

Hate Speech in the Political Discourse on Social Media: Disparities Across Parties, Gender, and Ethnicity

Kirill Solovev
kirill.solovev@wi.jlug.de
JLU Giessen
Germany

Nicolas Pröllochs
nicolas.proellocks@wi.jlug.de
JLU Giessen
Germany

ABSTRACT

Social media has become an indispensable channel for political communication. However, the political discourse is increasingly characterized by hate speech, which affects not only the reputation of individual politicians but also the functioning of society at large. In this work, we empirically analyze how the amount of hate speech in replies to posts from politicians on Twitter depends on personal characteristics, such as their party affiliation, gender, and ethnicity. For this purpose, we employ Twitter's Historical API to collect every tweet posted by members of the 117th U.S. Congress for an observation period of more than six months. Additionally, we gather replies for each tweet and use machine learning to predict the amount of hate speech they embed. Subsequently, we implement hierarchical regression models to analyze whether politicians with certain characteristics receive more hate speech. We find that tweets are particularly likely to receive hate speech in replies if they are authored by (i) persons of color from the Democratic party, (ii) white Republicans, and (iii) women. Furthermore, our analysis reveals that more negative sentiment (in the source tweet) is associated with more hate speech (in replies). However, the association varies across parties: negative sentiment attracts more hate speech for Democrats (vs. Republicans). Altogether, our empirical findings imply significant differences in how politicians are treated on social media depending on their party affiliation, gender, and ethnicity.

CCS CONCEPTS

• **Human-centered computing** → **Social media; Empirical studies in collaborative and social computing**; • **Applied computing** → **Sociology**.

KEYWORDS

Social media, political discourse, hate speech, sentiment analysis, disparities, computational social science, explanatory modeling

ACM Reference Format:

Kirill Solovev and Nicolas Pröllochs. 2022. Hate Speech in the Political Discourse on Social Media: Disparities Across Parties, Gender, and Ethnicity. In *Proceedings of the ACM The Web Conf (WWW '22)*, April 25–29, 2022, Lyon, France. ACM, New York, NY, USA, 5 pages.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WWW '22, April 25–29, 2022, Lyon, France

© 2022 Association for Computing Machinery.

1 INTRODUCTION

Social media has become an indispensable communication channel for politicians in the U.S. and around the world. Compared to traditional media, it provides a number of key benefits for politicians: (i) social media provides a tool to spread messages to the public at scale, thereby increasing people's awareness of their (political) agenda [25, 30, 46]. (ii) Social media encourages the dialogue between politicians and users, allowing for direct feedback from constituents and discussions of political ideas [18]. (iii) Due to its interactive nature, social media can be used as a tool for political mobilization [31, 34]. These benefits are further reinforced by the openness of social media as politicians are no longer restricted by geography, scope, or content and can reach significantly wider audiences [23].

However, the shift from traditional channels towards social media does not necessarily improve the quality of the political discourse. Instead, social media is known to foster echo chambers and “us versus them” rhetoric [36]. These factors correlate with cyberbullying, harassment, and, in particular, hate speech [19]. Broadly speaking, hate speech refers to abusive or threatening speech (or writing) that expresses prejudice against a particular group, often on the basis of ethnicity or sexual orientation [49]. Hate speech often originates from semi-anonymous trolls [28, 36], and is particularly frequent in discussions that cause a strong emotional response, such as in political topics [57]. The adoption of social media by politicians is thus a double-edged sword posing risks both to themselves and society as a whole [29]. At the individual level, hate speech can threaten reputations and may even lead to long-run mental health issues [56]. At the societal level, it fosters political polarization [32], which can have severe consequences. Examples include erosion of intergroup political relations and increased opportunities for the spread of ideologically branded misinformation [21, 45, 51].

Research Goal: In this study, we empirically analyze how the user base on Twitter responds to posts from members of the U.S. Congress. We are interested in understanding whether differences in the prevalence of hate speech can be explained by personal characteristics of politicians, such as their party affiliation, gender, and ethnicity. More precisely, we address the following research questions:

- **(RQ1)** *Are Twitter users more likely to respond with hate speech to tweets from U.S. representatives depending on party affiliation, gender, and ethnicity of the members of the U.S. Congress?*
- **(RQ2)** *Does hate speech in the replies to tweets depend on the sentiment of the source tweet? Does the strength of the association differ depending on their party, gender, and ethnicity?*

Data & Methods: To address our research questions, we employ the Twitter Historical API to collect all tweets from members of the 117th U.S. Congress between the first session on January 3, 2021 and the end of July 2021. In addition, we collect replies to each source tweet. We then use machine learning to determine the share of replies of each tweet that embeds hate speech. Subsequently, we implement a multilevel binomial regression model with random effects to estimate whether Twitter users are more likely to respond with hate speech depending on the party affiliation, gender, and ethnicity of the politician that has posted the tweet.

Contributions: To the best of our knowledge, this study is the first to empirically model how hate speech in replies to tweets from politicians depends on their personal characteristics (party affiliation, gender, ethnicity). All else being equal, we find that tweets are more likely to receive hate speech in replies if they are authored by (i) persons of color from the Democratic party, (ii) white Republicans, and (iii) women. As an additional contribution, our analysis reveals that more negative sentiment (in the source tweet) is associated with more hate speech (in replies). However, the association varies across parties: negative sentiment attracts more hate speech for Democrats (vs. Republicans). Altogether, our findings fuel new insights into ongoing discussions on political polarization on social media and highlight disparities in how politicians are treated depending on their party affiliation, gender, and ethnicity

2 BACKGROUND

Political communication on Twitter: The use of social media by U.S. politicians has experienced a rapid surge. At the start of 2009, only 69 individual members of Congress had a Twitter account [24]. Today, every member of the U.S. Congress has a professional Twitter account and oftentimes a second personal account being active at the same time. Existing studies suggest that there are three main reasons *why* politicians adopt social media [29]. First, social media allows for *unidirectional delivery* of information to the public. Compared to classical media, there is less moderation and real time scrutiny allowing politicians to freely express themselves [4]. Second, social media enables *dialogue* between politicians and the public. Politicians can use social media as a tool to connect with constituents to discuss political issues and receive feedback [18]. Engaged users may further spread the message with likes and/or reshares. Third, social media can be seen as a tool for *political mobilization*. Specifically, it allows politicians to rally for projects, events, and movements [55], though it does not guarantee success [35].

Hate speech: Although there is no all-encompassing definition [10], hate speech is typically considered to refer to abusive or threatening speech (or writing) that expresses prejudice against a particular group, often on the basis of ethnicity or sexual orientation [49]. While research on hate speech has received increasing attention lately [e.g., 3, 12, 13, 17, 37, 38, 41, 47, 58, 60], studies that analyze hate speech in the context of political communication are scant. The few existing works typically focus on qualitative insights or analysis of summary statistics. For instance, previous works have studied hate speech towards female Japanese politicians [22], far-right political party discourse in Spain [9], hateful propaganda towards politicians in Macedonia [9], hate speech against Members of Parliament in the U.K. [2], and hate against German politicians [15].

We are aware of only one paper analyzing hate speech and incivility in the context of tweets from members of the U.S. Congress [54]. However, this study again focuses on summary statistics. In particular, it does not model the effects of personal characteristics of politicians (e.g., ethnicity) on the likelihood of receiving hate speech.

Disparities across parties, gender, and ethnicity: Existing research suggest that political party leanings in the U.S. correlate with different speech patterns: Democrats tend to use more swear words and higher sentiment, while Republicans prefer to communicate more negative sentiment and group identity [52]. Besides party differences, a vast strand of studies has shown that there are discrepancies in communication behavior across genders. For instance, women are more likely to hide expressive and negative emotions [14], and are guided by a greater focus on care in moral dilemmas [39]. This is directly applicable to the domain of social media, where women are more likely to report messages targeting racial minorities and women [16]. Gender differences are further reinforced by widespread stereotypes regarding the role of women in society [42], who are perceived as less persuasive and are often outright dismissed when displaying aggressive and forceful behavior online [59]. Furthermore, survey studies suggest that women more often tend to be a target of cyber-bullying and hateful attacks [8], especially if they present an openly active stance, such as feminism [26]. Ethnicities and racial stereotypes play a similar role in offline and online discourse and differ greatly across countries [53]. For instance, for the U.S., existing studies suggest frequent hate speech against African Americans [33].

Research gap: Existing research on hate speech in the political discourse focuses either on qualitative insights or on summary statistics. We are not aware of previous works empirically modeling the effect of personal characteristics on the likelihood of a politician to receive hate speech. This presents our contribution.

3 DATASET

Members of the U.S. Congress: We analyze tweets from all 541 members of the 117th U.S. Congress that convened on January 3, 2021. Data on the members of Congress was gathered from the official webpage of the U.S. Congress [40], which provides links to personal and campaign web pages. By following these links, we collected the following information about each politician: (i) party affiliation, (ii) branch of Congress in which the politician serves, (iii) time served in Congress, (iv) gender, and (v) ethnicity. Fig. 1 provides an overview of the composition of the 117th U.S. Congress. Most voting seats are held by members of the two major political parties with 269 Democrats (D) and 263 Republicans (R), while 2 seats are occupied by independent senators. Women (W) hold 27% of all Congress Seats, accounting for 39% of all Democrats and 15% of all Republicans, respectively. Notably, the 117th U.S. Congress is the most ethnically diverse so far with 39% of Democrats and 8% of Republicans identifying as people of color (PoC).

For the sake of simplicity and interpretability, we focus our later empirical analysis on tweets from Republican and Democratic members; and exclude tweets from the two independent senators.

Collection of tweets: Twitter handles (user names) of every politician in the U.S. Congress are provided by the University of

California San Diego library [50]. We employed the Twitter Historical API to download the complete timelines of every politician between January 3, 2021 and the end of July 2021. Here we collected the entire tweet history of each person, excluding retweets and replies, resulting in a total number of 199,294 tweets. The average number of tweets per politician is 368.38. We additionally queried Twitter’s Historical API to gather the replies to every source tweet in our data set. To ensure feasibility, we restricted the data collection to up to 250 replies for each original tweet, starting with the earliest reply. The crawling process resulted in a total number of 8,362,555 replies.

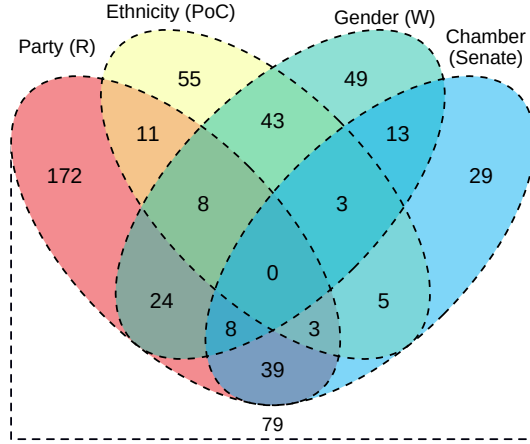


Figure 1: Venn Diagram visualizes the composition of the 117th U.S. Congress.

4 METHODS

4.1 Hate Speech Detection

In this work, we use machine learning to detect hate speech in replies to tweets. Compared to dictionary-based methods that merely count hate-related words [5], this approach is generally considered as being more accurate [6]. Nonetheless, as part of our robustness checks, we validate our results with the frequently-employed Hate-base dictionary [27], finding confirmatory results.

We implement machine learning for hate speech detection as follows: we employ the annotated twitter dataset from [13], containing 25,000 tweets labeled as hateful or not hateful. Each tweet was annotated by at least 3 users who were explicitly instructed to think about the context of the message and not only the words contained within [13]. We use the annotated tweets to implement a deep neural network classifier that predicts whether or not a tweet is hateful.¹ The hate speech classifier is then used to predict a binary label of whether or not a tweet is hateful (= 1 if true; otherwise = 0) for each reply tweet in our dataset. For each source tweet, we calculate the share of replies that are hateful. The resulting variable ranges from 0 to 1, with 0 indicating the lack of hate speech in replies, and 1 indicating that every reply is hateful.

¹We use Universal Sentence Encoder (USE) [11] as text representation. The machine learning classifier yields a weighted out-of-sample $F1$ score of 0.89, which is similar to previous works [13] and can be seen as reasonably accurate in the context of our study. The model is implemented in Python 3.8.5 using TensorFlow 2.6.0 [1].

4.2 Explanatory Regression Model

We implement a multilevel binomial regression to estimate the effects of party, gender, and ethnicity on the likelihood of a tweet receiving hate speech.

Formally, we model the number of hate speech replies, $HReplies$, as a binomial variable with probability parameter θ . The number of trials is given by the total number of replies a tweet receives ($Replies$). The key explanatory variables are the politicians’ party affiliation ($Party$; = 1 if Republican, otherwise 0), gender ($Gender$; = 1 if Man, otherwise 0), and ethnicity ($Ethnicity$; = 1 if Person of Color, otherwise 0). Furthermore, for each source tweet, we calculate a sentiment score ($SourceSentiment$) using SentiStrength. We also control for the congressperson’s age (Age), the number of years served ($YearsInOffice$), whether media was attached to the tweet ($AttachedMedia$; = 1 if true, otherwise 0), and the chamber of congress at which the politician serves ($Chamber$; = 1 if Senate, otherwise 0). Based on these variables, we specify the following regression model:

$$\begin{aligned} \text{logit}(\theta) = & \beta_0 + \beta_1 Party + \beta_2 Gender + \beta_3 Ethnicity \\ & + \beta_4 SourceSentiment + \beta_5 YearsInOffice + \beta_6 Age \\ & + \beta_7 AttachedMedia + \beta_8 Chamber \\ & + u_{\text{user}} + \varepsilon, \end{aligned} \quad (1)$$

$$HReplies \sim \text{Binomial}[Replies, \theta], \quad (2)$$

with intercept β_0 , error term ε , and user-specific random effects u_{user} . Note that the latter is important as it allows us to control for heterogeneity in users’ social influence (e.g., some accounts have many followers and reach different audiences) [43, 44].

We estimate Eq. 1 and Eq. 2 using MLE and generalized linear models. To facilitate the interpretability of our findings, we z -standardize all variables, so that we can compare the effects of regression coefficients on the dependent variable measured in standard deviations. Our regression analyses are implemented in R 4.0.5 using the `lme4` package [7].

5 EMPIRICAL ANALYSIS

5.1 Summary Statistics

We start our analysis by evaluating summary statistics. The average share of hateful replies per tweet in our dataset amounts to 1.99 %. We perform both t -tests and Kolmogorov-Smirnov (KS) tests to evaluate whether there are statistically significant differences across parties, genders, and ethnicities. Our findings are as follows: (i) tweets from Democrats (vs. Republicans) receive, on average, a 3.67% higher share of hate replies. (ii) Tweets from women (vs. men) politicians receive 7.71% higher share of hate replies. (iii) Tweets from persons of color (vs. whites) receive 37.75% higher share of hate replies. For each of these comparisons, two-sided t -tests confirm that the differences in means are statistically significant ($p < 0.01$). In Fig. 2, we visualize the complementary cumulative distribution functions (CCDFs) for the ratio of hate speech in replies. We again find that Democrats, women and persons of color receive more hate speech. KS-tests confirm that all differences in distributions are statistically significant ($p < 0.01$).

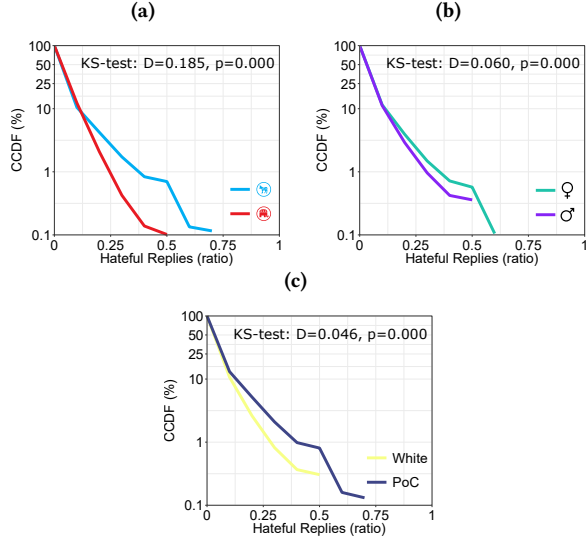


Figure 2: CCDFs for the ratio of hate speech in replies separated by (a) party, (b) gender, and (c) ethnicity.

5.2 Regression Analysis

We estimate a multilevel binomial regression to understand the effects of party affiliation, gender, and ethnicity on the likelihood of a tweet receiving hate speech (see model w/o interactions in Fig. 3). In contrast to summary statistics, this allows us to estimate effect sizes *after* controlling for confounding effects. The largest effect size is estimated for *Ethnicity* with a coefficient of 0.346 ($p < 0.01$), which implies that the odds of receiving hate speech for persons of color are $e^{0.346} \approx 1.41$ times the odds for whites. We further observe pronounced party and gender effects. Compared to Democrats, the odds for tweets from Republicans to receive hate speech are 22.02% higher ($\beta = 0.199$, $p < 0.01$). The odds for men to receive hate speech are 8.33% ($\beta = -0.087$, $p < 0.05$) lower than for women. We also find that a more negative sentiment in the source tweet is associated with more hate speech in replies. A one standard deviation increase in *SourceSentiment* is associated with a 25.99% ($\beta = -0.301$, $p < 0.01$) decrease in the odds of receiving hate speech. We find no statistically significant effects from a politician’s age, time in office, chambers, and media attachments.

We add interaction terms to test whether users react differently to gender, ethnicity, and sentiment depending on the party affiliation (see model w/ interactions in Fig. 3). Here we find a statistically significant interaction term between *Party* and *Ethnicity* ($\beta = -0.287$, $p < 0.01$). This implies that persons of color from the Democratic party have higher odds for receiving hate speech than persons of color from the Republican party. Furthermore, the strength of the association between sentiment in the source tweet and hate speech varies across parties ($\beta = 0.235$, $p < 0.01$). Specifically, negative sentiment attracts more hate speech for Democrats. The interaction between party affiliation and gender is not significant at common statistical significance thresholds.

Altogether, our analysis implies that three groups of politicians are particularly likely to receive hate speech in response to their

tweets: (i) persons of color from the Democratic party, (ii) white Republicans, and (iii) women.

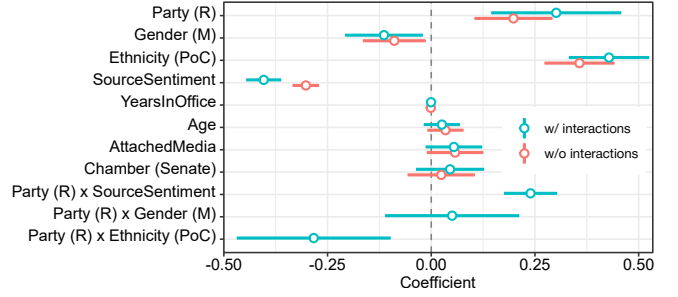


Figure 3: Coefficient estimates for binomial regression w/o (coral) and w/ (teal) interaction terms for political party. The horizontal bars represent 95% confidence intervals. User-specific random effects are included.

5.3 Robustness Checks

We conducted additional checks to validate the robustness of our analysis: (1) We repeated our analysis with a dictionary-based approach for hate speech detection, specifically the Hatebase dictionary [27]. (2) We calculated variance inflation factors for all independent variables in our regression model and found that all remain below the critical threshold of four. (3) We repeated our analysis with alternative estimators (e.g., beta regression), controlled for outliers, tested for quadratic effects, and added multiple interaction terms for each explanatory variable. In all cases, our results are robust and consistently support our findings.

6 DISCUSSION

Summary of findings: This work empirically models how the amount of hate speech in replies to tweets from politicians depends on their personal characteristics (party affiliation, gender, ethnicity). All else being equal, we find that Tweets are particularly likely to receive hate speech replies if they are authored by (i) persons of color from the Democratic party, (ii) white Republicans, and (iii) women. Furthermore, our analysis reveals that more negative sentiment (in the source tweet) is associated with more hate speech (in replies). However, the association varies across parties: negative sentiment attracts more hate speech for Democrats (vs. Republicans). Altogether, our empirical findings imply statistically significant differences in how politicians are treated on social media depending on their party affiliation, gender, and ethnicity.

Implications: Our findings are relevant both for politicians and from a societal perspective. Politicians should be aware that social media is a double-edged sword as it comes with the risk of receiving vast numbers of hate comments. This is concerning as hate speech can destroy reputations and may even lead to long-run mental health consequences [56]. Given that hate speech can affect peoples’ decision to participate in politics [48], this may also impede diversity in the composition of political institutions. Furthermore, hate speech goes hand in hand with increased polarization, hyper-partisanship, and less common ground between opposing political sides [20], thereby threatening the functioning of democracy itself.

REFERENCES

- [1] Martin Abadi, Ashish Agarwal, Paul Barham, et al. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. <https://www.tensorflow.org/>
- [2] Pushkal Agarwal, Oliver Hawkins, Margarita Amaxopoulou, Noel Dempsey, Nishanth Sastry, and Edward Wood. 2021. Hate speech in political discourse: A case study of UK MPs on Twitter. In *ACMHT*.
- [3] Sohail Akhtar, Valerio Basile, and Viviana Patti. 2020. Modeling annotator perspective and polarized opinions to improve hate speech detection. In *HCOMP*.
- [4] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31, 2 (2017), 211–236.
- [5] Ahlam Alrehili. 2019. Automatic hate speech detection on social media: A brief survey. In *AICCSA*.
- [6] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *WWW Companion*.
- [7] Douglas Bates, Deepayan Sarkar, Maintainer Douglas Bates, and L Matrix. 2021. *lme4*. <https://cran.r-project.org/web/packages/lme4/index.html> version 1.1.27.
- [8] Linda Beckman, Curt Hagquist, and Lisa Hellström. 2013. Discrepant gender patterns for cyberbullying and traditional bullying—An analysis of Swedish adolescent data. *Computers in Human Behavior* 29, 5 (2013), 1896–1903.
- [9] Anat Ben-David and Ariadna Matamoros Fernández. 2016. Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. *International Journal of Communication* 10 (2016), 27.
- [10] Susan Benesch. 2014. Defining and diminishing hate speech. *State of the World's Minorities and Indigenous Peoples* 2014 (2014), 18–25.
- [11] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv 1803.11175* (2018).
- [12] Shivang Chopra, Ramit Sawhney, Puneet Mathur, and Rajiv Ratn Shah. 2020. Hindi-english hate speech detection: Author profiling, debiasing, and practical perspectives. In *AAAI*.
- [13] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM*.
- [14] Teresa L. Davis. 1995. Gender differences in masking negative emotions: Ability or motivation? *Developmental Psychology* 31, 4 (1995), 660–667.
- [15] Tom de Smedt and Sylvia Jaki. 2018. The Polly corpus: Online political debate in Germany. In *Computer-Mediated Communication*.
- [16] Daniel M. Downs and Gloria Cowan. 2012. Predicting the importance of freedom of speech and the perceived harm of hate speech. *Journal of Applied Social Psychology* 42, 6 (2012), 1353–1375.
- [17] Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *ICWSM*.
- [18] Gunn Sara Enli and Eli Skogerbo. 2013. Personalized campaigns in party-centred politics: Twitter and Facebook as arenas for political communication. *Information, Communication & Society* 16, 5 (2013), 757–774.
- [19] Karmen Erjavec and Melita Poler Kovačić. 2012. “You don’t understand, this is a new war!” Analysis of hate speech in news web sites’ comments. *Mass Communication and Society* 15, 6 (2012), 899–920.
- [20] Eli J Finkel, Christopher A Bail, Mina Cikara, Peter H Ditto, Shanto Iyengar, Samara Klar, Lilliana Mason, Mary C McGrath, Brendan Nyhan, David G Rand, et al. 2020. Political sectarianism in America. *Science* 370, 6516 (2020), 533–536.
- [21] Deen Freelon, Alice Marwick, and Daniel Kreiss. 2020. False equivalencies: Online activism from left to right. *Science* 369, 6508 (2020), 1197–1201.
- [22] Tamara Fuchs and Fabian Schäfer. 2019. Normalizing misogyny: hate speech and verbal abuse of female politicians on Japanese Twitter. In *Japan Forum*.
- [23] Jason Gainous and Kevin M. Wagner. 2013. *Tweeting to power: The social media revolution in American politics*. Oxford University Press.
- [24] Jennifer Golbeck, Justin M. Grimes, and Anthony Rogers. 2010. Twitter use by the US Congress. *Journal of the American Society for Information Science and Technology* 61, 8 (2010), 1612–1621.
- [25] Todd Graham, Marcel Broersma, Karin Hazelhoff, and Guido Van’T Haar. 2013. Between broadcasting political messages and interacting with voters: The use of Twitter during the 2010 UK general election campaign. *Information, Communication & Society* 16, 5 (2013), 692–716.
- [26] Claire Hardaker and Mark McGlashan. 2016. “Real men don’t hate women”: Twitter rape threats and group identity. *Journal of Pragmatics* 91 (2016), 80–93.
- [27] Hatebase. 2021. A collaborative, regionalized repository of multilingual hate speech. <https://hatebase.org/>
- [28] Kenneth E Himma and Herman T Tavani. 2008. *The handbook of information and computer ethics*. John Wiley & Sons.
- [29] Sounman Hong, Haneul Choi, and Taek Kyu Kim. 2019. Why do politicians tweet? Extremists, underdogs, and opposing parties as political tweeters. *Policy & Internet* 11, 3 (2019), 305–323.
- [30] Sounman Hong and Daniel Nadler. 2012. Which candidates do the public discuss online in an election campaign? The use of social media by 2012 presidential candidates and its impact on candidate salience. *Government Information Quarterly* 29, 4 (2012), 455–461.
- [31] Nigel A. Jackson and Darren G. Lilleker. 2009. Building an architecture of participation? Political parties and Web 2.0 in Britain. *Journal of Information Technology & Politics* 6, 3–4 (2009), 232–250.
- [32] James A. Piazza. 2020. Politician hate speech and domestic terrorism. *International Interactions* 46, 3 (2020), 431–453.
- [33] Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *AAAI*.
- [34] Anders Olof Larsson. 2015. Pandering, protesting, engaging. Norwegian party leaders on Facebook during the 2013 ‘Short campaign’. *Information, Communication & Society* 18, 4 (2015), 459–473.
- [35] Helen Margetts, Peter John, Scott Hale, and Taha Yasseri. 2015. *Political turbulence: How social media shape collective action*. Princeton University Press.
- [36] Mainack Mondal, Leandro Araújo Silva, and Fabricio Benevenuto. 2017. A measurement study of hate speech in social media. In *ACMHT*.
- [37] Zewdie Mossie. 2020. Social media dark side content detection using transfer learning emphasis on hate and conflict. In *WWW Companion*.
- [38] Seema Nagar, Sameer Gupta, C. S. Bahushruth, Ferdous Ahmed Barbhuiya, and Kuntal Dey. 2021. Empirical assessment and characterization of homophily in classes of hate speeches. In *AAAI Workshop on Affective Content Analysis*.
- [39] Nhung T. Nguyen, M. Tom Basuray, William P. Smith, Donald Kopka, and Donald McCulloh. 2008. Moral issues and gender differences in ethical judgment using Reidenbach and Robin’s (1990) multidimensional ethics scale: Implications in teaching of business ethics. *Journal of Business Ethics* 77, 4 (2008), 417–430.
- [40] Library of Congress. 2021. Members of the U.S. Congress. <https://www.congress.gov/members>
- [41] Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush Varshney. 2018. The effect of extremist violence on hateful speech online. In *ICWSM*.
- [42] Deborah A. Prentice and Erica Carranza. 2003. Sustaining cultural beliefs in the face of their violation: The case of gender stereotypes. In *The Psychological Foundations of Culture*. Psychology Press, 268–289.
- [43] Nicolas Pröllochs, Dominik Bär, and Stefan Feuerriegel. 2021. Emotions explain differences in the diffusion of true vs. false social media rumors. *Scientific Reports* 11, 22721 (2021).
- [44] Nicolas Pröllochs, Dominik Bär, and Stefan Feuerriegel. 2021. Emotions in online rumor diffusion. *EPJ Data Science* 10, 1 (2021), 51.
- [45] Nicolas Pröllochs, Dominik Bär, and Stefan Feuerriegel. 2022. Community-based fact-checking on Twitter’s Birdwatch platform. In *ICWSM*.
- [46] Karen Ross, Susan Fountaine, and Margie Comrie. 2015. Facing up to Facebook: politicians, publics and the social media (ted) turn in New Zealand. *Media, Culture & Society* 37, 2 (2015), 251–269.
- [47] Punyajoy Saha, Binny Mathew, Kiran Garimella, and Animesh Mukherjee. 2021. “Short is the road that leads from fear to hate”: Fear speech in Indian WhatsApp groups. In *WWW*.
- [48] Jennifer Scott. 2019. Women MPs say abuse forcing them from politics. *BBC News* (2019).
- [49] Andrew Sellars. 2016. Defining hate speech. *Berkman Klein Center Research Publication* 2016-20 (2016), 16–48.
- [50] Kelly L. Smith. 2021. LibGuides: Congressional Twitter accounts. https://ucsd.libguides.com/congress_twitter
- [51] Kirill Solovev and Nicolas Pröllochs. 2022. Moral emotions shape the virality of COVID-19 misinformation on social media. In *WWW*.
- [52] Karolina Sylwester and Matthew Purver. 2015. Twitter language use reflects psychological differences between democrats and republicans. *PLOS ONE* 10, 9 (2015).
- [53] Beverly Daniel Tatum. 2017. *Why are all the Black kids sitting together in the cafeteria? And other conversations about race*. Hachette UK.
- [54] Yannis Theocharis, Pablo Barberá, Zoltán Fazekas, and Sebastian Adrian Popa. 2020. The dynamics of political incivility on Twitter. *Sage Open* 10, 2 (2020).
- [55] Yannis Theocharis, Will Lowe, Jan W. van Deth, and Gema García-Albacete. 2015. Using Twitter to mobilize protest action: Online mobilization patterns and action repertoires in the Occupy Wall Street, Indignados, and Aganaktismenoi movements. *Information, Communication & Society* 18, 2 (2015), 202–220.
- [56] Bertie Vidgen, Emily Burden, and Helen Margetts. 2021. *Understanding online hate: VSP regulation and the broader context*. Ofcom.
- [57] Angelia Wagner. 2020. Tolerating the trolls? Gendered perceptions of online harassment of politicians in Canada. *Feminist Media Studies* (2020), 1–16.
- [58] Maximilian Wich, Melissa Breitingier, Wienke Strathern, Marlena Naimarevic, Georg Groh, and Jürgen Pfeffer. 2021. Are your friends also haters? Identification of hater networks on social media: Data Paper. In *WWW Companion*.
- [59] Julia Winkler, Annabell Halfmann, and Rainer Freudenthaler. 2017. Backlash effects in online discussions: Effects of gender and counter-stereotypical communication on persuasiveness and likeability. *International Communication Association*.
- [60] Savvas Zannettou, Barry Bradlyn, Emiliano de Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. 2018. What is Gab: A bastion of free speech or an alt-right echo chamber. In *WWW Companion*.