Evaluation in Online Safety: a discussion of hate speech classification and safety measures

Economic Discussion Paper Series Issue 9

Published 12 March 2024

Ofcom discussion paper series in communications regulation

The discussion paper series

Ofcom is committed to encouraging debate on all aspects of media and communications regulation and to creating rigorous evidence to support its decision-making. One of the ways we do this is through publishing a series of discussion papers, extending across economics and other disciplines. The research aims to make substantial contributions to our knowledge and to generate a wider debate on the themes covered.

Disclaimer

Discussion papers contribute to the work of Ofcom by providing rigorous research and encouraging debate in areas of Ofcom's remit. Discussion papers are one source that Ofcom may refer to, and use to inform its views, in discharging its statutory functions. However, they do not necessarily represent the concluded position of Ofcom on particular matters.

Contents

Section

Overview

As the regulator of video-sharing platforms and online safety, we need evidence to understand the effectiveness of safety measures at reducing people's experiences of harmful content online.[1]

Online safety is a new regime in a dynamic industry where new services, harms and safety measures will emerge. We will need to continually update our understanding of harms, users, regulated services and safety measures, and update our approach to policy based on this evidence.

In this paper, we discuss the existing literature on the effectiveness of a variety of safety measures applied by some online platforms at reducing hate speech online. We summarise the key findings, highlight gaps, and offer suggestions to shape the direction of future research. In particular, we highlight the importance of assessing the accuracy of hate speech classification by conducting our own analysis of the accuracy of commonly used hate speech classifiers and exploring the implications for research on the effectiveness of a safety measure.

By sharing this paper, our aim is to stimulate discussions on the development of robust methodologies for evaluation of safety measures. We also seek to raise awareness of challenges associated with addressing hate speech online, fostering a more informed and nuanced public discourse on the impact of safety measures dealing with hate speech, and how these impacts may be measured.

What we have found – in brief

There is evidence within the academic literature to suggest that safety measures can reduce hate speech online, although effects vary widely and over time and the evidence is incomplete

Field experiments and quasi-experimental methodologies have demonstrated significant reductions in hate speech from a range of safety measures, from prompts, to bans of communities or individuals, to using the platform's reporting system to remove hateful content. The magnitude of the effects varies widely and appears to be contingent on the characteristics of each platform.

There remains a number of gaps in the literature. The literature has focused on a narrow set of platforms and safety measures. There is less understanding of how safety measures on one platform would affect user behaviour on another platform. The performance of 'off the shelf' hate speech classification techniques is under-assessed, which could weaken the robustness of the research findings.

Understanding the performance of automated hate speech classifiers in hate speech detection is critical

Hate speech classifiers are automated ways of identifying if some text is considered as hate speech or not. The majority of research we reviewed, and which applied these classifiers, did not assess whether the hate speech classifiers accurately identified hate speech.

Many studies use generic hate speech classifiers that have been trained on a wide range of datasets, which are known as off-the-shelf classifiers. The training data (i.e. the data which is used to teach

[1] We describe any intervention taken by online platforms to reduce users' experience of harms online as a safety measure, such as a prompt to nudge users into posting more respectful comments or banning an individual from the platform.

the software which language should be classified as hate speech) for these classifiers can differ to the data used in the evaluation study, which may lead to inaccurate hate speech classification. Inaccurate hate speech classification could then lead to misleading results on the effectiveness of a safety measure.

To address this, researchers would need to assess the performance of each of the classifiers used with respect to the language data on the platform of interest to confirm whether or not the off-the-shelf hate speech classifiers accurately identified hate speech. This would provide greater confidence that significant changes in the trend of hate speech are due to the safety measure, rather than a random trend generated from inaccurate hate speech classification; or that an absence of change does not arise due to the choice of classifier, the choice of language used to train that classifier, or bias inherent in the classifier.

We conducted our own assessment of how accurately two commonly-used classifiers identified hate speech, and we explored whether the errors classifiers made were random or systematic.

We applied the hate speech classifiers to the HateXplain test dataset – a dataset of social media comments with hate speech labels and labels identifying the target of the hate speech. We assessed the performance of two commonly-used hate speech classifiers: Google Perspective API – the most commonly used 'off-the-shelf' classifier and the HateXplain model, which was trained on similar data to the HateXplain test dataset. This case study aimed to compare the performance of an 'off-the-shelf' classifier to a classifier specifically designed for the dataset.[2]

We found the classifier specifically designed for the dataset, HateXplain, performed the best – precisely identifying most of the hate speech in the dataset, while the 'off-the-shelf' classifier failed to identify most of the hate speech comments in the dataset. We also found the performance of the classifiers varied depending on the target of the hate speech. The classifiers made more errors in identifying hate speech targeted at some ethnic groups compared to others.

We highlight that, if inaccuracies of 'off-the-shelf' hate speech classifiers are untested and not understood, then their use can introduce bias into causal inference analysis, making it harder to evaluate whether an intervention has had the desired impact. In our discussion, we point to methods that can be used to study and correct this bias.

[2] For details on the HateXplain dataset and model, see Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P. and Mukherjee, A., 2021, May. Hatexplain: A benchmark dataset for explainable hate speech detection. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 17, pp. 14867-14875); and, for details on the Google Perspective classifier, see https://perspectiveapi.com/

1. Literature Review

1.1. In this section, we present a review of the literature on evaluating the effectiveness of safety measures targeting hate speech.[3] The focus of this review is on empirical approaches to evaluate the effectiveness of safety measures.[4]

1.2. Our key findings are:

- Overall, this work tends to find evidence of a deterrence effect of content moderation: that is, that platform sanctions and nudges can be successful at decreasing the hate speech of posts by users who post hateful content, without substantially decreasing their engagement.
- Methodologically, the most convincing evidence relies on field experiments or on quasi-experimental methods that employ difference-in-difference specifications, using as a control group a set of users who are not subject to moderation; or, regression discontinuity designs using platforms' internal toxicity[5] scoring algorithms as running variables.
- There is a number of gaps in the literature. There is a narrow focus on a small number of platforms, researchers are limited to studying a few types of safety measures and there is a limited understanding of how safety measures on one platform affect user behaviour on other platforms.
- Most notably, many studies did not evaluate the accuracy of the hate speech classification tool. Without this evaluation, we can be less confident the trends which the researchers observed in hate speech are caused by the safety measure. We discuss this further in the next section.

Approaches to estimating causal effects

1.3. **Experimental techniques**, such as randomised control trials ("RCT"), aim to investigate causal effects by comparing the outcomes between a treatment group and a control group. The researcher randomly assigns participants to each group, helping to ensure that any differences between the groups are due to the treatment (e.g. the introduction of the safety measure).

1.4. To conduct experimental research, researchers may collaborate with the platform to control the random assignment of users to treatment and control groups. Katsaros et al. (2022) is one such study. With the advantage of access to the Twitter (now, X) platform, it uses a field experiment to show that just-in-time nudges can be effective but their impact is

[3] The literature on safety measures evaluation does not have a universally accepted definition for hate speech. Our review reflects this by encompassing studies with differing definitions of hate speech. We find the definitions used in the literature are broadly consistent with the definition used in the ISD report, 'Hate of the Nation – A Landscape mapping of observable, plausibly hateful speech in the UK", which defined hate speech as "Activity which seeks to dehumanise, demonise, express contempt or disgust for, exclude, harass, threaten, or incite violence against an individual or community based on a protected characteristic. Protected characteristics are understood to be race, national origin, disability, religious affiliation, sexual orientation, sex, or gender identity".

[4] We acknowledge commissioned work by Charles River Associates for this literature review.

[5] Toxicity is defined as "rude, disrespectful or unreasonable language that is likely to make someone leave a discussion". https://perspectiveapi.com/how-it-works/

low and decreases over time.[6] They find that Twitter users posted 6% fewer offensive Tweets in response to the prompt.

1.5. Jimenez-Duran (2022) employs two experiments on Twitter (now, X), one of which exploits the platform tool which lets users flag harmful content.[7] The study shows that, while the platform acts on these reports by removing content, the authors of hate speech do not change their engagement or the hatefulness of their posts. More precisely, the study finds that reporting increases by 66% the likelihood of the tweet being deleted by Twitter (now, X).

1.6. **Quasi-experimental techniques** attempt to simulate the random assignment mechanism in an experimental setting by exploiting a change that is outside of the control of the platform or incidental to the change but that assigns users to treatment and control groups. For example, a platform may introduce a threshold for an intervention and a researcher can then assume a similarity between users either side of, but close to, this threshold; or an outside event occurs which affects users differently. This acts as a proxy for the random assignment of users that a researcher would conduct in an RCT.

1.7. Chandrasekharan et al. (2017)[8] exploits a natural experiment on Reddit to study how a ban on certain communities affected the subsequent behaviour of users subscribed to those communities. Some users exited the platform following the ban; other users remained active on Reddit and migrated to other communities. The authors demonstrated that users remaining on Reddit reduced their usage of hate speech by at least 80%. Furthermore, in the communities these users had migrated to, there was no significant change in the average level of hate speech, i.e. other communities did not engage in more (or less) hate speech following the ban.

1.8. Ribeiro et al. (2023)[9] exploit Facebook's policy of automatically hiding or removing content with a toxicity score above certain thresholds, to study the user behaviour from comments just below and just above this threshold. Because the authors gather over 400 million comments, they are able to find users which are just below or just above the threshold, meaning that these users may be assumed to be similar but those users just above the threshold receive the treatment, while those just below do not and serve as a control. The authors find that the removal of a poster's comment can reduce subsequent rule-breaking behaviour in the follow-up period.

1.9. Müller and Schwarz (2022)[10] evaluate the effect of the suspension of Trump's Twitter (now, X) account on the toxicity of tweets posted by his followers. They find that the toxicity

---

[6] Katsaros, M., Yang, K. and Fratamico, L., 2022, May. Reconsidering tweets: Intervening during tweet creation decreases offensive content. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 16, pp. 477-487).

[7] Jiménez-Durán, R., 2023. The economics of content moderation: Theory and experimental evidence from hate speech on Twitter. George J. Stigler Center for the Study of the Economy & the State Working Paper, (324).

[8] Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J. and Gilbert, E., 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. Proceedings of the ACM on human-computer interaction, 1(CSCW), pp.1-22..

[9] Horta Ribeiro, M., Cheng, J. and West, R., 2023, April. Automated Content Moderation Increases Adherence to Community Guidelines. In Proceedings of the ACM Web Conference 2023 (pp. 2666-2676).

[10] Müller, K. and Schwarz, C., 2023. The Effects of Online Content Moderation: Evidence from President Trump's Account Deletion.

among his followers dropped by around 25% in response to the ban compared to a representative sample of US Twitter (now, X) users.

1.10. Apart from the study designs which aim to find a causal effect of a safety measure, some studies provide insights into the impact of safety measures with **descriptive evidence** (Chandrasekharan et al., 2017; Rauchfleisch & Kaiser, 2021).[11] For example, Chancellor et al (2016) [12] examined pro-eating disorder communities and found that they adapted their language to circumvent safety measures on Instagram, and that communities that adapted their language were correlated with an increase in user participation and support of eating disorders. Rauchfleisch & Kaiser (2021) examined the migration of deplatformed far-right channels from YouTube to BitChute. They show that deplatforming on YouTube reduces the reach of hate speech and disinformation as smaller platforms have less reach.

Gaps in the literature

1.11. While the literature on safety measures is growing, there remain gaps in our understanding of the impact of safety measures and how to evaluate them.

1.12. Firstly, the literature, with some exceptions, does not provide evidence on the accuracy of the classification of language. However, when assessing the impact of a safety measure, understanding the performance of the classifier used is important to make studies more comparable and understand the impact of quality of the classification on the estimated effect of the intervention. Specifically, we found that few papers address the following topics relating to the quality of the classifier:

a) **Performance**. Few papers evaluate how well the classifier performed at identifying hate speech in their datasets. When 'off-the-shelf' classifiers are used, no papers we reviewed evaluate the performance of the classifier on their dataset, even though using 'off-the-shelf' classifiers can result in lower performance.

b) **Bias.** Machine learning models make errors in predictions. If these errors are systematic, the results of the hate speech classification process are biased, which could in turn introduce bias into causal inference analysis. There was little discussion on how to detect bias in hate speech classification, nor was there any discussion of how one might correct for biases in the subsequent causal inference analysis.

1.13. We consider this an important area to explore. Machine learning is already the most used method to measure harms at scale on online platforms and in evaluation studies. To generate robust causal evidence from machine learning outputs, we need to develop better methods to investigate the nature of the prediction errors and consider how they will impact causal inference analysis. We discuss this further in the next chapter.

1.14. Moreover, there is a number of other gaps in the literature which would benefit from more exploration, and could be helped by greater transparency by platforms:

- Platform variety. Large platforms such as Twitter (now, X), Reddit and Facebook are the focus of most studies. A few studies explored alternative platforms such as Gab and Parler. Concentrating on large platforms has the advantage of focusing on safety

[11] Proceedings of the ACM on Human-Computer Interaction, 1(CSCW), 1-22; Rauchfleisch, A., & Kaiser, J. (2021). Deplatforming the far-right: An analysis of YouTube and BitChute. Available at SSRN 3867818.

[12] Chancellor, S., Pater, J. A., Clear, T., Gilbert, E., & De Choudhury, M. (2016, February). # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing (pp. 1201-1213).

measures that reach the largest number of users. However, this focus presents several potential issues. It may limit the generalisability of findings. Each platform may have unique issues; a safety measure on one platform may not be effective on another. There may be smaller platforms with less content moderation which can be overlooked. Furthermore, data accessibility can bias research toward platforms with readily available data. A more diverse selection of platforms is crucial to gain an understanding of the challenges and solutions across the online landscape.

- Safety measure variety. There is little public information about platform safety measures and the algorithms these platforms use. For example, on Twitter (now, X) it was possible to measure suspensions and deletions, but the literature shows the difficulty with studying other "unobservable sanctions" such as locking users' accounts or which triggers are used by the platforms to proceed with their sanctions.
- Establishing a control group. The main challenge in identifying causal inference is that a researcher typically does not have a control group which is not affected by the safety measure introduced. For instance, a platform may implement a message prompt urging users to be mindful of offensive language but there is no control which would receive the message at random or a variation of the message, or no message. As such, there is a difficulty in devising an experiment without working with a platform in order to perform A/B testing and manipulating moderation.[13] Platforms' lack of transparency makes it difficult to obtain detailed information about natural experiments.[14]
- Cross-platform effects. We recognise the potential for hate speech to spill over on to other platforms (e.g., a safety measure on one platform may decrease hate speech on it but "move" some of the toxic conversations to other platforms). Techniques to measure spill-over effects would help build a comprehensive understanding of all the effects of a safety measure.[15]

[13] A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drives business metrics. Platforms may regularly apply A/B testing to versions of the service to test user reactions.

[14] Although, it is possible that platforms deliberately limit transparency as a strategy to prevent bad actors from exploiting additional information to circumvent safety measures.

[15] Note that the presence of spill-over effects does not suggest the safety measure is ineffective, but these effects should be considered alongside the within-platform effects of a safety measure.

2. Hate Speech Classification

2.1. As we set out above, understanding the performance and possible bias of classifiers is important for causal inference. In this chapter, we first expand on the importance of these issues for causal inference. We then examine the performance of two hate speech classifiers: Google Perspective API, the most frequently used classifier in the literature; and, HateXplain – a highly-cited hate speech classification model designed by academics. We evaluate the performance of the classifiers against the "HateXplain" labelled dataset.[16] This provides a case study to understand the benchmark performance of an 'off-the-shelf' classifier against a classifier that has been specifically trained for the dataset under investigation.

Causal Inference and classification

2.2. For causal inference, more emphasis is put on measurement error in independent variables due to their potential for bias, but less attention is typically paid to measurement error in outcome variables (here, the frequency of hate speech in the data). However, measurement error is important for causal inference and has two important implications:

a) It introduces noise in the estimated relationship. From a statistical perspective, this means that we are more likely to conclude that there is no evidence of an effect while actually there is an effect (also known as type 2 error).
b) It may bias the estimated relationship. Bound et al. (1989) show that, if measurement error is systematically related to the 'true' outcome, then this will result in biased estimates. When using outcome variables generated by machine learning, measurement error of this sort may arise.[17] For example, if the machine learning classifier leads to misclassifying the outcome for a specific sub-group, this may introduce correlation between the misclassification and the 'true' outcome, thus resulting in a bias in the impact analysis.[18]

2.3. While both implications can reduce the precision and accuracy of causal estimates, bias in the measurement error can pose a greater challenge. Noise can lead to an underestimation of the true relationship. Bias in the causal estimate leads to an incorrect identification of the underlying relationship, possibly estimating a relationship that is directionally incorrect - a positive relationship is estimated when the true relationship is negative, or vice versa.

2.4. In the context of evaluating safety measures, measurement error can arise when the outputs of classification are used as the outcome variable, for example when a classifier is used to identify the incidence of hate speech (the outcome variable) in a sample of text.

---

[16] Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P. and Mukherjee, A., 2021, May. Hatexplain: A benchmark dataset for explainable hate speech detection. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 17, pp. 14867-14875). This dataset is generated from Twitter (now, X) and Gab data and includes labels on the targeted group of hate speech.

[17] More generally, Bound et al. (1989) show that, if there is correlation between the measurement error in the dependent variable (e.g., hate speech) and an explanatory variable (e.g. the introduction of content moderation), then the estimated treatment effect will be biased.

[18] To make the example more concrete, if hate speech is classified using machine learning, the classifier may not detect hate speech directed at a specific group. This likely results in correlation of the 'true' outcome and the error term, leading to a bias in the estimate of the impact of a safety feature on reducing hate speech.

While machine learning based classifiers have become very good, they are not perfect – we provide some evidence on this in the next section. Therefore, machine learning classifiers introduce noise, increasing the likelihood of not finding a statistically significant effect while actually there is an effect.[19]

2.5. Moreover, if the misclassification is systematic, there is the risk that bias is introduced in an impact evaluation of a safety feature. To illustrate this risk of systematic misclassification, consider the example of Milios and Behnam Ghader (2022) who show that BERT, a classifier frequently used for hateful language classification, has certain social biases, for instance related to gender or ethnicity. Assuming an underrepresentation in the classification of certain demographics, the estimated impact of a safety feature would likely be lower compared to the true impact of the safety measure.[20]

2.6. Generally, the misclassification can result in an over- or underestimate of the effect of interest depending on whether the classifier over- or underrepresents certain demographics in the predictions.[21] As such, the issue of bias in estimating the effect of a safety measure can be complex as a classifier may overrepresent some groups and underrepresent other groups. [22] Bound et al. (1989) suggest a data-driven approach to assess and potentially correct for this issue. For example, one could use an annotated dataset to estimate correlation between the annotated comments and the classified comments to estimate the size of the potential bias.[23] Bound et al. (1989) show this correlation can be used to correct for the potential bias.

2.7. The above example suggests that the presence of biases in classification may affect the measurement of the causal impact of safety features. Therefore, understanding the noise and bias introduced by machine learning models is important to ensure the estimated causal effects are precise and accurate. In the next section we investigate the performance and biases of two hate speech classifiers: Perspective API and HateXplain.

Findings on performance and bias

Performance: off-the-shelf classifiers can miss most of the hate speech in a dataset

2.8. Understanding the accuracy of a classifier is important in order to assess the degree of noise that is introduced into the estimated relationship when assessing the impact of a safety measure. One should therefore commission a manual classification of a small sample of the data to provide a comparison with the application of the classifier to the full sample and then report the following standard metrics to assess how accurately the classifiers identified hate speech in the data:

---

[19] This is called a type 2 error or false negative finding.
[20] Of course, the classifier may also be 'over-confident' on other demographics introducing and upwards bias in the estimate. This means careful evaluation of the classifier is needed.
[21] Moreover, this issue may be complex because some groups may be over, some others underrepresented in the classification, thus making it impossible to sign the bias without estimating it.
[22] When dealing with Hate Speech there may not be an objective truth. But we assume that human annotation of Hate Speech is comparatively more accurate.
[23] The authors suggest several measures that a researcher could compute to understand the direction and magnitude of the bias in estimated effect. Moreover, they suggest that those metrics can be used to correct for the bias in the estimated effect.

- *Precision*: Precision measures how often a classifier correctly identifies the type of speech out of the total instances of hate or normal speech it predicts. It is the ratio of true positive predictions to the total number of positive predictions (true positives + false positives). For example, 80% precision indicates that, when a classifier identifies a comment as hate speech, it is correct 80% of the time.
- *Recall*: Recall measures how often a classifier correctly identifies the type of speech out of the total instances of hate or normal speech in the data. It is the ratio of true positive predictions to the total number of actual positive labels (true positives + false negatives). For example, a 60% recall indicates that the classifier has correctly identified 60% of all the actual hate speech in the dataset.

2.9. Focusing efforts to achieve a better score of one of these measures can come at the expense of the other. We therefore also report the F1 Score, which combines precision and recall into a single value, providing a balanced measure of a model's performance. All three metrics we have used here are standard metrics used in machine learning and data science literature on the performance of machine learning models (although not in the literature on how these models are applied to online safety measures). Table 1 below shows the performance of the two classifiers of interest when applied to the HateXplain dataset with respect to these three metrics.

Table 1: Performance scores of hate speech classifiers on the HateXplain dataset

	Precision	Recall	F1
Perspective API	0.79	0.13	0.23
HateXplain	0.78	0.78	0.78

Note: scores indicate the performance of the classifier on the 'hate' subcategory. 1 indicates the model can make perfect classifications while 0 indicates that the model cannot make any correct classifications.

2.10. The dataset used to evaluate the two models is a subset of the HateXplain dataset[24], meaning the data is highly similar to the data used to train HateXplain. By contrast, Perspective API may not have been trained on similar data and may not be fine-tuned to detect hate speech in context. This test is intended to investigate the potential misclassification that can occur with using off-the-shelf classifier and applying it to a particular data set. It is not to say that Perspective API would generally perform worse than the HateXplain classifier when applied to other data sets of language (and, indeed, it may perform far better).

2.11. Table 1 shows the performance varied widely across the two classifiers. As expected, HateXplain demonstrates higher overall performance than the 'off-the-shelf' classifier, Perspective API. Perspective API is as precise as HateXplain, but identified only 13% of all hate speech in the dataset (compared to 78% with HateXplain). The disparity in performance highlights the greater accuracy arising from using a classifier that has been trained on a dataset that is from the same platform and user base from which data for the causal analysis is collected.

[24] Specifically, it is the evaluation dataset, which is used to test the performance of the HateXplain model but not used in its training.

2.12. We also explore whether the two classifiers agreed on the classification of a comment. Using an agreement score, we find some agreement between the two classifiers that a comment contained hate speech.[25] However, this agreement wasn't very strong or decisive – there were many cases where they disagreed on the classification.

2.13. The results presented in Table 1 and the agreement score suggest a high degree of measurement error can arise from using off-the-shelf classifiers (informed by machine learning models) to predict the outcome variable. In this case study, the off-the-shelf classifier Perspective API, failed to identify most of the hate speech in the dataset. The performance of the classifier that was specifically trained for the dataset in question is better, although this still produces a degree of measurement error.

2.14. Given the measurement errors that can arise, researchers need to be mindful that using machine learning based classifiers can, at the very least, introduce noise into the estimated relationship which may then result in an underestimation of the true effect. Where there is a high degree of measurement error (for example – using the off-the-shelf classifier in the case above), this can result in concluding there is no significant relationship between the explanatory and outcome variables while there actually is one.

Bias: classifiers were better at identifying hate speech for some protected characteristic groups than others

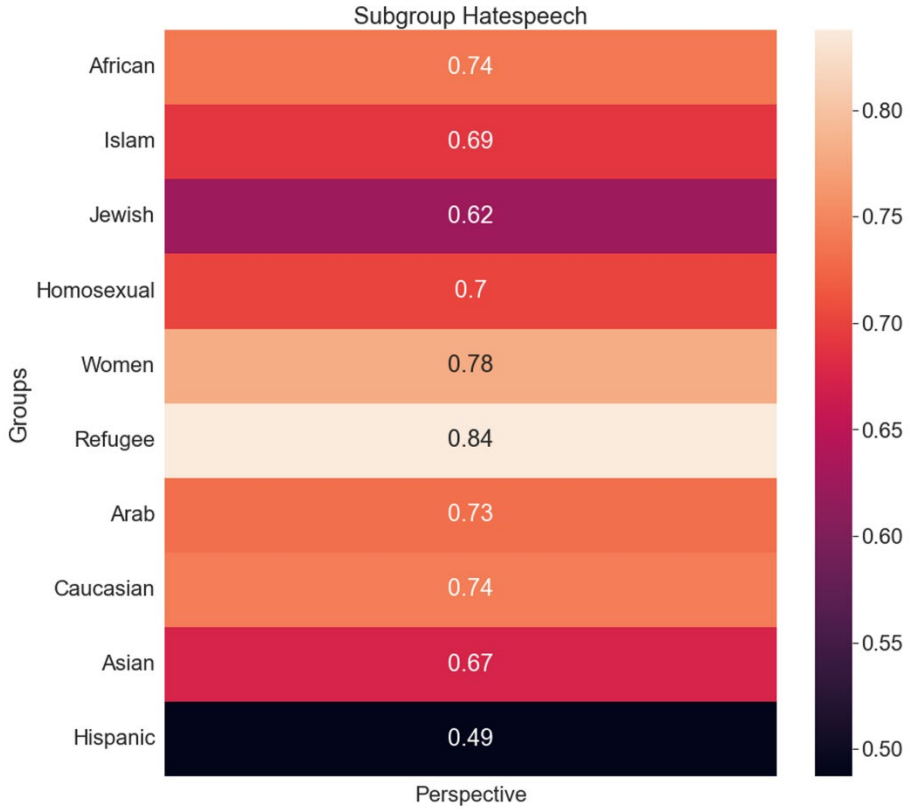
2.15. The next step we believe that researchers applying machine learning classifiers should undertake is to explore whether the measurement error is biased. As discussed, biased measurement errors can be more problematic than noise for the validity of causal inference analysis.

2.16. Systematic errors in classification can manifest in many forms in the context of hate speech classification. Machine learning models may perform better at identifying hate speech from one demographic or cultural group, while underperforming on others. Models may favour certain dialects or languages. This can arise from biases introduced at various stages of a classifier. For example, confirmation bias can occur when the training data incorporates existing biases from our society. Labelled training data could be biased depending on the training given to human annotators, or the cultural background of the annotators. These biases are then absorbed by the machine learning models, which, in turn, make predictions that can reinforce and perpetuate existing stereotypes.

2.17. The type of bias we aim to investigate is whether the models are better at identifying hate speech when certain groups with protected characteristics are targeted. The HateXplain labelled dataset includes information on which group was the target of hate speech. The figure below shows the performance of Perspective API by the target-group of the hate speech. Darker colours (and lower 'Area Under the Curve' scores) in the heatmap indicate that the model performs poorly at distinguishing between hate speech and normal posts, and lighter colours indicate higher scores and better performance.

[25] We used an agreement score, Cohen's kappa, to test the agreement between the two classifiers. We found that comparing HateXplain and Google Perspective labels generated an agreement score of 0.39, which Cohen suggests interpreting as 'fair agreement'. The scale ranges from -1 to 1. ≤ 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement. McHugh, M.L., 2012. Interrater reliability: the kappa statistic. Biochemia medica, 22(3), pp.276-282.

Figure 1: Heatmap of the performance of the Perspective API classifier at labelling hate speech, by target of hate speech



Note: The scores are "Area Under the Curve" (AUC) scores, a standard metric used to evaluate the performance of a classification model, with higher AUC values indicating better model discrimination between classes.

2.18. The scores in the heatmap show that how well the classifier distinguished between hate speech and non-hate speech comments made in relation to each protected group. The closer the score is to 1, the better the classifier can separate hate speech from non-hate speech, while a score lower than 0.5 means that the model is not better than random guessing.

2.19. The heatmap shows, for example, that the Perspective classifier is more likely to confuse between hateful and normal posts specific to the Hispanic group but perform better with respect to the Refugee group.

2.20. The results suggest that Perspective API may make biased errors when predicting hate speech targeted at certain groups in this case study. As this is only a case study, these results are not representative of the performance of Perspective API on all datasets. However, we found that literature using Perspective API did not examine the performance or potential bias of the classifier on the datasets to which the classifier was applied. If unaddressed, the measurement bias can bias the causal estimate, which may compromise the validity of the study.

2.21. Beyond the implications for the causal analysis, lower accuracy in labelling hate speech against specific protected characteristic groups has policy implications:

- Research relying on these classifiers may not fully capture the experiences of marginalised communities, potentially limiting our understanding of the impact of hate speech on these groups.

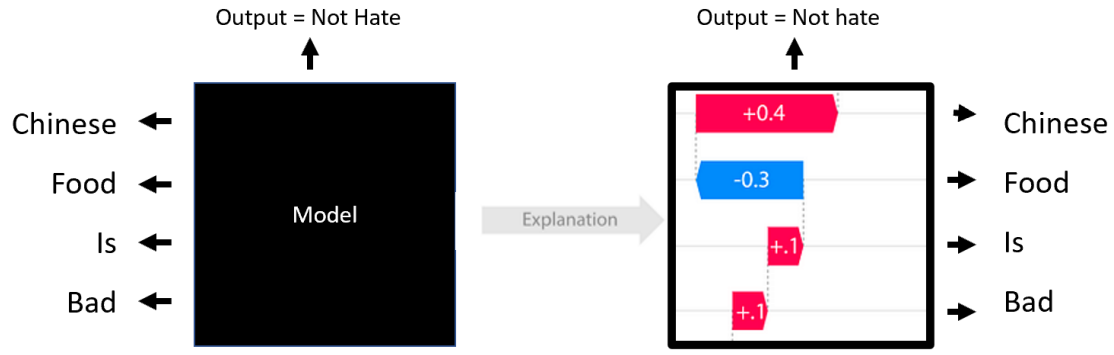
- If this bias is also present in content moderation algorithms used by platforms, it can result in undetected instances of harmful content targeting these groups, leading to insufficient moderation and protection within online platforms, or over-identification of hate speech which could have negative impacts on freedom of expression.
- Consequently, having awareness about these potential biases and considering strategies to identify and mitigate them would support similar levels of protection across user groups in online spaces. Strategies could include selecting classifiers that have been trained on comparable datasets and are regularly updated to keep up with changing trends in online language and hate speech.

Explain-ability: unravelling the decision-making processes of hate speech classifiers to understand sources of bias

2.22. Next, we considered whether a technique from the field of 'explainable AI' can be used to investigate the biases in measurement error.

2.23. Shapley values represent a technique to quantify the importance of each variable to a model's prediction. In the context of hate speech classification, we can use Shapley values to understand which words contribute the most to a classifier deciding to label a comment as hate speech.

Figure 2: Illustrative example of an Explainable AI method



Note: the illustration on the right shows that the word 'Chinese' contributed +0.4 to the probability of a hate speech prediction, while the word 'Food' contributed -0.3. The other words had negligible contributions to a classifier's decision-making.

2.24. If benign words have high positive Shapley values, that could be evidence to suggest the classifier is not accurately labelling hate speech and therefore introducing noise into the causal inference framework.

2.25. Furthermore, Shapley values could be used to gain insight into the bias and noise in the measurement error on individual comments. If a comment has been incorrectly labelled as hate speech, Shapley values can give some explanation as to which terms are driving the incorrect labelling. Consider the following stylized example, a benign comment has been falsely identified as hate speech and further analysis reveals the word 'female' has a high positive Shapley value. This indicates the classifier incorrectly labelled the comment as hate speech because it placed significant weight on the reference to 'female'. This could indicate that there were differences between the use of the word 'female' in the training data compared to the test dataset, which the researcher could investigate.

2.26. However, this can be a less reliable technique to explain the sources of bias in aggregate. Shapley values do not account for differences in the overall number of types of hate speech in the underlying data. Therefore, if a term related to a certain group has a disproportionately high Shapley value across the entire dataset, this could indicate that there is a high proportion of hate speech comments made in relation to that group in the underlying data, rather than because the classifier is biased towards finding hate speech against that group.

3. Conclusion

3.1. We found that there is a growing literature that describes a range of experimental and quasi-experimental approaches which have been used to evaluate safety measures. The most convincing evidence is from field experiments that have robustly demonstrated modest decreases in hate speech following the introduction of behavioural safety measures. In addition, researchers have applied hate speech classifiers, to large scale data. The intention of this is to enable efficient data analysis which can be used to measure evidence of harms at scale, and hence understand which safety measures are effective.

3.2. There remain a number of gaps in the literature. First, there is limited exploration of a diverse array of safety measures, and platforms, hindering the generalisability of findings. Second, the current literature predominantly focuses on within-platform effects, leaving questions about cross-platform impacts unanswered. Third, the accuracy of off-the-shelf hate speech detection methods is often not assessed and this can lead to misleading results.

3.3. In particular, our work above shows that outputs from off-the-shelf machine learning classifiers may be characterised by measurement error, particularly when the data to which they are applied is different to the data on which these classifiers have been trained. When these outputs are used as inputs into causal inference analysis and this measurement error is random, the noise introduced through the measurement error may lead researchers to wrongly conclude no relationship exists between the safety measure and the incidence of hate speech, based on a lack of statistical significance. Problems with identifying accurately a causal relationship may become even more severe if the hate speech classifier systematically misclassifies language relating to one or more characteristics (e.g. ethnicity, gender), as this may lead to a biased estimate of the relationship (including showing the opposite of the true relationship).

3.4. Looking ahead, we welcome future research that explores further issues and solutions related to combining machine learning and causal inference techniques, in particular – techniques to improve the validity and precision of causal estimates when measurement errors are present. In particular, researchers should make clear the performance of the machine learning classifier employed and consider using alternative classifiers to test the robustness of results. Additionally, there is a need to widen the scope of research into a greater variety of platforms, safety measures and to investigate spill-over effects to other platforms: recognising the interconnected nature of online spaces. In conclusion, while progress has been made, the complexity of evaluating safety measures and addressing hate speech requires ongoing research efforts, sparking discussions and collaborations for a safer and more inclusive online environment.