Women, politics and Twitter: Using machine learning to change the discourse

Lana Cuthberston*
ParityYEG
lanacuthbertson@gmail.com

Alex Kearney
University of Alberta
kearney@ualberta.ca

Riley Dawson
University of Alberta
rileydawson@ualberta.ca

Ashia Zawaduk
Dev Edmonton Society
mail@ashia.ca

Eve Cuthbertson
University of Toronto
evecuthbertson@gmail.com

Ann Gordon-Tighe
University of Alberta
ann.gordontighe@gmail.com

Kory W Mathewson
DeepMind
University of Alberta
korymath@google.com

Abstract

Including diverse voices in political decision-making strengthens our democratic institutions. Within the Canadian political system, there is gender inequality across all levels of elected government. Online abuse, such as hateful tweets, leveled at women engaged in politics contributes to this inequity, particularly tweets focusing on their gender. In this paper, we present *ParityBOT*: a Twitter bot which counters abusive tweets aimed at women in politics by sending supportive tweets about influential female leaders and facts about women in public life. ParityBOT is the first artificial intelligence-based intervention aimed at affecting online discourse for women in politics for the better. The goal of this project is to: 1) raise awareness of issues relating to gender inequity in politics, and 2) positively influence public discourse in politics. The main contribution of this paper is a scalable model to classify and respond to hateful tweets with quantitative and qualitative assessments. The ParityBOT abusive classification system was validated on public online harassment datasets. We conclude with analysis of the impact of ParityBOT, drawing from data gathered during interventions in both the 2019 Alberta provincial and 2019 Canadian federal elections.

1 Introduction

Our political systems are unequal, and we suffer for it. Diversity in representation around decision-making tables is important for the health of our democratic institutions [18]. One example of this inequity of representation is the gender disparity in politics: there are fewer women in politics than men, largely because women do not run for office at the same rate as men. This is because women face systemic barriers in political systems across the world [24]. One of these barriers is online harassment [19, 21]. Twitter is an important social media platform for politicians to share their visions and engage with their constituents. Women are disproportionately harassed on this platform because of their gender [12].

*Corresponding author

AI for Social Good workshop at NeurIPS (2019), Vancouver, Canada.

To raise awareness of online abuse and shift the discourse surrounding women in politics, we designed, built, and deployed *ParityBOT*: a Twitter bot that classifies hateful tweets directed at women in politics and then posts "positivitweets". This paper focuses on how ParityBOT improves discourse in politics.

Previous work that addressed online harassment focused on collecting tweets directed at women engaged in politics and journalism and determining if they were problematic or abusive [3, 7, 19]. Inspired by these projects, we go one step further and develop a tool that directly engages in the discourse on Twitter in political communities. Our hypothesis is that by seeing "positivitweets" from ParityBOT in their Twitter feeds, knowing that each tweet is an anonymous response to a hateful tweet, women in politics will feel encouraged and included in digital political communities[10]. This will reduce the barrier to fair engagement on Twitter for women in politics. It will also help achieve gender balance in Canadian politics and improve gender equality in our society.

2 Methods

2.1 Technical Details for ParityBot

In this section, we outline the technical details of ParityBot. The system consists of: 1) a Twitter listener that collects and classifies tweets directed at a known list of women candidates, and 2) a responder that sends out positivitweets when hateful tweets are detected.

We collect tweets from Twitter's real-time streaming API. The stream listener uses the open-source Python library Tweepy [22]. The listener analyses tweets in real-time by firing an asynchronous tweet analysis and storage function for each English tweet mentioning one or more candidate usernames of interest. We limit the streaming to English as our text analysis models are trained on English language corpora. We do not track or store retweets to avoid biasing the analysis by counting the same content multiple times. Twitter data is collected and used in accordance with the acceptable terms of use [23].

The tweet analysis and storage function acts as follows: 1) parsing the tweet information to clean and extract the tweet text, 2) scoring the tweet using multiple text analysis models, and 3) storing the data in a database table. We clean tweet text with a variety of rules to ensure that the tweets are cleaned consistent with the expectations of the analysis models (see Appdx 1.1).

The text analysis models classify a tweet by using, as features, the outputs from Perspective API from Jigsaw [17], HateSonar [8], and VADER sentiment models [11]. Perspective API uses machine learning models to score the perceived impact a tweet might have [17]. The outputs from these models (i.e. 17 from Perspective, 3 from HateSonar, and 4 from VADER) are combined into a single feature vector for each tweet (see Appdx 1.2). No user features are included in the tweet analysis models. While these features may improve classification accuracy they can also lead to potential bias [25].

We measure the relative correlation of each feature with the hateful or not hateful labels. We found that Perspective API's TOXICITY probability was the most consistently predictive feature for classifying hateful tweets. Fig. 1 shows the relative frequencies of hateful and non-hateful tweets over TOXICITY scores. During both elections, we opted to use a single Perspective API feature to trigger sending positivitweets. Using the single TOXICITY feature is almost as predictive as using all features and a more complex model 1.3. It was also simpler to implement and process tweets at scale. The TOXICITY feature is the only output from the Perspective API with transparent evaluation details summarized in a Model Card [15, 20].

2.2 Collecting Twitter handles, predicting candidate gender, curating "positivitweets"

Deploying ParityBOT during the Alberta 2019 election required volunteers to use online resources to create a database of all the candidates running in the Alberta provincial election. Volunteers recorded each candidate's self-identifying gender and Twitter handle in this database. For the 2019 federal Canadian election, we scraped a Wikipedia page that lists candidates [26]. We used the Python library *gender-guesser* [5] to predict the gender of each candidate based on their first names. As much as possible, we manually validated these predictions with corroborating evidence found in candidates' biographies on their party's websites and in their online presence.

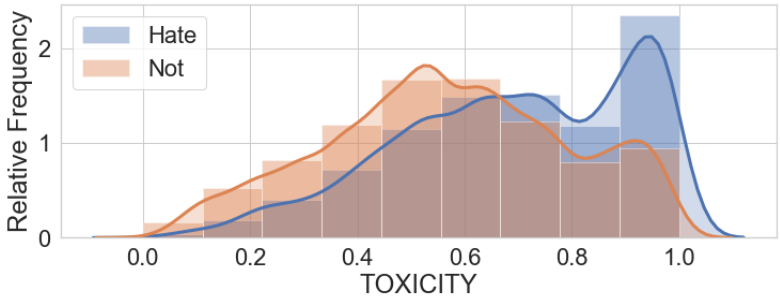

Figure 1: Visualizing the training data distribution. Relative frequency of hateful versus not hateful tweets for varying levels of the Perspective API [17] TOXICITY score. Normalized histograms are plotted underneath kernel density estimation (KDE) plots.

ParityBOT sent positivitweets composed by volunteers. These tweets expressed encouragement, stated facts about women in politics, and aimed to inspire and uplift the community. Volunteers submitted many of these positivitweets through an online form.[2] Volunteers were not screened and anyone could access the positivitweet submission form. However, we mitigate the impact of trolls submitting hateful content, submitter bias, and ill-equipped submitters by reviewing, copy editing, and fact checking each tweet. Asking for community contribution in this way served to maximize limited copywriting resources and engage the community in the project.

2.3 Qualitative Assessment

We evaluated the social impact of our system by interviewing individuals involved in government ($n = 5$). We designed a discussion guide based on user experience research interview standards to speak with politicians in relevant jurisdictions [14]. Participants had varying levels of prior awareness of the ParityBOT project. Our participants included 3 women candidates, each from a different major political party in the 2019 Alberta provincial election, and 2 men candidates at different levels of government representing Alberta areas. The full discussion guide for qualitative assessment is included in Appdx 3.1. All participants provided informed consent to their anonymous feedback being included in this paper.

3 Results and Outcomes

We deployed ParityBOT during two elections: 1) the 2019 Alberta provincial election, and 2) the 2019 Canadian federal election. For each tweet we collected, we calculated the probability that the tweet was hateful or abusive. If the probability was higher than our response decision threshold, a positivitweet was posted. Comprehensive quantitative results are listed in Appendix 2.

During the Alberta election, we initially set the decision threshold to a TOXICITY score above 0.5 to capture the majority of hateful tweets, but we were sending too many tweets given the number of positivitweets we had in our library and the Twitter API daily limit [23]. Thus, after the first 24 hours that ParityBOT was live, we increased the decision threshold to 0.8, representing a significant inflection point for hatefulness in the training data (Fig. 1). We further increased the decision threshold to 0.9 for the Canadian federal election given the increase in the number and rate of tweets processed. For the Alberta provincial election, the model classified 1468 tweets of the total 12726 as hateful, and posted only 973 positivitweets. This means that we did not send out a positivitweet for every classified hateful tweet, and reflects our decision rate-limit of ParityBOT. Similar results were found for the 2019 Canadian election.

[2]The full list of positivitweets can be found at `https://paritybot.com`

3.1 Values and Limitations

We wrote guidelines and values for this to guide the ongoing development of the ParityBOT project.[3] These values help us make decision and maintain focus on the goal of this project.

While there is potential to misclassify tweets, the repercussions of doing so are limited. With ParityBOT, false negatives, hateful tweets classified as non-hateful, are not necessarily bad, since the bot is tweeting a positive message. False positives, non-hateful tweets classified as hateful, may result in tweeting too frequently, but this is mitigated by our choice of decision threshold.

In developing ParityBOT, we discussed the risks of using bots on social media and in politics. First, we included the word "bot" in the project title and Twitter handle to be clear that the Twitter account was tweeting automatically. We avoided automating any direct "at (@) mention" of Twitter users, only identifying individuals' Twitter handles manually when they had requested credit for their submitted positivitweet. We also acknowledge that we are limited in achieving certainty in assigning a gender to each candidate.

3.2 User experience research results

In our qualitative research, we discovered that ParityBOT played a role in changing the discourse. One participant said, "it did send a message in this election that there were people watching" (P2). We consistently heard that negative online comments are a fact of public life, even to the point where it's a signal of growing influence. "When you're being effective, a good advocate, making good points, people are connecting with what you're saying. The downside is, it comes with a whole lot more negativity [...] I can always tell when a tweet has been effective because I notice I'm followed by trolls" (P1).

We heard politicians say that the way they have coped with online abuse is to ignore it. One participant explained, "I've tried to not read it because it's not fun to read horrible things about yourself" (P4). Others dismiss the idea that social media is a useful space for constructive discourse: "Because of the diminishing trust in social media, I'm stopping going there for more of my intelligent discourse. I prefer to participate in group chats with people I know and trust and listen to podcasts" (P3).

4 Future Work and Conclusions

We would like to run ParityBOT in more jurisdictions to expand the potential impact and feedback possibilities. In future iterations, the system might better match positive tweets to the specific type of negative tweet the bot is responding to. Qualitative analysis helps to support the interventions we explore in this paper. To that end, we plan to survey more women candidates to better understand how a tool like this impacts them. Additionally, we look forward to talking to more women interested in politics to better understand whether a tool like this would impact their decision to run for office. We would like to expand our hateful tweet classification validation study to include larger, more recent abusive tweet datasets [1, 4]. We are also exploring plans to extend ParityBOT to invite dialogue: for example, asking people to actively engage with ParityBOT and analyse reply and comment tweet text using natural language-based discourse analysis methods.

During the 2019 Alberta provincial and 2019 Canadian federal elections, ParityBOT highlighted that hate speech is prevalent and difficult to combat on our social media platforms as they currently exist, and it is impacting democratic health and gender equality in our communities [2]. We strategically designed ParityBOT to inject hope and positivity into politics, to encourage more diverse candidates to participate. By using machine learning technology to address these systemic issues, we can help change the discourse an link progress in science to progress in humanity.

[3] `https://paritybot.com`

References

[1] Valerio Basile et al. "Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter". In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. 2019, pp. 54–63.

[2] Danielle Keats Citron. "Cyber civil rights". In: *BUL Rev.* 89 (2009), p. 61.

[3] Laure Delisle et al. "A large-scale crowdsourced analysis of abuse against women journalists and politicians on Twitter". In: *arXiv preprint arXiv:1902.03093* (2019).

[4] Antigoni Maria Founta et al. "Large scale crowdsourcing and characterization of twitter abusive behavior". In: *Twelfth International AAAI Conference on Web and Social Media*. 2018.

[5] *Gender Guesser*. `https://pypi.org/project/gender-guesser/`. Acc: 2019-09-06.

[6] Jennifer Golbeck et al. "A large labeled corpus for online harassment research". In: *Proceedings of the 2017 ACM on Web Science Conference*. ACM. 2017, pp. 229–233.

[7] Mark A Greenwood et al. "Online Abuse of UK MPs from 2015 to 2019". In: *arXiv preprint arXiv:1904.11230* (2019).

[8] *HateSonar*. `https://github.com/Hironsan/HateSonar`. Acc: 2019-09-06.

[9] Haibo He et al. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning". In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE. 2008, pp. 1322–1328.

[10] Frank La Rue Heiner Bielefeldt and Githu Muigai. "OHCHR expert workshops on the prohibition of incitement to national, racial or religious hatred". In: *Expert workshop on Africa* (2011).

[11] Clayton J Hutto and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text". In: *8th AAAI on weblogs and social media*. 2014.

[12] Amnesty International. *Toxic Twitter: Women's experiences of violence and abuse on Twitter*. `http://bit.ly/2jZTb5w`. 2018.

[13] Guolin Ke et al. "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in Neural Information Processing Systems*. 2017, pp. 3146–3154.

[14] Steve Krug. *Rocket surgery made easy: The do-it-yourself guide to finding and fixing usability problems*. New Riders, 2009.

[15] Margaret Mitchell et al. "Model cards for model reporting". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM. 2019, pp. 220–229.

[16] Randal S Olson and Jason H Moore. "TPOT: A tree-based pipeline optimization tool for automating machine learning". In: *Automated Machine Learning*. Springer, 2019.

[17] *Perspective API*. `https://www.perspectiveapi.com/`. Acc: 2019-09-06.

[18] Anne Phillips. "Democracy and representation: Or, why should it matter who our representatives are?" In: *Feminism and politics* 224 (1998), p. 240.

[19] Ludovic Rheault, Erica Rayment, and Andreea Musulan. "Politicians in the line of fire: Incivility and the treatment of women on social media". In: *Research & Politics* 6.1 (2019).

[20] *Toxicity Model Card*. `https://github.com/conversationai/perspectiveapi/blob/master/2-api/model-cards/English/toxicity.md`. Acc: 2019-11-14.

[21] Linda Trimble. *Ms. Prime Minister: Gender, Media, and Leadership*. U. Toronto Press, 2018.

[22] *Tweepy*. `http://www.tweepy.org/`. Acc: 2019-09-06.

[23] *Twitter Developer Agreement and Policy*. `https://developer.twitter.com/en/developer-terms/agreement-and-policy.html`. Acc: 2019-09-06.

[24] Inter-Parliamentary Union. *Sexism, harassment and violence against women parliamentarians*. `https://www.ipu.org/file/2425/download?token=0H5YdXVB`. Oct. 2018.

[25] Zeerak Waseem and Dirk Hovy. "Hateful symbols or hateful people? predictive features for hate speech detection on twitter". In: *Proceedings of the NAACL student research workshop*. 2016, pp. 88–93.

[26] *Wikipedia: List of candidates by riding for the 43rd Canadian federal election*. `https://w.wiki/7zT`. Acc: 2019-09-06.

1 Tweet Cleaning and Feature Details

1.1 Tweet Cleaning Methods

We use regular expression rules to clean tweets: convert the text to lowercase, remove URLs, strip newlines, replace whitespace with a single space, and replace mentions with the text tag 'MENTION'. While these rules may bias the classifiers, they allow for consistency and generalization between training, validation, and testing datasets.

1.2 Tweet Featurization Details

Each tweet is processed by three models: Perspective API from Jigsaw [17], HateSonar [8], and VADER sentiment models [11]. Each of these models outputs a score between $[0, 1]$ which correlates the text of the tweet with the specific measure of the feature. The outputs from these models (i.e. 17 from Perspective, 3 from HateSonar, and 4 from VADER) are combined into a single feature vector for each tweet. Below we list the outputs for each text featurization model:

- **Perspective API**: 'IDENTITY_ATTACK', 'INCOHERENT', 'TOXICITY_FAST', 'THREAT', 'INSULT', 'LIKELY_TO_REJECT', 'TOXICITY', 'PROFANITY', 'SEXUALLY_EXPLICIT', 'ATTACK_ON_AUTHOR', 'SPAM', 'ATTACK_ON_COMMENTER', 'OBSCENE', 'SEVERE_TOXICITY', 'INFLAMMATORY'
- **HateSonar**: 'sonar_hate_speech', 'sonar_offensive_language', 'sonar_neither'
- **VADER**: 'vader_neg', 'vader_neu', 'vader_pos', 'vader_compound'

1.3 Validation and Ablation Experiments

For validation, we found the most relevant features and set an abusive prediction threshold by using a dataset of 20194 cleaned, unique tweets identified as either hateful and not hateful from previous research [6]. Each entry in our featurized dataset is composed of 24 features and a class label of *hateful or not hateful*. The dataset is shuffled and randomly split into training (80%) and testing (20%) sets matching the class balance (25.4% hateful) of the full dataset. We use Adaptive Synthetic (ADASYN) sampling to resample and balance class proportions in the dataset [9].

With the balanced training dataset, we found the best performing classifier to be a gradient boosted decision tree [13] by sweeping over a set of possible models and hyperparameters using TPOT [16]. For this sweep, we used 10-fold cross validation on the training data. We randomly partition this training data 10 times, fit a model on a training fraction, and validate on the held-out set.

We performed an ablation experiment to test the relative impact of the features derived from the various text classification models.

2 Quantitative analysis of elections

This table includes quantitative results from the deployment of ParityBOT in the Alberta 2019 provincial and Canadian 2019 federal elections.

	Alberta 2019 provincial election Apr. 1-15 2019	**Canada 2019 federal election** Sep. 11-Oct. 26 2019
Total positivitweets sent	973	2428
Total impressions	84,961	304,600
Total retweets	142	529
Total likes	412	1500
Total replies	n/a	30
Total tweets analysed	12,726	228,255
Total tweets scored abusive	1468	9987
Abusive rate	7.65%	4.38%
Total candidates tracked	90	314
Decision threshold	0.8 (80% likely to be abusive)	0.9 (90% likely to be abusive)

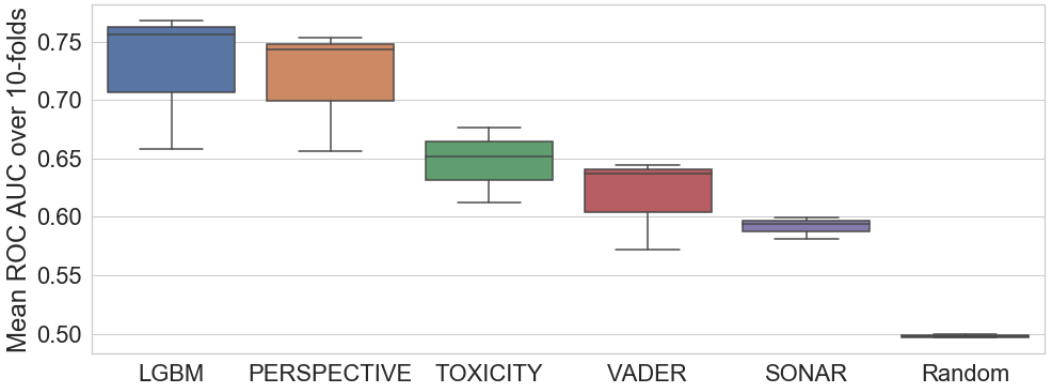

Figure 2: 10-fold cross validation ablation experiment showing the relative impacts of including various feature sets (PERSPECTIVE [17], VADER [11], HATESONAR [8], from left to right) in the feature vectors on performance. Performance is measured using binary classification area under the receiver operated characteristic curve (ROC AUC) and averaged over the 10-folds. These methods are compared with the best performing classifier from the validation study, a gradient-boosted decision tree (textscLGBM [13], left) and a stratified RANDOM classifier (right).

3 ParityBOT Research Plan and Discussion Guide

Overview Interviews will be completed in three rounds with three different target participant segments.

Research Objectives

- Understand if and how the ParityBOT has impacted women in politics
- Obtain feedback from Twitter users who've interacted with the bot
- Explore potential opportunities to build on the existing idea and platform
- Gain feedback and initial impressions from people who haven't interacted with the Bot, but are potential audience

Target Participants

- Round 1: Women in politics who are familiar with the Bot
- Round 2: Women who've interacted with the Bot (maybe those we don't know)
- Round 3: Some women who may be running in the federal election who haven't heard of the ParityBOT, but might benefit from following it
- All participants: Must be involved in politics in Canada and must be engaged on Twitter - i.e. have an account and follow political accounts and/or issues

Recruiting

- Round 1: [Author] recruit from personal network via text
- Round 2: Find people who've interacted with the bot on Twitter who we don't know, send them a DM, and ask if we can get their feedback over a 15- to 30-minute phone call
- Round 3: Use contacts in Canadian politics to recruit participants who have no prior awareness of ParityBOT

Method 15- to 30-minute interviews via telephone

Output Summary of findings in the form of a word document that can be put into the paper

3.1 Discussion Guide

Introduction

[Author]: Hey! Thanks for doing this. This shouldn't take longer than 20 minutes. [Author] is a UX researcher and is working with us. They'll take it from here and explain our process, get your consent and conduct the interview. I'll be taking notes. Over to [Author]!

[Author]: Hi, my name is [Author], I'm working with [Author] and [Author] to get feedback on the ParityBOT; the Twitter Bot they created during the last provincial election.

With your permission, we'd like to record our conversation. The recording will only be used to help us capture notes from the session and figure out how to improve the project, and it won't be seen by anyone except the people working on this project. We may use some quotes in an academic paper, You'll be anonymous and we won't identify you personally by name.

If you have any concerns at time, we can stop the interview and the recording. Do we have your permission to do this? (Wait for verbal "**yes**").

Round 1 (Women in Politics familiar with ParityBOT)

Background and Warm Up

- When you were thinking about running for politics what were your major considerations? For example, barriers, concerns?
- We know that online harassment is an issue for women in politics - have you experienced this in your career? How do you deal with harassment? What are your coping strategies?
- What advice would you give to women in politics experiencing online harassment?

Introduction to PartyBOT Thanks very much, now, more specifically about the ParityBOT:

- What do you know about the ParityBOT?
- What do you think it's purpose is?
- Did you encounter it? Tell me about how you first encountered it? Did it provide any value to you during your campaign? How? Do you think this is a useful tool? Why or why not? Did it mitigate the barrier of online harassment during your time as a politician?
- Is there anything you don't like about the Bot?

Next Steps If you could build on this idea of mitigating online harassment for women in politics, what ideas or suggestions would you have?

Conclusion Any other thoughts or opinions about the ParityBOT you'd like to share before we end our call?

Thank you very much for your time! If you have any questions, or further comments, feel free to text or email [Author].