Ex Machina: Personal Attacks Seen at Scale

Ellery Wulczyn *
Wikimedia Foundation
ellery@wikimedia.org

Nithum Thain *
Jigsaw
nthain@google.com

Lucas Dixon
Jigsaw
ldixon@google.com

ABSTRACT

The damage personal attacks cause to online discourse motivates many platforms to try to curb the phenomenon. However, understanding the prevalence and impact of personal attacks in online platforms at scale remains surprisingly difficult. The contribution of this paper is to develop and illustrate a method that combines crowdsourcing and machine learning to analyze personal attacks at scale. We show an evaluation method for a classifier in terms of the aggregated number of crowd-workers it can approximate. We apply our methodology to English Wikipedia, generating a corpus of over 100k high quality human-labeled comments and 63M machine-labeled ones from a classifier that is as good as the aggregate of 3 crowd-workers, as measured by the area under the ROC curve and Spearman correlation. Using this corpus of machine-labeled scores, our methodology allows us to explore some of the open questions about the nature of online personal attacks. This reveals that the majority of personal attacks on Wikipedia are not the result of a few malicious users, nor primarily the consequence of allowing anonymous contributions from unregistered users.

1. INTRODUCTION

With the rise of social media platforms, online discussion has become integral to people's experience of the internet. Unfortunately, online discussion is also an avenue for abuse. A 2014 Pew Report highlights that 73% of adult internet users have seen someone harassed online, and 40% have personally experienced it [5]. Platforms combat this with policies concerning such behavior. For example Wikipedia has a policy of "Do not make personal attacks anywhere in Wikipedia"[33] and notes that attacks may be removed and the users who wrote them blocked.[1]

The challenge of creating effective policies to identify and appropriately respond to harassment is compounded by the difficulty of studying the phenomena at scale. Typical annotation efforts of abusive language, such as that of Warner and Hirschberg [27], involve labeling thousands of comments, however platforms often have many orders of magnitude more; Wikipedia for instance has 63M English talk page comments. Even using crowd-workers, getting human-annotations for a large corpus is prohibitively expensive and time consuming.

The primary contribution of this paper is a methodology for quantitative, large-scale, longitudinal analysis of a large corpus of online comments. Our analysis is applicable to properties of comments that can be labeled by crowd-workers with high levels of inter-annotator agreement. We apply our methodology to personal attacks on Wikipedia, inspired by calls from the community for research to understand and reduce the level of *toxic discussions* [31, 30], and by the clear policy Wikipedia has on personal attacks [33].

We start by crowdsourcing a small fraction of the corpus, labeling each comment according to whether it is a personal attack or not. We use this data to train a simple machine learning classifier and experiment with features and labeling methods. The machine learning methods are not novel, but their application does validate and extend the findings of Nobata et al. [15]: character-level n-grams result in an impressively flexible and performant classifier for a variety of abusive language in English. We additionally note that using the empirical distribution of human-ratings, rather than the majority vote, produces a better classifier, even in terms of the AUC metric.

The classifier is then used to annotate the entire corpus of comments - acting as a surrogate for crowd-workers. To know how meaningful the automated annotations are, we develop an evaluation method for comparing an algorithm to a group of human annotators. We show that our classifier is as good at generating labels as aggregating the judgments of 3 crowd-workers. To enable independent replication of the work in this paper, as well as to support further quantitative research, we have made public our corpus of both human and machine annotations as well as the classifier we trained [36].

We use our classifier's annotations to perform quantitative analysis over the whole corpus of comments. To ensure that our results accurately reflect the real prevalence of personal attacks within different sub-groups of comments, we select a threshold that appropriately balances precision and recall. We also empirically validate that the threshold produces results on subgroups of comments commensurate with the results of crowd-workers.

This allows us to answer questions that our much smaller sample of crowdsourced annotations alone would struggle to. We illustrate this by showing how to use our method to explore several open questions about the nature of personal attacks on Wikipedia: What is the impact of allowing anonymous contributions, namely those from unregistered users? How do attacks vary with the quantity of a user's contributions? Are attacks concentrated among a few

*Equal contribution.

[1]This study uses data from English Wikipedia, which for brevity we will simply refer to as Wikipedia.

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3052591

users? When do attacks result in a moderator action? And is there a pattern to the timing of personal attacks?

The rest of the paper proceeds as follows: Sec. 2 discusses related work on the prevalence, impact, and detection of personal attacks and closely related online behaviors. In Sec. 3 we describe our data collection and labeling methodology. Sec. 4 covers our model-building and evaluation approaches. We describe our analysis of personal attacks in Wikipedia in Sec. 5. We conclude in Sec. 6 and outline challenges with our method and possible avenues of future work.

2. RELATED WORK

Definitions, Prevalence and Impact. One of the challenges in studying negative online behavior is the myriad of forms it can take and the lack of a clear, common definition [19]. While this study focuses on personal attacks, other studies explore different forms of online behavior including hate speech ([7], [13], [19], [27]), online harassment ([3], [40]), and cyberbullying ([17], [20], [25], [35], [38]).

Online harassment itself is sometimes further divided into a taxonomy of forms. A recent Pew Research Center study defines online harassment to include being: called offensive names, purposefully embarrassed, stalked, sexually harassed, physically threatened, and harassed in a sustained manner [5]. The Wikimedia Foundation Support and Safety team conducted a similar survey [23] using a different taxonomy (see Figure 1).

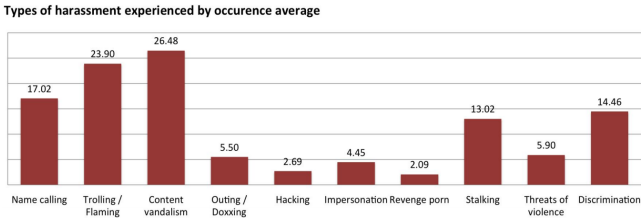


Figure 1: Forms of harassment experienced on Wikimedia [23].

This toxic behavior has a demonstrated impact on community health both on and off-line. The Wikimedia Foundation found that 54% of those who had experienced online harassment expressed decreased participation in the project where they experienced the harassment [23]. Online hate speech and cyberbullying are also closely connected to suppressing the expression of others [21], physical violence [29], and suicide [4].

Automated Detection. There have been a number of recent papers on detecting forms of toxic behavior in online discussions. Much of this work builds on existing machine learning approaches in fields like sentiment analysis [16] and spam detection [22]. On the topic of harassment, the earliest work on machine learning based detection is Yin et al.'s 2009 paper [40] which used support vector machines on sentiment and context features extracted from the CAW 2.0 dataset [6]. In [21], Sood et al. use the same algorithmic framework to detect personal insults using a dataset labeled via Amazon Mechanical Turk from the Yahoo! Buzz social news site. Dinakar et al. [4] decompose the issue of cyberbullying by training separate classifiers for attacks based on sexual orientation, race or intelligence in YouTube comments. Building on these works, Cheng et al. [3] use random forests and logistic regression techniques to predict which users of the comment sections of several news sites would become banned for antisocial behavior. Most recently, Nobata et al. [15] extract character n-gram, linguistic, syntactic, and distributional semantic features from a very large corpus of Yahoo! Finance and News comments to detect abusive language.

Data Sets. A barrier to further algorithmic progress in the detection of toxic behavior is a dearth of large publicly available datasets [19]. To our knowledge, the current open datasets are limited to the Internet Argument Corpus [26], the CAW 2.0 dataset provided by the Fundacion Barcelona Media [6], the "Hate Speech Twitter Annotations" corpus[28], and the "Detecting Insults in Social Commentary" dataset released by Impermium via Kaggle [10]. In past work, many researchers have relied on creating their own hand-coded datasets ([13], [21], [27]), using crowd-sourced or in-house annotators. These approaches limit the size of the labeled corpora due to the expense of labeling examples. A few authors have suggested alternative techniques that could be effective in obtaining larger scale datasets. In [19], Saleem et al. outline some of the limitations of using a small hand-coded dataset and suggest an alternative approach that uses all comments within specific online communities as positive and negative training examples of hate speech. Xiang et al. [37] use topic modeling approaches along with a small seed set of tweets to produce a training set for detecting offensive tweets containing over 650 million entries. Building on the work of [40], Moore et al. [14] use a simple rule-based algorithm for automatic labeling of forum posts on which they wish to do further analysis.

3. CROWDSOURCING

In this section we discuss our approach to identifying personal attacks in a subset of Wikipedia discussion comments via crowdsourcing. The crowdsourcing process involves:

1. generating a corpus of Wikipedia discussion comments,
2. choosing a question for eliciting human judgments,
3. selecting a subset of the discussion corpus to label,
4. designing a strategy for eliciting reliable labels.

To generate a corpus of discussion comments, we processed the public dump of the full history of English Wikipedia as described in Appendix A. The corpus contains 63M comments from discussions relating to user pages and articles dating from 2004-2015.

The question we posed to get human judgments on whether a comment contains a personal attack is shown in Figure 2. In addition to identifying the presence of an attack, we also try to elicit if the attack has a target or whether the comment quotes a previous attack. We do not, however, make use of this additional information in this study. Before settling on the exact phrasing of the question, we experimented with several variants and chose the one with the highest inter-annotator agreement on a set of 1000 comments.

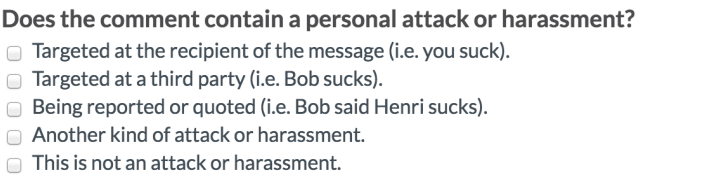


Figure 2: The question posed to our Crowdflower annotators.

To ensure representativeness, we undertook the standard approach of randomly sampling comments from the full corpus. We will refer to this set of comments as the *random* dataset. Through labeling a random sample, we discovered that the overall prevalence of per-

it's huge data imbalance

sonal attacks on Wikipedia talk pages is around 1% (see Section 5.1).

To allow training of classifiers, we need to create a corpus that contains a sufficient number and variety of examples of personal attacks. In order to obtain these, we enhance our random dataset by also sampling comments made by users who where blocked for violating Wikipedia's policy on personal attacks [33]. In particular, we consider the 5 comments made by these users around every block event. We call this the *blocked* dataset and note that it has a much higher prevalence of attacks (approximately 17%).

Sample Type	Annotated Comments	Percentage Attacking
Random	37611	0.9 %
Blocked	78126	16.9 %
Total	115737	11.7 %

Table 1: Summary statistics of labeled data. Each comment was labeled 10 times. Here we define a comment as an attack if the majority of annotators labeled it as such.

We labeled our subset of comments using the Crowdflower crowdsourcing platform.[2] Crowdsourcing as a data collection methodology is well studied ([2], [24]) and has proven effective for constructing corpora for machine learning in various contexts ([3], [13], [18], [21], [27]).

As a first step to ensuring data quality, each annotator was required to pass a test of 10 questions. These questions were randomly selected from a set that we devised to contain balanced representation of both attacking and non-attacking comments. Annotators whose accuracy on these test questions fell below a 70% threshold would be removed from the task. This improved our annotator quality by excluding the worst ~2% of contributors. Under the Crowdflower system, additional test questions are randomly interspersed with the genuine crowdsourcing task (at a rate of 10%) in order to maintain response quality throughout the task.

In order to get reliable estimates of whether a comment is a personal attack, each comment was labeled by at least 10 different Crowdflower annotators. This allows us to aggregate judgments from 10 separate people when constructing a single label for each comment. We chose 10 judgments based on experiments in Sec. 4.3 that showed that aggregating more judgments provided little further improvement. Finally, we applied several data cleaning steps to the Crowdflower annotations. This included removing annotations where the same worker labeled a comment as both an attack and not an attack and removing comments that most workers flagged as not being English.

We evaluated the quality of our crowd-sourcing pipeline by measuring inter-annotator agreement [11]. This technique measures whether a set of "common instructions to different observers of the same set of phenomena, yields the same data within a tolerable margin of error" [9]. We chose the specific inter-annotator agreement metric of Krippendorf's alpha due to our context, where multiple raters rate overlapping, but disparate sets of comments [12]. Our data achieves a Krippendorf's alpha score of 0.45. This result is in-line with results achieved in other crowdsourced studies of toxic behavior in online communities [3].

4. MODEL BUILDING

[2]https://www.crowdflower.com/

We now use the set of crowdsourced annotations to build a machine learning classifier for identifying personal attacks. We first discuss the set of machine learning architectures we explored and then describe our evaluation methodology.

4.1 Model Building Methodology

We treat the problem of identifying personal attacks as a binary text classification problem. We rely purely on features extracted from the comment text instead of including features based on the authors' past behavior and the discussion context. This makes it easy for Wikipedia editors and administrators, journalists and other researchers to explore the strengths and weaknesses of the models by simply generating text examples. It also allows the models to be applied beyond the context of Wikipedia.

In terms of model architectures, we explored logistic regression (LR), and multi-layer perceptrons (MLP). In future work, we plan to experiment with long short-term memory recurrent neural networks (LSTM) as well. For the LR and MLP models we simply use bag-of-words representations based on either word- or character-level n-grams. Past work in the domain of detecting abusive language in online discussion comments, showed that simple n-gram features are more powerful than linguistic and syntactic features, hand-engineered lexicons, and word and paragraph embeddings [15].

In all of the model architectures, we have a final softmax layer and use the cross-entropy as our loss function. The cross-entropy function is defined as:

$$H(y,\hat{y}) = -\sum_i y_i \log(\hat{y}_i) \quad (1)$$

where $\hat{y}$ is our predicted probability distribution over classes, and y is the true distribution.

In addition to experimenting with different model architectures, we also experimented with two different ways of synthesizing our 10 human annotations per comment to create training labels. In the traditional classification approach, there is only one true class and so the true distribution, y, is represented as a one-hot (OH) vector determined by the majority class in the comment's set of annotations.

For the problem of identifying personal attacks, however, one can argue that there is no single true class. Different people may judge the same comment differently. Unsurprisingly, we see this in the annotation data: most comments do not have a unanimous set of judgments, and the fraction of annotators who think a comment is an attack differs across comments.

The set of annotations per comment naturally forms an approximate empirical distribution (ED) over opinions of whether the comment is an attack. A comment considered a personal attack by 7 of 10 annotators can thus be given a true label of [0.3, 0.7] instead of [0,1]. Using ED labels is motivated by the intuition that comments for which 100% of annotators think it is an attack are probably different in nature from comments where only 60% of annotators consider it so. Since the majority class is the same in both cases, the OH labels lose the distinction. Hence, in addition to the OH labels, we also trained each architecture using ED labels.

Finally, we should note that the interpretation of a model's scores depends on whether it was trained on ED or OH labels. In the case of a model trained on ED labels, the attack score represents the predicted fraction of annotators who would consider the comment an attack. In the case of a model trained on OH labels, the attack score represents the probability that the majority of annotators would consider the comment an attack.

4.2 Model Building Evaluation

As discussed above, we considered three major dimensions in the model design space:

1. model architecture (LR, MLP)
2. n-gram type (word, char)
3. label type (OH, ED)

In order to evaluate each of the 8 possible modeling strategies we randomly split our set of annotations into train, development and test splits (in a 3:1:1 ratio). For each model, we performed 15 iterations of random search over a grid of relevant hyper-parameters[1].[3] During the model tuning process, each run was trained on the train split and evaluated on the development split. Table 2 shows two evaluation metrics for each of the 8 tuned models. The standard 2-class area under the receiver operating characteristic curve (AUC) score is computed between the models' predicted probability of being an attack and the majority class label in the set of annotations for each comment. To better evaluate the performance of models trained on ED labels, we also include the Spearman rank correlation between the models' predicted probability of being an attack and the fraction of annotators who considered the comment an attack.

Model Type	N-Gram Type	Label Type	AUC	Spearman
LR	Word	OH	94.62	53.16
		ED	95.55	65.2
	Char	OH	96.18	59.20
		ED	96.24	66.68
MLP	Word	OH	95.25	56.11
		ED	96.15	66.33
	Char	OH	95.90	58.77
		ED	**96.59**	**68.17**

Table 2: Evaluation metrics of different model architectures trained on the train split and evaluated on the development split. The hyper-parameters of each architecture were tuned using randomized search

Across all model and label types, we see that character n-grams outperform word n-grams, which is consistent with feature importance analysis in [15].[4] We suspect this is due to the higher robustness to spelling variations that char-n-grams exhibit, which are very common in online discussions, especially in expletives commonly used in personal attacks.

Another consistent pattern is the boost in the performance metrics for models trained using ED labels. The large boost in Spearman correlation is somewhat unsurprising because it is a function of the fraction of annotators who consider a comment to be a personal attack. The models trained using OH labels did not receive any supervision on how to estimate this fraction (they only see majority class). The interesting result is that even the AUC scores are consistently higher: this means that using training labels that encode what fraction of people think a comment is an attack helps in predicting what the majority of annotators think. Our results indicate that using ED labels may give a performance boost over the standard OH labels for other machine-learning tasks using multiple crowdsourced labels per training example.

[3]For details on the set of hyper-parameters explored, we refer the reader to the relevant notebook in our code repository: `https://github.com/ewulczyn/wiki-detox/blob/master/src/modeling/cv_ngram_architectures.ipynb`

[4]Note we explored word n-grams in the range of 1-2 and characters n-grams in the range 1-5. During the hyper-parameters search, we searched over the same range of values for the number of n-gram features to include.

4.3 Human Baseline Comparison

We developed a classifier for detecting personal attacks in order to score the full history of comments on Wikipedia in a cost and time effective manner. For our purposes, the model is an approximation of the crowdsourcing process. Hence, we want to be able to answer the question: How good of a surrogate is our model for crowdsourced annotations?

To answer this we will use one group of annotators, call them the prediction-group, to predict what another group of annotators, call them the truth-group, thinks about a comment. We treat the aggregated judgments of the truth-group as ground truth labels. We treat the prediction-group as a model: an ensemble of annotators, who pool their judgments to make predictions. Hence, we will refer to the prediction-group as the *annotator ensemble*. By comparing our machine learning model's predictive power to the predictive power of the annotator ensemble, we can get an estimate of how good of a surrogate our model is for a fixed size annotator ensemble. We will refer to this method of generating baselines as *annotator ensemble baselining*.

To be more specific, let's fix the size of the truth-group at n_t and the size of the prediction-group at n_p. Assume we have collected at least $n_t + n_p$ annotations per comment in our corpus. Now, for each comment c, we randomly split the full set of annotations for c into two non-overlapping sets, T_c for the truth-group and P_c for the prediction-group, of sizes n_t and n_p respectively. We split the set of annotators at the comment level, since the corpus of comments may be so large that not every annotator judges every comment, making a fixed split across all comments impossible in general. We will aggregate annotations in T_c using the function agg_{true} to get a ground truth label $y(c)$ for comment c. The choice of agg_{true} depends on the evaluation metric we want to use. For the AUC metric agg_{true} is the OH aggregation function, whereas for Spearman correlation agg_{true} is the ED aggregation function. We will take the average of the annotations in P_c to get a prediction $\hat{y}(c)_{AE}$. We apply our machine learning model, to comment c to get prediction $\hat{y}(c)_{ML}$. Finally, we compute the evaluation metrics of AUC and Spearman correlation over the entire corpus between y and $\hat{y}_{ML}$ and between y and $\hat{y}_{AE}$ to compare the machine learning model to the annotator ensemble. The comparison of these scores, tells us how good our model is compared to an ensemble of annotators of size n_p at predicting labels generated by pooling n_t annotations.

Note that, for each question, the annotators are randomly split into a prediction-group and a truth-group. As a result, there is some variability in the evaluation metrics stemming from this assignment step. By running the entire process several times, we can estimate this variability and average the evaluation metric results from each run to get a more stable estimate.

We applied our *annotator ensemble baselining* method to a set of 8,000 comments from the test split and had each comment labeled 20 times. We will refer to this special subset of comments as the baseline split. Out of the baseline split, 4000 comments come from our *random* dataset and 4000 come from our *blocked* dataset. We fix n_t, the number of annotations used to generate labels, at 10, since this is the number of annotations per comment used in training the model. Table 3 shows AUC scores and Spearman correlations for the aggregate prediction of the prediction-group as we vary its size n_p from 1 to 10. The final line of the table also reports

the values for the best LR model architecture from Table 2.[5] The reported mean scores and standard errors are the result of running the entire process 25 times.

For the annotator ensemble, both the AUC scores and Spearman correlations increase with diminishing returns as the size of the ensemble increases. On both of our metrics, our model outperforms an annotator ensemble of size $n_p = 3$. Thus, by these two metrics, running our model over the full history of comments in Wikipedia is as good as having each comment labeled by 3 annotators.

n_p	AUC		Spearman	
1	88.54	(0.42)	53.58	(0.79)
3	95.49	(0.31)	64.75	(0.44)
5	97.13	(0.23)	68.27	(0.46)
7	97.81	(0.15)	69.86	(0.60)
9	98.24	(0.14)	70.97	(0.44)
10	98.53	(0.12)	71.11	(0.36)
Model:	97.19	(0.14)	66.02	(0.44)

Table 3: Mean evaluation metrics (and standard errors) on the baseline split, fixing the truth-group size n_t at 10 and varying the prediction-group size n_p.

5. ANALYSIS

Using the best personal attack classifier from Sec. 4.2, we obtain a full corpus of machine-labeled discussions in Wikipedia. In this section, we use the fully annotated corpus to better understand the prevalence and nature of attacks in Wikipedia. For the following analyses, we focus on comments made in 2015 and exclude administrative comments and comments generated by bots as described in Appendix A.

5.1 Choosing a Threshold

Given a comment, our classifier outputs a continuous score in the interval $[0,1]$. To get a discrete label from this score, we pick a threshold t and let comments with a score above t have label 1, indicating an attack. Using discrete labels makes it possible to identify individual comments that are predicted to contain personal attacks and estimate the fraction of attacks within a set of comments.

To choose a threshold, we pick the point that strikes a balance between precision and recall on *random* evaluation data. A key property of this threshold for the purpose of using machine-generated labels for analysis, is that false positives are offset by false negatives. As a result, the fraction of comments that are labeled as attacks by the classifier is the same as the fraction of comments that are labeled as attacks by the human annotators. Hence, we refer to this threshold as the *equal-error threshold*.[6]

To see how well this property generalizes to new data, we used the *equal-error threshold* on the development set and used it to get model-generated labels for the test set. Fig. 3a shows that the estimated rate of attacks computed using model-generated labels lies within a 95% confidence interval for the rate of attacks computed from crowd-generated labels.

[5]Note that the reported performance of our model is slightly different in Table 2 than in Table 3, since in the former table the model is evaluated on the dev split, while in the latter table it is evaluated in the baseline split and ratios of random to blocked comments differ across the two splits.

[6]This threshold also maximizes the F1 score, since our precision and recall are monotonic functions of the decision threshold.

Even though the thresholded model-scores give good estimates of the rate of attacks over a random sample of comments, it is not given that they also give accurate estimates when partitioning comments into different groups. To provide empirical evidence that the thresholded scores give accurate estimates when subdividing the dataset into groups that will be important for later analysis, we shows various splits of the dataset in Fig. 5.1. We split on the year the comment was posted, by whether the author was logged-in, by the number of days the author has been active and by whether the comment contains the n-gram "thank", an important feature of the classifier. For the following analyses, we then use the *equal-error threshold* over the union of the development and tests sets. At this threshold ($t = 0.425$), the precision is 0.63, the recall (e.g true-positive rate) is 0.63 and the false-positive rate is 0.0034.

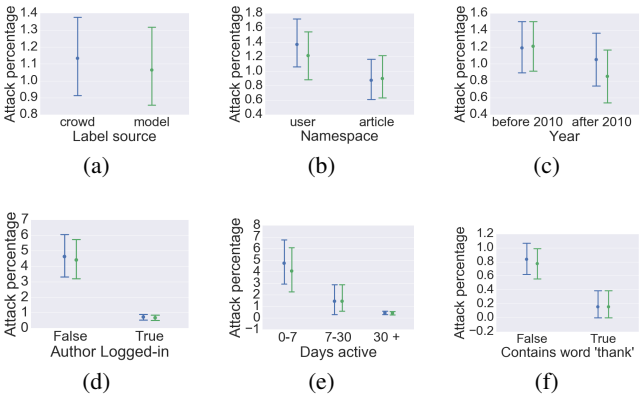


Figure 3: Bootstrapped confidence intervals for the fraction of attacks within the test set broken down by (b) discussion namespace, (c) year, (d) logged in status of the author, (e) number of days author made an edit as of 2015, (f) whether the comment contains the word "thank". Blue intervals come from ground-truth human labels, green intervals come from machine generated labels.

5.2 Understanding Attacks

In this section, we ask a number of questions about the nature of attackers, attack timing, and moderation on Wikipedia. We investigate the answers using our machine-labeled data at the equal-error threshold.

What is the impact of anonymity?

Wikipedia users can make edits either under a registered username or anonymously. In the latter case, the edits are attributed to the IP address from which they were made.[7] Table 4 shows that, last year, 43% of editing accounts were anonymous and these contributed 9.6% of the comments in our dataset.

It has been shown that anonymity provides psycho-social benefits to cyberbullies [14] and can lead to "heightened aggression and inappropriate behavior" [39]. We compare the prevalence of attacks for registered and anonymous users in Table 4.[8] This shows that the attack prevalence among comments by anonymous users is

[7]This means that, in principle, there can be multiple users editing under the same IP, and the same user editing under multiple IPs.

[8]A future analysis might also try to differentiate registered accounts with a long running reputation from so called sock-puppet accounts created by a user to appear as if their contributions are coming from multiple users.

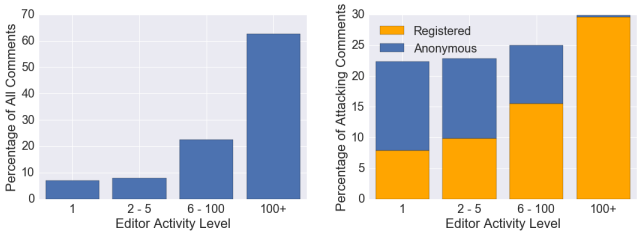
six times as high as that of registered users. Thus, while anonymous contributions are much more likely to be an attack, overall they contribute less than half of attacks. This difference of means is significant at a $p < 0.0001$ ($t=63.8$) level. These extreme values of significance are not surprising as our algorithm allows us to label data at a population level.

Anonymity	Number of Accounts	Number of Comments	Attack Prevalence
Anonymous	97,742	191,460	3.1%
Registered	129,394	2,023,559	0.5%
Totals	227,136	2,215,019	0.8%

Table 4: Comment statistics by user anonymity (2015).

How do attacks vary with the quantity of a user's contributions?

Editors on Wikipedia fall along a wide spectrum in terms of their engagement with discussions on the platform. Some comment a few times a year whereas others will comment several times a week. For our purposes, a user's *activity level* is the number of comments that they made in 2015. In Fig. 4a, we show how many comments were made by users with different activity levels. We see that over 60% of comments are made by users who made over 100 comments over the year. Users who made 5 or fewer comments are only responsible for 15% of total comments.



(a) (b)

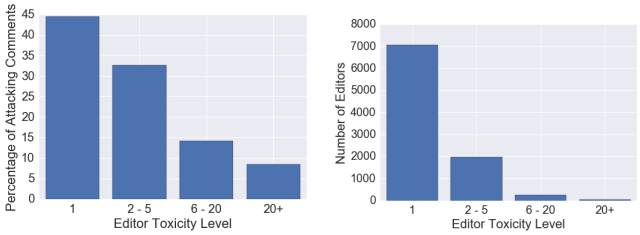
Figure 4: (a). A histogram of the percentage of total comments by user activity level. (b) A histogram of the percentage of total attacks by user activity level.

The story completely changes when we use these same user segments to understand attacking behavior. Fig. 4b shows the percentage of total attacks attributable to users with different activity levels. We find that almost half of all attacks are made by users with an activity level below 5. Even controlling for the effect of anonymity that we saw earlier, more than 18% of attacking comments are made by registered users with an activity level below 5. Users with an activity level of over 100 comments (almost all of whom are registered) are responsible for 30% of attacking comments. Thus, users at both low and high levels of contribution are responsible for a significant portion of attacks.

Are attacks concentrated among a few highly toxic users?

We define the *toxicity level* of a user to be the number of attacks written by that user in 2015. By segmenting users by toxicity level, we are able to uncover whether attacks are diffused among many low toxicity users or concentrated among a few users with high toxicity.

Fig. 5a describes the proportion of attacks made by users at different levels of toxicity. Fig. 5b provides the total number of users at each toxicity level. By comparing these figures, we see that almost 80% of attacks come from the over 9000 users who have made fewer than 5 attacking comments. However, the 34 users with a toxicity level of more than 20 are responsible for almost 9% of attacks. Thus, while the majority of Wikipedia's attacks are diffused infrequent attackers, significant progress could be made by moderating a relatively small number of frequent attackers.



(a) (b)

Figure 5: (a) A histogram of the percentage of total attacks by user toxicity level. (b) A histogram of the total number of users at each toxicity level.

When do attacks result in moderation?

Moderators and administrators can enforce the policy on personal attacks [33] by warning or blocking offending users for a period of time. Our analysis takes all attacking comments in the 2015 Wikipedia corpus and asks how many of these lead to a moderation event in the following 7 days. We find that 7.7% of attacks are followed by a warning and 7.0% of attacks are followed by a block within this period. 11.3% of attacks are followed by either a warning or a block.

As discussed in Sec. 5.1, at the equal-error threshold, our algorithm has a precision of 0.63. After normalizing by this precision, we find that 12.2% of the expected number of true attacks are followed by a warning, 11.1% are followed by a block and 17.9 % are followed by either. Thus, a high proportion of attacking comments remain unmoderated.

There are a number of factors that affect the chances of moderation, including repeated attacks and having been moderated in the past. Fig. 6 shows us that the chance of being blocked or warned increases with the number of personal attacks a user makes.

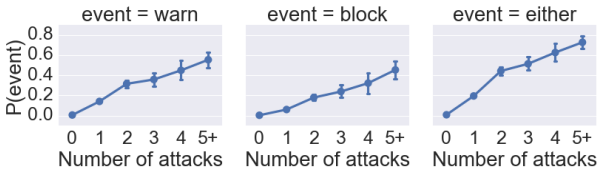


Figure 6: Probability of being warned or blocked in 2015 as a function of the number of personal attacks made in 2015

Finally, we see in Fig. 7 that the likelihood of a new attack leading to a block increases with the number of times a user has been blocked in the past. This may be due to heightened scrutiny of previously blocked users. Alternatively, it may be that blocked users make more frequent or more toxic attacks, and are hence more

likely to be warned and moderated in the future.

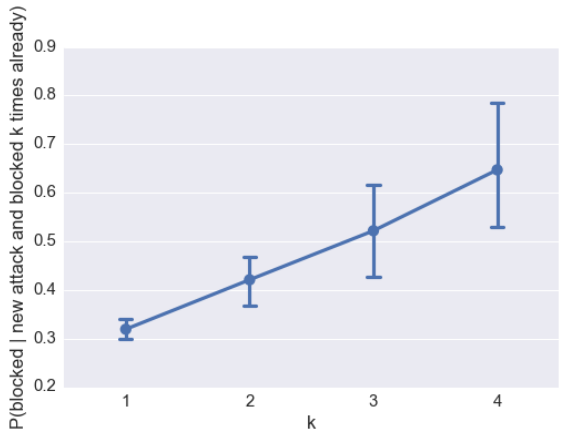


Figure 7: Probability of being blocked again after a new attack as function of the number of times the user has been blocked in the past.

Is there a pattern to the timing of attacks?

With machine labeled longitudinal data, we can generate a time-series of attacks in conversations as they occur on each page of Wikipedia. This allows us to ask whether there is a pattern to the timing of attacks in Wikipedia comments. To answer this question, we segment every comment in our corpus by whether or not it is a personal attack. We then build a neighborhood around each comment consisting of the n comments that occurred on the same page immediately before and after it, excluding the central comment. For each central comment, we compute the fraction of attacks that occur in this neighborhood and call this the *neighboring attack fraction* of the central comment.

Table 5 shows the average neighboring attack fraction around attacking and non-attacking comments at different values of n. We see that even for small n, there is a significant difference ($t = 56.2$, $p < 0.0001$) in neighboring attack fractions. Indeed, the neighboring attack fraction for $n = 1$ is twenty-two times higher around attacking comments than non-attacking comments. This is a strong indication personal attacks cluster in time on Wikipedia discussions.[9] It also suggests that early intervention by moderators could be an effective means of curbing the prevalence of personal attacks.

n	Attacking	Non-Attacking
1	15.6 %	0.7%
3	10.6 %	0.7 %
5	8.3 %	0.8 %

Table 5: Average neighboring attack fraction around attacking comments and non-attacking comments.

6. DISCUSSION & CONCLUSION

We have introduced a methodology for generating large-scale, longitudinal data on personal attacks in online discussions. After crowdsourcing the identification of personal attacks within a sample of discussion comments, machine learning classification is leveraged to scale the identification process to the whole corpus. In so doing, we explored methods for aggregating multiple human judgments per comment into training labels, compared different model architectures and text features, and introduced a technique for comparing the performance of machine learning models to human annotators.

We illustrated our methodology by applying it to Wikipedia, generating an open dataset of over 100k high-quality human-labeled comments, 63M machine-labeled comments, and a classifier that approximates the aggregate of 3 crowd-workers. We believe this provides the largest corpus of human-labeled comments supporting the study of online toxicity to date.

By calibrating our classifier's threshold we can then perform large scale longitudinal analysis of the whole corpus of Wikipedia discussions along a wide variety of dimensions. We illustrate this by exploring some open questions about the nature of personal attacks on Wikipedia. This leads to several interesting findings: while anonymously contributed comments are 6 times more likely to be an attack, they contribute less than half of the attacks. Similarly less than half of attacks come from users with little prior participation; and perhaps surprisingly, approximately 30% of attacks come from registered users with over 100 contributions.

These results suggest the problems associated with personal attacks do not have an easy solution. However, our study also shows that less than a fifth of personal attacks currently trigger any action for violating Wikipedia's policy. Moreover, personal attacks cluster in time - perhaps because one personal attacks triggers another. If so, early intervention by a moderator could have a disproportionately beneficial impact. Moreover, automated classifiers may then be a valuable tool, not only for researchers, but also for moderators on Wikipedia. They might be used to help moderators build dashboards that better visualize the health of Wikipedia conversations, or to develop systems to better triage comments for review.

Perhaps the biggest challenge with our methodology is illustrated by our case study with Wikipedia: we used a relatively small set of annotators, 4053 in total, whom we know little about. While they have reasonable levels of inter-annotator agreement, their interpretation of a comment being a personal attack may differ from that of the Wikipedia community. Moreover, the crowdsourced data may also result in other forms of unintended bias.

This brings up key questions for our method and more generally for applications of machine learning to analysis of comments: who defines the truth for the property in question? How much do classifiers vary depending on who is asked? What is the subsequent impact of applying a model with unintended bias to help an online community's discussion?

The methodology and data sets we have developed also open many other avenues for further work. The corpus of human-labeled comments can be used train more sophisticated machine learning models. It can also be used to analyze Wikipedia with traditional statistical inference, while our corpus of machine-labels can be employed to carry out further analysis that require large scale and longitudinal data. While there are many such questions to analyze, some notable examples include:

1. What is the impact of personal attacks on a user's future contributions?
2. What interventions can reduce the level of personal attacks on a conversation?
3. What are the triggers for personal attacks in Wikipedia comments?

[9] A follow up analysis could investigate to what extent an initial attack sparks retaliation.

Finally, we remark that our methodology can easily be applied to different characteristics of comments, not just personal attacks. As mentioned in Sec. 2, there are many taxonomies by which one can analyze the positive and negative properties of a comment. There are also many other discussion corpora to be considered.

Acknowledgements. The authors would like to thank CJ Adams, Dario Taraborelli and Patrick Earley for fruitful feedback and discussions.

APPENDIX

A. WIKIPEDIA COMMENT CORPUS

Here we describe our approach to generating a corpus of discussion comments from English Wikipedia. Every Wikipedia page, including articles and user pages, has an accompanying "talk page" that can be used for communicating with other users. Discussion pages pertaining to user pages are said to belong to the *user talk namespace*, while discussions pertaining to articles belong to the *article talk namespace*. Although there are 35 talk namespaces in total, we focus on these two throughout the paper since they contain at least an order of magnitude more discussion pages and comments than the others.

MediaWiki, the software underlying Wikipedia, does not impose any constraints on editing talk pages. However, an edit on a talk page typically consists of a user adding a comment to a discussion in accordance with a set of formatting conventions. Figure 8 gives an example of a conventionally formatted discussion.

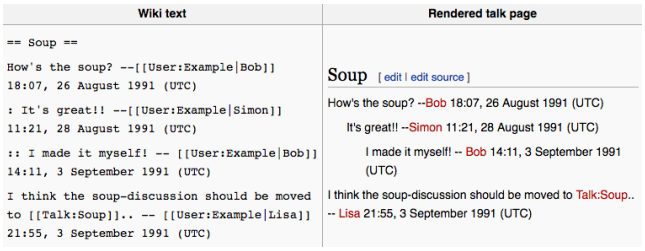


Figure 8: Example of a discussion thread taken from [32]. Includes the raw MediaWiki markdown or "Wiki text" and the corresponding rendering.

One approach to generating a corpus of comments, is to take a current snapshot of all talk pages and parse each page into discussions and comments. The downside of this approach is that comments with personal attacks are usually quickly removed and that comments on user talk pages are often removed after they have been read to reduce clutter. As a result, no single snapshot of talk pages will contain a representative or complete collection of comments made on Wikipedia.

We pursue an alternative approach, which involves processing the "revision history", which represents the history of edits on a page as a sequence of files. There is a separate file, called a revision, corresponding to the state of the article after each edit. We can compute a diff between successive revisions of a talk page to see what text was added as a result of each edit. The benefit of this approach is that it captures all content that has been added to a talk page.[10] The downside is that the content added during a talk page edit is not always a full, new comment. The content added can also represent a modification of an existing comment. For completeness, we include the text added in these types of edits in our corpus.

In practice, we processed the revision history from a public dump of English Wikipedia made available on 2016-01-13. To generate the set of diffs from the revision history, we used the existing *mwdiffs* [8] python package along with the standard longest-common-substring diff algorithm. For the purpose of this study we define a talk page comment as the concatenation of the MediaWiki markup added during an edit of a talk page. We also compute a clean, plain-text version of each comment by stripping out any html or MediaWiki markup, which we use in the crowd-sourcing task described below.

While manually inspecting the data, we found that a large portion of comments left on talk pages (20%-50%, depending on the namespace) were clearly administrative in nature and generated using a bot or template. In this study we are interested in comments made by human users in the context of discussions. After using a regular expression to filter out all messages from these users, we still observed a large number of administrative comments with little or nor modification on many user talk pages. We generated a another set of regular expressions to remove the most commonly occurring comments of this nature. Table 6 gives the number of comments in the user and article talk namespaces after each filtering step.

Namespace	All	No Bot	No Bot/Admin
User	47.3M	36.5M	24.2M
Article	47.8M	39.2M	39.2M
Totals	95.1M	75.7	63.4M

Table 6: Summary statistics of comment corpus broken down by namespace. We first filter out all comments from bot accounts and then filter out messages containing templates used for administrative purposes.

[10]Note that there is a mechanism for removing revisions from the public record, and that personal attacks are a valid reason for doing so [34]. Since this work is based entirely on publicly available data, comments introduced on deleted revisions are not included in our corpus.

7. REFERENCES

[1] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 13:281–305, Feb. 2012.

[2] M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, 6(1):3–5, 2011.

[3] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec. Antisocial behavior in online discussion communities. In *ICWSM*, 2015.

[4] K. Dinakar, R. Reichart, and H. Lieberman. Modeling the detection of textual cyberbullying. *The Social Mobile Web*, 11:02, 2011.

[5] M. Duggan. *Online harassment*. Pew Research Center, 2014.

[6] Fundacion Barcelona Media (FBM). Caw 2.0 training datasets, 2009. `http://caw2.barcelonamedia.org/`.

[7] I. Gagliardone, D. Gal, T. Alves, and G. Martinez. *Countering online hate speech*. UNESCO Publishing, 2015.

[8] A. Halfaker. mwdiffs. `https://github.com/mediawiki-utilities/python-mwdiffs`.

[9] A. F. Hayes and K. Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89, 2007.

[10] Impermium. Detecting insults in social commentary dataset, 2012. https://www.kaggle.com/c/detecting-insults-in-social-commentary.

[11] K. Krippendorff. *Content analysis: An introduction to its methodology*. Sage, 2004.

[12] K. Krippendorff. Reliability in content analysis. *Human communication research*, 30(3):411–433, 2004.

[13] I. Kwok and Y. Wang. Locate the hate: Detecting tweets against blacks. In *AAAI*, 2013.

[14] M. J. Moore, T. Nakano, A. Enomoto, and T. Suda. Anonymity and roles associated with aggressive posts in an online forum. *Computers in Human Behavior*, 28(3):861–867, 2012.

[15] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive language detection in online user content. In *WWW*, 2016.

[16] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[17] S. Pieschl, C. Kuhlmann, and T. Porsch. Beware of publicity! perceived distress of negative cyber incidents and implications for defining cyberbullying. *Journal of School Violence*, 14(1):111–132, 2015.

[18] B. Plank, D. Hovy, and A. Søgaard. Learning part-of-speech taggers with inter-annotator agreement loss. In *EACL*, pages 742–751, 2014.

[19] H. M. Saleem, K. P. Dillon, S. Benesch, and D. Ruths. A web of hate: Tackling hateful speech in online social spaces. In *TA-COS*, 2016.

[20] A. Schrock and D. Boyd. Problematic youth interaction online: Solicitation, harassment, and cyberbullying. *Computer-Mediated Communication in Personal Relationships*, pages 368–398, 2011.

[21] S. O. Sood, E. F. Churchill, and J. Antin. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology*, 63(2):270–285, 2012.

[22] N. Spirin and J. Han. Survey on web spam detection: principles and algorithms. *ACM SIGKDD Explorations Newsletter*, 13(2):50–64, 2012.

[23] Support and Safety Team. *Harassment Survey*. Wikimedia Foundation, 2015. https://upload.wikimedia.org/wikipedia/commons/5/52/Harassment_Survey_2015_-_Results_Report.pdf.

[24] J. R. Tetreault, E. Filatova, and M. Chodorow. Rethinking grammatical error annotation and evaluation with the amazon mechanical turk. In *NAACL-HLT*, 2010.

[25] R. S. Tokunaga. Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in human behavior*, 26(3):277–287, 2010.

[26] M. A. Walker, J. E. F. Tree, P. Anand, R. Abbott, and J. King. A corpus for research on deliberation and debate. In *LREC*, pages 812–817, 2012.

[27] W. Warner and J. Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26. Association for Computational Linguistics, 2012.

[28] Z. Waseem and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of NAACL-HLT*, pages 88–93, 2016.

[29] D. Wiener. Negligent publication of statements posted on electronic bulletin boards: Is there any liability left after zeran. *Santa Clara L. Rev.*, 39:905, 1998.

[30] Wikimedia. Harassment consultation 2015. https://meta.wikimedia.org/wiki/Harassment_consultation_2015.

[31] Wikimedia. Machine-learning tool to reduce toxic talk page interactions. https://meta.wikimedia.org/wiki/2015_Community_Wishlist_Survey/Bots_and_gadgets#Machine-learning_tool_to_reduce_toxic_talk_page_interactions.

[32] Wikipedia. Help:Talk pages. https://www.mediawiki.org/wiki/Help:Talk_pages.

[33] Wikipedia. Wikipedia:No personal attacks. https://en.wikipedia.org/wiki/Wikipedia:No_personal_attacks.

[34] Wikipedia. Wikipedia:Revision_deletion. https://en.wikipedia.org/wiki/Wikipedia:Revision_deletion.

[35] N. E. Willard. *Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress*. Research Press, 2007.

[36] E. Wulczyn, N. Thain, and L. Dixon. https://figshare.com/articles/Wikipedia_Detox_Data/4054689.

[37] G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *CIKM*, 2012.

[38] J.-M. Xu, B. Burchfiel, X. Zhu, and A. Bellmore. An examination of regret in bullying tweets. In *HLT-NAACL*, pages 697–702, 2013.

[39] M. L. Ybarra and K. J. Mitchell. Youth engaging in online harassment: Associations with caregiver–child relationships, internet use, and personal characteristics. *Journal of adolescence*, 27(3):319–336, 2004.

[40] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards. Detection of harassment on web 2.0. In *WWW*, 2009.