Characterizing Twitter Users Who Engage in Adversarial Interactions against Political Candidates

Yiqing Hua, Mor Naaman, Thomas Ristenpart
Cornell Tech, Cornell University
yiqing@cs.cornell.edu, mor@jacobs.cornell.edu, ristenpart@cornell.edu

ABSTRACT

Social media provides a critical communication platform for political figures, but also makes them easy targets for harassment. In this paper, we characterize users who adversarially interact with political figures on Twitter using mixed-method techniques. The analysis is based on a dataset of 400 thousand users' 1.2 million replies to 756 candidates for the U.S. House of Representatives in the two months leading up to the 2018 midterm elections. We show that among moderately active users, adversarial activity is associated with decreased centrality in the social graph and increased attention to candidates from the opposing party. When compared to users who are similarly active, highly adversarial users tend to engage in fewer supportive interactions with their own party's candidates and express negativity in their user profiles. Our results can inform the design of platform moderation mechanisms to support political figures countering online harassment.

Author Keywords

Online Harassment; Twitter; User Behavior; Political Candidates

CCS Concepts

•**Human-centered computing** → **User studies; Empirical studies in HCI;**

INTRODUCTION

Social media is a natural place for political figures to connect to the public [30, 38], both by broadcasting and — as importantly — listening to their constituents and others [14]. At the same time, social media services have been struggling with the prevalence of online abuse and harassment [4, 8, 21, 22, 32]. Politicians in particular are common targets of negativity online, with reports consistently showing that politicians face overwhelming amounts of online harassment [5, 45]. This trend has discouraged some from engaging in conversations online [59], caused others to quit seeking public office [45], and may have chilling effects on those who would otherwise

engage in democracy and public service. It is therefore crucial to understand how to ensure a reasonable level of discourse on social media for political figures and the broader public.

Making sense of incivility is complicated, with no clear definitions of what types of content may cross acceptable boundaries in which contexts. A small but growing body of research uses data-driven analyses to increase our understanding of incivility towards politicians [29, 33, 56], complementary to similar studies of harassment in other contexts [9, 10, 23, 40]. These works use various definitions of "uncivil" behavior. For our purposes, we broadly refer to messages as adversarial if they are intended to hurt, embarrass, or humiliate a targeted individual, although specific guidelines and accurate judgment of such a definition is challenging, as we expand on below.

In this work, we use a combination of automated tools and qualitative coding techniques to provide the richest characterization to date of Twitter users who engage in adversarial interactions with political candidates. Fueling this analysis is a dataset we collected[1], capturing Twitter interactions with political candidates during the run-up to the U.S. midterm election in November 2018. Our dataset consists of all tweets by 1,110 accounts manually validated as used by 756 candidates for the U.S. House of Representatives. Additionally, it includes 1.2 million replies to the candidates' tweets made by 0.4 million unique Twitter users, and these users' Twitter profile information. We focus our analysis on replies to candidates' tweets. Twitter replies are, by default, visible to other users interacting with the candidates' tweets.

With this data, we first identify users who post adversarial replies to any of the candidates by using Perspective API [36], a general language tool for toxic content detection. We show that while a significant number of adversarial interactions are generated by users who only interact with candidates a few times, over 35% of the adversarial replies are created by just 10% of the users who repeatedly post adversarial contents towards candidates. We also explore the factors that are associated with higher levels of adversarial interactions with candidates, showing that the less a user is embedded in the political discussions with candidates in general, the more they tend to use adversarial language when replying to candidates.

We then focus on users who engage in significantly more adversarial activity than others. We perform qualitative coding

---

[1]Available at: `https://figshare.com/articles/U_S_Midterm_Election_Twitter_Dataset_2018/11374062`

to better understand how highly-adversarial users (those who posted more than 10 adversarial tweets to candidates in our dataset) are different than other user groups in terms of the content they post and the information they expose in their user profile. We show that, the highly adversarial users post more personal attacks and generate more off-topic replies, namely ones that are irrelevant to the candidate's tweet. In addition, these users' Twitter profiles are more likely to contain partisan and adversarial attacks, potentially foreshadowing their negative activity. Finally, we evaluate how well the automated tool we used for much of our analysis achieves good agreement with our human coders, validating Perspective API's effectiveness in identifying highly-adversarial actors. If anything, the methods we used slightly under-estimate the amount of adversarial content these users shared.

Our work characterizes adversarial social media users in an important societal context, those who target political candidates. To this end, we perform context-specific analysis and show findings that both extend, and conflict with, those from previous works that also focus on adversarial users on Twitter but in different settings [9, 10, 23, 54]. Our work provides a better understanding of the activities and characteristic of these "adversarial users", which can help inform design of social platforms, policies against adversarial interactions, as well as mechanisms to allow for better moderation and interactions with political figures in social media.

RELATED WORK
A significant amount of previous work has focused on measuring and understanding adversarial behaviors online, including studies on victims' experiences [41, 64] and attackers' motivations [11, 12, 40]. Other studies have developed machine learning tools that can be deployed at a large scale to support detecting personal attacks [20, 46, 65]. We focus on three themes of related work that are pertinent to our work here: (1) studies of online harassment on Twitter, (2) characterizing abusive Twitter users, and (3) studies of adversarial interactions against political figures.

Online harassment on Twitter. Twitter has a particularly severe problem with the amount of online harassment on the platform [26, 41]. Victims who are targets of ongoing harassment reported that frequent encounters with harassing content disrupts their day-to-day lives [39]. Yet, various institutions [4, 41] have reported that harassment reports from victims often are not addressed properly.

Detecting and characterizing abusive users on Twitter. A number of recent projects attempt to characterize and automatically identify abusive users on Twitter [9, 10, 23, 54]. Some of the features considered by these projects include Twitter account metadata [10, 54], personality traits [23], tweeting patterns [40], and social network structures [9, 10, 23, 54] associated with abusive behaviors. These studies used computational methods to provide quantitative insights on behavioral patterns of abusive Twitter users in different contexts [9, 10, 23, 40, 54]. Many of these works sample abusive or hateful speech by searching for a dictionary of explicitly offensive words [23, 40, 54], or mentions of known harassment campaigns [9, 10]. Our work differs from these previous

efforts in several ways. First, we focus on an important societal context — political candidates — constructing our data by collecting all Twitter interactions with this specific set of users. We leverage unique aspects of this political context (e.g., political leanings, political engagement on social media) to better analyze adversarial activities. Further, we complement our quantitative measurements with context-specific analysis based on qualitative coding to provide a deeper and more detailed understanding of user behaviors.

Adversarial interactions against politicians. Adversarial interactions against political figures, including politicians and candidates, have been studied long before the existence of social media (e.g., [17]). Recent surveys at politicians from multiple countries show that most of them have experienced both online and offline harassment [1, 24, 35, 52]. The rise of social media provides political figures with an important channel to connect with citizens [2, 14, 51], as well as a platform where they can be easily addressed and targeted by online harassment [5, 24, 29, 33, 45]. It has been reported that social-media-based abuse discourages politicians from engaging in online conversations [59]. At the extreme, some reported quitting seeking public office due to the amount of harassment on social media [45]. Naturally, many political figures have professional teams running their social media accounts, insulating them to some extent from harassment. However, this does not obviate the ill effects of adversarial content, for example such content might have negative impact on other people who engage with the candidate's tweets [11].

Closest to our work is research from Gorrell et al. [28, 29], examining Twitter harassment towards United Kingdom parliament members with emphasis on understanding what attributes of politicians – such as party and gender – and what political topics attract abusive replies. Our own work used the same dataset we analyze here to study how this set of U.S. House candidates experienced harassment on Twitter in that time period [33]. Both works focus on measuring trends in the adversarial interactions experienced by the candidates, such as the different forms that adversarial interactions may take [33], or topics that adversarial interactions focus on [28, 29]. In contrast, our focus here is on characterizing the *users* that are making adversarial posts.

DATA COLLECTION AND ANALYTIC APPROACH
We use a data-driven approach to gain a better understanding of the users who post adversarial replies to political candidates online. We define adversarial interactions as messages posted by users which are intended to hurt, embarrass or humiliate a target user, following a notion that has been largely accepted in the community [13, 18, 19, 31, 33, 40, 57]. Of course, determining whether a message satisfies this definition can be highly subjective, and we will use a combination of existing machine learning techniques, and, in a following section, human annotations, to assess whether messages are adversarial.

Our dataset. We built on a set of $1,110$ Twitter accounts used by 786 candidates, 431 Democrats (D) and 355 Republicans (R) running in 2018 for the U.S. House of Representatives developed in our earlier work [33]. Our study is focused on 1.2 million Twitter replies to the candidates' posts, made by

0.4 million users between September 17th and November 6th, 2018 (the day of the U.S. midterm elections). We collected the data using the Twitter Streaming API, including all tweets posted by, replying to, or retweeting any of the candidate accounts. On Twitter, replies to public accounts directly engage with the target account and are visible by default to the public who are viewing the original posts. Thus, replies to candidates directly contribute to the candidate's own "public sphere" [60]. We therefore focus on replies to candidate posts in this work. We do not include Twitter "mentions" of these accounts and tweets that include a link to tweets by these accounts (often called "quote tweets"), which may be visible to others on Twitter but are not immediately available to those interacting with a candidate's account. We alternatively refer to tweets in our dataset as "replies" or simply "tweets"—but note that all of these are direct replies to tweets by the candidates.

We augment the dataset of replies with the Twitter following network for the users in the dataset. On Twitter, analyzing who users follow has proven to be informative for inferring user interests and political preferences [7, 55]. We therefore retrieve the 5,000 most-recently followed accounts by each user in our dataset (i.e., their "Twitter friends") using the Twitter API [62]. Due to API rate limits, this data collection lasted over a longer time period, from September 2018 until March 2019. This network data is limited in at least three ways. First, by the time of data collection, some users have left Twitter or set their profiles to be private, preventing us from retrieving their friends list. However, since our data collection prioritized the more active users in our dataset, we do not think these omissions are critical. Second, users might follow new accounts between the time they made a reply to a candidate and their data being collected by us, resulting in some inaccuracies. Last, we were limited to retrieve only the 5,000 most recent followers from any user. As a result, we obtained the complete friends list for 92% of the users. For an additional 4% of users who followed more than 5,000 accounts, we obtained the partial friends list. The remaining 4% of accounts were deleted or suspended at the time of our network data collection.

Note that it is possible that organized *information operations* took place on Twitter during our data collection period, whether using bots powered by automated algorithms [16, 25] or through state-operated campaign [6, 34]. We compared our data with a list of state controlled accounts with over 5,000 followers published by Twitter [43] (as the information of this set of accounts was not anonymized): none of them showed up in our dataset. Further, manual review of highly-active and highly-adversarial users as discussed later indicates that these users exhibit sophisticated behaviors that were not likely generated by automated algorithms. Finally, most of our analyses would remain relevant even if the accounts are automated or are part of organized campaigns since the resultant adversarial activities are nevertheless perceived by, and have impact on others, e.g. the candidates and the observers.

Automated labeling of adversarial interactions. The scale of this dataset suggests the need for a computational approach to analyze adversarial interactions. We use machine learning

τ	0.5	0.6	0.7	0.8	0.9
τ-adversarial tweets (%)	28%	21%	14%	9%	4%
Average per user	2.2	2.0	1.8	1.6	1.3

Table 1: Fraction of τ-adversarial tweets in our dataset, for varying thresholds τ, and the average number of τ-adversarial tweets made by a user that made at least one such tweet.

tools for flagging toxic content in the first part of this work, and as a mechanism to identify content to analyze qualitatively in the second part of this work. Specifically, we use Perspective API [36] to assign toxicity scores to tweets. The API has been used before in different contexts, including Twitter [23, 33]. The scores generated by Perspective API indicate whether a general language model, trained with Wikipedia discussions and multiple other online conversation sources [36, 65], believes the utterance to be "discouraging participation in conversations". The API outputs a "toxicity" score in the range $[0, 1]$, with 1 being highly toxic and 0 being non-toxic. We define a tweet as τ-*adversarial* if its content has a toxicity score more than τ. Note that although similar, toxicity is in fact a different concept than being adversarial (see the definition we use at the beginning of this section), as one can be toxic without being adversarial (e.g., using foul language as expression), and adversarial without being toxic (e.g., combative speech). Here, we use toxicity in replies as an approximation of adversarial actions. In our earlier work, we validated the precision of this technique for 0.7-*adversarial* tweets, showing this method offers high precision [33].

Nevertheless, Perspective API as well as other existing automated approaches that use general language models to flag adversarial content [20, 46] have inherent limitations, in particular having potentially low recall for some forms of adversarial interactions. We mitigate the limitations inherent in using a general language model in two key ways. First, we provide analysis showing that the trends in our results are robust to the toxicity thresholds we use in our computational analysis. Second, we report on a large-scale qualitative coding of tweets that helps expose whether the method is prone to specific biases in the sources of abuse it detects.

VOLUME OF ADVERSARIAL ACTIVITY

To gain an initial understanding of the use of toxic language by different users, we analyze the distribution of τ-adversarial activities per user observed in our dataset. The analysis is geared towards exposing the levels of activity of adversarial users, and how those levels relate to the general activity levels of users interacting with candidates. A secondary goal of the analysis is to understand the sensitivity of these trends to the values used for the threshold τ. In general, changing the threshold results in a significant change to the number of tweets classified as toxic: Table 1 provides some indication on how raising the threshold τ decreases the number of tweets considered as toxic.

Regardless of the specific τ, a relatively small number of users generate a large fraction of adversarial tweets, in line with prior studies on online communities [47, 53]. Figure 1a
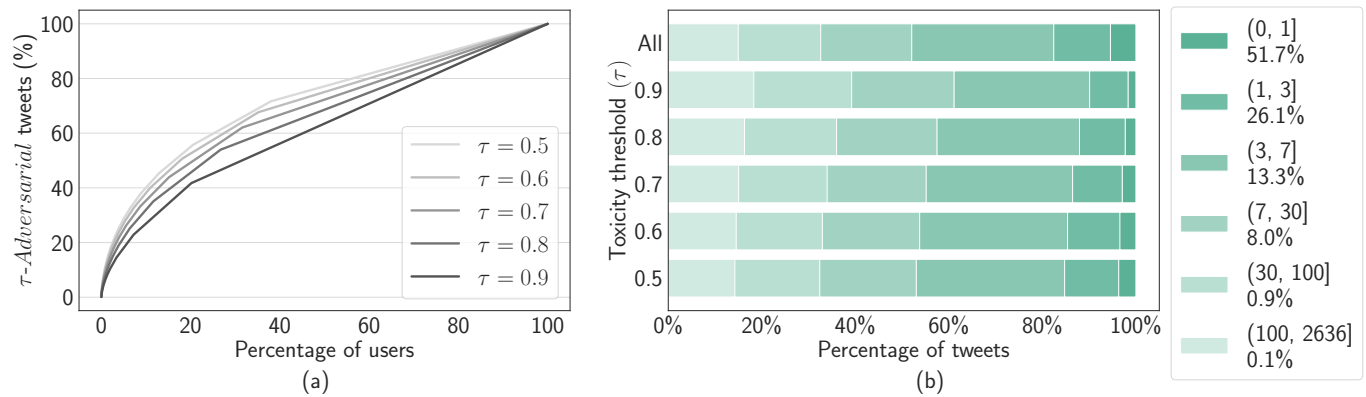
Figure 1: **(a)** CDF of τ-adversarial tweets contributed by users. **(b)** Breakdown of replies by users with different *overall* activity (top row), with the fraction of adversarial tweets made by each user group, for various τ (following rows). The legend specifies user overall activity level for each group, with the group size in percentage.

shows the cumulative fraction of τ-adversarial tweets (Y-axis) contributed by an increasing fraction of the users (X-axis), for various values of τ. Users are sorted in descending order of the number of τ-adversarial tweets they made. For example, 49% of the 0.7-adversarial tweets were posted by 20% of the users and 35% of the 0.7-adversarial tweets were generated by just 10% of the users. Note that the total number of tweets is different for each curve, as indicated by Table 1.

Users who had fewer overall interactions with the candidates contribute more adversarial replies than their overall contribution predicts. Figure 1b breaks down the proportion of adversarial activity with different τ (separated by rows), contributed by users with different levels of *overall* activity in our dataset (shades of green). The lightest shade (left) are users who are at or below the median (51.7% of the users in our dataset that posted a total of 1 tweet reply to candidates). The following darker shades are the 75th percentile (2 or 3 tweets), 90% (4–7 tweets), 99% (8–30 tweets) and 99.9% (31–100 tweets). The most active 0.1% (over 100 tweets) is in darkest shade of green. The top row of the chart shows the breakdown of all tweets by these groups, while the remaining rows show a breakdown of adversarial tweets by activity level groups for different levels of τ-adversarial content. For example, the figure shows that 51.7% of users who replied to candidates only once (leftmost, lightest shade) contributed 15% of all replies (top row), but over 18% of the 0.9-adversarial replies (second from top). Grouped together, the top 10% most active users (the three rightmost, darkest groups) contributed 48% of all replies (top bar) but only 39% of the 0.9-adversarial replies (second from top bar). In summary, while "heavy" adversarial users still carry most of the weight, Figure 1b shows that "casual" offenders reply more with adversarial content than their overall activity predicts.

CHARACTERIZATION OF ADVERSARIAL USERS

We are interested in the tendency of users to share more adversarial content than expected given their basic user characteristics, engagement with political topics and adversarial behavior of Twitter friends. To this end, we perform an analysis of a set of moderately active users — those who posted between 4

to 30 replies to candidates. The $[4, 30]$ user group comprises only 21% of the user base (N=74,533), but authored 50% of all tweets and 52% of the adversarial tweets in our data. As the distribution of user activity is highly skewed, this group excludes the under- (users who produced fewer than 4 tweets) and over-producers (users who produced more than 30 tweets) of content. The over-producers were removed because their behaviors are likely to be materially different than less active users. The under-producers do not have enough content in our dataset to provide meaningful analyses. We performed a similar analysis that included the under-producers in addition to the $[4, 30]$ user group, and the regression results reflected a similar trend.

Specifically, we focus on three categories of features in our analysis: the basic characteristics of the user as reflected in their Twitter account metadata; the specific characteristics of the user's Twitter "political activity" (in our data); and features that capture adversarial behaviors by a user's friends on Twitter. We fit these features into a linear regression to the number of τ-adversarial tweets by the user as calculated using Perspective API. For simplicity, we use $\tau = 0.7$ and refer to 0.7-adversarial tweets as "adversarial tweets" from now on. We set this threshold based on prior work [33] and validation of precision associated with the 0.7 threshold in the later discussion via manual labeling. Also, as shown above, the analysis trends are robust to small changes in τ.

Regression Features

Basic user characteristics. We start with a set of basic attributes about users inspired by previous studies: their number of followers, the age of their account, whether their account is verified or not, and their level of activity. Prior works in other contexts have shown that the number of followers is correlated with abusive behavior. Interestingly, the effect is context dependent, with some studies indicating that more followers a user has is positively [10] correlated with adversarial behavior while other studies show the opposite trend [9, 54]. Those prior works also use the age of an account (i.e., how long has this user been registered on Twitter), and, again, show differing correlations with adversarial behavior [10, 54]. Pre-

vious studies on online forums show that anonymity can be used as a criteria to distinguish attackers from defenders of aggressive behavior [44, 66]. We do not have a robust measure of anonymity of Twitter accounts, as even those who seem to be using their real names could be using others' identities. Instead, we use the Twitter "verified" [63] status as a (limited) replacement. This feature is offered by Twitter to users who are public figures, and marks the fact that the user identity had been verified by Twitter. Finally, we include the number of replies posted by user to candidates in our dataset as a control, to account for the heightened user activity level that may give them more opportunity to behave in an uncivil manner.

Engagement with political topics. A second set of features we explore captures various aspects of a user's political activity on Twitter, including interactions with candidates, as well as the candidate accounts followed by the user and the amount of retweeting candidates' posts. We use the number of candidates followed and number of candidate tweets retweeted by a user as approximate measures of a user's interest in politics. Using methods described in [33], we label each user as either pro-Democrat or pro-Republican based on their retweeting of candidates, except for 1% of the moderately active users for whom we do not have enough retweet information to classify partisanship, who are excluded from the analyses in this section. Opinion conflicts are indicative features of incivility [40]. To measure the potential frequency of opinion conflicts, we include both the number of candidates of the user's opposing party that the user has replied to, and the percentage of the user's replies that were directed towards candidates from the opposite party.

Another measure of the user's level of political engagement is in their self-description. For example, some users include partisan hashtags such as #MAGA in their Twitter profile. Using a list of partisan hashtags compiled from [33], we identify users who express their political preferences in their user profiles. This feature is used in the following analysis as a signal of a user's partisanship. A final measure we use for political engagement is the centrality of the users in the discussions involving the candidates. To compute this measure, we calculate the page rank score [50] of each user in our dataset in the network of the following relationship of all the users who reply to political candidates (i.e., all users in our full dataset). In other words, the network is the partial graph of the Twitter following network where the set of nodes is limited to the users in our data.

We have also considered other factors, including absolute number of replies to candidates from the opposing party, number of followers supporting the same party, and number of candidates followed by the user from the party they support — but ultimately excluded them from the analysis due to multicollinearity [49] with other variables we are already considering.

Adversarial behavior of Twitter friends. It is known that a user's adversarial activity level may be associated with the activity of their Twitter friends (i.e., the accounts a user is following on Twitter) [3, 42]. We consider two attributes to measure this effect: the total number of adversarial tweets posted by a user's Twitter friends, and the percentage of one's

	Model 1	Model 2	Model 3	Mean	Std
(Intercept)	1.237***	1.237***	1.237**		
	(0.006)	(0.006)	(0.006)		
Number of replies	1.49***	1.314***	1.307***	8.14	5.21
	(0.012)	(0.013)	(0.013)		
Account age (days)	-0.04**	-0.052***	-0.004	1818	1201
	(0.012)	(0.012)	(0.012)		
Number of followers	-0.052***	-0.036*	-0.138***	2.36	0.91
	(0.015)	(0.016)	(0.022)		
Verified on Twitter (99.8% not verified)	-0.4**	-0.006	-0.218		
	(0.132)	(0.012)	(0.136)		
Number of retweets		-0.032*	-0.044**	0.71	0.67
		(0.013)	(0.013)		
Candidates followed		-0.048***	-0.051***	11.4	19.6
		(0.012)	(0.013)		
Page rank		-0.068***	-0.118***	0.09	0.08
		(0.015)	(0.015)		
Partisan hashtag in profile (bool) (80% don't have partisan hashtag)		0.085***	0.051**		
		(0.013)	(0.017)		
% replies to opponent candidates		0.429***	0.426***	53%	38%
		(0.015)	(0.015)		
Opponent candidates replied to		0.43***	0.433***	2.14	1.93
		(0.016)	(0.016)		
Adversarial tweets posted by friends			0.076**	1.49	0.99
			(0.025)		
% friends posted ≥ 3 adversarial tweets			0.204***	1.1%	1.2%
			(0.017)		
R^2	0.18	0.227	0.231		

Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table 2: Results of the regression analysis predicting the number of adversarial tweets posted by a user (number of users is $N = 74,533$). Histograms represent the distribution of the variable. Log transformed variables are listed in italic, with mean, standard deviation and distribution in log scale.

friends that posted at least x adversarial tweets. We use $x = 3$ in the following analysis, but verified that the results of $x = 1, 2, 4, 5$ are similar.

Results

We analyze the power of the attributes listed above in predicting the dependent variable — the number of 0.7-adversarial tweets made by the user — using three increasingly complete linear regression models that incrementally add each set of features described above. *Model 1* uses the basic user Twitter features (the first group described above), including controlling for the user's overall activity level. *Model 2* combines these features with the features capturing the user's political activity in our dataset. Finally, *Model 3* adds the features capturing adversarial behaviors of the user's Twitter friends.
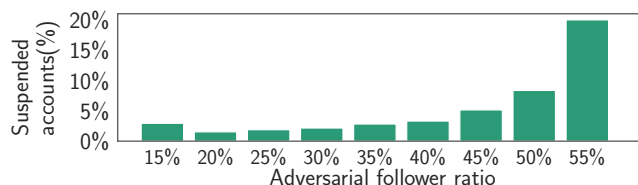
Figure 2: Users with higher adversarial follower ratio are generally more likely to have been suspended.

The histograms in Table 2 show the data distribution of each variable used in the regression models, along with its average and standard deviation (rightmost two columns). As the table histograms show, not all variables are normally distributed. To account for this, we log-transform the number of followers, number of retweets, page rank score, and amount of adversarial activities of Twitter friends (the italic features in Table 2). Furthermore, all continuous variables were standardized by centering and scaling by two standard deviations following recommended practice [27]. The binary variables were centered in order to have the same scale for comparison. Hence the magnitude of the standardized coefficients reflects the significance of the variables. Table 2 also shows the results of the regression analysis, including the standardized beta coefficients (b) and their standard errors (SE) for all three models. The p values are computed using two-tailed t-tests. All variance inflation factors (VIFs) are less than three, indicating that multicollinearity is not an issue in our independent variables. Overall, the three models explain $R^2 = 0.18$, $R^2 = 0.227$, and $R^2 = 0.231$ of the variance of the dependent variable (bottom row of the table).

Among the features included in *Model 1*, the number of replies has the strongest correlation with adversarial replies ($b = 1.49$): users who make more adversarial replies to candidates are more active in general. It retains its strength and significance across the other two models. In *Model 3*, for example, each additional reply is related to an increase of 0.13 in the number of adversarial replies (a standardized coefficient of $b = 1.307$, scaled by twice the standard deviation of 5.21). Even after controlling for the overall number of replies, the scale of adversarial activity of the user is correlated with other attributes. For example, a higher number of followers is associated with less adversarial activity. According to *Model 3*, every tenfold increase in followers is associated with a 0.08 decrease in adversarial activity. On the other hand, there is inconsistent significance for account age and verified status across the three models, suggesting these are not particularly powerful predictors.

A user's political activity attributes are strongly correlated with level of adversarial activity. The most significant factor is the tendency to engage with candidates from the opposing party on Twitter. This trend was measured by the number of opponent candidates the user replied to and the percent of their tweets dedicated to opponent candidates, both of which correlated with an increase in adversarial activity. An increase of 1% in the number of replies towards opponent candidates is correlated with a 0.006 increase in the quantity of adversarial replies, while each additional opponent candidate replied to

raises the adversarial activity by 0.11. Including a partisan hashtag into one's user profile correlates with a 0.051 increase in the predicted amount of adversarial activity. In contrast, the centrality in the political user network measured by page rank is negatively correlated with adversarial activity. The less "important" users in that network tend to engage in more adversarial interactions. Other indicators of political activity — number of retweets and candidates followed, both associated with partisan support — are also *negatively* correlated with adversarial activity, though with less significance. For example, according to *Model 3*, tenfold increases in the number of retweets and each additional candidate followed are correlated with 0.033 and 0.001 decrease in number of adversarial tweets, respectively.

Finally, *Model 3* shows that adversarial behaviors by friends has strong correlation with a user's behavior in posting adversarial posts. A 1% increase in friends who posted at least three adversarial tweets is correlated with 0.08 more adversarial tweets posted by the user.

Overall, our findings demonstrate a number of key trends correlated with higher adversarial activity towards candidates on Twitter. Users who engaged in more adversarial interactions with candidates tend to have less followers and are less central in the network of users interacting with candidates. In other words, adversarial activity was higher among "fringe" users. Further, the findings indicate that adversarial activity is correlated with more attention to opponent candidates and reduced attention to the user's own candidates. Moreover, adversarial activity is negatively correlated with (likely) support-oriented activity on Twitter: the more a user retweets or follows candidates (presumably, of their partisan alignment), the less adversarial activity they participate in. Conversely, users who pay relatively more attention to and interact with more candidates from the opposing party produce more adversarial replies.

Accounts followed by Adversarial Users
The analysis above has shown that a user's adversarial replies to political candidates are associated with both the number of friends who adversarially interact with candidates and the magnitude of their friends' adversarial activities. In this section we explore whether the behaviors of the accounts followed by disproportionally more adversarial users are indeed more problematic. We performed an analysis by collecting data about accounts followed by our adversarial users. Note that these accounts may not necessarily be included in our dataset, as they may have not replied to any House candidate during the data collection period. In the following comparison, we selected two groups of users from our dataset: (1) the *Adversarial* group: users who posted at least 3 adversarial tweets, and (2) the *Baseline* group: users who posted at least 3 tweets and never posted adversarial tweets. We define the *Adversarial Follower Ratio* (AFR) of an account as the proportion of followers from the *Adversarial* group over the number of followers from both groups. In March 2019, we queried Twitter for the user statuses — i.e. still active on Twitter or not — of a set of 21 million accounts who are followed by users in our dataset. By the time of data collection, 2% of the accounts had
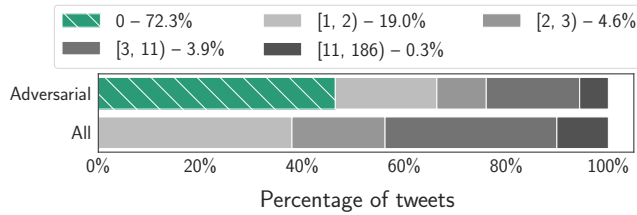
Figure 3: Breakdown of tweets by different user groups, as divided by the amount of adversarial interactions made (specified in the legend with group size in percentage).

been either deleted or suspended from Twitter. We focus on the 97 K accounts that had at least 100 followers in total.

Figure 2 depicts the rate at which accounts with different AFR were suspended by Twitter. Each bin represents at least 100 accounts. Interestingly, the chart shows that accounts with higher AFR are more likely to be suspended. For example, 20% of the accounts with AFR in the range 50–55% were suspended (rightmost bar), which is significantly higher than the 2% rate of suspension for all collected accounts. Account suspension on Twitter may result from activities like spamming, showing signs of account being compromised, and abusive tweets or behavior [61]. We have no information about why these accounts were suspended. It is possible that Twitter is using AFR or similar metrics to prioritize moderation. More likely, the other behaviors of these users that led to suspension were correlated with their AFR.

ANALYSIS OF HIGHLY ADVERSARIAL USERS

It is important to consider and understand in more depth the group of highly adversarial users, defined here as those who made at least 11 posts that are 0.7-adversarial. Figure 3 depicts the proportion of all tweets (top bar) and of adversarial tweets (bottom bar) contributed by each user group, where the groups are based on the number of *adversarial* replies made per user (unlike Figure 1b that grouped users by total reply volume). The highly-adversarial group (dark, rightmost area of both bars) makes up only 0.3% (1,010 users) of the total population of users interacting with the candidates, but these users contributed 10% of the adversarial replies to candidates (bottom bar). This group was very active in general, contributing 5.6% of the total replies to candidates (top bar).

The small size of this user group allows us to use qualitative coding methods to better understand these users' activities. We first evaluate the type of content they produce, in particular in comparison to a control of replies from other highly-active users. We also examine whether the 0.7-adversarial threshold is in some way biased in detecting and interpreting the activities of this adversarial group compared to others. In addition we perform an analysis based on qualitative coding of the information in the users' profiles, extending the "use of political hashtag" findings from the last section.

Content analysis of users' replies to candidates. To better understand the activities of highly adversarial users who are responsible for an outsized amount of hostile content, we performed qualitative coding of a sample of tweets from both *highly adversarial* and *highly active* users. The analysis was

geared to address two questions: (1) Does our use of Perspective API bias our view of who is an adversarial user? For example, users may exploit different vocabularies or engage in actions that are easier, or harder, for general language models to detect. (2) Does the content of negative interactions with candidates differ between highly adversarial and highly active users? Moreover, does the content of positive interactions differ? Note that we use the terminology of negative/positive here instead of adversarial/non-adversarial, as we are not only focusing on hurtful language, but rather any kind of negative interaction, including civil disagreement and criticism. We make the distinction between negative and adversarial interactions as many of the interactions such as attacks against the candidate's policy or political view may not be adversarial, i.e., not intending to hurt, embarrass or humiliate. Rather, negative but civil interactions are necessary for democratic discussions.

We engaged in qualitative coding of a sample of 800 tweets from both user groups. Specifically, we sampled 400 tweets each from highly adversarial users with at least 11 0.7-adversarial replies in our dataset (U_{adv}), and highly active users with at least 11 replies in our dataset (U_{active}, 19,797 users in total). Note that $U_{adv} \subset U_{active}$. For each group, we sample 200 tweets that are 0.7-adversarial and 200 tweets with adversarial score ≤ 0.7 (i.e., non-0.7-adversarial or "clean" according to the threshold-based analysis). We denote these sets of tweets $T_{adv,adv}$ (adversarial users' adversarial tweets) and, accordingly, $T_{adv,clean}$, $T_{active,adv}$ and $T_{active,clean}$.

The coding scheme was developed by the research team in a sequence of coding sessions, where two of the authors independently coded a subset of tweets, then discussed their codes to simplify and streamline the task. We ended up with a coding scheme that asks coders to rate:

- *Does the tweet contains civil, mildly uncivil, or severely uncivil content?* As rater confusion was concentrated around the "mildly uncivil" and "severely uncivil" categories, we merged them into one "uncivil" label in the analysis.
- *What type of negative interaction is in the tweet, if any?* Raters are allowed to assign multiple labels including personal attacks, accusations, other attacks against the candidate, attacks against the candidate's affiliation or leadership, and attacks against the candidate's policy or political view.
- *Does the tweet express support for the candidate, the candidate's party affiliation, or his/her political views/policies?* Raters are allowed to assign multiple labels.
- *Is the reply relevant to the original tweet?* We asked raters to choose a relevance label for each tweet among the following options: irrelevant, same/similar topic, somewhat relevant, and directly responding. In our analyses we converted the last three labels into a single "relevant" category.

Tweets were coded individually using a web-based interface. In each screen we present coders not only the reply tweet to be annotated, but also the original tweet by the candidate for context. In addition, the display included the candidate's political affiliation (Democrat or Republican). We did not expand URLs or images in the original or reply tweets. The 800 tweets were assigned to coders in random order. In total, five different members of the research team labelled 800 tweets, with each
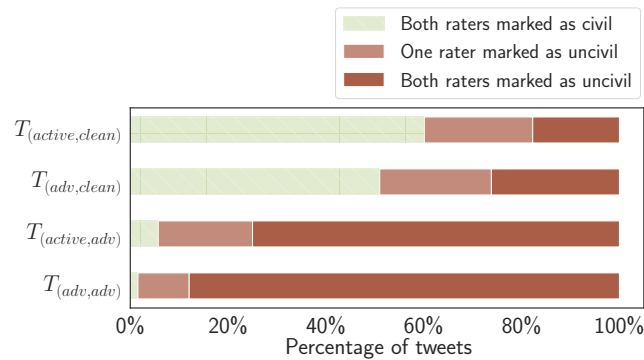
Figure 4: Aggregating the scores of two human raters shows significant agreement with the automatic classification results.

tweet being coded by two team members for a total of 1600 labels. We note that, in our estimate, coders exhibit liberal leaning in terms of political preference and have higher than average exposure to U.S. politics, though most of them were U.S. residents but not U.S. citizens. Familiarity with politics was required to allow the coders to label more complicated questions, such as whether a tweet contains a political topic, and whether comments on specific topics might be considered subtly adversarial [33]. We excluded all tweets that coders were not able to evaluate, most frequently when a tweet only contained a URL (which we did not expand). Our final data used in this analysis includes 690 tweets in total, 350 of them come from the highly active group ($T_{active,adv} \bigcup T_{active,clean}$).

Human annotation exhibited significant agreement with algorithm-labeled adversarial-ness. The raters were reasonably consistent in evaluating whether a given reply to a candidate contained content that can be labeled as "civil" or "uncivil", with 0.6 based on Cohen's Kappa, high end of the moderate agreement range. Figure 4 aggregates the *civility* ratings. Tweets that were given "uncivil" scores by both raters are marked in dark red while those that were given "civil" scores by both raters are marked in light green and with a stripe pattern. Tweets that were assigned with disagreeing labels are marked in light red. The rows represent the sets of tweets analyzed based on the group that produced them (active or adversarial users), and whether the tweet was marked as adversarial by the algorithm. From the top, the rows represent $T_{active,clean}$, $T_{adv,clean}$, $T_{active,adv}$, and $T_{active,adv}$.

The figure indicates the differences between the two groups, as well as the performance of the 0.7-adversarial threshold method. First, the chart indicates that 0.7-adversarial tweets from both groups (bottom two rows) are indeed often labeled as adversarial, with over 75% of the tweets in either row marked as "uncivil" by both raters, and close to 100% marked as "uncivil" by at least one rater. The trend is opposite for the top two rows, showing that, while some tweets may be missed by the algorithm, it is not likely to affect the classification of adversarial users — the algorithm may miss more of their tweets compared to other users' adversarial tweets. Taken together, we conclude that the 0.7-adversarial threshold method is an effective means to identify highly adversarial users. In

the following discussion, we take a deeper dive into a detailed analysis on the diversity of tweets posted by these users.

We now compare the type of contents in our sample of tweets by adversarial users ($T_{adv,adv} \bigcup T_{adv,clean}$) and our sample of tweets by active users ($T_{active,adv} \bigcup T_{active,clean}$) in Figure 5. For each tweet, we take the union of labels assigned by raters. We use χ^2 tests to compare the two groups, running tests separately for each category as the categories are not mutually exclusive; the *p*-values we report on are significant even after correcting for multiple tests. Note that similar to the above analysis in Figure 4, our results likely underestimate the magnitude of the difference between the groups, as the highly active sample is drawn from a set that also contains the highly-adversarial users as well as other users who are adversarial but did not make the 11-tweet threshold for U_{adv}. Moreover, in this analysis we use a *balanced sample* of tweets of each kind (clean or adversarial) made by each group of users. In other words, we are not just comparing the groups' overall activities, we are directly comparing "clean" and "adversarial" tweets made by each group, showing the differences in types even within these categories.

Figure 5a shows the manually-labeled categories assigned to the users' tweets. It is clear that the users in U_{adv} are more likely to engage in personal attacks against candidates (leftmost bar, $\chi^2(1, N = 690) = 6.25$, $p < 0.05$). On the other hand, Figure 5b shows that U_{active} users are more likely to engage in supportive interactions with candidates (of their own partisan affiliation, naturally). Particularly they express support more frequently to the candidate themselves ($\chi^2(1, N = 690) = 6.31$, $p < 0.05$). Finally, Figure 5c shows that when considering all tweets with agreeing labels, highly active users were more likely to respond in a manner that is relevant to the candidate tweet that they are replying to ($\chi^2(1, N = 492) = 6.95$, $p < 0.01$).

We also compared the sets of candidates that U_{adv} and U_{active} replied to, as well as the sets these groups posted adversarial replies to. Both groups seem to focus on the most "visible" candidates. There were no significant differences in the set of candidates being replied to between the two groups.

In summary, the results presented in Figure 5 expose different types of adversarial engagements and interactions performed by the highly-adversarial users. In particular, adversarial users are more likely to engage in personal attacks and are *less* likely to engage in other supportive interactions or respond with relevant content with candidate tweets. In other words, the difference in motivations of the adversarial users is manifested in all their behaviors.

Analysis of user profiles. Table 2 shows indications that the profiles of highly adversarial users' accounts may differ from that of other groups — specifically, the presence of a partisan hashtag in the user profile has significant positive association with the amount of adversarial interaction with political candidates. To understand and elaborate on this trend we performed a qualitative coding study of Twitter user profiles, to examine how partisanship is represented differently in adversarial users'
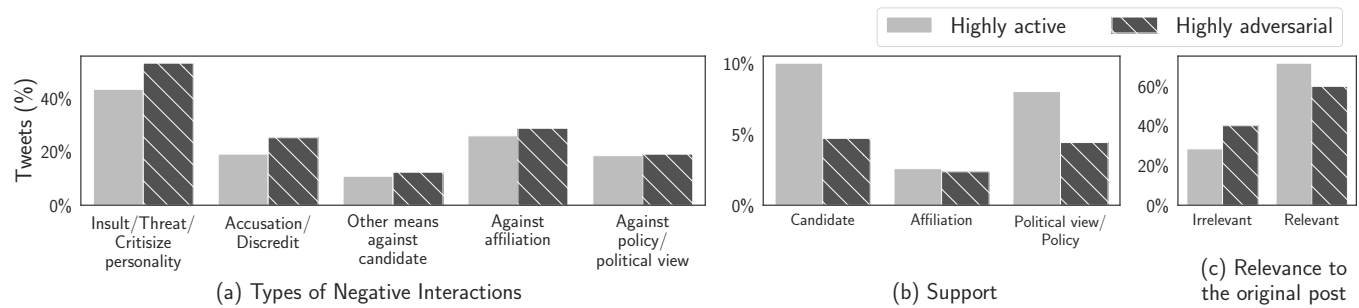
(a) Types of Negative Interactions

(b) Support

(c) Relevance to the original post

Figure 5: Highly adversarial users exhibit different activity patterns in terms of **(a)** types of negative interactions and **(b)** types of supportive interactions they engaged in, as well as **(c)** reply relevance to original posts compared to highly active users.
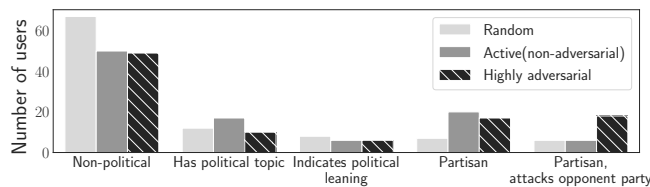

Figure 6: Highly adversarial users are more likely to include attacks on the opposite party in their user profiles.

profiles. In addition, we examined whether these profiles are more likely to appear anonymous.

In particular, we randomly selected 100 user profiles from each of the following three groups: (1) *random users*: a set of users chosen at random (denoted U_r); (2) *highly adversarial*: users who have posted at least 11 adversarial tweets (U_{adv} as defined above); and (3) *active non-adversarial*: a subset from *highly active* users, who have posted at least 11 tweets and never posted any adversarial tweet ($U_{active,clean}$, 4,867 users in total). The last group is different than U_{active} in the previous discussion from Figure 4 in that it specifically excludes users posting adversarial content. We developed a coding scheme in a sequence of coding sessions, where two of the authors coded a subset of user profiles, then discussed their codes to focus on the key differences between accounts in respect to this task. Our coding scheme asks coders to rate user profiles as:

- *Non-political*: A profile without political topics or empty.
- *Mentions political topic*: User profiles that mention interest in political topics, but do not suggest political leaning.
- *Shows political leaning*: User profiles that mention political views and opinions ("liberal", "Blue Lives Matter") that do not explicitly include partisan support.
- *Partisan*: User profiles explicitly expressing partisan support ("#MAGA", "#BlueWave").
- *Partisan attack of opponents*: User profiles attacking the political opponent, their party or leadership ("Anti Trump").

Two annotators independently coded each profile of our sample of 300 accounts. The rating was based on the user profile details including Twitter handle, profile description, location and name. Disagreements were resolved by a third annotator. One of the user profiles was excluded from the data as no consensus was reached.

Figure 6 shows the distribution of each category of user profiles for different groups. The histogram shows the categories on the X-axis, and the number of users of each group whose profile is in each category (Y-axis). The three groups are significantly different ($\chi^2(8, N = 299) = 21.99$, $p < 0.005$). For example, the random users were more likely (lightest column in the leftmost group) to have non-political accounts, though a majority of accounts across all the three groups were in that category. On the other hand, the active and adversarial users were more likely to show partisan support. However these differences are not significant in a post-hoc analysis. More significantly, highly adversarial users were more likely to express their opposition to the opposing party in their user profile (the rightmost dark column, $\chi^2(2, N = 299) = 10.56$, $p < 0.01$).

We also explored whether highly-adversarial users tend to use accounts that appear anonymous (i.e. without clear identifying information) more frequently than the other two groups. To this end, we coded every profile as "presenting as anonymous" or not. In other words, based on the profile handle, description, and published name, we coded as "anonymous" those profiles who cannot be connected to a specific individual (note that we naturally cannot detect users who offer identifying details that are not correct). While 48 out of the 100 random users were labeled as anonymous, the active and the high-adversarial groups had 63 out of 100 and 64 out of 100 anonymous profiles, respectively. The difference between these groups was statistically significant ($\chi^2(2, N = 300) = 6.6103$, $p < 0.05$).

DISCUSSION

Our results provide some reflection on earlier findings on adversarial behaviors in social media. One hypothesis from the literature is that arbitrary users can turn adversarial given an appropriately triggering context [11]. Our results do not refute (or support) this hypothesis, but we do see an alternative pattern of behavior: users that seem to seek out opportunities for adversarial interactions. Two aspects of our analysis can be interpreted as supporting this viewpoint. First, highly adversarial users reply more in general to candidates from the party the user opposes, suggesting that they are actively seeking out these opportunities for cross-party communications. Second, highly adversarial users are significantly more likely to have profiles that include partisan attacks, suggesting an adversarial disposition independent of any specific conversation.

Future research on improving online conversations can build upon our observations. For example, our findings from both qualitative and quantitative analyses show that users with higher adversarial activity level engage in fewer supportive interactions with political candidates. One possible direction as a future study is to further understand if encouraging supportive interactions would decrease user participation in adversarial behaviors. In addition, our analysis shows that users with lower activity level have a higher participation in adversarial behaviors as compared to their general posting activities. While prior research often focus on how to encourage user engagement from inactive users [48,58], it is as important to understand how to encourage *high-quality* engagement.

Our findings offer design implications for tools assisting human-in-the-loop content moderation in social media. For example, our results show that highly adversarial users are more likely to attack the opponent party in their user profiles, suggesting that tools can be developed to surface users who have expressed negativity explicitly in their profiles. This would enable moderators to prioritize reviewing content generated by these users. Further, social media platforms could take more granular measures to discourage users from seeking out conflicts. For example, down-ranking toxic interactions with public figures can both limit public exposure to the adversarial behaviors and create more friction for conflict-seeking users.

Finally, we urge platforms to take stronger actions to help political figures against adversarial interactions on social media, because of the ramifications for our democracy and society. In line with prior findings on social media participation [47, 53], "repeat" offenders contributed a large proportion of the adversarial contents. These users might be good targets for punitive actions by the platform, such as banning. Yet in our measurements we saw that only 15% of the highly adversarial users were later banned by Twitter.

Limitations. Our measurement study naturally suffers from some limitations. First, our study, being observational, cannot speak to the causal relationships between the variable we studied. Second, our focus was on the U.S. political system, and in particular the elections for the U.S. House of Representatives; the results may not generalize to other settings. Likewise, our findings may also be skewed by the specific topics which were in the news during the data collection period. For example, the confirmation hearings for U.S. Supreme Court justice Brett Kavanaugh were at the center of attention in that time period, and may have triggered more or less adversarial interactions than would be observed at other periods of time. In fact, as we pointed out, previous studies on characterizing adversarial behaviors under various scenarios on Twitter [9, 10, 23, 54] reported differing trends on certain features in term of correlations with adversarial activities. For example, [10] shows that the number of followers a user has is positively correlated with adversarial behaviors while [9,54] and our study suggest the opposite trend. This observation might suggest that adversarial behaviors are contributed by different social media populations and models need to be developed tailored to the context. Therefore these results are not ready to be applied directly on fully automated detection.

Third, our qualitative analysis is based on answers from raters who exhibit liberal leaning. Recent research [37] has shown that although people with different political leanings disagree on the severity of uncivil tweets, they agree when distinguishing uncivil tweets from civil ones. Nevertheless, it is possible that a group of coders with more diverse political backgrounds could bring a different perspective to the analysis. Fourth, when annotating tweets, coders were exposed to the political affiliation of the politicians. Although this design was intentional in order to improve annotation quality, as many tweets are hard to evaluate without this context, we acknowledge that this may introduce bias into our analysis.

Finally, as previous works have pointed out [15, 33], harassment detection based on machine learning tends to focus on usage of offensive language, while having a high false negative rate when classifying subtle adversarial content. Further, the tool used in our analysis, Perspective API, does not include Twitter in its training data. While we validated the precision of our method in this and earlier work [33], our method may nevertheless misses or ignores uncivil content not being flagged and underestimate the amount of adversarial activities. In fact, we showed in our qualitative analysis that our approach is more likely to *underestimate* the amount of adversarial activities produced by highly adversarial users, which has minimal impact for our results.

CONCLUSIONS

We studied users who engage in adversarial activities against political candidates, with a dataset collected during the run-up to the 2018 U.S. midterm elections. Using existing machine learning tools, we found that while a significant number of adversarial replies are made by those who tweeted to candidates very few times over the data collection period, a small minority of active users is responsible for an outsized portion of adversarial interactions. Using a combined approach with both quantitative and qualitative analysis that used features specific to the context of U.S. politics, we showed that for moderately active users, those who are more adversarial are both less central in the social graph consisting of users who interact with candidates, and pay less attention to candidates of their own partisan affiliation. When compared to highly active users, highly adversarial users tend to focus their activities more on negative interactions with candidates and less on supportive ones. Our results also indicate that highly-adversarial users are more likely to express negativity towards political opponents in their user profiles. With these findings we hope to inform future research in this area and encourage platform efforts in protecting political figures from adversarial interactions.

ACKNOWLEDGEMENT

REFERENCES

[1] Susan J Adams, Tracey E Hazelwood, Nancy L Pitre, Terry E Bedard, and Suzette D Landry. 2009. Harassment of members of parliament and the legislative assemblies in Canada by individuals believed to be mentally disordered. *The Journal of Forensic Psychiatry & Psychology* 20, 6 (2009), 801–814.

[2] Pushkal Agarwal, Nishanth Sastry, and Edward Wood. 2019. Tweeting MPs: Digital Engagement between Citizens and Members of Parliament in the UK. In *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media*. AAAI Press.

[3] Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. 2012. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *Proceedings of the Sixth International AAAI Conference on Web and Social Media*. AAAI Press.

[4] Amnesty International. 2018. TROLL PATROL FINDINGS: Using Crowdsourcing, Data Science & Machine Learning to Measure Violence and Abuse against Women on Twitter. `https://decoders.amnesty.org/projects/troll-patrol/findings#inf_12`. (2018). [Online; accessed 21-Dec-2018].

[5] Amnesty International UK. 2018. Black and Asian women MPs abused more online. `https://www.amnesty.org.uk/online-violence-women-mps`. (2018). [Online; accessed 5-Jan-2019].

[6] Ahmer Arif, Leo Graiden Stewart, and Kate Starbird. 2018. Acting the part: Examining information operations within #BlackLivesMatter discourse. *Proceedings of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing* (2018).

[7] Pablo Barberá. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis* 23, 1 (2015), 76–91.

[8] Andrea Caumont. 2013. Twitter Reaction to Events Often at Odds with Overall Public Opinion. `https://www.pewresearch.org/2013/03/04/twitter-reaction-to-events-often-at-odds-with-overall-public-opinion/`. (2013). [Online; accessed 3-Apr-2019].

[9] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017a. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on Web Science Conference*. ACM.

[10] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017b. Measuring# gamergate: A tale of hate, sexism, and bullying. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee.

[11] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM.

[12] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial behavior in online discussion communities. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*. AAAI Press.

[13] Lucie Corcoran, Conor Guckin, and Garry Prentice. 2015. Cyberbullying or cyber aggression?: A review of existing definitions of cyber-based peer-to-peer aggression. *Societies* 5, 2 (2015), 245–255.

[14] Kate Crawford. 2009. Following you: Disciplines of listening in social media. *Continuum* 23, 4 (2009), 525–535.

[15] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*. AAAI Press.

[16] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flamini, and Filippo Menczer. 2016. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee.

[17] Park E Dietz, Daryl B Matthews, Daniel Allen Martell, Tracy M Stewart, Debra R Hrouda, and Janet Warren. 1991. Threatening and otherwise inappropriate letters to members of the United States Congress. *Journal of Forensic Science* 36, 5 (1991), 1445–1468.

[18] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2, 3 (2012), 18.

[19] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *Proceeding of the Fifth International AAAI Conference on Web and Social Media*. AAAI Press.

[20] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*. ACM.

[21] Jack Dorsey. 2019. `https://twitter.com/jack/status/919028949434241024`. (2019). [Online; accessed 3-Apr-2019].

[22] Maeve Duggan. 2017. Online harassment in focus: Most recent experience. `https://www.pewinternet.org/2017/07/11/online-harassment-in-focus-most-recent-experience/`. (2017). [Online; accessed 3-Apr-2019].

[23] Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018. Peer to peer hate: Hate speech instigators and their targets. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*. AAAI Press.

[24] Susanna Every-Palmer, Justin Barry-Walsh, and Michele Pathé. 2015. Harassment, stalking, threats and attacks targeting New Zealand politicians: A mental health issue. *Australian & New Zealand Journal of Psychiatry* 49, 7 (2015), 634–641.

[25] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* 59, 7 (2016), 96–104.

[26] R Stuart Geiger. 2016. Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society* 19, 6 (2016), 787–803.

[27] Andrew Gelman. 2008. Scaling regression inputs by dividing by two standard deviations. *Statistics in medicine* 27, 15 (2008), 2865–2873.

[28] Genevieve Gorrell, Mark Greenwood, Ian Roberts, Diana Maynard, and Kalina Bontcheva. 2018a. Online Abuse of UK MPs in 2015 and 2017: Perpetrators, Targets, and Topics. *arXiv preprint arXiv:1804.01498* (2018).

[29] Genevieve Gorrell, Mark A Greenwood, Ian Roberts, Diana Maynard, and Kalina Bontcheva. 2018b. Twits, Twats and Twaddle: Trends in Online Abuse towards UK Politicians. In *Proceedings Of the Twelfth International AAAI Conference on Web and Social Media*. AAAI Press.

[30] Will J Grant, Brenda Moon, and Janie Busby Grant. 2010. Digital dialogue? Australian politicians' use of the social network tool Twitter. *Australian Journal of Political Science* 45, 4 (2010), 579–604.

[31] Dorothy Wunmi Grigg. 2010. Cyber-aggression: Definition and concept of cyberbullying. *Journal of Psychologists and Counsellors in Schools* 20, 2 (2010), 143–156.

[32] Alex Hern. 2015. Twitter CEO: We suck at dealing with trolls and abuse. `https://www.theguardian.com/technology/2015/feb/05/twitter-ceo-we-suck-dealing-with-trolls-abuse`. (2015). [Online; accessed 3-Apr-2019].

[33] Yiqing Hua, Thomas Ristenpart, and Mor Naaman. 2020. Towards measuring adversarial Twitter interactions against candidates in the US midterm elections. (2020).

[34] Jane Im, Eshwar Chandrasekharan, Jackson Sargent, Paige Lighthammer, Taylor Denby, Ankit Bhargava, Libby Hemphill, David Jurgens, and Eric Gilbert. 2019. Still out there: Modeling and Identifying Russian Troll Accounts on Twitter. *arXiv preprint arXiv:1901.11162* (2019).

[35] David V James, Frank R Farnham, Seema Sukhwal, Katherine Jones, Josephine Carlisle, and Sara Henley. 2016. Aggressive/intrusive behaviours, harassment and stalking of members of the United Kingdom parliament: a prevalence study and cross-national comparison. *The Journal of Forensic Psychiatry & Psychology* 27, 2 (2016), 177–197.

[36] Jigsaw. 2018. Perspective API. `https://www.perspectiveapi.com/`. (2018). [Online; accessed 3-Jan-2019].

[37] Kenneth Joseph, Briony Swire-Thompson, Hannah Masuga, Matthew A Baum, and David Lazer. 2019. Polarized, Together: Comparing Partisan Support for Trump's Tweets Using Survey and Platform-Based Measures. In *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media*. AAAI Press.

[38] Anders O Larsson and Bente Kalsnes. 2014. 'Of course we are on Facebook': Use and non-use of social media among Swedish and Norwegian politicians. *European Journal of Communication* 29, 6 (2014), 653–667.

[39] Kaitlin Mahar, Amy X Zhang, and David Karger. 2018. Squadbox: A tool to combat email harassment using friendsourced moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM.

[40] Suman Kalyan Maity, Aishik Chakraborty, Pawan Goyal, and Animesh Mukherjee. 2018. Opinion Conflicts: An Effective Route to Detect Incivility in Twitter. *Proceedings of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing* (2018).

[41] J Matias, Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jaclyn Friedman, and Charlie DeTar. 2015. Reporting, reviewing, and responding to harassment on Twitter. *Available at SSRN 2602018* (2015).

[42] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 27, 1 (2001), 415–444.

[43] Carlos Monje Jr. 2019. 2018 US midterm elections review. `https://blog.twitter.com/en_us/topics/company/2019/18_midterm_review.html`. (2019). [Online; accessed 20-Aug-2019].

[44] Michael J Moore, Tadashi Nakano, Akihiro Enomoto, and Tatsuya Suda. 2012. Anonymity and roles associated with aggressive posts in an online forum. *Computers in Human Behavior* 28, 3 (2012), 861–867.

[45] New York Times. 2018. For Female Candidates, Harassment and Threats Come Every Day. `https://www.nytimes.com/2018/08/24/us/politics/women-harassment-elections.html?action=click&module=Top%20Stories&pgtype=Homepage`. (2018). [Online; accessed 5-Jan-2019].

[46] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee.

[47] Blair Nonnecke and Jenny Preece. 2000. Lurker demographics: Counting the silent. In *Proceedings of the 2000 CHI Conference on Human Factors in Computing Systems*. ACM.

[48] Blair Nonnecke and Jenny Preece. 2001. Why lurkers lurk. *Proceedings of the Americas Conference on Information Systems* (2001).

[49] Robert M O'brien. 2007. A caution regarding rules of thumb for variance inflation factors. *Quality & quantity* 41, 5 (2007), 673–690.

[50] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank citation ranking: Bringing order to the web*. Technical Report. Stanford InfoLab. `http://ilpubs.stanford.edu:8090/422/`

[51] John H Parmelee and Shannon L Bichard. 2011. *Politics and the Twitter revolution: How tweets influence the relationship between political leaders and the public*. Lexington Books.

[52] Michele Pathé, Jane Phillips, Elke Perdacher, and Ed Heffernan. 2014. The harassment of Queensland Members of Parliament: A mental health concern. *Psychiatry, Psychology and Law* 21, 4 (2014), 577–584.

[53] Jenny Preece, Blair Nonnecke, and Dorine Andrews. 2004. The top five reasons for lurking: improving community experiences for everyone. *Computers in human behavior* 20, 2 (2004), 201–223.

[54] Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on twitter. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*. AAAI Press.

[55] Daniel M Romero, Chenhao Tan, and Johan Ugander. 2013. On the Interplay between Social and Topical Structure.. In *Proceedings of the Seventh International AAAI Conference on Web and Social Media*. AAAI Press.

[56] Alexandra Siegel, Evgenii Nikitin, Pablo Barberá, Joanna Sterling, Bethany Pullen, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. 2019. Trumping Hate on Twitter? Online Hate in the 2016 US Election and its Aftermath. `https://smappnyu.org/wp-content/uploads/2019/04/US_Election_Hate_Speech_2019_03_website.pdf`, (2019).

[57] Vivek K Singh, Marie L Radford, Qianjia Huang, and Susan Furrer. 2017. They basically like destroyed the school one day: On Newer App Features and Cyberbullying in Schools. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM.

[58] Steven JJ Tedjamulia, Douglas L Dean, David R Olsen, and Conan C Albrecht. 2005. Motivating content contributions to online communities: Toward a more comprehensive theory. In *Proceedings of the 38th annual Hawaii international conference on system sciences*. IEEE.

[59] Yannis Theocharis, Pablo Barberá, Zoltán Fazekas, Sebastian Adrian Popa, and Olivier Parnet. 2016. A bad workman blames his tweets: the consequences of citizens' uncivil Twitter use when interacting with party candidates. *Journal of communication* 66, 6 (2016), 1007–1031.

[60] Twitter. 2019a. About replies and mentions. `https://help.twitter.com/en/using-twitter/mentions-and-replies`. (2019). [Online; accessed 1-April-2019].

[61] Twitter. 2019b. About suspended accounts. `https://help.twitter.com/en/managing-your-account/suspended-twitter-accounts`. (2019). [Online; accessed 3-Apr-2019].

[62] Twitter. 2019c. Twitter Standard API. `https://developer.twitter.com/en/docs/tweets/filter-realtime/overview`. (2019). [Online; accessed 15-Jan-2019].

[63] Twitter. 2019d. Verified account FAQs. `https://help.twitter.com/en/managing-your-account/twitter-verified-accounts`. (2019). [Online; accessed 1-April-2019].

[64] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. 2017. Identifying Women's Experiences With and Strategies for Mitigating Negative Effects of Online Harassment. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM.

[65] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee.

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