Universidad del Rosario
Facultad de Economía

Tweeting for Peace: Experimental Evidence from the 2016 Colombian Plebiscite

Jorge Gallego

Juan D. Martínez

Kevin Munger

Mateo Vásquez

SERIE DOCUMENTOS DE TRABAJO

No. 208

Noviembre de 2017

Tweeting for Peace: Experimental Evidence from the 2016 Colombian Plebiscite

Jorge Gallego[*] Juan D. Martínez [†]

Kevin Munger [‡] Mateo Vásquez [§]

November 9, 2017

Abstract

The decades-long Colombian civil war nearly came to an official end with the 2016 Peace Plebiscite, which was ultimately defeated in a narrow vote. This conflict has deeply divided Colombian civil society, and non-political public figures have played a crucial role in structuring debate on the topic. To understand the mechanisms underlying the influence of members of civil society on political discussion, we performed a randomized experiment on Colombian Twitter users shortly before this election. Sampling from a pool of subjects who had been frequently tweeting about the Plebiscite, we tweeted messages that encouraged subjects to consider different aspects of the decision. We varied the identity (a general, a scientist, and a priest) of the accounts we used and the content of the messages we sent. We found little evidence that any of our interventions were successful in persuading subjects to change their attitudes. However, we show that our pro-Peace messages encouraged liberal Colombians to engage in significantly more public deliberation on the subject.

[*] Facultad de Economía, Universidad del Rosario.

[†] Departamento Nacional de Planeación, Colombia.

[‡] Department of Politics, New York University.

[§] Department of Politics, New York University.

1 Introduction

As part of the 2016 peace negotiations between the guerrillas and the Colombian government, the Conflict and Victims Historical Commission (CHCV) was asked to review several studies about the origins of violence in Colombia. In the twelve documents presented in La Havana during the negotiations, the authors agreed that one of the most characteristic features of Colombia during the nineteenth century and the first half of the twentieth century was the confrontation between the conservative and the liberal visions of the world (CHCV, 2015). Ideology is a fundamental feature of Colombian politics, and elites on either side of the debate have sanctified their respective priority—peace or justice—as the ultimate criterion to support or oppose the agreement between the government and the FARC.

Several studies argue that the early consolidation of the Liberal and Conservative Parties facilitated the elite-based political competition that excluded the political participation of other groups (Wills, 2015).[1] Moreover, Colombians have long recognized the importance of non-political public figures, like priests or generals, in structuring political debate—even compared to other Latin American countries—but there has been little academic research about the precise mechanisms that underly this type of influence.

After 60 years of conflict, in which different segments of society became participated in some way, it is not clear who has the authority to talk about peace. Most institutions in the country have participated in different ways in the civil conflict, and few of them were effective at mobilizing the public to support peace (Duncan, 2015).

We want to understand how different non-political public figures can persuade and change citizens' attitudes about conflict, or at least encourage them to deliberate on this matter. Do endorsements by public figures matter when frames are ideologically aligned? What type of speaker would have a greater impact for different segments of the ideological spectrum? What type of figures and messages are more effective in changing deliberative decisions such as participating in the debate?

This paper addresses these questions in the context of the 2016 Colombian Peace Plebiscite, an election in which Colombians had to decide whether they supported the peace agreement between the FARC and the government. We want to test if messages

[1] During The National Front (Spanish: Frente Nacional 1958-1974) the two main political parties agreed to rotate power, alternating for a period of four presidential terms and restricting the participation of other political movements. The FARC was founded in 1964 as a response to the limited political opportunities available at the time.

related to the process, from like-minded public figures, cause positive reactions and increased engagement from citizens. We are able to test this causally by conducting an experimental study using Twitter "bots" that we control to randomize the messages sent by accounts that appeared to be Colombian non-political public figures (Munger, 2017*a*,*b*). This approach allows us to perform the experiment on the sample of interest—Colombian Twitter users who frequently posted comments about the plebiscite—in a naturalistic setting.

These subjects were interested in the peace process, but they expressed a wide range of opinions about it; some were strongly in favor, others strongly against. This heterogeneity was partially the product of an information deficit and partially the product of differences in fundamental beliefs and values. While the second issue is outside the scope of our intervention, we study different ways to motivate an informative discussion about the agreement. This is is an example of a "hard case" of social influence; the debate over the peace process was central to many peoples' political worldviews, and our subjects were those who had already expressed strong views. This paper investigates whether elite influence in the form of a single message from an account that appears to be a potentially influential figure in Colombian society (a priest, a scientist or a general) can change the way that people talked about this important political decision.

We find little evidence of persuasion from this intervention—very few people switched their attitude towards the peace process as a result, which is somewhat unsurprising given our highly motivated sample and the low intensity of our treatment. However, we do find that liberals (who advocated for the peace agreement) were motivated to send more messages in favor of the process after receiving a message arguing in favor of the process. Conservatives did not send more of these positive messages, but neither did they send more negative messages. On balance, we find robust evidence that a variety of cultural figures were able to spur increased participation in the online discussion of this important political event.

2 Background

2.1 The 2016 Peace Plebiscite

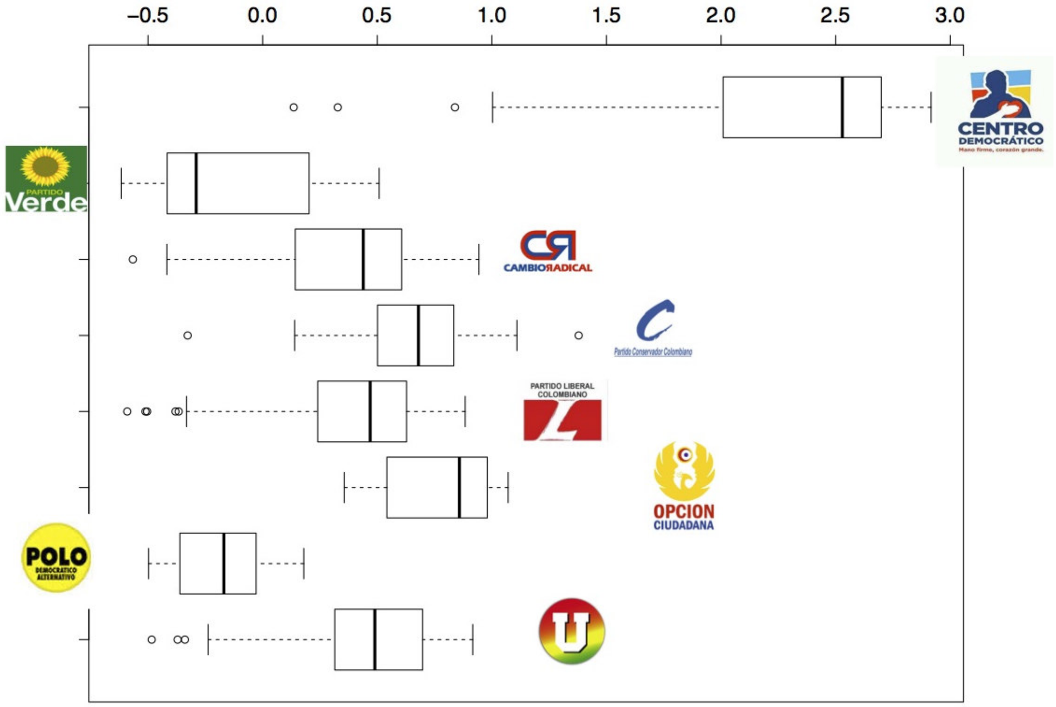
In this study we focus on messages related to the Peace Agreement negotiated in La Havana during 2012-2016. After more than 50 years of war, the Colombian government and guerrilla group FARC reached an accord. Citizens had the opportunity of validating or rejecting this deal through a plebiscite that took place on October 2nd, 2016. The referendum to ratify the final peace agreement failed with 50.2% of people voting 'No' and 49.8% voting 'Yes'.

The referendum consisted of a single question that voters had to approve or reject: "Do you support the final agreement to end the conflict and build a stable and lasting peace?" Two clear sides campaigned during the weeks preceding the referendum. The 'Yes' campaign was supported by many members of the Colombian community from the political left, center-left and center, led by President Juan Manuel Santos. The political parties in favor were the Alternative Democratic Pole, the Social Party of National Unity, Radical Change, the Independent Movement of Absolute Renovation, the Indigenous Social Alliance Movement, the Green Party of Colombia, and the Liberal Party of Colombia. This group emphasized peace over justice.

The most prominent campaigner for the 'No' vote was the Democratic Center Party, a right wing party led by current senator and former president Alvaro Uribe. The Democratic Center Party presented several arguments against the peace deal, among them that the guerrillas would not serve enough time in prison, that they would automatically be awarded seats in Congress, and that in pursuing the negotiations President Santos had gone beyond the terms of the Colombian Constitution; in general, this group prioritized justice over peace.

There was a strong connection between ideology in the traditional left-right political spectrum and support for the peace process. We validate this conventional wisdom by estimating the ideology of the parties involved in the campaign according to their Twitter networks. Figure 1 plots these estimates for each of member of Congress in Colombia. These estimates do not use labeled data or expert coding, but are derived entirely from the Bayesian Spatial Following Model (Barberá, 2015). This model looks only at the accounts that each account follows and iteratively updates the closeness of each account in the network. The main intuition behind this method is that the probability of following someone on Twitter increases with the ideological closeness between the two accounts. The estimates in Figure 1 pass the test of face validity of the relative positions of the parties in the referendum campaign: Centro Democratico is located at the right of the graph, which represents a more conservative ideology and consequently a negative view of the referendum. President Santos' main coalition—Partido de la U, Cambio Radical, Liberal and Conservador—are located towards the center right of the scale. Polo Democratico and Partido Verde, which are the left parties in Colombia and supported the peace process, are located at the left of the graph.

Figure 1: Ideological Position of Political Parties in Colombia



Estimates of the ideological position of each Member of Congress in Colombia, sorted by their party. These estimates are derived from Barberá (2015)'s Bayesian Spatial Following Model.

2.2 Twitter and Public Figures in Colombia

As in recent elections in other contexts, social media proved to be one of the most important platforms to express opinions on both sides of the debate. Twitter use is very common in Colombia,[2] and it is widely 1utilized by political and media elites.[3] This election provided an ideal opportunity to analyze the effects on sentiments and opinions about the Colombian peace process of different persuasion strategies on Twitter. The referendum was conducted without explicit party labels on the ballot, and concerned the single most important issue in Colombian politics.

We presented our bots as representative of people or institutions trusted by Colombian citizens; some of these identities are associated with conservative values, and others with liberal values. For instance, according to a Gallup poll conducted in October of 2016, about 60% of the respondents had a favorable opinion of the Catholic Church. This is relatively high, compared to other institutions.[4] Catholicism is, by far, the most popular religion in Colombia, and throughout history it has been associated with the Conservative Party.[5] Therefore, for a huge segment of the population, priests' political positions shape public opinion. In the context of this plebiscite, the Catholic Church remained officially neutral, while some Protestant groups clearly declared their opposition to the peace accords.[6] Still, priests tend to have much more influence in shaping opinions among conservatives.

The same poll reveals that 71% of the population has a favorable opinion of the military, whose reputation can be explained by their role in the longstanding civil conflict with the FARC. Even though the military tend to be associated with values linked to conservative parties, in this case the perception is at least ambiguous, as the soldiers killed by this war came from cross-cutting segments of society, generating sentiments of sympathy and acknowledgment among all Colombians. Finally, even

[2] According to the Colombian Minister of Technology and Communication, in recent years, users registered in social networks in Colombia are increasing. The most popular platforms are Facebook and Twitter. According to their study, it is estimated that in Colombia there are 5.2 million users, or over 10% of the population (MINTIC, 2015).

[3] While the two leaders of the campaign during the Plebiscite were particularly savvy Twitter users, virtually all elected congressmen, mayors, and governors in the country have an account.

[4] The percentage of respondents with a favorable opinion of other institutions are: 21% for the Congress, 41% for unions, 51% for the media, 15% for the judicial system, and 14% for political parties.

[5] In fact, some of the 19th and 20th century civil wars were associated with religious issues or had some sort of connection to the Catholic Church.

[6] See, for instance, this news coverage: http://www.bbc.com/mundo/noticias-america-latina-37560320

though opinion polls usually do not ask about citizens' perceptions of scientists, it is well known that academics tend to be more inclined towards liberal values. In fact, in the context of this plebiscite, several of the most renowned Colombian professors signed petitions supporting the peace deal with FARC.[7]

In fact, a recent study (British Council, 2017) reveals that the most trusted institutions by the Colombian youth are: professors (54%), the army (48%), and the Catholic Church (45%), in sharp contrast with their perceptions of the government, political parties, and illegal armed groups (refer to Figure 15 in the Appendix for a graphical illustration). Consequently, we consider that in the Colombian context public figures and institutions associated with the church, the military, and academia, exert influence on citizens' political opinions. We theorized that the priest would be most associated with conservative values and thus the "No" vote and the scientist with liberal values and thus the "Yes" vote, while the soldier would be more moderate.

3 Persuasion: Moral Values and Social Cues

Political psychologists recognize that moral values play an important role in determining political attitudes and subsequent electoral behavior (Feinberg and Miller, 2015; Janoff-Bulman, Sheikh, and Baldacci, 2008; Morgan, Skitka, and Wisneski, 2010; Volkel and Feinberg, 2016). Evidence suggests that ideology shapes people's response to information and affects how they make political decisions (Campbell, 1960), but in this process moral convictions are crucial as they shape political attitudes (Graham, Haidt, and Nosek, 2009; Lakoff, 2002; Morgan, Skitka, and Wisneski, 2010). In other words, a person's moral values act as 'perceptual screens' that influence the position that they take with respect to a salient issue (Kernell, 2013).

Consequently, in order to persuade someone to adopt a certain position or to support a particular political action, it is necessary to appeal to the counterparts' moral convictions –something called "moral reframing" (Feinberg and Miller, 2015; Volkel and Feinberg, 2016). If a liberal wants to convince a conservative about a certain issue, she must elaborate an argument based on the moral values in which conservatives believe. For example, recent research shows how conservatives framed under the argument that same-sex couples are "*proud* and *patriotic* Americans", are more likely to support same-sex marriage, compared to those who are framed under the argument that these couples

[7] See, for instance, http://www.elespectador.com/noticias/politica/academicos-el-si-el-plebiscito-articulo-648447

"should be treated *equally* to opposite-sex couples" (Feinberg and Miller, 2015).

Central to this argument is the idea that liberals and conservatives have different moral profiles and do not necessarily share the same view of what is correct and what is not (Caprara et al., 2006; Graham, Haidt, and Nosek, 2009; Thorisdottir et al., 2007). The Moral Foundations Theory (Graham et al., 2011; Graham, Haidt, and Nosek, 2009), based on surveys of thousands of respondents around the world, suggests that there are five primary moral foundations and that liberals and conservatives differ in the weight they place on these issues: harm/care,[8] fairness/reciprocity,[9] in-group/loyalty,[10] authority/respect,[11] and purity/sanctity[12] (Volkel and Feinberg, 2016).

Evidence suggests that while conservatives put more weight on in-group/loyalty, authority/respect, and purity/sanctity, liberals respond more to the harm/care and fairness/ reciprocity foundations. This explains why conservatives framed under the "*patriotic* Americans" argument are more likely to support same-sex marriage. Similarly, Volkel and Feinberg (2016) find that liberals are less likely to support Hillary Clinton when critiques against her are framed within the fairness foundation, in contrast to loyalty values. Therefore, this bulk of evidence suggests that to persuade citizens to take a particular attitude or behavior, arguments should be grounded in moral terms that appeal to those in which the counterpart believes.

This motivated our use of two types of messages—one for each set of moral foundations in which liberals or conservatives tend to believe. The liberal message highlights the fact that the war has victimized thousands (harm), especially among the poor (reciprocity). The conservative message, in contrast, emphasizes the fact that the agreement represents the victory of Colombian compatriots (in-group loyalty and authority) and the will of God (sanctity). We test whether these messages have differential effects on liberal/conservative Twitter users. Note that we are aware that large-scale studies of persuasion efforts find them to be minimal (Broockman and Green, 2014; Kalla and Broockman, 2017). Part of our motivation for this study was to see if these minimal effects obtain in an online setting as well; we do not expect to find strong persuasive effects with this experiment.

While the literature on moral reframing has emphasized the *content* of the message,

[8] The harm/care foundation's main preoccupation is to prevent and alleviate the suffering of the least advantageous in a society.

[9] This moral value refers to justice and equality.

[10] Under the ingroup/loyalty foundation, it is morally acceptable to prioritize one's in-group welfare.

[11] According to this value, there should be respect for tradition and higher-ranked individuals.

[12] This foundation is concerned with purity and sacredness.

it has largely ignored the role played by the *identity* of the messenger. A wealth of literature in political science shows that elites and important public figures mold how people perceive different events, situations or messages by using different frames. This phenomenon, known as "elite-issue framing", occurs when in the course of describing an issue, a speaker emphasizes on a set of considerations that causes individuals to focus on them when processing the information and forming an opinion.[13]

Extensive research suggests that citizens rely on simple and reliable cues from elites and public figures in order to make policy judgments (Lupia, 1994, 2015; Lupia and McCubbins, 1998; **?**). Elites affect the public's perceptions of their political in-group and shape attitudes and behavior towards the out-group. A number of studies show that framing effects—how an issue is emphasized—can substantially shape and alter opinions. This work isolates a variety of factors that moderate the impact of a given frame. One of the most important factors is a frame's strength (Aarøe, 2011; Chong and Druckman, 2007; Druckman and Leeper, 2012). However, it has been difficult to conduct research in this area that is both causally identified and outside of a lab setting. Our approach allows us to address this gap by randomly assigning the identity of an important character providing information *and* modifying the framing of that information, all in the realistic context of social media.

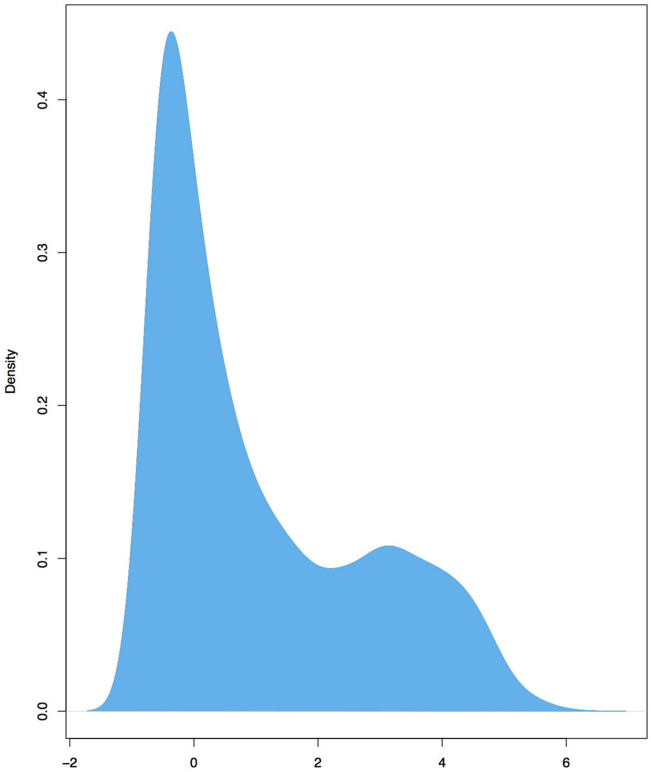
We focus here on three identities as sources of information: a priest, a general, and a scientist. These three figures are associated with a conservative, a moderate, and a liberal vision of the issue at hand. The selection of these figures is motivated by the ideological distribution of our sample. As seen in Figure 2, the distribution of ideology of Twitter users is slightly skewed. There is a tall, thin cluster of liberals and a broader cluster of conservatives, as well as a range of moderate individuals between them.

Accordingly, we divide the political spectrum into three segments: liberal, conservatives, and moderates. We test the extent to which people decide to increase their participation in response to a message from a like-minded elite. Additionally, we evaluate the effect of the treatment on the tone and use of language used in the discussion. We hypothesize that the message of a priest (scientist) will generate a more active and civil participation from a conservative (liberal). The hypothesis follows the results in Dickson, Hafer, and Landa (2008) where people 'overspeak' in a debate when the discussion is likelier to alienate rather than persuade.

Further, we hypothesize that the effect will vary according to the source of informa-

[13] See Druckman (2014) for a review of this literature.

Figure 2: Ideology Distribution of Subjects



Histogram of the ideological position of each of the subjects in our experiment. These estimates are derived from Barberá (2015)'s Bayesian Spatial Following Model.

tion for different segments of the users. We expect stronger reactions for non-moderate citizens. First, we expect opponents of the peace process (conservatives) to have a negative reaction to messages that come from a liberal source.

4 Deliberation in Social Media

Deliberation is an inclusive process of reciprocal argumentation, which aims to be rational, public, and free of any form of coercion (Habermas, 1996). Informal political talk, which is a particular form of deliberation (Graham, 2015) helps citizens become aware of others opinions, discover important matters, develop their preferences and test new ideas. Therefore, deliberation is essential for any democracy, as it prepares citizens for further political action, which may include the vote. Deliberation can increase levels of political knowledge, civic engagement, and tolerance (Coleman and Blumer, 2009; Eveland, 2004; Graham, 2015; Johnston et al., 2001; Kim, Wyatt, and Katz, 1999).

Does social media increase the levels of deliberation and political talk in a society? This is an open question, and evidence from Facebook (Robertson, Vatrapu, and Medina, 2010), Twitter (Yardi and Boyd, 2010), Weblogs (Papacharissi, 2009), Youtube (Halpern and Gibbs, 2013), and Wikipedia (Black et al., 2011) are decidedly mixed. This contrasts with the initial optimism about the possibility of the interent to revitalize the public sphere of debate (Papacharissi, 2004).

There is also a pessimistic strand of research on deliberation and social media. Some scholars argue that most popular online forums (and platforms like Twitter) are not very deliberative (Janssen and Kies, 2005).[14] In part because of the multifaceted nature of reason-giving, the platforms where debate takes place create challenges: issues like anonymity and increased polarization make an ideal deliberative process in online debates, at least, challenging (Hartz-Karp and Sullivan, 2014). Twitter allows for anonymity and therefore the quality of the discussion is lower in terms of some users being more uncivil while others experiencing more harassment. As a consequence, social media has grown as a platform for both political communication and incivility (Munger, 2017*a*).

And second, homophily is a major reason to remain skeptical about the premise that social media platforms will enhance the deliberative process. Citizens participating in

[14] One sub-branch of online deliberation research is actually dedicated to developing new platforms that facilitate deliberative experiences that surpass currently available options (Chaudoin and Tingley, 2017; Muhlberger, 2005)

online discussions about politics tend to prefer discussions with like-minded individuals; the extent of online "echo chambers" remains hotly debated, but some degree of homophily is undeniable. Hence, online deliberation may become a way of reinforcing preexisting views and not a real platform for reciprocal argumentation, in the sense of Habermas (1996).

Our experiment connects these disparate threads in the literature. By tweeting at people interested in the peace process and using different types of arguments in favor of the deal, we aim to encourage deliberation. We test under what circumstances the combination of content/sender is most effective at promoting discussion.

5 Experimental Design

We conducted a field experiment on Twitter during the 2016 Colombian Plebiscite. We first collected all tweets related to the peace process from March through September of 2016. We identified accounts that were been active on this topic two months prior to the plebiscite. Using text analysis and machine learning techniques, we constructed a sentiment score for each account, which specifies if the account supports or opposes the process.[15] We then selected a random sample of 4,500 of these accounts. Using block randomization, with two blocks differentiating between supporters and opponents, we constructed seven groups (six treatment groups and a control group).

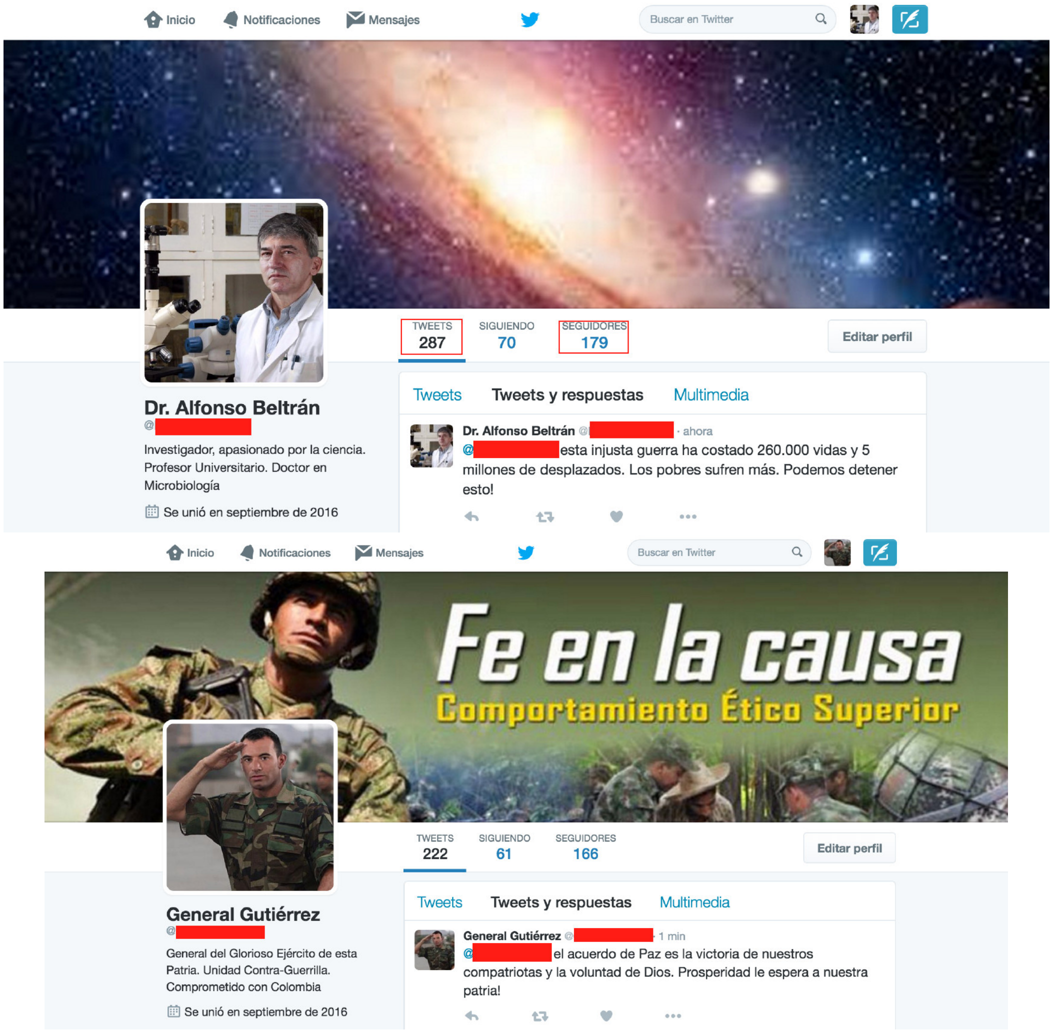
The actual experimental manipulation was to send public messages to subjects. All of the messages were in favor of the peace process, for reasons we discuss below. We varied the treatment on two dimensions: the identity of the sender and the ideological framing of the message. To manipulate identity, we created "bots" that had public profiles identifying them as one of three figures: a general, a priest, or a scientist. Figure 3 shows the accounts of the liberal scientist and the conservative general.

We sent two types of messages: a conservative message that emphasized typical conservative values such as patriotism, authority, and sanctity; and a liberal message that emphasized liberal values such as harm, fairness, and reciprocity. We rotated through the bots and tweeted the messages:

Conservative: "`@[subject] The peace agreement is a victory of our compatriots and the will of God. Prosperity awaits for our homeland`"

[15] Details of this process can be found in Appendix 1.

Figure 3: Treatments—Scientist (Liberal Message) and General (Conservative Message)



Dr. Alfonso Beltrán
@[redacted]

Investigador, apasionado por la ciencia. Profesor Universitario. Doctor en Microbiología

Se unió en septiembre de 2016

Dr. Alfonso Beltrán @[redacted] · ahora
@[redacted] esta injusta guerra ha costado 260.000 vidas y 5 millones de desplazados. Los pobres sufren más. Podemos detener esto!

General Gutiérrez
@[redacted]

General del Glorioso Ejército de esta Patria. Unidad Contra-Guerrilla. Comprometido con Colombia

Se unió en septiembre de 2016

General Gutiérrez @[redacted] · 1 min
@[redacted] el acuerdo de Paz es la victoria de nuestros compatriotas y la voluntad de Dios. Prosperidad le espera a nuestra patria!

Liberal: "`@[subject] This war has taken 260,000 lives and 5 million missing. The poor suffer more.  We can stop this`"

Note that while these two messages emphasize different values, they both argue in favor of the peace process. Although we could have varied this dimension and included messages that argued against the peace process, ultimately, we decided not to. Including this variation would have meant that our treatments would have varied in three ways: type of bot, content of the message, and political position of the message. We would have needed a larger sample size to support this $3 \times 2 \times 2$ design, or sacrifice one of the other two dimensions. Given the hypotheses that we wanted to test, we opted to sacrifice the political position dimension and keep it constant throughout our treatments. There is also an ethical consideration to this decision—all else equal, it is better to minimize the amount of deception involved in an experiment like this, and we were ourselves in favor of the peace process. The messages we sent were factual, and we would happily have sent them from personal accounts; the only deception was in the identities of the bots.[16]

6 Data

We first need to know whether the subject replied directly to the bot's tweet. This behavior indicates whether or not the subject accepts the message and sender as a legitimate authority, and could explain the mechanism by which future behavior changes. Overall, 158 subjects (4% of those in a treatment group) sent a tweet directly in reply to our bots. We coded these as either positive or negative reactions.

The primary behavior targeted in this experiment is the frequency and sentiment of tweets about the peace process. To capture this behavior, we scraped each subject's Twitter history before and after the treatment and restricted the sample to the tweets that were about the peace process. We used a conservative approach to identifying these tweets: a dictionary of popular phrases and hashtags. Any tweet containing one of these key terms was coded as being about the peace process.[17]

[16] The research described in this paper was approved by the IRB at NYU and Universidad de Rosario.

[17] We selected the most popular hashtags related to the discussion of the peace process, as well as terms that tended to co-occur with those hashtags. There are undoubtedly some tweets that we miss with this approach, but unless this classification covaries

To control for each subjects' pre-treatment behavior, we calculated their rate of tweeting about the peace process in the three months before the experiment. This measure was included as a covariate in all of the following analysis.

To test our hypothesis that the effect of our treatments would be moderated by the ideology of the subjects, we need to be able to say which of subjects were liberal or conservative. We implemented the method developed by Barberá (2015) to estimate subjects' ideological ideal points. This was possible for 3,500 of the 4,500 subjects who followed enough Colombian political elites.

In addition to the raw number of tweets about the peace process, we were interested in their orientation: was the tweet in favor of "Si" or "No"? We call this the *sentiment* of the tweet. We began by hand-coding a balanced sample of 2,000 tweets as in favor of "Si" (*pro*) or "No" (*con*). After pre-processing the text of the tweets, we then trained a Support Vector Machine (SVM) on these labeled tweets. SVM is a commonly-used and computationally efficient machine learning technique; our model performed well, with a cross-validated out-of-sample prediction accuracy of 77%.[18] We then applied the trained SVM to the rest of the tweets, generating binary sentiment scores for the 70,000 subject tweets we identified as being about the peace process.

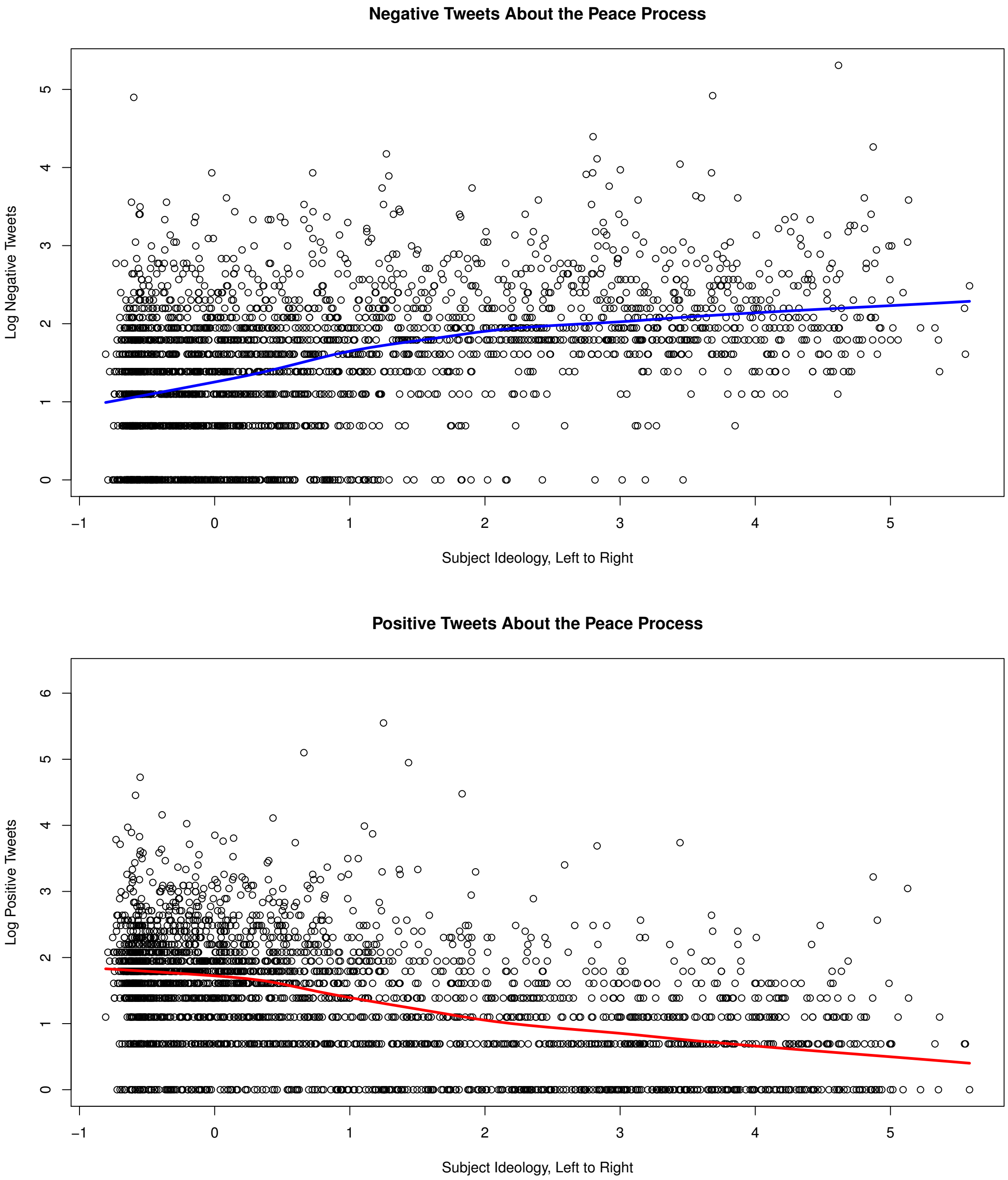
As a validity check of our application of these two machine learning classifiers, we plot the log (plus one) of the number of positive and negative tweets each subject sent about the peace process against their estimated ideology score, in Figure 4. The top panel plots the log number of negative tweets; as expected, liberal subjects (with negative ideology scores) sent far fewer negative tweets about the peace process.

The bottom panel plots the log number of positive tweets, and the trend is reversed: liberal subjects sent more positive tweets about the peace process. In both cases, the trend is steepest (and the points densest) among subjects with ideology scores ranging from -1 to 1. This is because the ideology scores generated by the Barberá (2015) algorithm in this case exhibited a long right tail. Roughly half of the subjects were estimated to be on either side of the 0 midpoint, the correct proportion. However, the

with our randomly assigned treatment, this should not be a problem for our analysis. These are the key terms we selected: "#AdiósALaGuerra","#PazCompleta","Acuerdo Gobierno FARC","Acuerdo FARC","Firma paz","paz Santos","proceso de paz","#PazenColombia","FARC Habana","paz Colombia","acuerdo Habana","#SíALaPaz","Uribe paz","plebiscito si","#ProcesoDePaz","#SiALaPaz","plebiscito paz","gobierno FARC","paz FARC","acuerdo paz","#FirmaDeLaPaz","#EnCartagenaDecimosNo","#FeliSídad","plebiscito no","#Plebiscito","diálogos de paz","#SantosElTal23NoExiste","Habana paz","negociaciones Habana","conversaciones Habana","Gobierno paz FARC","Gobierno Habana FARC","#NoMarcho"

[18] Details about the implentation of the SVM model can be found in Appendix 1.

Figure 4: Validating Tweet Sentiment and Estimated Subject Ideology



network of conservatives was much more segregated than that of the liberals, enabling the algorithm to give finer-grained estimates for extreme conservatives.

As a result, we cannot use these continuous ideology scores as covariates in the model: the marginal change in the subjects' ideology is not constant throughout the range of this variable. A change from -1 to 1 indicates a switch from a liberal to a moderate conservative subject, but a change from 4 to 5 indicates a switch from an extreme conservative to an even more extreme conservative. We thus create a categorical Ideology Score variable that takes the value 0 for subjects estimated below the 25th percentile ("Liberals"); the value 1 for subjects between the 25th and 75th percentile ("Moderates"); and the value 2 for subjects above the 75th percentile ("Conservatives").

7 Results

7.1 Direct Replies

We first analyze the direct replies to the bots' tweets. For this purpose, keeping constant the identity of the bot, we estimate the effect of a liberal message —versus a conservative one— on the probability of reacting to the direct mention made to each account. Formally, for each bot k, where k = General, Priest, Scientist, we estimate models of the type:

$$Reaction_i = \beta_{k0} + \beta_{k1} Liberal\_Message_i + \varepsilon_i$$

where $Reaction_i$ is a dummy variable indicating whether subject i reacts to the message sent by the bot or not, $Liberal\_Message_i$ is a dummy variable that indicates if subject i received a liberal message from bot k, and ε_i is the error term. We estimate separate models for any type of reaction, as well as for exclusively positive or negative reactions. Positive reactions correspond to likes, retweets, or positive replies to the bot. On the other hand, negative reactions are associated with negative replies.[19] The coefficients of interest in this set of regressions are β_{k1}. If this coefficient is positive for bot k, it means that subjects tweeted by such bot tend to respond more (positively or negatively) when the message has a liberal content, as compared to the conservative message.

[19] Manual coding for these replies was performed, to determine whether the subject responded positively or negatively to the bot.

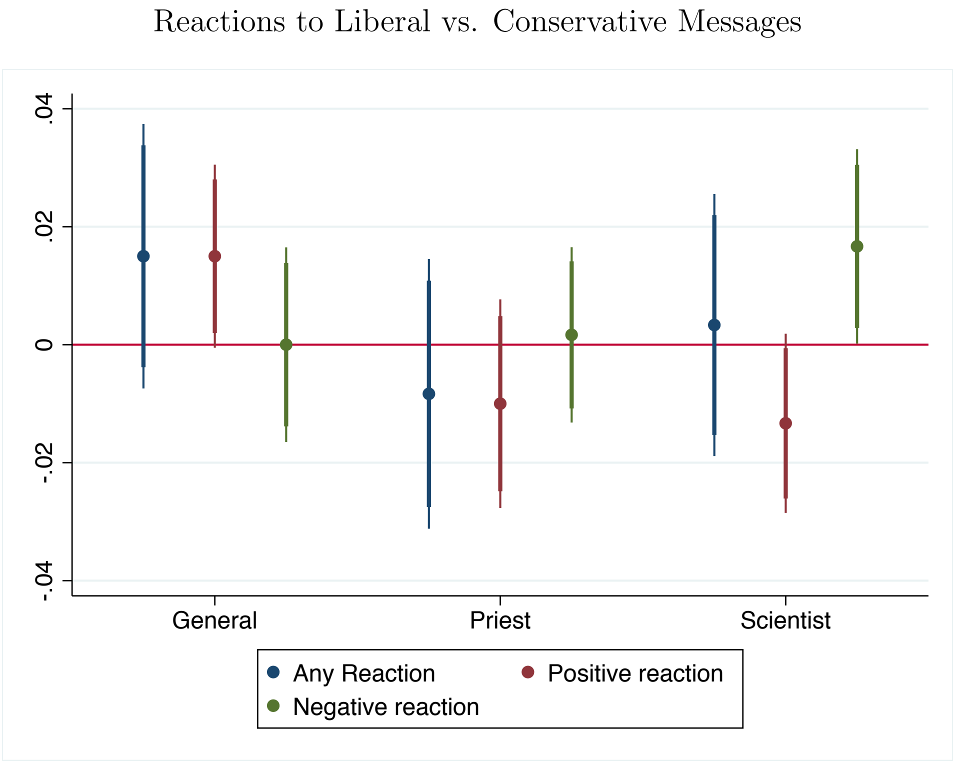


Figure 5: Each of the three outcomes are the result of a separate OLS regression. Values above 0 in the figure mean that outcome was more likely to be caused by a message from the liberal version of that type of bot, while values below 0 indicate higher likelihood for the conservative version.

Keeping constant the identity of the bot, these coefficients indicate if liberal or conservative messages produce more reactions.

Figure 5 plots the regression coefficients —and the associated confidence intervals— of these models for the liberal versus conservative versions of each of the three types of bots. Each of the three outcomes in the Figure (any reaction, positive reaction and negative reaction) are the result of a separate OLS regression—results are substantively the same if a Logit model is used instead. Values above 0 in Figure 5 mean that outcome was more likely to be caused by the liberal message that type of bot, while values below 0 indicate higher likelihood for the conservative message.

The results in Figure 5 indicate that the liberal general caused more positive reactions than the conservative general, and that the liberal scientist caused fewer positive reactions and more negative reactions. In both cases, then, the bots that sent messages "against type" (liberal messages sent by the general and conservative messages sent by the scientist) were more likely to engender positive reactions than messages "with

Heterogeneous Effects: Liberal vs. Conservative Messages from the General

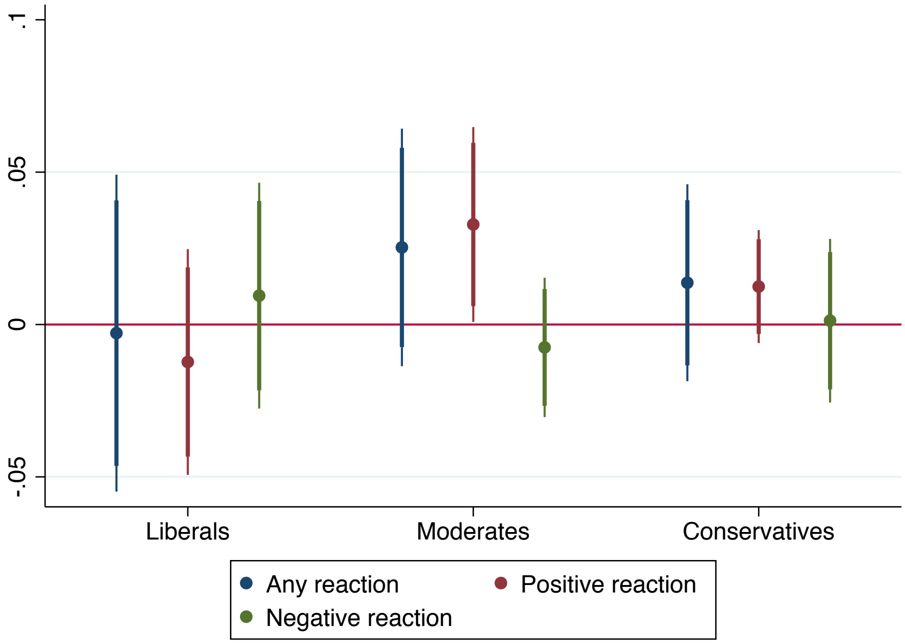


Figure 6: Each of the three outcomes are the result of a separate OLS regression. Values above 0 in the figure mean that outcome was more likely to be caused by a message from the liberal version of that type of bot for each subgroup of subjects, while values below 0 indicate higher likelihood for the conservative version.

type." As we expected, there were no differential effects of the liberal priest compared to the conservative priest. In order to understand the channels driving these results, we disaggregate the effects of these messages along the ideology dimension: we test whether there are differential effects for liberal, moderate, and conservative subjects.

The results in Figures 6, 7, and 8 represent heterogeneous effects at the ideology level. These results reflect that the positive effects of liberal messages sent by the General are mainly driven by moderate subjects (Figure 6). Additionally, the increase in negative reactions to liberal messages sent by the scientist are driven by conservative subjects disliking these messages (Figure 8). Finally, in the case of the priest, conservative subjects are more likely to react positively when they receive a conservative message from this type of bot. Note that in some cases there is no point estimate. This occurs when there is no variation in the outcome variable. For example, in the case of liberals who received a message from the scientist, none of them had a negative reaction.

Heterogeneous Effects: Liberal vs. Conservative Messages from the Priest

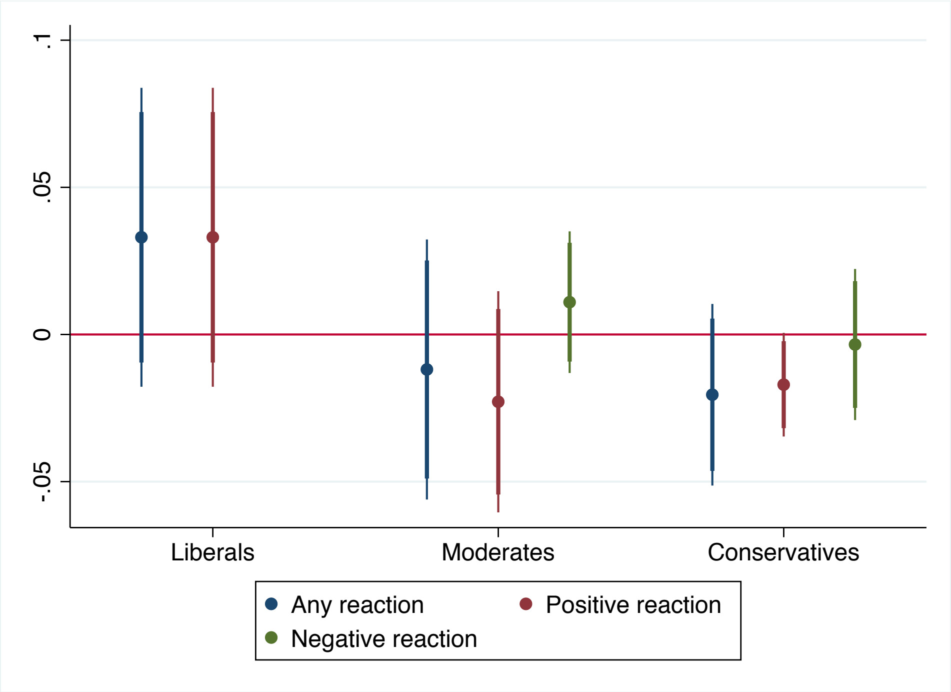


Figure 7: Each of the three outcomes are the result of a separate OLS regression. Values above 0 in Figure mean that outcome was more likely to be caused by a message from the liberal version of that type of bot for each subgroup of subjects, while values below 0 indicate higher likelihood for the conservative version.

Heterogeneous Effects: Liberal vs. Conservative Messages from the Scientist

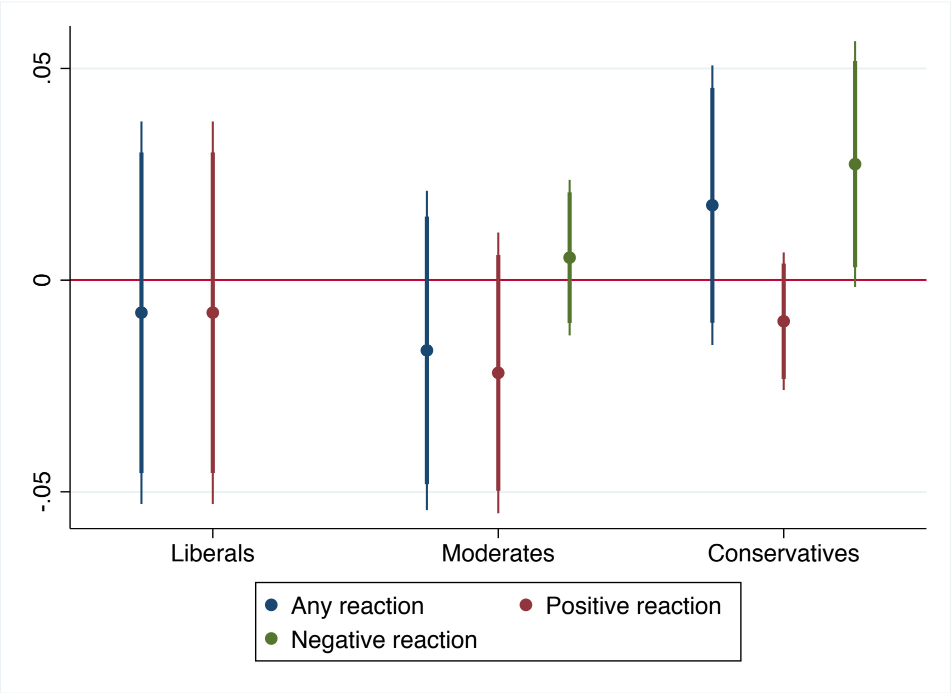


Figure 8: Each of the three outcomes are the result of a separate OLS regression. Values above 0 in the figure mean that outcome was more likely to be caused by a message from the liberal version of that type of bot for each subgroup of subjects, while values below 0 indicate higher likelihood for the conservative version.

Overall, these results reveal that in general conservative subjects are the ones who react differentially to these messages and tend to respond directly to our bots, especially depending if the message is "against" of "with" type. They dislike liberal messages sent by the scientist–relative to conservative messages from same—and are more likely to react in a positive way when contacted by the conservative priest compared to the liberal one.

7.2 Post-Treatment Sentiment and Peace Tweets

Our main analysis uses the subjects' of pre- and post-treatment tweets, categorized as discussed in the "Data" section above as pertaining to the peace process and being either positive or negative about the vote. These data are count data, so OLS would be inappropriate. A chi-squared test indicates that the counts are overdispersed, so following Munger (2017*a*), we used negative binomial regression. The negative binomial specification is estimated using the following model:

$$ln(Tweets_{post}) = x_{int} + \beta_1 Tweets_{pre} + \beta_2 T_{priest} + \beta_3 T_{soldier} + \beta_4 T_{scientist} + \beta_5 Ideology$$

$$+\beta_6(T_{priest} \times Ideology) + \beta_7(T_{soldier} \times Ideology) + \beta_8(T_{scientist} \times Ideology)$$

To interpret the relevant treatment effects implied by the coefficients estimated by these models, the exponent of the estimated $\hat{\beta}_k$ for each of the treatment conditions needs to be added to the corresponding $\hat{\beta}$ for the interaction term, evaluated at each level of Ideology Score (Hilbe, 2008). For example, the effect of the Priest treatment on Conservative subjects (Ideology score 2) is:

$$IRR_{priest \times Ideology_2} = e^{\hat{\beta}_2 + \hat{\beta}_6 \times 1}$$

Before presenting the results on post-treatment peace tweets, it is important to acknowledge that we also estimated this model using as an outcome variable the post-treatment sentiment score of our subjects, to test if the intervention can persuade people in favor of the accord. We find null effects for all of our treatments; in no case we are able to alter mean sentiment scores (results not shown). This is not surprising, given the highly polarized context in which the experiment took place and the low intensity of our intervention. However, we do find effects on post-treatment tweeting behavior.

The experimental results on the sample of liberal subjects for each of the first four weeks (excluding day 1, in which there were many direct reactions to the tweets) after treatment are displayed in Figure 9; in all of the analysis that follows, the dependent variable is the number of tweets (either positive or negative) the subject sent in the specified time period.

$IRR_{scientist \times Ideology_0} \approx 1.5$, the effect of the conservative scientist treatment on liberal subjects during the first week after the intervention, can be seen in the black line in the left section of the plot. This Incidence Ratio implies that the average subject with Ideology Score 0 (liberal) who received the conservative scientist treatment tweeted about 150% as many positive tweets about the peace process as the average subject with Ideology Score 0 in the control condition.[20] The confidence intervals in Figure 9 are calculated from the estimated variance of this estimator:

$$V_{priest \times Ideology_1} = V(\hat{\beta}_2) + Ideology^2 V(\hat{\beta}_6) + 2Ideology \times Cov(\hat{\beta}_2 \hat{\beta}_6)$$

These are ratios: going from .5 to 1 represents the same effect size (a 100% increase) as going from 1 to 2, so the upper half of the confidence intervals appear longer than the lower half. Also, recall that the Liberal and Conservative samples each comprise 25% of the overall sample compared to 50% for the moderate sample. Because the sample is twice as big, the standard errors for the moderate sample are smaller.

In general, all six of the treatment conditions had similar effects on liberals' rate of sending positive tweets: Liberal respondents were encouraged to send more positive tweets during the first week after treatment. As can be seen on the right hand portion of this figure, this effect disappears after the first week.

Figure 10 shows that our treatment conditions have no effect, in general, on the subsample of moderate subjects. In the case of conservative users, as shown in Figure 11, for the first week the point estimates are negative but non-significant. In sum, our treatments encourage liberals to tweet more about the peace process during the first week, but no effects are produced on conservatives or moderates at any given time.

[20] Note that this approach assumes that treatment effects are constant, and holds the pre-treatment level of pre-treatment tweets about the peace process constant at its mean level.

Positive Tweets About the Peace Process From Liberals (*N=3,516*)

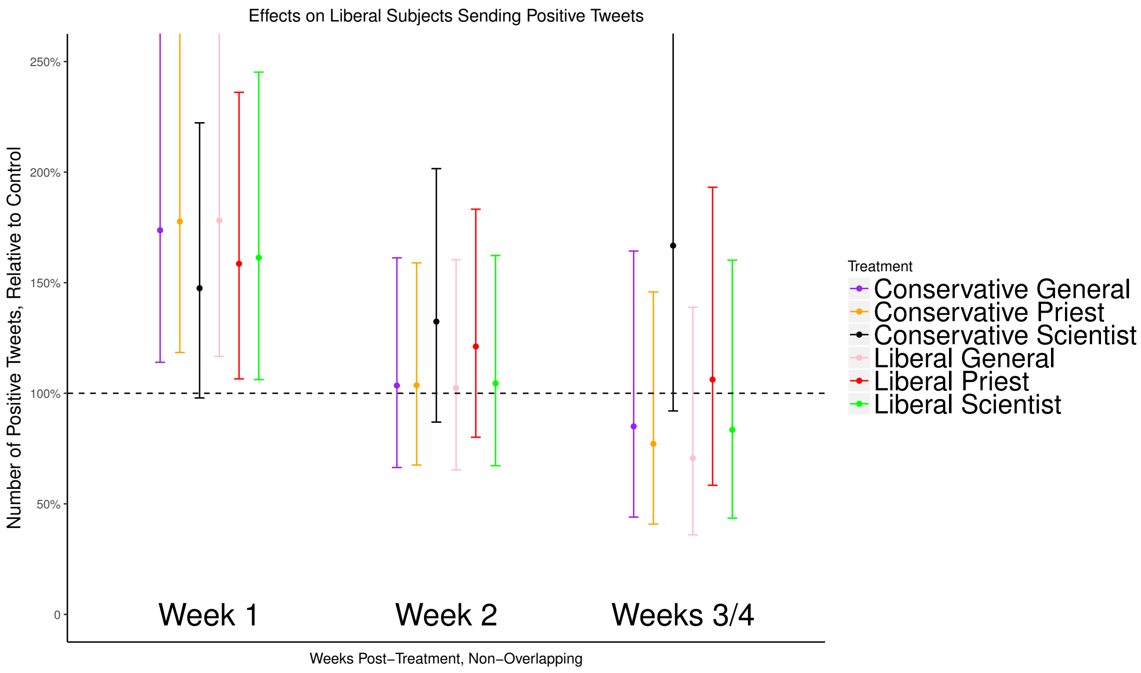


Figure 9: The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model in the first week after treatment, excluding day 1. For example, the Incidence Ratio of 1.3 associated with the liberal priest treatment on subjects with Ideology Score 1 in the middle of the plot means that these subjects sent 130% as many positive tweets about the peace process as the subjects with Ideology Score 1 in the control group. The bars represent 95% confidence intervals.

Positive Tweets About the Peace Process From Moderates(*N=3,516*)

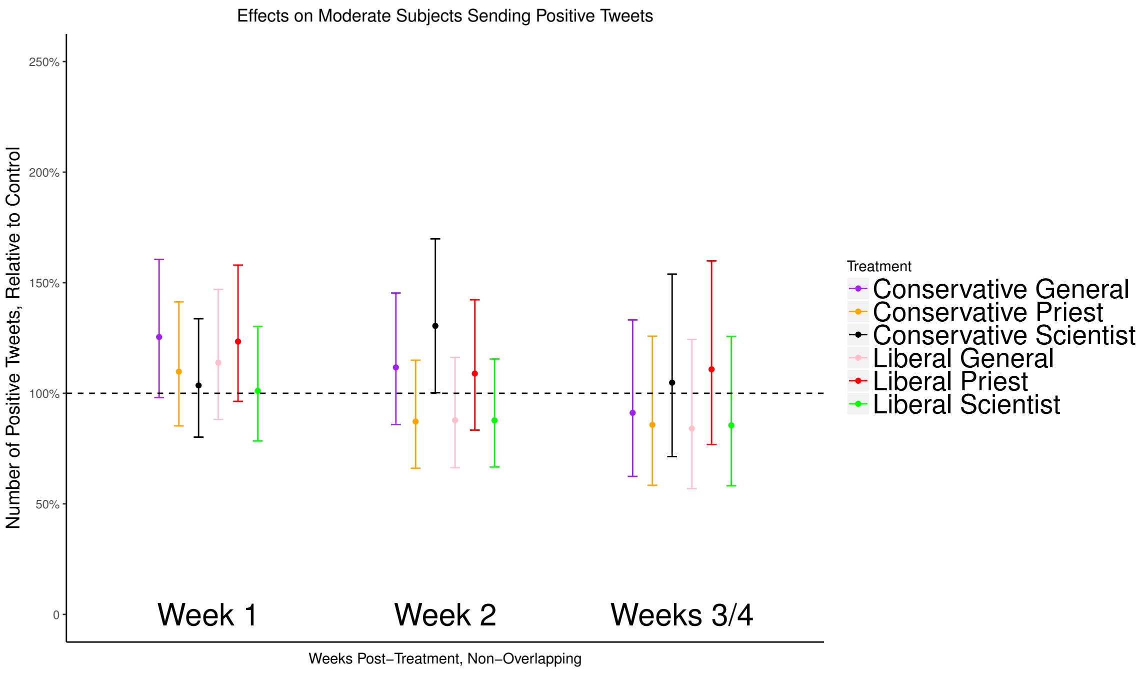


Figure 10: The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model in the first week after treatment, excluding day 1. For example, the Incidence Ratio of 1.3 associated with the liberal priest treatment on subjects with Ideology Score 1 in the middle of the plot means that these subjects sent 130% as many positive tweets about the peace process as the subjects with Ideology Score 1 in the control group. The bars represent 95% confidence intervals.

Positive Tweets About the Peace Process From Conservatives (*N=3,516*)

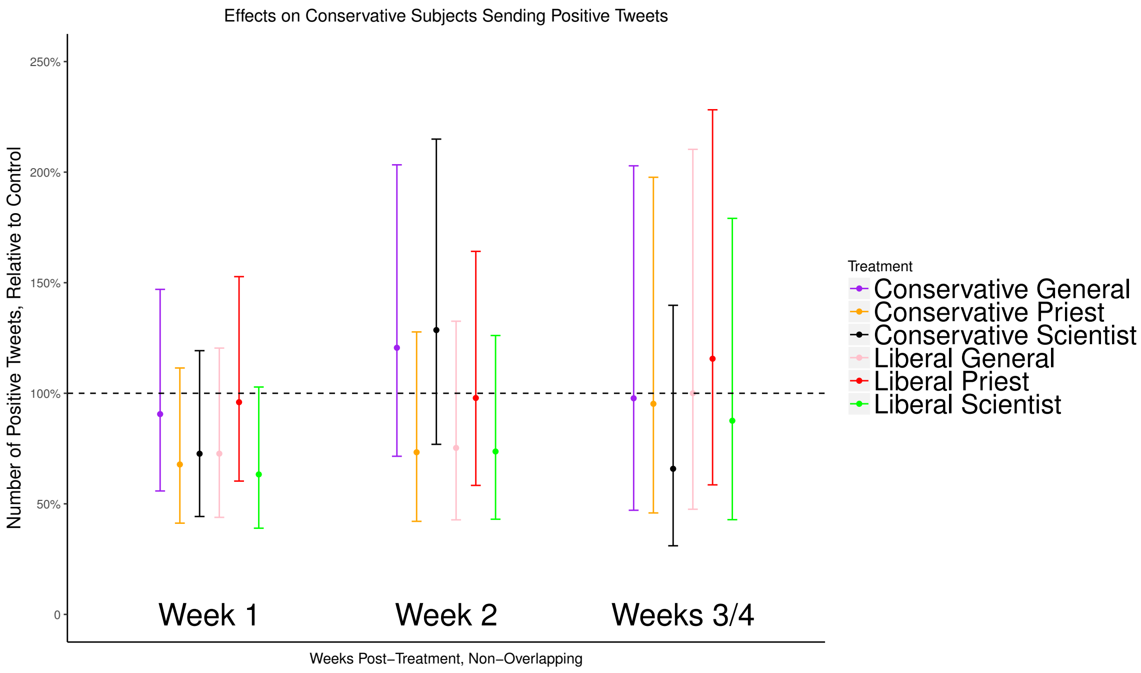


Figure 11: The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model in the first week after treatment, excluding day 1. For example, the Incidence Ratio of 1.3 associated with the liberal priest treatment on subjects with Ideology Score 1 in the middle of the plot means that these subjects sent 130% as many positive tweets about the peace process as the subjects with Ideology Score 1 in the control group. The bars represent 95% confidence intervals.

We also need to see if our interventions caused any change in the rate of sending *negative* tweets—tweets that argued against the peace process. Figures 12-14 plot those results.

Encouragingly, across all treatment conditions interacted with subject ideology and the post-intervention week analyzed, only three showed a statistically significant increase in the rate of sending negative tweets about the peace process—precisely the number that we would expect to see by chance. Again, these results support our argument that conservatives do not feel encouraged to send more negative tweets, and perhaps—with the exception of those who reply directly to the bots–simply ignore the messages. Overall, we do in fact find evidence that we were able to promote deliberation. But only among those that to begin with are aligned with the position defended by the bots and independent of its identity and moral values used. After confirming their original preconception, they simply tweet more about it.

Negative Tweets About the Peace Process From Liberals (*N=3,516*)

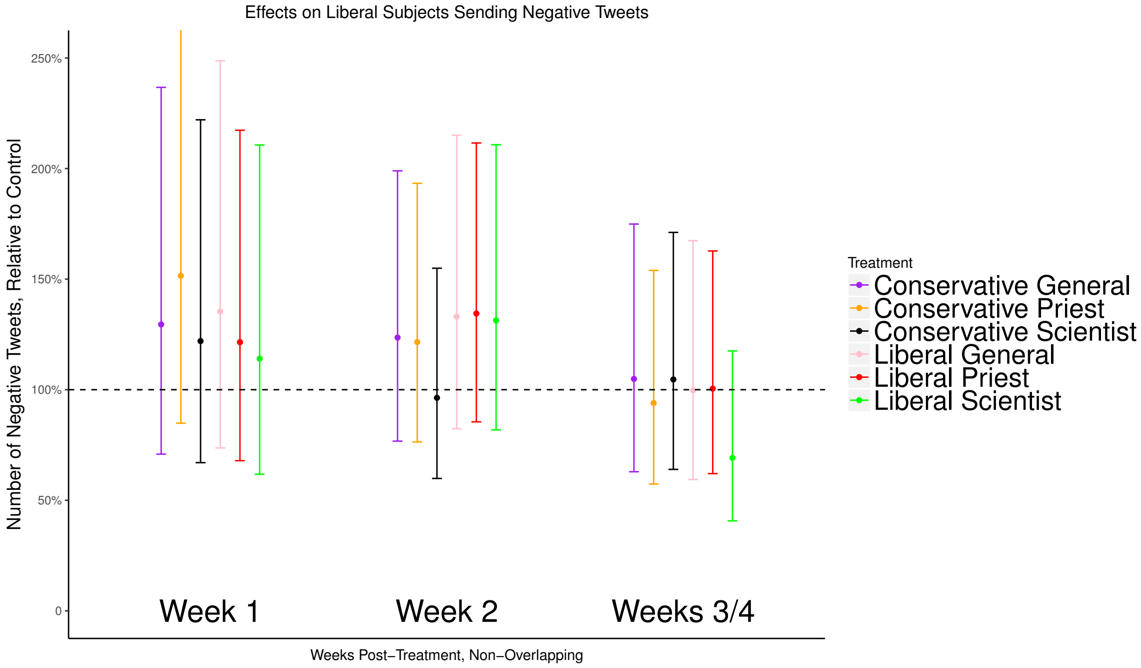


Figure 12: The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model in the time period from 2 to 7 days after treatment. For example, the Incidence Ratio associated with the liberal priest treatment on subjects with Ideology Score 1 in the middle of the plot means that these subjects sent 140% as many tweets about the peace process as the subjects with Ideology Score 1 in the control group. The thick bars represent 90% confidence intervals and the thin lines represent 95% confidence intervals.

Negative Tweets About the Peace Process From Moderates (*N=3,516*)

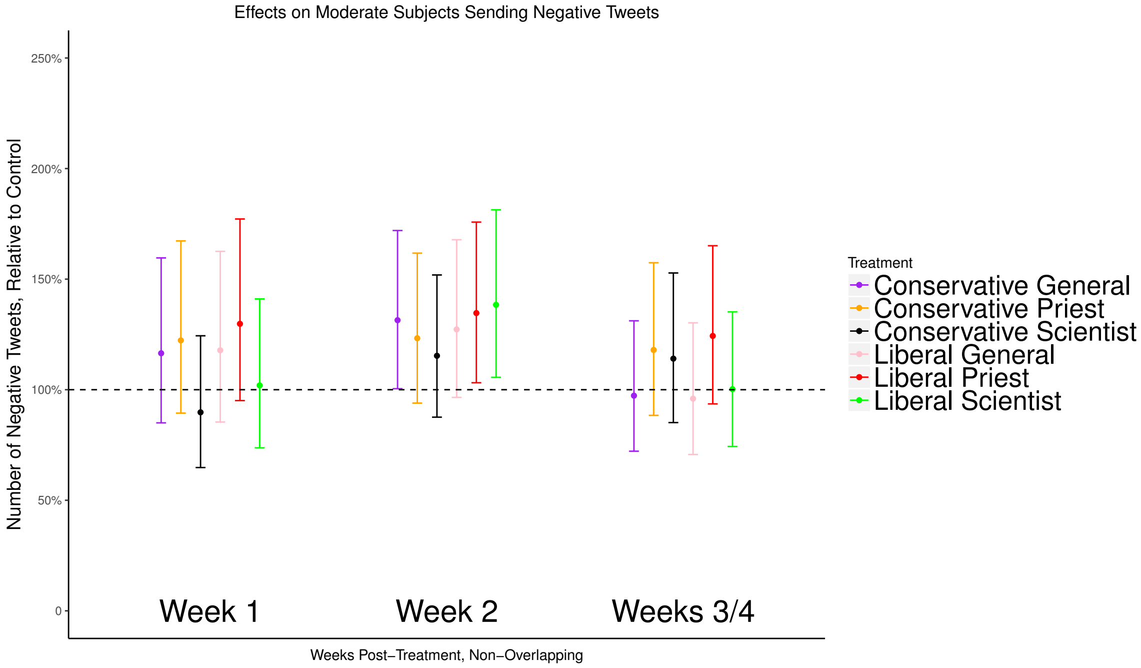


Figure 13: The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model in the time period from 2 to 7 days after treatment. For example, the Incidence Ratio associated with the liberal priest treatment on subjects with Ideology Score 1 in the middle of the plot means that these subjects sent 140% as many tweets about the peace process as the subjects with Ideology Score 1 in the control group. The thick bars represent 90% confidence intervals and the thin lines represent 95% confidence intervals.

Negative Tweets About the Peace Process From Conservatives (N=*3,516*)

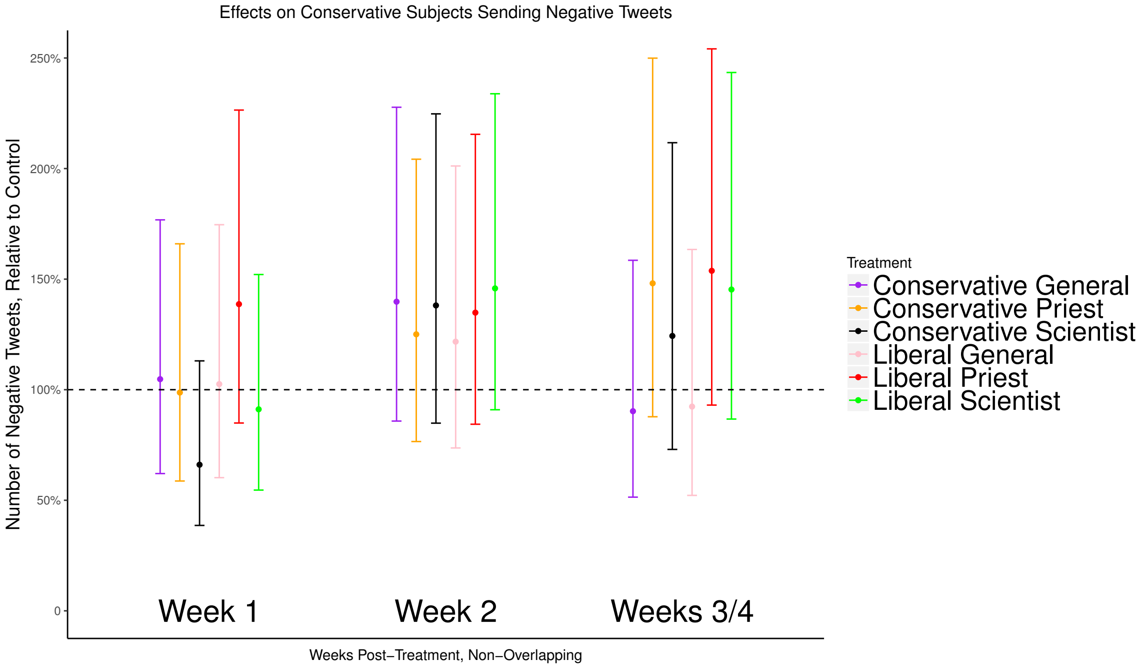


Figure 14: The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model in the time period from 2 to 7 days after treatment. For example, the Incidence Ratio associated with the liberal priest treatment on subjects with Ideology Score 1 in the middle of the plot means that these subjects sent 140% as many tweets about the peace process as the subjects with Ideology Score 1 in the control group. The thick bars represent 90% confidence intervals and the thin lines represent 95% confidence intervals.

8 Conclusions

We performed a randomized experiment on Twitter users who we identified as interested in the peace process in Colombia. To our knowledge, this is the largest-scale social science Twitter bot experiment conducted to date. We tested several hypotheses about persuasion and deliberation. To do so, ahead of the plebiscite we sent public messages to users encouraging them to support the peace process, varying the identity of the information source and the content. We sent two types of messages, a conservative message and a liberal message, from three different accounts, namely that of a scientist, a priest and a general, which are respected non-political public figures in Colombia.

With this experiment we aimed to test hypotheses related to political persuasion and public deliberation. Our goal was to learn if public figures and messages more aligned with subjects ideological preconceptions, would be more effective at encouraging people support and talk more about the peace process.

Our results show that we could not persuade subjects to change their minds to support the peace process. We are not surprised by these null effects, similar to findings in alternative contexts (Broockman and Green, 2014; Kalla and Broockman, 2017). The peace plebiscite took place in a highly polarized environment and the result of the elections reflects it. Information flows ahead of the election were massive, so that it would be hard to change people's opinions with a simple tweet.

But we also have evidence that certain combinations of senders and messages might backfire, as conservatives had differential reactions when approached by a liberal scientist or a conservative priest. Hence, in terms of the moral reframing theory (Feinberg and Miller, 2015; Volkel and Feinberg, 2016), we have learned that not only the content of messages matter, but also the identity of the sender and if it is aligned with the receiver's ideological position.

In terms of deliberation, the fact that our treatments encouraged certain group of subjects to tweet positively more about the peace process could be interpreted as good news. However, a careful inspection of these results reveals that the effect is driven by liberal subjects who were encouraged by *any* type of pro-peace message, independent of sender and content. We interpret this results as evidence of a confirmation bias effect (Sunstein, 2002). People talk more when they read what they want to read.

In many ways, the election studied in this paper is quite unusual. A plebiscite to endorse the agreement signed by a central government and a guerrilla group in a developing country, is a rare event. However, many of the characteristics of this election resemble what has happened –an is going to happen– elsewhere. A deeply polarized society in which social media, elites, and public figures play a key role at shaping citizens' opinions and their subsequent political decisions. In such context, non-political public institutions, like the ones used in this experiment, might increase the debate and get people to deliberate more, but in highly unexpected ways.

References

Aarøe, Lene. 2011. "Investigating frame strength: The case of episodic and thematic frames." *Political Communication* 28 (2): 207–226.

Barberá, Pablo. 2015. "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data." *Political Analysis* 23 (1): 76–91.

Black, Laura, Howard Welser, Dan Cosley, and Jocelyn DeGroot. 2011. "Self-governance through group discussion in Wikipedia: Measuring deliberation in online groups." *Small Group Research* 42 (5): 595–634.

British Council. 2017. "Next Generation: Aplificando la Voz de los Jóvenes en Colombia." Reporte Preliminar.

Broockman, David, and Donald Green. 2014. "Do Online Advertisements Increase Political Candidates' Name Recognition or Favorability? Evidence from Randomized Field Experiments." *Political Behavior* 36 (2): 263–289.

Campbell, Angus. 1960. *The American Voter.* University of Chicago Press.

Caprara, Gian, Shalom Schwartz, Cristina Capanna, and Claudio Barbaranelli. 2006. "Personality and Politics: Values, Traits, and Political Choice." *Political Psychology* 27 (1): 1–28.

Chaudoin, S., Shapiro J., and D. Tingley. 2017. "Revolutionizing Teaching and Research with a Structured Debate Platform." *Working Paper* .

CHCV. 2015. "Comisión Histórica del Conflicto y sus Víctimas: Contribución al entendimiento del conflicto armado en Colombia.".

Chong, Dennis, and James N Druckman. 2007. "Framing theory." *Annu. Rev. Polit. Sci.* 10: 103–126.

Coleman, Stephen, and Jay Blumer. 2009. *The Internet and Democratic Citizenship: Theory, Practice and Policy.* Cambridge University Press.

Dickson, Eric S, Catherine Hafer, and Dimitri Landa. 2008. "Cognition and strategy: a deliberation experiment." *The Journal of Politics* 70 (4): 974–989.

Druckman, James N. 2014. "Pathologies of studying public opinion, political communication, and democratic responsiveness." *Political Communication* 31 (3): 467–492.

Druckman, James N, and Thomas J Leeper. 2012. "Learning more from political communication experiments: Pretreatment and its effects." *American Journal of Political Science* 56 (4): 875–896.

Duncan, Gustavo. 2015. "Exclusioin, insurreccion y drimen." *Contribución al entendimiento del conflicto armado en Colombia* .

Eveland, William. 2004. "The Effect of Political Discussion in Producing Informed Citizens: The Roles of Information, Motivation, and Elaboration." *Political Communication* 21 (2): 177–193.

Feinberg, Matthew, and Robb Miller. 2015. "From Gulf to Bridge: When Do Moral Arguments Facilitate Political Influence." *Personality and Social Psychology Bulletin* 41 (12): 1665–1681.

Graham, Jesse, Brian A. Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H. Ditto. 2011. "Mapping the Moral Domain." *Journal of Personality and Social Psychology* 101 (2): 366–385.

Graham, Jesse, Jonathan Haidt, and Brian Nosek. 2009. "Liberals and Conservatives Rely on Different Sets of Moral Foundations." *Journal of Personality and Social Psychology* 96 (5): 1029–1046.

Graham, Todd. 2015. *Everyday Political Talk in the Internet-Based Public Sphere.* Edward Elgar Publishing chapter 14.

Habermas, Jurgen. 1996. *Between facts and norms. Contributions to a discourse theory of law and democracy.* MIT Press.

Halpern, Daniel, and Jennifer Gibbs. 2013. "Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression." *Computers in Human Behavior* 29 (3): 1159–1168.

Hartz-Karp, Janette, and Brian Sullivan. 2014. "The unfulfilled promise of online deliberation." *Journal of Public Deliberation* 10 (1).

Hilbe, Joseph M. 2008. "Brief overview on interpreting count model risk ratios: An addendum to negative binomial regression.".

Janoff-Bulman, Ronnie, Sana Sheikh, and Kate Baldacci. 2008. "Mapping moral motives: Approach, avoidance, and political orientation." *Journal of Experimental Social Psychology* 44 (4): 1091–1099.

Janssen, Davy, and Raphaël Kies. 2005. "Online forums and deliberative democracy." *Acta política* 40 (3): 317–335.

Johnston, Pamela, Conover, Donald, Searing, and Ivor Crewe. 2001. "The Deliberative Potential of Political Discussion." *British Journal of Political Science* 32 (1): 21–62.

Kalla, Joshua, and David Broockman. 2017. "The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments." *American Political Science Review* .

Kernell, Georgia. 2013. The Scope of the Partisan ?Perceptual Screen? PhD thesis Northwestern University.

Kim, Joohan, Robert Wyatt, and Elihu Katz. 1999. "News, Talk, Opinion, Participation: The Part Played by Conversation in Deliberative Democracy." *Political Communication* 16 (4): 361–385.

Lakoff, George. 2002. *Moral Politics: How Liberals and Conservatives Think.* University of Chicago Press.

Lupia, Arthur. 1994. "Shortcuts versus encyclopedias: information and voting behavior in California insurance reform elections." *American Political Science Review* 88 (01): 63–76.

Lupia, Arthur. 2015. *Uninformed: Why people seem to know so little about politics and what we can do about it.* Oxford University Press.

Lupia, Arthur, and D McCubbins. 1998. *The Democratic Dilemma: Can citizens learn what they need to know?* Cambridge University Press.

Morgan, Scott, Linda Skitka, and Daniel Wisneski. 2010. "Moral and Religious Convictions and Intentions to Vote in the 2008 Presidential Election." *Analyses of Social Issues and Public Policy* 10 (1): 307–320.

Muhlberger, Peter. 2005. "The Virtual Agora Project: A research design for studying democratic deliberation." *Journal of Public Deliberation* 1 (1).

Munger, Kevin. 2017*a*. "Don?t@ Me: Experimentally Reducing Partisan Incivility on Twitter.".

Munger, Kevin. 2017*b*. "Tweetment effects on the tweeted: Experimentally reducing racist harassment." *Political Behavior* 39 (3): 629–649.

Papacharissi, Zizi. 2004. "Democracy online: Civility, politeness, and the democratic potential of online political discussion groups." *New media & society* 6 (2): 259–283.

Papacharissi, Zizi. 2009. "The virtual geographies of social networks: a comparative analysis of Facebook, LinkedIn and ASmallWorld." *New media & society* 11 (1-2): 199–220.

Robertson, Scott, Ravi Vatrapu, and Richard Medina. 2010. "Off the wall political discourse: Facebook use in the 2008 U.S. presidential election." *Information Polity* 15: 11–31.

Sunstein, Cass. 2002. *Republic.com.* Princeton University Press.

Thorisdottir, Hulda, John Jost, Ido Liviatan, and Patrick Shrout. 2007. "Psychological Needs and Values Underlying Left-Right Political Orientation: Cross-National Evidence from Eastern and Western Europe." *Public Opinion Quarterly* 71 (2): 175–203.

Volkel, Jan, and Matthew Feinberg. 2016. "Morally Reframed Arguments Can Affect Support for Political Candidates." Working Paper.

Wills, Maria E. 2015. "Los tres nudos de la guerra colombiana." *Contribucin al entendimiento del conflicto armado en Colombia* .

Yardi, Sarita, and Danah Boyd. 2010. “Dynamic Debates: An Analysis of Group Polarization Over Time on Twitter.” *Bulletin of Science, Technology and Society* 30 (5): 316–327.

A Details of constructing the sentiment classifier

The data collection process was the fundamental element of the experiment, given the fact that the identification of the subjects, its sentiment and its ideology depended purely on evidence related with data from twitter. Twitter grants access to the public interested in its data through an API which can be used to scrape and organize tweets from a user, a hashtag, a topic or even a location. The methodology of data collection was focused on fetching data from twitter based on a dictionary of words and n-grams related to the peace process in Colombia. This dictionary of terms included words such as: *timochenko,farc*, *plebiscite*. As well as n-grams such as: *proceso de paz*, *santos paz*, *habana paz*, *fin conflicto*, among others. In total, the list of terms had more than 30 entries related to the topic of peace in Colombia.

After the dictionary was built, the list of terms was filled into the twitter API as search keys. Then, the twitter API returned the list of tweets that mentioned the search keys in any possible order. This process, is able to retrieve data in a near-real-time basis, having of course the limit in the number of tweets the user is allowed to download. We programmed an algorithm that repeated this process automatically every hour stacking the tweets in a database. This first stage ran from March 2016 to October 2016, and resulted in a 1million tweets with twitter users regarding the peace process in Colombia. This database contains not only the text of every tweet, but also the name of the user, the twitter user name, the number of retweets, the number of favorites, the self-reported location, the geocoded location (not always available), the biography of the user, among other variables.

This database was used afterwards to identify the users that were sharing more comments about the peace process. The information was filtered in terms of number of tweets, number of followers and location since the target users were supposed to have the following profile: Not too many followers, active in terms of sharing comments about the peace process from July to September of 2016 and located in Colombia. Furthermore, the second purpose of the database was to use labelled tweets as examples to train a machine learning model able to tell the sentiment of a tweet in the context of the peace process of Colombia.

This process of building the machine learning classifier had four main steps: Data cleaning, Data labelling (in this case was not possible to have a labelled data set), feature extraction and selection, training the algorithm and calibration of parameters. The final product is a method to calculate the sentiment score of any tweet from negative to positive. The first step involves making the text of each tweet readable for a computer, this means taking away uncommon symbols, accents, icons, upper case letters and extra spaces. After this, we identify a set of words called stop words inside every tweet and delete them, this is necessary given that not every word contributes information about the sentiment of a tweets, for example: an, any, or, to, the. The third step consists on performing stemming to the words of each tweet. This process seeks to homologate the words with same meaning, but different conjugations, for instance, the word negotiating has the same base meaning as the word negotiated, therefore it would be useful if the computer understands these two words as the same one. In this case, both terms would be converted to its base word or stem, which in this case corresponds to the word negotiate.

The second phase is responsible for the development of a set of examples whose main purpose is to teach the machine learning model to classify correctly. We manually labelled a random set of tweets according its sentiment towards de peace process (positive or negative). As we decided to use labelled examples to feed the machine learning model it should be used a supervised learning algorithm, specifically for this research we choose the Naive Bayes binary classifier.

Before the training stage, the text present inside every tweet needs to be expressed in a structured form (units of observation with a set of characteristics expressed as rows and columns). Our approach is to use the method bag of words, expressing words inside a tweet as dichotomic variables. In this sense, every tweet represents a row of the data frame with as many variables as possible words in a tweet. This means that the number of variables depends on the size of the vocabulary present in the corpus used. At the end of this process we had every tweet expressed as a row of zeros and ones (one if the word appears in the tweet and zero if not).

Finally, having all the text structured in a database, we trained a Naive Bayes binary classifier with the set of labelled tweets. The basic premise of this algorithm is to use the words present in a tweet to estimate the probability that its sentiment is positive or negative. This method is based on the work of Bayes (1763), and was first applied to text classification by Mosteller and Wallace (1964).

$$\hat{c} \quad = \quad argmaxP(c|d) \tag{1}$$

Essentially, the intuition behind the use of Bayes theorem to classify text is to simplify the equation (1) and make a naive assumption regarding the interaction between the words inside a document. In equation (1) $\hat{c}$ expresses the estimated class c given its probability,. Having the Bayes theorem expressed in the equation (2) with d as the document and c as the class (in our case either positive or negative), we can substitute equation (1) into equation (2) to have the expression (3).

$$P(c|d) \quad = \quad \frac{P(d|c)P(c)}{P(d)} \tag{2}$$

$$argmaxP(c|d) \quad = \quad \frac{P(d|c)P(c)}{P(d)} \tag{3}$$

In this sense, since the marginal probability $P(d)$ is equal for all classes, it can be disregarded of the equation (3) and we can simplify the equation to:

$$argmaxP(c|d) \quad = \quad P(d|c)P(c) \tag{4}$$

Then, the probability of class c is given by the multiplication of the prior probability of the class c and the likelihood of the document d given the class c. After this, the document can be represented as a set of features (in this case words). At this point, we assume that the words of a document are independent from each other. This let us express the likelihood as the multiplication of the probabilities of every single word in the document (words expressed as w). Therefore, we are calculating for each word, in a set of a labelled documents, the probability it appears given the class of the document, and then these results are multiplied by the prior probability of the class c as seen in the equation (5). In this case, the training database is used to calculate both the prior probability of each class and the conditional probability of every single word inside our corpus.

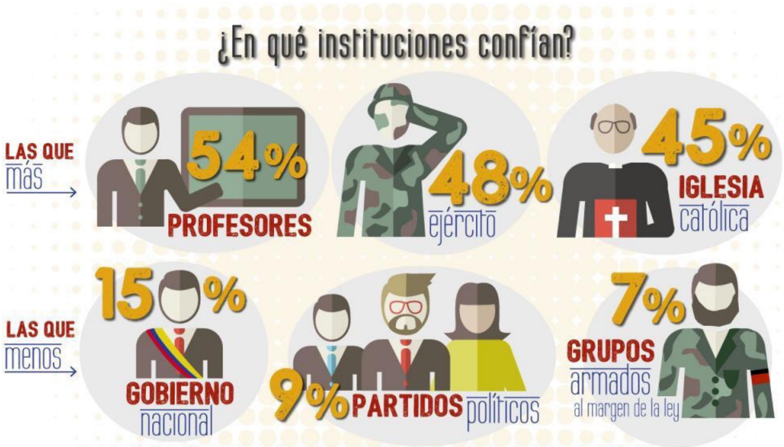
$$\hat{c} = \quad argmax \quad P(c)\prod P(w|c) \tag{5}$$

Ultimately, we used a method called *Cross validation* to evaluate the results and the performance of the classification algorithm. This means that we randomly divided our labelled database in 10 equally sized folds, and then for each one we calculated and evaluated our Naive Bayes classifier. For the purpose evaluating the results, each fold divides into a training sample and a test sample. The test samples provide the ability to compare the true classes for every tweet versus the predicted ones. At the end, to compute the overall precision, a simple average between folds is calculated.

B Trust in Colombian institutions among the youth

Opinion polls like the ones conducted by Gallup reveal that the Catholic Church and the army are among the most trusted institutions and public figures in Colombia. These surveys usually do not include respondents' opinions about professors or scientists in general. However, a recent study conducted on young Colombians (British Council, 2017), reveals that the most trusted institutions, in that order, are professors, the army, and the Catholic Church. These are precisely the public figures used in our experiment and the clear motivation on relying on characters that might influence citizens' opinions and behavior.

Figure 15: Most and Least Trusted Institutions Among Young Colombians



Fuente: Investigación Next Generation Colombia 2017, realizada por British Council con el apoyo de la Universidad de los Andes y la Universidad del Rosario.

Source: British Council (2017). Each percentage represents the proportion of respondents that acknowledge trusting each figure or institution.