Running Influence Campaigns on Twitter

Qi Yang
Massachusetts Institute of Technology
Cambridge, MA 02139
yangqi@mit.edu

Khizar Qureshi*
Massachusetts Institute of Technology
Cambridge, MA 02139
kqureshi@mit.edu

Tauhid Zaman
Yale School of Management
New Haven, CT 06511
tauhid.zaman@yale.edu

Abstract

We conduct a field experiment on Twitter and study the effect of race, opinion, and closeness on persuasion. Our goal is to persuade users against immigration to be more open towards immigration. We find that "pacing" users with a similar political stance before exposing them to opposing views pushes them further away from the desired outcome, and induces a backfire effect. However, opinion pacing while closely interacting with the users helps mitigate the backfire effect. The results hold for both white and brown bots, with a higher significance for the former. Despite several limitations, this study has important implications for the emerging field of computational social science and ongoing efforts to reduce political polarization online.

1 Introduction

Social media has become an increasingly important avenue of political discourse in the modern world, and Twitter has emerged as a leading platform. Because tweets and interaction are visible to the public by default, the usage of Twitter as a platform by social scientists studying influence has recently increased.

The rise of ISIS and the beginning of the European refugee crisis in 2015 brought migration to the forefront as a political issue. Consequently, several studies have used sentiment analysis (Öztürk and Ayvaz [2018], Backfried and Shalunts [2016], Coletto et al. [2016]) to examine opinion on the refugee crisis on Twitter, but few have considered experimentally attempting to influence Twitter users. We synthesize the topics of opinion analysis and peer influence and build upon the results of this earlier work.

We consider to what degree the opinions of Twitter users can be influenced by mutual followers. Often in political campaigns, proponents of each division campaign for influence, and actively "pull" their targets towards a common consensus, typically their own. An instance of this is "pacing and leading", holistically reviewed by O'Connor [1990, 1994], in which the influencer provides verifiable statements (pace), followed by slight suggestions (lead). While such techniques are often successful, literature suggests that there is often a backfire effect (Bail [2015], Lord [2015], Nyhan [2010]), polarizing the targets' opinions further, if the targets' political stance is opposite of that the political campaign. In this study, we want to compare (i) the traditional way of influencing–directly exposing users to opposing views and (ii) the pacing influence campaign method. The latter involves (i) sharing

*Current: Millennium Management, New York City, New York 10103

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

a similar political stance with the targets, then (ii) slowly shift to the other political camp, and (iii) observe whether the targets (a) follow the bot or (b) backfire.

We report the results of a systematic experiment in which we use four bots (influencers) to interact with and collect data from a subset of Twitter users who have been active in discussing immigration policies and are anti-immigration. Specifically, we consider both White and Brown bots, and while one set of bots begins with the opposite stance – pro-immigration, another set of bots shares the same initial (anti-immigration) sentiment as the target users, intending to gradually shift their opinion over time. The second set of bots present themselves initially as members of a shared ideological and social group and then gradually shift to the ideological out-group. Furthermore, we randomly assign half of the users a state of "closeness", in which a bot signals friendship through means of liking statuses to test whether social influence will be stronger for those with closer relationships. Through this, we draw conclusions on the effectiveness of peer influence on Twitter in changing opinions on a politically heated topic; and in how peer influence varies in social media based on the level of shared group identity and social rapport between users.

2 Experiment Design In a Nutshell

We wished for anti-immigration users to interact with so it was essential to have an active subset of anti-immigration users. To do this we began by constructing a list of both hashtags and keywords that conveyed anti-immigration sentiment. We scoured twitter, and found hashtags that were anti-immigration like #CloseThePorts, #BanMuslim, #BuildTheWall, etc.. To avoid translation issues, we limit tweets only in English.

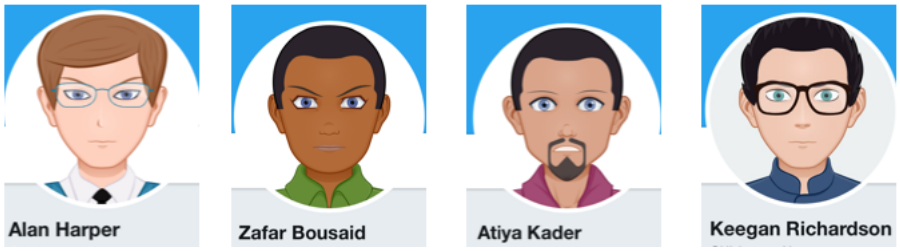

Figure 1: Profile pictures and names of the four bots (pacing/white, pacing/brown, no-pacing/brown, no-pacing/white).

To test if race would affect the result, we varied the race of the 4 bots we created, as in figure 2. We gave two of them an avatar of a white male and a traditional European name, and two of them an avatar of a Middle Eastern male and a traditional Arabic or Persian name (based on most European immigrants coming from Syria or Iran). We used cartoon avatars for the profile pictures since if we had used real photos, there would exist the possibility that the particular people pictured varied on some dimension other than race.

The bots will then start to follow the users we identified as anti-immigration people to gain follow backs. We made sure that no two bots were following the same target as this could arouse suspicion. After the targets follow back, we initialize two bots with anti-immigration stance and two with pro-immigration. But they will not clarify their stance until we gathered all followers and started the experiment. Moreover, we randomly assign half of the follow back targets to closeness conditions. Overall, we have four bots: a White no-pacing bot with initial pro-immigration stance, a White pacing bot with initial anti-immigration stance, a Brown no-pacing bot with initial pro-immigration stance, and a Brown pacing bot with initial anti-immigration stance.

We then observe how our targets interact with our tweets. We record whether they remain friends with our bots despite bots tweeting more pro-immigration materials. We also take note of whether they retweet or like any of the pro-immigration material we posted opposing their original opinions. Lastly, we measure the change of usage of classifying keywords such as "illegals" to approximate how much their attitude has changed. Based on the literature, we hypothesize that the pacing bot will be the most influential in changing users' opinion, especially for those who the bot interacts with closely.

3 Data

We fetched all data in an automated fashion using multi-processing across eight cores with 64 GB virtual memory. Using the set of users "so far" as a state, we updated the queue of users after every run. For each user, we flushed the buffer out in order to handle the excess memory capacity. This process varied by user according to the content on their wall, and totaled 28 calendar days. Our exception handling suggested there were three reasons why a user's data was not available: (i) privacy settings, (ii) account deletion by user, (iii) account suspension by Twitter. Whenever an attempt for a given user failed, we re-added the user to the tail of the queue, allowing for a maximum of ten attempts.

3.1 Interaction data

Interaction data was collected via a cron job at midnight eastern time daily, and cached on a user and bot level. On a daily basis, we recorded the set of all users following each bot. This allowed us to create a transition matrix of friendship across time, where a change in state between consecutive dates for a given user would indicate either a follow/unfollow/suspension. On a tweet level, we recorded actions including likes, retweets, comments, and mentions for each bot.

3.2 Feature/tweet data

In addition to interaction data, we collected (i) features for every user in every group, (ii) features for every tweet. Specifically, for every user, we collect number of followers, number following, number of likes, and number of status updates. For every tweet (retweets excluded), we collect the number of retweets and likes. The feature summary tables for these can be found in table 3.

4 Key Results and Discussion

Using the Twitter API, we obtained all tweets with anti-immigration hashtags used in a month, which included #BanMuslims, #BuildTheWall, etc. and this results in 149,304 total tweets. After checking for uniqueness users across tweets and across dates, we attained 38,981 unique users who tweeted with anti-immigration hashtags. To make sure they were active users, since we would want to observe reactions in a weekly span, we further filtered out users whose tweeting frequency was less than a week and were left with 12,468 unique users. We successfully follow 12,182 users who were not banned or had disabled their accounts at the time of following. Across four bots, 2289 users followed back and continued to follow back before the start the experiment, with a follow-back rate 19.4%. This follow-back corroborated with our expectations of literature, which suggested 14% follow-back rate (Rajagopalan [2016]).

Bot	Identify	Follow Backs
Alan Harper	White	636
Keegan Richardson	White	717
Atiya Kader	Brown	454
Zafar Bousaid	Brown	482

We see that white bots have a higher follow-back number than brown bots. At this point, the bots have not taken a stance on immigration related topics yet, and so users choose to follow back only based on bots' profiles, which differ only in race. This also proves that we targeted the right users who are biased toward immigrants. Because users self-select to follow back the bots, there might be a difference in user identity between the white bots and the brown bots. For instance, brown bots' followers might be inherently less anti-immigration because they choose to follow back a brown user on Twitter.

Overall, we do not see a significant difference in follow rate across bots and across conditions. First, we ruled out the possibility that we did not reach users who remain following since we checked their tweeting frequency, and verified that they are active users on Twitter. There are a few plausible explanations. It could be that the cost of unfollowing was higher than simply ignoring the bot, given our bot did not spam. Otherwise, another possibility is that after following, users felt obligated to remain in the friendship and not unfollow.

Bot	Close Follow Rate	Standard Follow Rate
White Pacing	91%	93%
White No-Pacing	94%	92%
Brown Pacing	91%	89%
Brown No-pacing	94%	94%

Moreover, we want to evaluate whether there is any sentiment change among the users we target. We use the term "illegals" to infer anti-immigration sentiment, which is often seen as dehumanizing and hence the common rallying cry heard among immigrant rights activists: "No one is illegal". We calculated the frequency of the usage of word for each user before and after the experiment to check if there is any difference among users following different bots under different conditions.

In our study, we wanted to estimate the causal effect on users who follow back, since influence only happens among friends on Twitter. We used a two factorial model, with one variable belonging to {pacing, no-pacing},and the other one pertaining to closeness {close, standard}.

In our study, we randomly assign users to pacing/no-pacing conditions as well as closeness conditions after they follow back. Let, $t \in T$ denote a treatment within the set of treatments used. The unconfoundedness assumptions for causal interpretation hold such that the treatment indicator W_t is independent of the respective potential outcome Y_t given the pre-treatment variables X.

Assumption 1 $W_t \perp\!\!\!\perp Y_t|X, \forall t \in T$

Since the assumption holds, the the expected value of Y_t can be estimated by adjusting for X:

$$E(Y_t|X) = E(Y_t|W_t = 1, X) = E(Y|W_t = 1, X) = E(Y|T = t, X) \tag{1}$$

and thus

$$E(Y_t) = E[E(Y_t|X)] = \mu_t. \tag{2}$$

We used a multiple linear regression to predict the change in usage of anti-immigration words. We collected features such as number of friends, number of followers, number of friends following the same bot, number of friends following another bot for each user and later controlled for these variables in our model. The regression equation is:

$$Y = \alpha + \beta_1 Close + \beta_2 Pacing + \beta_3 Close \times Pacing + \beta_4 Covariates + \epsilon \tag{3}$$

Table 6 contains the regression results for the White bots. We see a similar result for the Brown bots, but with lower significance. The table below contains the coefficient and estimated everage for each treatment.

Treatment	Coefficient	$\hat{\mu}$ (%)
no pace/not close	$\hat{\alpha}$	0.13
no pace/close	$\hat{\alpha} + \hat{\beta}_1$	0.31
pace/not close	$\hat{\alpha} + \hat{\beta}_2$	0.34
pace/close	$\hat{\alpha} + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$	0.12

Table 1: Estimated averages for each treatment level

The baseline outcome is 0.13%, which is positive due to two potential reasons. First, there was an anti-immigration movement related to the construction of the border wall during the time the experiment. Second, there has been a backfire effect when directly exposing users to opposing views. We see that the pacing:close treatment has the lowest estimated average in increase of anti-immigration word usage, leading to the lowest increase in polarization.

4.1 Treatment effects

The effects of each treatment can be found in the table below. The main effect of pacing is significant at the $(p = 0.097)$ level. Compared with the baseline, there was a 27% increase in usage of the word "illegals". Additionally, the interaction effect equals half the interaction term and is significant at the $(p = 0.034)$ level, implying (i) the effect of pacing also depends on closeness, and (ii) pacing when closely interacting with targets helps offset the increase in polarization.

Treatment	Calculation	$\hat{\tau}$
close	$\hat{\beta}_1 + \frac{1}{2}\hat{\beta}_3$	0.005
pacing	$\hat{\beta}_2 + \frac{1}{2}\hat{\beta}_3$	0.035
interaction	$\frac{1}{2}\hat{\beta}_3$	-0.175

Table 2: Effects of each treatment

5 Conclusion

We find that (i) pacing induces a larger backfire effect compared with directly exposing targets to opposing views, (ii) the effect of pacing also depends on closeness, which can help mitigate the backfire and polarization.

References

Gerhard Backfried and Gayane Shalunts. Sentiment analysis of media in german on the refugee crisis in europe. In *International Conference on Information Systems for Crisis Response and Management in Mediterranean Countries*, pages 234–241. Springer, 2016.

C. Bail. Terrified: How anti-muslim fringe organizations became mainstream. *Princeton University Press*, 2015.

Mauro Coletto, Andrea Esuli, Claudio Lucchese, Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego, and Chiara Renso. Sentiment-enhanced multidimensional analysis of online social networks: Perception of the mediterranean refugees crisis. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1270–1277. IEEE Press, 2016.

Ross L. Lepper M.R. Lord, C.G. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Princeton University Press*, 37(11):2098–2109, 2015.

Reifler J. Nyhan, B. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, 2010.

Seymour J. O'Connor, J. *Introducing Neurolinguistic Programming.* Aquarian Press, London, UK, 1990.

Seymour J. O'Connor, J. *Training with NLP.* Aquarian Press, London, UK, 1994.

Nazan Öztürk and Serkan Ayvaz. Sentiment analysis on twitter: A text mining approach to the syrian refugee crisis. *Telematics and Informatics*, 35(1):136–147, 2018.

Krishnan Rajagopalan. Interacting with users in social networks: The follow-back problem. Technical report, MASSACHUSETTS INST OF TECH LEXINGTON LEXINGTON United States, 2016.

6 Appendix

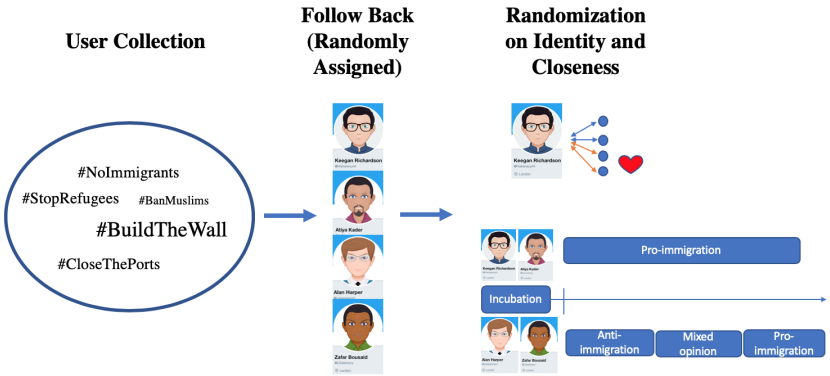

Figure 2: Experiment design

target users		
	retweets	favorites
count	8175	8175
mean	4655.208073	9413.289297
std	73470.34481	170748.9158
min	0	0
0%	0	0
10%	1	10
20%	4	24
30%	9	45
40%	21	91
50%	47	198
60%	117	477.4
70%	283	1129.6
80%	755.2	2679
90%	2597.6	7771.6
max	3807484	10335892

Table 3: Contains tweet statistics per feature.

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	0.0013	0.001	1.471	0.142	-0.000	0.003
Close	0.0018	0.001	1.584	0.113	-0.000	0.004
Pacing	0.0021	0.001	1.660	0.097 *	-0.000	0.005
Close:Pacing	-0.0035	0.002	-2.122	0.034 **	-0.007	-0.000
followers_count	2.89e-09	8.93e-08	0.032	0.974	-1.72e-07	1.78e-07
friends_count	-5.505e-08	1.4e-07	-0.392	0.695	-3.31e-07	2.21e-07
spillover_same_bot	6.999e-05	6.05e-05	1.157	0.247	-4.87e-05	0.000
spillover_other_bots	-1.422e-05	2.53e-05	-0.561	0.575	-6.39e-05	3.55e-05

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$