Automated content moderation and political bias: balancing ideological diversity with civility*

Extended Abstract†

Reuben Binns
Jun Zhao
Max Van Kleek
Nigel Shadbolt
University of Oxford
Oxford, United Kingdom

ABSTRACT

The web has become a key forum for political debate, as people take to platforms to comment and express their political viewpoints. However, online comment platforms sometimes contain abusive comments. Recent work has explored automated means of flagging abusive comments, by automatically classifying them as either toxic or civil. While such algorithmic content moderation may promote diversity by encouraging participation from those who would otherwise face abuse, it might also reproduce political biases inherent in training data, resulting in disparate impacts across partisan divides. This paper aims to better understand the potential risks of algorithmic ideological bias in such contexts. We train a simple text classifier using an existing data set of 100,000 Wikipedia comments rated by humans for 'toxicity'. This classifier is applied to a corpus of 4,000 sentences which have been labelled for ideological bias. We find that conservative sentences are more likely than liberal sentences to be automatically classified as toxic. We discuss the potential for and desirability of methods to mitigate such biases, and the implications of such systems for ideological diversity in the public sphere of the web.

KEYWORDS

ACM proceedings, LaTeX, text tagging

ACM Reference format:
Reuben Binns
Jun Zhao
Max Van Kleek
Nigel Shadbolt
University of Oxford
Oxford, United Kingdom
. 2017. Automated content moderation and political bias: balancing ideological diversity with civility. In *Proceedings of ACM Web Science Conference, Troy, New York, USA, June 2017 (WebScience'17),* 2 pages.
DOI: 10.475/123_4

1 KEYWORDS

political science algorithmic accountability machine learning online abuse discussion platforms

2 INTRODUCTION

recent proposals suggest sanitizing online conversations using algorithmic models to classify comments as either toxic or non-toxic.

3 BACKGROUND

Generic work on online discussions.

Abuse.

3.1 Automated detection

3.2 Discovering and mitigating bias in machine learning models

DADM and FAT-ML.

mostly concerned with fairness in terms of non-discrimination. race, gender. as yet, not applied to political diversity.

4 DATA SOURCES AND METHODOLOGY

We 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)

we then used a corpus of sentences which had been labelled for partisan bias (ideological books corpus) to examine whether the detox classifier had a disparate impact on liberal or conservative writing.

4.1 Data sources

4.1.1 Wikipedia Talk Annotations. 100k annotations. toxic or not. png

4.1.2 Ideological books corpus. iyyer et al

4.2 Methodology

train a simple bag of words text classifier.

thresholds

feed lib / con through it.

fiitfis bias all the way downfi

[*] Produces the permission block, and copyright information

[†] The full version of the author's guide is available as `acmart.pdf` document

5 RESULTS

We found differences in toxicity rates between different ideologies. Conservative comments were more often classified as toxic (18%) than liberal comments (13%)

cons toxic is .18 lib toxic is 0.13 netural is 0.19

6 DISCUSSION

we then discuss the significance of our findings.

the consequences go deep into methodological questions in political philosophy.

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

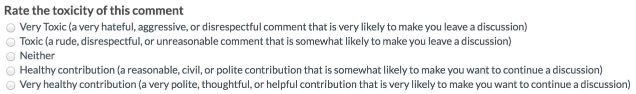
first, can toxicity be separated from partisanship in a way which does not beg the question? second, what is the right amount of diversity in a free society? How can we get past biases in the training data?

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).

6.1 Future work

7 CONCLUSIONS



Rate the toxicity of this comment

○ Very Toxic (a very hateful, aggressive, or disrespectful comment that is very likely to make you leave a discussion)
○ Toxic (a rude, disrespectful, or unreasonable comment that is somewhat likely to make you leave a discussion)
○ Neither
○ Healthy contribution (a reasonable, civil, or polite contribution that is somewhat likely to make you want to continue a discussion)
○ Very healthy contribution (a very polite, thoughtful, or helpful contribution that is very likely to make you want to continue a discussion)

Figure 1: blah blah

REFERENCES