Understanding the political representativeness of Twitter users

Pablo Barberá [*] New York University

Gonzalo Rivero [†] YouGov

October 15, 2014

Forthcoming in *Social Science Computer Review*

Abstract

In this article we analyze the structure and content of the political conversations that took place through the micro-blogging platform Twitter in the context of the 2011 Spanish legislative elections and the 2012 US presidential elections. Using a unique database of nearly 70 million tweets collected during both election campaigns, we find that Twitter replicates most of the existing inequalities in public political exchanges. Twitter users who write about politics tend to be male, to live in urban areas, and to have extreme ideological preferences. Our results have important implications for future research on the relationship between social media and politics, since they highlight the need to correct for potential biases derived from these sources of inequality.

[*]Pablo Barberá (www.pablobarbera.com) is a Ph.D. Candidate in the Department of Politics at New York University. He can be reached at pablo.barbera@nyu.edu.

[†]Gonzalo Rivero (griverorz.net) is a Senior Analyst at YouGov. He can be reached at gonzalo.rivero@yougov.com.

1 Introduction. Twitter's potential as a source of information about public opinion and electoral preferences

The micro-blogging service Twitter has become one of the most important social networking sites, with 200 million montly active users worldwide[1], including 18% of all online Americans[2] and 15% of online Spaniards. While most of the messages that are sent through this platform are personal updates, its use for political purposes has increased in the past few years. Virtually all candidates and elected officials have a presence on Twitter; and many users rely on Twitter to stay informed about political events. The content and structure of the political discussion that takes place on this platform, easily accessible through their API, represents a unique opportunity for researchers interested in the study of elections and public opinion.

The increase in the use of social media has led many social scientists to examine whether specific patterns in the stream of tweets might be able to predict real-world outcomes. Asur and Huberman (2010), for example, show how a simple model measuring chatter from Twitter about movies predicts box-office revenues, outperforming market-based predictors[3]. Applying a similar method, Lampos et al. (2010) are able to accurately track the prevalence of Influenza-like illnesses in several regions of the United Kingdom. Paul and Dredze (2011) extend this analysis into a broader range of illnesses in the United States, opening a whole new agenda in the field of public health research. In two highly publicized articles, Golder and Macy (2011) and Dodds et al. (2011) study the temporal patterns of happiness of millions of people in real time based on their tweets. An inno-

[1]Source: Twitter's account on Twitter, December 18 2012.

[2]Source: The Pew Research Center's Internet & American Life Project, August 2013.

[3]However, a recent study by Wong et al. (2012) contradicts this result and argues that opinions expressed via Twitter tend to be more positive than those that can be found on other website, which reduces their predictive potential.

vative study conducted by Hannak et al. (2012) builds up on this research to show how weather affects aggregated sentiment. Measurements of collective mood states derived from Twitter feeds had already been found to be correlated with stock market indexes by Bollen et al. (2011). Finally, exploiting the geographic information that Twitter users provide has allowed researchers to estimate the epicenter of earthquakes in Japan (Sakaki et al., 2010).

Given the accuracy of these predictions, and the consolidation of Twitter as a source of political information, a battlefield for campaigning, and a public forum of political expression, some researchers have wondered whether "tweets" validly mirror offline public opinion. "Can we analyze publicly available data to infer population attitudes in the same manner that public opinion pollsters query a population?" (O'Connor et al., 2010, p.122). Were this approach to be successful, its advantages would be obvious: Twitter provides (relatively) easy and free access to millions of public messages in real-time and from most countries around the world. Inferring public opinion from social media messages is challenging, but also potentially very rewarding, given the wealth of this information.

The first studies of this kind, in the context of the German legislative elections of 2009 (Tumasjan et al., 2010) and the first two years of the Obama presidency in the United States (Cummings et al., 2010; O'Connor et al., 2010) gave reasons to be optimistic. Tumasjan et al. (2010) found that "the mere number of messages [mentioning each German political party] reflect[ed] the election result and even [came] close to traditional electoral polls". O'Connor et al. (2010), on the other hand, showed that "a relatively simple sentiment detector based on Twitter data replicate[d] presidential job approval polls. The results highlight the potential of text streams as a substitute for traditional polling". Similar studies conducted in Singapore (Choy et al., 2011; Skoric et al., 2012), United Kingdom (Lampos, 2012), Denmark (Sang and Bos, 2012), Spain (Congosto et al., 2011) and recent

elections in the United States (DiGrazia et al., 2013) have also found that, during the elections, "the Twittersphere represents a rich source of data for gauging public opinion and that the frequency of tweets mentioning names of political parties, political candidates and contested constituencies could be used to make predictions about the share of votes at the national level" (Skoric et al., 2012, p.2583). With these results, Cummings et al. (2010) even wondered "who needs polls?".

The response to this set of papers arrived in two recent research articles by Metaxas et al. (2011) and Gayo-Avello (2012). These authors warn against "turning social media into another 'Literary Digest' poll" and claim that the "predictive power of Twitter regarding elections has been greatly exaggerated". They illustrate their concerns with analyses of several Senate races in the 2008 and 2010 US Congressional elections, and find that electoral predictions applying similar methods as those used by the previous authors do not perform better than chance. These authors also criticize the previous results in this literature. For example, they point out that if Tumasjan et al. (2010) had not restricted their analysis to parties with parliamentary representation, the Pirate Party would have won the 2009 German elections – it was the party with the highest number of mentions in Twitter (Jungherr et al., 2011). In their view, an accurate prediction can only come through "correctly identifying likely voters and getting an un-biased representative sample of them". Self-selection biases, overrepresentation of younger, more educated citizens on Twitter, and the simplistic assumptions of the existing sentiment analysis techniques are the three most important methodological challenges to overcome.

On a similar note, it is important to distinguish prediction from causality: the debate about whether Twitter metrics are or not correlated with public opinion metrics is orthogonal to the discussion about the effectiveness of online campaign strategies in vote choice. In fact, it is usually the case that the candidate that is mentioned the most on Twitter is

the one that receives more criticism. For this reason, any method developed to measure sentiment should be capable of distinguishing praise from criticism, a positive message from a sarcastic message, and a relevant tweet from just "spam."

To sum up, this set of research papers evidences some of the challenges that the use of Twitter data presents. The average internet user is younger, more interested in politics, and comes from a higher socioeconomic background than the average citizen, which raises concerns about external validity (Mislove et al., 2011; Gong, 2011). Furthermore, the voice of political minorities tend to be under-represented in the public debate on Twitter, and differences in party strategies regarding their presence in social media can also bias any measure of public opinion that relies on the number and content of tweets. It is therefore necessary to obtain more background information about each individual user, so that it is possible to stratify them and weight public opinion estimates. And even if most studies implicitly assume that Twitter is not a representative sample of the entire population, none of them examines to what extent participants on political discussions on Twitter are self-selected.

That is precisely the purpose of our study. Focusing on the 2011 Spanish legislative election and then 2012 US presidential election, we characterize the content and structure of the political conversations about each election that took place on Twitter. These two elections took place in countries with diverse sociopolitical characteristics, and with different adoption rates of Twitter at the time of our analysis, which allows us to examine to what extent the patterns that emerge in our analysis could be generalized to other countries and elections.

We use a unique dataset that includes all tweets sent in the 70 days prior to the election with references to the two major candidates that were running, which represents a total of 3 million tweets in Spain, and 62 million tweets in the US. To analyze inequality in

political discussions, we focus on a random sample of 12,000 and 50,000 users respectively that participated actively in the political discussion about each election, for which we estimated sociodemographic and political variables of interest.

Our analysis yields four main conclusions. First, we find that political debate on Twitter in both countries has a male name: a majority of users who tweet about politics are men. Second, the geographic distribution of Twitter users approximates that of population, with a slight bias towards urban areas. Third, we find that political discussion on Twitter is highly polarized: users with clear ideological leaning are much more active and generate a majority of the content. Finally, we find significant differences in the patterns of discussion over our period of analysis: in both countries, campaign events increase or diminish inequality in participation on Twitter. These results have important implications for the use of Twitter data as a source of information about public opinion.

2 Data and method. Twitter in the 2012 legislative election in Spain and the 2013 US presidential election.

2.1 Dataset

To analyze the structure of content of political discussion on Twitter about our two study cases, we captured all tweets mentioning the two most important candidates in each election, Rajoy and Rubalcaba, and Romney and Obama, over the 70 days prior to the date of the election (November 20th, 2011 and November 6th, 2012).[4] During this period, we systematically captured this set of tweets, at regular intervals.[5] In particular, in the first case

[4] Our period of study spans the 70 days between September 10, 2011 and November 20, 2011 in Spain, and the 70 days between August 28, 2012 and November 7, 2012 in the US.

[5] In the Spanish case, tweets were captured every each hour before the campaign, every half hour during the campaign, and in intervals of one minute during the debate (November 7), using Twitter's REST API

we included all tweets that mention at least one of the following four keywords: "rubalcaba", "rajoy", "conrubalcaba", "marianorajoy".[6] This represents a total of 2.8 million tweets sent by 375,000 unique users. For the the US presidential election, we included all tweets that mentioned at least one of the following four keywords: "obama", "romney", "barackobama", "mittromney". In this case, our dataset contains a total of 62 million tweets and 8.8 million unique users.

As it has been shown before, production of tweets by users is profoundly unequal: not all participants in the global discussion contribute to the same extent, and the differences are similar across our two case studies. While 47% of Spanish users and 48% of US users published only one tweet mentioning one of the candidates over our 70 days of study, others ?94 in Spain, 4,873 in the US? tweeted more than 1,000 times.

Similarly, while the 1% most active users published 34.8% of all tweets in Spain (and 44.6% of all tweets in the US), and the top 10% generated 70.9% of all tweets (70.6% in the US), the remaining 90% sent only around 29% of all tweets. This distribution corresponds to a Gini coefficient of 0.764 in Spain and 0.755 in the US, which denote a high concentration of the publication of tweets in a very limited number of users.

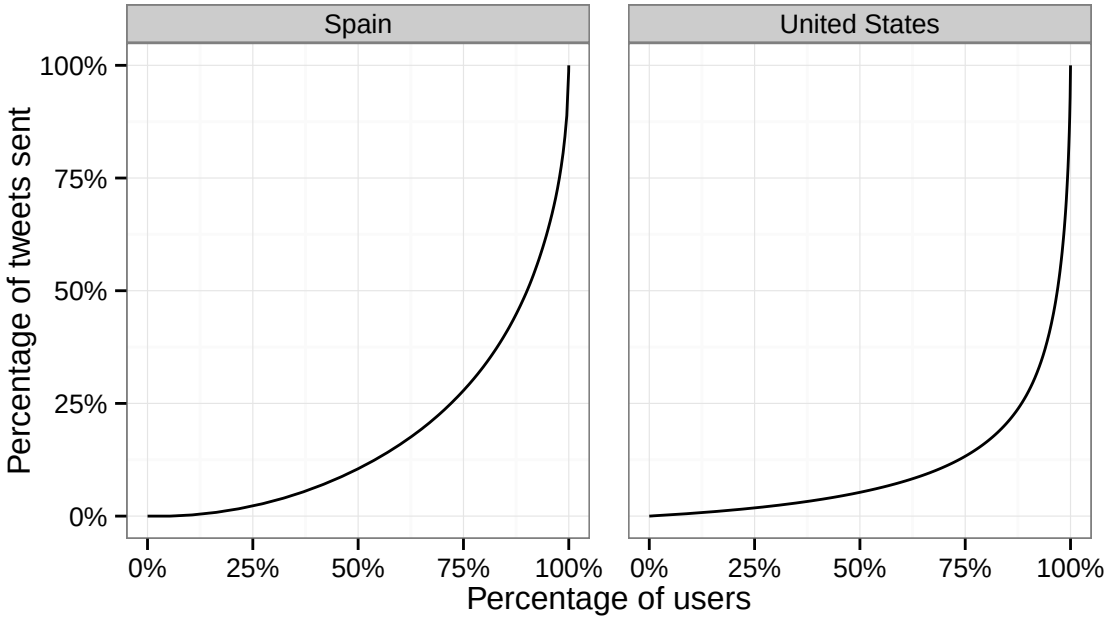
Given the magnitude of our dataset, and with the purpose of providing a closer look at political discussion about these two elections, most of our analysis in this article focuses on a random sample of 12,000 active users[7] in Spain and 50,000 active users in the US.

and the twitteR R package (Gentry, 2013). In the US case, we relied on Twitter's Streaming API and the streamR R package (Barber, 2013). Although the use of different methods of data collection can introduce bias in our analysis (Gonzlez-Bailn, 2014), in both cases the number of potentially missed tweets is small in comparison with the total size of our dataset, which reduces the probability that any difference that we observe is due to the collection method.

[6]The last term includes both tweets that mention the name of the candidate and their screen name on Twitter.

[7]In both cases, we have considered as "active users" those that sent at least five tweets during our period of analysis and follow at least three politicians, journalists or political news outlets. In the US, we further restricted our sample to users with at least 25 followers and located inside the United States, with the purpose of avoiding the inclusion of spam or fake users and also individuals tweeting from other countries.

Figure 1: Lorenz Curve. Inequality in the number of tweets



At the end of our period of analysis, we captured the list of friends and followers of each of these users, as well as their description and location, as provided in their personal profile on Twitter. As we show below, this information will allows us to characterize their activity on this online networking site. Figure 1 shows how, even within these two groups of active users, there is a high degree of inequality in participation: for example, we see that the bottom 50% of least active users sent only around 10 to 15% of the total of tweets.

2.2 Dependent variables

In our analysis of the structure and content of the political discussion on Twitter, we will focus on three metrics of interest for theoretical or methodological reasons: the total volume of tweets, the volume of retweets, and the volume of @-replies. Each of these indicators is relevant for the reasons we now discuss.

2.2.1 Total volume of tweets

Most studies that aim at predicting election outcomes using Twitter data rely on this metric, by quantifying the number of tweets that mention a party or candidate. The problem with this approach is that, as we show in the following selection, some users publish more tweets than others, and these differences are systematically correlated with key individual-level characteristics.

The total volume of tweets is also relevant because it affects the probability that a campaign message becomes "trending topic". Even if other factors are also considered in the algorithm that Twitter uses to select the top 10 most important "topics" of the moment, it is obvious that an important strategy that candidates can use to increase their visibility is to generate "noise" on social media. Again, it is important to understand to what extent changes in the volume of tweets over an electoral campaign are due to a small group of very prolific users, or a generalized increase in political activism of ordinary citizens.

2.2.2 Volume of retweets and @replies

These two indicators are relevant because they allow us to estimate to what extent individuals engage in political conversations interactively (Wu et al., 2011). On one hand, a high volume of @-replies[8] between users about a candidate or party can be considered a sign of the horizontality of conversations about that political actor, therefore indicating a high degree of interactivity.

The number of retweets that users make of tweets with political content, on the other hand, is relevant for two additional reasons. First, this type of tweets help keep "alive" messages from other users (see for example the study by Barash and Kelly, 2012, about

[8]We consider as "@-replies" all tweets that begin with a mention to another user. This type of tweets are used as a tool to send directed messages publicly.

Twitter use in Russia). When a tweet sent by a party or candidate is retweeted, that contributes to their slogans and campaign messages to be generated larger cascades of information, since the rapidly decreasing visibility of tweets is the main obstacle for the propagation of messages (Oken Hodas and Lerman, 2012). Furthermore, authors such as Cha et al. (2010) have emphasized that the number of retweets is a better predictor of online influence than the number of followers.

2.3 Independent variables

The indicators we presented in the previous section vary at the individual and aggregate levels as a function of a long series of factors, many of them difficult to measure empirically. In our analysis, this variation becomes particularly relevant when it is systematically associated with other variables that can affect citizens' electoral behavior. When that is indeed the case, all electoral predictions that do not control in some way (either by weighting or calibrating based on the characteristics of the sample of individuals included in the analysis) will be systematically biased. In this article, we will focus on four of such variables: gender, geographic location, ideology, and the moment of the electoral campaign.

2.3.1 Gender

We have inferred the gender of each user in our random sample of 62,000 individuals based on the name they report in the "full name" field using a Nave Bayes classifier (Bird et al., 2009). Our training dataset was a list of names labeled by gender in each country. In Spain, our source is the National Instituce of Statistics; in the US, we rely on a large dataset of anonymized names available in the RandomNames R library (Betebenner, 2012). In

both cases, the accuracy of our classifier is higher than 70%.[9]

This variable is relevant because gender has a significant effect on political behavior, even after controlling for other confounding factors (see for example Chaney et al., 1998; Kellstedt et al., 2010; Martinez and Calvo, 2010). Similarly, in the specific context of Twitter, different comparative studies have found that women are underrepresented (Mislove et al., 2011; Parmelee and Bichard, 2011; Bode et al., 2011). As we will show with greater detail in the following section, this appears to be the case also in Spain and the US: we estimate that only between 35 and 40% of users who tweet about politics are women.

2.3.2 Geographic location

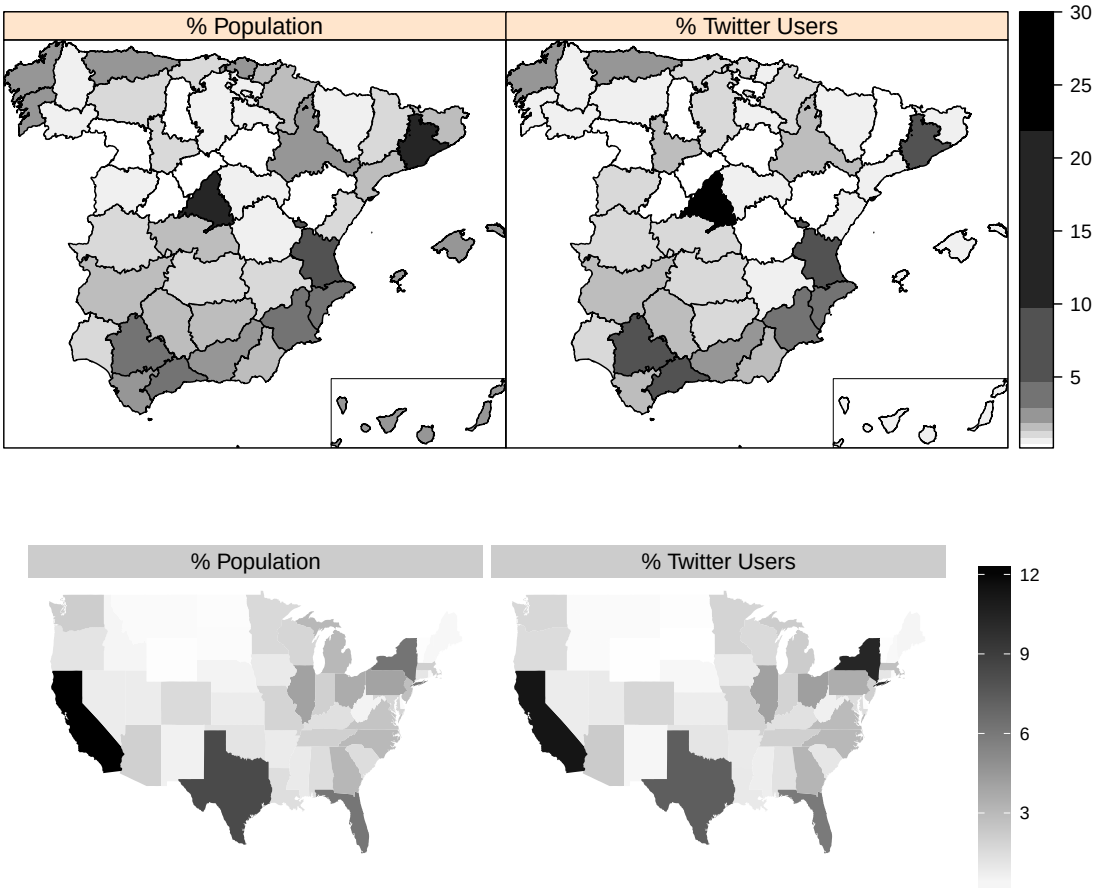
In our analysis we also examine the geographic location of Twitter users. Several studies suggest that Twitter users are located predominantly in urban areas (see for example Mislove et al., 2011). However, there is some controversy about this finding, since several authors (Hale et al., 2012; Hecht et al., 2011) argue that most Twitter users cannot be located accurately.

Our results also point in this direction. Using the Yahoo Maps API, we tried to locate each user in our random sample of 62,000 individuals using the "location" field in Twitter users' profiles. In Spain, we found that 39% of them do not provide enough information to determine the province in which they live. This proportion increases to 47% if we try to place users in cities. In the United States we find similar results: it is possible to identify the state from which users are tweeting in 84% of the cases, and the county for 67% of them.

With this caveat, our analysis of the geographic distribution by province or state of

[9] In comparison, Al Zamal et al. (2012) achieved a 80.2% accuracy using information about each user's most recent tweets, their network of followers, and their propensity to retweet particular types of content.

Figure 2: Geographic distribution of Twitter users in Spain and the United



the 7,307 Spanish Twitter users and 42,060 US Twitter users that we were able to geolocate shows that this sample is not systematically biased with respect to the geographic distribution of the whole population.[10] As we show in Figures 2 and 3, the proportions of Twitter users and citizens living in each province/state are very similar (Pearson's correlation coefficient between the two variables is 0.866 with N=52 in Spain, and 0.959 with N=50 in the US), with the single exception of Madrid, with 30% of Twitter users but only 13.7% of the total population.

[10]The source of population data is the local census in Spain (National Institute of Statistics) and the population estimates of the national census in the US.

2.3.3 Ideology/Party identification

The political science literature that studies electoral behavior has shown the importance of ideology and party identification in anchoring the vote in general and in the two cases of study in particular (Campbell et al., 1960; Torcal and Medina, 2002).[11] An analysis of the representativeness of Twitter users thus requires an understanding of the extent to which voters that identify with different parties or ideological orientations are present on Twitter, and whether they participate at similar levels.

To examine this issues, we implement a Bayesian item-response model which allows us to infer the orientation and intensity of ideological identification of each user as a function of the political orientation that is predominant in their network of "friends" (those that they decided to follow on Twitter). Hence, in the model we assume that ideology is a latent variable that can be inferred from the distribution of political actors that each user is following, and it is based on the assumption that users prefer to follow politicians whose ideological viewpoint is similar to theirs. One of the main advantages of the statistical model producing the results is that it naturally incorporates the possibility that some accounts are less discriminative than others for learning the ideology of a given user by reason of the high profile or visibility of the politicians behind them. A full description of the method can be found in Barberá (2014).

2.3.4 Timing of the electoral campaign

Finally, we also analyze how different metrics evolve throughout the electoral campaign. Timing is important for two reasons. First, it has been shown that voters with low in-

[11] We use ideology and party identification indistinctively throughout the rest of the text. Although they are conceptually different, in our analysis they both stand for whichever political dimension explains the observed political homophilia in the network of Twitter users. In this interpretation, the underlying concept captured by the ideal point estimation is expected to overlap with the main dimension of political conflict in each case of study (ideology in Spain and party identification in the US).

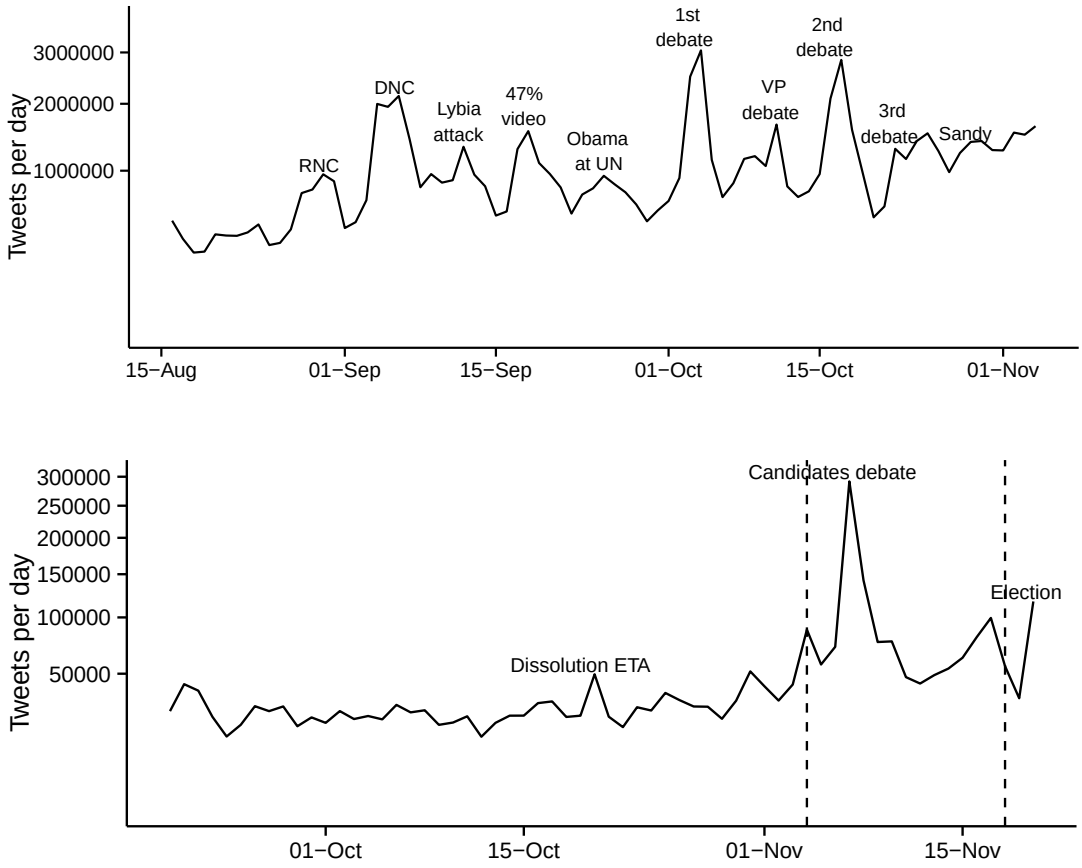
terests in politics decide who to vote for during the electoral campaign. In Spain, for example, the post-election survey conducted by the Center of Sociological Research in 2011 showed that 28.2% of voters made that decision in the last two weeks before the election. Second, in both elections there is an important asymmetry in the online strategies of the two main parties. While Rubalcaba opened his Twitter account in August of 2011 and sent personal message since then; Rajoy didn't start until the end of September of that year. As a result, the PP (the main right-wing party) concentrated their online activity during the electoral campaign, while the PSOE (the main left-wing party) had had an active presence on Twitter since earlier in the year. On a similar fashion, Obama had more than 15 million followers on Twitter since the beginning of 2012, while Romney did not reach one million followers until a few weeks before the election.

As we show in Figure 3, the volume of tweets included in our dataset largely varies over the electoral campaign in both countries. Furthermore, we find that the intensity of the debate increases during specific political events, such as the candidate debates, the dissolution of the terrorist group ETA, or the attack on the American embassy in Lybia.

3 Results

As indicated in the pages above, one of the most remarkable results in our analysis relates to the deep inequality in the proportion of men and women that are participating in the political discussion on Twitter. As shown in Table 1, a large mayority (60-65%) of the users that were active during our period of study are men and we can clearly reject the hypothesis of an equal representation by gender in both Spain and the US. Similar results have been reported in previous studies (Mislove et al., 2011; Parmelee and Bichard, 2011;

Figure 3: Volume of tweets in our dataset, by day, over our period of analysis



Bode et al., 2011) that also extend this finding to other social networks and contexts.[12]

However, regardless of the descriptive distribution of users, in this paper we are interested in the extent to which gender is a variable affecting behavior on the platform. Table 1 shows that women tend to be more active and they also show a level of sophistication slightly higher than men, as suggested by the number of retweets and @-replies. However, the size of those differences is not large enough to compensate for the general underrepresentation of women on Twitter. In fact, in the case of Spain the difference between the two groups is not statistically significant.

Table 1: Patterns of behavior on Twitter, by gender

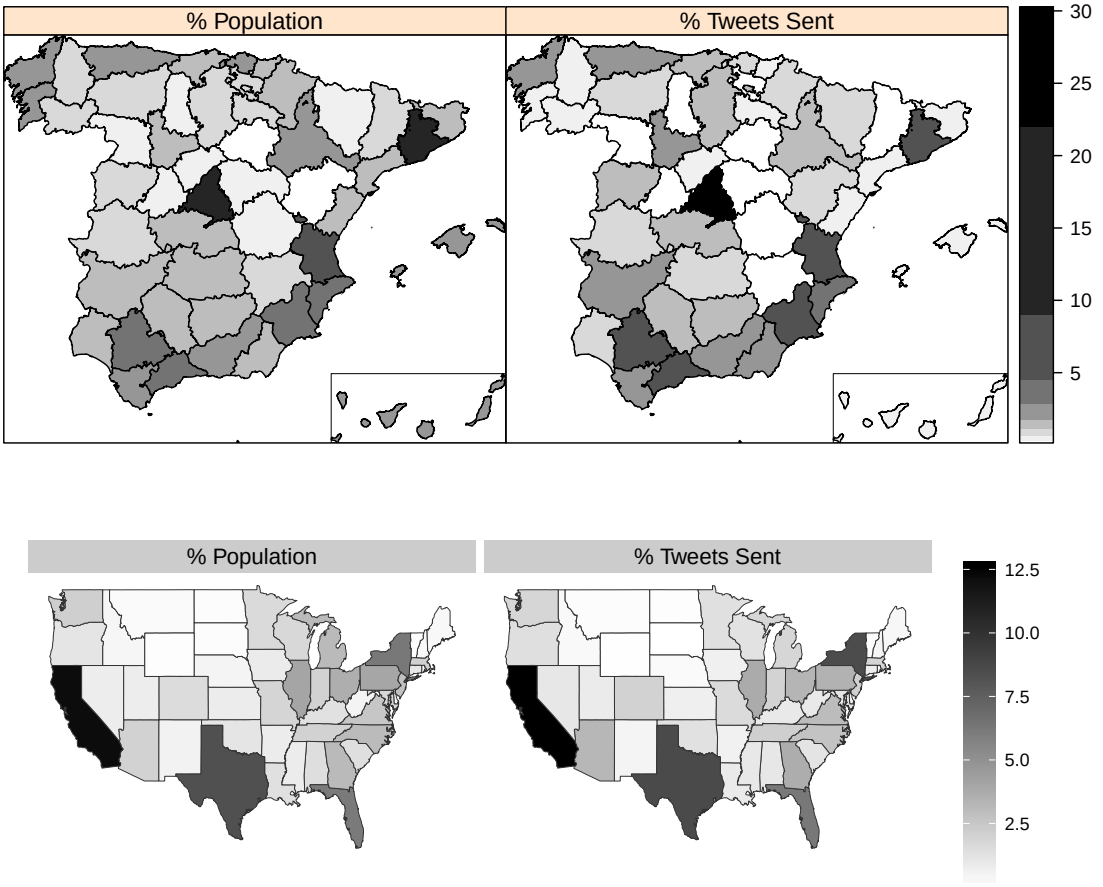
Spain	Variable	Men	Women	All	t-stat
	Number of users	6,192	3,389	9,581	
		[65%]	[35%]	[100%]	29.9*
	Mean of tweets	19.6	21.4	20.2	1.9
	Mean of retweets	8.8	9.5	9.0	1.2
	Mean of replies	3.6	3.9	3.7	1.1
US	Variable	Men	Women	All	t-stat
	Number of users	25,452	17,385	42,837	
		[60%]	[40%]	[100%]	39.6*
	Mean of tweets	74.3	91.0	81.1	5.5*
	Mean of retweets	38.1	55.7	45.2	8.7*
	Mean of replies	7.1	6.7	6.9	0.6

These results are based on the analysis of a random sample of size 12,000 and 50,000 active users in the conversation on Twitter about the elections in Spain and the US. Gender of the users was identified based on the username, although the procedure did not allow to infer the gender of 1,773 users (15%) in Spain and 7,163 (14%) in the US. Significance (equality of means or proportions): * 5%.

Our next variable of interest is the geographical distribution of Twitter users. We see in figure 2 that there are no clear differences in the distribution of the general population and the distribution of active Twitter users at least when we aggregate data at the

[12]Note that our result is only based on tweets of political content. The distribution of male and female users in Twitter at the aggregate level will probably be more homogeneous if we analyzed a different kind of tweets.

Figure 4: Geographical distribution of tweets sent during the electoral campaign



province level in Spain or the state level in the US: more populated provinces/states also show a higher number of users in a roughly similar proportion, with perhaps the only exception of Madrid. This result is robust to other metrics, like the total volume of tweets (figure 4) with a very high correlation (Pearson's ? = 0.846 with N = 52 in Spain, and Pearson's ? = 0.977 with N = 50 in the US). Therefore, it seems clear that Twitter users are a representative sample of the geographical diversity of both countries.

But geography may affect behavior through a different dimension. In particular, the urban/rural cleavage is usually found to be one of the structural factors of political com-

petition (Lipset and Rokkan, 1967) and in fact, the kind of habitat appears as one of the most solid regularities in the prediction of political behavior. More specifically, both in Spain and in the US, large urban centers are usually found to be associated with left-wing/liberal electoral outcomes, and it is a commonplace to point out that voters in rural areas are more likely to vote for conservative parties than those living in large cities.

In order to analyze whether there are differences by kind of habitat, we focus on the town/"municipio" from which each Twitter user has sent messages to the platform. We split the sample in two groups: individuals living in cities with populations larger than 500,000 in Spain or 1,000,000 in the US, and the rest of users. Results are shown in table 2. Although we do not have data for our full sample,[13] our data suggest that localities of smaller size are clearly underrepresented when political issues are debated online. Although less than 17% of the Spanish population lives in one of the six larger cities (Madrid, Barcelona, Valencia, Sevilla, Zaragoza y Málaga), the users in our sample who live in one of them represent around 32% of the total volume of tweets. Similarly, 7% of the US population lives in the 9 most populated cities, but tweets sent from them represent around 21% of the total.

If we focus on different metrics, we do not find significant differences between these two groups in our population: the average number of tweets, retweets and "@-replies" is similar for users located in big cities and the rest for both countries. As a consequence, city size is an important source of inequality, as the smaller number of users is not compensated by a larger number of tweets.

We now present our results in relation to ideology as an explanatory factor. As indicated above, this variable has been estimated using an iterative method based on the

[13]Unfortunately, it is likely that the probability of an individual not indicating his place of residence is correlated with town size (i.e., it is likely that variable is not missing at random). This possibility limits the validity of our analysis and therefore our results should be taken with a grain of salt.

Table 2: Patterns of behavior on Twitter, by size of habitat

Spain	Variable	Cities	Other	All	t-stat
	Number of users	2,007	4,288	6,295	
		[32%]	[68%]	[100%]	
	Mean of tweets	20.4	21.8	21.3	1.1
	Mean of retweets	8.9	9.8	9.6	1.2
	Mean of replies	3.7	3.8	3.7	0.1
US	Variable	Cities	Other	All	t-stat
	Number of users	7,042	26,300	33,342	
		[21%]	[79%]	[100%]	
	Mean of tweets	72.3	68.0	81.1	1.2
	Mean of retweets	38.2	37.3	45.2	0.3
	Mean of replies	5.4	5.7	5.6	0.3

These results are based in a random sample of size 12,000 and 50,000 active users in the conversation on Twitter about the elections in Spain and the US. The location of the tweets was recovered using the information provided by each user, using the geolocation API of Yahoo!, although this procedure did not allows to infer the location of 5,705 users (47%) in Spain and 16.658 users (33%) in the US. Significance (equality of means or proportions): * 5%.

political identification dominant in the network of friends for each user (Barberá, 2014). This procedure generates a score for each individual in a continuous scale with mean 0 and standard deviation 1, in which smaller values are associated with higher intensity in the political self-identification with the political left (liberalism), and higher values are associated with the political right (conservatism). In the results shown below, we restrict our analysis to the hard core of each ideological orientation in Twitter – i.e., to users in our sample whose values in our scale is statistically different from zero.

Results are shown in table 3. Firstly, and with respect to the distribution of users by ideological orientation, we see that users are divided in similar proportions between left and right in both countries. Secondly, we find that the activity of users shows a remarkable ideological bias with respect to the number of tweets, retweets, and "@-replies" published during our period of analysis. In fact, the total participation of users in the extremes of the ideological scale is higher than for users that we could classify as belonging to the

Table 3: Patterns of behavior on Twitter, by party identification

Spain	Variable	Right	Left	All	t-stat
	Number of users	2,895	2,626	12,000	
		[25%]	[23%]	[100%]	
	Mean of tweets	26.4	26.6	20.1	0.1
	Mean of retweets	10.9	13.0	9.0	2.2*
	Mean of replies	4.7	5.1	3.8	1.0
US	Variable	Right	Left	All	t-stat
	Number of users	12,513	12,193	50,000	
		[25%]	[25%]	[100%]	
	Mean of tweets	145.8	95.0	84.0	10.0*
	Mean of retweets	81.4	58.8	46.3	6.6*
	Mean of replies	12.5	7.5	7.0	5.0*

These results are based in a random sample of size 12,000 and 50,000 active users in the conversation on Twitter about the elections in Spain and the US. Party identification was recovered applying an item-response model to the network of friends of the users (Barber, 2014). Significance (equality of means or proportions): * 5%. All differences between those identifying with left and right are statistically significant at the 5% level.

center. To put it differently, this result suggests that the political discussion in Twitter is mainly driven by citizens with extreme values in the ideological scale, a situation that certainly favors the level of polarization of the political discussion on Twitter. This result is consistent with the growing literature on the political consequences of the Internet that claims that these new communication platforms are creating polarized "echo-chamber" environments where individuals are exposed to political information that reinforces their previous beliefs (Adamic and Glance, 2005; Prior, 2007; Sunstein, 2001).

When dividing the sample between right-wing and left-wing users (or conservatives and liberals), we also find other relevant asymmetries. For instance, although in the Spanish case the total number of tweets sent by each group is roughly similar, conservative users retweet at a considerably higher rate than liberal ones. This kind of behavior suggests that the political discussion among right-wing voters is more hierarchical and is more structured, at least if we consider that most tweets from this group come from a rel-

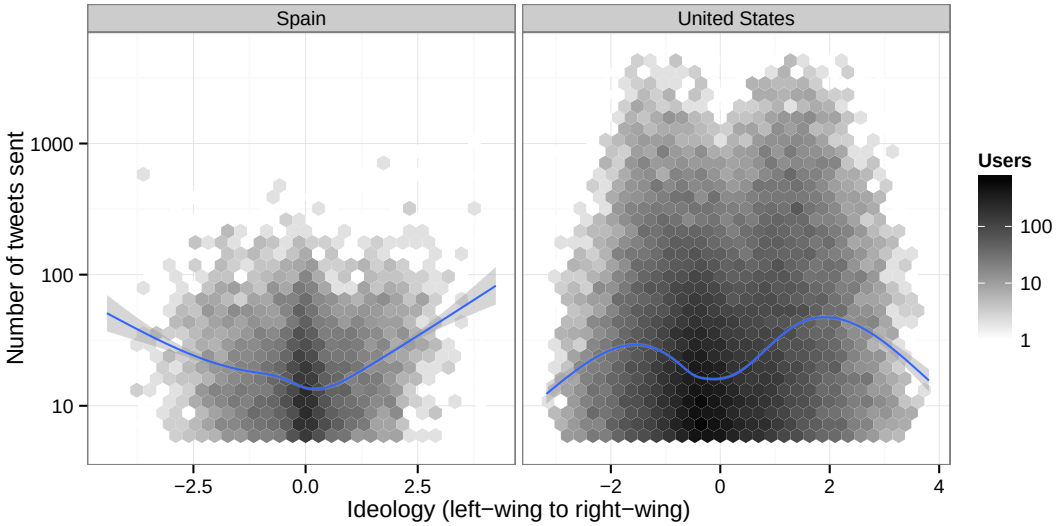
atively reduced number of users that are echoed by individuals politically more closely related to them.

In the US we also find significant differences between ideological groups, but at a deeper level. The users we classify as conservatives tend to be more active in general, with a higher number of tweets coming from this side of the ideological spectrum. This result is consistent with the findings reported in Conover et al. (2012), who showed that Twitter users closer to the Republican party have a more polarized behavior that those leaning towards the Democrats.

Figure 5 complements this analysis of the different patters of behavior arising from party identification by exploiting the fact that our measurement of ideological leaning not only allows us to classify individuals, but also locates them in a continuous scale. In the figure, the color saturation of each hexagon represents the number of users whose ideology and number of tweets are located in that region of the graphic. The image that arises confirms our previous conclusions. Again, we find that individuals with more extreme values in the ideological scale are also associated with a higher number of tweets. Therefore, it is more than reasonable to expect that the political discussion on Twitter is more polarized than it would have been had the number of tweets been more homogeneously spread among users.

Finally, our last dimension of analysis is the relation between participation and the moment of the political campaign. In this case, we are interested on the inequality in participation at the aggregate level. Thus, we returned to the full dataset and we calculated the Gini index for the number of users and tweets mentioning each of the two main candidates in each election. In the Gini index, higher values are associated to higher levels of inequality and it thus allows us to study the extent to which the discussion related to candidates is limited to a few users or, on the contrary, is open to more users with a simi-

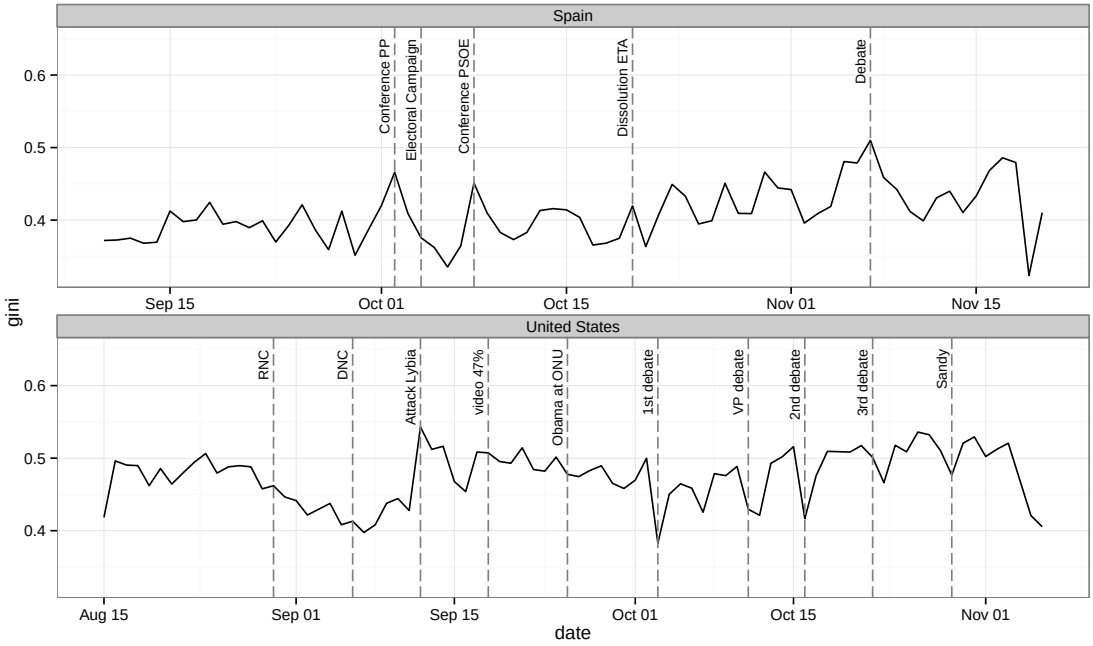
Figure 5: Number of tweets sent during the period of analysis, by ideology



lar level of participation. Moreover, a longitudinal analysis allows us to observe how this index changes as we move into the political campaign. One would expect that, as election day approaches, political promises and campaign events should increase the volume of messages, be that because the most involved users increase their participation or because new users join the discussion about current events.

Figure 6 presents the results of our analysis. The main finding is that inequality in participation changes substantially throughout the campaign as a consequence of the main political events (indicated by dashed lines). In the Spanish case, we find that these events increase inequality or, to put it differently, they increase the rate of participation of the most active users. The picture is more complicated for the US and the effect of those events on the distribution of participation among users seems to depend on the nature of the event. On the one hand, electoral debates attract a higher number of participants, which makes the distribution of tweets more homogeneous. On the other, events such as the attack against the US embassy in Libya show the opposite effect, most likely because that particular event generated an intense response from voters from the Republi-

Figure 6: Evolution of the inequality in the participation on Twitter.



can party.[14]

Hence, the evolution of inequality throughout the campaign suggests that any analysis based on Twitter data not accounting for the changes in behavior due to these exogenous events risks introducing relevant biases arising from the unequal participation of different types of users.

4 Conclusions

We started this article by underlining the opportunities offered by Twitter for the analysis of public opinion: messages are exchanged by numerous users in a public forum and they

[14]To demonstrate that this result is not driven by media outlets and journalists being more active on Twitter during these campaign events, we replicated our analysis excluding all "verified" users. (Verification is granted by Twitter to public figures, including journalists and media outlets, in order to certify that their profile corresponds to their real identity.) The resulting plot (available upon request from the authors) shows that our interpretation is valid.

may contain valuable information about individual preferences and reactions to different political events in an environment that is fully accessible to the researcher. However, the validity of the generalizations that one can make from that potential data source is conditional on our ability to overcome the limitations arising from the fact that participation, at least in the political discussion in this platform, is not homogeneously distributed among users.

Our study about the use of Twitter in the legislative elections of Spain of 2011 and the presidential elections of the United States of 2012 analyzes some of these possible shortcomings. Users participating in the political discussion were mostly men, living in urban areas, and with strong ideological preferences. Even more, our results remark the deep inequality that exists in the behavior of the groups overrepresented in Twitter, namely, that the followers of political parties are more active in the political discussion than the rest. We also found similar patterns along ideological lines, with conservative users being more active than liberal ones.

Beyond these specific results relative to our cases of study, our analysis has relevant, more general implications for the use of Twitter as a tool for electoral prediction. An analysis based on the volume of tweets or their "sentiment" will inherently be biased if it is doe not account for the factors indicated in the pages above. In this sense, our analysis should not be considered as a criticism, but as an opportunity. The same way public opinion research significantly improved after taking more seriously the biases that arise from missing data, sample selection, and social desirability, we also expect researchers using information based on tweets to be able to improve their estimates by taking advantage of a deeper understanding of the patterns of inequality that arise in the political discussion in the Internet.

References

ADAMIC, L. AND N. GLANCE (2005): "The political blogosphere and the 2004 US election: divided they blog," in *Proceedings of the 3rd international workshop on Link discovery*, ACM, 36–43.

AL ZAMAL, F., W. LIU, AND D. RUTHS (2012): "Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors," in *Proceedings of the International Conference on Weblogs and Social Media*.

ASUR, S. AND B. HUBERMAN (2010): "Predicting the future with social media," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, IEEE, vol. 1, 492–499.

BARASH, V. AND J. KELLY (2012): "Salience vs. Commitment: Dynamics of Political Hashtags in Russian Twitter," *Berkman Center Research Publication*.

BARBERÁ, P. (2014): "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data," *Political Analysis*.

BETEBENNER, D. W. (2012): "randomNames: Function for creating gender and ethnicity correct random names," R package available on CRAN.

BIRD, S., E. KLEIN, AND E. LOPER (2009): *Natural language processing with Python*, O'reilly.

BODE, L., A. HANNA, B. SAYRE, J. YANG, AND D. SHAH (2011): "Mapping the Political Twitterverse: Finding Connections Between Political Elites," Unpublished manuscript.

BOLLEN, J., H. MAO, AND X. ZENG (2011): "Twitter mood predicts the stock market," *Journal of Computational Science*.

CAMPBELL, A., P. E. CONVERSE, W. E. MILLER, AND E. DONALD (1960): *The American Voter*, New York: Wiley.

CHA, M., H. HADDADI, F. BENEVENUTO, AND K. GUMMADI (2010): "Measuring user influence in twitter: The million follower fallacy," in *4th International AAAI Conference*

on Weblogs and Social Media (ICWSM).

CHANEY, C. K., R. M. ALVAREZ, AND J. NAGLER (1998): "Explaining the gender gap in US presidential elections, 1980-1992," *Political Research Quarterly*, 51, 311–339.

CHOY, M., M. CHEONG, M. LAIK, AND K. SHUNG (2011): "A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction," *Arxiv preprint arXiv:1108.5520*.

CONGOSTO, M., M. FERNÁNDEZ, AND E. MORO EGIDO (2011): "TWITTER Y POLÍTICA: INFORMACIÓN, OPINIÓN Y PREDICCIÓN?" *Cuadernos de Comunicación Evoca*.

CONOVER, M. D., B. GONÇALVES, A. FLAMMINI, AND F. MENCZER (2012): "Partisan Asymmetries in Online Political Activity," *EPJ Data Science*, 1, 1–19.

CUMMINGS, D., H. OH, AND N. WANG (2010): "Who Needs Polls? Gauging Public Opinion from Twitter Data," Unpublished manuscript.

DIGRAZIA, J., K. MCKELVEY, J. BOLLEN, AND F. ROJAS (2013): "More tweets, more votes: Social media as a quantitative indicator of political behavior," *PloS one*, 8, e79449.

DODDS, P., K. HARRIS, I. KLOUMANN, C. BLISS, AND C. DANFORTH (2011): "Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter," *Arxiv preprint arXiv:1101.5120*.

GAYO-AVELLO, D. (2012): "I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper A Balanced Survey on Election Prediction using Twitter Data," *Arxiv preprint arXiv:1204.6441*.

GOLDER, S. AND M. MACY (2011): "Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures," *Science*, 333, 1878–1881.

GONG, A. (2011): "An automated snowball census of the political web," .

HALE, S., D. GAFFNEY, AND M. GRAHAM (2012): "Where in the world are you? Geolocation and language identification in Twitter," Tech. rep., Working paper.

HANNAK, A., E. ANDERSON, L. BARRETT, S. LEHMANN, A. MISLOVE, AND M. RIEDE-

WALD (2012): "Tweetinin the Rain: Exploring societal-scale effects of weather on mood," .

HECHT, B., L. HONG, B. SUH, AND E. CHI (2011): "Tweets from justin bieber's heart: the dynamics of the location field in user profiles," in *Proceedings of the 2011 annual conference on Human factors in computing systems*, ACM, 237–246.

JUNGHERR, A., P. JURGENS, AND H. SCHOEN (2011): "Why the pirate party won the german election of 2009 or the trouble with predictions: A response to "predicting elections with twitter: What 140 characters reveal about political sentiment"," Social Science Computer Review.

KELLSTEDT, P. M., D. A. PETERSON, AND M. D. RAMIREZ (2010): "The macro politics of a gender gap," *Public opinion quarterly*, nfq003.

LAMPOS, V. (2012): "On voting intentions inference from Twitter content: a case study on UK 2010 General Election," *Arxiv preprint arXiv:1204.0423*.

LAMPOS, V., T. DE BIE, AND N. CRISTIANINI (2010): "Flu detector-tracking epidemics on twitter," *Machine Learning and Knowledge Discovery in Databases*, 599–602.

LIPSET, S. AND S. ROKKAN (1967): *Party systems and voter alignments: Crossnational perspectives*, New York: Free Press.

MARTINEZ, A. AND K. CALVO (2010): "Un análisis del efecto de la Ley de igualdad en la representación electoral, parlamentaria y en el comportamiento electoral de las mujeres en las elecciones generales de 2008," *Estudios de Progreso*, 48.

METAXAS, P., E. MUSTAFARAJ, AND D. GAYO-AVELLO (2011): "How (Not) To Predict Elections," Unpublished manuscript.

MISLOVE, A., S. LEHMANN, Y. AHN, J. ONNELA, AND J. ROSENQUIST (2011): "Understanding the Demographics of Twitter Users," in *Proc. 5th International Conference on Weblogs and Social Media*.

O'CONNOR, B., R. BALASUBRAMANYAN, B. ROUTLEDGE, AND N. SMITH (2010): "From

tweets to polls: Linking text sentiment to public opinion time series," in *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 122–129.

OKEN HODAS, N. AND K. LERMAN (2012): "How Visibility and Divided Attention Constrain Social Contagion," .

PARMELEE, J. AND S. BICHARD (2011): *Politics and the Twitter Revolution: How Tweets Influence the Relationship Between Political Leaders and the Public*, Lexington Books.

PAUL, M. AND M. DREDZE (2011): "You are what you Tweet: Analyzing Twitter for public health," in *Barcelona, Spain: 5th International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*.

PRIOR, M. (2007): *Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections*, Cambridge Univ Pr.

SAKAKI, T., M. OKAZAKI, AND Y. MATSUO (2010): "Earthquake shakes Twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*, ACM, 851–860.

SANG, E. AND J. BOS (2012): "Predicting the 2011 Dutch Senate Election Results with Twitter," *EACL 2012*, 53.

SKORIC, M., N. POOR, P. ACHANANUPARP, E. LIM, AND J. JIANG (2012): "Tweets and Votes: A Study of the 2011 Singapore General Election," in *2012 45th Hawaii International Conference on System Sciences*, IEEE, 2583–2591.

SUNSTEIN, C. (2001): *Republic. com*, Princeton University Press.

TORCAL, M. AND L. MEDINA (2002): "Ideología y voto en España 1979-2000: los procesos de reconstrucción racional de la identificación ideológica," *Revista Española de Ciencia Política*, 6, 57–96.

TUMASJAN, A., T. SPRENGER, P. SANDNER, AND I. WELPE (2010): "Predicting elections with twitter: What 140 characters reveal about political sentiment," in *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 178–185.

WONG, F., S. SEN, AND M. CHIANG (2012): "Why Watching Movie Tweets Won't Tell the Whole Story?" *Arxiv preprint arXiv:1203.4642*.

WU, S., J. HOFMAN, W. MASON, AND D. WATTS (2011): "Who says what to whom on twitter," in *Proceedings of the 20th international conference on World wide web*, ACM, 705–714.