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.proellochs@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

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
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 Hatebase 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. 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.

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 \(\theta\). 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:

\[
\logit(\theta) = \beta_0 + \beta_1 Party + \beta_2 Gender + \beta_3 Ethnicity + \beta_4 SourceSentiment + \beta_5 YearsInOffice + \beta_6 Age + \beta_7 AttachedMedia + \beta_8 Chamber + \mu_{user} + \epsilon, \tag{1}
\]

\(HReplies \sim Binomial[Replies, \theta] \tag{2}\)

with intercept \(\beta_0\), error term \(\epsilon\), and user-specific random effects \(\mu_{user}\). 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 (\(p < 0.01\)).
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.

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.
