# Troll Patrol: Using Crowdsourcing, Data Science & Machine Learning to Measure Violence and Abuse against Women on Twitter

## **Introduction**

**Welcome to the findings of our Troll Patrol project: a joint effort by human rights researchers, technical experts and thousands of online volunteers to build the world’s largest crowd-sourced dataset of online abuse against women.** Our findings reveal the sheer scale and nature of online abuse faced by women and provides a resource to researchers and engineers interested in exploring the potential of machine learning in content moderation.

These findings are the result of a collaboration between Amnesty International and [Element AI](https://www.elementai.com/), a global artificial intelligence software product company. Together, we surveyed millions of tweets received by 778 journalists and politicians from the UK and US throughout 2017 representing a variety of political views, and media spanning the ideological spectrum. Using cutting-edge data science and machine learning techniques, we were able to provide a quantitative analysis of unprecedented scale of online abuse against women in the UK and USA.

Our findings reinforce Amnesty International’s previous research into online abuse and violence against women online. Specifically, we found that:

* 7.1% of tweets sent to the women in the study were “problematic” or “abusive”. This amounts to 1.1 million tweets mentioning 778 women across the year, or one every 30 seconds.
* Women of colour, (black, Asian, Latinx and mixed-race women) were 34% more likely to be mentioned in abusive or problematic tweets than white women.
* Black women were disproportionately targeted, being 84% more likely than white women to be mentioned in abusive or problematic tweets.
* Online abuse targets women from across the political spectrum - politicians and journalists faced similar levels of online abuse and we observed both liberals and conservatives alike, as well as left and right leaning media organisations, were targeted.

Note: This analysis is specific to the group of women in the study and would likely differ if applied to other professions, countries or the wider population.

Online abuse against women on this scale should not and does not have to exist on social media platforms. Companies like Twitter have a responsibility to respect human rights, which means ensuring that women using the platform are able to express themselves freely and without fear.

**Amnesty International has repeatedly asked Twitter to make available meaningful and comprehensive data regarding the scale and nature of abuse on their platform, as well as how they are addressing it**. Such data will be invaluable for anyone seeking to understand and combat this barrier to women’s human rights online. In light of Twitter’s refusal to do so, it is our hope that this project will shed some insight into the scale of the problem.

## **Twitter - a toxic place for women**

For the past two years, Amnesty International has documented the scale and nature of abuse and violence against women online. In 2017, we commissioned an online [polling of women in 8 countries](https://www.amnesty.org/en/latest/news/2017/11/amnesty-reveals-alarming-impact-of-online-abuse-against-women/) about their experiences of abuse on social media platforms and [used data science](https://medium.com/@AmnestyInsights/unsocial-media-tracking-twitter-abuse-against-women-mps-fc28aeca498a) to analyse the abuse faced by female Members of Parliament (MPs) on Twitter prior to the UK’s 2017 snap election.

In March 2018, we released our report [Toxic Twitter: Violence and abuse against women online](https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1/) which outlines the human rights harms facing women on Twitter, and proposed concrete steps that the company can take to address these harms. The report found that as a company, Twitter is failing in its responsibility to respect women’s rights online by failing to adequately investigate and respond to reports of violence and abuse in a transparent manner which leads many women to silence or censor themselves on the platform.

We are experiencing a watershed moment with women around the world using their collective power to speak out about the abuse they face and amplify their voices through social media platforms. However, Twitter’s failure to effectively tackle violence and abuse on the platform has a chilling effect on freedom of expression online and undermines women’s mobilisation for equality and justice – particularly groups of women who already face discrimination and marginalisation.

**Amnesty International has repeatedly urged Twitter to publicly share comprehensive and meaningful information about reports of violence and abuse against women, as well as other groups, on the platform, and how they respond to it.**

On 12 December 2018 Twitter released an updated [Transparency Report](https://transparency.twitter.com) in which it included for the first time a section on '[Twitter Rules Enforcemen](https://transparency.twitter.com/en/twitter-rules-enforcement.html)t'. This was one of Amnesty International’s key recommendations to Twitter and we see the inclusion of this data as an encouraging step. The disclosure of this data sheds some light on levels of abuse reported on the platform, and how the company responds to it.

We are disappointed, however, that the information provided in the transparency report does not go far enough. For instance, the data Twitter includes on enforcement is not disaggregated by the category of abuse reported but is simply aggregated in two categories of ‘abuse’ and ‘hateful conduct’. Meaningful data on how Twitter responds to the different types of reports of abuse available in its reporting mechanism will increase understanding of how different categories of abuse, in particular tweets that direct hate against a protected category, are responded to. The data in Twitter’s Transparency report also fails to include any information about how long it takes moderators (on average) to respond to reports of abuse, how many decisions were overturned in an appeals process or any information about the number of content moderators employed per region and language to respond to reports of abuse on the platform.

In addition, Twitter's data states that 2.8 million unique accounts were reported for abuse of which Twitter actioned 248,000 – approximately 9%. However, the data published only reflects unique accounts that were reported for abuse and actioned. Twitter should also publish the total number of tweets reported for abuse and hateful conduct - disaggregated by category - in order to avoid potentially underplaying the true scale of abuse on the platform. Knowing more complete and robust figures on reports of abuse and hateful conduct on Twitter will ultimately allow policy makers, civil society and technologists to develop more concrete and comprehensive solutions to tackle the problem.

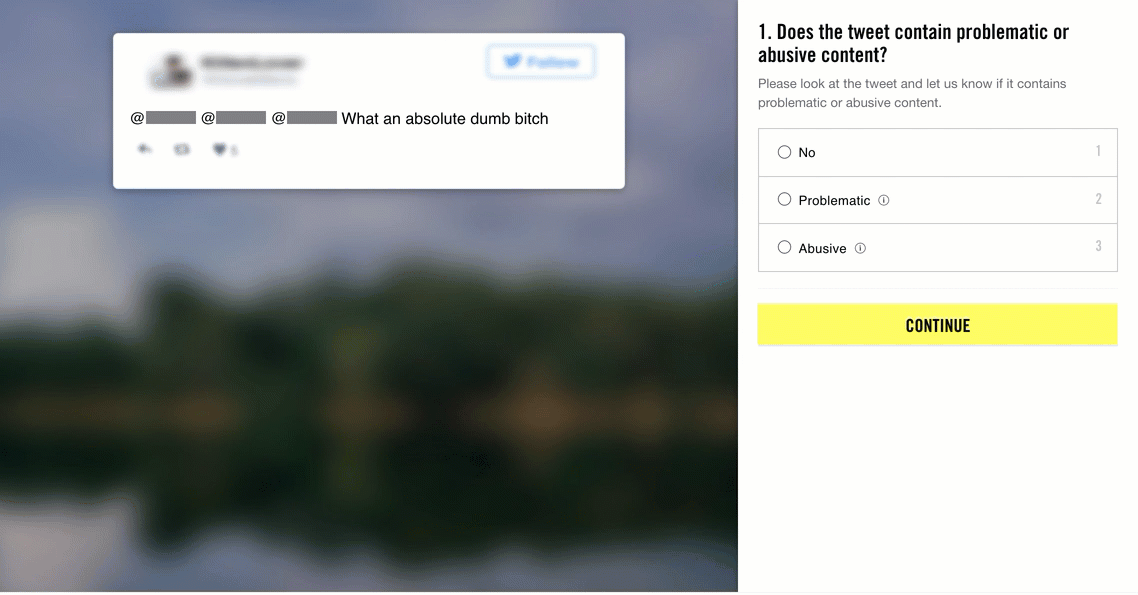
This study is an attempt to - in part - remedy this lack of data.

## **Crowd-sourcing the world’s biggest dataset of online abuse targeting women**

Following the launch of our Toxic Twitter report in March 2018, Amnesty International launched Troll Patrol, a global crowdsourcing effort to demonstrate the scale and nature of abuse that women continue to experience on Twitter.

Working with [Element AI,](https://www.elementai.com/) we first designed a large, unbiased dataset of thousands of tweets mentioning 778 women politicians and journalists from the UK and US. The sample of women selected included all parliamentarians from the UK Parliament and US Congress and Senate as well as women journalists from publications like the Daily Mail, Gal Dem, the Guardian, Pink News, the Sun in the UK and Breitbart and the New York Times in the USA.

More than **6,500** digital volunteers from around the world then took part in Troll Patrol, analysing 288,000 unique tweets to create a labelled dataset of abusive or problematic content. The volunteers were shown an anonymized tweet mentioning one of the women in our study, then were asked simple questions about whether the tweets were abusive or problematic, and if so, whether they revealed misogynistic, homophobic or racist abuse, or other types of violence. Each tweet was analysed by multiple people. The volunteers were given a tutorial and definitions and examples of abusive and problematic content, as well as an online forum where they could discuss the tweets with each other and with Amnesty International’s researchers.



The volunteers collectively dedicated an incredible **2,500 hours** analysing tweets - that’s the equivalent of someone working full-time for 1.5 years. They were a very diverse group of people- aged between 18 to 70 years old and from over 150 countries.

Three experts on violence and abuse against women also categorized a sample of 1,000 tweets, to ensure we were able to assess the quality of the tweets labelled by our digital volunteers.

Using a subset of the Decoders and experts’ categorization of the tweets, Element AI used data science to extrapolate the abuse analysis to the full 14.5 million tweets that mentioned the women journalists and politicians selected for our study. The results we are publishing are based on this.

The tweets in this study dated from January-December 2017, and were collected for analysis in March 2018. Crucially, because we could only download historical Twitter data, our sample did not include tweets that had already been deleted or tweets from accounts that were suspended or disabled during 2017, but only tweets that were still available on the platform in March 2018. Therefore, we can only assume the true scale of abuse was even higher than our results show. Some of the tweets in our sample have been deleted since March 2018.

Read our full methodology.

## 

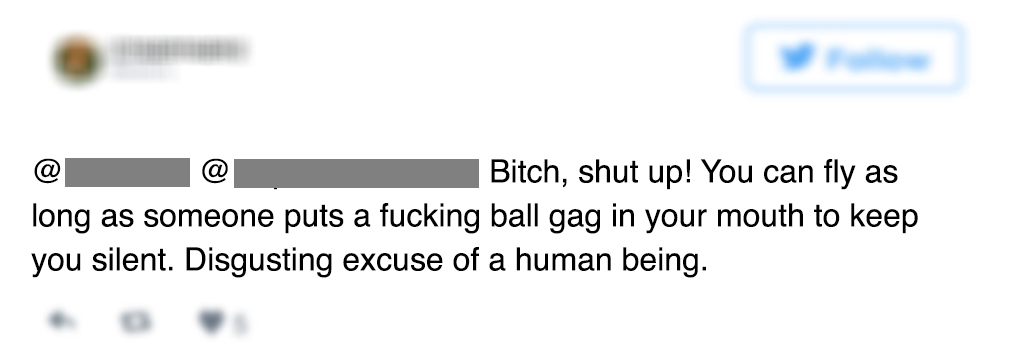
## **What did we find?**

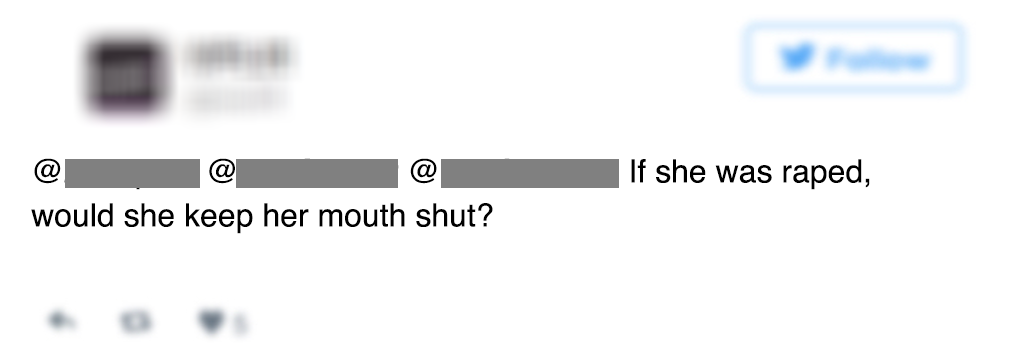
**1.1 million abusive or problematic tweets** sent to **778 women** journalists and politicians in 2017

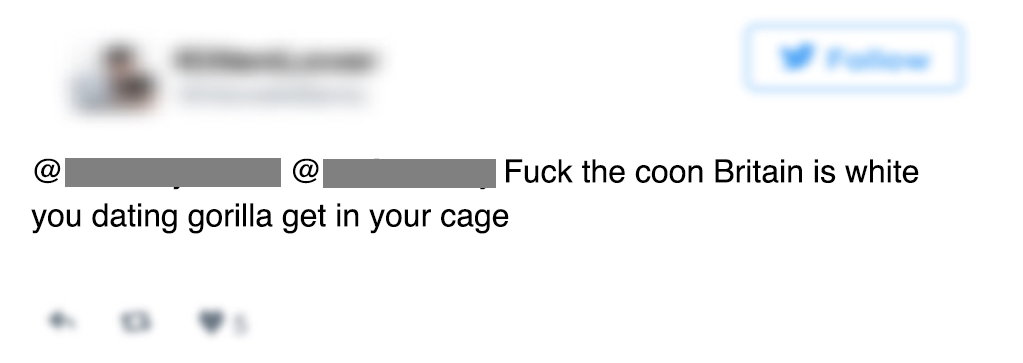
Our study found that **7.1%** of tweets sent to the women in the study were problematic or abusive. This amounts to **1.1 million problematic or abusive mentions of these 778 women** across the year, or **one every 30 seconds** on average. Women of colour were more likely to be impacted - with black women disproportionately targeted with problematic or abusive tweets. The study also reiterated that online abuse against women cuts across the political spectrum and does not pay heed to political party divisions.

All estimates (including percentages here) are subject to margins of error. For details, see methodology note.

**Abusive tweets** violate Twitter’s own rules and include content that promotes violence against or threats to people based on their race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. Examples include physical or sexual threats, wishes for the physical harm or death, reference to violent events, behaviour that incites fear or repeated slurs, epithets, racist and sexist tropes, or other content that degrades someone.

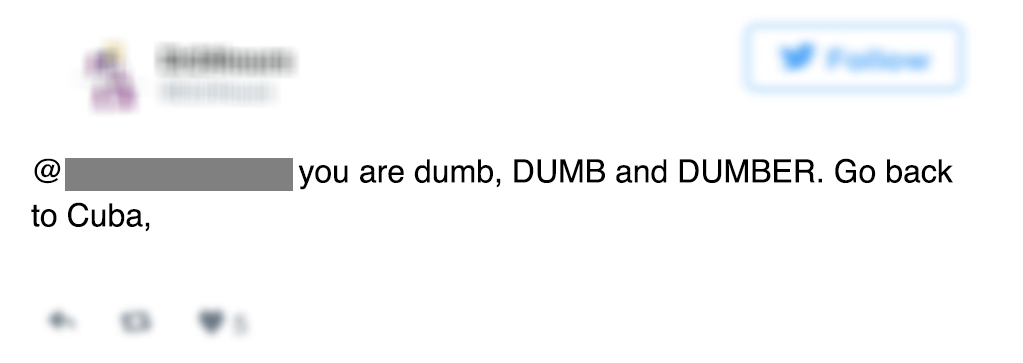


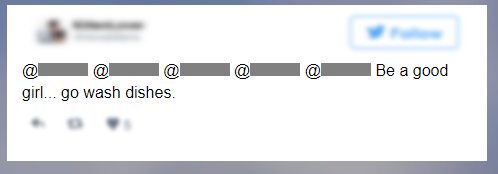




**Problematic tweets** contain hurtful or hostile content, especially if repeated to an individual on multiple or occasions, but do not necessarily meet the threshold of abuse. Problematic tweets can reinforce negative or harmful stereotypes against a group of individuals (e.g. negative stereotypes about a race or people who follow a certain religion). Such tweets may still have the effect of silencing an individual or groups of individuals.

It’s important to acknowledge that problematic tweets may qualify as legitimate speech and would not necessarily be subject to removal from the platform. We included problematic tweets because it is important to highlight the breadth and depth of toxicity on Twitter in its various forms and to recognize the cumulative effect that problematic content may have on the ability of women to freely expressing themselves on the platform. Twitter, itself, has committed to increase the 'health' of public conversations on the platform and the problematic content labelled in our study will be useful for Twitter to consider.

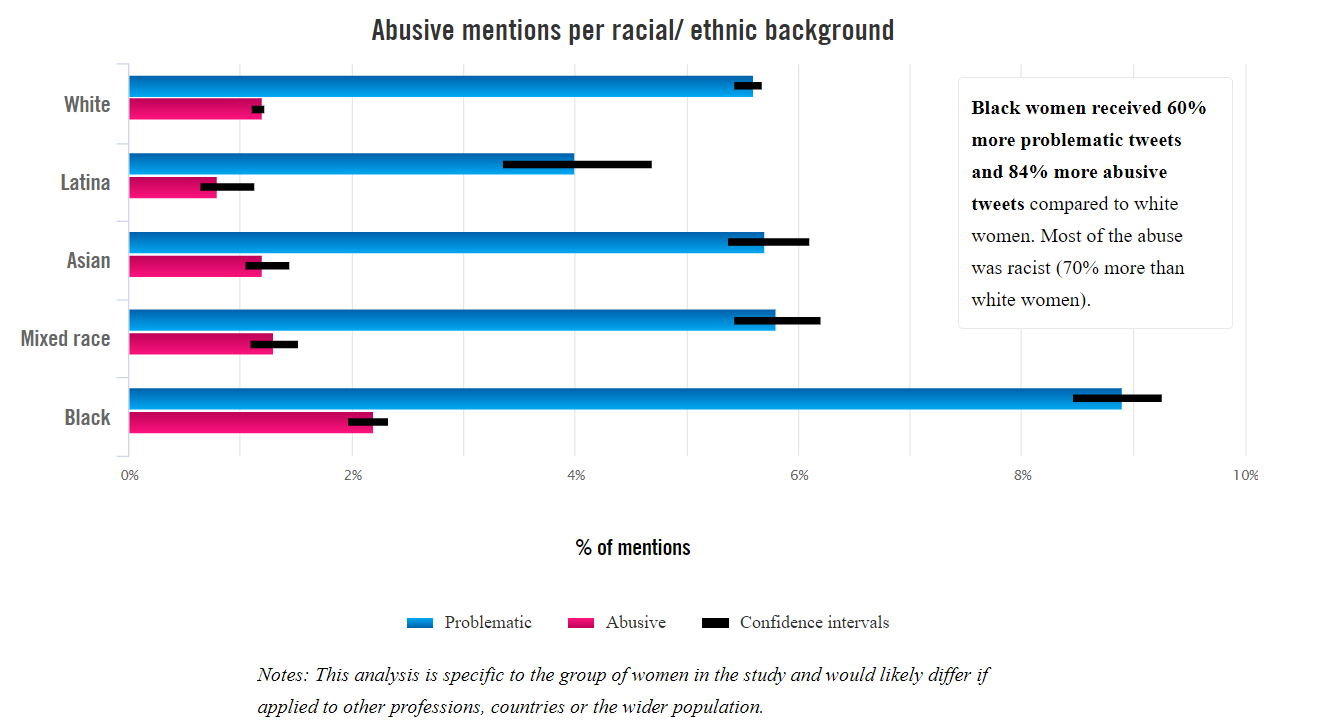




## **Black women were disproportionately targeted with online abuse**

Our study shows that as a group, women of colour, (black, Asian, Latinx and mixed-race women) were 34% more likely to be mentioned in abusive tweets than white women. Black women were particularly affected, being 84% more likely than white women to be mentioned in abusive tweets. This included all kinds of abuse - sexual and physical threats and misogyny, as well as racial slurs.

To better understand if and how women of colour were targeted as compared to white women on Twitter, we classified our sample of 778 journalists and politicians by race/ ethnic background. To determine race/ethnic background, we used official sources such as ethnic diversity studies but also information in the public domain such as public profiles, articles written by or about the women, Wikipedia pages, etc. It is important to note that this is a crude classification for the purpose of this analysis and is not necessarily a reflection of how each of the 778 women self-identify.



|  |  |  |
| --- | --- | --- |
|  | Problematic | Abusive |
| White | 5.6% | 1.2% |
| Latinx | 4.0% | 0.8% |
| Asian | 5.7% | 1.2% |
| Mixed | 5.8% | 1.3% |
| Black | 8.9% | 2.2% |

* White women - 6.7% or approximately 1 in 15 tweets received by white women in the sample were problematic or abusive
* Although in proportion Latinx women got less abuse than white women, they were 81% more likely to get specific threats of physical threats compared to white women.
* Mixed race women received abuse across all categories, including sexism, racism, physical and sexual threats.
* Asian women in the study were the most likely to be mentioned in tweets containing ethnic, racial and religious slurs - 70% more than white women.
* Black women received 60% more problematic tweets and 84% more abusive tweets compared to white women. Most of the abuse was racist (70% more than white women).

Twitter recently updated their [hateful conduct policy](https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy) recognising that some groups are disproportionately targeted with abuse online - this includes “women of colour, lesbian, gay, bisexual, transgender, queer, intersex, asexual individuals, marginalized and historically underrepresented communities”. We did not replicate this analysis for other aspects of women's identity such as religion, sexual orientation or gender identity. However, our previous research shows that women of colour, religious or ethnic minority women, lesbian, bisexual, transgender or intersex (LBTI) women, and other women with different identities will often experience abuse that targets them in unique or compounded way.

## **Online abuse against women cuts across the political spectrum**

### Journalists

Our analysis found that 7% of Twitter mentions of women journalists were problematic (5.8%) or abusive (1.2%) - that is 1 in every 14 mentions. In total, an estimate of 225,766 mentions sent to 454 journalists in our study were found to be abusive or problematic.

### Politicians

In terms of percentage, politicians received a similar rate of abuse to journalists - 5.85% of mentions were problematic and 1.27% abusive (7.12% in total). The overall estimated volumes

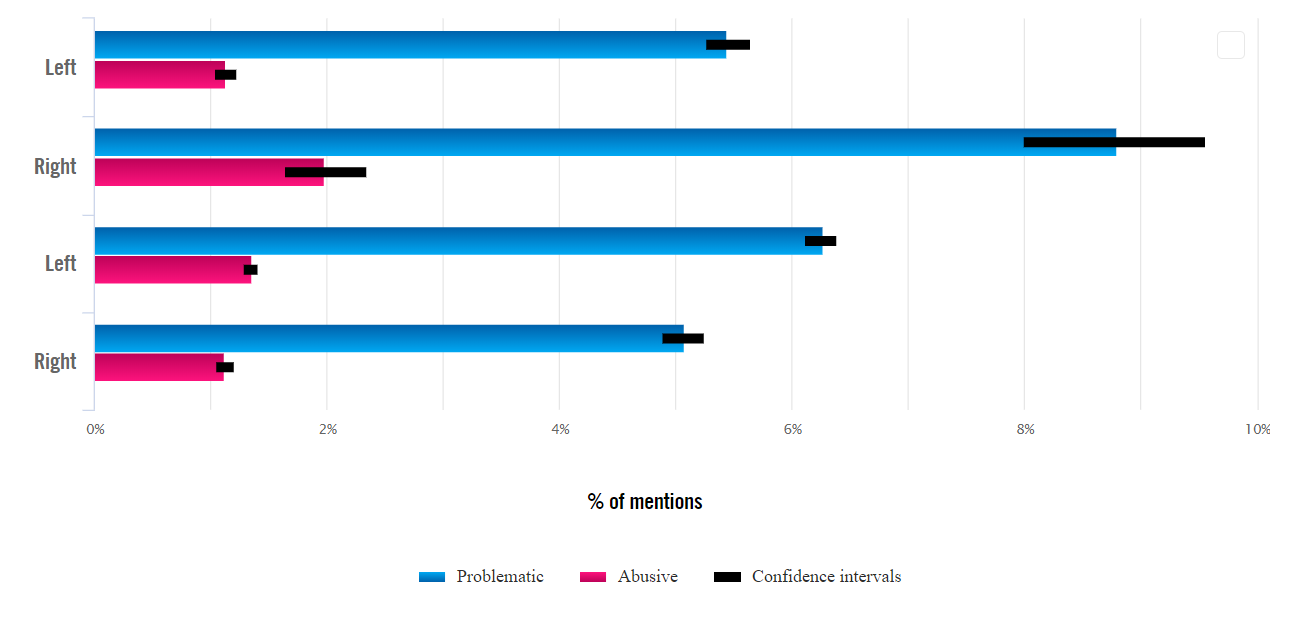
for 324 politicians studied was 867,136 (on average, the politicians got many more mentions than journalists).

Online abuse against women cuts across all political parties and is a bi-partisan issue that does not pay heed to political boundaries. Women politicians and journalists faced similar levels of online abuse and we observed that both liberals and conservatives alike, as well as left and right leaning media organisations, were targeted ([source for classification of media bias](https://www.allsides.com/media-bias/media-bias-ratings)).

On average, women working for left leaning parties and media organisations received slightly more problematic and abusive mentions (15% more) than women working for right leaning causes.

When disaggregating the data by profession we could observe that left leaning **politicians** – e.g. Democrat in the USA or Labour in the UK - are on the receiving end of 23% more problematic and abusive mentions that their counterparts in right-leaning parties - e,g. Republican (USA) or Conservative (UK). However, **journalists** working for right leaning media groups like Daily Mail, the Sun or Breitbart were mentioned in 64% more problematic and abusive tweets than journalists working at left leaning organisations like New York Times or the Guardian.

Abuse per political affiliation and profession



|  |  |  |
| --- | --- | --- |
|  | Problematic | Abusive |
| Journalist; Left leaning | 5.44% | 1.13% |
| Journalist; Right leaning | 8.80% | 1.98% |
| Politician; Left leaning | 6.27% | 1.35% |
| Politician; Right leaning | 5.07% | 1.12% |

The political association analysis is a crude classification for the purpose of this study. This was determined based on the institution (political parties, media organisations) that women were associated with, not based on analysis of the individuals involved ([source for classification of media bias](https://www.allsides.com/media-bias/media-bias-ratings)).

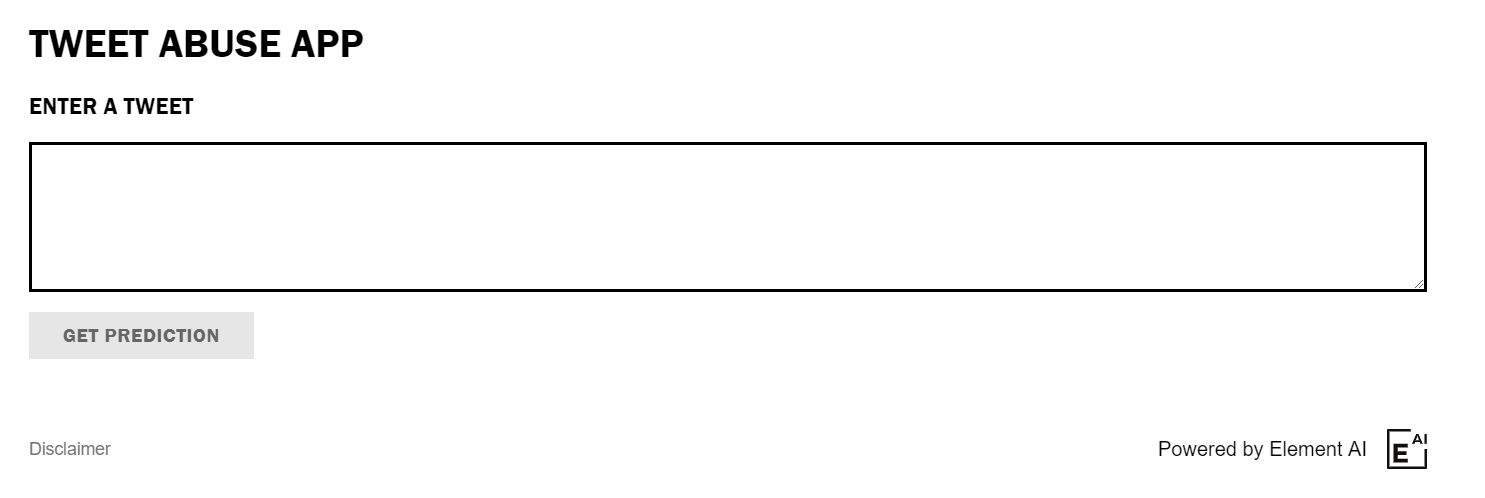
## **The role for machine learning in content moderation**

Large social media platforms have millions of users, host vast quantities of content, and must contend with massive volumes of abuse. As a result, they are increasingly turning to automated systems to help manage abuse on their platforms. To better understand the potential and risks of using machine learning in content moderation systems, we worked with Element AI to develop a machine learning model that would attempt to automate the process of detecting violence and abuse against women online.

While it is far from perfect, the model has advanced the state of the art compared to existing models and on some metrics, achieves results comparable to our digital volunteers at predicting abuse. Even so, it still achieves about a 50% accuracy level when compared to the judgement of our experts, meaning it identifies 2 in every 14 tweets as abusive or problematic, whereas our experts identified 1 in every 14 tweets as abusive or problematic.

Element AI is making its model available to try out here for three weeks to demonstrate the potential and current limitations of AI technology in this field.

## Demo



Machine learning systems are already being widely used to flag potentially problematic content to a human workforce: For example, [Facebook](https://motherboard.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works) and YouTube are using machine learning-powered software to scan and flag content to human moderators. Meanwhile, [Perspective API](https://www.perspectiveapi.com/), developed by Google Jigsaw, has been [used to flag potentially inappropriate content](https://www.technologyreview.com/s/603072/if-only-ai-could-save-us-from-ourselves/) for review on both Wikipedia and the New York Times comments section.

In a letter to Amnesty International, Twitter has called machine learning “one of the areas of greatest potential for tackling abusive users”. Twitter CEO Jack Dorsey has similarly [said](https://www.theguardian.com/technology/live/2018/sep/05/facebook-sheryl-sandberg-twitter-congress-russia?page=with:block-5b9023c5e4b0c59f70c18e44#liveblog-navigation) that “We think that we can reduce the amount of abuse and create technology to recognize it before a report has to be made.” Twitter has also said that it is focused on machine learning in an effort [to combat spam and automated accounts](https://blog.twitter.com/official/en_us/topics/company/2018/how-twitter-is-fighting-spam-and-malicious-automation.html), and that it has [begun acting against abusive accounts](https://slate.com/human-interest/2017/03/twitters-proactive-new-anti-harassment-effort-will-flag-users-as-potentially-abusive.html) that have not yet been reported.

However, the trend towards using machine learning to automate content moderation online also poses risks to human rights. For example, David Kaye, the UN Special Rapporteur on Freedom of Expression, has [noted](https://documents-dds-ny.un.org/doc/UNDOC/GEN/G18/096/72/PDF/G1809672.pdf?OpenElement) (paras 32-25) that “automation may provide value for companies assessing huge volumes of user-generated content.” He cautions, however, that in subject areas dealing with issues which require an analysis of context, such tools can be less useful, or even problematic.

We have already seen that there can be serious human rights consequences when algorithms mistakenly censor content. In June 2017, [Google announced four steps intended to fight terrorism online](https://www.blog.google/around-the-globe/google-europe/four-steps-were-taking-today-fight-online-terror/), among them more rigorous detection and faster removal of content related to violent extremism and terrorism. The automated flagging and removal of content resulted in the [accidental removal](https://www.amnesty.org/en/latest/news/2017/11/youtube-removals-threaten-evidence-and-the-people-that-provide-it/) of hundreds of thousands of YouTube videos uploaded by journalists, investigators, and human rights organizations.

The simple reality is that the use of machine learning necessarily accepts working within margins of error. For example, the decision to weight an algorithm towards greater *precision* will result in increased detection of genuinely abusive tweets, at the risk of missing abusive content which is more subtle (equivalent to casting the net too narrow). On the other hand, weighting an algorithm towards greater *recall* would capture a wider range of abusive content, at the risk of also capturing *false positives* - that is to say, content that should be protected as legitimate speech (equivalent to casting the net too wide). These trade-offs are value-based judgements with serious implications for freedom of expression and other human rights online.

Amnesty International and Element AI’s experience using machine learning to detect online abuse against women highlights the risks of leaving it to algorithms to determine what constitutes abuse. As it stands, automation may have a useful role to play in assessing trends or flagging content for human review, but it should, at best, be used to assist trained moderators, and certainly should not replace them. Human judgement by trained moderators remains crucial for contextual interpretation, such as examination of the intent, content and form of a piece of content, as well as assessing compliance with policies. It is vital that companies are transparent about how exactly they are using automated systems within their content moderation systems and that they publish information about the algorithms they develop.

## Moving forward

Amnesty International has repeatedly asked Twitter to make available meaningful and comprehensive data regarding the scale and nature of abuse on their platform, as well as how they are addressing it. Such data will be invaluable for anyone seeking to understand and combat this barrier to women’s human rights online. In light of Twitter’s refusal to do so, it is our hope that this project with Element AI will help shed some insight into the scale and nature of abuse on the platform, and also provide tools to others who wish to conduct statistical research on this topic.

This research also provides valuable insights into the potential role of automation in content moderation processes. As companies including Twitter embrace the use of machine learning to flag content for moderation, it is more important than ever that they are transparent about the algorithms they use. They should publish information about training data, methodologies, moderation policies and technical trade-offs (such as between greater precision or recall) for public scrutiny. At best, automation should be part of a larger content moderation system characterized by human judgement, greater transparency, rights of appeal and other safeguards.

Amnesty International’s full set of recommendations to Twitter are available [here](https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-8/#topanchor).

## Take action

[AI USA campaign](https://act.amnestyusa.org/page/22071/action/1?ea.tracking.id=messagingcategory_GEN)