepancies in the positive class: The first two findings are quite counter-intuitive since both the *EA* and *CH* datasets were collected to address the task of COVID-related anti-Asian hate speech detection, and are expected to be more similar to each other than to the general-purpose abusive datasets. To further investigate these findings, we manually annotate instances for *explicitness of abuse* from the positive class in the *EA* dev set and the *CH* dataset. Instances that include profanity, insult or rudeness directed at Asian people or entities, and 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 79% of the *CH-positive-antiAsian* class is categorized as explicit, whereas only 8% of the *EA-positive* class in the *EA* dev set is labeled as explicit. Therefore, the *Wiki* and *Founta* classifiers, which have been exposed to large amounts of generally explicit abuse, perform well on the mostly explicit *CH-positive* sub-classes, but experience difficulty with the COVID-specific implicit abuse in the *EA-positive* class. For example, the tweet *'the chinavirus is a biological attack initiated by china'* is misclassified as non-abusive. On the other hand, the *EA* dataset contains fewer explicitly abusive examples than the general datasets, therefore the *EA* classifier underperforms in detecting the explicit examples in the *CH-positive-antiAsian* class, despite being trained for the same task. We also notice that many instances in the *CH-positive-antiAsian* class labeled as 'implicitly abusive' are correctly classified by the *EA* classifier, but are not recognized as abusive by the *Wiki* and *Founta* classifiers. Further, as expected the *EA* classifier performs poorly on the *CH-positive-other* class as it was not exposed to such data.

Discrepancies in the negative class: As mentioned above, the negative class is defined very differently in the general and COVID-related datasets; therefore, it is perhaps not surprising that the *Wiki* and *Founta* classifiers do not perform well on the *CH-negative* class. These classifiers tend to assign the positive label to instances of the *CH-negative* class due to the presence of obscene words and expressions (e.g.,, *'any racist in america talks sh*t about asians'*). Unlike *CH-negative*, there is not much obscenity in *EA-negative*.

4 Sensitivity to Implicit and Explicit Abuse to Explain Generalizability

In Section 3, we showed that the generalizability of an abusive language classifier to a new, fine-grained category is highly dependent on the relative balance between implicit and explicit abuse in the train and test sets. 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 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, with only a small set of examples, and use these sensitivities to 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. The concept is defined with a set of *concept examples*. 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 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 $f_{emb} : \mathbb{R}^{k \times n} \to \mathbb{R}^m$, 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 $S_{C,p}(x)$:

$$\begin{aligned} S_{C,p}(x) &= \lim_{\epsilon \to 0} \frac{h(f_{emb}(x) + \epsilon \upsilon_C^p) - h(f_{emb}(x))}{\epsilon} \\ &= \bigtriangledown h(f_{emb}(x)).\upsilon_C^p \end{aligned} \tag{2}$$

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:

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

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 importance of C to the prediction.

4.2 Classifier's Sensitivity to a Concept

We define each concept C with $N_C = 100$ 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-positive-antiAsian*. We define two implicit anti-Asian concepts taken from the *EA-positive* and *CH-positive-antiAsian*, 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.

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-EA abuse: tweets labeled as implicit from *EA* dev
Implicit-CH abuse: 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.

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 three classifiers trained on *Wiki*, *Founta*, and *EA* 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.[7] 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.

[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.

[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.

[7] Note that a zero TCAV score can be due to the absence of that concept in train 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* classifier), 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.

Classifier	Concept					
	non-coherent	COVID-19	explicit anti-Asian	implicit-EA	implicit-CH	generic hate
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)
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)
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)

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

Observation 1: The *Wiki* and *Founta* classifiers are significantly more sensitive to the explicit concept than the implicit concepts. Given that the majority of examples in *CH-positive-antiAsian* are explicitly abusive and the majority of examples in *EA-positive* are implicitly abusive, the differences in sensitivities predict that the *Wiki* and *Founta* classifiers are more accurate in detecting *CH-positive-antiAsian* than in detecting *EA-positive*. This is consistent with Finding 1 in Section 3.

Observation 2: The sensitivity of the *Wiki* and *Founta* classifiers to the explicit and the generic hate concepts are higher than the sensitivity of the *EA* classifier to these concepts. The difference is significant in the case of the generic hate concept. Also, the *EA* classifier is more sensitive to the implicit-EA concept than the implicit-CH concept, suggesting that the two datasets do have different underlying distributions in content. These observations are in line with Finding 2.

Observation 3: In Table 4 we see that the *EA* classifier shows sensitivity to both explicit anti-Asian abuse and the two implicit anti-Asian concepts. Unlike the *EA* classifier, the *Wiki* and *Founta* classifiers are only sensitive to explicit anti-Asian abuse and are not able to differentiate texts about Asian entities at a nuanced level. The *Wiki* and *Founta* classifiers would label any example that includes profane words as positive due to their high sensitivity to explicit abuse. Given that the *CH-negative* class includes many texts with profane words, we would expect the general classifiers misclassify a large portion of this class. This is consistent with Finding 3, where on the *CH-negative* class we observed low recall for the general classifiers and high recall for the *EA* classifier.

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 if we define a new concept by adding one utterance to the explicit concept examples, the sensitivity of the classifier to the new concept will stay close to 1, only if the new utterance is explicitly abusive. Here, we modify Equation 1 and calculate each CAV by averaging the RoBERTa representations of $N_v - 1$ explicit concept examples, and the new utterance for which we want the degree of explicitness, x_{new}, with representation r_{new}. Thus,

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

We then calculate the average TCAV score for each x_{new} as its DoE score. If the new utterance, x_{new}, is explicitly abusive, v^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 v^p_{new} will diverge from representations of explicit abuse, and the average score will drop. We use $N_v = 3$ in the following experiments.

DoE analysis on COVID-related abusive data: We then 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 *EA* dev sets (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. We calculate the classification confidence (maximum output probability) for these examples as the baseline. Figure 1 shows the comparison of DoE with classification confidence. We observe that for both classifiers, the distribution of DoE scores of implicit examples is different from the distribution of DoE scores of explicit examples, but

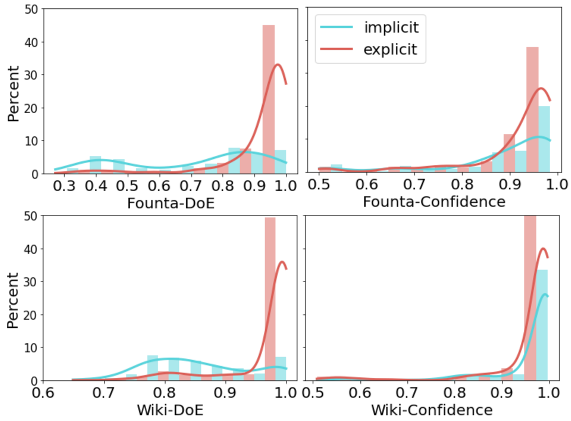

Figure 1: Comparison of classification confidence and DoE score for distinguishing between implicit and explicit abusive utterances.

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 B.

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-purpose abusive language classifier trained on *Wiki*. 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.

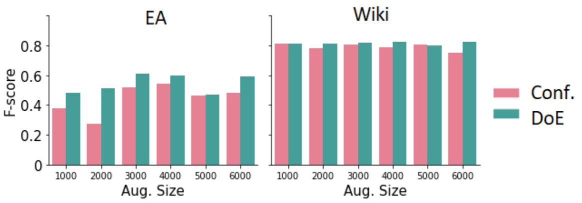

Figure 2: Classification F-score of the augmented *Wiki* classifier on the *EA* and *Wiki* test sets.

6.1 Results

Figure 2 shows the F-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 C.) 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 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 F-score of the classifier trained just on *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: F1-score of 0.61 for the DoE-based augmentation obtained with 3K added examples, and F1-score of 0.54 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 with 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 method maintains the classification performance

Method	Training Dataset	*EA*	*CH*	*Wiki*
DoE	*Wiki*+3K *EA*	**0.61**	**0.73**	**0.82**
confidence	*Wiki*+ 4K *EA*	0.54	0.71	0.79
merging data	*Wiki+EA*	0.58	0.72	0.78
baseline	*Wiki*	0.27	0.69	0.82

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

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*.

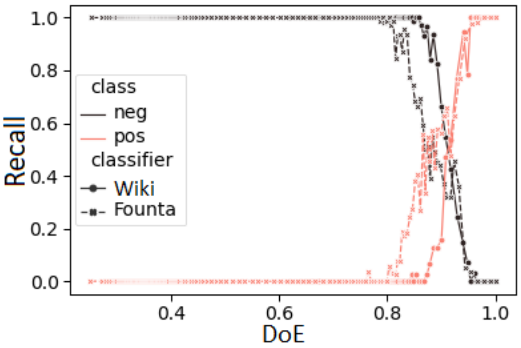

Figure B.1: Recall per class for varying DoE scores on the *EA* train set

A 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.

B 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 B.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).

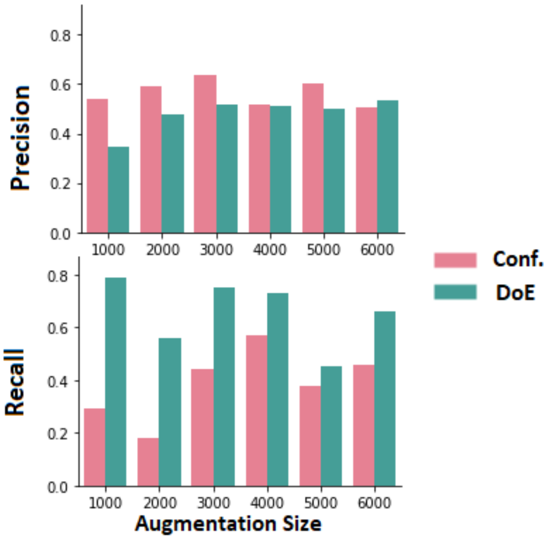

Figure C.1: Precision and recall of the augmented *Wiki* classifier on the *EA* test set.

C 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 F-score. Here, we provide a more fine-grained analysis based on recall and precision.

Figure C.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. In Appendix B, 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 'toxi-

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

city' 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.