Understanding Messages to Underrepresented Racial, Ethnic, Gender, and Sexual Groups on Social Media by Democratic Politicians and their Electoral Implications

Henry Smith

Google Summer of Code Final Project Report

1 Research Overview:

Social media continues to proliferate in the United States as a popular mode of communication, oftentimes replacing face-to-face interactions. Responding to this societal paradigm, politicians have turned to social media in order to communicate with, and ultimately garner the support of, their constituents. While many have already critically examined politicians' use of social media, my Google Summer of Code Project considered specifically how Democratic politicians make appeals to groups of underrepresented voters. Nearly half of the Democratic party is made up of voters belonging to one or more underrepresented racial or ethnic groups. Moreover, in recent elections, the LGBTQ+ community has emerged as one of the most solid Democratic voting blocs. A complete project proposal and research index justifying the project in greater depth are available in the project Github repository.

The dataset used for the project consists of Facebook images and captions ($N = 65,355$) posted during the 2018 U.S. election cycle from the profiles of politicians who won their primary elections. Images were collected from January 1, 2018 to a month following election day (December 6, 2018). A corresponding dataset was constructed including the portrait photographs of politicians whose Facebook images were collected.

From the data collected prior to election day, a semi-random[1] sample of Facebook images and captions were collected separately from Democratic ($n_{D1} = 3,264$) and Republican ($n_{R1} = 2,910$) politicians. For each post included in the two samples, we wanted to measure whether or not the politician intended to appeal to **four underrepresented voter groups**: Black, Hispanic, Asian, and LGBTQ+ voters. Each of these samples were annotated according to the following scheme by U.S. workers on Amazon Mechanical Turk, an online crowdsourcing service. Annotators were simultaneously presented with the Facebook post (i.e. image/caption pair) as well as the politician's portrait image. These individuals were instructed that the politician's race/ethnicity alone should not determine whether an image appeals to a certain underrepresented group.

[1] Sampling techniques are thoroughly described in the project blog

Following completion of the manual annotations, Dr. Kunwoo Park, a postdoctoral researcher at the University of California, Los Angeles Department of Communications, constructed an image-based residual neural network and simply-connected text-based neural network by which to predict appeal to the four measured voter groups from the Facebook images and captions, respectively. These networks were built from pretrained models and subsequently trained on the Amazon Mechanical Turk annotations from the two semi-random samples. The performance of the models on both the training and test sets were assessed using the F1 score. Having achieved acceptable performance, the models were then inferred on the entire Facebook dataset. Predicted scores (0-1) from the text and image-based models for each appeal variable (coded 'Black_appeal, 'Hispanic_appeal', 'Asian_appeal', 'LGBTQ_appeal') were averaged.

So to improve the performance of the machine learning models, two additional samples of Democratic ($n_{D2} = 3,886$) and Republican ($n_{R2} = 3,759$) images were collected and thereafter annotated using Amazon Mechanical Turk. This time, however, images were sampled semi-uniformly by predicted score for each appeal variable. Provided that the distributions of predicted 'Black_appeal', 'Hispanic_appeal', 'Asian_appeal', and 'LGBTQ+_appeal' scores were each strongly right-skewed for both Democratic and Republican Facebook images/captions, this uniform sampling technique achieved a greater positive annotation rate across all four variables compared to when the images were randomly sampled. Once again, the text and image-based machine learning models were trained on these annotations and inferred on the entire Facebook dataset. Provided the models attained very high performances, particularly for 'Black_appeal', these predicted scores were then used for analysis.

2 Technical Analysis of Appeal:

Upon inferring the image and text-based machine learning models on the entire dataset of $N = 65,355$ Facebook images and captions, respectively, the predicted scores, averaged across the image and text models, were analyzed according to the following statistical techniques. Implementation of these techniques can be found in the score analysis pipeline.

2.1 Political Party Appeal Comparison:

Although Republican politicians' appeal to underrepresented voter groups is not the primary focus of our research, we would like to establish a point of comparison between Democrats and Republicans. To do so, we consider $(X_1, Y_1), ..., (X_l, Y_l)$, where each (X_i, Y_i) $(1 \leq i \leq l)$ represents a pair consisting of the **mean** inferred [Black/Hispanic/Asian/LGBTQ+][2] appeal scores across all Facebook posts shared by the Democratic (X_i) and Republican (Y_i) politicians participating in the ith general election competition. Only races in which there were both a Democratic and a Republican competitor are considered.

[2]The phrasing "[Black/Hispanic/Asian/LGBTQ+]" indicates that the mentioned procedure was performed separately for each of the four underrepresented groups

Subsequently, we examine the distribution of $X_1 - Y_1, ..., X_l - Y_l$ and test $H_0 : \tilde{\mu} = 0$ against $H_A : \tilde{\mu} > 0$, where $\tilde{\mu}$ is the median of the underlying distribution $S = X_j - Y_j$ for $\forall j$. A matched design is used to account for the positive covariance in appeal scores between politicians competing in the same district/state (i.e. politicians in pair i have the same constituency, similar donation/endorsement opportunities, etc.). Provided we should not assume that S is normally distributed, a non-parametric test is used, namely the Wilcoxon signed-rank test. This test assumes that S is symmetrically distributed, and the test statistic W is asymptotically normally distributed under H_0. A large value of W provides evidence that Democratic politicians share Facebook posts with a greater predicted appeal to [Black/Hispanic/Asian/LGBTQ+] voters than do the Republicans running against them.

2.2 Predicting Democratic Appeal:

Beyond understanding differences in Democrat-Republican appeal to underrepresented voters, we also seek to explain intra-party discrepancies: which Democratic politicians are more likely to make appeals to these voters? To answer this question, we construct a linear model with response variable $Y = Y_1, ..., Y_n$, the average predicted [Black/Hispanic/Asian/LGBTQ+] appeal score across all Facebook posts shared by each Democratic politician i ($1 \leq i \leq n$). Alternatively, we can use the proportion of predicted [Black/Hispanic/Asian/LGBTQ+] appeal images, calculated using the cutoff score value that achieved the greatest F1 score in the machine learning models. To predict politician appeal, we use the proportion of [Black/Hispanic/Asian] constituents[3] in politician i's state (governor and senate candidates) or district (congressional candidates), $x_{11}, ..., x_{n1}$ as well as whether politician i is [Black/Hispanic/Asian], $x_{12}, ..., x_{n2}$.

Moreover, we address concerns of multicollinearity between predictors by observing the sample covariance matrix and variance inflation factors prior to regression. Once the maximum likelihood estimators $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ are calculated for the linear model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon$, $1 \leq i \leq n$, fit is assessed by examining the residuals, which are used to approximate the error ϵ. The linear regression model assumes that the variance, σ^2, of the error in approximation, ϵ, is constant for all values of the predictors, X_1, X_2, and that these $\epsilon \sim N(0, \sigma^2)$ for fixed $(X_1, X_2) = (x_1, x_2)$.

2.3 Predicting Election Outcome

Rather than inferring Democratic politician appeal, we now implement a logistic regression model to predict a Democratic politician's election outcome (W/L) from their use of [Black/Hispanic/Asian/LGBTQ+] appeals, X_1; X_1 is measured either by average predicted score or predicted proportion of appeals. In addition to a Democratic politician's appeal, we also integrate the proportion of [Black/Hispanic/Asian] constituents in their electoral district or state, X_2, and whether or not the politician is [Black/Hispanic/Asian], X_3, as predictors in the model. To minimize the impact of multicollinearity, we implement four separate regression models corresponding to the four appeal variables.

[3] 2015 U.S. Census Bureau estimates

The logistic regression model $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$, $1 \leq i \leq n$ is fit to $Y_1, ..., Y_n$ using maximum likelihood estimation, where $Y_i \sim$ Bernoulli(p_i), $p_i \in (0, 1)$. The appropriateness of the model is considered by inspecting the relationship between each predictor and the log-odds ratio. Additionally, the Pearson residuals, calculated $\frac{Y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$, should be distributed with mean = 0, variance = 1 in a well-fit logistic model, and there should be no observable pattern between the predictors and residuals.

2.4 Temporal Analysis

Furthermore, the timing of when Democratic politicians choose to make appeals to underrepresented racial, ethnic, sexual, and gender voter groups through Facebook may be strategic. First, we consider the average (i.e. normalized) predicted [Black/Hispanic/Asian/LGBTQ+] appeal score of Facebook posts shared by all Democratic politicians grouped by month and day. Through these visualizations, we consider when Democratic politicians are more likely to post images appealing to a specific underrepresented voter group, paying particular attention to Black History Month (February), Hispanic Heritage Month (September-October), Asian American and Pacific Islander Heritage Month (May), and LGBTQ+ Pride Month (June).

Democratic politicians may also exhibit different strategies related to voter engagement among underrepresented populations during each the primary and general election campaign. Considering that the median voter participating in the Democratic primary election is likely much more liberal and diverse than the median voter in the general election, we hypothesize that Democratic candidates share images appealing to the four measured underrepresented groups with a greater probability during the primary election campaign than during the general campaign.

In order to evaluate this hypothesis, we consider two pairs of one-month periods: the first one month prior to and one month following the politician's primary election date, and the second one month prior to and one month following the general election date (November 6, 2018). For each of these two pairs, we let $(X_1, Y_1), ..., (X_n, Y_n)$ represent politician i's ($1 \leq i \leq n$) average predicted [Black/Hispanic/Asian/LGBTQ+] appeal score of Facebook images/captions shared during the periods one month prior to, X_i, and after, Y_i, the election.

Similar to the Democratic-Republican appeal comparison, we then utilize a paired design from $Y_1 - X_1, ..., Y_n - X_n$ to test $H_0 : \tilde{\mu} = 0$ versus $H_A : \tilde{\mu} < 0$ (primary election pair) and $H_A : \tilde{\mu} > 0$ (general election pair), where $\tilde{\mu}$ is the median of the underlying distribution S ($S = Y_j - X_j$ for $\forall j$). In the case that the underlying distributions of X and Y are each normally distributed, the paired T-statistic is used; if not, the nonparametric Wilcoxon signed-rank statistic assumes only that the distribution of S is symmetric. Both statistics are asymptotically normally distributed and reject the null hypothesis for small values in the primary election pair or for large values in the general election pair.

3 Acknowledgements:

The project "Understanding Messages to Underrepresented Racial, Ethnic, Gender, and Sexual Groups on Social Media by Democratic Politicians and their Electoral Implications" is contributed by Henry Smith with Red Hen Lab. It was completed under the guidance of Jungseock Joo, PhD, and in collaboration with Kunwoo Park, PhD and Danni Chen.

A paper detailing the project is forthcoming.

Any questions or comments regarding the project can be emailed to henry.smith@yale.edu.

4 Research Resources:

The following project resources are available to view:

- GitHub page:
    1. Project repository
    2. Project blog
- Research proposal
- Review of research related to politician appeal to underrepresented racial, ethnic, sexual, and gender groups
- Facebook image and caption annotation scheme
- Python inferred score analysis pipeline