Like trainer, like bot: inheritance of bias in algorithmic content moderation

Reuben Binns, Max Van Kleek, Nigel Shadbolt[1] and Michael Veale[2]

[1] Department of Computer Science, University of Oxford
...
[2] StEAPP, UCL
...

Abstract. The internet has become a central medium through which 'networked publics' express their opinions and engage in debate. But offensive, aggressive comments and personal attacks can inhibit participation. Recent proposals for automated content moderation aim to overcome this with machine learning classifiers trained on large corpora of texts manually labelled by human raters as offensive, aggressive or personally attacking. While such systems could help encourage more civil debate, the normative boundaries they aim to navigage are inherently contestable, subject to the idiosyncratic norms of the human raters who provide the training data. This paper provides an initial investigation of the effects of such normative differences on algorithmic content moderation, by way of a case study using an existing dataset of comments labelled for offence. We train classifiers on comments labelled by different demographic subsets (men and women) to understand how differences in conceptions of offence between these groups might affect the performance of the resulting models on various corpora. We introduce a novel concept of 'unfairness' for automated text classification tasks where the concepts at stake are inherently normative and contestable, and use it to assess the bias of the 'gendered' classifiers. Based on this case study, we discuss some of the ethical choices facing designers of algorithmic moderation systems, and the platforms that deploy them.

Keywords: political science algorithmic accountability machine learning online abuse discussion platforms

1 Introduction

Online platforms are increasingly the place where we express our opinions and engage in debate. They have been called the new 'curators of public discourse' [8], and the digital extension of the public sphere [4]. Social media, news websites, and question / answer forums provide significant opportunities allow their users to express views and engage in deliberation with a diverse range of peers [9]. However, aggressive, offensive or bullying comments can stifle debate and drive people away from participating in such discussions. Left unchecked, they may also lead to calls for intervention by regulators or law enforcement authorities. As a result, many platforms have terms of use and content policies to define the bounds of acceptable discourse, and employ active measures to enforce them [12].

Defining company policies and the standards and norms that underpin them might be within the remit of company executives, or undertaken in consultation with users and other stakeholders. Platforms with large volumes of video, like Facebook and Youtube have teams dedicated to identifying and removing offensive content. Many platforms also rely on users flagging content, both as a means of detection and as a 'rhetorical justification' for censorship [3]. In some cases, moderation privileges are granted to particular volunteers, who may be self-appointed (e.g. Reddit), appointed through semi-democratic processes (e.g Wikipedia), or implicitly through reputation developed on a platform (e.g. StackExchange). Community norms defining acceptability are rarely static, consistent or uncontested. The discussion forum Reddit features many sub-fora in which different norms hold, leading to frequent arguments between users, staff and executives across sub-fora over what kinds of posts should be allowed [2]. What counts as acceptable is therefore always a subjective matter, particular to a platform and even to particular sub-groups within it.

Whether it is performed by employees, external agencies, users or volunteers, moderation is a primarily human endeavour. However, with the quantity of content appearing on many platforms, manually vetting each individual item can be very costly, driving interest in technical solutions which could automate harmful content detection. A common approach in the past has been to apply automatic detection of abusive terms based on manually curated blacklists of banned words, but maintaining effective lists proved difficult, particularly as language and norms change, and as users learn to game deployed systems. As a response, researchers and technology companies have proposed novel algorithmic means of moderating such content. By training machine learning algorithms on large corpora of texts manually labelled for aggression, offence or abuse, they aim to create automatic classification systems to flag harmful comments. One such initiatives is Perspective API, from Google.[3] According to the project description, a platform could receive a score which predicts the âĂŸimpact a comment might have on a conversationâĂŹ, which could be used âĂŸto give realtime feedback to commenters or help moderators do their jobâĂŹ. Similar efforts are being pursued by other platforms, including Twitter, a microblogging platform, and Disqus, a third party comment plugin provider.[4].

While such automated content moderation might help lighten the burden on human moderators, automated content moderation with manually labelled data can only be based on the collective judgements regarding norms of offense of the people whose ratings provide training data for the classifier. Where multiple implicit or explicit communities exist âĂŤ particularly where participation in

[3] See https://www.perspectiveapi.com/

[4] https://blog.disqus.com/first-steps-to-curbing-toxicity,
https://www.recode.net/2017/2/7/14528084/twitter-abuse-safety-features-update

labelling is not balanced âĂŤ this might penalise particular types of content or communication, such as political views or vernacular. The norms of these raters risk being imposed on all users of the platform, potentially affecting the balance and diversity of participation.

This is compounded in cases where training data is de-contextualised from the domain of application. Such decontextualisation might occur as, in an effort to build more sophisticated classifiers, data from multiple platforms are combined. New companies interested in using automated content moderation may have no choice but to use models built from data collected elsewhere. Yet even if the training data is only taken from where the moderation is occurring, it might introduce historical biases or patterns incompatible with the changing nature of community norms, which would then be reproduced. Finally, in cases where the community standards are in flux, or contested between different stakeholders, the question of whether an automatic abuse classifier is âĂŸaccurateâĂŹ is likely to impinge upon pre-existing platform conflicts.

This paper provides an initial investigation of the effects of potential bias in algorithmic content moderation. As an illustration of the potential risks, we experiment with a series of text classifiers using an existing dataset of 100,000 Wikipedia comments manually scored for 'toxicity', âĂŸaggressionâĂŹ and âĂŸpersonal attacksâĂŹ. In order to examine how differences between people's norms of offence might result in different classifiers, we built different classifiers from demographically distinct subsets of the population responsible for labelling the training data. Specifically, we focus on gender as a demographic variable which may be associated with differences in judgements about offence [5].

This case study aims to illustrate methods and metrics for exploring bias in text classification tasks where the learned concept is inherently contestable; we also use it to reflect on a range of ethical considerations that should be taken into account by designers of algorithmic moderation systems, and the platforms that deploy them. As algorithmic content moderation approaches become more pervasive, the platforms deploying them will face ethical choices with significant implications for the development of community discussions and the digital public sphere.

2 Background and Related work

league of legends online gaming. started on wikipedia - 63m english talk pages. wishlist. Generic work on online discussions.

hate speech [7]

online harassment, and cyberbullying of kids [14,19]. Ä recent Pew Research Center study defines online harassment to include being: called offensive names, purposefully embarrassed, stalked, sexually harassed, physically threatened, and

[5] We chose gender as a relevant demographic primarily due to the ease with which statistically balanced samples can be drawn compared to the other variables (age, education level); we do not assume or intend to establish any general conclusions about gender and offence

harassed in a sustained manner - [21]. personal attacks and abuse can suppress the free speech of the victim.

2.1 Automated detection

detecting hate speech [15, 20, 22]

SVM on sentiment and context

cyberbullying detection based on attack type (sexuality, race, intelligence) [6]

character-level ngrams [13]

2.2 Algorithmic bias

These automated systems rely on machine learning. machine learning can be biased. DADM and FAT-ML. mostly concerned with fairness in terms of non-discrimination. race, gender. as yet, not applied to disparities in abuse detection.

What kinds of bias might apply in this case?

What some people think is 'abuse' is just partisan disagreement. in a study of news platform comment section moderation, Diakopoulos and Naaman found that media organisations acknowledge that their moderators may bring their own biases to the evaluation of standards [5]. Sometimes users flag comments as abusive when other users might judge them as OK; the differences may be due to political partisan divides. As one moderator said; "what one person thinks racially prejudiced, another may think is a criticism of culture. ThatâĂŹs usually the toughest question".

An important difference between this case and the typical contexts that have previously been studied in DADM is that the bias is not against particular demographic groups per se (e.g. gender, race, age), but but rather against particular definitions of offense. However, it is also likely that individuals within certain demographic groups are likely to share definitions of offense, at least to some extent. For example, various studies report gender differences regarding offensive language [10,11,18]. Therefore, in talking about bias in offense classification, we are not talking about bias directly affecting individuals from a certain protected group (for instance, disportionate numbers of women being denied loans). Rather, we are talking about bias towards or against definitions of offense which are more strongly associated with one group over another. For instance, a classifier might more often classify comments as 'offensive' according to definitions of offense that are more prevalent amongst men, than alternative definitions that are more prevalent amonst women. If certain populations have particular definitions of offense, and an automated system more often classifies comments in accordance with one such definition rather than another, we might say that the system is unfair to those who favour the other definition.

Furthermore, conversational environments have norms of acceptability which do not exist in a vacuum - they are reinforced by prior norms, but also malleable. Previous research on online comments has found that by intervening in certain ways, news organizations can affect the deliberative behavior of commenters [16]. By altering the kinds of comments a user, they can determine what kinds of comments they post (e.g. thoughtful or thoughtless) [17]. Finally, nasty comments can lead to negative perceptions of the content they comment on [1].

3 Key questions and hypotheses

Our general question is: how do latent norms and biases in training data affect the operation of offense detection systems? In order to measure this empirically, we need to define what we mean by bias and fairness. What would it mean for an offence classifier to be fair / unfair?

The usual way to evaluate a classifier is to define a loss function, which measures the extent to which its predictions are the same as the ground truth of the phenomena of interest. Normally, we are only interested in one version of the ground truth. In this case, we want to measure biases between classifiers relative to different definitions of offense; i.e. multiple 'ground truths'. To this end, we define the following terms.

For any corpus of comments C in natural language [c1, c2, ... cn], and for any definition of offence D, there is some distribution of offense labels L (0 = 'not offensive', 1 ='offensive'), over C, that are a function of D. For example, definition D might have a distribution of labels L over corpus C, D(L,C) = L(c1) : 0, L(c2) : 1, ... L(cN) : 0.

A classifier CL is biasedâĂŃ against definition D1 and in favour of definition D2, with regards to corpus C, to the extent that CL's offense labels CL(L,C) are 'closer' to D1(L,C) than they are to D2(L,C). That is, we define compare two different loss functions, corresponding to D1 and D2, to determine if CL is biased towards D1 or D2. There are various different ways to measure loss functions, and therefore how 'close' a distribution of offence labels is to a given definition of offense. These are discussed in the results section below.

4 Datasets

Multiple data sets were used in this study, for training and bias detection purposes.

4.1 Wikipedia Talk Annotations

In order to train our classifiers, we used an existing dataset from the Wikipedia Detox project. It features 100,000 annotations of Wikipedia talk page comments manually labelled by crowd-workers using the Crowdflower platform. Each comment is labelled by 10 workers on three features; âĂŸtoxicityâĂŹ, âĂŸpersonal attackâĂŹ and âĂŸaggressionâĂŹ. Each worker gives a score between -2 (very toxic / aggressive / personal attack) and 2 (benign). If the average of 10 workersâĂŹ scores for a particular comment is below 0, that comment is classified as toxic. Generalisation test data We also used another dataset, in order to evaluate the generalisation of our classifiers to a different context. We used the Impermium dataset released on Kaggle, which features posts by users from a range of sources, including âĂIJnews commenting sites, magazine comments, message boards, blogs, text messagesâĂİ. These have been manually rated as âĂŸinsultâĂŹ (1) or âĂŸneutralâĂŹ (0). kaggle.com/c/detecting-insults-in-social-commentary/data

5 Methodology

We began with exploratory analysis of the data.

original authors report high level of inter-annotator agreement - KrippendorfâĂŹs alpha score of 0.45"

we measued agreement within and bvetween genders.

We then trained using a dataset of manually scored comments.

we trained a classifier along the lines of the wikipedia detox project (maybe: we also benchmarked against the jigsaw perspective API)

evaluated using the ROC / AUC. https://en.wikipedia.org/wiki/Receiver$_{o}perating_{c}haracteristicArea_{u}n$

We used a bootstrapping method to sample annotators. For each comment which had both male and female raters, we selected 10 male / female annotators at random, with replacement. We then took the average toxicity/personal attack/abuse rating for these 10 sampled raters.

used this to generate 10 different sets of training data for each gender.

These 10 different sets of training data were used to train 10 different text classifiers to identify comments that are toxic/personal attack/abuse.

We then tested these classifiers against four different sets of test data.

- previously unseen comments withheld from the detox dataset, labelled by a mixture of males and females.

- previously unseen comments withheld from the detox dataset that had been labelled by males, randomly sampled using the same bootstrap method.

- previously unseen comments withheld from the detox dataset that had been labelled by females, randomly sampled using the same bootstrap method.

-previously unseen comments from a different dataset, of twitter posts labelled 'offensive' or 'inoffensive' by a different set of crowd workers whose demographics are unknown.

NB: we aren't claiming that gender is necessarily a strong determinant of norms about offense, nor are we making any claim about the origins (whether environmental, genetic or otherwise) of any putative gender differences in norms about offense, nor any normative endorsement of a particular set of norms about offense. Gender is just an easily understood and accessible demographic attribute of the labelling population, which we hypothesised would exhibit some differences between demographics. Our aim is not to essentialise gender differences. just using gender as one example of socially constructed distinction which may correlate with different norms about offence.

6 Results

-

structure

1. exploratory data analysis.

krippendorffs alpha - demographic subgroups differ in their judgements. women have more diversity in what they consider offensive

2. differences in strength of coefficients between models

we compared the coefficients between the male and female classifiers. We took the features used by the classifiers, and calculated their average coefficient across the 10 classifiers created for each gender.

3. classifier performance

-

and 'fairness'

explanation of the error rate comparison: we ask 'how well does this classifier do in a world where offense is defined by women, compared to a world where offense is defined by men?'. the comparison between error rates can be used to define a gender-specific notion of classifier 'fairness'. If a classifier performs worse according to female-defined offense, than it does compared to male-defined offense, it can be said to be 'unfair to women' in the sense that it is worse at tracking their collective judgements about offense.

so looks like both male and female-trained classifiers have a higher error rate on female-labelled test data than male-labelled test data. In other words, both are 'unfair' to women, in the sense that they are worse at replicating women's collective judgements about offense than men's. However, the male-trained classifiers have a higher disparity in type I error rates between female vs male-labelled test data than female-trained classifiers. in other words, compared to bots trained by women, bots trained by men are more likely to mis-classify comments that women think are inoffensive as offensive. Both male and female-trained bots are worse at capturing women's collective judgements about offense than men's; but for male-trained bots, the disparity is greater.

As established in the previous section, the female raters broader divergence in labelling probably explains why the classifiers are generally worse at predicting female labels.

7 Discussion

we then discuss the significance of our findings.

the consequences go deep into methodological questions in political philosophy.

maybe the crowd-workersâĂŹ notions of toxicity are partisanal? but maybe lib / con are just more offensive?

first, can toxicity be separated from partisanship in a way which does not beg the question? second, what is the right amount of diversity in a free society? How can we get past biases in the training data?

to this end, we re-trained the model using only demographic subsets of the crowd-workers (gender, age, education). These alternative models yielded differential disparate impacts on partisan groups. (to test their significance, we also trained multiple randomly drawn subsets to obtain an expected observation for the null hypothesis).

potential partisan bias mitigation strategies (sub-class of discrimination-aware data mining).

our conclusion is not that automated systems are necessarily a bad idea. but rather that because they are never neutral, there are difficult choices to be made.

8 Conclusion and Outlook

Acknowledgments ...

In the bibliography, use `\textsuperscript` for "st", "nd", ...: E.g., "The 2nd conference on examples". When you use JabRef, you can use the clean up command to achieve that. See `https://help.jabref.org/en/CleanupEntries` for an overview of the cleanup functionality.

References

1. Anderson, A.A., Brossard, D., Scheufele, D.A., Xenos, M.A., Ladwig, P.: The âĂIJnasty effect:âĂİ online incivility and risk perceptions of emerging technologies. Journal of Computer-Mediated Communication 19(3), 373–387 (2014)
2. Centivany, A.: Values, ethics and participatory policymaking in online communities. Proceedings of the Association for Information Science and Technology 53(1), 1–10 (2016)
3. Crawford, K., Gillespie, T.: What is a flag for? social media reporting tools and the vocabulary of complaint. New Media & Society 18(3), 410–428 (2016)
4. Dahlberg, L.: The internet and democratic discourse: Exploring the prospects of online deliberative forums extending the public sphere. Information, Communication & Society 4(4), 615–633 (2001)
5. Diakopoulos, N., Naaman, M.: Towards quality discourse in online news comments. In: Proceedings of the ACM 2011 conference on Computer supported cooperative work. pp. 133–142. ACM (2011)
6. Dinakar, K., Reichart, R., Lieberman, H.: Modeling the detection of textual cyberbullying. The Social Mobile Web 11(02) (2011)
7. Gagliardone, I., Gal, D., Alves, T., Martinez, G.: Countering online hate speech. UNESCO Publishing (2015)
8. Gillespie, T.: The politics of âĂŸplatformsâĂŹ. New Media & Society 12(3), 347–364 (2010)
9. Halpern, D., Gibbs, J.: Social media as a catalyst for online deliberation? exploring the affordances of facebook and youtube for political expression. Computers in Human Behavior 29(3), 1159–1168 (2013)
10. Jay, T.: Cursing in America: A Psycholinguistic Study of Dirty Language in the Courts, in the Movies, in the Schoolyards, and on the Streets. John Benjamins Publishing (1992)

11. Johnson, F.L., Fine, M.G.: Sex differences in uses and perceptions of obscenity. Women's Studies in Communication 8(1), 11–24 (1985)
12. Ksiazek, T.B.: Civil interactivity: How news organizations' commenting policies explain civility and hostility in user comments. Journal of Broadcasting & Electronic Media 59(4), 556–573 (2015)
13. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.: Abusive language detection in online user content. In: Proceedings of the 25th International Conference on World Wide Web. pp. 145–153. International World Wide Web Conferences Steering Committee (2016)
14. Schrock, A., Boyd, D.: Problematic youth interaction online: Solicitation, harassment, and cyberbullying. Computer-mediated communication in personal relationships pp. 368–398 (2011)
15. Sood, S.O., Churchill, E.F., Antin, J.: Automatic identification of personal insults on social news sites. Journal of the American Society for Information Science and Technology 63(2), 270–285 (2012)
16. Stroud, N.J., Scacco, J.M., Muddiman, A., Curry, A.L.: Changing deliberative norms on news organizations' facebook sites. Journal of Computer-Mediated Communication 20(2), 188–203 (2015)
17. Sukumaran, A., Vezich, S., McHugh, M., Nass, C.: Normative influences on thoughtful online participation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 3401–3410. ACM (2011)
18. Sutton, L.A.: Bitches and skankly hobags. Gender articulated: Language and the socially constructed self pp. 279–296 (2001)
19. Tokunaga, R.S.: Following you home from school: A critical review and synthesis of research on cyberbullying victimization. Computers in human behavior 26(3), 277–287 (2010)
20. Warner, W., Hirschberg, J.: Detecting hate speech on the world wide web. In: Proceedings of the Second Workshop on Language in Social Media. pp. 19–26. Association for Computational Linguistics (2012)
21. Wolak, J., Mitchell, K.J., Finkelhor, D.: Does online harassment constitute bullying? an exploration of online harassment by known peers and online-only contacts. Journal of adolescent health 41(6), S51–S58 (2007)
22. Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A., Edwards, L.: Detection of harassment on web 2.0. Proceedings of the Content Analysis in the WEB 2, 1–7 (2009)

All links were last followed on October 5, 2014.