Research Article

Research and Politics
January- March 2019: 1–7

DOI: 10.1177/2053168018816228
journals.sagepub.com/home/rap
SAGE

Politicians in the line of fire: Incivility and the treatment of women on social media

Ludovic Rheault, Erica Rayment and Andreea Musulan

Abstract

A seemingly inescapable feature of the digital age is that people choosing to devote their lives to politics must now be ready to face a barrage of insults and disparaging comments targeted at them through social media. This article represents an effort to document this phenomenon systematically. We implement machine learning models to predict the incivility of about 2.2 m messages addressed to Canadian politicians and US Senators on Twitter. Specifically, we test whether women in politics are more heavily targeted by online incivility, as recent media reports suggested. Our estimates indicate that roughly 15% of public messages sent to Senators can be categorized as uncivil, whereas the proportion is about four points lower in Canada. We find evidence that women are more heavily targeted by uncivil messages than men, although only among highly visible politicians.

Keywords

Gender, machine learning, political communication, political incivility, social media, Twitter

Politicians in the line of fire: Incivility and the treatment of women on social media

To what extent are politicians, particularly women, subjected to incivility on social media? The question not only touches on unresolved puzzles in the literature on women in politics, but is also relevant to understanding how the digital age is transforming the practice of democracy. Recent evidence from civil society suggests that women in politics are targeted disproportionately by uncivil comments online. Reports published in traditional news outlets indicate that female leaders in the UK, the USA, and Australia are on the receiving end of a particularly abusive form of harassment through social media (Bowles, 2016; Carter and Sneesby, 2017; Hunt et al., 2016; Saner, 2016). For instance, an analysis performed by a data analytics firm revealed that Hillary Clinton received nearly twice as many tweets containing abusive words as her opponent Bernie Sanders during the 2016 Democratic primaries, whereas in Australia, Julia Gillard was also disproportionately targeted by online incivility compared to her Australian Labor Party leadership rival Kevin Rudd (Hunt et al., 2016). In Canada, news reports have highlighted a stream of hate-fueled comments directed at elected women through social media (Crawley, 2017; Huncar, 2015; Rushowy, 2017; Sturino and O'Brien, 2017; The Canadian Press, 2017).

Despite recent news coverage and analyses conducted by research firms, we still know very little about the phenomenon of incivility directed at politicians online. Our study tackles the question by examining a collection of over 2 m messages directly addressed to politicians on the Twitter platform. We introduce machine learning models trained to predict uncivil messages with high levels of accuracy. We then provide estimates of the levels of incivility directed at public officials, and test hypotheses about the gendered distribution of uncivil comments. Our main contribution is to show that gender effects do exist, but are conditional on levels of public recognition.

Are women politicians treated differently?

Our study relates to a rich literature on public perceptions of women in politics, and the question of whether or not women are evaluated more harshly than their male counterparts—a topic for which there is no clear consensus among scholars (see Brooks, 2013; Dolan, 2004). The expectation

University of Toronto, Canada

Corresponding author:
Ludovic Rheault, Department of Political Science, University of Toronto, 100 St. George Street, Toronto, ON M5S 3G3, Canada.
Email: ludovic.rheault@utoronto.ca

that women who occupy leadership roles in politics will trigger negative reactions from the public is grounded in influential models in social psychology. Gender role theory (Eagly, 1987; Shimanoff, 2009) posits the existence of perceived appropriate norms of behavior and roles associated with each gender that exert influence on individuals' career decisions as well as on public perceptions of these decisions. Eagly et al. (1992) suggest that women who occupy leadership positions are at odds with stereotypically female characteristics and, as a result, are often devalued. A number of empirical studies drawing on this theory suggest that women in traditionally male positions face resistance (Eagly and Karau, 2002; Puwar, 2004; Rudman and Phelan, 2008). In this sense, incivility toward women who participate in politics can be viewed "as a form of gender role enforcement" (Krook and Restrepo, 2016: 466).

Consistent with the existence of gender stereotypes, a body of research suggests that women and men in politics are judged according to different standards. In particular, several studies find that voters perceive women candidates as warm and generous—stereotypically feminine traits associated with a lower level of political competence and suitability for elective office (Banwart, 2010; Higgle et al., 1997; Huddy and Terkildsen, 1993). Researchers also reported evidence of female candidates receiving less media coverage or coverage that undermines the viability of their candidacy (Dunaway et al., 2013; Gidengil and Everitt, 2003; Kahn and Goldenberg, 1991; Lawrence and Rose, 2009). Although perceptions of women politicians on social media have not received nearly as much attention, a recent study of Twitter engagement in senatorial and gubernatorial campaigns found gender differences in terms of the focus of online discussions (McGregor and Mourao, 2016).[1]

On the other hand, a recent stream of literature challenged these findings, concluding that the effect of gender stereotypes on public perceptions of women politicians has waned, if not disappeared altogether (Hayes and Lawless, 2015). For instance, Dolan (2014) observes that gender stereotypes have no impact on the vote for female candidates once the role of party affiliation is accounted for, echoing the optimistic conclusions reported by Brooks (2013). Similarly, some studies suggest that the coverage of women in the media has increased and become more balanced, if not positive (Jalalzai, 2006; Smith, 1997). In their study of congressional elections, Hayes and Lawless (2016) find no significant differences in public evaluations or in the media coverage of female political candidates, explaining the disappearance of double standards in part by claiming that women are no longer a novelty in US politics.

Although women in politics may no longer be a novelty in the strictest sense, nonetheless, they remain a significant minority, particularly in peak positions of power. Few studies have specifically considered the role of status—the visibility or public profile of a politician—as a condition for the occurrence of negativity toward women. Yet, such a variable appears essential to assess the theoretical claim that women in politics face reprisals because of gender stereotypes. If women who occupy influential positions associated with masculine traits are perceived as violators of traditional gender roles, reactions to this transgression should be stronger when they achieve public recognition, because their role incongruity is then more visible. As a result, it stands to reason that the backlash against women politicians on social media, if it exists, should intensify as they gain in status and visibility.

Based on expectations from gender role theory and the discussion above, we consider two hypotheses: (1) female politicians receive more uncivil messages than men; and (2) female politicians are more targeted by incivility the higher their status. We operationalize the status of each politician using an indicator of visibility already integrated into the Twitter platform and capturing network ties within the online community: the number of followers. This indicator not only facilitates measurement and replication, but also allows us to distinguish between politicians holding the same position yet enjoying different levels of public visibility. Finally, our analysis also accounts for possible confounder variables, such as party and ethnicity, identified in the literature.

Data collection

We selected two samples of public officials from Canada and the USA. The Canadian sample consists of cabinet members of the federal government and of the 10 provinces and is, therefore, comprised of the most influential elected officials in the country. This sample contains substantial variation in terms of politician attributes, such as gender and visibility, which are independent variables of interest. At the time of data collection, 3 of the 10 provincial Premiers and half of the federal cabinet ministers were women. The Canadian sample contains 195 politicians with an active account on the Twitter platform, 37% of whom are female. We also replicate results with the 100 US Senators. The Senate has interesting properties in that it comprises high-profile politicians distributed equally across states. At the time of data collection, 21 of the 100 Senators were women.

We collected our corpus from the Twitter microblogging platform, on which messages are called statuses or tweets, with the streaming API. The collection took place over a period of one month for each country between April and July 2017. The API allows developers to collect up to 1% of all public tweets posted at any given time, using a list of filters containing up to 400 keywords. Because we use specific search criteria, our data comprise the near totality of messages meeting these criteria, rather than a sample. To collect statuses addressed to politicians, we retrieved the official Twitter *handles* (user names preceded by the "@" symbol) of each politician on the site, which we utilized as

Table 1. Inferring the level of incivility by gender.

	USA			Canada		
	Women	Men	Total	Women	Men	Total
Fitted proportions	12.95%	14.54%	14.13%	8.55%	11.66%	10.69%
Corpus size	530,663	1,545,175	2,075,838	53,195	116,919	170,114

Proportions predicted with a balanced bagging model using 50 replications of support vector machine estimators.

filters for the stream. The presence of a handle inside the raw text of the document indicates that the message was addressed to a specific politician. Hence, our corpus does not merely comprise messages *about* a politician, it contains precisely those messages addressed to them. Public messages addressed to a Twitter user should not be confused with Direct Messages—private messages that can only be sent to someone who "follows" a user on the site. In the language of the platform, public messages addressed to a user with the handle could be original tweets, replies to a thread in which the politician was marked as a recipient, or quotes added to a message in which the politician was directly referred to. An Online Appendix provides readers with specific details on data collection and preprocessing. In particular, we removed duplicates and shared messages (*retweets*) from our corpus.

Predicting uncivil messages using textual data

Our analysis required a very large corpus to ensure a sufficient quantity of messages sent to each politician and, thus, reliable generalizations. Therefore, we relied on supervised machine learning models to classify the 2.2 million tweets in our main corpus as either civil or uncivil. Supervised learning is widespread in applications involving textual analysis, and expands on statistical techniques commonly used in the social sciences (see Hastie et al., 2009). The predictive models use training data—a set of examples coded by humans—to predict incivility in the full corpus. To maximize the accuracy of our methods, we created training samples by randomly selecting 10,000 tweets from the full corpus (5000 for Canada and 5000 for the USA). Because the training set is a random sample from the population of interest, covering the entire period, we avoided several limitations associated with supervised learning pointed out in the literature (Hand, 2006; Hopkins and King, 2010). In particular, we increased the confidence that the distribution of words in the full corpus would be similar to the distribution in the training data.

We relied upon the FigureEight platform (formerly CrowdFlower) to annotate each tweet in the training set. This crowd-sourcing website allows researchers to hire workers for coding text documents, and recent publications have documented its accuracy for applications in political science and for the detection of latent categories (Benoit et al., 2016; Lind et al., 2017). We provided workers with detailed instructions about the coding scheme. An uncivil text was defined as a tweet containing at least one of the following elements: (1) swear words; (2) vulgarities; (3) insults; (4) threats; (5) personal attacks on someone's private life; or (6) attacks targeted at groups (hate speech). To ensure a high level of quality, we retained judgments only from annotators who scored 85% or more on a test set of over 50 questions for which we provided the ground truth.[2] The average pairwise agreement across all text documents is 89.9% in the Canadian sample, and 86.5% in the US sample. With regard to the US training set, 15.4% of statuses were coded as uncivil, compared to 10.6% for the Canadian sample.

We assessed model performance in relation to predicting the incivility of new, unseen tweets using tenfold cross-validation. The selected models are support vector classifiers fitted using 50 bootstrap aggregating (bagging) replications (Breiman, 1996), which reached an accuracy rate of 91.7% for Canada and 89.3% for the USA during the validation stage. Put simply, bagging consists of running the predictions multiple times after randomly resampling the training examples, and choosing the class (civil/uncivil) predicted the most often by the models. This method has been shown to improve the quality of prediction and reduce sensitivity to outliers (Bauer and Kohavi, 1999). Our models rely on a total of 2002 features: unigrams and bigrams (single words and sequences of two words); an indicator of semantic similarity with a list of common insults; and a measure of sentiment. The Online Appendix presents a full description of the linguistic features used in the models and a detailed assessment of accuracy statistics.

Empirical findings

We begin by reporting basic estimates of the proportion of uncivil messages in the full corpus, broken down by the gender of the politician at whom the messages were directed (Table 1). These proportions are computed by aggregating tweets predicted as uncivil using the bagging classifier. Given the high accuracy of the models and because we were able to compare their results with random subsamples of human-coded tweets, we are confident in the reliability of the overall proportions we report. The proportion of

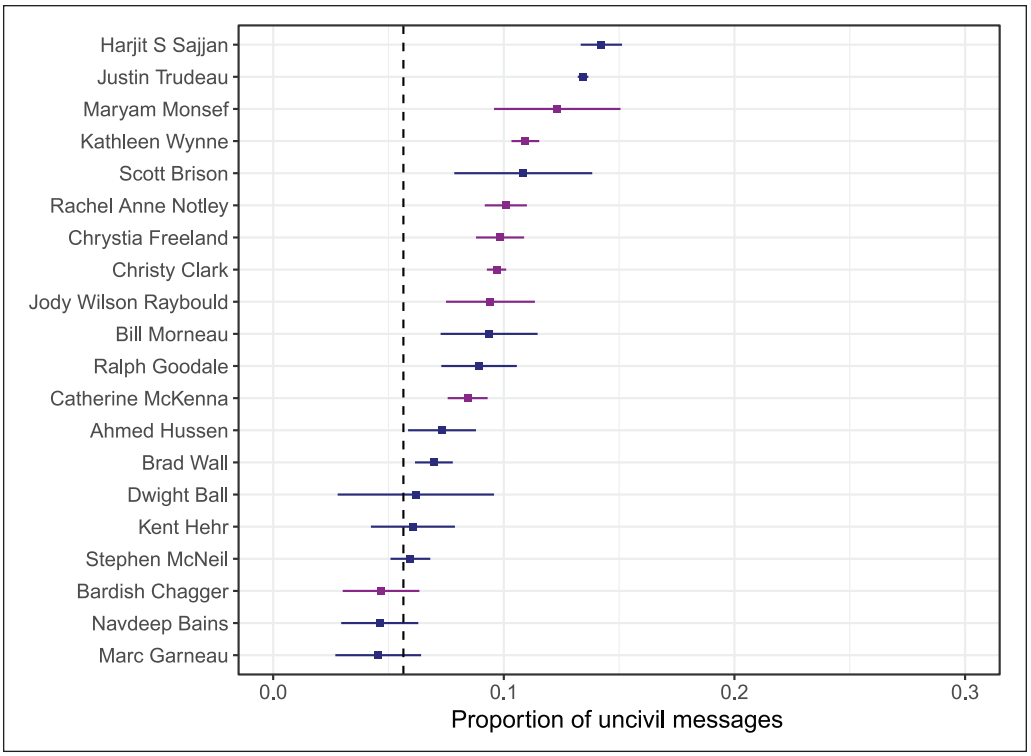


Figure 1. Canadian politicians most targeted by uncivil messages.
The vertical line indicates the average proportion of uncivil messages received by federal ministers and Premiers. We use a color code to distinguish between female and male politicians.

uncivil tweets addressed to politicians in the full sample is estimated at 10.69% in Canada and 14.13% in the USA, close to those observed in the two annotated samples, 10.6% and 15.4%, respectively. Without accounting for the visibility of politicians, the estimates in Table 1 run against the first of our hypotheses. The proportion of uncivil tweets directed at men is actually slightly higher than the proportion of uncivil tweets directed at women in both Canada and the USA. In Canada, for instance, approximately 8.6% of tweets targeted at female office holders were uncivil, compared to approximately 11.7% for men.

Because explanations of the negativity toward women in politics are often rooted in the challenge they pose to traditional gender roles, the likelihood of being targeted by uncivil remarks should be affected by the visibility women enjoy. To illustrate evidence in support of this phenomenon, Figure 1 reports the top 20 Canadian politicians most heavily targeted by uncivil remarks on Twitter. The figure aggregates the share of tweets classified as uncivil, along with a 95% confidence interval for sample proportions. For the comparisons to be more meaningful, we limited the ranking to federal politicians and provincial Premiers. The variety of cabinet positions helps to demonstrate the role of status and visibility in shaping the number of uncivil tweets a politician receives. For instance, Justin Trudeau, the current Prime Minister of Canada, received 85,153 of the tweets in our corpus, and he accounts for over 11,000 uncivil messages by himself. When looking beyond the Prime Minister, several politicians among the top targets of abusive messages are women, but they are women who occupy high-profile positions—the Premiers of Alberta, Ontario, and British Columbia are included in this list, along with federal cabinet ministers with high-profile portfolios such as foreign affairs, justice, democratic reform, and environment.

We examine the hypotheses more thoroughly with a multivariate analysis accounting for other attributes of the message recipients, for instance, party affiliation. Using the binary class variable of uncivil tweets created earlier, which equals one if a tweet is uncivil and zero otherwise, we fit logistic regression models explaining the probability of an uncivil tweet according to the attributes of the recipients. In addition to gender, we test an interaction effect between that variable and an indicator of visibility based on the log count of followers for each politician. Because the random components of the model are not independent, that is, each recipient is observed multiple times in the sample, we rely on robust standard errors clustered by politician. Tables 2 and 3 report the main results for Canada and the USA, respectively. An Online Appendix presents additional results and analysis.

Starting with the Canadian sample, as soon as we control for the visibility of the politician, the association between

Table 2. Incivility as a function of politician attributes (Canada).

	(1)	(2)	(3)	(4)
Gender (Female = 1)	-0.344*	0.166	-2.197**	-2.559***
	(0.159)	(0.126)	(0.795)	(0.776)
Log follower count		0.180***	0.166***	0.197***
		(0.027)	(0.026)	(0.022)
Gender × log follower count			0.206**	0.250***
			(0.064)	(0.068)
Visible minority				0.529**
				(0.175)
Party (Liberal = 1)				-0.164*
				(0.064)
Intercept	-2.026***	-4.534***	-4.334***	-4.706***
	(0.129)	(0.420)	(0.405)	(0.242)
Observations	170,114	170,114	170,114	170,114

Note: Dependent variable: 1 = uncivil tweet; 0 = otherwise. The models are logistic regressions with clustered standard errors on politicians. The last model includes fixed effects for federal level, hours of day, and days of the week.
*p <0.05; **p <0.01; ***p < 0.001.

Table 3. Incivility as a function of politician attributes (USA).

	(1)	(2)	(3)	(4)
Gender (Female = 1)	-0.134	-0.161	-1.992***	-3.247***
	(0.112)	(0.082)	(0.541)	(0.738)
Log follower count		0.097***	0.081***	0.062
		(0.025)	(0.023)	(0.071)
Gender × log follower count			0.136***	0.241***
			(0.040)	(0.058)
Visible minority				-0.122
				(0.118)
Party (Democrat = 1)				-0.106
				(0.100)
Intercept	-1.771***	-3.038***	-2.824***	-2.574***
	(0.061)	(0.316)	(0.306)	(0.736)
Observations	2,075,838	2,075,838	2,075,838	2,075,838

Note: Dependent variable: 1 = uncivil tweet; 0 = otherwise. The models are logistic regressions with clustered standard errors on Senators. The last model includes fixed effects for state, hours of day, and days of the week.
*p <0.05; **p <0.01; ***p <0.001.

being female and the probability of receiving an uncivil message becomes positive (although not statistically significant). We also find a significant interactive effect between visibility and gender: women receive more uncivil messages as their visibility increases. Panel (a) from Figure 2 depicts the change in predicted probabilities of receiving an uncivil message by gender, using two realistic sample values of low and high follower counts. Female officials turn out to be more likely to receive an uncivil tweet in both scenarios. Only for very low numbers of followers (roughly 30,000 and fewer) do we observe male politicians with a higher predicted probability of being targeted by hostile messages. In short, female politicians at the bottom of the political hierarchy fare well relative to men when it comes to online incivility. However, they appear to be more heavily targeted when they become more visible. When replicating the process with US data (Table 3), we find a similar interaction effect, suggesting that the moderating role of visibility may be generalizable. However, in the US case the trend reversal only occurs for out-of-sample values of the visibility variable, as shown in Figure 2(b). We explain the weaker results for the US case by the lack of high-profile female politicians in the Senate, which limits variation in the dataset. We replicated this analysis using aggregated data and alternative indicators of incivility, in particular the frequency of swear words from the LIWC 2015 dictionary (Tausczik and Pennebaker, 2010). We report these tests in the Online Appendix. The additional findings are consistent with those

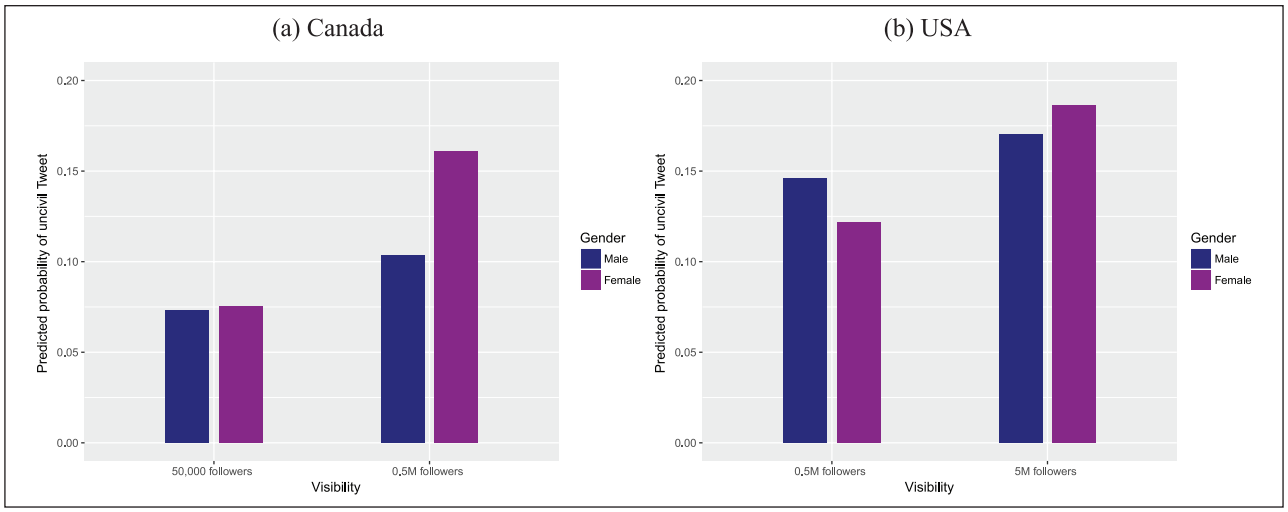


Figure 2. Predicted probability of uncivil tweet by popularity and gender.

reported here, and suggest that the interaction between gender and visibility is robust.

Conclusion

According to our estimates, close to 11% of messages addressed to Canadian politicians on social media can be categorized as uncivil, either because they rely upon explicit profanities or because they represent more fundamental and personal attacks. That proportion is higher when considering US Senators (about 15%). Our main objective was to test the claim of disproportionate levels of incivility toward women politicians on social media. In particular, we assessed whether women breaking the glass ceiling by achieving high levels of public recognition in politics are more often subjected to uncivil messages. Although the baseline rates of incivility are higher for male politicians, we find that the association between gender and the likelihood of being targeted is conditional on visibility: women who achieve a high status in politics are more likely to receive uncivil messages than their male counterparts. We find a similar interaction effect in both the Canadian and US samples, although the results are clearer in Canada, where the proportion of women in high-profile positions provides more observations to conduct such an analysis. This interaction with visibility may help to explain why some recent media reports mentioned in the introduction that focused on female leaders or candidates for leadership positions have found a disproportionate amount of abusive messages targeting them. More generally, the findings suggest that differences in status and visibility may be a relevant factor to consider in future research on women in politics.

Declaration of conflicting interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Social Sciences and Humanities Research Council of Canada [Grant Number 430-2017-0012].

Supplementary materials

The supplementary files are available at http://journals.sagepub.com/doi/suppl/10.1177/2053168018816228. The replication files are available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/B97TGX.

Notes

1. To our knowledge, very few (if any) academic studies focused on the gendered nature of political incivility expressed online, although recent studies have discussed more general trends in cyberbullying targeted at women on the web (Jane, 2014b, 2014a; Megarry, 2014; Vickery and Everbach, 2018).
2. We allowed coders to select an "unsure" category when they were uncertain; for simplicity, we recoded the unsure documents as uncivil as the class was seldom chosen by a majority of coders.

References

Banwart MC (2010) Gender and candidate communication: Effects of stereotypes in the 2008 Election. *American Behavioral Scientist* 54(3): 265–283.

Bauer E and Kohavi R (1999) An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning* 36(1–2): 105–139.

Benoit K, Conway D, Lauderdale BE, et al. (2016) Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review* 110(2): 278–295.

Bowles N (2016) Former lawmaker Wendy Davis: "Trolls want to diminish and sexualize you." *The Guardian*, 12 March. Available at: https://www.theguardian.com/culture/2016/mar/12/wendy-davis-sxsw-2016-texas-trolls-online-abuse (accessed 9 August 2017).

Breiman L (1996) Bagging predictors. *Machine Learning* 24(2): 123–140.

Brooks DJ (2013) *He Runs, She Runs: Why Gender Stereotypes Do Not Harm Women Candidates*. Princeton, NJ: Princeton University Press.

Carter A and Sneesby J (2017) Mistreatment of women MPs revealed. BBC News, 25 January. Available at: https://www.bbc.com/news/uk-politics-38736729 (accessed 9 August 2017).

Crawley M (2017) Premier Kathleen Wynne bombarded on social media by homophobic, sexist abuse. CBC News, *25 January*. Available at: https://www.cbc.ca/news/canada/toronto/kathleen-wynne-twitter-abuse-1.3949657 (accessed 9 August 2017).

Dolan K (2004) *Voting For Women: How The Public Evaluates Women Candidates*. Boulder, CO: Westview Press.

Dolan K (2014) Gender stereotypes, candidate evaluations, and voting for women candidates: What really matters? *Political Research Quarterly* 67(1): 96–107.

Dunaway J, Lawrence RG, Rose M, et al. (2013) Traits versus issues: How female candidates shape coverage of Senate and gubernatorial races. *Political Research Quarterly* 66(3): 715–726.

Eagly AH (1987) *Sex Differences in Social Behavior: A Social-Role Interpretation*. Hillsdale, MI: Lawrence Erlbaum.

Eagly AH and Karau SJ (2002) Role congruity theory of prejudice toward female leaders. *Psychological Review* 109(July): 573–598.

Eagly AH, Makhijani MG and Klonsky BG (1992) Gender and the evaluation of leaders: A meta-analysis. *Psychological Bulletin* 111(1): 3–22.

Gidengil E and Everitt J (2003) Talking tough: Gender and reported speech in campaign news coverage. *Political Communication* 20(3): 209–232.

Hand DJ (2006) Classifier technology and the illusion of progress. *Statistical Science* 21(1): 1–14.

Hastie T, Tibshirani R and Friedman J (2009) *The Elements of Statistical Learning*. Berlin: Springer.

Hayes D and Lawless JL (2015) A non-gendered lens? Media, voters, and female candidates in contemporary congressional elections. *Perspectives on Politics* 13(1): 95–118.

Hayes D and Lawless JL (2016) *Women on the Run: Gender, Media, and Political Campaigns in a Polarized Era*. New York: Cambridge University Press.

Higgle EDB, Miller PM, Shields TG, et al. (1997) Gender stereotypes and decision context in the evaluation of political candidates. *Women & Politics* 17(3): 69–88.

Hopkins DJ and King G (2010) A method of automated nonparametric content analysis for social science. *American Journal of Political Science* 54(1): 229–247.

Huddy L and Terkildsen N (1993) Gender stereotypes and the perception of male and female candidates. *American Journal of Political Science* 37(1): 119–147.

Huncar A (2015) Alberta female politicians targeted by hateful, sexist online attacks. CBC News, 20 October. Available at: https://www.cbc.ca/news/canada/edmonton/alberta-female-politicians-targeted-by-hateful-sexist-online-attacks-1.3281275 (accessed 9 August 2017).

Hunt E, Evershed N and Liu R. (2016) From Julia Gillard to Hillary Clinton: Online abuse of politicians around the world. *The Guardian*, 27 June. Available at: https://www.theguardian.com/technology/datablog/ng-interactive/2016/jun/27/from-julia-gillard-to-hillary-clinton-online-abuse-of-politicians-around-the-world (accessed 9 August 2017).

Jalalzai F (2006) Women candidates and the media: 1992–2000 elections. *Politics & Policy* 34(3): 606–633.

Jane EA (2014a) "Your a ugly, whorish slut." *Feminist Media Studies* 14(July): 531–546.

Jane EA (2014b) "Back to the kitchen, cunt": Speaking the unspeakable about online misogyny. *Continuum* 28(July): 558–570.

Kahn KF and Goldenberg EN (1991) Women candidates in the news: An examination of gender differences in U.S. Senate campaign coverage. *The Public Opinion Quarterly* 55(2): 180–199.

Krook M and Restrepo J (2016) Violence against women in politics. A defense of the concept. *Política y gobierno* 23(December): 459–490.

Lawrence RG and Rose M (2009) *Hillary Clinton's Race for the White House: Gender Politics and the Media on the Campaign Trail*. Boulder, CO: Lynne Rienner Publishers.

Lind F, Gruber M and Boomgaarden HG (2017) Content analysis by the crowd: Assessing the usability of crowdsourcing for coding latent constructs. *Communication Methods and Measures* 11(3): 191–209.

McGregor SC and Mourao RR (2016) Talking politics on Twitter: Gender, elections, and social networks. *Social Media + Society* 2(3): 1–14.

Megarry J (2014) Online incivility or sexual harassment? Conceptualising women's experiences in the digital age. *Women's Studies International Forum* 47(November): 46–55.

Puwar N (2004) *Space Invaders: Race, Gender and Bodies Out of Place*. Oxford: Berg.

Rudman LA and Phelan JE (2008) Backlash effects for disconfirming gender stereotypes in organizations. *Research in Organizational Behavior* 28(January): 61–79.

Rushowy K (2017) Twitter and Facebook are a minefield of threats, hate and anger for many female politicians. *The Toronto Star*, 24 February. Available at: https://www.thestar.com/news/canada/2017/02/24/twitter-and-facebook-are-a-minefield-of-threats-hate-and-anger-for-many-female-politicians.html (accessed 9 August 2017).

Saner E (2016) Vile online abuse against female MPs "needs to be challenged now". *The Guardian*, 18 June. Available at: https://www.theguardian.com/technology/2016/jun/18/vile-online-abuse-against-women-mps-needs-to-be-challenged-now (accessed 9 August 2017).

Shimanoff SB (2009) Gender role theory. In: Littlejohn SW and Foss KA (eds) *Encyclopedia of Communication Theory*. Thousand Oaks, CA: SAGE Publications, pp.434–436.

Smith KB (1997) When all's fair: Signs of parity in media coverage of female candidates. *Political Communication* 14(1): 71–82.

Sturino I and O'Brien L (2017) Female politicians speak out about sexist, violent cyberbullying. CBC Radio, 14 February. Available at: https://www.cbc.ca/radio/thecurrent/the-current-for-february-14-2017-1.3981592/female-politicians-speak-out-about-sexist-violent-cyberbullying-1.3981651 (accessed 9 August 2017).

Tausczik YR and Pennebaker JW (2010) The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29(1): 24–54.

The Canadian Press (2017) "I became a target": The difficult tenure of women politicians in Canada. CBC News, 2 August. Available at: https://www.cbc.ca/news/canada/newfoundland-labrador/women-politics-canada-1.4231672 (accessed 9 August 2017).

Vickery JR and Everbach T (2018) *Mediating Misogyny: Gender, Technology, and Harassment*. London: Palgrave Macmillan.

Supplemental Materials (Online Appendix)

Politicians in the Line of Fire: Incivility and the Treatment of Women on Social Media

Additional Information on Data Collection

We retrieved messages from the Twitter platform using the public streaming API during a period of one month for each country. For Canada, data collection took place from 7:00 AM to 3:00 AM, Eastern time, using a script launched automatically every day. The period of collection ranges from April 24 to May 26, 2017. For the US Senators, the platform was streamed in real-time between May 27 and July 5, 2017, during the same hours each day. In total, we collected 551,373 tweets addressed to Canadian politicians, and 5.6 million targeted at US Senators. There were no interruptions of service during that period. The Twitter streaming API is limited to 1% of the total quantity of statuses posted on the site at any given point (for detailed discussions, see Morstatter et al. 2013; Morstatter, Pfeffer, and Liu 2014; Joseph, Landwehr, and Carley 2014). Since we relied upon very specific search filters—the handles used by each politician—we generally remain well below the rate limits. This means that our data represent not a sample of tweets during that period, but virtually all the messages matching our search criteria. More specifically, Twitter reports the number of statuses that could not be retrieved when exceeding the rate limit. In total, 1,952 messages were not retrieved in Canada due to rate limits (about 0.4% of the total corpus size), and 81,322 for the United States (about 1.4% of the total). In other words, we were

able to collect roughly 99% of all messages meeting our criteria. Finally, note that some politicians had more than one Twitter account, in which case we used the one associated with their official function.

The statuses were processed to extract the displayed text using custom scripts. We considered statuses with at least three tokens (words or punctuation marks). We removed external links (URLs) from these messages, and after associating them to the politician they target, we removed all handles from the text. We removed all duplicate texts and purged the corpus from shared messages (retweets). Hence, our data collection is restricted to unique messages addressed directly to politicians. Moreover, we restricted the data to messages sent to a unique politician (that is, we exclude messages addressed to more than one recipients from our sample of politicians). Finally, we exclude a few tweets sent by the politicians themselves, to restrict our focus on the general public. The curated datasets contain 170,114 and 2.1 million tweets, respectively for Canada and the USA. We coded the gender and other attributes of politicians in our sample using their official biographies. Our measure of politician visibility is a variable measuring the count of followers on the site. This information was extracted from the website using the REST API between June 8 and June 10, 2017.

Defining Uncivil Tweets

Our training data annotated by human coders was described in the text, but we provide additional information here. Workers were provided with specific guidelines to identify uncivil tweets based on the six criteria mentioned in the text and discussed below. We also included specific examples and advice to interpret these criteria. FigureEight (formerly CrowdFlower) uses test questions—questions for which we provided the ground truth—to create a trust score for each coder. The platform's algorithm then computes a confidence in each judgment as the proportion choosing the majority category weighted by the individual trust scores. The average confidence is 93.0% for the American sample, and 94.3% for the Canadian sample.

When devising instructions, we defined as uncivil those messages containing explicit forms

of offensive language that human coders can readily detect—namely swear words, vulgarities, and direct insults (for instance, "idiot", "stupid")—as well as forms of incivility along the lines of those used in Papacharissi (2004)—namely threats, personal attacks directed at one's private life, and attacks toward groups (hate speech). This choice differs from a body of literature in political science that adopts broad definitions of incivility for the study of elite discourse—including negative campaign advertisements or an adversarial tone during televized debates. For example, in their experiment on televized incivility, Mutz and Reeves (2005) organized a mock debate between politicians, with subjects exposed to either a civil or an uncivil version of the same exchange. The uncivil tone was characterized with phrasings such as "You're really missing the point" (Mutz and Reeves 2005, 199), which represent mild violations of social norms yet were sufficient to affect the subjects' levels of political trust. Brooks and Geer (2007, 5) adopt a slightly different definition, identifying incivility in terms of discursive behaviours resorting to "animosity and derision" and the addition of "inflammatory comments that add little in the way of substance to the discussion." These conceptions of what constitutes civility have merits for studying elites, but they establish a high bar when analyzing political debates among members of the public on social media, where transgressions tend to be more extreme and more common. For example, it would be unlikely to witness a politician using profanity in public statements, yet such forms of incivility are part of the linguistic register in social media.

By establishing a higher threshold for what counts as incivility, we allow for the adversarial tone and heated exchanges to be expected in online debates. Actual examples from our corpus may help to illustrate the implications of our definition. The following two examples contain direct insults:

> *I bet your sick & twisted mind gets off on it. I know ppl like you; chip on shoulder, rejected by the opposite sex. You have a"loser" aura.*
>
> *How about you put a sock in it and go away!!!! You and your pant suit sisters need to ride off into the sunset. You are a b***h.* [*expletive blurred*]

In both cases, the nature of the message goes beyond the expression of political opinions during a heated exchange. Since they comprise one or more elements of our above definition (direct

insults, personal attacks), we view them as uncivil. On the other hand, we consider the following example to fall within the boundaries of civility:

You have no idea what rights and freedoms even mean to Canadians. Youre out of touch.

This tweet expresses a forceful criticism of a politician's character, and the statement rests on subjective assumptions. Unlike the two previous examples, however, the message does not contain offensive language or attacks referring to someone's private life. Notice that such a comment could be considered uncivil using Mutz and Reeves (2005)'s definition, if it were used in the context of a debate between political candidates. But it exemplifies a common type of criticism on social media. Conflating statements of that nature with the previous two would seriously boost our estimates about the prevalence of incivility, and in the process we would risk overlooking the severity of the more abusive comments. We prefer to rely on a more conservative approach.

Description of Machine Learning Models

Our models use three types of linguistic features as predictors for the category of a tweet. First, we make use of the 2,000 unigrams and bigrams (sequences of one and two words) most predictive of the class of a tweet in the annotated sample, based on chi-square values. Occurrences for these 2,000 expressions are converted into numerical values using a term-frequency/inverse document frequency (TF-IDF) weighting scheme, which gives additional importance to less common utterances. Prior to this step, we lemmatized the training sample (that is, we reduced each noun and verb to its root form) and removed English stop words, user handles, as well as mentions of the names of politicians in our main sample. These last steps avoid the reliance on clues too closely related to the recipient of the tweets when predicting their category. A few tweets with no textual content left after these steps were removed from the sample.

Second, we devise an indicator measuring the semantic similarity of a tweet with respect to a reference list of insults and swear words. This reference list is a filter for inappropriate content on the web, namely the *swearjar* JavaScript library.[1] We use a dataset of word embeddings—

[1]The list contains 247 common swear words and vulgarities for which we can compute similarity metrics.

the numerical coefficients of neural network models predicting word co-occurrences in large collections of texts (Mikolov et al. 2013; Pennington, Socher, and Manning 2014)—to compute the cosine similarity between any new lemma and those contained in the reference list. Specifically, we rely on publicly released word embeddings, pre-trained on a corpus of 27 billion tokens from the Twitter platform, fitted using the GloVe algorithm (Pennington, Socher, and Manning 2014). Our indicator is the maximum cosine similarity with the reference list of abusive words, for each tweet: the higher this maximum value, the more likely a tweet contains an offensive word.[2]

Third, we measure the sentiment of each tweet as a numerical value. We rely upon the *vader* library for Python (Gilbert and Hutto 2014), which was designed for social media data. The library computes a compound score ranging from -1 to 1 representing the emotional polarity of a document, from negative to positive. Since uncivil messages are more likely to be negative in tone, we expect sentiment to be a relevant predictor, even though this feature would be insufficient by itself.

Our objective is to fit a model that can both predict the incivility of individual tweets and the aggregate proportions of uncivil tweets accurately in the full corpus. To find the most suitable model, we compared the performance of classifiers commonly used for the analysis of text documents: support vector machines (SVM), decision trees, and logistic regressions. Our most accurate model is a SVM classifier fitted using 50 bootstrap aggregating (bagging) replications (Breiman 1996). Put simply, bagging consists of running the predictions multiple times after randomly resampling the training examples, and choosing the class (civil/uncivil) predicted the most often by the models. This method reduces concerns about overfitting (Bauer and Kohavi 1999). We also rely on a bagging estimator that accounts for the imbalance between the classes using random undersampling of the majority category.[3]

Table A1 reports accuracy statistics for our models, comparing SVMs with and without the bagging algorithm. Following conventions in the field of machine learning, we evaluate each model by first separating the sample into training and testing sets, to emulate the accuracy in

[2]This indicator accounts for obfuscation spellings and neologisms commonly used as insults on social media.

[3]We fit all models using the *sklearn* and *imblearn* libraries for Python. Our models will be made available to researchers upon publication.

the prediction of unseen documents. The statistics in Table A1 are averaged over 10 replications, using stratified 10-fold cross-validation (i.e. randomly splitting the sample into 10 parts and repeating the training and prediction stages 10 times using a different testing sample each time). The first two statistics evaluate the accuracy of individual class predictions in the testing sample: the percent correctly predicted and the area under the receiver operating characteristic curve (AUROC). As can be seen, the models using balanced bagging correctly predict the class of a tweet for close to 90% of cases in the American sample, and about 92% of cases in the Canadian sample. The distribution of tweets across the two classes being unbalanced, the AUROC statistic represents a more reliable metric since it assesses the capacity of each model to avoid both false positives and false negatives (the closer to 1, the better the model). Once again, the bagging estimators outperform the standard models. Finally, the proportion error is the absolute difference in the aggregate proportions of tweets in each class, that is, the difference between the percentage of tweets deemed to be uncivil by human coders and the percentage predicted to be uncivil by the model. The lower the error, the more accurate the prediction. We compare this last statistic to the one computed using Hopkins and King (2010)'s estimator (*ReadMe*), which we fit on the first part of a random 50/50 split of the annotated sample, and evaluate on the other part.[4] This model is not designed to predict individual documents, so the first two accuracy metrics cannot be computed. However, the *ReadMe* estimator tends to be more accurate at fitting proportions. As a result, it represents a useful benchmark to assess our models. Our final models generate proportions close to those achieved by this estimator.

Additional Results

Table A2 compares the baseline rates of uncivil tweets reported in the main text, along with additional word frequencies based on popular lexicons. These are frequencies by 1,000 words of expressions contained in the *swearjar* lexicon introduced earlier, and in two categories from the 2015 dictionaries of the popular psycholinguistic software LIWC (Tausczik and Pennebaker

[4]We use random subsets of 20 words and 300 repetitions. We fitted the model using the same 2,000 unigrams and bigrams as for the other classifiers. For information on these parameters, see (Hopkins and King 2010).

Table A1: Accuracy Results

Sample	Model	Accuracy	AUROC	Proportion Error
USA	SVM	87.05%	0.711	0.031
	SVM (Balanced Bagging)	89.27%	0.763	0.024
	ReadMe			0.020
Canada	SVM	90.65%	0.704	0.023
	SVM (Balanced Bagging)	91.68%	0.766	0.010
	ReadMe			0.007

Accuracy statistics are computed using stratified 10-fold cross-validation. We report average statistics over the 10 folds. The accuracy is the percent correctly predicted in the testing sets. AUROC stands for the area under the receiver operating characteristic (ROC) curve. We use Platt's method to retrieve the probability of a positive outcome with SVMs (Platt 1999). The proportion error is the absolute difference between the predicted and the true proportions of civil tweets.

2010), namely swear words and negative words. As is the case for the predicted proportions of uncivil tweets, the additional frequencies suggest that swear terms and negative words are used more frequently in messages sent to male politicians than in messages sent to female politicians. Again, these comparisons ignore the differences in status between politicians, which are relevant to derive substantively meaningful conclusions. Since men tend to be overrepresented among visible politicians, a multivariate analysis taking into account this confounder is justified.

Table A2: Inferring the Level of Incivility by Gender

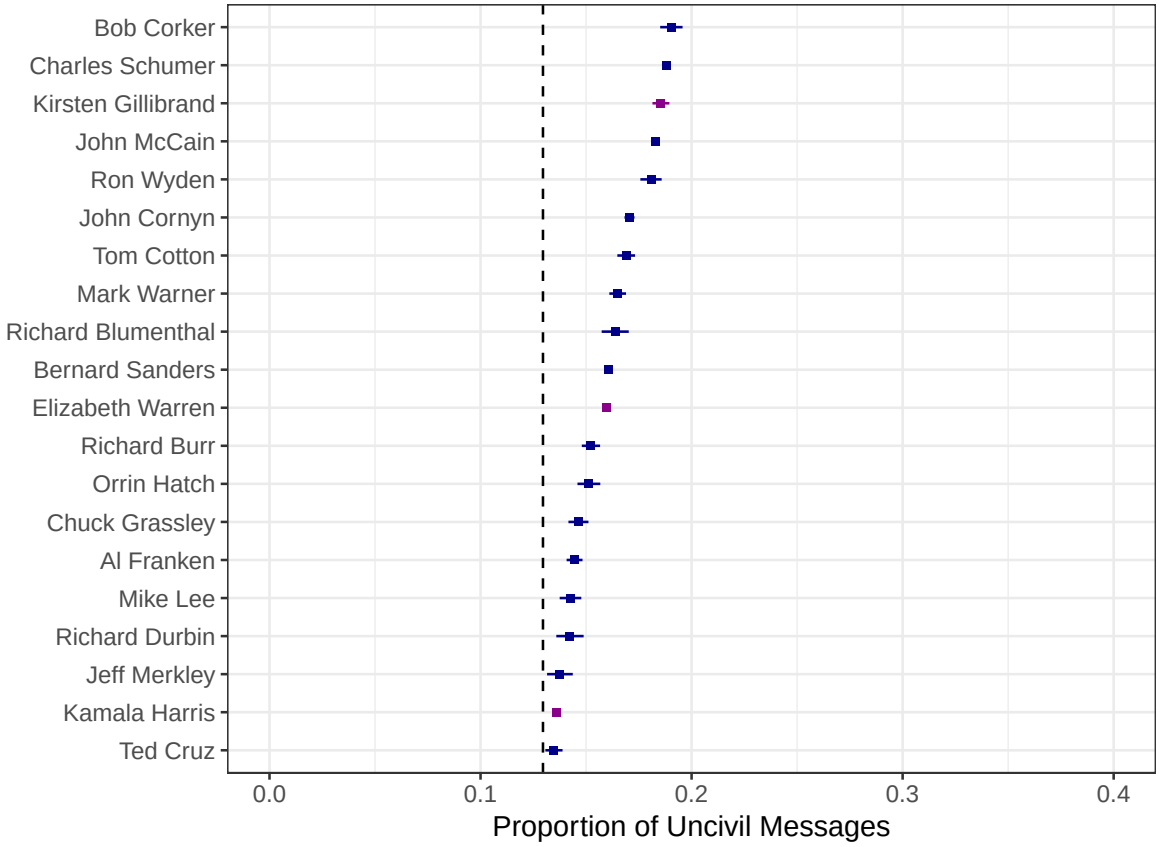
		United States			Canada		
	Method/Lexicon	Women	Men	Total	Women	Men	Total
Fitted Proportions	Classifier	12.95%	14.54%	14.13%	8.55%	11.66%	10.69%
Frequencies by 1,000 Words	Swear Jar	6.67	7.44	7.25	3.97	7.44	6.32
	LIWC Swear Words	11.35	12.63	12.31	7.01	11.98	10.38
	LIWC Negative Words	71.74	73.48	73.05	49.29	61.23	57.38
Corpus Size		530,663	1,545,175	2,075,838	53,195	116,919	170,114

Proportions predicted with a balanced bagging model using 50 replications of support vector machine estimators.

Figure A1 replicates the figure presented in the main text for Canada, and shows the 20 US Senators most often targeted by uncivil messages, restricting to those having received at least 10,000 tweets. As can be seen, the primary targets tend to occupy important positions in the upper house. For instance, the Democratic minority leader Chuck Schumer ranks in second

position, and Senators with a large follower count on Twitter such as John McCain and Bernard Sanders also feature in the Top 20. There are few women in the Senate to begin with, and there are even fewer of them enjoying a high status. But for the women who do have visibility, for instance New York Senator Kristen Gillibrand and Elizabeth Warren, uncivil messages are frequent.

Figure A1: US Senators Most Targeted by Uncivil Messages



The vertical line indicates the average proportion of uncivil messages received by Senators with at least 10,000 messages addressed to them in the corpus. We use a color code to distinguish between female and male politicians.

Aggregate Empirical Models

To assess the robustness of the multivariate results presented in the main text, we replicated the analysis by aggregating the count of uncivil tweets received by each politician. This considerably reduces the sample size (to 195 politicians in Canada, and 100 Senators in the United States). The aggregate dependent variable also accumulates prediction errors, and as a result this transformation may inflate standard errors. Nonetheless, we show that the main finding is repli-

cated with an aggregate dataset, for both countries. Moreover, we replicate the results using the counts of swear words contained in the tweets sent to each politician as the dependent variables. The counts are based on the LIWC dictionary and the swearjar list mentioned above. We show that the relationship emphasized in the main text is supported when using these alternative dependent variables. These replications suggest that the interaction of gender and visibility is not simply an artifact of the methodology used to predict uncivil tweets. For all models, we rely on quasi-Poisson regressions accounting for overdispersion.[5] All models use an offset of the log of the total number of tweets received to account for exposure.

To begin, Tables A3 and A4 report models with only two covariates and an interaction term, using each of the three different aggregated count variables as the outcome. As can be seen, the interaction between the female gender and the measure of visibility remains positive and statistically significant, as was the case in the main models. Again, this suggests that women politicians are more likely to become targets of uncivil messages, but conditional on gaining visibility. Without a high level of visibility, however, men are more likely to face incivility. As was the case for the main models, the results appear more robust, in terms of statistical level of confidence, when considering the sample of Canadian politicians.

Tables A5 and A6 report the output of count models including control variables. We account for party affiliation and visible minority status, as well as a variable relevant for each country. For Canada, we include a binary variable accounting for the level of government (which equals 1 for the federal level). For the United States, we include instead the seniority of a Senator in logged number of years. As can be seen, the main finding holds after accounting for these control variables.

[5] On the properties of quasi-Poisson regressions, see Ver Hoef and Boveng (2007).

Table A3: Aggregate Models of Incivility (Canada)

	Dependent variable:		
	Uncivil Tweets	LIWC Swear Words	Swearjar Words
Gender (Female = 1)	-2.078^{***}	-2.998^{***}	-2.614^{***}
	(0.500)	(0.569)	(0.713)
Log Follower Count	0.150^{***}	0.189^{***}	0.205^{***}
	(0.010)	(0.011)	(0.013)
Gender × Log Follower Count	0.195^{***}	0.270^{***}	0.233^{***}
	(0.043)	(0.049)	(0.062)
Intercept	-4.248^{***}	-5.079^{***}	-5.790^{***}
	(0.149)	(0.155)	(0.196)
Observations	195	195	195

Notes: Quasi-Poisson regression models using the count of uncivil tweets (model 1), or the count of lexicon words based on the resource indicated in the column headers (models 2 and 3). Each model includes an offset for exposure, using the log of the total number of tweets received during the period.
*p<0.05; **p<0.01; ***p<0.001

Table A4: Aggregate Models of Incivility (United States)

	Dependent variable:		
	Uncivil Tweets	LIWC Swear Words	Swearjar Words
Gender (Female = 1)	-1.757^{**}	-1.790^{**}	-1.836^{**}
	(0.546)	(0.561)	(0.683)
Log Follower Count	0.069^{***}	0.056^{***}	0.053^{**}
	(0.014)	(0.014)	(0.017)
Gender × Log Follower Count	0.120^{**}	0.122^{**}	0.125^{*}
	(0.040)	(0.041)	(0.050)
Intercept	-2.828^{***}	-3.097^{***}	-3.580^{***}
	(0.181)	(0.184)	(0.224)
Observations	100	100	100

Notes: Quasi-Poisson regression models using the count of uncivil tweets (model 1), or the count of lexicon words based on the resource indicated in the column headers (models 2 and 3). Each model includes an offset for exposure, using the log of the total number of tweets received during the period.
*p<0.05; **p<0.01; ***p<0.001

Table A5: Aggregate Models of Incivility, with Controls (Canada)

	Dependent variable:		
	Uncivil Tweets	LIWC Swear Words	Swearjar Words
Gender (Female = 1)	-2.440^{***}	-4.930^{***}	-4.009^{***}
	(0.535)	(0.752)	(0.977)
Log Follower Count	0.184^{***}	0.179^{***}	0.207^{***}
	(0.015)	(0.016)	(0.022)
Gender × Log Follower Count	0.240^{***}	0.461^{***}	0.373^{***}
	(0.048)	(0.067)	(0.087)
Visible Minority	0.497^{***}	0.134	0.226
	(0.076)	(0.091)	(0.122)
Party (Liberal = 1)	-0.155^{*}	-0.453^{***}	-0.351^{**}
	(0.078)	(0.088)	(0.121)
Federal Level	0.063	0.416^{***}	0.311^{**}
	(0.072)	(0.087)	(0.117)
Intercept	-4.674^{***}	-4.900^{***}	-5.777^{***}
	(0.178)	(0.196)	(0.270)
Observations	195	195	195

Notes: Quasi-Poisson regression models using the count of uncivil tweets (model 1), or the count of lexicon words based on the resource indicated in the column headers (models 2 and 3). Each model includes an offset for exposure, using the log of the total number of tweets received during the period.

*p<0.05; **p<0.01; ***p<0.001

Table A6: Aggregate Models of Incivility, with Controls (United States)

	Dependent variable:		
	Uncivil Tweets	LIWC Swear Words	Swearjar Words
Gender (Female = 1)	-2.170^{***}	-1.882^{**}	-2.136^{**}
	(0.551)	(0.608)	(0.753)
Log Follower Count	0.058^{***}	0.055^{***}	0.046^{*}
	(0.014)	(0.015)	(0.018)
Gender × Log Follower Count	0.150^{***}	0.124^{**}	0.145^{*}
	(0.041)	(0.046)	(0.057)
Visible Minority	0.194^{*}	0.136	0.179
	(0.089)	(0.098)	(0.120)
Party (Democrat = 1)	0.122^{*}	0.112^{*}	0.062
	(0.048)	(0.053)	(0.066)
Log Seniority	0.118^{***}	0.049	0.067
	(0.031)	(0.034)	(0.042)
Intercept	-3.006^{***}	-3.221^{***}	-3.671^{***}
	(0.174)	(0.189)	(0.233)
Observations	100	100	100

Notes: Quasi-Poisson regression models using the count of uncivil tweets (model 1), or the count of lexicon words based on the resource indicated in the column headers (models 2 and 3). Each model includes an offset for exposure, using the log of the total number of tweets received during the period.

$^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

References

Bauer, Eric, and Ron Kohavi. 1999. "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants." *Machine learning* 36(1-2): 105–139.

Breiman, Leo. 1996. "Bagging predictors." *Machine learning* 24(2): 123–140.

Brooks, Deborah Jordan, and John G. Geer. 2007. "Beyond Negativity: The Effects of Incivility on the Electorate." *American Journal of Political Science* 51(January): 1–16.

Gilbert, C.J., and Eric Hutto. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM-14).* pp. 216–225.

Hopkins, Daniel J, and Gary King. 2010. "A method of automated nonparametric content analysis for social science." *American Journal of Political Science* 54(1): 229–247.

Joseph, Kenneth, Peter M Landwehr, and Kathleen M Carley. 2014. Two 1% s Don't Make a Whole: Comparing Simultaneous Samples from Twitter's Streaming API. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction.* Springer pp. 75–83.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at ICLR.*

Morstatter, Fred, Jürgen Pfeffer, and Huan Liu. 2014. When Is it Biased? Assessing the Representativeness of Twitter's Streaming API. In *Proceedings of the 23rd International Conference on World Wide Web.* ACM pp. 555–556.

Morstatter, Fred, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media.*

Mutz, Diana C., and Byron Reeves. 2005. "The New Videomalaise: Effects of Televised Incivility on Political Trust." *The American Political Science Review* 99(1): 1–15.

Papacharissi, Zizi. 2004. "Democracy online: civility, politeness, and the democratic potential of online political discussion groups." *New Media & Society* 6(April): 259–283.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*.

Platt, John C. 1999. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *Advances in Large Margin Classifiers*. MIT Press pp. 61–74.

Tausczik, Yla R., and James W. Pennebaker. 2010. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods." *Journal of Language and Social Psychology* 29(1): 24–54.

Ver Hoef, Jay M., and Peter L. Boveng. 2007. "Quasi-Poisson vs. Negative Binomial Regression: How Should We Model Overdispersed Count Data?" *Ecology* 88(11): 2766–2772.