Troll Patrol Methodology Note

Laure Delisle* †
Alfredo Kalaitzis* †
Krzysztof Majewski†
Archy de Berker†
Milena Marin‡
Julien Cornebise†

v1.1 - December 19, 2018

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.

Changelog

Version	Date	Changes
1.1	2018-12-19	Added table of contents and changelog, updated title.
1.0	2018-12-17	Initial release.

Contents

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

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 (Dhrodia, 2017a; International, 2018) 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 (on Standards in Public Life, 2017). A first smaller-scale analysis of online abuse against women UK Members of Parliament (MPs) on Twitter (Dhrodia, 2017b; Stambolieva, 2017) 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 (Guardian, 2018).

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 (González-Ibáñez et al., 2011).

The social impact Amnesty International is aiming for is ultimately to influence social media companies like Twitter into increasing investment and resources 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

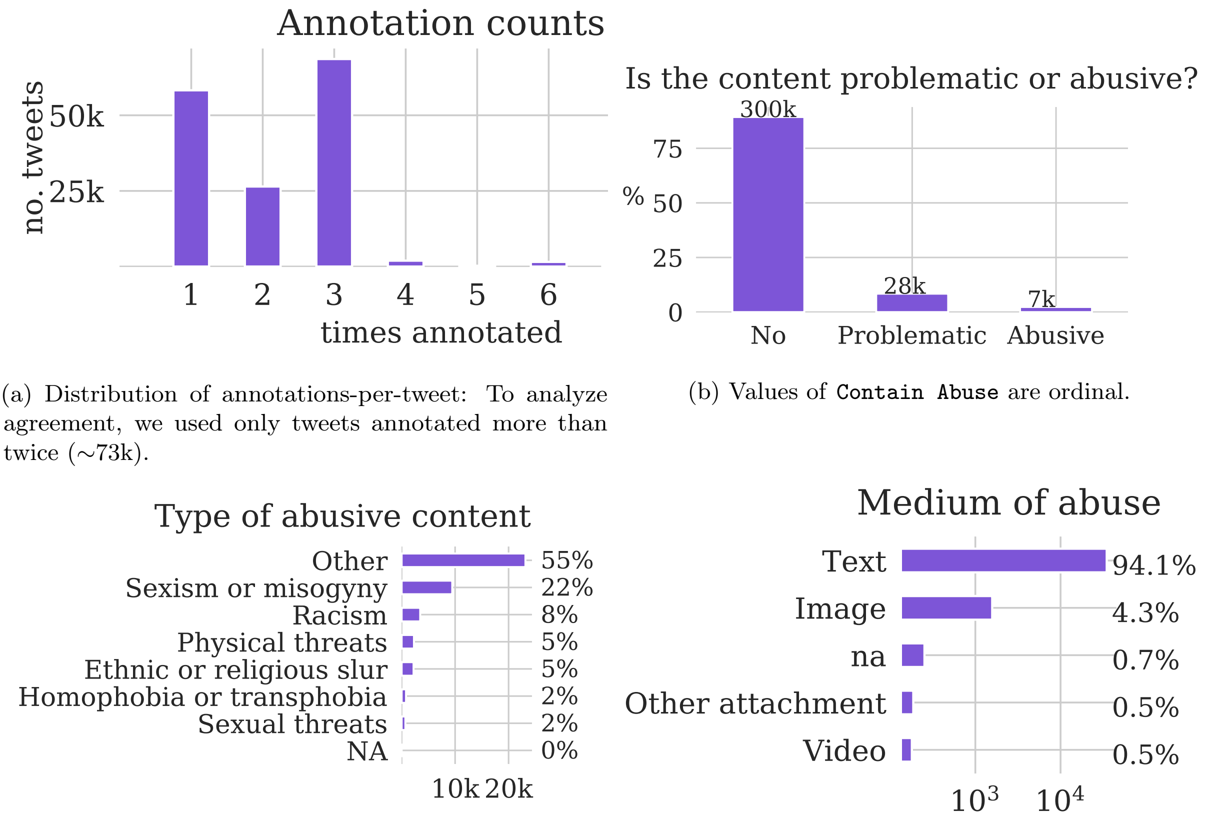
Core to this study is the careful crafting of a large set of tweets followed by a massive crowd-sourced data labeling effort.

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 political affiliations. 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 by Stambolieva (2017). 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 (Labs, 2014) and Discourse (Discourse) 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 Tro (2018) – 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.



(a) Distribution of annotations-per-tweet: To analyze agreement, we used only tweets annotated more than twice (∼73k).

(b) Values of `Contain Abuse` are ordinal.

(c) `Type` conditioned on `Contain Abuse` ≠ `No`: the majority of abuse is not easily classified.

(d) `Medium` conditioned on `Contain Abuse` ≠ `No`: the vast majority of abuse is textual.

Figure 1: Distribution of annotations.

By August 2018, 288K unique tweets had been categorized at least once, totalling 631K labels, thanks to the contribution of 6, 524 online volunteers. Focusing on the whole year 2017 and discarding early 2018, to avoid seasonal effects, and considering politicians and print journalists (for comparable exposure), we used a subset of the 157K unique labeled tweets, amounting to 167K mentions and 337K labels for our analysis.

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 1, 000 tweets, of which we used a subset of 568 tweets for the reasons discussed in the previous paragraph. Those tweets were sampled from tweets labeled by 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 284 tweets uniformly from those labeled as "Abusive" by the Naive-Bayes classifier mentioned in Stambolieva (2017) and "Basic Negative" by Crimson Hexagon's sentiment analysis (our Firehose access provider), and 284 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 Agreement analysis

We quantified the agreement among raters – within crowd and within experts – using *Fleiss' kappa* (κ), a statistical measure of inter-rater agreement (Fleiss, 1971). κ is designed for *nominal* (non-ordinal categorical) variables, e.g. Fig 1c, 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) (Shrout and Fleiss, 1979) for ordinal categorical annotations, like `Contains Abuse`: *No* < *Problematic* < *Abusive*. Further explanations, as well as the definition of κ and icc, are available in Appendix B.

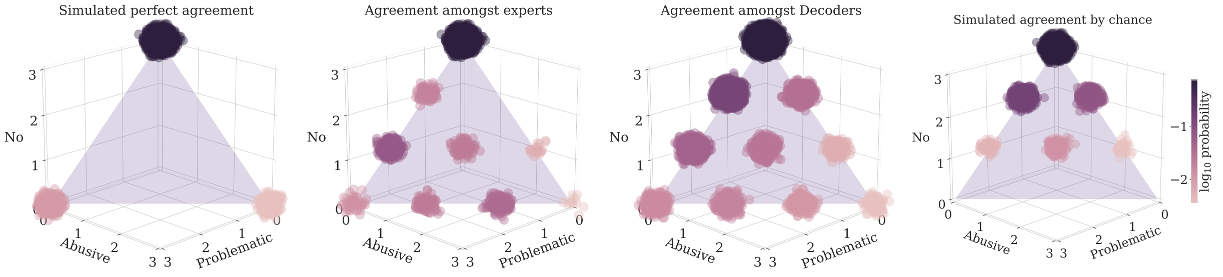


Figure 2: Visualizing the distribution of annotations $a^{(t)}$ (+ jitter for clarity) in the multinomial 2-simplex. The corners are events of *complete* agreement. The center is *no agreement* with the non-ordinal assumption, but partial agreement with ordinality. Left to right: Simulated perfect agreement, $a_c^{(t)} = 3, c \sim \hat{P}(C)$; 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: $a^{(t)} \sim \mathit{Multinomial}(N = 3, p = \hat{P}(C))$. The multinomial assumes independence between trials. See Appendix B for notations. A hierarchical modeling approach can capture inter-rater dependence (Kalaitzis and Silva, 2013; Silva and Kalaitzis, 2015).

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.

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	-

4 Analysis of scale, typology and intersectionality

4.1 Accounting for sampling uncertainty: Bootstrapping

Every statistical analysis based on a sample must evaluate the robustness of its findings against the randomness induced by its sampling mechanism. Such concerns are often the topic of abundant conversation shortly after any election whose outcome some widely reported pre-election polls failed to predict. Statisticians then point out that their outcome estimates came with margins of errors that had not always been reported. That lack of reporting for sake of simplicity often seems to assign an unwarranted absolute confidence to statistical analyses.

In this study, one major but controllable source of uncertainty stems from the limitation in the data collection: maximum 10,000 tweets per day, sampled by the provider at random (uniformly, it is assumed) among all the tweets of that day mentioning the journalists and politicians requested. This is but a small portion of the much larger volume of such tweets from that day. Therefore, luck of the draw must be accounted for: What if luck of the draw in this particular limited sample had concentrated all of the abusive tweets actually sent that day? Or, inversely, if none of the abusive tweets actually sent that day happened to be part of this random sample? These are but two extreme cases.

The unfeasible but ideal way to measure the robustness of the findings would be to reproduce the study several thousand of times and compare the results. This is obviously impractical, since each replication would require to follow the exact same methodology – from the sampling of the tweets all the way to the analysis, including the crowd-sourcing step. A alternative and basic way to measure the reliability of the findings is with parametric confidence intervals based on the asymptotic distribution of the estimator. However, this is impractical for estimators of proportions close to 0 or 1 – several variants exist, none of which are unanimously accepted.

Hence, we resort to the more computationally intensive, but altogether more robust method of *bootstrapping* (Efron, 1992; Efron and Tibshirani, 1994), which relies on much fewer assumptions and gracefully handles confidence intervals for small proportions. Bootstrapping generates multiple versions of the estimates, each on a set of tweets resampled with replacement among the actual labeled tweets. The rationale is that sampling from the empirical distribution of the observed tweets is the closest approximation to sampling from the real-world distribution. The estimates resulting from repeated applications of this sampling procedure therefore provide a close approximation to the distribution in the "unfeasible but ideal way" mentioned above. Therefore, we can sensibly measure the uncertainty of our estimates.

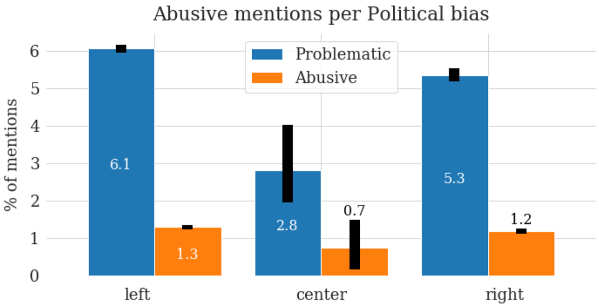


Figure 3: Example of estimates (colored bars) and their confidence intervals obtained via bootstrapping (black bars). The confidence interval for problematic tweets mentioning women at the center is ±1%, compared to ±.1% at the left. This is due to the smaller number of women in our study being at the center of the political spectrum, which thus provides us with less information.

This has allowed the human rights researchers to focus their analysis on the most robust of the findings.

4.2 Design choices and possible refinements

4.2.1 Bayesian methods

A more refined alternative to bootstrapping would be a fully Bayesian approach (Gelman and Hill, 2006; Gelman et al., 2013; Robert, 2001). Bayesian methods provide a full probability distribution on the estimates. However, we made the choice to keep the analysis as simple as possible with a relatively straightforward approach. The same holds true about the analysis of disagreement: a full hierarchical model could be used for more subtle assessment.

4.2.2 Aggregating proportions across tweets rather than considering the majority vote

For the descriptive statistics given in each of the findings of this study, we chose to aggregate across tweets the proportions of votes, as opposed to aggregating the majority vote. This means that the following two scenarios on 10 tweets would lead to the same aggregated estimate of 30% abuse (simplified as “3 abusive tweets”):

- 10 tweets, each labeled as abusive by 30% of its graders, and as neutral by the remaining 70% of its graders.
- 10 tweets, 3 of which are labeled as abusive by 100% of their graders, and the other 7 are labeled as neutral by 100% of their graders.

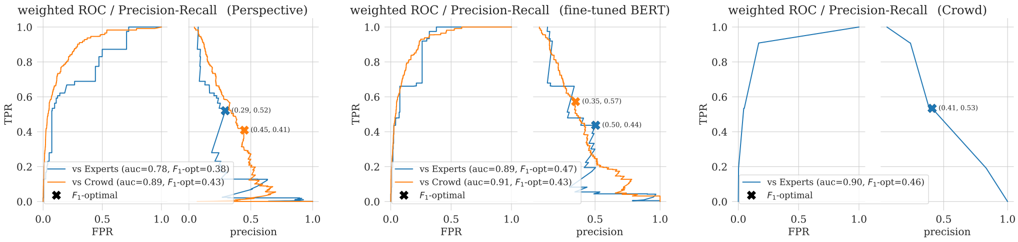


Figure 4: **Left & center**: performance of Perspective API and fine-tuned BERT classifiers, with respect to the experts' and crowd's labels. **Right**: performance of the crowd-as-a-classifier against the experts' labels. **Note**: *Recall* is equivalent to *TPR*, hence the y-axes (TPR<>Recall) of the two plots are aligned.

Table 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**

This is justified by the disagreement analysis discussed in Section 3. Unlike a visual classification task with a clear and objective ground truth, labelers are often divided, especially when the abuse in a tweet is subtle or contextual. The aggregation of proportions accounts for this variability of opinions and reflect the most accurate picture, without resorting to the extreme solution of a majority vote (which dismisses the minority's perception of abuse) or considering a tweet abusive if at least one of its graders votes as such (which overestimates the overall perception of abuse).

One refinement to our analysis would be to estimate the abuse through *ordinal regression*. This approach requires a more intricate methodology such as accounting for variable thresholds amongst graders, and therefore more design choices. We chose to keep the analysis reasonably simple.

4.2.3 Impossibility to aggregate at a thinner level

It would be interesting to analyze the results at a more granular aggregation level. For example, to estimate the volume of abuse received by members of diverse sexual orientations, or with finer cross-variable intersections. However, the more granular an aggregation is, the fewer women and fewer tweets it is comprised of, which inflated the variance (and uncertainty) of our estimates. Hence, we restricted our analysis at a high aggregation level to ensure robust findings. Also, we note that the information about each journalist and politician was gathered from public sources, and some of the variables needed for finer aggregations are not publicly available.

4.2.4 Accounting for heteroscedasticity

The variance of the answers differs amongst graders and amongst tweets. To reduce the variance of our estimators, it would be interesting to take consider weighing each tweet by the inverse of the variance of the answers on that tweet, as is commonly done in the case of regression in heteroscedastic settings. This should give rise to narrower confidence intervals – potentially allowing us to study even finer aggregation levels. Note that this would not change the results presented in this study since we considered only robust findings, but instead would refine them and possibly uncover more findings.

5 Generalization: investigating automated detection

5.1 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 Stambolieva (2017). Crimson Hexagon refers to sentiment labels – Category and Emotion – from Crimson Hexagon. We also benchmarked the pre-trained classifier from Davidson et al. (2017). Perspective API refers to the public toxicity scoring API provided by Jigsaw (Hosseini et al., 2017; Jigsaw). We also trained our own model, which combined a pre-trained BERT embedder (Devlin et al., 2018) and an abuse-specific embedding trained from scratch. For details see 5.2.

Methodology For this part of the analysis, we focus on a somewhat simpler binary classification problem, by conflating the labels *Problematic* and *Abusive* into one positive (Problematic) class. The crowd labels are the majority votes over these conflated labels on tweets labeled by exactly three volunteers – which prevents any ties, since we now have only two classes. The expert labels are majority votes over labels from the three domain experts mentioned in Section 2. For Crimson Hexagon, we define Problematic 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.

5.2 Custom-trained deep learning model

While the models described in the previous section are generic models pre-trained on unrelated data, it is natural to wonder how much improvement can be obtained by training directly on part of the labeled dataset, so as to best fit the problem.

Model architecture We used a pretrained BERT model (Devlin et al., 2018) (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 (Devlin et al., 2018)). The model was implemented in Pytorch (Paszke et al., 2017), and made use of the BERT implementation provided at (Team), 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 (von Ahn), 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 fully-connected layer of 64 units 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.

Results The numerical results are visible in the row *Fine-tuned BERT* of Table 2. Unsurprisingly, training on the labeled tweets does lead to the best average precision (Column AP). This model is trained for the very question we are trying to solve, as opposed to conflating e.g. "Toxicity" and "Abuse" which is the underlying assumption when applying Jigsaw Perspective to abuse detection. Comparison between the two can be seen in Figure 4 What is especially interesting is that, when evaluating on the expert labels, the Fine-tuned BERT classifier performs comparatively to the crowd in terms of $F1$ error (the geometric average between precision and recall). This suggests that such tools could potentially be used as an assistance to trained human moderators.

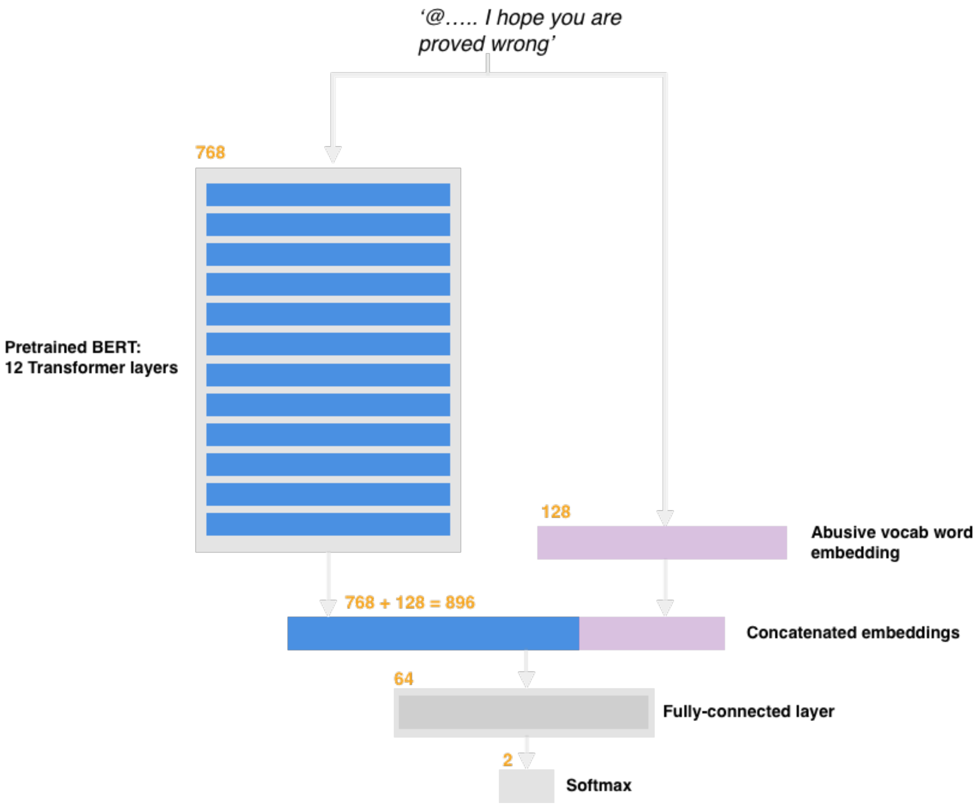

'@..... I hope you are proved wrong'
768
Pretrained BERT: 12 Transformer layers
128
Abusive vocab word embedding
768 + 128 = 896
Concatenated embeddings
64
Fully-connected layer
2
Softmax

Figure 5: We combined a pre-trained BERT model with a word embedding exclusively including abusive words. The two embeddings were concatenated and passed through a fully-connected layer before a softmax layer returned the prediction. Numbers in orange are layer widths.

6 Discussion

6.1 Dataset availability and reproducibility

Amnesty International is looking into how to publish the dataset in an ethical manner, to encourage replication and further research on the topic. 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.

6.2 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 (Gambäck and Sikdar, 2017). The moderation of abusive messages at scale is pushing companies in the direction of using some form of automated assistance. Our results highlight the double challenge of automatic abuse classification: the subjectivity in the labels and the limitations of current state-of-the-art classifiers. Amnesty International has warned of the real-world censorship consequences when automated content moderation systems get it wrong. This all points toward the need for systems where human subtlety and context awareness remain at the centre of content moderation decisions, even if they are assisted 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.

Bibliography

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

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.

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

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

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

Discourse. Discussion forum. URL https://www.discourse.org/.

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.

Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992.

Bradley Efron and Robert J Tibshirani. *An introduction j the bootstrap*. CRC press, 1994.

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

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

Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

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.

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.

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.

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

Jigsaw. Perspective API. URL https://www.perspectiveapi.com.

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, 2013.

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

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.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

Christian Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer, 2001.

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

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

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

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

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 (218, including 22 women who left parliament during the June 2017 elections and after excluding one politician with over 1 million followers);
- All women in the United States Congress (105, after excluding 3 politicians with more than 1 million followers);
- And 455 women journalists working at the following news organizations, selected to represent a diversity of political affiliations:
  - Breitbart,
  - Daily Mail,
  - The Sun,
  - The Guardian,
  - The New York Times,
  - Gal-Dem,
  - PinkNews.

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

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

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 Shrout and Fleiss (1979). 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 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:

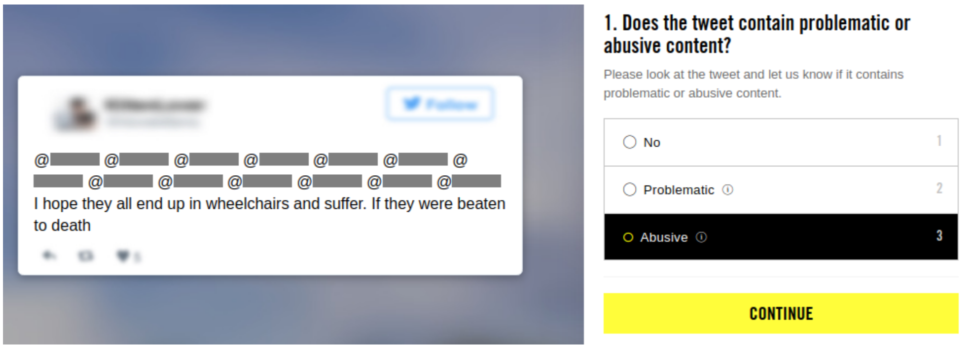
$$\hat{\mathbb{E}}_W[f(t, k, d)] = \sum_{(t_i,k_i,d_i)\in A} \frac{w_i}{\sum_j w_j}\ f(t_i, k_i, d_i)\,. \tag{19}$$

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 Douc and Moulines (2008), but is outside of the scope of this article.

D Labeling tool screenshots

The workflow presented to each grader by the labeling tool is illustrated in Figure 6 and Figure 7.



@[redacted] @[redacted] @[redacted] @[redacted] @[redacted] @[redacted] @[redacted] @[redacted] @[redacted] @[redacted] @[redacted] @[redacted] @[redacted]
I hope they all end up in wheelchairs and suffer. If they were beaten to death

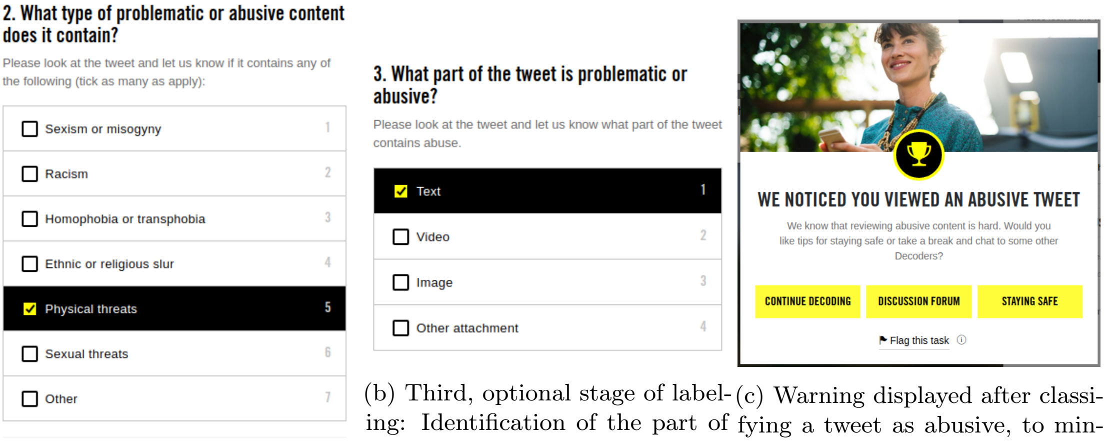
1. Does the tweet contain problematic or abusive content?

Please look at the tweet and let us know if it contains problematic or abusive content.

○ No	1
○ Problematic ⓘ	2
○ Abusive ⓘ	3

CONTINUE

Figure 6: First stage of labeling: Initial screen showing an anonymized tweet, with anonymized handles and first question.



2. What type of problematic or abusive content does it contain?

Please look at the tweet and let us know if it contains any of the following (tick as many as apply):

□ Sexism or misogyny	1
□ Racism	2
□ Homophobia or transphobia	3
□ Ethnic or religious slur	4
☑ Physical threats	5
□ Sexual threats	6
□ Other	7

3. What part of the tweet is problematic or abusive?

Please look at the tweet and let us know what part of the tweet contains abuse.

☑ Text	1
□ Video	2
□ Image	3
□ Other attachment	4

WE NOTICED YOU VIEWED AN ABUSIVE TWEET

We know that reviewing abusive content is hard. Would you like tips for staying safe or take a break and chat to some other Decoders?

(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.

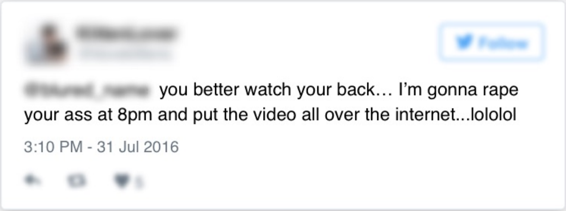
Figure 7: Follow-up stages, conditional on the first stage of labeling.

E Definitions and examples used in Troll Patrol - Trigger Warning

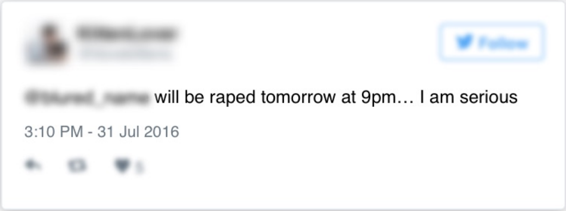
Abusive content Abusive content violates Twitter's own rules and includes tweets that promote violence against or threaten people based on their race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.

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).

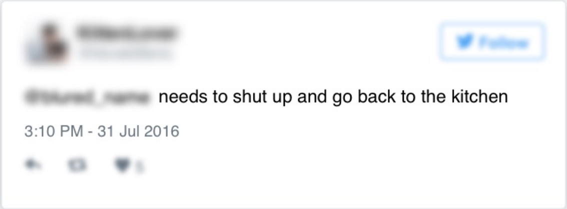

[redacted] [redacted] you better watch your back… I'm gonna rape your ass at 8pm and put the video all over the internet...lololol

3:10 PM - 31 Jul 2016


[redacted] [redacted] will be raped tomorrow at 9pm… I am serious

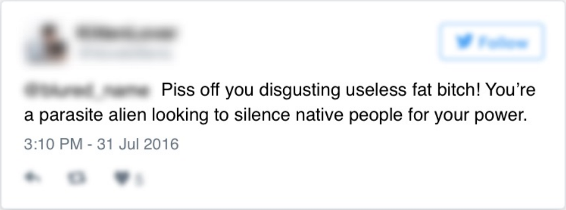
3:10 PM - 31 Jul 2016

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.

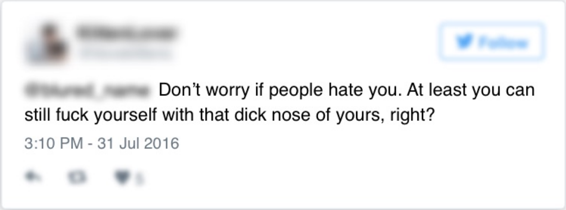

[redacted] [redacted] needs to shut up and go back to the kitchen

3:10 PM - 31 Jul 2016

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.

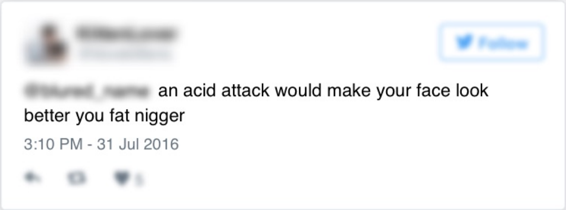

[redacted] [redacted] Piss off you disgusting useless fat bitch! You're a parasite alien looking to silence native people for your power.

3:10 PM - 31 Jul 2016

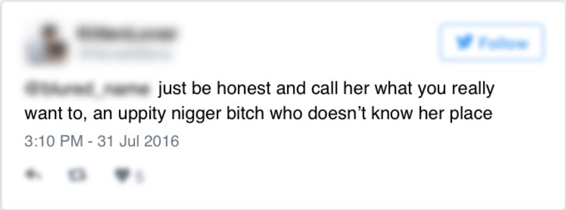

[redacted] [redacted] Don't worry if people hate you. At least you can still fuck yourself with that dick nose of yours, right?

3:10 PM - 31 Jul 2016

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.

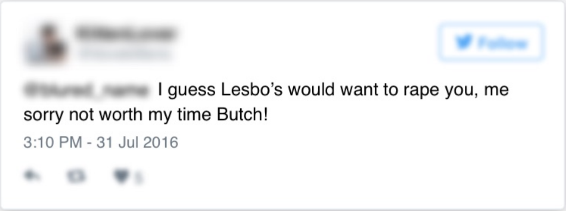

[redacted] [redacted] an acid attack would make your face look better you fat nigger

3:10 PM - 31 Jul 2016


[redacted] [redacted] just be honest and call her what you really want to, an uppity nigger bitch who doesn't know her place

3:10 PM - 31 Jul 2016

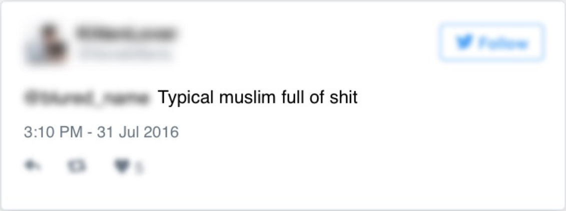
Homophobia or transphobia Discriminatory, offensive or insulting content directed at a woman based on her sexual orientation, gender identity or gender expression. This includes negative comments towards bisexual, homosexual and transgender people.



[redacted] I guess Lesbo's would want to rape you, me sorry not worth my time Butch!

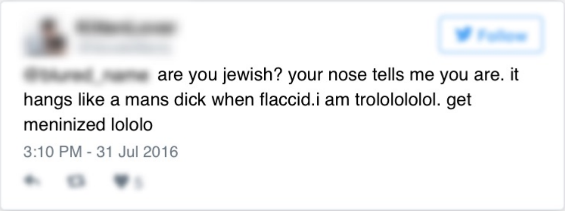
3:10 PM - 31 Jul 2016

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



[redacted] Typical muslim full of shit

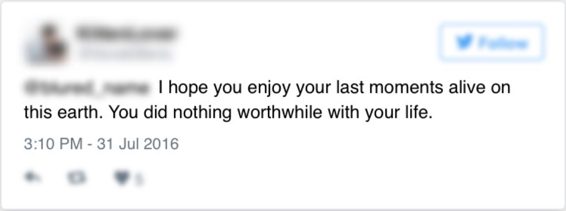
3:10 PM - 31 Jul 2016



[redacted] are you jewish? your nose tells me you are. it hangs like a mans dick when flaccid.i am trololololol. get meninized lololo

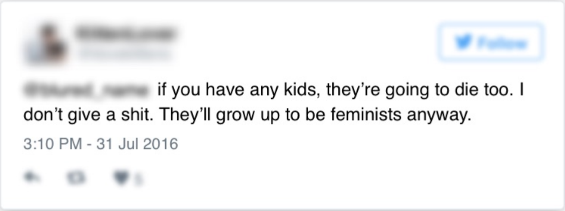
3:10 PM - 31 Jul 2016

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



[redacted] I hope you enjoy your last moments alive on this earth. You did nothing worthwhile with your life.

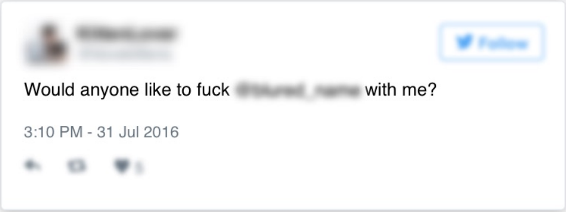
3:10 PM - 31 Jul 2016



[redacted] if you have any kids, they're going to die too. I don't give a shit. They'll grow up to be feminists anyway.

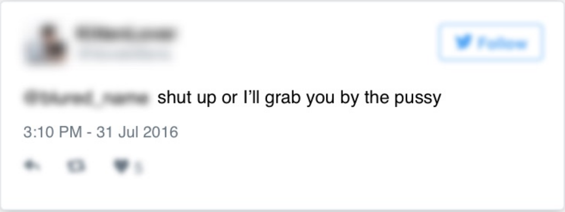
3:10 PM - 31 Jul 2016

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



Would anyone like to fuck [redacted] with me?

3:10 PM - 31 Jul 2016



[redacted] shut up or I'll grab you by the pussy

3:10 PM - 31 Jul 2016

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.