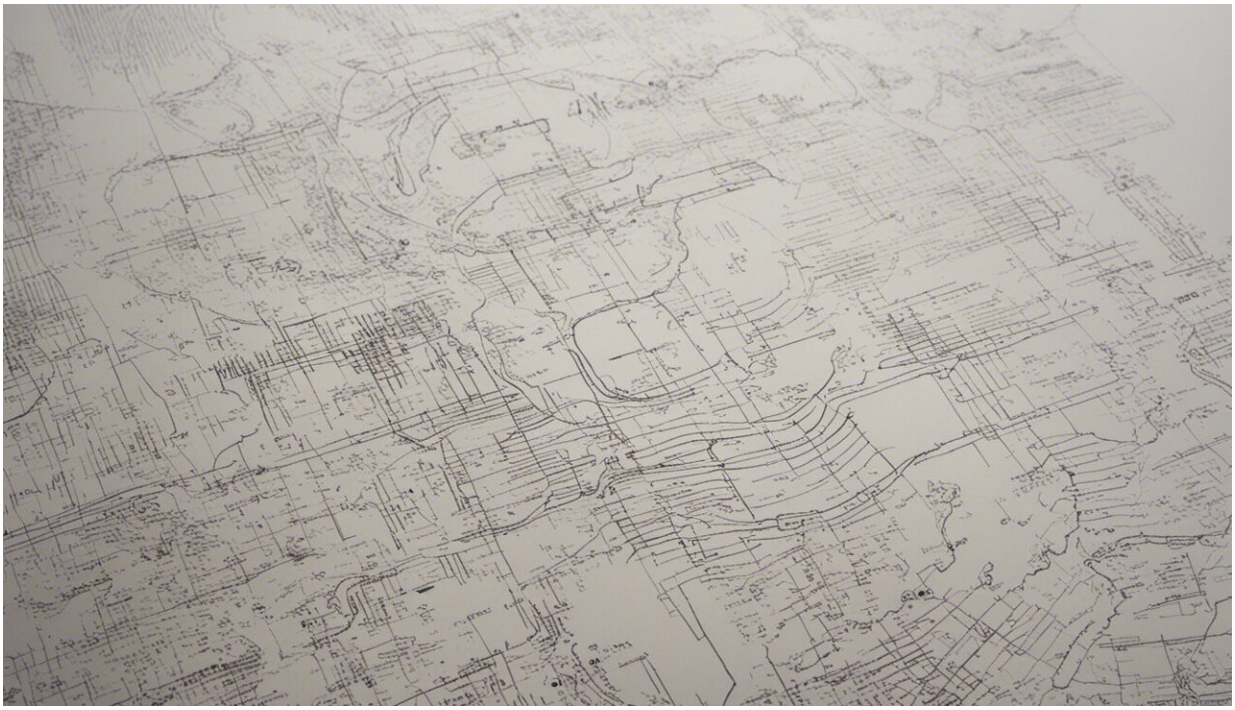
How we built a tool that detects the strength of Islamophobic hate speech on Twitter

January 2 2019, by Bertie Vidgen And Taha Yasseri



Credit: AI-generated image (disclaimer)

In a landmark move, a group of MPs recently published a working definition of the term Islamophobia. They defined it as "rooted in racism", and as "a type of racism that targets expressions of Muslimness or perceived Muslimness".

In our latest working paper, we wanted to better understand the prevalence and severity of such Islamophobic hate speech on social media. Such speech harms targeted victims, creates a sense of fear among Muslim communities, and contravenes fundamental principles of fairness. But we faced a key challenge: while extremely harmful, Islamophobic hate speech is actually quite rare.

Billions of posts are sent on social media every day, and only a very small number of them contain any sort of hate. So we set about creating a classification tool using machine learning which automatically detects whether or not tweets contain Islamophobia.

Detecting Islamophobic hate speech

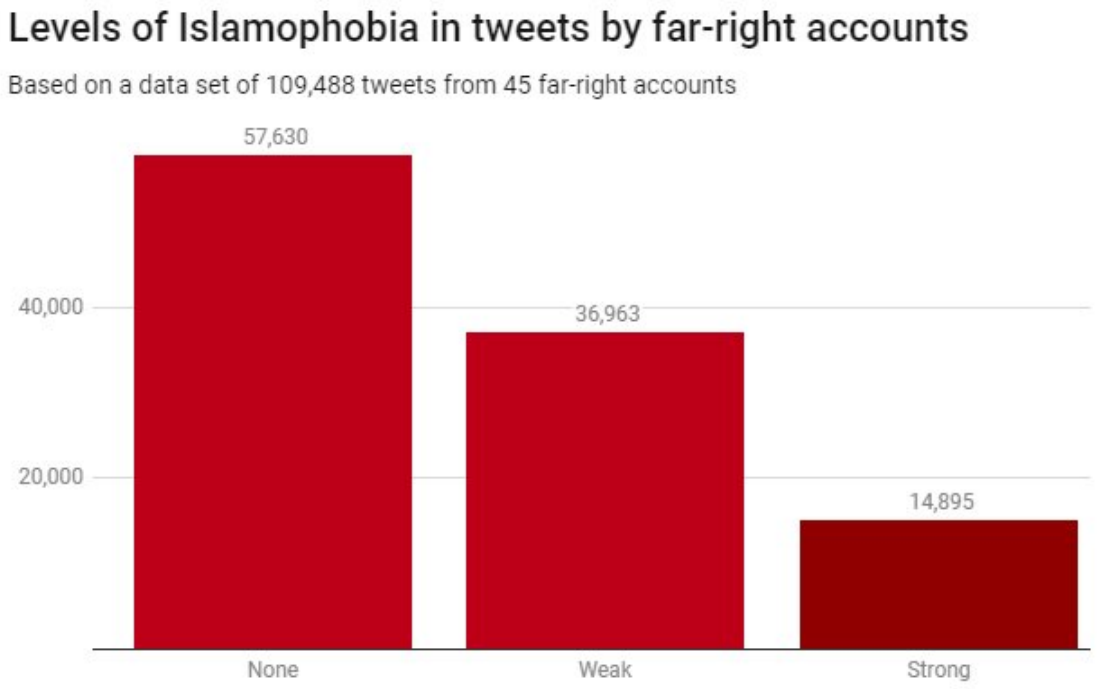
Huge strides have been made in using machine learning to classify more general hate speech robustly, at scale and in a timely manner. In particular, a lot of progress has been made to categorise content based on whether it is hateful or not.

But Islamophobic hate speech is much more nuanced and complex than this. It runs the gamut from verbally attacking, abusing and insulting Muslims to ignoring them; from highlighting how they are perceived to be "different" to suggesting they are not legitimate members of society; from aggression to dismissal. We wanted to take this nuance into account with our tool so that we could categorise whether or not content is Islamophobic and whether the Islamophobia is strong or weak.

We defined Islamophobic hate speech as "any content which is produced or shared which expresses indiscriminate negativity against Islam or Muslims". This differs from but is well-aligned with MPs' working definition of Islamophobia, outlined above. Under our definitions, strong Islamophobia includes statements such as "all Muslims are barbarians", while weak Islamophobia includes more subtle expressions, such as

"Muslims eat such strange food".

Being able to distinguish between weak and strong Islamophobia will not only help us to better detect and remove hate, but also to understand the dynamics of Islamophobia, investigate radicalisation processes where a person becomes progressively more Islamophobic, and provide better support to victims.



Credit: Vidgen and Yasseri

Setting the parameters

The tool we created is called a supervised machine learning classifier. The first step in creating one is to create a training or testing dataset –

this is how the tool learns to assign tweets to each of the classes: weak Islamophobia, strong Islamophobia and no Islamophobia. Creating this dataset is a difficult and time-consuming process as each tweet has to be manually labelled, so the machine has a foundation to learn from. A further problem is that detecting hate speech is inherently subjective. What I consider strongly Islamophobic, you might think is weak, and vice versa.

We did two things to mitigate this. First, we spent a lot of time creating guidelines for labelling the tweets. Second, we had three experts label each tweet, and used statistical tests to check how much they agreed. We started with 4,000 tweets, sampled from a dataset of 140m tweets that we collected from March 2016 to August 2018. Most of the 4,000 tweets didn't express any Islamophobia, so we removed a lot of them to create a balanced dataset, consisting of 410 strong, 484 weak, and 447 none (in total, 1,341 tweets).

The second step was to build and tune the classifier by engineering features and selecting an algorithm. Features are what the classifier uses to actually assign each tweet to the right class. Our main feature was a word embeddings model, a deep learning model which represents individual words as a vector of numbers, that can then be used to study word similarity and word usage. We also identified some other features from the tweets, such as the grammatical unit, sentiment and the number of mentions of mosques.

Once we'd built our classifier, the final step was to evaluate it, which we did by applying it to a new dataset of completely unseen tweets. We selected 100 tweets assigned to each of the three classes, so 300 in total, and had our three expert coders relabel them. This let us evaluate the classifier's performance, comparing the labels assigned by our classifier with the actual labels.

The classifer's main limitation was that it struggled to identify weak Islamophobic tweets as these often overlapped with both strong and none Islamophobic ones. That said, overall, its performance was strong. Accuracy (the number of correctly identified tweets) was 77% and precision was 78%. Because of our rigorous design and testing process, we can trust that the classifier is likely to perform similarly when it is used at scale "in the wild" on unseen Twitter data.

Using our classifier

We applied the classifier to a dataset of 109,488 tweets produced by 45 far-right accounts during 2017. These were identified by the charity Hope Not Hate in their 2015 and 2017 State of Hate reports. The graph below shows the results.

While most of the tweets – 52.6% – were not Islamophobic, weak Islamophobia was considerably more prevalent (33.8%) than strong Islamophobia (13.6%). This suggests that most of the Islamophobia in these far-right accounts is subtle and indirect, rather than aggressive or overt.

Detecting Islamophobic hate speech is a real and pressing challenge for governments, tech companies and academics. Sadly, this is a problem that will not go away – and there are no simple solutions. But if we are serious about removing hate speech and extremism from online spaces, and making social media platforms safe for all who use them, then we need to start with the appropriate tools. Our work shows it's entirely possible to make these tools – to not only automatically detect hateful content but to also do so in a nuanced and fine-grained manner.

Provided by The Conversation