Mitigating the Backfire Effect Using Pacing andLeading

Tauhid Zaman (tauhid.zaman@yale.edu)
Yale University
Qi Yang
Massachusetts Institute of Technology
Khizar Qureshi
Massachusetts Institute of Technology

Research Article

Keywords: social networks, social media, political polarization , computational social science

Posted Date: March 2nd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1402967/v1

Mitigating the Backfire Effect Using Pacing and Leading

Qi Yang[1], Khizar Qureshi[2], Tauhid Zaman[3]

[1] Massachusetts Institute of Technology, Cambridge, MA 02139, USA, yangqi@mit.edu,
[2] Massachusetts Institute of Technology, Cambridge, MA 02139, USA, kqureshi@mit.edu,
[3] Yale University, New Haven , CT 06520, USA tauhid.zaman@yale.edu,

Abstract. Researchers have found that individuals in online social networks become more resolute in their beliefs when shown opinions opposite their own, a phenomenon referred to as the *backfire effect*. In this work we present a persuasion technique we call *pacing and leading*, which can mitigate the backfire effect. This dynamic technique has one gradually evolve their opinion over time to initially pace the persuasion target, and then lead them to the desired opinion. To test this technique, we conduct a field experiment in Twitter using artificial bot accounts where the goal is to persuade users who oppose immigration to support it. Neural network based sentiment analysis of the experimental subjects' tweets identifies a strong negative sentiment for tweets containing the word "illegals". Based on this, we use the frequency of tweets containing illegals as a measure of the treatment effect. We find that pacing and leading is more effective than simply presenting an opposing view, which we refer to as *arguing*, which actually results in a backfire effect. Our results suggest that dynamic persuasion techniques can be more effective than static approaches which present a constant opinion.

Keywords: social networks, social media, political polarization , computational social science

1 Introduction

When one wants to advocate for a position, point of view, or opinion, one typically presents arguments in favor of the desired opinion. This is a standard persuasion technique which we refer to as *arguing*. Classic opinion dynamics models indicate that arguing will result in consensus, i.e. the persuasion targets will eventually hold the same opinion [8]. More modern models state that persuasion can only occur if the desired opinion is within the confidence bound of the target, otherwise, the opinion of the target will remain unchanged. [12]. These bounded confidence models suggest that there are limitations to the arguing persuasion technique. Persuasion campaigns that are aimed at people whose opinions are in opposition to the desired opinion would have no effect at all.

Empirical studies have shown that reality is more complex than the classic models or bounded confidence models. Researchers have found that when presented with opposing opinions, people become more steadfast in their initial belief. There is no movement toward the desired opinion, as predicted by classic models. The target opinion does not remain unchanged, as predicted by bounded confidence models. Instead, the target opinion moves away from the desired opinion. The resulting effect is that a persuasion campaign leaves the targets in greater opposition to the desired opinion than if there had been no campaign conducted. This phenomenon is referred to as the *backfire effect* and has been demonstrated in multiple experiments [13],[17],[2].

The backfire effect is of greater concern in online social networks. In these networks, edges are constructed by individuals choosing with whom to connect. The result is that network neighbors very often share similar opinions, which is known as homophily [3]. Experiments have shown that the formation process of these networks are heavily influenced by homophily. One study found that Twitter users formed network connections with users of opposing political opinion at a rate three times lower than with co-partisans [14]. As a result, these networks typically consist of echo chambers where there is opinion homogeneity and contrasting opinions are not common. Users in these environments do exhibit the backfire effect, as shown in a study conducted in Twitter [2]. In this work, Twitter users of different political partisanships were exposed to counter-partisan content from an automated Twitter account, which we refer to as a bot. It was found that after repeated exposure to the bot content, the initial partisanship of the subjects had grown stronger. In other words, the persuasion campaign of the bots had pushed the subjects' opinions away from the counter-partisan point of view.

The presence of the backfire effect means that one must be careful when designing persuasion campaigns. One needs to present opinions to targets in a manner that does not push their opinions in the opposite direction. One approach to this was shown in a study concerning racism in social networks [15]. In this study, subjects were identified who used racist language on Twitter. The experiment had two different bots reply to these subjects with a message suggesting that such language can be hurtful. The bots profile pictures indicated that they had different races, with one appearing to be a white male, and the other appearing to be a black male. It was found that the white bot was able to reduce the frequency of racist language in its subjects' tweets compared to the black bot. Because the subjects were white, the white bot was in the same racial group as them, while the black bot was not. The fact that the in-group persuasion of the white bot was more effective than the black bot's out-group persuasion is supported by theories of inter-group conflict [21].

Another method to enhance the effectiveness of persuasion is to form a more significant connection or rapport with the target. This can put the target into a positive affective state. It has been shown that persuasion is impacted by affective states [19, 20]. Being liked by the target and having some sort of social rapport with them can enhance persuasion efficacy [4], [5].

1.1 Our Contributions

In-group persuasion may be a way to mitigate the backfire effect. The study in [15] showed that race can function as an in-group feature. The question is whether there is a more general way to create an in-group situation where persuasion can be effective. We propose a new persuasion technique called *pacing and leading* which is more effective than standard arguing and can mitigate the backfire effect. Pacing and leading is a dynamic technique, where the opinion of the persuader evolves over time. Initially, the persuader's opinion aligns with the target to form an in-group bond and emotionally pace the target. Then the persuader gradually shifts his opinion to lead the target to the desired opinion. The idea is that by starting aligned with the target and gradually evolving the opinion, the persuader remains within the confidence bound of the target. As a result, the persuader can avoid the backfire effect and persuade effectively.

To test the pacing and leading technique, we conduct a five month field experiment in Twitter. The experiment focuses on the subject of immigration. Immigration a charged political issue and is actively discussed on social networks. Several studies have measured population level sentiment on this topic in Twitter [18], [1], [6]. It was found in [18] that English posts about the refugee crisis were more likely to have a negative opinion on the topic. A similar result was found for Twitter users in the United Kingdom [6]. Given the level of interest in the topic and its geo-political importance, immigration is an ideal topic to test persuasion methods.

Our goal is to persuade individuals who exhibit xenophobic feelings to have more positive sentiment towards immigration. We compare pacing and leading to the standard arguing technique. In addition, we also test an interaction method that we refer to as *contact* where the persuader likes the social media posts of its targets. This interaction can help the persuader form a stronger rapport with the target and potentially improve persuasion.

An abridged version of our experimental results is found in [22]. Here we present a variety of new results and analyses. We include more details about the experiment design in Section 2. In Section 3 we use neural network based tweet sentiment to determine the response variable for the experiment. Section 4 presents a more detailed analysis of the experimental data. Finally, in Section 5 we perform additional statistical analysis to compare the different treatment effects.

2 Experiment Design

2.1 Experiment Overview

An illustration of our experiment is shown in Figure 1. We now present a high-level overview of the experiment, with details provided in subsequent sections. We use automated Twitter accounts, also known as bots, to test different persuasion methods. Our experiment subjects are Twitter users who actively discuss immigration issues, have anti-immigration sentiment, and follow one of the bots.

Each bot implements a different persuasion method. One bot is a control which posts no content and does not interact with the subjects. One bot applies the arguing method by posting content which is pro-immigration. The third bot applies pacing and leading by posting content that is initially anti-immigration and then gradually becomes more pro-immigration.

We also test a contact treatment, where the bots like the tweets of the subjects to form a closer bond with them. When the bot liked a subject's tweet, the subject is notified. Liking tweets would make the bot more visible to the subject and potentially give the subject a greater trust or affinity for the bot. We randomly selected half of the subjects from each non-control bot and had the bots like the posts of these subjects. The control bot did not apply the contact treatment to any of its subjects.

To assess the effectiveness of the different persuasion methods, we analyze the sentiment of content posted by these subjects over the course of the experiment. We discuss more about how sentiment is used to identify a response variable for the experiment in Section 3.

All subjects voluntarily chose to follow the bots, which may lead to a selection bias in our subjects. Therefore, our conclusions are limited to Twitter users willing to follow the bots and do not necessarily generalize to all Twitter users. However, since a follow-back is required for a Twitter account to implement a tweet based treatment, this is not a strong limitation of our conclusions. This experiment was approved by the Institutional Review Board (IRB) [4] for the authors' institution and performed in accordance with relevant guidelines and regulations.

2.2 Experiment Timeline

The experiment had four different phases. We denote the incubation period as phase zero. During this phase the bots engage in activity in order to appear human. We discuss more about this phase in Section 2.4. Phases one, two, and three are the main active phases of the experiment. During these phases, the non-control bots were both tweeting and applying the contact treatment. The control bot did nothing for these phases. The argue bot posted a pro-immigration tweet once a day in these phases. The pacing and leading bot also posted tweets once a day in these phases, but the tweet opinion varied. In phase one the tweets were anti-immigration. In phase two the tweets expressed uncertainty about immigration or potential validity of pro-immigration arguments. In phase three the tweets were pro-immigration, similar to the argue bot. We constructed the tweets based on what we deemed a proper representation of the opinion for each phase. In phase four of the experiment the bots tweeted nothing. We used this phase to measure any persistent effect of the treatments. The incubation phase began on September 27th, 2018 and the fourth phase was completed on March

[4] The protocol has been approved following full board review by the Committee on the Use of Humans as Experimental Subject (COUHES) under protocol number 1808496544.

1st, 2019. The experiment timeline is shown in Figure 1 and precise dates for the phases are provided in Table 1.

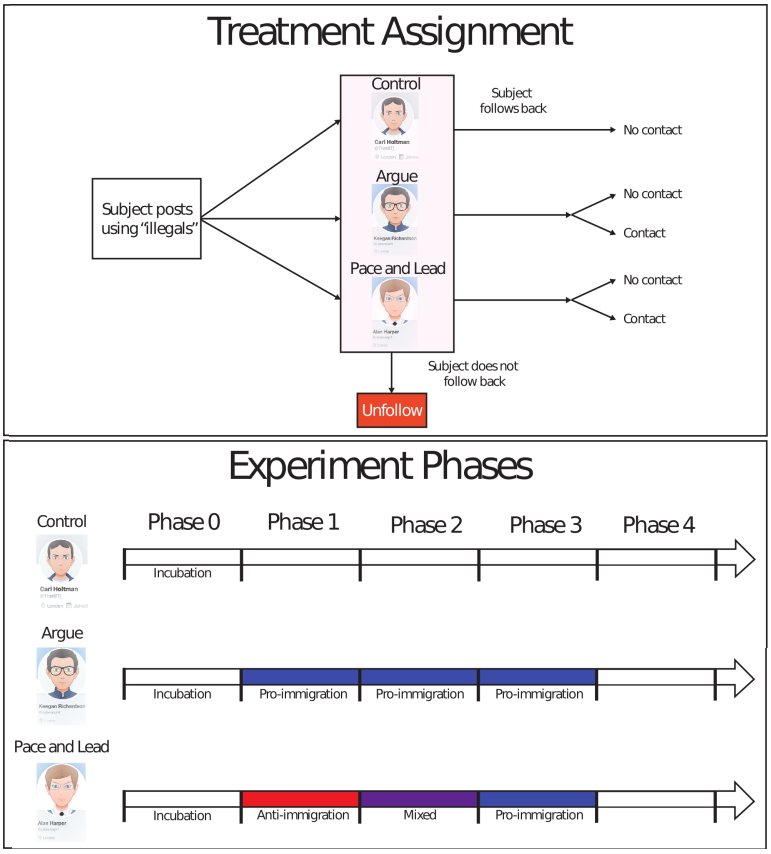

Fig. 1. (top) Diagram illustrating the subject acquisition procedure for the experiment. (bottom) Timeline of experiment phases.

2.3 Subject Selection

The subjects for our experiment were Twitter users who exhibited anti-immigration sentiment. To find potential subjects we began by constructing a list of terms that conveyed strong anti-immigration sentiment, such as #CloseThePorts, #BanMuslim, and #RefugeesNotWelcome. We show in Table 2 the complete list of terms used to find subjects. We used the Twitter Search API to find posts, known as tweets, that contained at least one of these terms. We then collected

Phase	Start Date	End Date
0	September 27, 2018	October 27, 2018
1	October 27, 2018	November 27, 2018
2	November 27, 2018	December 25, 2018
3	December 25, 2018	January 29, 2019
4	January 29, 2019	March 1, 2019

Table 1. Start and end dates for each phase of the experiment.

the screen names of the users who posted these tweets. These users became potential subjects for the experiment.

We required each potential subject to satisfy the following criteria. First, their tweets must be in English and must not contain only punctuation or emojis. Second, the user should not be an automated bot account. The text conditions on the tweet were checked using simple pattern matching. Bot accounts were identified using the machine learning based Botometer algorithm [7]. Users who Botometer identified as being the most bot-like were manually reviewed and eliminated if they were indeed bots.

Our search procedure had the potential to find users who do not have anti-immigration sentiment. For instance, to convey support for immigrants, a user could post a tweet critical of an anti-immigration phrase. To make sure that there were not many users who fall in this category, we manually investigated 100 random users collected by our search procedure. We found that none of the users were pro-immigration, giving us confidence that the overwhelming majority of our potential subjects were anti-immigration.

1 RefugeesNotWelcome	12 StopIslam
2 Rapefugees	13 ISLAMIZATION
3 BanMuslims	14 UnderwearBomber
4 WhiteGenocide	15 NoRefugees
5 StopRefugees	16 StopIllegalMigration
6 CloseThePorts	17 AntiImmigration
7 ImmigrationInvasion	18 Reimmigration
8 MigrantCrime	19 NoRefugees
9 FreeTommy	20 NoIslam
10 QAnon	21 ProtectOurBorder
11 MAGA	

Table 2. Terms used to identify subjects in Twitter for the experiment.

2.4 Details of Bot Operation

We created Twitter accounts for the control, argue, and pacing and leading treatments. One of the goals of our experiment was to test persuasion strategies

in a realistic setting. Therefore, we wanted the bots to resemble human Twitter users, in contrast to the study in [2] where the subjects were told in advance the Twitter account they were following was a bot. To have the bots exhibit human behavior, we had them be active on Twitter before we started the experiment. This incubation period, which is phase zero of the experiment, was designed to let the bots develop an online presence. Each of the bots location was set to London, and they followed a number of popular British Twitter accounts. The bots were designed to look like white males with traditional European names. We used cartoon avatars for the profile pictures, similar to what was done in [15]. We show the profile images for the bots and list their treatment type in Figure 1. During the incubation period, once or twice a day the bots posted tweets about generic, non-immigration topics and shared tweets about trending topics on Twitter, an act known as retweeting. We provide examples of incubation tweets in Table 3. The bots also tweeted articles or videos talking about immigration, but not yet taking a stance on the issue. This was done to show that the bots had some interest in immigration before the experiment began.

Phase	Bot Tweets
0	What an incredible experience #RyderCup18
0	Newcastle become the first team in #PL history to score twice against Man Utd at Old Trafford in the opening 10 minutes #MUNNEW
0	GOAAALLLL!! Shaqiri again playing a big part in the goal. Salah with a smashing finish to make it two!
0	Looking forward to Saturday already! #MondayMotivation

Table 3. Tweets posted by the bots in phase zero of the experiment.

During the incubation period, we began obtaining subjects for the experiment from our list of potential subjects with anti-immigration sentiment. To participate in the experiment, the potential subjects needed to follow one of the bots so that the bots' tweets would be visible in their Twitter timelines. We randomly assigned each of the users in the subject pool to the bots. The bots then liked a recent tweet of their assigned users and followed them. The liking of the tweet and following were done to increase the follow-back rate of the potential subjects. We made sure that no two bots were following the same user as this could arouse suspicion. To avoid bias before the experiment, all tweets the bots liked were not immigration related.

The bots also unfollowed the users after some given time to prevent our following count from being inflated, and to keep a better ratio of followers to following which is desirable for appearing human and gaining followers. The "unfollow time" depends on user tweet frequency and was calculated in the following way. Let W_u denote how long the bot waits between following and unfollowing user u. The wait time should reflect how often a user checks Twitter and it should be shorter for more active users because we want to give the user

time to log in and see that the bot had interacted with and followed them and then make the choice on whether or not to follow it. Also, we want the bot to wait at least one day before unfollowing and at most seven days to ensure that it would not wait too long or unfollow too soon. Let μ_u and σ_u be the mean and standard deviation of the inter-tweet time for user u. Small values for μ_u indicate that u is a active Twitter user and checks the application often. Then W_u is given by

$$W_u = \min(7\ days, \max(1\ day, \mu_u + 4\sigma_u)) \tag{1}$$

The bot would unfollow the user if the user was not following the bot when the wait time had elapsed. Users who followed the bot were not unfollowed and became subjects for the experiment. After liking and following their assigned subjects' tweets, the bots were able to achieve an average follow back rate of 19.3%. Table 4 shows the number of users each bot attempted to follow and the number of users who followed back and were available throughout the experiment.

Bot Name	Treatment Type	Followed	Followed Back	Available
Alan Harper	Pacing and leading	3045	636	578
Keegan Richardson	Arguing	3051	717	651
Carl Holtman	Control	817	125	107

Table 4. Number of users who were followed by, followed back, and remained available for all phases of the experiment for each bot.

Our experiment segmented the subjects into five different treatments groups. There is the control group which consisted of the followers of the control bot. The arguing and pacing and leading bots had their followers split into contact and non-contact groups, depending on who received the contact treatment. The primary features we could measure for the subjects who followed the bots is their follower and following counts on Twitter. We wanted to make sure these values were distributed similarly across the treatment groups. Table 5 shows the followers and following count of the subjects. We performed a pair-wise t-test for all groups and we found that there was no statistically significant difference between any group means ($p < 0.05$).

The tweets of the bots were constructed by humans and posted roughly once a day. The arguing bot tweets all expressed pro-immigration sentiment for all active experiment phases. The pacing and leading bot posted tweets that were anti-immigration in phase one, nuanced in phase two, and pro-immigration in phase three. Tables 6, 7 and 8 show randomly selected tweets posted by the bots in phases one, two, and three.

Treatment	Followers Count		Following Count	
	mean	st. dev.	mean	st. dev.
A	6854	18510	6793	12962
AC	5184	9738	5622	9891
P	6057	11368	6155	9992
PC	4754	20319	4621	13950
Control	6367	7860	6474	7645

Table 5. Statistics of the follower and following count for the followers of each bot. The treatment groups are labeled as follows: control is the control bot, A is argue without contact, AC is argue with contact, P is pacing and leading without contact, and PC is pacing and leading with contact.

Phase	Argue Bot	Pacing and Leading Bot
Phase 1	Former Calais Jungle child refugee who was unlawfully refused safe passage to join his aunt in Britain still in France two years from the closure of the camp. Can we reunite him with his aunt?	Immigrants strike again. Muslim Uber driver Khaled Elsayedsa Ali charged in California with kidnapping four passengers. This needs to be stopped.
Phase 1	Unbelievable. A revised estimate of 56,800 migrants have died/gone missing over the past four years.	Muslims attempt to derail high-speed train in Germany using steel wire. Threats in Arabic were found thereafter.
Phase 1	A win for refugees! Former refugee elected to US congresswoman.	Unacceptable. After mass Muslim migration into Germany, sex attacks are up 70% in Freiburg alone.

Table 6. Tweets posted by the bots in phase one of the experiment.

Phase	Argue Bot	Pacing and Leading Bot
Phase 2	Chancellor Angela Merkel defends UN migration pact. A step in the right direction.	UK Government to sign UN migration pact. Interesting that Angela Merkel defends it, and rejects "nationalism in its purest form". I believe in her.
Phase 2	It's human rights day, and refugees across Europe face widespread human rights violations. Europe needs to do more to uphold natural human rights.	"Muslim imam performed call to worship during a Church of England cathedral's Armistice without permission." Crossed the line. However, it would probably be overlooked if it were the other way around, am i right?".
Phase 2	Now that's efficient and socially productive! Germany sets out new law to find skilled immigrants.	The UN migration pact, which would criminalize criticism of mass migration and redefine a refugee, will be signed by world leaders next week." Though not through public consent, the #ImmigrationMatters initiative did deliver guiding messages to the public.

Table 7. Tweets posted by the bots in phase two of the experiment.

Phase	Argue Bot and Pacing and Leading Bot
Phase 3	Pathetic. At the height of the Syrian refugee crisis in 2015, Syria‘s neighbors took in 10,000 refugees per DAY. Yet the UK Home Secretary just called the arrival of 75 asylum seekers by boat in 3 days a major incident.
Phase 3	Appalling? In 2018 at least 2,242 people have died in the Mediterranean Sea trying to reach Europe.
Phase 3	The sole survivor said he was left alone in the water for at least 1 day before a fishing boat found and rescued him.

Table 8. Tweets posted by the bots in phase three of the experiment.

3 Response Variable

The treatments in the experiment are designed to shift the subjects' sentiment towards immigration. To measure this we need to define a response variable for the subjects. Our only data for the sentiments is the content of their tweets. We need to choose what feature of the tweet content the best represents sentiment towards immigration.

We begin by measuring the sentiment of the tweets using a Bidirectional Encoder Representations from Transformers (BERT) neural network [9]. We selected a pre-trained model that was fine-tuned on 629,000 online product reviews in multiple languages [16]. The sentiment ranges from one to five, with one being the most negative, and five being the most positive. Our dataset consisted of over 2.8 million tweets. To parallelize the computation of the tweet sentiment, we used eight NVIDIA Quadro RTX A6000 GPUs with 2 TB of SSD memory.

We constructed a list of words which both represented sentiment towards immigration and also occurred frequently enough in the data to allow us to measure a response. We then calculated the mean sentiment for the tweets containing each of these words. The results are shown in Figure 2. As can be seen, the word associated with the lowest sentiment is *illegals* with a mean sentiment of 2.0. A Z-test comparing the mean sentiment of tweets with and without illegals used shows that the difference in mean is significant (z-statistic = -66.70, p-value $< 10^{-6}$). The term illegals is a pejorative term used by people with anti-immigration sentiment. For instance, there are tweets such as "*I want a refund on all the tax money spent on illegals!!!*" which shows strong anti-immigration sentiment. The usage frequency of such extreme language can be used to gauge sentiment, as was done in [15]. This suggests that the frequency of tweets containing illegals can be a good representation of anti-immigration sentiment. Therefore, we will use this as our response variable when analyzing the experimental results. This approach was taken in [22], but we now use sentiment analysis to further justify this choice for the response variable.

It is interesting to see how the aggregate fraction of tweets containing illegals varies over the course of the experiment. We show this in Figure 3, along with the value for Google Trends results for illegals. Google Trends provides a normalized volume of global Google searches for keywords [11]. It can be used to gauge

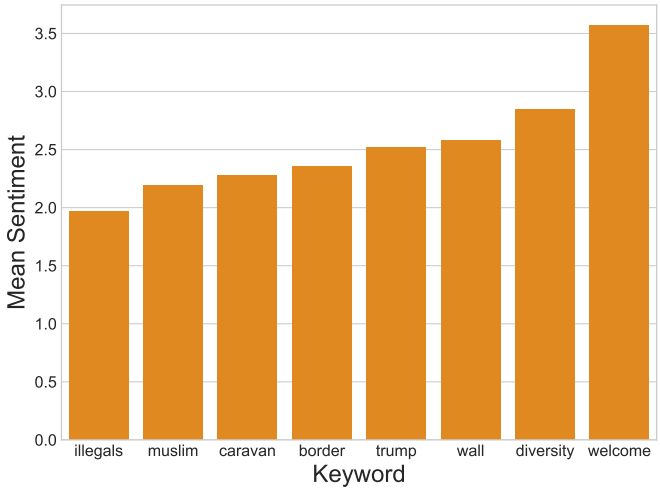

Fig. 2. Bar plot of the mean sentiment of tweets containing immigration related keywords.

general interest in a topic. We see that the tweet frequency and Google Trends for illegals track each other very closely. Further evidence of this relationship is found in the correlation coefficient of the daily tweet fraction and Google Trends for illegals, which is 0.633 (p-value $< 10^{-6}$). Therefore, we see that Twitter activity and Google searches are highly correlated for the keyword illegals.

There is an increase in both Google Trends volume and tweet fraction of the word illegals in phase three of the experiment, which covers December 25th 2018 to January 25th, 2019. During this time period one of the major immigration related news stories was a caravan of migrants from Central America approaching the U.S. southern border [10]. U.S. President Donald Trump expressed his anger with the caravan on Twitter multiple times, tweeting warnings such as "A big new Caravan is heading up to our Southern Border from Honduras," and "Only a Wall will work. Only a Wall, or Steel Barrier, will keep our Country safe!". The approach of the caravan and Trump's aggressive tweets likely caused an increase in Google searches for the keyword illegals and an increase in the frequency of the keyword on Twitter.

Exogenous events such as the caravan likely stir strong negative emotions in users, who then post content with similar sentiment. This suggests that periods of higher tweeting activity should be accompanied by a more negative sentiment. We find evidence to support this claim in our data. Figure 4 shows a scatter plot of the daily tweet rate of all experiment subjects over all phases of the experiment versus the daily mean sentiment of these tweets. As can be seen, there is a trend where the sentiment decreases as the tweet rate increases. The

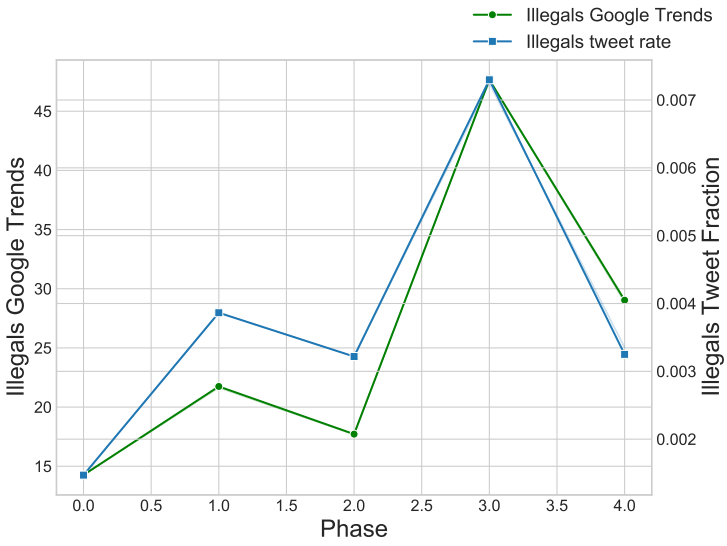

Fig. 3. Plot of the frequency of tweets containing 'illegals' and Google Trends for 'illegals' versus experiment phase.

correlation coefficient of these variables is -0.554 (p-value $< 10^{-6}$), showing that this trend is significant. This observation suggests that the individuals in the experiment were driven to tweet negative content more frequently that positive content.

4 Empirical Results

Our analysis of tweet sentiment motivates us to use the frequency of extreme anti-immigration language in the subjects' tweets to measure any persuasion effect the bots had. Because of the strong negative sentiment of the word illegals, we use its frequency as our measure of immigration sentiment.

We show in Table 9 the number of tweets and number of tweets with the word illegals in each phase and treatment group. These is the raw data used in our statistical analysis. We plot the illegals usage frequency in each phase and treatment group in Figure 5. This frequency is defined as the number of tweets containing illegals divided by the total number of tweets for all subjects in each phase and treatment group. We note that the overall frequency is very low, but in phase two, we see that the pacing and leading with contact treatment has a much lower frequency than the other treatments, while arguing with contact has the highest frequency. Recall that in phase two pacing and leading has tweets that are slightly pro-immigration. This suggests that perhaps pacing and leading with contact is mitigating the backfire effect. We next perform a statistical analysis of this data to obtain a more quantitative assessment of the different treatments.

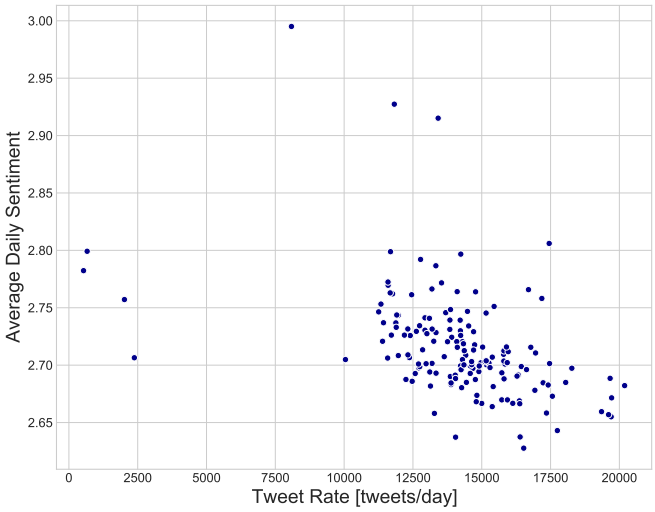

Fig. 4. Scatter plot of the daily mean sentiment and daily tweet rate of tweets from the experimental subjects during all phases of the experiment.

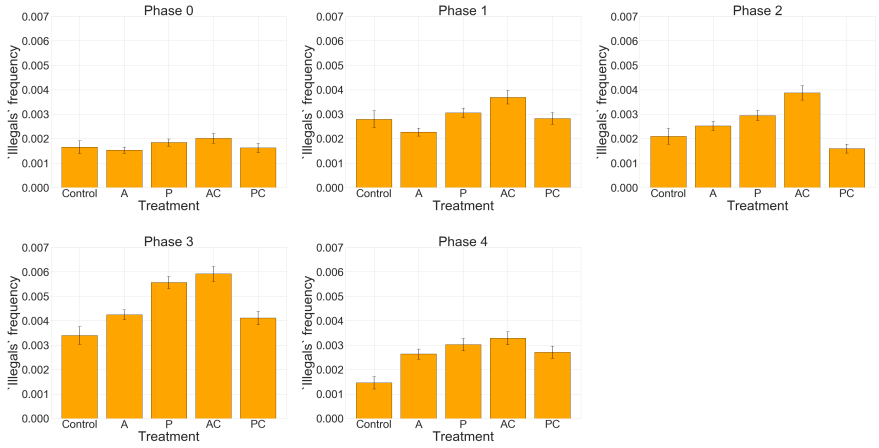

Fig. 5. Plot of the frequency and standard error of usage of the word "illegals" in tweets for each phase and treatment group. The treatments are labeled as follows: A is argue without contact, AC is argue with contact, P is pacing and leading without contact, and PC is pacing and leading with contact.

Phase	Treatment	Number of tweets	Number of tweets containing "illegals"
0	Control	24,156	40
0	A	97,277	149
0	P	86,236	159
0	AC	50,474	102
0	PC	47,467	77
1	Control	23,212	65
1	A	85295	194
1	P	83408	255
1	AC	47880	177
1	PC	49110	139
2	Control	19986	42
2	A	70877	179
2	P	68151	201
2	AC	44114	171
2	PC	47086	75
3	Control	25863	88
3	A	100079	425
3	P	90942	506
3	AC	63382	375
3	PC	56851	234
4	Control	23205	34
4	A	60262	159
4	P	47571	144
4	AC	44462	146
4	PC	42059	114

Table 9. The number of tweets and tweets containing the word illegals in each phase and treatment group of the experiment. The treatments are labeled as follows: A is argue without contact, AC is argue with contact, P is pacing and leading without contact, and PC is pacing and leading with contact.

5 Statistical Analysis

We treat each tweet as a binary outcome that equals one if the tweet contains the word illegals. The probability of such an outcome is modeled using logistic regression. For a tweet i the probability is

$$\log\left(\frac{p_i}{1-p_i}\right) = \sum_{t=0}^{4}\beta_t x_{t,i} + \sum_{t=0}^{4}\beta_{a,t}x_{a,i} + \sum_{t=0}^{4}\beta_{p,t}x_{p,i} + \sum_{t=0}^{4}\beta_{ac,t}x_{ac,i} + \sum_{t=0}^{4}\beta_{pc,t}x_{pc,i} + \epsilon_i. \tag{2}$$

The coefficients β_t for $t = 0, 1, ..4$ model exogenous factors that may impact the probability during each phase. For instance, news stories related to immigration may increase the probability. By regressing out the phase effect we can isolate the different treatments. We use separate treatment coefficients for each phase because the pace and lead treatment varies by phase. Recall that this treatment shifts the opinion of its tweets from anti- to pro-immigration over phases one to three. The treatment coefficients are indexed by subscripts indicating the treatment and phase. We use the subscript t for the phase, a for argue, and p for pace and lead. The subscript c indicates the contact treatment where the bots like the subjects' tweets. The x variables are binary indicators for the treatment group of the subject posting the tweet and in which phase the tweet occurred. User heterogeneity and other unobserved factors are modeled using a zero mean normally distributed random effect ϵ_i.

Phase	Coefficient
Phase 0	-6.40^{**}
Phase 1	-5.88^{**}
Phase 2	-6.16^{**}
Phase 3	-5.67^{**}
Phase 4	-6.52^{**}

Table 10. Regression coefficients for each phase. * indicates significance at a 5% level and ** indicates significance at a 1% level.

We summarize the regression coefficients for the phases in Table 10 and for the treatments in Table 11. For the phase coefficients we see that the largest value occurs in phase three, which was also the phase when Google Trends volume was highest, as seen in Figure 3.

For the treatment coefficients, we see in phases two and three that arguing with contact is positive and significant at a 1% level. This is the backfire effect,

Phase	Treatment	Coefficient
Phase 0	A	-0.078
Phase 0	AC	0.200
Phase 0	P	0.108
Phase 0	PC	-0.021
Phase 1	A	-0.209
Phase 1	AC	0.279
Phase 1	P	0.088
Phase 1	PC	0.011
Phase 2	A	0.184
Phase 2	AC	0.614^{**}
Phase 2	P	0.340
Phase 2	PC	-0.278
Phase 3	A	0.226
Phase 3	AC	0.556^{**}
Phase 3	P	0.494^{**}
Phase 3	PC	0.191
Phase 4	A	0.589^{**}
Phase 4	AC	0.809^{**}
Phase 4	P	0.727^{**}
Phase 4	PC	0.616^{**}

Table 11. Regression coefficients for each treatment. The treatments are labeled as follows: A is argue without contact, AC is argue with contact, P is pacing and leading without contact, and PC is pacing and leading with contact. * indicates significance at a 5% level and ** indicates significance at a 1% level.

as it means the treatment is worse than the control bot, which posted nothing. When there is no contact, arguing does not show a backfire effect. It is possible that without the contact treatment the bots' tweets are not seen by the subjects. This is one explanation as to why without contact, arguing did not have a backfire effect. Pacing and leading without contact is phase three also showed a backfire effect, while with contact it did not. In this phase, the pacing and leading treatment is posting pro-immigration tweets, the same as arguing. It is not clear why there is backfire without the contact treatment. In phase four all treatment groups show a backfire effect. The source of this can be found in Figure 5 where we see that the control group has a much lower illegals frequency than the other groups. It is not clear what causes this drop in the control group. In phase zero we see that all groups are balanced in their illegals frequency, suggesting that there is not a bias between the groups. Also, in phase four the bots stop tweeting. There must be another factor apart from bias that is responsible for the phase four backfire effect.

In Figure 6 we plot the treatment coefficients versus phase for each treatment group. The plot have the treatments grouped in different ways in order to makes differences in each individual treatment over the phases more visible. We conduct a two-sample proportions z-test for the pairs of treatments to assess the statistical significant of the differences in the treatment effects. Differences that are significant at a 1% level are indicated on the plots.

We first look at the effect of the contact treatment. In the top left plot of Figure 6 we see that the argue with contact coefficient is greater than argue without contact, and the difference does not vary much over the phases. The difference is significant for phases one, two, and three. In phases zero and four, where the bots do not tweet about immigration, there is no significant difference. The contact treatment may be making the bots' pro-immigration tweets more visible to the subject, resulting in a backfire effect where the subjects use the word illegals more frequently.

For pacing and leading in the top right plot of Figure 6, we see that the non-contact coefficient is greater than contact. In phases two and three the difference is significant. Contact appears to enhance the effectiveness of pro-immigration tweets in the later stages of the pacing and leading treatment. This is in contrast to arguing, where contact degrades the effectiveness of pro-immigration tweets.

We next look more closely at arguing versus pacing and leading when the contact treatment is fixed. In the bottom left plot of Figure 6 we see that without contact, the tweet based treatment coefficients have a small difference which does not vary appreciably across phases. Arguing has a smaller coefficient, but the difference is statistically significant only for phases one and three.

For the contact group in the bottom right plot of Figure 6, the difference changes sign. Argue has the larger coefficient and the difference varies across the phases. Phase two shows a large significant difference. The difference is smaller in phase three, but still significant. The moderately pro-immigration tweets of the phase two pace and lead treatment seem to be more effective than the argue tweets when the bot has contact with the subject. The same can be said of fully

pro-immigration tweets in phase three, but the advantage of pacing and leading over arguing is less than in phase two.

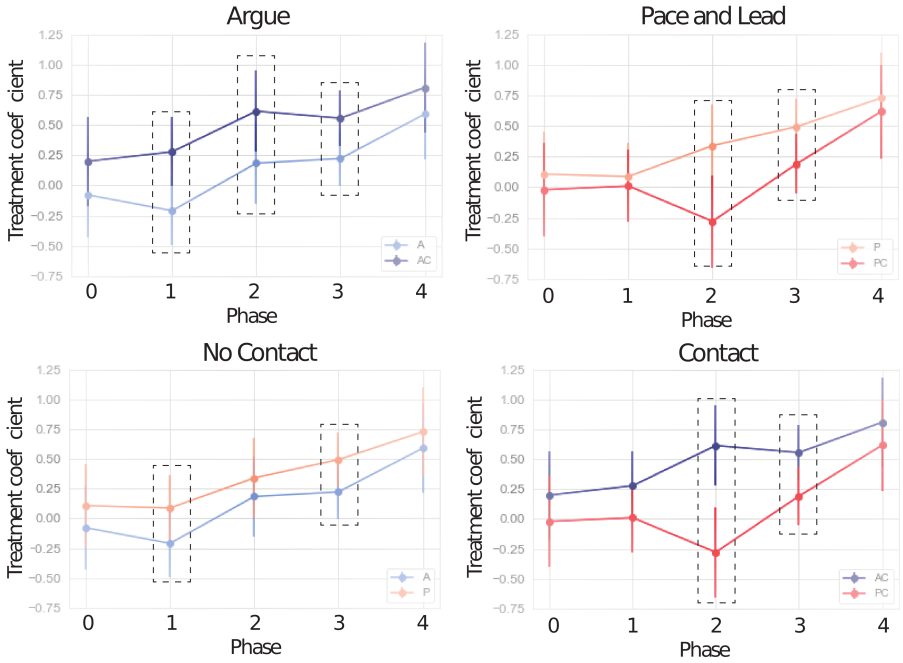

Fig. 6. Plots of the regression coefficients (with standard errors) for the treatments in each phase. The title of each plot indicates the treatment component that is held fixed. The treatments are labeled as follows: A is argue without contact, AC is argue with contact, P is pacing and leading without contact, and PC is pacing and leading with contact. The dashed boxes indicate which coefficients have a difference that is statistically significant at a 1% level.

One potential source of contamination in our experiment is if a subject retweeted the bot he followed, and then this retweet was seen by his follower who also followed a different treatment bot. This would cause the follower to receive treatments from two different bots, which is known as a spillover effect. Though retweets happen very rarely in our experiment, we still wanted to make sure the spillover effect does not affect our results.

In total, 18 users retweeted the bots during the experiment. After checking the follower network of the subjects, we identified 213 users (including the 18 retweeters) in the experiment who may have experienced the spill over effect. We excluded these users from our analysis to cross-validate our result. We run logistic regression on both the whole user set, as well as the refined user set

without spill over subjects. The resulting regression coefficients and significance levels were similar. Therefore, we do not see any appreciable spillover effects.

6 Discussion and Conclusion

We have presented here a field experiment which tested different persuasion techniques. We compared a static technique (arguing) with a dynamic technique (pacing and leading). We also tested the effect of having the bots make contact with the subjects by liking their tweets. The most effective treatment appeared to be pacing and leading with contact. In the later phases of the experiment, pacing and leading outperformed arguing when the contact treatment was applied. In fact, arguing with contact caused a backfire effect and increased the usage of xenophobic terms compared to the control group.

Pacing and leading with contact appeared to be most effective in phase two when the tweet sentiment was slightly pro-immigration, but not strongly pro-immigration. This suggests that to overcome the backfire effect, one should continuously like the posts of the users, and present arguments that are more nuanced and moderate in their tone. This softer approach appears to be more effective than standard arguing.

There are several interesting questions raised by our findings. One question concerns the phases for pacing and leading. We found that the moderate posts were most effective. It is not clear if this treatment would work in isolation or if the phase one pacing treatment is necessary. We hypothesize that this period allows greater trust to be built between subject and bot, but our experiment does not confirm this. Another question is whether the phase three pacing and leading treatment where the posts strongly advocate the target position is necessary. It may be that the moderate posts are sufficient to mitigate the backfire effect and potentially persuade the subject.

Finally, we note that care should be taken when trying to apply our results to more general settings. This study focused on the topic of immigration, which is an important political and policy issue. Discussion on this topic has split along traditional conservative and liberal fault lines. We expect our findings to extend to similar political issues, but further study is needed. However, our subjects were Twitter users with anti-immigration sentiment who were willing to follow our bots. This represents a limited population in a very specific social setting. More work is needed to determine whether our findings replicate in different populations or within varied social settings.

7 Declarations

7.1 Ethics Approval and Consent to Participate

This experiment was approved by the Institutional Review Board (IRB) of the Massachusetts Institute of Technology and performed in accordance with relevant guidelines and regulations. The protocol has been approved following full

board review by the Committee on the Use of Humans as Experimental Subject (COUHES) under protocol number 1808496544.

7.2 Consent for Publication

Not applicable.

7.3 Availability of Data and Material

Please contact author for data requests.

7.4 Competing Interests

The authors declare that they have no competing interests.

7.5 Funding

There are no funding sources to report for this work.

7.6 Authors' Contributions

All authors contributed equally to the manuscript.

7.7 Acknowledgements

The authors would like to thank Dean Eckles for helpful discussions on the experiment design.

References

1. G. Backfried and G. 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.
2. C. A. Bail, L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018.
3. E. Bakshy, S. Messing, and L. A. Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.
4. J. M. Burger, S. Soroka, K. Gonzago, E. Murphy, and E. Somervell. The effect of fleeting attraction on compliance to requests. *Personality and Social Psychology Bulletin*, 27(12):1578–1586, 2001.
5. R. B. Cialdini and M. R. Trost. Social influence: Social norms, conformity and compliance. 1998.

6. M. Coletto, A. Esuli, C. Lucchese, C. I. Muntean, F. M. Nardini, R. Perego, and C. 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.
7. C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th international conference companion on world wide web*, pages 273–274. MISSING PUBLISHER, 2016.
8. M. H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.
9. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2019.
10. J. Ernst and K. Sempler. Migrant caravan departs honduras. *The New York Times*, 2019.
11. Google. Google trends, 2022. https://trends.google.com/trends/.
12. R. Hegselmann, U. Krause, et al. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of artificial societies and social simulation*, 5(3), 2002.
13. C. G. Lord, L. Ross, and M. R. Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology*, 37(11):2098, 1979.
14. M. Mosleh, C. Martel, D. Eckles, and D. G. Rand. Shared partisanship dramatically increases social tie formation in a twitter field experiment. *Proceedings of the National Academy of Sciences*, 118(7), 2021.
15. K. Munger. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 39(3):629–649, 2017.
16. nlptown. bert-base-multilingual-uncased-sentiment. https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment. Accessed: 2022-02-22.
17. B. Nyhan and J. Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, 2010.
18. N. Öztürk and S. Ayvaz. Sentiment analysis on twitter: A text mining approach to the syrian refugee crisis. *Telematics and Informatics*, 35(1):136–147, 2018.
19. B. Rind. Effects of interest arousal on compliance with a request for help. *Basic and Applied Social Psychology*, 19(1):49–59, 1997.
20. B. Rind and D. Strohmetz. Effect on restaurant tipping of presenting customers with an interesting task and of reciprocity. *Journal of Applied Social Psychology*, 31(7):1379–1384, 2001.
21. H. Tajfel, J. C. Turner, W. G. Austin, and S. Worchel. An integrative theory of intergroup conflict. *Organizational identity: A reader*, 56:65, 1979.
22. Q. Yang, K. Qureshi, and T. Zaman. Mitigating the backfire effect using pacing and leading. In *International Conference on Complex Networks and Their Applications*, pages 156–165. Springer, 2021.