

Tweeting for Peace: Experimental Evidence from the 2016 Colombian Plebiscite

Jorge Gallego* Juan D. Martínez †

Kevin Munger ‡ Mateo Vásquez §

Working paper; prepared for APSA 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-governmental elites have played a crucial role in structuring public debate on the topic. To understand the mechanisms underlying elite influence on political discussion, we performed a randomized experiment on Colombian Twitter users shortly before citizens voted in this election. Sampling from a pool of subjects who had been frequently tweeting about the Plebiscite, we tweeted messages from several different accounts to encourage subjects to consider different aspects of the decision. By varying the identity of our accounts (a soldier, a scientist, and a priest) and the content of the messages we sent, we show that some elites are more effective at changing citizens' propensity to participate than are others. The biggest factor in determining how people talked about the Plebiscite, though, was their ideology, suggesting that there are limits to the effectiveness of elite activism.

*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 a conservative, religious vision and a liberal vision 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 goal—peace or justice—as ultimate goal of Colombian politics. The most active debate, then, is to determine which means are best able to balance and achieve these ends.

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 Colombians have long recognized the importance of elites 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 elite influence.

After 60 years of conflict, in which different segments of society became invested, participated or were affected 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 elites can mobilize, inform and change citizens' attitudes about conflict, and especially how these effects interact with citizens' ideology. Do elite endorsements 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 elites are more effective in changing deliberative decisions such as participating in debate or the tone used in political discussions?

This paper addresses these questions in the context of the 2016 Colombian Peace Plebiscite. We argue that messages from liked-minded elites should have caused positive

¹During 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.

reactions and increased engagement from citizens. We are able to demonstrate 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 elites (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 peace process vote—and 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 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. 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 elite 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’

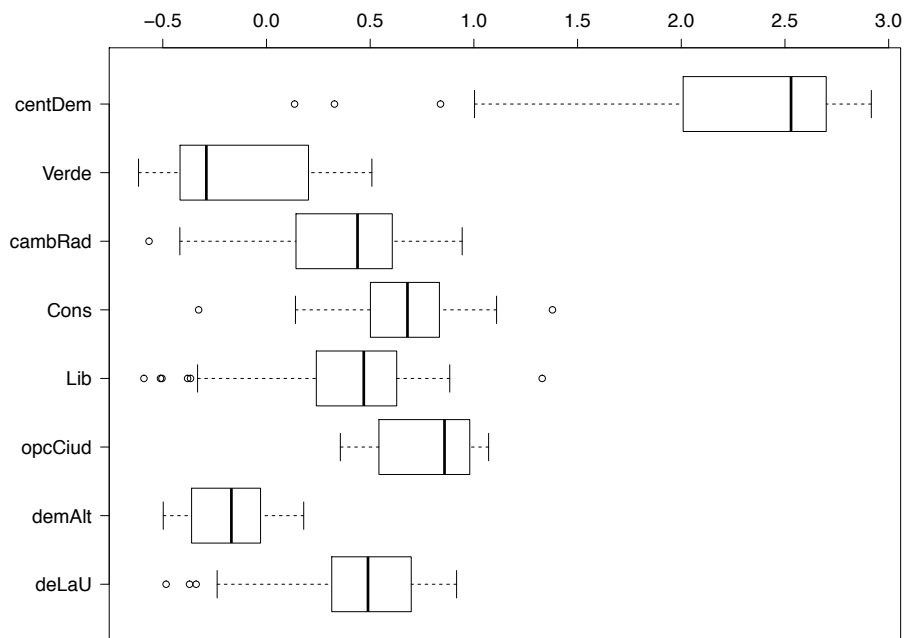
and 49.8% voting ‘Yes’.

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.

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, 2014). 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 Democrático 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 Democrático 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 thier party. These estimates are derived from Barberá (2014)'s Bayesian Spatial Following Model.

2.2 Twitter and Elites 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 utilized 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 in October 2016, about 60% of the respondents had a favorable opinion of the Catholic Church (Gallup, 2017). 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 to 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 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 given their role in the longstanding civil conflict with FARC. Even though the military tend to be associated to values linked to conservative parties, in this case the perception is at least ambiguous, as soldiers represent a large proportion of non-civilians killed by this war, which generates sentiments of sympathy and acknowledgment among Colombians. Finally, even though opinion polls do not ask about citizens' perceptions of scientists, it is well known that academics tend

²According 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).

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

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

⁵In 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.

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

to be more aligned 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 Consequently, we consider that in the Colombian context elites and institutions associated to 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 Positive Political Discourse in Social Media

An informative discussion is central to politics, as individuals in this realm are constantly making arguments and engaging in debate. On social media platforms like Twitter, discussion plays a central role since individuals give opinions and arguments for the position they take.

Ideally, deliberation is based on respecting diversity of opinions and alternatives in order to arrive at an informed solution and as such, it requires openness: an active participation and a sense that all contributions can be considered equally. Historically, the ability to participate in political discussions was limited by social standing and technology; social media presented the possibility of overcoming these problems. Indeed, there was initially optimism that social media would enable people to communicate their opinions easily, unconstrained by geography and the power imbalance of the physical world Papacharissi (2009). The overall effect of social media (and the internet more generally) on democratic politics was hoped to be to revitalize the public sphere of debate Papacharissi (2004).

However, deliberation per se is not always desirable: a more pessimistic approach to deliberation reminds us that there are forms of discourse that are not deliberative and that could help people avoid conflict. In other words, when we evaluate deliberation in reality, we should remember that it is not the only way for people to settle their differences peacefully, and that it may not always work to the good. The consensus in online political communication is that most popular online forums (and platforms like Twitter) are not very deliberative Janssen and Kies (2005).⁸ In part because of

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

⁸One 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,

the multifaceted nature of reason giving, the platforms where debate takes place create challenges: Issues like anonymity ⁹ and increased polarization make an ideal deliberative process in online debates 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 (2017a) argues that the presence of incivility has both a *compositional* and a *direct* effect on online discourse: only people with a high tolerance for uncivil discourse can hope to engage in public discussions; discussants who join an online forum and observe that an uncivil discussion is taking place, are more likely to be uncivil themselves.

Scholars and practitioners of politics have advocated for more and better opportunities for citizens to deliberate about matters of politics (Mendelberg (2002) makes a review of up to the date attempts to increase the quality of political debates). Recently, there are increasing calls for more civility in political discourse. These developments have come hand in hand with a growing sense that an open debate will help to reduce polarization and incivility in other spheres. The argument is basically that an informed and heterogeneous discussion leads to a larger 'argument repertoire' and more political knowledge. More active participation is then positively related to more political knowledge Scheufele et al. (2006).

Deliberation, which, among other qualities, celebrates civility, reason giving, and communication across lines of political difference, is the most salient norm in political communication research (Davies and Gangadharan, 2009; Stromer-Galley, 2007; Wright and Street, 2007)). Rather than the ideals of deliberative democracy, James Fearon (1998) calls for reason to increase positive discourse and active participation in political discussions. Deliberation refers to a particular sort of discussion - one that involves the careful and serious weighting of reasons for and against some propositions. A discussion, on the contrary, needs not to be careful nor serious. Nonetheless, if it is reason-centered, informed discussion is expected to produce a variety of positive democratic outcomes Fishkin and Luskin (2005).

What is the point or value of discussing things before making political decisions? The theoretical literature on deliberative democracy highlights at least three reasons

2017; Muhlberger, 2005; Price, Cappella, and Nir, 2002)

⁹Studies have found that more anonymous platforms experience less sophisticated political discussions Omernick and Sood (2013).

or arguments for discussing a matter before reaching a decision on what to do. One is the quality of the decision, a second one is the enforcement of the decision taken, and finally, the revelation of private information. Instead of focusing on the normative aspect of increasing participation in political debate, we explore how some relevant actors can affect the deliberative process. Specifically, we look under what conditions the elite is able to activate civil political participation.

4 Elite Cues and Framing Effects in Social Media

Since Campbell (1960), scholars have shown how ideology shapes people’s response to information and how it forms political decisions. More recently, attention has been given to studying the conditions that moderate the effect of these ‘perceptual screens’ (Kernell, 2013).

On the one hand, a wealth of literature shows that elites mold how people perceive different events situations or messages by using different frames. The phenomenon, known as “elite-issue framing”, occurs when in the course of describing a issue, a speaker emphasizes a set of considerations that causes individuals to focus on those when processing the information and forming an opinion¹⁰

Extensive research in both the American and comparative contexts suggests that citizens rely on simple and reliable cues from elites in order to make policy judgments (Druckman, 2001; Druckman, Peterson, and Slothuus, 2013; Lupia, 1994, 2015; Lupia and McCubbins, 1998). Elites affect the public’s perceptions of their political ingroup and shape attitudes and behavior towards the outgroup. 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; O’keefe, 2002). 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 elite providing information *and* modifying the framing of that information, all in the realistic context of social media.

Social media is now a major platform of mass political discussion that allows us to manipulate the type and source of information that people receive about an issue. We

¹⁰see (Druckman, 2014), and (Druckman, Peterson, and Slothuus, 2013) for a review of this literature.

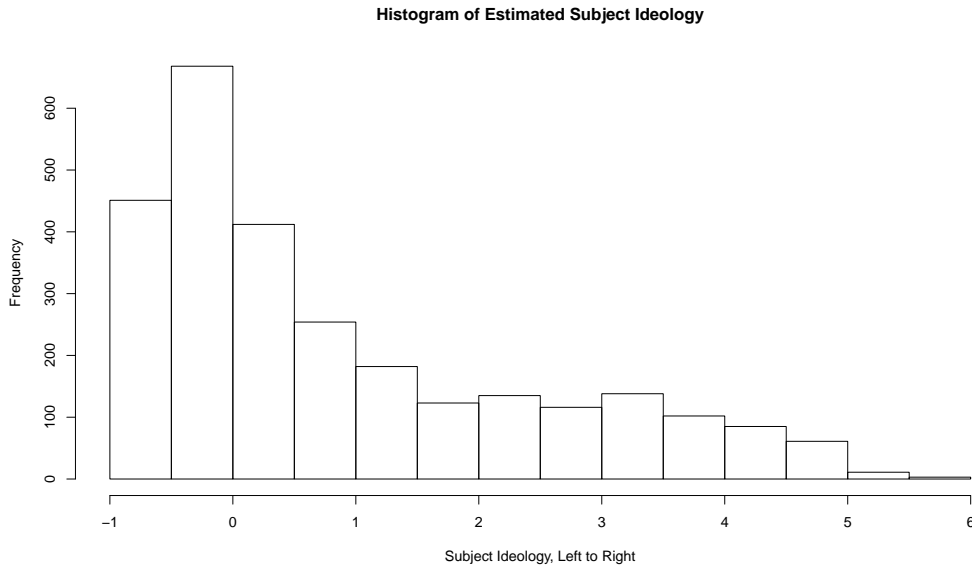
focus on Twitter for several reasons. Recent literature suggests that one of the ways citizens learn about party platforms is by directly communicating with politicians on Twitter (Munger et al., 2016).

Substantial literature in psychology and political science argue that people protect their world views through a series of complementary belief-preserving mechanisms. Examples of this perspective include the tendency to perceive evidence and arguments that support one’s views more persuasive than others; another example that we develop in this paper is the tendency to be more responsive to some interlocutors than to others, based on their ideological position. Both the content and the identity of the interlocutor may affect the effectiveness of a message. It is well known that political leaders attempt to identify themselves with their constituents in a effort to persuade them. Plutzer and Zipp (1996), for example, argue that women are more likely to vote for female candidates when they identify with women’s issues and were registered as independents. One questions raised by such findings is how identification with aspects of oneself affects behavior. A potential problem in this type of research is that many of the studies expose all subjects to the same information or to the same interlocutor. Without randomizing the content of the information itself and the source of information, we argue, it is impossible to properly test the predictions of this theory.

We focus here on three identities as sources of information: a priest, a soldier, 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, and their assumed ideological position, is based on the ideological distribution of our population of interest. As seen in Figure 2, the distribution of ideology of Twitter users is slightly skewed. The histogram shows that while liberals are clustered around one point of the ideology spectrum, there is a long tail of conservatives spreading out to extreme positions.

Accordingly, we divide the political spectrum in three segments: liberal, conservatives, and moderates. Building in the attitude polarization literature in American politics we test whether liberals (conservatives) will react positively to a liberal (conservative) message. In other words, 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

Figure 2: Ideology Distribution



Histogram of the ideological position of each of the subjects in our experiment. These estimates are derived from Barberá (2014)’s Bayesian Spatial Following Model.

the discussion is likelier to alienate rather than persuade.

Additionally, we hypothesize that the effect will vary according to the source of information for different segments of the users. That is to say, heterogenous effects are quite important in our analysis. We expect stronger reactions for non-moderate citizens. First, we expect strong opponents of the peace process (conservatives) to have a negative reaction to messages that come from a liberal source. When the source of information is a scientist, conservatives are less likely to have a positive reaction.

5 Experimental Design

We conducted a field experiment on Twitter during the 2016 Colombian Plebiscite. For this purpose, we collected all tweets related to the peace process from March through September of 2016. We identified accounts that have been active on this topic two months prior to the plebiscite. Using text analysis and machine learning, we construct a sentiment score for each account, which specifies if the account supports or opposes the process.¹¹ We then selected a random sample of 4,500 of these accounts. Using block

¹¹Details of this process can be found in Appendix 1.

Figure 3: Treatments–Liberal Scientist and Conservative General



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. Recall that all of the messages were in favor of the peace process, for the reasons discussed above. 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 for the liberal scientist and the conservative general:

In terms of content, 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”

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 dimensions 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.¹²

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

¹²The research described in this paper was approved by the IRB at NYU and la Universidad de Rosario.

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.¹³

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% .¹⁴ 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.

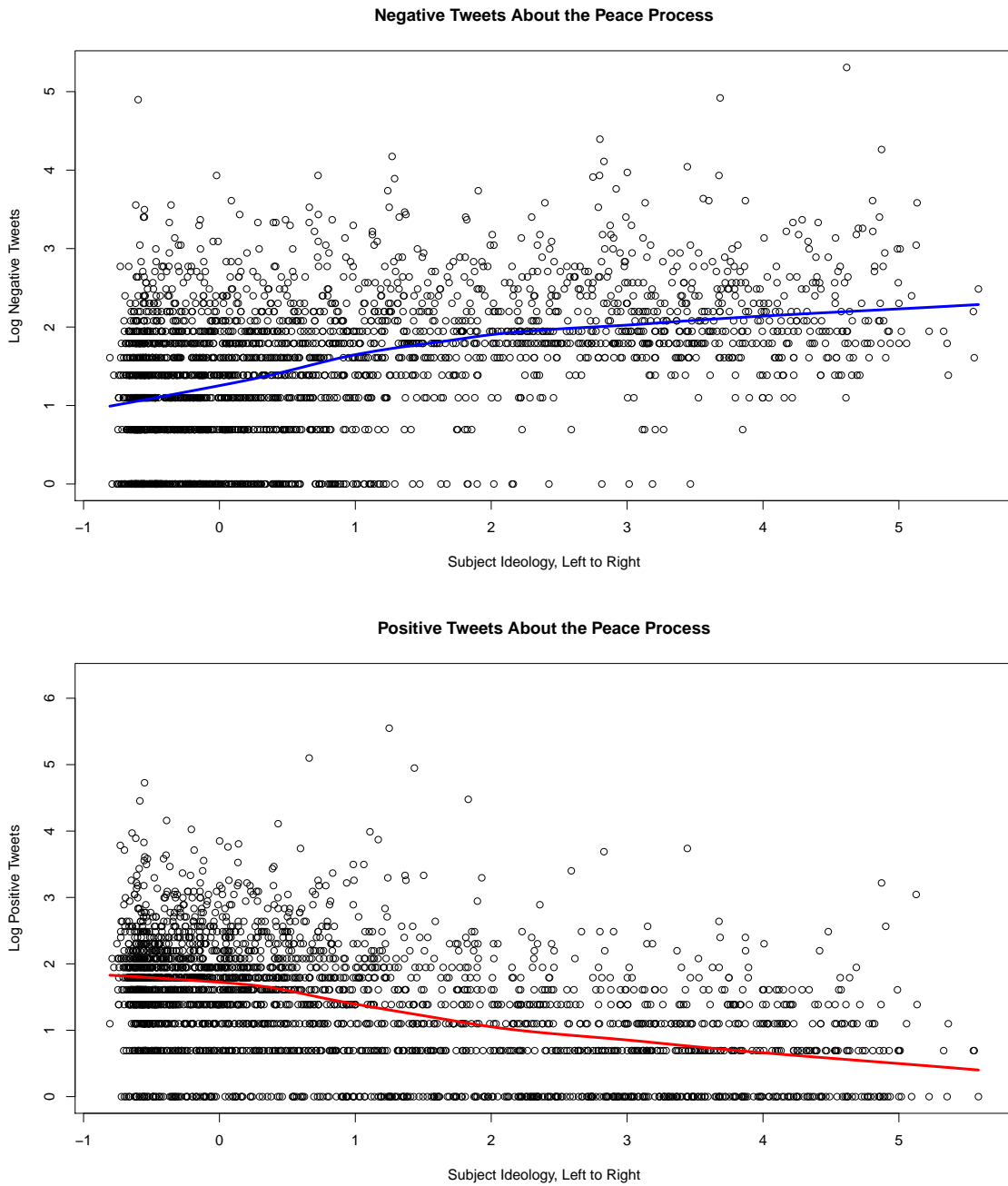
As a validity check of our application of these two machine learning classifiers, we plot the log (plus one) of the number of positive and negative tweets each subject sent about the peace process against their estimated ideology score, in Figure 4. The top panel plots the log number of negative tweets;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

¹³We 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 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", "#FeliSidad", "plebiscito no", "#Plebiscito", "diálogos de paz", "#SantosElTal23NoExiste", "Habana paz", "negociaciones Habana", "conversaciones Habana", "Gobierno paz FARC", "Gobierno Habana FARC", "#NoMarcho"

¹⁴Details about the implemtation of the SVM model can be found in Appendix 1.

Figure 4: Validating Tweet Sentiment and Estimated Subject Ideology



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 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 (“Conservatives”); and the value 2 for subjects above the 75th percentile (“Extreme 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 $k = \text{General, Priest, Scientist}$, 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 to negative replies.¹⁵ 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.

Hence, keeping constant the identity of the bot, these coefficients determine if liberal or conservative messages produce more reactions. In other words, through these models we are determining, for example, if the liberal priest generates more reactions than the conservative priest. Something similar for the general and the scientist.

Figure 5 plots the regression coefficients –and its 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 a message from the liberal version of that type of bot, while values below 0 indicate higher likelihood for the conservative version.

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 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. For this purpose, we split the sample of subjects in three, according to their ideology scores. Liberals are subjects with an ideology score below the 25th percentile of the distribution of such variable, conservatives are those above the 75th percentile, while the rest of subjects are classified as moderates.

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

¹⁵Manual coding for these replies was performed, to determine whether the subject responded positively or negatively to the bot.

Reactions to Liberal vs. Conservative Messages

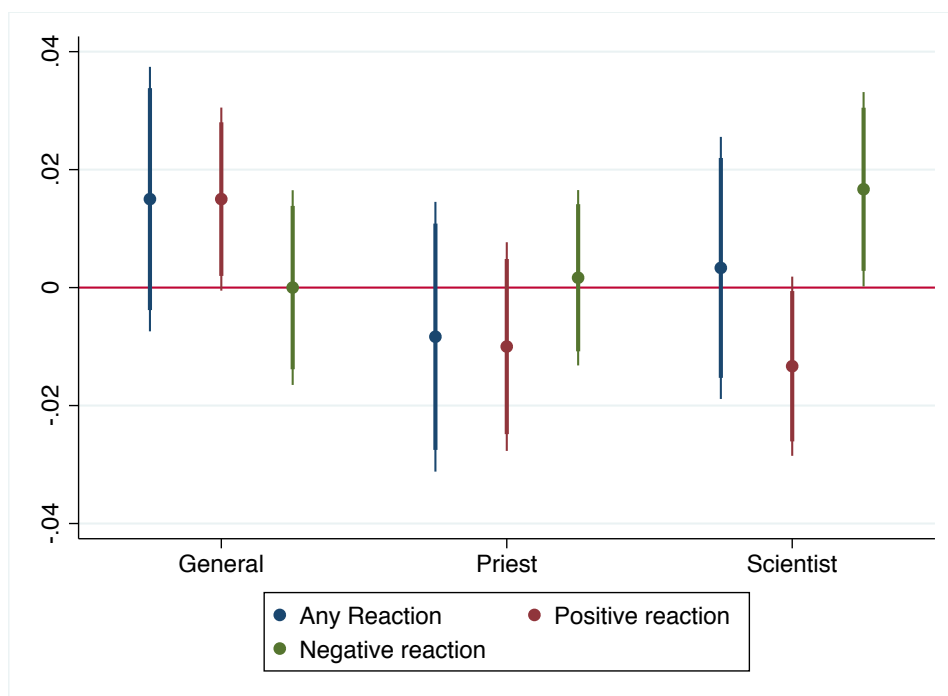


Figure 5: Each of the three outcomes are the result of a separate OLS regression. Values above 0 in the figure mean that outcome was more likely to be caused by a message from the liberal version of that type of bot, while values below 0 indicate higher likelihood for the conservative version.

Heterogeneous Effects: Liberal vs. Conservative Messages from the General

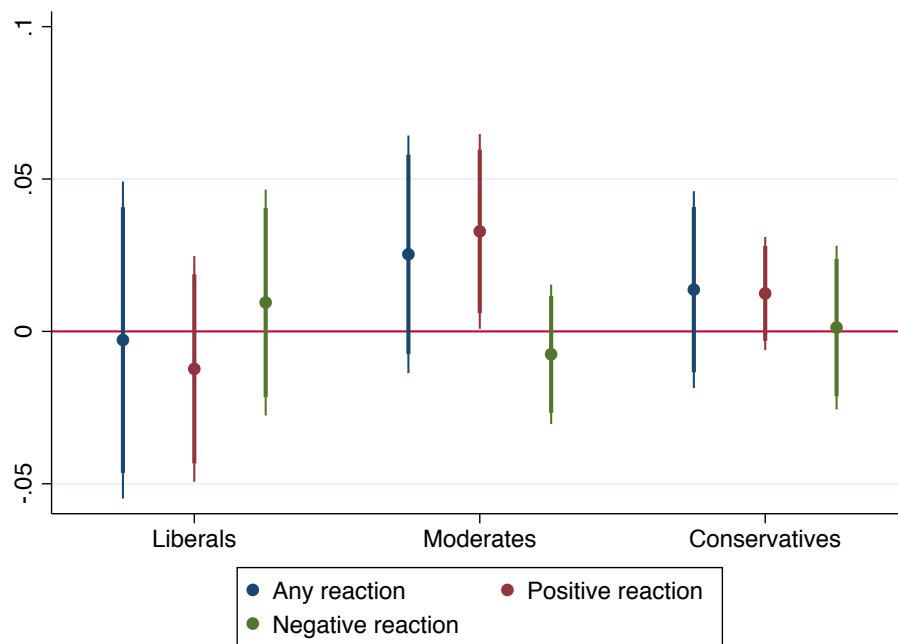


Figure 6: Each of the three outcomes are the result of a separate OLS regression. Values above 0 in the figure mean that outcome was more likely to be caused by a message from the liberal version of that type of bot for each subgroup of subjects, while values below 0 indicate higher likelihood for the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

Heterogeneous Effects: Liberal vs. Conservative Messages from the Priest

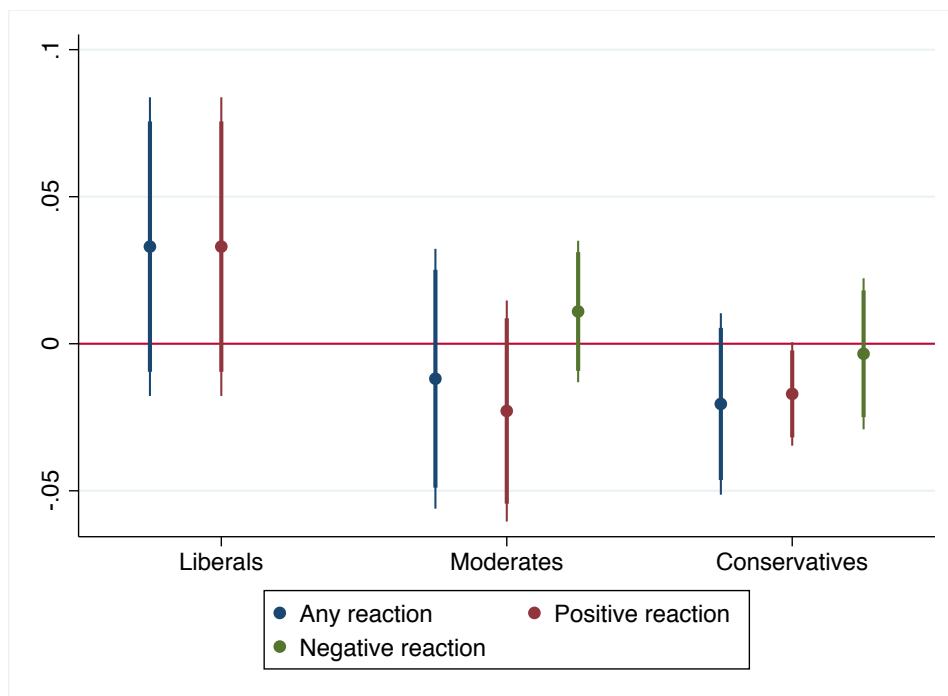


Figure 7: Each of the three outcomes are the result of a separate OLS regression. Values above 0 in Figure mean that outcome was more likely to be caused by a message from the liberal version of that type of bot for each subgroup of subjects, while values below 0 indicate higher likelihood for the conservative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

Heterogeneous Effects: Liberal vs. Conservative Messages from the Scientist

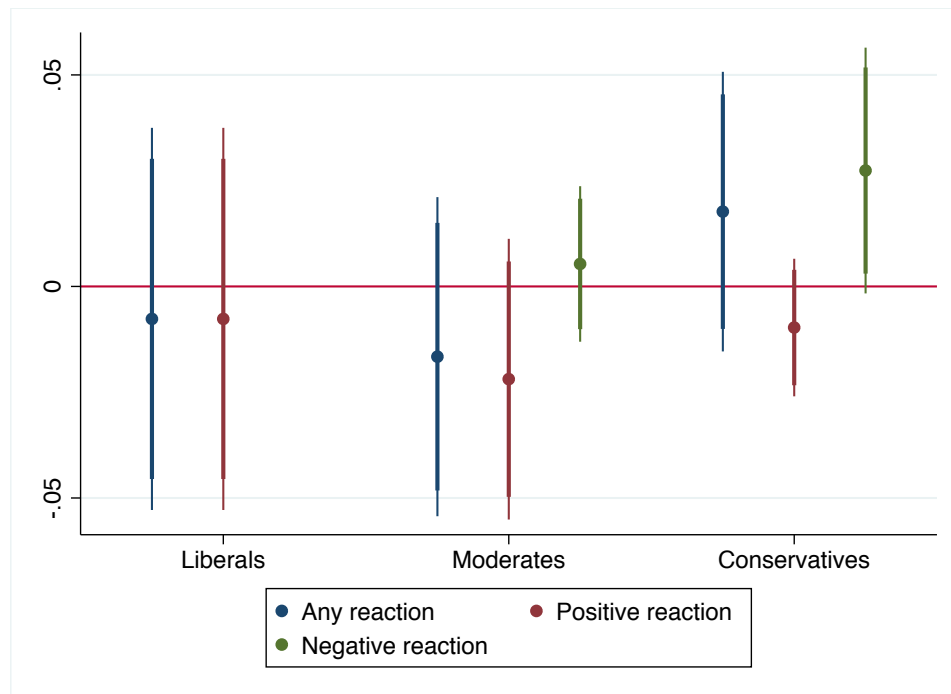


Figure 8: Each of the three outcomes are the result of a separate OLS regression. Values above 0 in the figure mean that outcome was more likely to be caused by a message from the liberal version of that type of bot for each subgroup of subjects, while values below 0 indicate higher likelihood for the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

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 by this type of bot.

7.2 Post-Treatment 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 (2017a), we used negative binomial regression. The negative binomial specification is estimated using the following model:

$$\begin{aligned} \ln(Tweets_{post}) = & x_{int} + \beta_1 Tweets_{pre} + \beta_2 T_{priest} + \beta_3 T_{soldier} + \beta_4 T_{scientist} + \beta_5 Ideology + \beta_6 (T_{priest} \times Ideology) \\ & + \beta_7 (T_{soldier} \times Ideology) + \beta_8 (T_{scientist} \times Ideology) \end{aligned}$$

To interpret the relevant treatment effects implied by the coefficients estimated by this model, 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:

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

The experimental results on the full sample in the first week (excluding day 1, in which there were many direct reactions to the tweets) after treatment are displayed in Figure ??; 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.

$IRR_{priest \times Ideology_1} = 1.3$ can be seen in the lime green line in the middle of the plot. This Incidence Ratio implies that the average subject with Ideology Score 1 who received the liberal priest treatment tweeted 130% as many positive tweets about the peace process as the average subject with Ideology Score 1 in the control condition.¹⁶

¹⁶Note 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, 2-7 Days Post-Treatment ($N=3,516$)

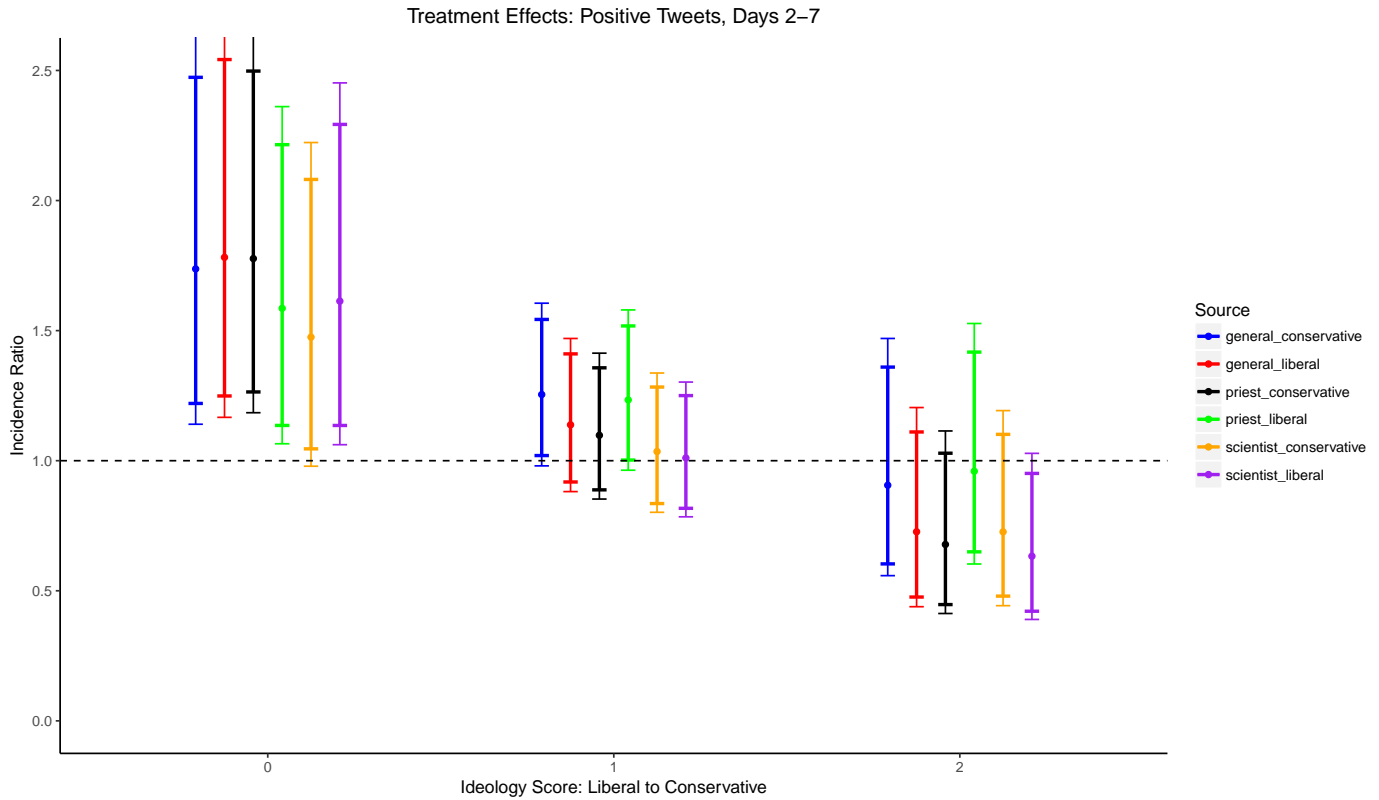


Figure 9: The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model in the first week after treatment, excluding day 1. For example, the Incidence Ratio of 1.3 associated with the liberal priest treatment on subjects with Ideology Score 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 thick bars represent 90% confidence intervals and the thin lines represent 95% confidence intervals.

The confidence intervals in Figure ?? 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 respondents' rate of sending positive tweets; the largest heterogeneity was in their effects on respondents with different ideologies. Liberal respondents were encouraged to send more positive tweets, as were some moderates. Two of the treatment conditions (the conservative general and the liberal priest) caused a significant increase in the rate of sending positive tweets among moderates, and the point estimate of the other 4 conditions was also positive.

These two conditions were also most effective on the conservative sample: they had an effect estimated at 0, while the other 4 conditions had a *negative* effect on the rate of sending positive tweets. In one condition, the liberal scientist, this effect was statistically significant.

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. Figure ?? plots those results.

Encouragingly, across all 18 treatment conditions interacted with subject ideology, only one showed a statistically significant increase in the rate of sending negative tweets about the peace process—precisely the number that we'd expect to see by chance. The point estimates tend to be positive, even for the liberal sample. This is somewhat surprising

8 Conclusions

We performed a randomized experiment on Twitter users who we identified as interested in the peace process in Colombia. We tested several hypotheses about elite cues and attitude polarization. To do so, we sent public messages to Twitter users encouraging

Negative Tweets About the Peace Process, 2 to 7 Days Post-Treatment ($N=3,516$)

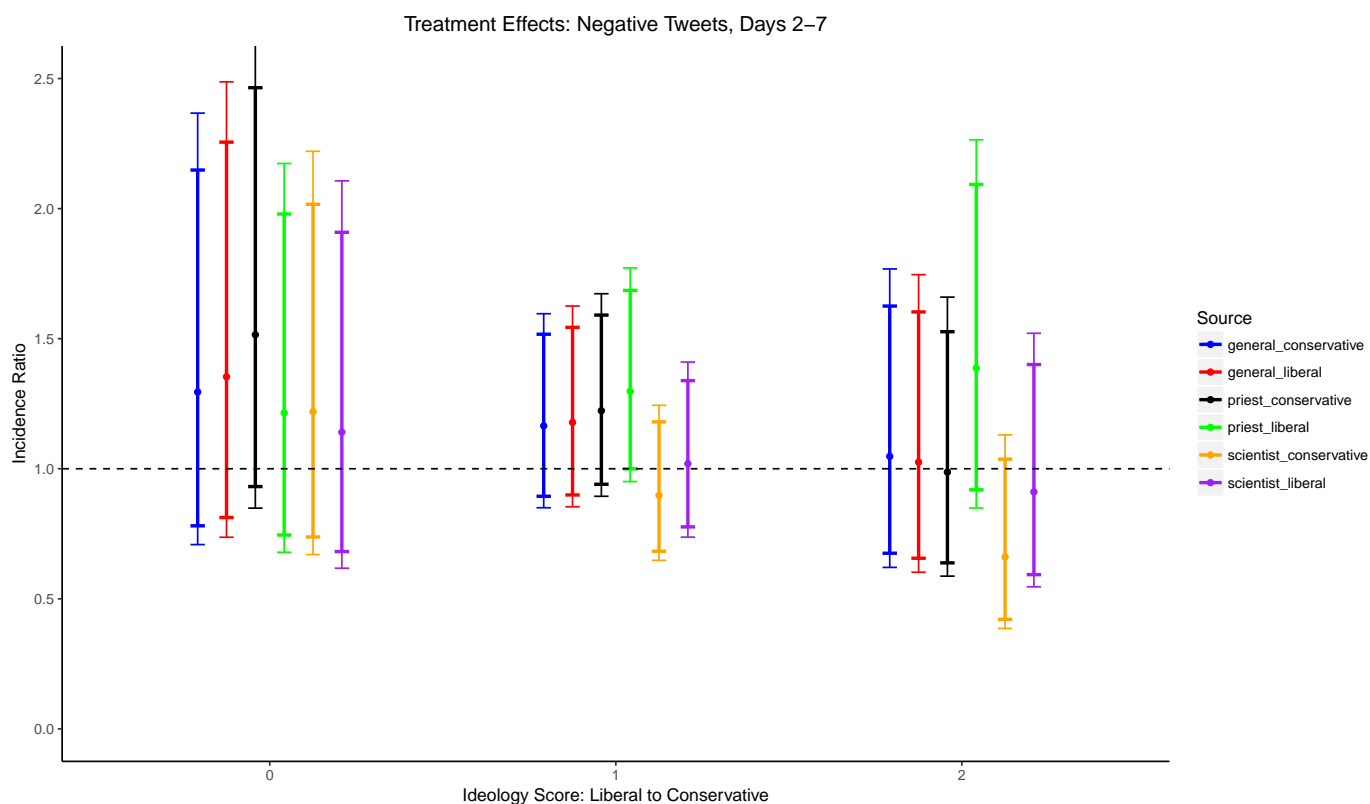


Figure 10: The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model in the time period from 2 to 7 days after treatment. For example, the Incidence Ratio associated with the liberal priest treatment on subjects with Ideology Score 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.

people 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.

Our results include two behavioral outcomes of interest. First, we find 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. If we analyze the impact of these messages conditional on the ideology of the user, we see that the positive effects of the liberal message sent by the general are driven by moderate subjects whereas conservative users disliked the liberal sent by the scientist and reacted positively to the conservative counter-part.

The main outcome of interest is in the change in behavior. Here again we find heterogeneous results based on the identity of the subject and bot. In general, liberal subjects were more likely to send more positive tweets, especially in response to the liberal priest. On the other hand, conservatives were less likely to change the sentiment of their tweets, except to send fewer positive tweets after interacting with the conservative scientist.

References

- Aarøe, Lene. 2011. "Investigating frame strength: The case of episodic and thematic frames." *Political Communication* 28 (2): 207–226.
- Barberá, Pablo. 2014. "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data." *Political Analysis* 23 (1): 76–91.
- Barberá, Pablo. 2015. "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data." *Political Analysis* 23 (1): 76–91.
- Campbell, Angus. 1960. *The american voter*. University of Chicago Press.
- Chaudoin, S., Shapiro J., and D. Tingley. 2017. "Revolutionizing Teaching and Research with a Structured Debate Platform." *Working Paper* .
- CHCV. 2015. "Comisin Histrica del Conflicto y sus Vctimas: Contribucin al entendimiento del conflicto armado en Colombia."
- Chong, Dennis, and James N Druckman. 2007. "Framing theory." *Annu. Rev. Polit. Sci.* 10: 103–126.
- Davies, Todd, and Seeta Peña Gangadharan. 2009. "Online deliberation: Design, research, and practice."
- 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. 2001. "The implications of framing effects for citizen competence." *Political behavior* 23 (3): 225–256.
- Druckman, James N. 2014. "Pathologies of studying public opinion, political communication, and democratic responsiveness." *Political Communication* 31 (3): 467–492.
- Druckman, James N, Erik Peterson, and Rune Slothuus. 2013. "How elite partisan polarization affects public opinion formation." *American Political Science Review* 107 (01): 57–79.
- 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.” *Contribucin al entendimiento del conflicto armado en Colombia* .
- Fishkin, James S, and Robert C Luskin. 2005. “Experimenting with a democratic ideal: Deliberative polling and public opinion.” *Acta Politica* 40 (3): 284–298.
- 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.”.
- Janssen, Davy, and Raphaël Kies. 2005. “Online forums and deliberative democracy.” *Acta política* 40 (3): 317–335.
- Kernell, Georgia. 2013. *The Scope of the Partisan ?Perceptual Screen?* PhD thesis Northwestern University.
- Lupia, Arthur. 1994. “Shortcuts versus encyclopedias: Information and voting behavior in California insurance reform elections.” *American Political Science Review* pp. 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.
- Mendelberg, Tali. 2002. “The deliberative citizen: Theory and evidence.”.
- Muhlberger, Peter. 2005. “The Virtual Agora Project: A research design for studying democratic deliberation.” *Journal of Public Deliberation* 1 (1).
- Munger, Kevin. 2017a. “Dont@ Me: Experimentally Reducing Partisan Incivility on Twitter.”.
- Munger, Kevin. 2017b. “Tweetment effects on the tweeted: Experimentally reducing racist harassment.” *Political Behavior* 39 (3): 629–649.

- Munger, Kevin, Jonathan Ronen, Jonathan Nagler, Pat Egan, and Joshua Tucker. 2016. The Impact of Social Media Use on Voter Knowledge and Behavior in the 2015 UK Election: Evidence from a Panel Survey. In *Unpublished Manuscript*.
- O’keefe, Daniel J. 2002. *Persuasion: Theory and research*. Vol. 2 Sage.
- Omernick, Eli, and Sara Owsley Sood. 2013. The impact of anonymity in online communities. In *Social Computing (SocialCom), 2013 International Conference on*. IEEE pp. 526–535.
- 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.
- Plutzer, Eric, and John F Zipp. 1996. “Identity politics, partisanship, and voting for women candidates.” *Public Opinion Quarterly* 60 (1): 30–57.
- Price, Vincent, Joseph N Cappella, and Lilach Nir. 2002. “Does disagreement contribute to more deliberative opinion?” *Political communication* 19 (1): 95–112.
- Scheufele, Dietram A, Bruce W Hardy, Dominique Brossard, Israel S Waismel-Manor, and Erik Nisbet. 2006. “Democracy based on difference: Examining the links between structural heterogeneity, heterogeneity of discussion networks, and democratic citizenship.” *Journal of Communication* 56 (4): 728–753.
- Stromer-Galley, Jennifer. 2007. “Measuring deliberation’s content: A coding scheme.” *Journal of public deliberation* 3 (1).
- Wills, Maria E. 2015. “Los tres nudos de la guerra colombiana.” *Contribucin al entendimiento del conflicto armado en Colombia* .
- Wright, Scott, and John Street. 2007. “Democracy, deliberation and design: the case of online discussion forums.” *New media & society* 9 (5): 849–869.

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).

$$\hat{c} = \operatorname{argmax} P(c|d) \tag{1}$$

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).

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

$$\operatorname{argmax} P(c|d) = \frac{P(d|c)P(c)}{P(d)} \quad (3)$$

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:

$$\operatorname{argmax} P(c|d) = P(d|c)P(c) \quad (4)$$

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.

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

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.