A large-scale crowd-sourced analysis of abuse against women journalists and politicians on Twitter

Laure Delisle[*,†] Element AI
Alfredo Kalaitzis[*,†] Element AI
Krzysztof Majewski[†] Element AI
Archy de Berker[†] Element AI
Milena Marin[‡] Amnesty International
Julien Cornebise[†] Element AI

Abstract

We report the first, to the best of our knowledge, hand-in-hand collaboration between human rights activists and machine learners, leveraging crowd-sourcing to study online abuse against women on Twitter. On a technical front, we carefully curate an unbiased yet low-variance dataset of labeled tweets, analyze it to account for the variability of abuse perception, and establish baselines, preparing it for release to community research efforts. On a social impact front, this study provides the technical backbone for a media campaign aimed at raising public and deciders' awareness and elevating the standards expected from social media companies.

1 Introduction

Social media platforms have become a critical space for women and marginalized groups to express themselves at an unprecedented scale. Yet a stream of research by Amnesty International [1, 2] showed that many women are subject to targeted online violence and abuse, which denies them the right to use social media platforms equally, freely, and without fear. Being confronted with toxicity at a massive scale leaves a long-lasting effect on mental health, sometimes even resulting in withdrawal from public life altogether [3]. A first smaller-scale analysis of online abuse against women UK Members of Parliament (MPs) on Twitter [4, 5] proved the impact such targeted campaigns can have: it contributed to British Prime Minister Theresa May publicly calling out the impact of online abuse on democracy [6].

This laid the groundwork for the larger-scale *Troll Patrol* project that we present here: a joint effort by human rights researchers and technical experts to analyze millions of tweets through the help of online volunteers. Our main research result is the development of a dataset that could help in developing tools to aid online moderators. To that end, we *i)* Designed a large, enriched, yet unbiased dataset of hundreds of thousands of tweets; *ii)* Crowd-sourced its labeling to online volunteers; *iii)* Analyzed its quality via a thorough agreement analysis, to account for the personal variability of abuse perception; *iv)* Compared multiple baselines with the aim of classifying a larger dataset of millions of tweets. Beyond this collaboration, this should allow researchers worldwide to push the envelope on this very challenging task – one of many in natural language understanding [7].

The social impact Amnesty International is aiming for is ultimately to influence social media companies like Twitter into increasing investment and resources – under any form – dedicated to tackling online abuse against women. With this study, we contribute to this social impact by providing the research backbone for a planned media campaign in November 2018.

2 Crowd-sourcing an importance-sampled enriched set

Core to this study is the careful crafting of a large set of tweets followed by a massive crowd-sourced data labeling effort.

[*]Equal contribution. [†]{laure,freddie,km,archy,julien}@elementai.com; [‡]milena.marin@amnesty.org

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

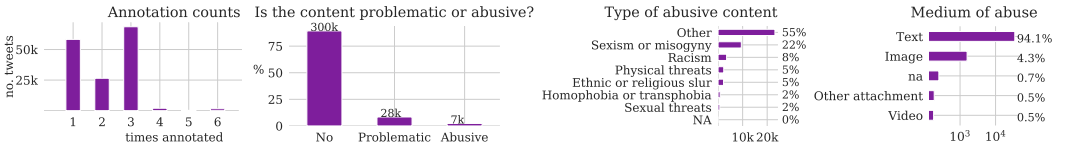


Figure 1: Left to right: Distribution of annotations-per-tweet: To analyze agreement, we used only tweets annotated more than twice ($\sim$73k); Values of `Contain Abuse` are ordinal; `Type` and `Medium` conditioned on `Contain Abuse` $\neq$ `No`: the majority of abuse is not easily classified, and the vast majority of abuse is textual.

Studied population: We selected 778 women politicians and journalists with an active, non-protected Twitter account, with fewer than 1 million followers, including most women British MPs and all US Congresswomen, and journalists from a range of news organizations representing a diversity of media bias. Full details are in Appendix A.

Tweet collection: 14.5M tweets mentioned at least one woman of interest during 2017. We obtained a subset of 10K per day sampled uniformly from Twitter's Firehose, minus tweets deleted since publication, totaling 2.2M tweets.

Pre-labeling selection: Taking into account the average labeling time per tweet from a pilot study, the expected duration of the campaign, and the expected graders' engagement, we targeted labeling at most 275K tweets in triplicate. We first selected 215K tweets, correcting the 10K daily cap using per-day stratified sampling proportional to each day's actual volume. While this sample is statistically representative of the actual tweet distributions, its class imbalance would induce high variance into any estimator, and waste the graders' engagement. We therefore enriched the dataset with 60K tweets pre-filtered through the Naive-Bayes classifier pre-trained in [5]. To maintain statistical non-bias, we keep track of the importance sampling weights.

Volunteers labeling via crowd-sourcing: Finally, these tweets, properly randomized, were deployed through Amnesty Decoders, the micro-tasking platform based on Hive [8] and Discourse [9] where Amnesty International engages digital volunteers (mostly existing members and supporters) in human rights research. Great effort was put into designing a user-friendly, interactive interface, accessible at [10] – see Appendix D for screenshots. After a video tutorial, volunteers were shown an anonymized tweet from the randomized sample, then were asked multiple-choice questions: 1) "Does the tweet contain problematic or abusive content?" (*No*, *Problematic*, *Abusive*). Unless their answer was *No*, the follow-up questions were "What type of problematic or abusive content does it contain?" (at least one of six) and (optional question) "What is the medium of abuse?" (one of four). See Fig. 1 for details and summary statistics. At all times they had access to definitions and examples of abusive and problematic content, and the typologies thereof – see Appendix E.

By August 2018, 157K unique tweets containing 167K mentions of the studied individuals had been categorized at least once, totalling 337K labels, thanks to the contribution of 4,537 online volunteers.

Experts labeling: In addition to engaging digital volunteers, Amnesty also asked three experts (Amnesty's researcher on online abuse against women, Amnesty's manager of the *Troll Patrol* project and an external expert in online abuse) to label a sub-set of 568 tweets. Those tweets were sampled from tweets labeled by exactly three volunteers as of June 8, 2018. To ensure low variance in the estimates, we once again used importance sampling, inflating the proportion of potentially abusive tweets by sampling 500 tweets uniformly from those labeled as "Abusive" by the Naive-Bayes classifier mentioned in [5] and "Basic Negative" by Crimson Hexagon's sentiment analysis (our Firehose access provider), and 500 tweets uniformly sampled on the remainder.

Re-weighting after importance sampling: To ensure that any inference or training based on the enriched sample is representative of the Twitter distribution, we use importance sampling to re-weight the tweets in the empirical distribution. The weights are defined as the ratio of the target distribution (as estimated by the daily counts) and the enriched distribution – see Appendix C for the full derivation of the weights.

3 Analysis and generalization

3.1 Agreement analysis

We quantified the agreement among raters – within crowd and within experts – using *Fleiss' kappa* (κ), a statistical measure of inter-rater agreement [11]. κ is designed for *nominal* (non-ordinal categorical) variables, e.g. Fig 1(c), whereas in ordinal variables κ tends to underestimate the agreement because it treats the disagreement between *Problematic* <> *Abusive* the same as *No* <> *Abusive*. We also use the *intra-class correlation* (ICC) [12] for ordinal categorical annotations, like `Contains Abuse`: *No* < *Problematic* < *Abusive*. We define κ and icc below, and further explain in Appendix B.

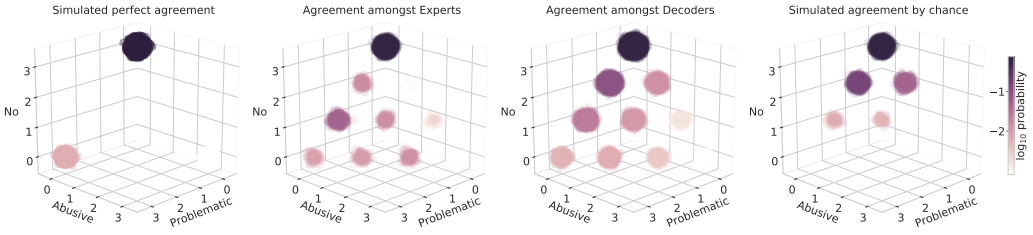


Figure 2: Visualizing the distribution of annotations $a^{(t)}$ (+ jitter for clarity) in the multinomial 2-simplex. The corners are events of *complete* agreement. The center is *no agreement* with the non-ordinal assumption, but partial agreement with ordinality. Left to right: Simulated perfect agreement, $a_c^{(t)} = 3, c \sim \hat{P}(C)$; Agreement among 3 experts on 1000 tweets: empirical probabilities are visually amplified by over-sampling $a^{(t)}$ to 20k; Agreement among $N = 3$ Decoders per tweet: if $N > 3$, raters are chosen randomly; Simulated agreement-by-chance only: $a^{(t)} \sim Multinomial(N = 3, p = \hat{P}(C))$. The multinomial assumes independence between trials. A hierarchical modeling approach can capture inter-rater dependence [13, 14].

Fleiss' kappa:A rater can annotate a tweet as class $c \in C = \{\text{No}, \text{Problematic}, \text{Abusive}\}$. The annotation $a = (a_{No}, a_{Pr}, a_{Ab}), a_c \in \{0, 1, 2, 3\}, \Sigma_c a_c = N$, contains the class counts for a tweet annotated by N raters. The overall agreement for a set of tweets T is $\kappa = \frac{1}{|T|}\Sigma_t \kappa^{(t)}$, where $\kappa^{(t)} = \frac{\Sigma_c r_c^{(t)} - \Sigma_c p_c^2}{1 - \Sigma_c p_c^2}$ is the *within-tweet* agreement, and $r_c = \frac{a_c(a_c-1)}{N(N-1)} \in [0, 1]$ is the fraction of pairs of raters that agree on c. $\hat{P}(C = c) = p_c$ is the empirical probability of c, hence $\Sigma_c p_c^2$ is the probability of agreement-by-chance.

ICC (intra-class correlation): Let $\mathbf{A} \in \mathbb{R}^{|T| \times N}$ be the matrix of annotations, $A_{i,j} \in \{0, 1, 2\}$ (ordinal values raters can assign), and each row is a tweet annotated by N random raters. The tweet-specific mean is $\mu_i = \frac{1}{N}\Sigma_j A_{i,j}$ and the overall mean is $\mu = \frac{1}{|T|}\Sigma_i \mu_i$. The within-tweet disagreement for tweet i is $V^{(i)} = \frac{1}{(N-1)}\Sigma_j \ (A_{i,j} - \mu_i)^2$, and its average $V_w = \frac{1}{|T|}\Sigma_i V^{(i)}$ is the overall within-tweet variance. Similarly, the *between*-tweet variance is $V_b = \frac{N}{|T|-1}\Sigma_i (\mu_i - \mu)^2$. The icc can now be expressed in terms of a one-way ANOVA [12]: $\text{icc} = \frac{V_b - V_w}{V_b + (N-1)V_w}$, the fraction of variation in annotations that is not explained by between-tweet disagreements.

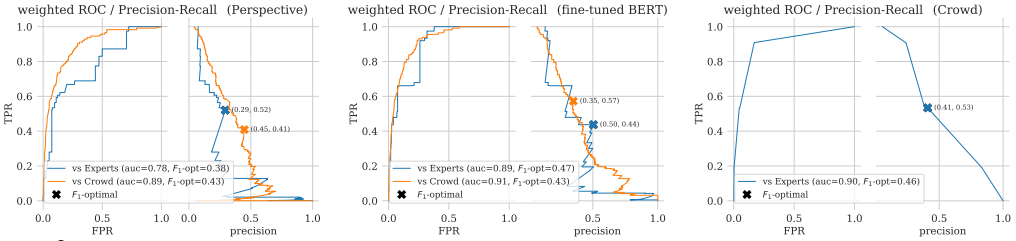


Figure 3: **Left & center**: performance of Perspective API and fine-tuned BERT classifiers, with respect to the experts' and crowd's labels. **Right**: performance of the crowd-as-a-classifier against the experts' labels. **Note**: *Recall* is equivalent to *TPR*, hence the y-axes (TPR<>Recall) of the two plots are aligned.

Table 1: Agreement per variable and per labeling cohort.

Labels from	Crowd		Experts	
	κ	ICC	κ	ICC
Contain Abuse	.26	.35	.54	.70
Type of Abuse	.16	-	.74	-

Table 2: Classifier performance vs. crowd vs. expert labels

Labels from	Crowd				Experts			
	Precision	*Recall*	F_1^*	*AP*	*Precision*	*Recall*	F_1^*	*AP*
Naive Bayes	.13	.25	.17	.11	.40	.27	.32	.21
Crimson Hexagon	.14	.40	.20	-	.05	.04	.04	-
Davidson et al.	.53	.27	.36	.25	.35	.46	.39	.25
Perspective API	.45	.41	**.43**	.34	.29	.52	.38	.25
Fine-tuned BERT	.35	.57	**.43**	**.40**	.50	.44	**.47**	.36
Crowd	-	-	-	-	.41	.53	.46	**.39**

Results: Table 1 and Figure 2 (mid left / mid right) show more agreement among the experts than among the volunteers – higher κ and ICC among the former. There is also more agreement when assessing the presence of abuse than when assessing the type of abuse.

3.2 Comparison of baseline classifiers

The core focus in this study is to build and analyze the dataset, with a view to extend that analysis to the remaining 2M unlabeled tweets using state of the art models. We prepare this follow-up research community effort by establishing baselines on various classification models.

Classifiers: In Table 2, Naive Bayes refers to the classifier from [5]. Crimson Hexagon refers to sentiment labels – Category and Emotion – from Crimson Hexagon. We also benchmarked the pre-trained classifier from Davidson et al. [15]. Perspective API refers to the public toxicity scoring API provided by Jigsaw [16, 17]. We also trained our own model, which combined a pre-trained BERT embedder [18] and an abuse-specific embedding trained from scratch. For details see F.

Methodology: For this analysis, we conflate the labels *Problematic* and *Abusive* into one positive (Abusive) class. The crowd labels are the majority votes over labels on tweets labeled by exactly three volunteers. The expert labels are majority votes over labels from the three domain experts mentioned in Section 2. For Crimson Hexagon, we define Abusive as the intersection of `Category = Basic Negative` and `Emotion = Anger | Disgust`.

Results: Table 2 shows the F_1^* (optimal F_1 score), corresponding precision and recall, and the Average Precision (AP), to evaluate several abuse detection classifiers with respect to labels from the crowd and from the experts.

4 Discussion

Dataset availability and reproducibility: Amnesty International intends to publish as much of the dataset as possible to encourage replication and further research on the topic. At the very least the URLs of the tweets and the grades will be made public. Publishing the actual tweets is more delicate due to Twitter Terms and Conditions. Publishing the meta-data on the graders (gender, location) would be of great interest, but is still under discussion from an ethical point of view.

Future work: We aim to eventually apply different classifiers to the whole unlabeled dataset, so as to scale up the human rights researchers' work by sifting through the huge amount of tweets. As shown in Section 3.2, this will require a careful tuning of models to increase the precision beyond its current performance. In parallel, we also want to analyze the scale, typology and intersectionality of abuse, either on the labeled set or on the classified extra 2M tweets, for the planned media campaign.

Social impact: The sheer volume of hateful speech on social media has recently prompted governments to put strong pressure on social media companies to remove such speech [19]. The moderation of abusive messages at scale requires some form of automated assistance. Our results highlight the double challenge of automatic abuse classification: the subjectivity in the labels and the limited ability of state-of-the-art classifiers to generalize beyond training data. This all points toward the need for systems where human subtlety and context awareness are empowered by automatic pre-screening.

Whether the companies themselves should be trusted with (or required to implement) such moderation, or whether they should fund or be supervised by a third-party neutral watchdog, goes far beyond a purely technical conversation. This is why collaboration between technical experts (machine learners, data scientists) and domain experts (human rights researchers, anti-censorship activist, etc.), as well as society in a broader sense, is so important for genuinely impactful AI for Social Good efforts.

Acknowledgements

We are extremely grateful to all the volunteers from Amnesty Decoders for their hard work. We would like to thank Nasrin Baratalipour, Francis Duplessis, Rusheel Shahani and Andrei Ungur for modelling support. We also thank Jerome Pasquero for his support, and the Perspective team for access to their API.

References

[1] Azmina Dhrodia. Unsocial media: The real toll of online abuse against women, November 2017. URL www.medium.com/amnesty-insights/unsocial-media-the-real-toll-of-online-abuse-against-women-37134ddab3f4. [Online; posted 20-November-2017].

[2] Amnesty International. Toxic Twitter, 2018. URL https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1.

[3] Committee on Standards in Public Life. Intimidation in public life. 2017. URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/666927/6.3637_CO_v6_061217_Web3.1__2_.pdf.

[4] Azmina Dhrodia. Unsocial media: Tracking Twitter abuse against women MPs, September 2017. URL www.medium.com/@AmnestyInsights/unsocial-media-tracking-{T}witter-abuse-against-women-mps-fc28aeca498a. [Online; posted 04-September-2017].

[5] Ekaterina Stambolieva. Technical report, Amnesty International, 2017. URL https://drive.google.com/file/d/0B3bg_SJKE9GOenpaekZ4eXRBWk0/view.

[6] The Guardian. Theresa may calls abuse in public life 'a threat to democracy', 2018. URL https://www.theguardian.com/society/2018/feb/05/theresa-may-calls-abuse-in-public-life-a-threat-to-democracy-online-social-media.

[7] Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. Identifying sarcasm in Twitter: A closer look. pages 581–586, 2011. URL http://dl.acm.org/citation.cfm?id=2002736.2002850.

[8] New York Times Labs. Hive: Open-source crowdsourcing framework, 2014. URL http://nytlabs.com/blog/2014/12/09/hive-open-source-crowdsourcing-framework/.

[9] Discourse. Discussion forum. URL https://www.discourse.org/.

[10] Troll patrol, 2018. URL https://decoders.amnesty.org/projects/troll-patrol.

[11] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[12] Patrick E Shrout and Joseph L Fleiss. Intra-class correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

[13] Alfredo Kalaitzis and Ricardo Silva. Flexible sampling of discrete data correlations without the marginal distributions. In *Advances in Neural Information Processing Systems, pages=2517–2525, year=2013, keywords=sampling, discrete data, copulas*.

[14] Ricardo Silva and Alfredo Kalaitzis. Bayesian inference via projections. *Statistics and Computing*, 25(4):739–753, 2015.

[15] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language, 2017. URL https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665.

[16] Jigsaw. Perspective API. URL https://www.perspectiveapi.com.

[17] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*, 2017.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[19] Björn Gambäck and Utpal Kumar Sikdar. Using convolutional neural networks to classify hate-speech. 2017.

[20] Randal Douc and Eric Moulines. Limit theorems for weighted samples with applications to sequential Monte Carlo methods. *Annals of Statistics*, 36(5):2344–2376, 2008.

[21] Hugging Face Development Team. PyTorch pretrained BERT. URL https://github.com/huggingface/pytorch-pretrained-BERT.

[22] Luis von Ahn. List of abusive words. URL https://www.cs.cmu.edu/~biglou/resources/.

A Population definition

We selected politicians and journalists with an active, non protected Twitter account, with fewer than 1 million followers. The group included:

- All women members of the British Parliament (220, including 22 women who left parliament during the June 2017 elections and excluding one politician with over 1 million followers);
- All women in the United States Congress (107, excluding 3 politicians with more than 1 million followers);
- And women journalists working at the following news organizations, selected to represent a diversity of media bias:
  - Breitbart (16),
  - Daily Mail (78),
  - The Sun (54),
  - The Guardian (124),
  - The New York Times (278),
  - Gal-Dem (23),
  - PinkNews (9).

B Agreement analysis

B.1 Fleiss' Kappa

Notation A rater can annotate a tweet as class $c \in C = \{\text{No, Problematic, Abusive}\}$. The annotation tuple $a = (a_{No},\ a_{Pr},\ a_{Ab})$, $a_c \in \{0, 1, 2, \dots\}$, $\Sigma_c a_c = N$, contains the class-specific counts for a tweet annotated by N raters.

Estimation of κ The *within-class* agreement-ratio $r_c = \frac{a_c\ (a_c - 1)}{N\ (N-1)} \in [0, 1]$ is the ratio of pairs of raters that agree on c, over the total of pairs of N raters. $\hat{P}(C = c) = p_c$ is the empirical marginal probability of class c. Hence, $\Sigma_c\ p_c^2$ is the overall probability of agreement by chance across a dataset of tweets. For a specific tweet t we can compute the *within-class* agreement $\kappa_c^{(t)} = \frac{r_c^{(t)} - p_c^2}{1\ -\ \Sigma_c p_c^2}$, where the numerator is the *agreement-above-chance* attained on c, and the denominator is the *best-case-scenario* (maximal) agreement-above-chance attainable across classes. Hence $\kappa_c^{(t)}$ is the fraction that the attained agreement in c contributes to the best-case scenario, while accounting for agreement-by-chance. Finally, the *within-tweet* agreement for a tweet t is the sum across classes, $\kappa^{(t)} = \Sigma_c \kappa_c^{(t)}$, and the overall agreement across a set of tweets T is the expectation

$$\kappa = \mathbb{E}_T[\kappa^{(.)}] \approx \tfrac{1}{|T|} \Sigma_t \kappa^{(t)}. \tag{1}$$

B.2 ICC (intra-class correlation)

Notation We denote the matrix of annotation as $\mathbf{A} \in \mathbb{R}^{|T|\ \times\ N}$. In this work, $A_{i,\ j} \in \{0, 1, 2\}$ (ordinal values raters can assign), where each row represents a tweet annotated by N random raters.

Algorithm The tweet-specific mean is $\mu_i = \frac{1}{N} \Sigma_j A_{i,j}$ and the overall mean is $\mu = \frac{1}{|T|} \Sigma_i \mu_i$. We can express the *within-tweet* disagreement as the within-tweet variance $V^{(i)} = \frac{1}{(N-1)} \Sigma_j\ (A_{i,j} - \mu_i)^2$. Then the average of within-tweet disagreements expresses the overall within-tweet variance, $V_w = \frac{1}{|T|} \Sigma_i V^{(i)}$. Similarly, the *between-tweet variance* is $V_b = \frac{N}{|T|-1} \Sigma_i (\mu_i - \mu)^2$. Note that the i-th tweet is *polarized* when $V^{(i)}$ is maximized, i.e. half of the raters choose 0 and the other half choose 2. In the extreme scenario that all tweets are maximally polarizing, $V_b = 0$. Therefore V_b expresses the overall tendency for *disagreement-by-chance*. All classes of ICC are equivalent to a type of ANOVA (*ANalysis Of VAriance*) in linear mixed-effects model of annotations [12]. In our case,

$$\text{icc}(1, k) = \frac{V_b - V_w}{V_b + (N-1) V_w} \tag{2}$$

Intuitively, the ANOVA framework defines agreement as the fraction of variation in annotations that is not explained by between-tweet disagreements.

Systemic disagreement As mentioned above, in extreme scenarios where V_b is small, the ICC can be negative. Negative κ and icc values might seem like an artifact of degenerate or extreme data, only to be dismissed as *no agreement* in the downstream analysis. At closer inspection, the numerator shows that subtracting the agreement-by-chance yields a measure of systemic disagreement: e.g. expecting $P(\text{agreement-by-chance}) = 0.9$ but observing agreement only 20% of the time, implies a systemic cause for polarizing opinions (e.g. controversial content, raters annotating with different rules).

C Importance sampling analysis

Population and Crimson sets W and C: We denote the population (*World set*) as W, and the sample obtained from the Twitter firehose (*Crimson set*) as C. Members of the sets W and C are observation tuples (t, k, d), where t is the text content of a tweet, $k \in \{0, 1\}$ is the output of a Naive Bayes Classifier $\text{NBC} : t \mapsto k$, and d is the day in 2017 that a tweet was published:

$$C \subset W = \{(t, k, d)\} \tag{3}$$

Distributions p_W and p_C: We define p_W and p_C as the probability mass over sets W and C, respectively, and any marginals and conditionals thereof:

$$p_W(t, k, d) = p_W(t, k|d)\, p_W(d) \tag{4}$$
$$p_C(t, k, d) = p_C(t, k|d)\, p_C(d) \tag{5}$$

The density $p_W(d)$ is directly available from the daily total volumes n_d of tweets matching the query, total that is provided by Crimson Hexagon alongside the smaller sampled set C:

$$n_d = |\{(t, k, d') \in W : d' = d\}| \text{ provided as metadata,}$$
$$p_W(d) = \frac{n_d}{\sum_{d'} n_{d'}}\,. \tag{6}$$

The *Crimson set* C is constructed by uniform sampling over tweets in W, such that for any day d, the conditional probabilities over both sets are equal:

$$p_C(t, k|d) = p_W(t, k|d) \tag{7}$$

Then, using eq. (7) in (5):

$$p_C(t, k, d) = p_W(t, k|d)\, p_C(d) \tag{8}$$

Constructed set A: The final set A is defined as the union

$$A = B \cup F, \tag{9}$$

where

$$B = \{(t, k, d) \sim p_B(t, k, d) \simeq p_W(t, k|d)\, \hat{p}_W(d) \simeq\ p_W(t, k, d)\} \tag{10}$$

approximates the world joint distribution through stratified sampling per day, and

$$F = \{(t, k, d) \in C \backslash B : k = 1\} \tag{11}$$

is an enriched sample resulting from pre-filtering by a simple Naive Bayes classifier. The cardinalities of these sets are: $|C| = 2.2M$, $|B| = 215k$, $|F| = 60k$ and $|A| = 275k$.
With $\beta = \frac{|F|}{|B|+|F|}$, and $z(d)$ a normalizing constant depending on d:

$$p_A(t, k|d) = \frac{\beta \mathbb{I}(k = 1)\, p_A(t, k|d) + (1 - \beta)\, p_A(t, k|d)}{z(d)} \tag{12}$$

where $\mathbb{I}(.)$ is the indicator function.
The conditional probabilities of a tweet are identical in W and A:

$$p_W(t|k, d) = p_A(t|k, d)\,. \tag{13}$$

Combining equations (13) and (12) leads to:

$$p_A(t, k|d) \propto \beta \mathbb{I}(k = 1)\, p_W(t|k, d)\, p_W(k|d) + (1 - \beta)\, p_W(t|k, d)\, p_W(k|d)\,. \tag{14}$$

Importance weights w_i: Estimating statistics on the world set W using the samples in set A can be achieved using importance sampling, i.e. assigning a specific weight $p_W(t,k,d)/p_A(t,k,d)$ to each triplet (t,k,d) in A.

For each tweet $(t,k,d) \in A$, we define the weighting function w

$$\begin{aligned} w(t,k,d) &= \frac{p_W(t,k,d)}{p_A(t,k,d)} \\ &= \frac{p_W(t|k,d)\, p_W(k,d)}{p_A(t|k,d)\, p_A(k,d)} \,. \end{aligned} \tag{15}$$

Injecting equation (13) in equation (15), we can simplify by $p_W(t|k,d)$:

$$\begin{aligned} w(t,k,d) &= \frac{p_W(k,d)}{p_A(k,d)} \\ &= \frac{p_W(k|d)\, p_W(d)}{p_A(k|d)\, p_A(d)} \,. \end{aligned} \tag{16}$$

Since A is a finite set, the probability mass functions $p_A(k|d)$ and $p_A(d)$ in equation (16) are directly accessible by simple counting.

The probability mass functions $p_W(d)$ is known from (6). The term $p_W(k|d)$ is not available in closed form, but can be estimated straightforwardly. Indeed from equation (7) we have $p_W(k|d) = p_C(k|d)$, and the latter can be estimated by simple counting on C, leading to empirical estimate:

$$\hat{p}_W(k|d) = \frac{|\{(t,k',d') \in C : k' = k, d' = d\}|}{|\{(t,k',d') \in C : k' = k\}|} \,. \tag{17}$$

This leads to the final plug-in estimator of the importance weights:

$$\hat{w}(t,k,d) = \frac{\hat{p}_W(k|d) p_W(d)}{p_A(k|d) p_A(d)} \,. \tag{18}$$

For any given function $f(t,k,d)$, we therefore estimate its expectation in the whole population W using the self-normalized importance estimator:

$$\hat{\mathbb{E}}_W[f(t,k,d)] = \sum_{(t_i,k_i,d_i) \in A} \frac{w_i}{\sum_j w_j} f(t_i,k_i,d_i) \,. \tag{19}$$

where for any tweet (t_i,k_i,d_i) with GUID (Globally Unique Identifier) i we use the estimated unnormalized importance weight $w_i := \hat{w}(t_i,k_i,d_i)$.

Note that for full mathematical rigour, the asymptotic consistency of the importance sampling estimator $\hat{\mathbb{E}}_W$ could be proven by showing that the replacement of the density estimator $\hat{p}_W$ in the plug-in estimator $\hat{w}$ is asymptotically valid. Such a proof could proceed along the lines of [20], but is outside of the scope of this article.

D Labeling tool screenshots

The workflow presented to each grader by the labelling tool is illustrated in Figure 4 and Figure 5.

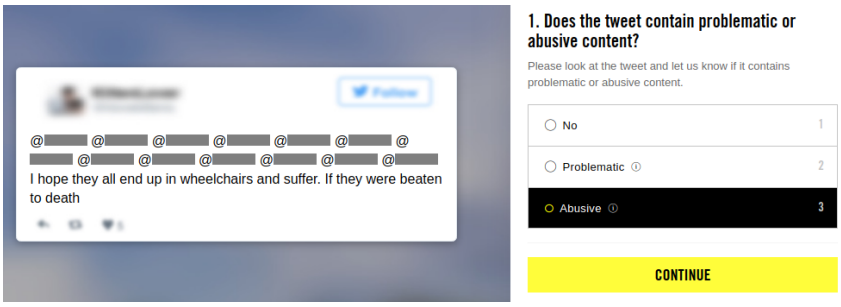
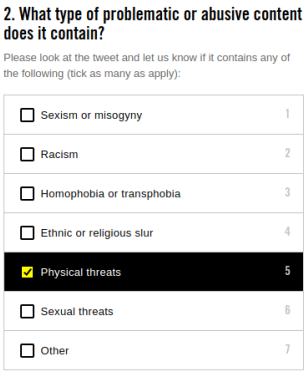
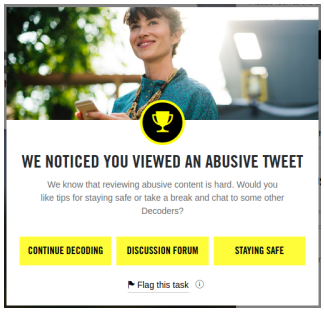


Figure 4: First stage of labeling: Initial screen showing an anonymized tweet, with anonymized handles and first question.



(a) Second stage of labeling: Identification of the type of abuse.

(b) Third, optional stage of labeling: Identification of the part of the tweet that carries the abuse.



(c) Warning displayed after classifying a tweet as abusive, to minimize the impact on the labelers' mental health.

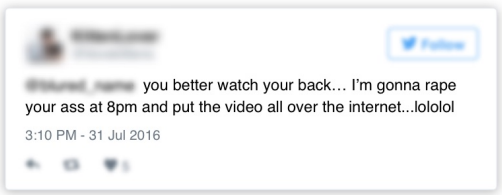
Figure 5: Follow-up stages, conditional on the first stage of labeling.

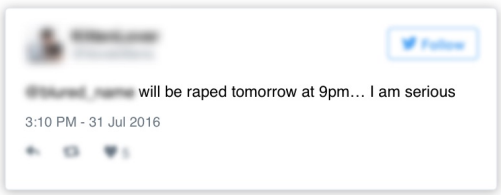
E Definitions and examples used in Troll Patrol - Trigger Warning

Abusive content Abusive content violates Twitter's own rules and includes tweets that promote violence against or threaten people based on their race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.

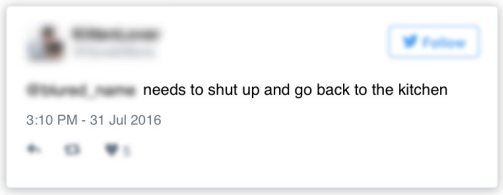
Examples include physical or sexual threats, wishes for the physical harm or death, reference to violent events, behaviour that incites fear or repeated slurs, epithets, racist and sexist tropes, or other content that degrades someone. For more information, see Twitter's hateful conduct policy.

In examples shown below, tweets were anonymized and only show a standard template (incl. the author handle, the author profile picture, the tweet date and time, likes and retweets).



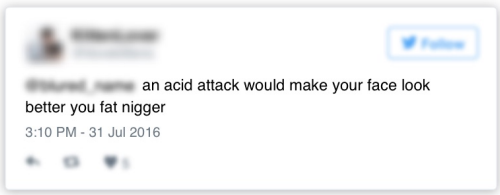


Problematic content Hurtful or hostile content, especially if it were repeated to an individual on multiple or cumulative occasions, but not as intense as an abusive tweet. It can reinforce negative or harmful stereotypes against a group of individuals (e.g. negative stereotypes about a race or people who follow a certain religion). Such tweets may have the effect of silencing an individual or groups of individuals.



Sexism or misogyny Insulting or abusive content directed at women based on their gender, including content intended to shame, intimidate or degrade women. It can include profanity, threats, slurs and insulting epithets.

Racism Discriminatory, offensive or insulting content directed at a woman based on her race, including content that aims to attack, harm, belittle, humiliate or undermine her.

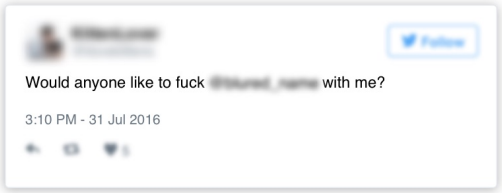


Homophobia or transphobia Discriminatory, offensive or insulting content directed at a woman based on her sexual orientation, gender identity or gender expression. This includes negative comments towards bisexual, homosexual and transgender people.

Ethnic or religious slur Discriminatory, offensive or insulting content directed at a woman based on her ethnic or religious identities.

Physical threats Direct or indirect threats of physical violence or wishes for serious physical harm, death, or disease.

Sexual threats Direct or indirect threats of sexual violence or wishes for rape or other forms of sexual assault.



Other There will be some tweets that fall under the 'other category' that are problematic and/or abusive. For example, statements that target a user's disability, be it physical or mental, or content that attacks a woman's nationality, health status, legal status, employment, etc.

F Abuse Classification Model

Model architecture We used a pretrained BERT model [18] (12 layers, 768 units per layer) as the basis for our classification model. We took the final-layer representation of the first token in the sequence as a fixed-length tweet embedding (see Figure 3 of [18]). The model was implemented in Pytorch, and made use of the BERT implementation provided at [21], which in turn utilizes a pre-trained model provided by Google.

To account for out-of-vocabulary abusive words, we added a second single-layer word embedding (128 units), which we trained from scratch with a limited abusive vocabulary. This vocabulary included a list of 1300 'possibly abusive' words available online [22], and the 1000 words which occurred most disproportionately in the abusive class of the training data. To obtain a fixed-length representation from this embedder, we took the mean across words in each tweet.

We concatenated these two representations to obtain a fixed-length tweet representation (of length 896), and passed through a pair of fully-connected layers of 64 units each before returning a decision via a binary softmax layer.

Data We split the crowd-sourced data into train, validation, and test sets (90 : 5 : 5). We adjust all reported performance metrics for the original importance sampling (see C).

Training We trained the model end-to-end with stochastic gradient descent (learning rate = 0.0001, momentum=0.9) for 11 epochs, minimizing a cross-entropy loss.

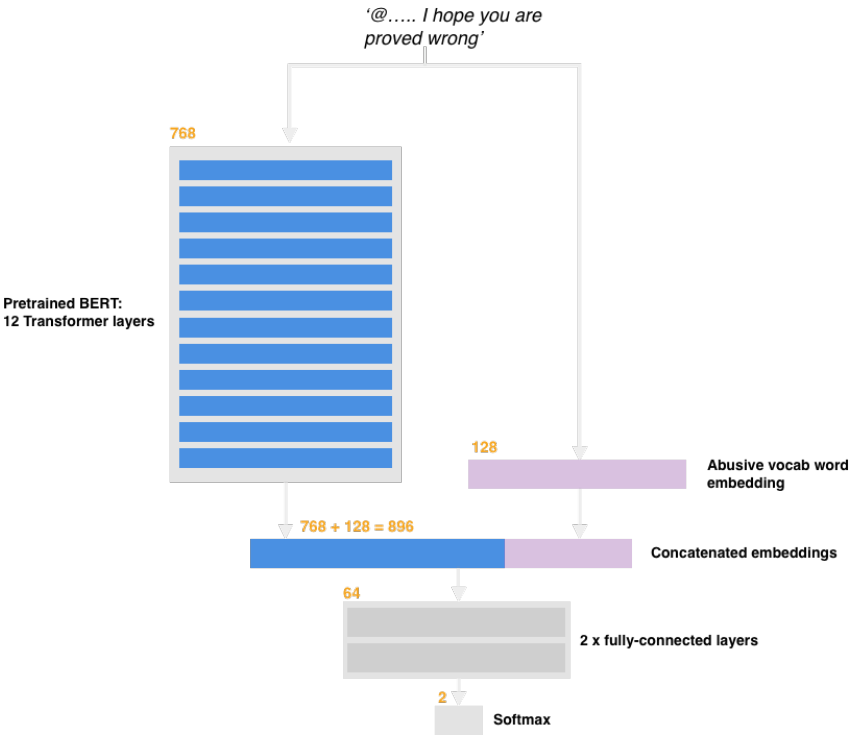


Figure 6: We combined a pre-trained BERT model with a word embedding exclusively including abusive words. The two embeddings were concatenated and passed through a pair of fully-connected layers before a softmax layer returned the prediction. Numbers in orange are layer widths.