Recalibrating classifiers for interpretable abusive content detection

Bertie Vidgen
The Alan Turing Institute

Sam Staton
University of Oxford

Scott A. Hale
University of Oxford

Ohad Kammar
University of Edinburgh

Helen Margetts
The Alan Turing Institute

Tom Melham
University of Oxford

Marcin Szymczak
RWTH Aachen University

Abstract

We investigate the use of machine learning classifiers for detecting online abuse in empirical research. We show that uncalibrated classifiers (i.e. where the 'raw' scores are used) align poorly with human evaluations. This limits their use for understanding the dynamics, patterns and prevalence of online abuse. We examine two widely used classifiers (created by Perspective and Davidson et al.) on a dataset of tweets directed against candidates in the UK's 2017 general election. A Bayesian approach is presented to recalibrate the raw scores from the classifiers, using probabilistic programming and newly annotated data. We argue that interpretability evaluation and recalibration is integral to the application of abusive content classifiers.

1 Introduction

Computational tools for automatically detecting and categorizing abusive online content are now widely used for content moderation, to enforce and monitor regulatory and legal standards, and to study the dynamics of online abuse (Williams, 2019; Vidgen et al., 2019; Fortuna and Nunes, 2018). These tools enable abusive content to be assessed quantitatively, scalably and efficiently.

Recent research has drawn attention to several biases with existing classifiers and the datasets they are trained on, such as racial biases (Davidson and Weber, 2019; Sap et al., 2019), and evidence that they may be more attuned to detecting abuse against certain targets than others (Garg et al., 2019). Other research shows that existing tools can be fooled by 'obfuscatory' content, in which small changes are made so that the abuse is 'masked', even though it is clearly discernible to humans (e.g. changing 'niggas' to 'n!gg@z') (Gröndahl et al., 2018). Equally, many classifiers struggle with contextual statements, irony, humour and content that is non-abusive but 'incivil' (Vidgen and Derczynski, 2020).

Evaluating the explainability of classification systems for online abuse detection has become an important focus of research (Aluru et al., 2020; Wang, 2018; Švec et al., 2018). Explainable classifications can help to ensure systems are accountable, social biases are identified and addressed, and that model performance and generalisability is improved (Wachter et al., 2017; Biran and Cotton, 2017; Doran et al., 2018). In practice, explainability often requires statistical modelling to uncover the complex interactions between different input features that led to a result (Ribeiro et al., 2016). It may even require complete rethinking of how classifiers are developed, given the difficulties of post-hoc rationalisation and the potential for poor explanations to confuse end users (Rudin, 2018).

A related but previously under-researched problem in abuse classification is whether the scores returned by classifiers meaningfully encode differences in the likelihood that content is abusive. This issue can be seen as a problem of *interpretation*. In contrast with explainability this does not involve showing *why* a particular classification is given but, rather, ensuring that the classification itself is presented in understandable terms (Gilpin et al., 2019; Narayanan et al., 2018). To our knowledge only one piece of research has investigated this problem. The Perspective team at Jigsaw calibrated the scores of their toxicity classifier using isotonic regression (PerspectiveScoreNorm). The full details of the method are not published, but we understand that this calibration is primarily intended to ensure that, even as the production models are updated, the threshold of 0.8 remains a useful cutoff which for content moderation.

Ensuring that the scores from abusive content classifiers are interpretable would pose several benefits. First, the actual probabilities returned by

Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science, pages 132–138

https://doi.org/10.18653/v1/P17

models can be used for empirical analyses, reducing the information lost from only using a categorical label decided by a threshold. This is crucial in cases where a lot of content lies near the cutoff, and is the primary motivation behind this work. Second, well-calibrated scores could help host platforms to curate and filter content. For instance, companies may only want their adverts near content that has a 99.99% chance of *not* being hateful. Well-calibrated models could be used to ensure this. Third, users who want to understand why their content has been taken down (or not) may want to review the scores. If they are poorly calibrated it could generate distrust and confusion.

We investigate the use of machine learning classifiers for detecting and analysing online abuse in empirical research. We present three contributions. First, we show that the scores from uncalibrated abusive content classifiers align poorly with human evaluations. Second, we present a method for recalibration which uses probabilistic programming, which also gives an indication of the confidence in the recalibration. Third, we show that not using a calibrated classifier can severely impact empirical analysis through a case study of abuse directed against MPs in the 2017 general election. All of our code and data is made available for other researchers to use.[1]

2 Research design

We evaluate the toxicity classifier from Perspective and the hate speech classifier provided by Davidson et al. (Davidson et al., 2017). Both classifiers, and the datasets they were trained on, have been extensively researched in machine learning and computational social sciences (e.g. Sap et al., 2019; Gröndahl et al., 2018; Davidson and Weber, 2019) and the Perspective classifier is widely used for content moderation. Better understanding of their limitations and flaws will help to inform responsible use of them , and support development of better systems in the future. To examine the classifiers, we study tweets directed at candidates on Twitter in the run up to the 2017 UK general election. We collected all mentions and replies to the 2,620 candidates from 16 May to 8 June 2017, creating a dataset of 8.93 million tweets. We apply both classifiers to the dataset, from which we construct two samples to be annotated.

2.1 Annotation of tweets for recalibration

Sample 1 contains 1,000 tweets, sampled uniformly from the probability distribution for the Perspective classifier scores (i.e. with 50 tweets from each 0.05 increment). Sample 2 contains 1,000 tweets, sampled uniformly from the probability distribution for the Davidson et al. classifier scores, also with 50 tweets from each 0.05 increment. The 1,000 tweets in each sample were given to annotators with the definitions of toxicity and hate provided by the original authors.

Annotators were not given the classifier scores, tweets were presented in random order, and they were not told the distribution of scores. Each sample was annotated by 5 different independent annotators (i.e. 10 in total). Annotators had all taken part in at least 6 weeks of hateful content annotation as part of other projects, and received 2 additional weeks of training and underwent regular discussion/training sessions. Annotators were all fluent in English (8 out of 10 were native speakers). They were equally split between male and female genders, all aged between 18 and 30, university educated and from a range of European countries.

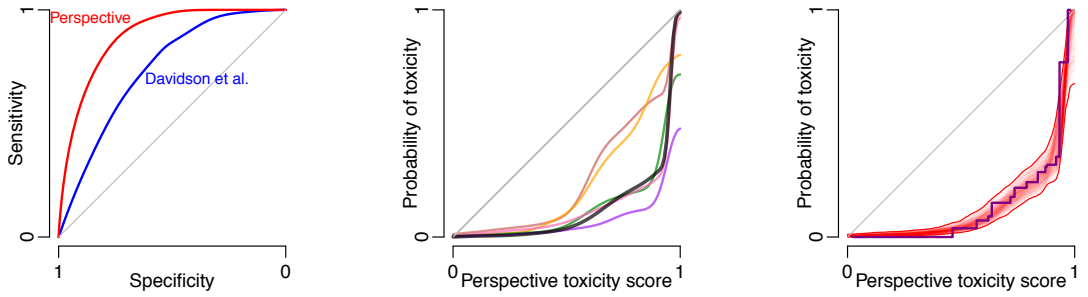
Annotators differ in their perception of toxicity and hate. For Sample 1, the number of toxic tweets identified by annotators ranges from 12 to 103 and the inter-rater reliability, as measured by Fleiss' Kappa, is 0.37. For Sample 2, the number of hateful tweets identified by annotators ranges from 68 to 243 and the inter-rater reliability is 0.46. These low to moderate levels of agreement are in line with other annotation studies of abusive content, reflecting the difficulty of such tasks (Vidgen et al., 2019). They also reflect the relatively short guidelines provided by the creators of the Davidson classifier (Davidson et al., 2017) and the fact that for the Perspective classifier we had only the definition of toxicity: "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."[2] This gives considerable scope for annotators' interpretation of the content to differ.[3]

Figure 1a shows the receiver operating characteristic curves for the classifiers from Perspective and Davidson et al. over our annotated 1,000 tweet samples. The label for each tweet is decided by taking the majority vote across the 5 annotators. The AUC is 0.899 for Perspective and 0.745 for Davidson.

[1] See: https://zenodo.org/record/4075461#.X4Cc9i2ZOu4

[2] https://www.perspectiveapi.com/#/home

[3] See our online appendix for full annotation instructions.



(a) ROC curves (b) Individual & majority calibration (c) Stan vs isotonic regression

Figure 1: (a) Receiver operating characteristic curves for the Perspective (Toxicity) and Davidson et al. classifiers on our samples. (b) Different recalibration curves for Perspective's toxicity classifier (colours for each individual and the majority vote in black). (c) The posterior distribution over calibration curves found by the Stan model (red), versus a standard piecewise-constant isotonic regression (purple).

3 Recalibrating the classifiers

3.1 Bayesian recalibration

In general, to recalibrate a classifier, one finds a recalibration curve that minimizes a particular cost function. The cost function is parameterized by true/false annotations $(a_1, \ldots, a_n)$ and classifier outputs $(p_1, \ldots, p_n)$, and it associates to each recalibration candidate $f : [0,1] \to [0,1]$ a cost (e.g. (Niculescu-Mizil and Caruana, 2005)). This has received considerable attention in machine learning and NLP research (Guo et al., 2017; Pleiss et al., 2017; Nguyen and O'Connor, 2015). We consider two methods for recalibration:

- isotonic regression as implemented in R and scikit-learn (Isotonic regression in R; SciKit), which finds a piecewise constant isotonic function minimizing the Brier score: $\frac{1}{n}(\sum_{a_i=\mathsf{true}}(f(p_i))^2+\sum_{a_i=\mathsf{false}}(f(1-p_i))^2)$;
- a custom spline regression which uses Stan's Hamiltonian Monte Carlo simulator (Carpenter et al., 2017) to maximize the log of the likelihood of observing the true/false annotations $\left(\prod_{a_i=\mathsf{true}} f(p_i) \cdot \prod_{a_i=\mathsf{false}} f(1-p_i)\right)$ with respect to an uninformative prior distribution over splines.

To decide the true/false labels for recalibration, we take the majority vote from the five annotators for both Samples 1 and 2. (Our majority vote is intended to balance the background, identity and training of annotators, but we note that more advanced methods such as MACE (Hovy et al., 2013) could be used.)

Figure 1b shows a curve fit to the annotations provided by each annotator for Sample 1 against the classification scores returned by Perspective's toxicity classifier, using the spline regression (in Stan). The bold black curve is the recalibration curve from a majority vote. A well-calibrated curve lies close to the diagonal line, which means that the inferred probabilities are similar to the classifier's scores. All annotators give scores which are substantially lower than the Perspective classifier. We observed a similar result for the Davidson et al. classifier, which is not shown for space. This suggests that using the classifiers' raw scores will lead to an overestimation of the probability that content is abusive.

The two methods give similar results; Fig. 1c shows that the isotonic regression lies within the Stan confidence interval. Isotonic regression is faster to implement. However, the Stan implementation is better motivated from several perspectives. The maximization of log-likelihood is better motivated statistically, and for this application a focus on smooth recalibration curves is more useful. There are other calibration methods that use log-likelihood, such as Platt's recalibration (Platt, 1999) and temperature-based recalibration (Guo et al., 2017). Since our method is fitting a spline, rather than a sigmoid function, it is yet more flexible than these approaches. Moreover, our Stan implementation, being Bayesian, gives us a posterior distribution over possible calibration curves, indicating how confident we can be in the choice of calibration with the given annotations.

3.2 Recalibrated abuse

Figure 2 presents the recalibrated curves for both Perspective and Davidson et al. It shows that the original classifiers match poorly with human inter-

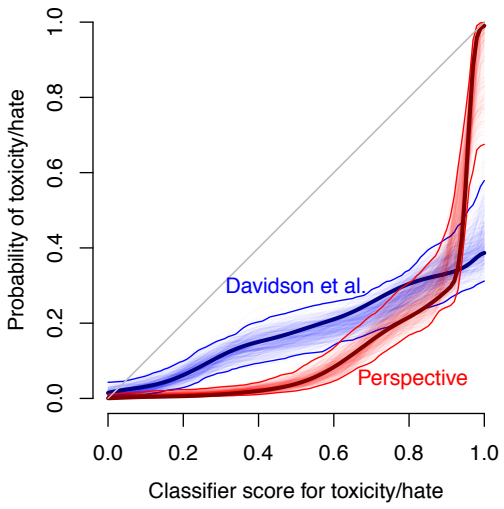


Figure 2: Recalibration curves for the Perspective and Davidson et al. classifiers. For each, the bold curve is maximum likelihood, and the 1%–99% interval is indicated.

pretations. For instance, a score of approximately 0.7 from Perspective aligns with only an inferred actual 0.2 probability of toxicity. The confidence interval is very tight for low toxicity scores, which is because nearly all the low-scored tweets were annotated as non-toxic; the confidence is less tight for low Davidson et al. scores because several low-scored tweets were annotated as hateful.

The Perspective classifier needs greater recalibration in the lower range of values than Davidson et al. but it has far better coverage of the inferred actual probabilities in the upper range. Notably, at no point does the inferred 'true' probability of abuse for the Davidson et al. classifier exceed 0.4. This suggests that, at least for this use case, the Davidson et al. classifier is a flawed way of measuring hate. This low upper limit may reflect how Davidson et al. constructed their training dataset, which involved sampling content through keywords and has been shown to contain several biases (Wiegand et al., 2019). Likely, the keywords and linguistic strategies used to express abuse differ in this new setting (i.e. tweets directed against UK candidates in the 2017 general election) compared with what the classifier was trained on.

The choice of 1,000 tweets is somewhat arbitrary, and we examine how the number of annotations that are used impacts recalibration (retaining a uniform distribution over the classification probabilities). Figure 3 shows how the maximum 98% confidence interval from Stan decreases as the number of annotations increases, showing the benefit of having more annotations. However, it also indicates that even relatively few annotations can be used for recalibration, with the rate of improvement slowing by 1,000 annotations. Confidence also varies across the calibration curves, with lower confidence in regions with greater annotator disagreement (approximately, in the 0.5 to 0.8 range for the Perspective classifier). It could be worth targeting annotator's efforts to these areas.

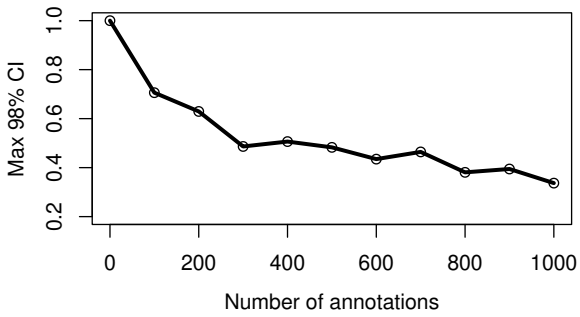


Figure 3: The maximum 98% confidence interval slowly decreases with the number of annotations.

As a further validation, we held out 20% of the annotated tweets and found a calibration curve f for the remaining 80%. The Brier score for the uncalibrated held-out tweets was 0.25, but it fell to 0.06 after recalibrating with f, a vast improvement.

4 Analysis of online abuse in the 2017 UK election

To illustrate the importance of classifier calibration, we look at the temporal dynamics of abuse directed at two successful candidates in the 2017 UK election, Ivan Lewis and Diane Abbott. We show that recalibration is important for understanding the dynamics of abuse, such as when abuse 'events' take place, especially for candidates with a low volume of tweets. In this section we focus only on Perspective's toxicity classifier.

Figure 4 plots the number of toxic tweets directed at Lewis, using different thresholds (set at 0.5, 0.6, 0.7 and 0.8). As expected, the estimated prevalence of toxic tweets directed at Lewis depends on where this arbitrary threshold is set—he received only two tweets with toxicity > 0.8, but 10 with toxicity > 0.7, a 5-fold increase. More concerningly, the thresholds give a very different view of *when* he receives abuse. In our dataset, this is a problem for all candidates who receive few tweets. More broadly, this problem will occur for

any segments of a dataset (e.g. groups, individuals or time periods) which have few entries.

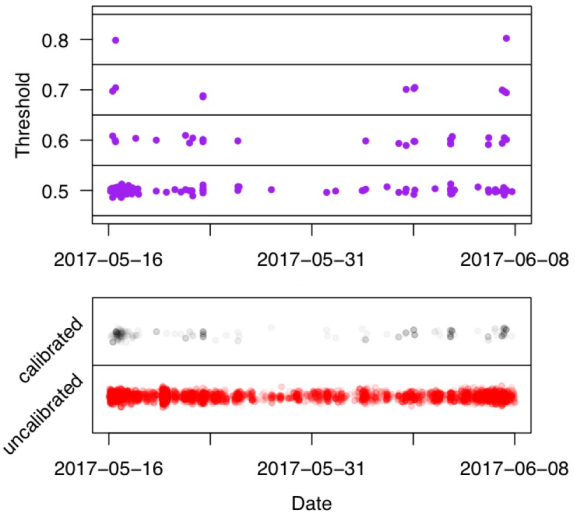


Figure 4: Timelines of toxicity received by Ivan Lewis. The upper panels use thresholds, the lower panels use the raw toxicity scores and the recalibrated scores. Each dot represents a toxic tweet. Opacity is proportional to the score. We deseason by stretching time by overall volume and use a vertical random jitter.

An alternative to applying a threshold is to directly analyse the classification probabilities. This is far more desirable as it is less biased and leads to less information loss. It can be achieved by summing the toxicity probabilities within each unit of time (e.g. every hour). The principle behind this is that if there are 50 tweets each with 0.2 probability of toxicity then it should be likely that ∼10 will be toxic — and summing the probabilities best capture this. However, as the second panel of Figure 4 shows, the uncalibrated scores substantially overemphasize non-toxic tweets, and using them for this purpose would inflate the estimated prevalence of abuse. The probabilities can only be used if the classifier is calibrated, allowing for far more flexible and insightful social scientific analysis.

Lewis received only 3,700 tweets during the run up to the 2017 election. In contrast, Diane Abbott (Fig. 5) received 126,000. For candidates that receive many tweets, such as Abbot, the recalibrated probabilities show similar dynamics compared with using a 0.8 threshold (recommended by Perspective) and thus could be a reasonable choice (although they would still be biased by exactly where the threshold is set). However, for smaller samples this would lead to far less reliable results and could severely distort analyses.

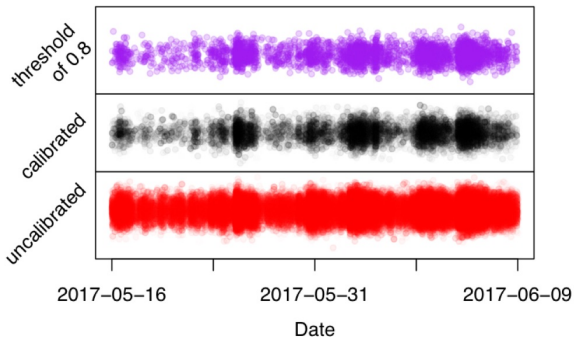


Figure 5: Timelines of toxicity received by Diane Abbott using different methods. (Because of the very large volume, the opacity of dots has been reduced by 80% overall.)

5 Discussion

Abusive content classifiers are increasingly being used for empirical analysis. Yet we show that they should be deployed with caution as their scores are often not interpretable. Although they are usually on an ordinal scale (i.e. a higher value means there is a higher chance of abuse), they are not interval (i.e. a score that is twice as great is twice as likely). If the scores do not meaningfully encode differences then content with a score that is twice as high is not necessarily twice as likely to be abusive—nor can it be interpreted as the 'strength' of abuse is twice as great. The probabilistic programming method we have presented addresses this problem, ensuring that classifiers better reflect human interpretations. Note that this procedure does not improve the 'performance' of classifiers, as measured by metrics such as AUROC, but is important because it makes them far more useable.

We propose that evaluation of interpretability (and recalibration) should be integral to abusive content classifier creation and application. One simple way of ensuring this is: (1) researchers apply their chosen classifier to the dataset they are analysing, (2) uniformly sample across the probabilities, (3) annotate the content based on the original guidelines, (4) evaluate the classifier scores using probablistic programming and (5) recalibrate them as needed. This process could be used for any similar NLP classification task, such as classification of incivil or aggressive language.

References

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep Learning

Models for Multilingual Hate Speech Detection. *Arxiv:2004.06465v2*, pages 1–16.

Or Biran and Courtenay Cotton. 2017. Explanation and Justification in Machine Learning: A Survey. In *IJCAI Workshop on Explainable Artificial Intelligence*, pages 1–5.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. Stan: A probabilistic programming language. *J. Statistical Software*, 76.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *ICWSM*, pages 1–4.

Thomas Davidson and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *3rd Workshop on Abusive Language Online (ACL)*, pages 1–11.

Derek Doran, Sarah Schulz, and Tarek R. Besold. 2018. What does explainable AI really mean? A new conceptualization of perspectives. *CEUR Workshop Proceedings*, 2071:1–8.

Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, 51(4):1–30.

Sahaj Garg, Ankur Taly, Vincent Perot, Ed H. Chi, Nicole Limtiaco, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226.

Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2019. Explaining explanations: An overview of interpretability of machine learning. *Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018*, pages 80–89.

Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. 2018. All You Need is "Love": Evading Hate-speech Detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, pages 2–12.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. *34th International Conference on Machine Learning, ICML 2017*, 3:2130–2143.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning Whom to Trust with MACE. In *Proceedings of NAACL-HLT*, pages 1120–1130.

Isotonic regression in R. 2015. isoreg function in R. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/isoreg.

Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, and Sam Gershman. 2018. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *ArXiv pre-print*, pages 1–21. ArXiv:1802.00682v1.

Khanh Nguyen and Brendan O'Connor. 2015. Posterior calibration and exploratory analysis for natural language processing models. In *Proc. EMNLP 2015*, pages 1587–1598. Long version at arxiv:1508.05154.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proc. ICML 2005*.

PerspectiveScoreNorm. 2020. Perspective API documentation: Score normalization and feedback. https://github.com/conversationai/perspectiveapi/blob/master/3-concepts/score-normalization.md.

John Platt. 1999. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. On fairness and calibration. *Advances in Neural Information Processing Systems*, 2017-December(Nips):5681–5690.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-Agnostic Interpretability of Machine Learning. In *ICML Workshop on Human Interpretability in Machine Learning*, pages 91–96, New York.

Cynthia Rudin. 2018. Please Stop Explaining Black Box Models for High Stakes Decisions. In *NIPS*, pages 1–15.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, Noah A Smith, and Paul G Allen. 2019. The Risk of Racial Bias in Hate Speech Detection. In *ACL Proceedings*, pages 1668–1678.

SciKit. 2020. Scikit isotonic regression. https://scikit-learn.org/stable/modules/isotonic.html.

Andrej Švec, Matúš Pikuliak, Marián Šimko, and Mária Bieliková. 2018. Improving Moderation of Online Discussions via Interpretable Neural Models. In *Proceedings of the Second Workshop on Abusive Language Online (ACL)*, pages 60–65, Brussels. ACL.

Bertie Vidgen and Leon Derczynski. 2020. Directions in Abusive Language Training Data: Garbage In, Garbage Out. *Arxiv:2004.01670v2*, 1(1):1–26.

Bertie Vidgen, Rebekah Tromble, Alex Harris, Scott Hale, Dong Nguyen, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *3rd Workshop on Abusive Language Online*.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the Gdpr. *Harvard Journal of Law & Technology*, 31:841–887.

Cindy Wang. 2018. Interpreting Neural Network Hate Speech Classifiers. In *Proceedings of the Second Workshop on Abusive Language Online (ACL)*, pages 86–92, Brussels. ACL.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *NAACL-HLT*, pages 602–608, Minneapolis. ACL.

Matthew Williams. 2019. *Hatred behind the scenes: a report on the rise of online hate speech*. Mishcon de Reya, London.