Polarized Opinion Detection Improves the Detection of Toxic Language

John Pavlopoulos[1,3] **Aristidis Likas**[2]
[1] Department of Informatics, Athens University of Economic and Business, Greece
[2] Department of Computer Science & Engineering, University of Ioannina, Greece
[3] Archimedes/Athena RC, Greece
annis@aueb.gr arly@cs.uoi.gr

Abstract

Distance from unimodality (DFU) has been found to correlate well with human judgment for the assessment of polarized opinions. However, its un-normalized nature makes it less intuitive and somewhat difficult to exploit in machine learning (e.g., as a supervised signal). In this work a normalized version of this measure, called nDFU, is proposed that leads to better assessment of the degree of polarization. Then, we propose a methodology for K-class text classification, based on nDFU, that exploits polarized texts in the dataset. Such polarized instances are assigned to a separate K+1 class, so that a K+1-class classifier is trained. An empirical analysis on three datasets for abusive language detection, shows that nDFU can be used to model polarized annotations and prevent them from harming the classification performance. Finally, we further exploit nDFU to specify conditions that could explain polarization given a dimension and present text examples that polarized the annotators when the dimension was gender and race. Our code is available at https://github.com/ipavlopoulos/ndfu.

1 Introduction

Annotations for subjective tasks are often aggregated, to form ground truth labels and allow supervised learning algorithms to be trained for these tasks. Given a text, for example, annotations are averaged to yield binary labels reflecting whether the text is misogynous or not (Kirk et al., 2023). These annotations, however, are not always described by a single mode. Specific data items may lead to non-unimodal annotations, increasing the inter-annotator disagreement (Baan et al., 2022). This point is clearer if we consider a post classified as -1 by half of the annotators and as 1 by the other half, assuming a 3-point scale. No point is suitable to represent this item due to the two polarized ratings.

Current machine learning conventions reduce the annotations for a given text into a single label (most often, the mode) and consider the inter-annotator agreement (Artstein and Poesio, 2008) as an indicator of the quality of the ground truth, or the task difficulty. In this work, we argue that polarized annotations may be beneficial in machine learning for subjective tasks and that inter-annotator agreement is not necessarily reflective of the ground-truth quality. The negative impact of the information loss due to such aggregations can be higher when the annotators come from different social groups. Language that is offensive to specific groups at risk for discrimination will be obscured in datasets with aggregated annotations and consequently be ignored by algorithms trained on those datasets.

In this work, we focus on polarized opinions of annotators about the label to be assigned,[1] suggesting their detection prior to supervised learning, to remove ambiguous annotations and improve the classification performance. Recently, the distance from unimodality (DFU) measure has been found to to have a strong correlation with human judgment when used as an index of polarization (Pavlopoulos and Likas, 2022). Although effective, this measure is un-normalized, a fact that limits the measure's interpretability. To this end, we propose in this work a normalization which directly improves the measure. By employing the normalized DFU, then, we propose a classification methodology where we introduce a new class comprising data with polarized annotations. Despite the fact that, in principle, a new class increases the task's difficulty, our approach outperforms the binary baseline in three datasets for abusive language classification. Furthermore, the probability for the added class, assigned for a text, serves as an estimate of the polarization of the text annotations.

The contribution of this work is threefold.

[1] The same post may be classified quite differently depending, for example, on the cultural background of the annotator. Tables 4 and 6 (Appendix) show examples in the domain of toxic language detection, where this is a realistic scenario.

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1946–1958

- First, we introduce a normalized variant of the DFU measure of polarized opinions, called nDFU, that also correlates well with human judgment and allows for better interpretation.
- Second, we propose *unpolarized learning*, an approach that introduces and exploits a new class that contains polarized (not simply ambiguous) data. Experimenting on the subjective and of high social impact task of toxic language detection,[2] we show that our approach outperforms the baseline in three datasets.
- Third, we present conditions based on nDFU, which can be used to detect polarized items that are unimodally-annotated by specific groups of annotators. Using gender and race, we present texts that satisfy those conditions, attempting to explore the roots of polarization.

2 Related Work

For many NLP tasks, a diversity of valid beliefs exist about what the correct data labels should be (Röttger et al., 2021). Such tasks comprise the detection of toxic language (Sap et al., 2021; Salminen et al., 2019), harassment (Al Kuwatly et al., 2020), and stance (Luo et al., 2020). Due to the lack of measures assessing polarized opinions, however, no published work to date aimed at *detecting and classifying polarized annotations*, which is the goal of this study. Instead, the focus is more broadly on ambiguous instances (Otani et al., 2020), with current approaches reducing the number of classification labels (Campagner et al., 2021; Thierry et al., 2019), or modelling the distribution of annotations using a Gaussian distribution (Wan and Chan, 2020; Chang et al., 2020), or learning the histogram of annotations (Fornaciari et al., 2021; Prabhakaran et al., 2021).

Reducing the number of class labels Campagner et al. (2021) showed that the quality of the ground truth (e.g., the inter-annotator agreement) impacts the performance of machine learning models and should not be taken for granted. The authors studied different ways to yield a single target label from multi-rater settings, which is a common approach in supervised learning for NLP. The standard reduction method is majority voting from crowdsourced opinions or the fraction of raters who said yes (in a yes/no question), binarized. Although common, this approach fails to encode uncertainty (Thierry et al., 2019).

Uncertain ground truth Uncertainty can be tackled by considering the annotations for a data item as noisy observations that can be modeled by a Gaussian distribution (Wan and Chan, 2020). Chang et al. (2020) attempted to learn simultaneously the mean and the variance of the normal distribution showing that this approach outperforms ground-truthing methods that disregard uncertainty. Although encoding uncertainty is useful, the use of a unimodal distribution (e.g., a Gaussian) imposes severe limitations, since it disregards the possibility of polarized opinions (multiple modes). Such an assumption may be harmful in tasks with subjective opinions, such as sentiment analysis and toxic language detection, where annotators with different personal, cultural, or demographic backgrounds may perceive differently commonsense knowledge (what they will assign as target label) of the same item (Akhtar et al., 2021).

Soft labels Instead of a noisy unidimensional target label, one may attempt to learn a multivariate probability density function (i.e., the normalized histogram). Such a ground truth model allows the maintenance of polarized annotated opinions in the supervised signal when using machine learning algorithms. Peterson et al. (2019), for example, showed that predicting the whole distribution of the class annotations improves robustness in image classification. Another example is the work of Gordon et al. (2021), who encoded human disagreement to improve the quality of social computing datasets, building on prior findings showing that annotators' disagreement is not noise (Chung et al., 2019; Kairam and Heer, 2016). These studies, however, treat polarized opinions as a special case of disagreement (Prabhakaran et al., 2021).

3 Assessing Opinion Polarization

3.1 The DFU Measure

DFU estimates the extent of polarization on a distribution of annotations (opinions) and it has been defined by Pavlopoulos and Likas (2022) for an opinion histogram as the deviation from unimodality. Let a set $X = \{x_1, ..., x_n\}$ of n opinions, each of which can take K ordinal ratings: $x_i \in$

[2] We use this term universally, to cover what researchers refer to as 'abusive', 'offensive', 'hateful' or otherwise harmful. Besides the social impact of this task, texts with polarized annotations (Tables 4 and 6) and unaggregated annotations exist in this domain, making it an ideal ground for our study.

$\{O^1, ..., O^K\}$. We assume that $f = (f_1, ..., f_K)$ are the relative frequencies of the K ratings defining the opinion distribution (histogram) of X. The discrete opinion distribution f is unimodal if it has a single mode, which means that there exists a maximum value f_m and that the values f_i monotonically decrease while moving away from m. More formally, $f_{i-1} \leq f_i$ for $i < m$ and $f_{i+1} \leq f_i$ for $i > m$. DFU is defined as the maximum of the differences d_i between successive f_i values that are computed as:

$$d_i = \begin{cases} f_i - f_{i-1} & m < i < K \\ f_i - f_{i+1} & 2 < i < m \\ 0 & i = m. \end{cases} \quad (1)$$

$$DFU = max(d) \quad (2)$$

3.2 The Normalized DFU Measure

As shown in Equation 2, DFU is defined as the maximum d_i value. This is also shown in line 9 of Algorithm 1. It can be observed that $d_i \leq f_m$, which means that DFU, which is the maximum d_i, is always smaller than the highest peak of the histogram (the mode). Therefore, we can produce a normalized variant by dividing DFU with the mode f_m (line 10 of Algorithm 1).

Algorithm 1: Calculation of nDFU

Data: Opinions X: $x \in \{O_1, ..., O_K\}$
Result: A score $nDFU \in [0, 1]$
1 **for** $i = 1$ *to* K **do**
2 $\quad f_i = \frac{\sum_{x=1}^{|X|} 1^{O_i = x}}{N}$;
3 $m = \underset{i}{\text{argmax}} f$;
4 $d_m = 0$;
5 **for** $i = m + 1$ *to* K **do**
6 $\quad d_i = f_i - f_{i-1}$;
7 **for** $i = 2$ *to* $m - 1$ **do**
8 $\quad d_i = f_i - f_{i+1}$;
9 $DFU = max(d)$;
10 $nDFU = \frac{DFU}{f_m}$;
11 **return** $nDFU$

It should be noted that the special case $f_m = max(d)$ (i.e., $nDFU = 1$) occurs when at least two non-consecutive bins are of equal height (e.g., in uniform distributions). In a simple 3-point Likert scale (e.g., 'agree' and 'disagree' at the poles, 'neutral' in the middle), this case regards equal height for the bins at the poles while the bin for 'neutral' is zero. As can be seen in Figure 1, both DFU variants, normalized or not, yield a zero score for the unimodal Gaussian. The scores of the normalized variant (nDFU), however, are considerably higher, close to 1, for the multimodal Gaussian mixtures. On the contrary, the un-normalized score (DFU) is neither intuitive nor interpretable.

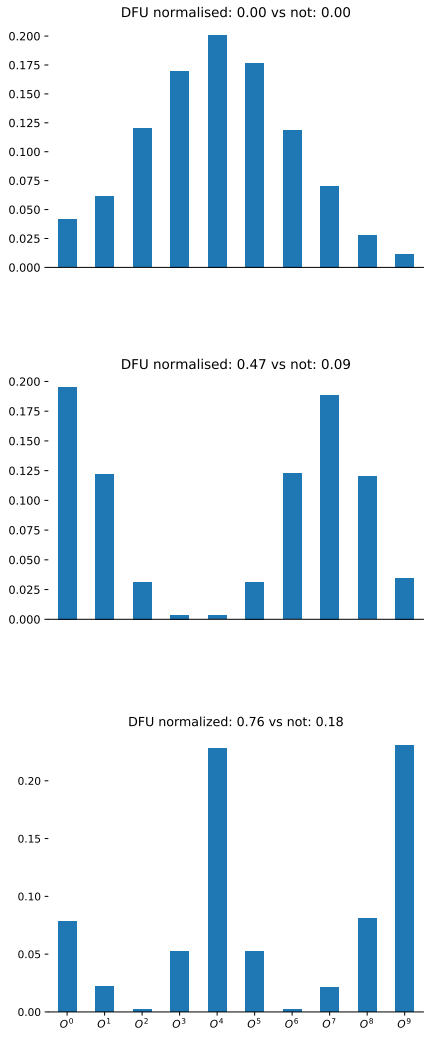


Figure 1: Histograms of three synthetic annotation distributions. Starting from the top, a unimodal, a bimodal, and a trimodal Gaussian mixture are shown, along with their corresponding DFU and nDFU values. Horizontally are the ordinal ratings, as these are defined in §3.1, and vertically are their relative frequencies.

4 Unpolarized Learning

Supervised learning is often applied to subjective tasks, such as toxic language classification, by transforming the set of ordinal annotations $X = \{x_1, ..., x_n\}$, $x_i \in \{O^1, ..., O^K\}$ into a binary label. Such a binarization is implemented by

first thresholding each annotation x_i and then applying majority voting. In other words, a threshold h is defined (where $O_1 \leq h \leq O_K$) that is used to binarize x_i (i.e., per annotator per item):

$$x_i = \begin{cases} 0, & x_i < h \\ 1, & x_i \geq h. \end{cases} \quad (3)$$

A probability is then computed, as the fraction of positive ratings, rounded (or thresholded) to produce the final binary label (i.e., the majority vote) assigned to the instance.

The assumption of binarized thresholded ratings is problematic, because items with polarized ratings will get a noisy signal. For example, assume the case where annotators rate a post's toxicity from 1 (clearly civil) to 5 (very toxic). A post that is rated as 5 by 49% of the raters and as 1 by the rest will be assigned a binary label of 0, meaning civil. Therefore, similar posts (i.e., causing polarized ratings) may end up in both binary classes, introducing noise to the dataset. We argue that introducing a $K+1$ class (e.g., a 3rd class in binary classification), comprising data with polarized annotations, is advantageous in a supervised learning setting. That is because only unimodal data will be used to learn the original K-class task while polarized items will form a class on their own. We call this strategy *unpolarized learning* because only unpolarized data (i.e., data with unimodal annotations) are used to learn the original K classes.

4.1 Training with $K+1$ Classes

In the following, without loss of generality, we assume $K = 2$, classifying an instance either to the negative (0) or the positive (1) class, and we provide more details of the proposed strategy.

First, we detect polarized items, which are the items that have an nDFU value that is greater than a threshold.[3] Unpolarized items, which are characterized by a single mode, are classified to the positive or the negative class, normally, based on majority voting. The rest, on the other hand, are not. Instead, we introduce a third ($K+1$) class label which we assign to all polarized instances. Next, we train the network for the 3-class classification task. The resulted network will learn to classify an item as positive, negative, or to the 3rd class with the polarized annotations.

In principle, training the classifier becomes harder when a class is added, reducing the accuracy of a random baseline from $\frac{1}{K}$ to $\frac{1}{K+1}$. At the same time, however, the supervised signal with which the network is learning the task becomes clearer, because each actual class is learned using items with unpolarized (unimodal) annotations. Therefore, a more accurate K-class classification is expected.

[3] A natural choice for this threshold is 0, but this is tunable.

4.2 Class Reduction at Inference

It should be noted that during inference, it is possible to exploit the $K+1$ classifier outputs in two ways. The first possibility is to refrain from assigning class labels to items that are identified as polarized (e.g., with a very high $K+1$ output value). The other possibility (considered in this work) is to ignore the $K+1$ output value and always classify an item to one of the original K classes, i.e., the one with the highest output value.

5 Datasets

We investigated one resource comprising what human experts perceive as polarized and three datasets comprising annotations for toxic language, a subjective task with high social impact. Regarding the latter, we limited our search to datasets whose annotations are released without any aggregation, i.e., one label per annotator is provided.

5.1 OPGT

The Opinion Polarization Ground Truth (OPGT) dataset was introduced by Koudenburg et al. (2021) to approximate what humans perceive as a distribution of polarized opinions. Sixty researchers of opinion polarization judged with a five-point scale the extent of polarization of fifteen opinion distributions. The average judgment per distribution was then used by Pavlopoulos and Likas (2022) to build the ground truth regarding the extent to which the participants thought that the respective histogram represented a polarized state.

5.2 Toxicity Detection

Several datasets exist for toxic language detection but the vast majority of them has only released an aggregated label (e.g., toxic) or score (e.g., 70% for being perceived as toxic) of the annotations. In this study, we opt for the two publicly available datasets that provide access to their raw annotations, viz. the Civil Comments (CCTK) and the Ex-Machina (XMACH) datasets. Also, we were granted access to another dataset (Attitudes) by Sap et al. (2021).

CCTK comprises comments posted from 2015 to 2017 on several English-language news sites.

Multiple annotators from several countries rated each post with a 4-point Likert scale, from non-toxic (68.7%), to "hard to say" (0.5%), to "toxic" (29.4%), and "very toxic" (1.5%). Pavlopoulos and Likas (2022) used this dataset to predict the not-normalized DFU score. They found that posts with high (not-normalized) DFU were annotated by people coming from more countries compared to ones with low DFU, revealing cultural context as a possible reason behind polarized opinions. We followed the authors' suggested split and we yielded a binary ground truth (when needed) by forming a single class of toxic and very toxic posts (17%).

Attitudes was introduced to study how the annotators' identities affect their text toxicity annotations (Sap et al., 2021).[4] The authors studied the annotators' race, gender, and political leaning. A small dataset was formed by giving fifteen posts to 641 participants and asking for their toxicity ratings, combined with their identities and their attitudes. The participants led to different proportions regarding their race (13% Black, 85% White), political (29% conservative, 59% liberal), and gender identities (54% women, 45% men, 1% non-binary). A 5-point Likert scale was used for the rating, from 1 (civil) to 5 (very toxic).[5] We formed the toxic class in a binary setting using posts assumed by the majority of the voters as very toxic (23%).

XMACH was developed by Wulczyn et al. (2017) who crowd-annotated 100k comments focusing on personal attacks or harassment. This was a subset of 63M Wikipedia comments from discussions relating to user pages and articles dating from 2004 to 2015. To address the imbalance of the toxic class (1%), the authors extended their resource by collecting and adding comments of users who were blocked or violating Wikipedia's policy. Five comments per user were added "around every block event", leading to an increased balance for this resource (17%) and overall (12%). A 5-point Likert scale was used for the rating, from -2 (very toxic) to 2. A binary toxic class was formed by merging toxic and very toxic posts (32%).

5.2.1 Exploratory Analysis

In Table 1, we summarise the statistics of our datasets' texts and annotations, computed on the training subsets. In all three datasets, we assume an equal train/test split. CCTK posts are the lengthier on average, followed by XMACH, and Attitudes. The latter, not only has the shortest posts, but also the fewer instances and the fewer annotations per text. XMACH, on the other hand, is the dataset with the most annotations on average per text.

	LENGTH	SIZE	CODES (#)
CCTK	309.3 (276.6)	10k	6.1 (2.8)
ATTIT.	125.4 (85.9)	313	5.6 (0.8)
XMACH	194 (128.3)	2k	8.4 (1.3)

Table 1: The average text length in characters (st. deviation), the number of train instances, and the average number of annotations (st. deviation) per dataset.

Figure 2 shows that for all three datasets the number of posts with zero nDFU is significantly greater than that of the rest. This means that the majority of posts comprise unimodal annotations. We also observe that for the two smaller datasets, there are nDFU zones for which there are no posts, as for example: $0.8 \leq nDFU \leq 0.9$.

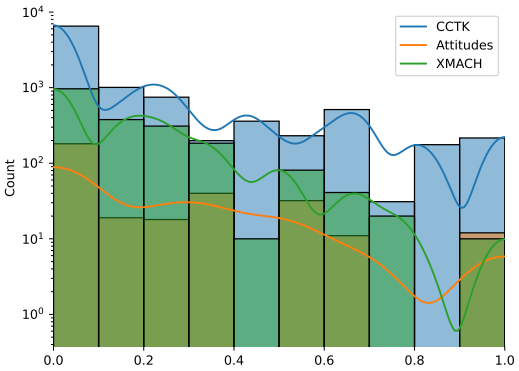


Figure 2: The number of instances (vertically, in log scale) per dataset per nDFU score (horizontally). Each dataset is represented with a different colour (one per line) but the colour of Attitudes appears only in the rightmost bar (0.9-1.0) where it exceeds XMACH.

6 Experiments

Using OPGT, we measured the correlation between our proposed nDFU and human judgment. Then, we used nDFU to introduce the additional class of polarized opinions in three toxicity datasets, and we compare the performance in toxicity detection with and without the added polarized class.

6.1 Correlation with Human Judgment

We computed the correlation between our nDFU measure and what humans perceive as a distribu-

[4]Only participants from the U.S.A. were considered to restrict the perceptions of race and political attitudes.

[5]Two criteria were used, toxic according to the annotator or to any. We used the former.

Figure 3: Histograms of fifteen opinion distributions. The average judgment of the extent to which sixty polarization experts (§5.1) thought the respective histogram represented a polarized state (Koudenburg et al., 2021) is shown in the horizontal axis (Gold). Transparency is reversely related to the respective normalized DFU score (shown in parentheses) per histogram.

tion of polarized opinions using OPGT (§5.1). Figure 3 shows the average judgment of each of these fifteen histograms, along with their nDFU score. The Pearson correlation between the score and human judgment is 0.90, which is on par with what has been reported by Pavlopoulos and Likas (2022) using the un-normalized DFU (0.89). By being limited in [0,1], the proposed measure facilitates also the tuning process, which is not straight-forward with the unconstrained (in upper limit) DFU.

In Figure 4, we show the correlation between nDFU and human judgment on subgroups of polarization experts, which were created by sampling participants, from three (on the left) to fifty (on the right) per subgroup. We can observe that a high correlation, yet less stable, is established with fewer participants in the survey (§5.1). This finding shows that nDFU is able to capture a polarized state even when only ten or fewer annotations are provided for a data item, which is most often the case in subjective machine-actionable datasets (Leonardelli et al., 2023).

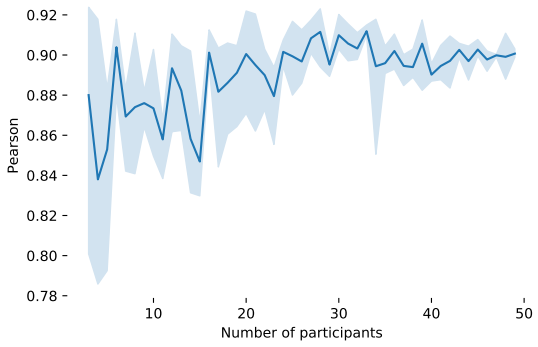


Figure 4: Pearson correlation between nDFU and subgroups of polarization experts of varying size.

	F1			AUC
	CIVIL	TOXIC	POLARIZED	
CCTK	0.82	0.13	0.58	0.80
ATTITUDES	0.49	0.35	0.54	0.65
XMACH	0.49	0.37	0.46	0.62

Table 2: F1 per class, along with one vs. rest AUC, of a BERT-based unpolarized learning classifier.

6.2 Benchmarking Unpolarized Learning

We opted for BERT features (Devlin et al., 2019), using the uncased base model, and training a logistic regression model on top of the [CLS] pseudo token.[6] Class weights were set according to the class balance of the dataset. By yielding binary toxicity labels per dataset (§5), and by introducing the class of polarized opinions, we trained and assessed this classifier. The results are shown in Table 2. The prediction of the toxic class is the most difficult task, especially in the heavily-imbalanced CCTK. The performance of predicting the polarized ($K+1$) class ranged from 0.46 (XMACH), to 0.54 (ATTITUDES), to 0.58 (CCTK) and the $K+1$ class was the easiest (ATTITUDES) or the second easiest (CCTK, XMACH) to predict among the three. In order to better understand the benefits of adding the polarized class, we experimented with a hypothesis where we ignore the predictions for the polarized class during inference, which we describe next.

6.3 Polarized Class Reduction

As discussed in §4, class reduction allows the evaluation of a $K+1$-class classifier in a K-class setting. Hence, we used it to compare our 3-class classifier

[6] This is a decent approach for classification tasks (Reimers and Gurevych, 2019, Table 5). We also experimented with fine-tuning, but that was time-consuming, especially for the two largest datasets.

	Test subset	Size	P+	R+	F1	AUC
CCTK	nDFU=0	6,712	0.59	0.33	**0.71**	**0.92**
	nDFU>0	3,287	0.63	0.18	0.54	0.71
ATTIT.	nDFU=0	89	0.45	0.39	0.62	**0.75**
	nDFU>0	68	0.50	0.43	0.62	0.73
XMACH	nDFU=0	472	0.63	0.67	**0.76**	**0.84**
	nDFU>0	528	0.58	0.62	0.71	0.79

Table 3: Precision and Recall of the toxic class, macro-F1, AUC in binary toxicity classification of a BERT baseline assessed on test data with zero (unimodal) and non-zero (non-unimodal) nDFU.

(§6.2) with a binary classification baseline.

The baseline is a binary classification model, that is the same BERT-based logistic regression classifier we used for unpolarized learning, but trained to classify a text as civil or toxic, which is a typical approach in this field (Hartvigsen et al., 2022; Zhou et al., 2023). To assess this classifier, we focused on evaluation posts with both, zero and non-zero nDFU and we present the results in Table 3. In all datasets, the classifier performs equally or better with unpolarized data, with a more clear difference in CCTK. The performance drop on polarized data (i.e., AUC being consistently lower when nDFU is positive) shows that they set a harder classification target, probably explained by the fact that their ground truth is formed by aggregating polarized opinions, i.e., far away from the two edges.[7]

The reduced predictions of our unpolarized learning method were compared to those of the binary baseline, but we tuned the threshold above which a text is classified to the polarized ($K+1$) class.[8] We opted for a development set per dataset to select the optimum threshold, based on the macro-averaged F1 when performing the class reduction step for the binary evaluation (Appendix A.2). Then, we sampled randomly test texts, comparing the predictions of the binary baseline with the reduced ones provided by our tuned model. A one-sided Mann-Whitney test (Mann and Whitney, 1947) showed that the latter had a better performance with a statistically significant difference across datasets ($p < 0.05$).[9] On average, the macro-averaged F1 score increased from 0.58 to 0.64 for Attitudes (+6 percent units), from 0.77 to 0.78 for XMACH (+1), and from 0.71 to 0.72 for CCTK (+1). Putting this result in a wider context, unpolarized learning has led to a better binary classification outcome.

7 Polarized Class Prediction Analysis

As shown with the reduced class hypothesis (§6.3), unpolarized learning can lead to a performance improvement in binary toxicity classification. Networks trained using the unpolarized learning strategy, also provide an additional benefit, which is the ability to estimate the probability for the $K+1$ class. In other words, for a new text input the $K+1$ output estimates the *probability that the text is going to receive polarized annotations*. In order to better assess the ability of the model to provide such predictions, we used the polarized class probability along with the model-agnostic explainability framework of Ribeiro et al. (2016), which has been found to be the best option for text classification tasks (Jeyakumar et al., 2020).

Local interpretable model-agnostic explanations (LIME) approximate the effect of features (words) on the model's output by training local surrogate models. The words suggested as explanations were not always easy to interpret or distinguish from the other two classes. One example from the CCTK dataset is "I'm not black, but there's a whole lotta times I wish I could say I'm not white - frank zappa", where the words 'black' and 'white' contribute toward the decision for the $K+1$ class while the surname of Frank Zappa contributed reversely.

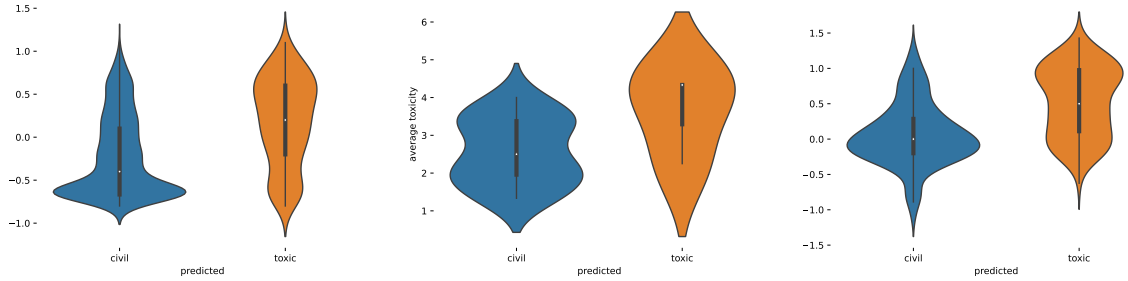
To gain more insights into the $K+1$ class of polarized annotations, we used error analysis as a proxy. A common mistake of models trained with unpolarized learning (i.e., using posts with non-zero nDFU to define the $K+1$ class) concerns the misclassification of $K+1$ posts to the civil class (confusion matrices in Appendix A.3). This information, however, is not useful on its own. Therefore, we focused solely on posts of the $K+1$ class, exploring their average toxicity without the step of binarization, which we analyze next.

As is shown in Fig 5, posts of the $K+1$ class that are misclassified as civil (in blue, on the left) are often annotated as civil by the annotators. On the other hand, posts of the $K+1$ class that are misclassified as toxic (in orange, on the right) are annotated more often as toxic. In other words, the

[7] In Appendix B, we discuss an alternative nDFU-based binary classification method that may perform on par while significantly reducing training time.

[8] We did not tune the classification threshold neither for the binary nor the $K+1$ classifiers. We only tuned the number of high nDFU posts removed from the training data. Doing a sanity check with the (small) Attitudes dataset and a Random Forest binary classifier, tuned from 0 to 0.9 with step 0.1, yielded 0.5 as the best threshold.

[9] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html



(a) CCTK (b) ATTITUDES (c) XMACH

Figure 5: Violin plots of the non-binarized average toxicity (vertically, higher means more toxic) of the $K+1$ class ($nDFU > 0$) that were predicted as civil (left, in blue) or toxic (right, in orange).

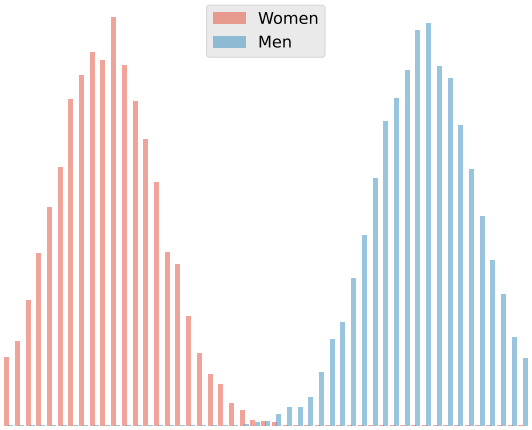


Figure 6: Synthetic bimodal histogram of annotations, where annotators agree only conditioned on their gender.

annotations of these posts were considered polarized (hence, the $K+1$ class) but the mode of the annotations was aligned with the model's prediction. This explains why unpolarized learning has led to a well-performing model in binary toxicity classification despite the burden of an added class.

8 A Posteriori Unimodality Explanations

Polarization may be due to various reasons, such as political beliefs, social dimensions, gender, age, etc. Although nDFU estimates polarization (§3.2), it does not suggest its cause. As an example, a bimodal annotation histogram is shown in Figure 6, where the colour reflects the gender of the annotator. Although polarization is easily estimated in this histogram, its root cause (here, gender) is not revealed. To address this important issue, we next propose an approach that could possibly explain polarization given a dimension.

Let the set of opinions X of Algorithm 1 for a non-zero nDFU post, and let G be the values for a dimension D that characterizes the opinion holder i, that is $D_i \in \{d^1, ..., d^G\}$.[10] Based on the value of D corresponding to each annotator, the set X can be partitioned into G subsets X^k $(k = 1, ..., G)$. This means that each subset X^k contains the annotations of the annotators with value d_k. Then, we consider that D explains the polarization of X ($nDFU(X) > 0$), if the following *a posteriori unimodality* conditions hold:

$$nDFU(X^k) = 0, \forall k = 1, \ldots, G \qquad (4)$$

In the simple case of Figure 6, where only male and female annotators are considered, let X^1 denote the set of annotations from men (red histogram) and X^2 the set of annotations from women (blue histogram). It is obvious that $nDFU(X^1) = 0$ and $nDFU(X^2) = 0$ since each of them forms a unimodal histogram. Consequently, it can be inferred that gender constitutes the source of polarization in the whole set of opinions X.

We explored CCTK posts that satisfy Equation 4 regarding two dimensions, viz. gender and race.[11] Table 4 presents two posts resulted from this exploration. The first claims that homosexuality deviates from normality. LGBTQ annotators considered this post toxic, but annotators from the control-group did not. The second post, touched on a racist stereotype, considered as civil by African Ameri-

[10] We assume a single dimension for simplicity, but a combination of dimensions could be used as well.

[11] We used the richer re-annotation of Goyal et al. (2022), who collected annotations for posts from three groups of annotators: African American, LGBTQ, and a control group.

Source	Toxic	Civil	Text
CCTK	LGBTQ	CNTRL	Homosexuality, transgender sex-all are deviations from normal human behavior as well.
CCTK	CNTRL	AA	In other words, he's not brown enough for ya to be labeled a terrorist.

Table 4: Posts satisfying Equation 4 found as toxic (civil) only by LGBTQ, African American (AA), or control-group (CNTRL) annotators.

can annotators, but not from the control group. A more thorough analysis of the detected posts and beyond (more examples are shown in Table 6 of the Appendix) is left for future work.

9 Ethical considerations and bias

The detection of toxic language can be susceptible to data and algorithmic biases, unfairing underrepresented groups or preferring over-represented perspectives. Our work makes a step toward recognising polarized opinions (nDFU) and toward understanding the poles in terms of annotator characteristics (aposteriori unimodality). Ignoring polarized instances during inference can improve the classification performance, but it may also perpetuate or reinforce existing biases.

In a real-world application Bias perpetuation/reinforcement may be addressed by employing methodologically aposteriori unimodality testing. That is, it could be used to highlight topics in which the opinions of annotators from underrepresented groups (e.g., at risk for discrimination) deviate from those of annotators from other groups. For posts related to these topics, then, more annotators could be added (perhaps from focused groups), which can potentially lead also to debiasing.

10 Conclusions

In this study we have focused on DFU, a measure that correlates well with human judgment for the assessment of polarized opinions. We have presented a normalized version, called nDFU, which not only correlates well with human judgment but is also more intuitive and interpretable that is important for tuning purposes. Using nDFU, we suggested the unpolarized learning method for text classification, which introduces a new class that contains the items detected as polarized. In this way the original classes are trained using unimodal (unpolarized items) and classification performance is improved. Experimenting with toxic language detection, an important and challenging task due to the subjective annotations, we showed that it outperforms the baseline with a statistically significant difference.

Finally, besides estimating polarization, we have shown that nDFU can also be used to trace the possible cause of polarization, by checking aposteriori unimodality conditions. Putting gender and race under the microscope, we presented texts per feature for which annotators were polarized only in an inter-dimension setting. In future work, we will apply aposteriori unimodality to more datasets, developing a corpus of polarized texts, and facilitating the study of polarization. Also, extensions of unpolarized learning will be investigated, exploring the path towards more accurate and fair NLP.

Limitations

- The proposed approach is potentially applicable to any classification task with subjective annotations (e.g., sentiment analysis). The experiments of this study, however, were limited regarding the modality (text input), the language (English), and the domain (toxicity). Future work will investigate such extensions.
- Aposteriori unimodality (§8) has already revealed posts with polarized annotations (Tables 4 and 6), but their analysis is limited in this study. A thorough investigation of each such post should follow, by also taking into consideration the post's context (e.g., conversational) in order to draw more robust conclusions regarding the roots of polarization.
- The application of unpolarized learning and aposteriori unimodality requires datasets with un-aggregated annotations. Such datasets, however, are scarce. In future work, we will investigate whether the ATTITUDES dataset can also become publicly available, assisting towards that end with one more dataset.

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *arXiv preprint arXiv:2106.15896*.

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias

based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernández. 2022. Stop measuring calibration when humans disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915.

Andrea Campagner, Davide Ciucci, Carl-Magnus Svensson, Marc Thilo Figge, and Federico Cabitza. 2021. Ground truthing from multi-rater labeling with three-way decision and possibility theory. *Information Sciences*, 545:771–790.

Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. 2020. Data uncertainty learning in face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5710–5719.

John Joon Young Chung, Jean Y Song, Sindhu Kutty, Sungsoo Hong, Juho Kim, and Walter S Lasecki. 2019. Efficient elicitation approaches to estimate collective crowd answers. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–25.

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

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, Massimo Poesio, et al. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Nitesh Goyal, Ian D Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–28.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, and Mani Srivastava. 2020. How can i explain this to you? an empirical study of deep neural network explanation methods. *Advances in Neural Information Processing Systems*, 33:4211–4222.

Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowd-sourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1637–1648.

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. Semeval-2023 task 10: Explainable detection of online sexism. *arXiv preprint arXiv:2303.04222*.

Namkje Koudenburg, Henk AL Kiers, and Yoshihisa Kashima. 2021. A new opinion polarization index developed by integrating expert judgments. *Frontiers in psychology*, page 4575.

Elisa Leonardelli, Alexandra Uma, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, and Massimo Poesio. 2023. Semeval-2023 task 11: Learning with disagreements (lewidi). *arXiv preprint arXiv:2304.14803*.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting stance in media on global warming. *arXiv preprint arXiv:2010.15149*.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Naoya Otani, Yosuke Otsubo, Tetsuya Koike, and Masashi Sugiyama. 2020. Binary classification with ambiguous training data. *Machine Learning*, 109:2369–2388.

John Pavlopoulos and Aristidis Likas. 2022. Distance from unimodality for the assessment of opinion polarization. *Cognitive Computation*, pages 1–8.

Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9617–9626.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:2110.05699*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B Pierrehumbert. 2021. Two contrasting data annotation paradigms for subjective nlp tasks. *arXiv preprint arXiv:2112.07475*.

Joni Salminen, Hind Almerekhi, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J Jansen. 2019. Online hate ratings vary by extremes: A statistical analysis. In *Proceedings of the 2019 conference on human information interaction and retrieval*, pages 213–217.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

Constance Thierry, Jean-Christophe Dubois, Yolande Le Gall, and Arnaud Martin. 2019. Modeling uncertainty and inaccuracy on data from crowdsourcing platforms: Monitor. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 776–783. IEEE.

Jia Wan and Antoni Chan. 2020. Modeling noisy annotations for crowd counting. *Advances in Neural Information Processing Systems*, 33:3386–3396.

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.

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D. Hwang, Swabha Swayamdipta, and Maarten Sap. 2023. COBRA frames: Contextual reasoning about effects and harms of offensive statements. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6294–6315, Toronto, Canada. Association for Computational Linguistics.

Appendix

A Unpolarized learning

A.1 Benchmark

Table 5 presents the Precision and Recall of the binary baseline, assessed on unimodal and non-unimodal evaluation data.

	P(uni/non)	R(uni/non)
CCTK	0.59/0.63	0.33/0.18
ATTITUDES	0.45/0.50	0.39/0.43
XMACH	0.63/0.58	0.67/0.62

Table 5: Precision and Recall in binary classification of the BERT baseline, assessed on evaluation data with zero (unimodal) and non-zero (non-unimodal) nDFU.

A.2 Tuning

We sampled 500 posts per threshold for CCTK and XMACH and 50 for the smaller ATTITUDES. We repeated the experiment ten times to compute 95% confidence intervals. We only used zero nDFU posts, which are clearly correct. Similar results but on a smaller scale were observed for multimodal data. The green solid line in Figure 7 depicts the F1 of the model trained with unpolarized learning for the different thresholds when we ignored the predictions to the K+1 class during inference (i.e., class reduction). The optimum threshold in our study was between 0.4 and 0.5, but this depends on the fraction of posts with polarized annotations and is expected to vary across datasets and depend on the annotators.

A.3 Confusion

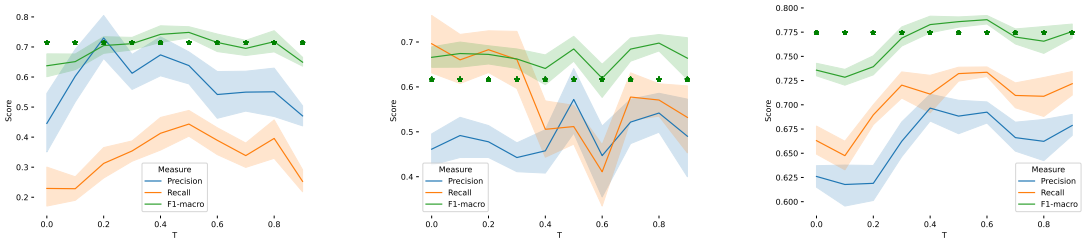
By focusing on the second row of each confusion matrix in Figure 8, we observe that K+1 posts are often (mis)classified as civil.

B Binary Classification with nDFU

The current binary classification formulation uses all the data, inferring a label for polarized annotations. Discarding high nDFU posts from the binary classifier's training data, however, sets another possible nDFU-based method. Our experiments in this direction showed that high-nDFU posts confuse the binary classifier. That is, by removing from 35% (CCTK) or 50% (Attitudes, XMACH) of the training instances (speeding up considerably training time), the performance remains the same in two out of three datasets (i.e., Attitudes, XMACH). Further investigation of this method is left for future work.

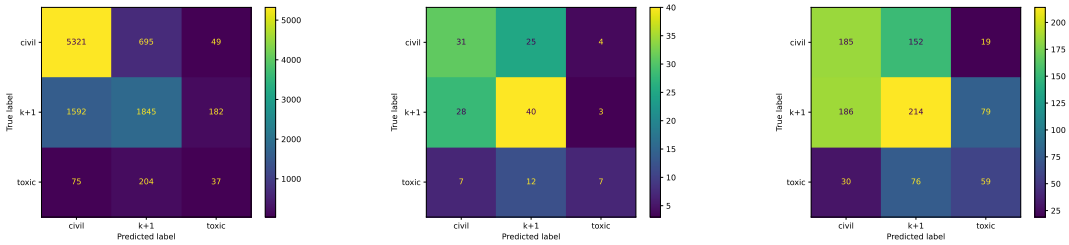
C Aposteri unimodal CCTK posts

Table 6 presents CCTK posts (§5) using the re-annotations provided by (Goyal et al., 2022) which come from three groups of annotators. Five annotators were African American, five were from the



(a) CCTK (b) ATTITUDES (c) XMACH

Figure 7: Precision, Recall, and macro-averaged F1 (in green) of a text classifier trained with the unpolarized learning approach for different thresholds T (horizontally). Predictions to the K+1 class ($nDFU > T$) are ignored during inference. The F1 score of a binary text toxicity classifier (ignoring the K+1 class) is shown in stars.



(a) CCTK (b) ATTITUDES (c) XMACH

Figure 8: Confusion matrix for our BERT-based text toxicity classifier trained with unpolarized learning where the K+1 class is defined for posts for which $nDFU > 0$.

LGBTQ community, and five were selected beyond these groups to act as a control group.

Source	Toxic	Civil	Text
CCTK	CNTRL	AA	Nothing in the Bible about Jesus saying anything about immigration rights. [...] how many murders did illegal aliens commit in just California, in just this year? Just California. How many murders by illegal aliens? You answer that.
CCTK	CNTRL	AA	Are you high??? Selling someone a cake? Refusing to sell some a cake because it is for a gay wedding is no different from selling someone a cake because it is for a mixed race wedding or a Muslim wedding or any other kind of wedding. IT IS DISCRIMINATION! Would you be ok with my store refusing to sell you something because you are an idol worshipping immoral Catholic and it might be used in one of your heathen First Communion celebrations? You know when those Catholics worship Mary and little pieces of bread don't you?
CCTK	AA	CNTRL	How do you know he is not proud of his cakes? Artists do not take pride in their work? Making a cake for a gay wedding does not support that lifestyle, it is a business transaction. Period. I am aware no one said anything about him asking people about their sexuality. I am sorry that was hard for you to understand. Is he going to ask everyone that comes in if the cake is for a gay wedding? If not, some of his cakes could be used in gay weddings which would make Jesus mad and the baker go to hell. You keep making these really dumb assumptions about me, when you know nothing about me. I am not confused, you are rude. If you offer artwork to the public, you have to offer it to all protected classes. Why would black people be discriminated against? Precedent. Ridiculous? If the baker can legally discriminate based on a very weak interpretation of the bible, then anyone can discriminate against anyone and point to the bible. Satanists can discriminate against Christians...
CCTK	AA/LGBTQ	cntrl	well thats a no brainer hillary clinton gave huma abdein a security clearance when she has ties to a known terrorist group the muslim brotherhood, and her mother runs an anti american news paper in the middle east, debbie washed up crook shultz got the awan famaily security clearances and they were recent immigrants, had absolutely no IT experience and possible ties to terrorist groups in pakistan. its pretty clear our liberal ran government is a complete and total failure when it comes to national security. 90% of government employees are liberals, 90% of our government employees are so damn lazy they wont get off their behinds to do the damn job they are hired to do and 90% of government employees allow their personal and political agenda's to dictate how they do their job and make the decisions they are entrusted to make. our government needs a douche and all public employees sent to the unemployment line union contracts negated and the whole thing started over again with out union
CCTK	LGBTQ	cntrl	All men are sex offenders? Really? A sexual predator is a person who attacks a victim. Typical men don't rape or use force on women. You are obviously a person who hates men and or healthy, normal sex.

Table 6: CCTK posts from Goyal et al. (2022) that satisfied Equation 4 and which were found as toxic (civil) only by LGBTQ, African American (AA), or control-group (CNTRL) annotators.