The Automated Speech Police

Teaching computers to weed out online hate speech is a terrible idea.

by Paula Boddington

There are plenty of reasons to worry about the concept of ‘hate speech’. There are also specific concerns about the notion of Islamophobia, especially in light of produce a definition. Both concepts are subjective and hard to pin down. But it gets worse. For around the globe, a cottage industry is springing up, attempting to devise ways to automate the detection of online ‘hate speech’ in general, and of ‘Islamophobia’ in particular.

The aura of scientific objectivity that goes along with the computerised detection of ‘hate’ online is very dangerous. You can’t make a loose and fuzzy idea rigorous by getting complicated algorithms and sophisticated statistical analysis to do your dirty work for you. But you can make it look that way. And worryingly, many of those working to automate ‘hate speech’ detection have direct influence on governments and tech firms.

Those working on such tools often see ‘hate speech’ as a problem worsened by technology. Hence they assume that the solution is more technology. For example, the Anti-Defamation League (ADL) is ‘teaching machines to recognise hate’ by working to produce an recent article, two researchers at the Oxford Internet Institute, Bertie Vidgen and Taha Yasseri, discuss a tool they have built to ‘detect the strength of Islamophobic hate speech on Twitter’. Their work merits more scrutiny, not least because anything produced within prestigious universities, like Oxford, may have disproportionate influence on policy and practice. While, again, they nod in the piece to the difficulty of defining and detecting Islamophobia, they steam on regardless.

The researchers took samples from the Twitter accounts of four mainstream British political parties: UKIP, the Conservatives, the Liberal Democrats and Labour. It then incorporated 45 additional ‘far right’ groups, drawn from anti-fascist group Hope Not Hate’s calling people ‘wallies’, which hardly makes their work appear rigorous or impartial.

Islamophobia is defined, in this study, as ‘any content which is produced or shared which expresses indiscriminate negativity against Islam or Muslims’. Attempting to introduce a degree of nuance, a distinction is made between ‘strong Islamophobia’ and ‘weak Islamophobia’.

The methodology Vidgen and Yasseri use is similar to that of the ADL – they had humans assess tweets, then used machine learning to train computers to continue the work. The first weak spot is, of course, the human assessors. The authors report that three unnamed ‘experts’ graded tweets from ‘strong Islamophobia’ to ‘weak Islamophobia’ to ‘no Islamophobia’. I’d be willing to bet a fiver that not one of these ‘experts’ is critical of the concept of hate speech. Broad agreement on grading between these ‘experts’ is hailed as proof of their rigour – but it may simply be proof that they share certain biases. The subsequent application of machine learning would only magnify such bias.

Worse still, there are no examples given here of tweets and their classification. Instead we just have an illustration of ‘weak’ Islamophobia, as ‘sharing a news story about a terrorist attack and explicitly foregrounding the fact that the perpetrator is a Muslim’. This is flawed. After all, in the wake of a terrorist attack, it is a reflex of some on social media to deny that it has any connection to Islam, even when the evidence suggests otherwise. In response, other social-media users often point out that the attack definitely does have something to do with Islam. And besides, simply highlighting the apparent ideology of a terrorist is hardly hateful in itself.

What is also absent from Vidgen and Yasseri’s analysis are accounts of any prominent atheists, secularists, Muslim reformers or ex-Muslims. Accounts devoted to scholarly critique of Islam might reasonably be presumed to have some basis in fact and reason, and would surely be useful in training data for machine learning. There are plenty of generalised truths about any religion which can be expressed in negative terms. In the case of Islam, these could appear as ‘strong Islamophobia’. But no attempt appears to have been made to exempt such legitimate criticisms.

Vidgen and Yasseri, like so many others, fail to distinguish between people and ideas, between Muslims and Islam. This is a strange, but widespread, phenomenon. From this perspective, to attack someone’s beliefs is to attack their very essence. People are ideas, ideas are people, and critiquing one is a body blow to the other. This subjectivist, relativist position feeds into the concept of hate speech, and the claim that offensive speech is harmful. But in the context of Islam this produces a particularly curious spectacle: a subjectivist worldview being used to defend a religion that is supposedly based on the teachings of an eternal, unchanging deity.

In the end, the issue of hate speech is far more complex than many researchers might like to make out. Policing hate speech is often about deciding whose opinions need protection and whose don’t. At the very least, let’s not hand over that process to machines.

First published in *Towards a Code for Artificial Intelligence*.