Political Polarization over Global Warming: 
Analyzing Twitter Data on Climate Change

A. Uthor¹, Author²
Author 1 affiliation i.e. department and University
Author 2 affiliation i.e. department and organization
Author1@mail.edu, Author2@mail.org

Abstract

The widespread adoption of social media for political communication creates unprecedented opportunities to monitor the opinions of large numbers of politically active individuals in real time. Specifically, several methods have been recently proposed for predicting the political alignment of Twitter users based on the content and structure of their political communication. In this study, we examine political polarization over climate change within the American public by analyzing data containing tweets from 315,862 users which have been talking climate change over the course of time.

1 Introduction

From the perspective of many countries the political system in the United States is unusual because of its simple two-party nature. Whereas countries as diverse as Switzerland, India, and the United Kingdom have at least occasionally had governing multi-party coalitions in power, U.S. politics is defined by a one-dimensional Democrats vs. Republicans, left vs. right bipolar configuration. The groups primarily associated with a 'left' political identity are democrats and progressives; those primarily associated with a 'right' political identity are Republicans, conservatives, libertarians, and the Tea Party.

Politicians worldwide and in the U.S. in particular have realized the power that social media carries for campaigning. Here, Twitter is on the frontline as it engages users in political debates and, ultimately mobilizes them for grassroots movements. Twitter¹ is a massive social networking site tuned towards fast communication. More than 140 million active users publish over 400 million 140-character "Tweets" every day. Twitter’s speed and ease of publication have made it an important communication medium for people from all walks of life. Twitter has played a prominent role in socio-political events, such as the Arab Spring² and the Occupy Wall Street movement³.

To effectively organize and search tweets, users have to depend on appropriate hashtags inserted into tweets. Within Twitter, hashtags play an important role as labels for ongoing debates that other users can "link to". Hashtags are used consciously by key influencers to frame a political debate and to define the vocabulary used in such debates. There are several examples of "hashtag wars" between Democrats and Republicans⁴.

Given the importance of political opinion formation, researchers have over the last years turned to online data to study the phenomenon of polarization at scale, not only for politicians but for engaged citizens. Adamic et al. [1] observed two strongly separated communities in the political blogosphere, with hyperlinks rarely crossing ideological boundaries. A similar pattern was observed by Conover et al. [2] for the case of Twitter, where users would be unlikely to retweet others with a different party affiliation. Using a combination of network clustering algorithms and manually-annotated data, they demonstrate that the network of political retweets exhibits a highly segregated partisan structure with extremely limited connectivity between left- and right-leaning users.

In 2011, McCright and Dunlap [3] examined political polarization over climate change within the American public by analyzing data from 10 nationally representative Gallup Organizations annual environment poll between 2001 and 2010. They consider six hypotheses one of which is:

- Self-identified liberals/Democrats are more likely to express personal concern about global warming than are self-identified conservatives/Republicans.

Their findings suggest that liberals and Democrats are more likely to report scientific beliefs and express personal concern about global warming than are conservatives and Republicans. In addition to that, their research documents political polarization between elites and organizations identifying the negative environmental consequences of industrial capitalism represented by climate change (e.g., environmental organizations, science advocacy organizations, and Democratic policymakers on the Left) and those defending the economic system from such charges (e.g., conservative think tanks, industry associations, and Republican policymakers on the Right).

Research objective. In this paper, our analysis aims to answer the following research question: Do self-identified liberals/Democrats express personal concern about global warming more than self-identified conservatives/Republicans in social networks such as Twitter?

¹http://Twitter.com
³http://nyti.ms/SwzkvD
⁴See, e.g., bit.ly/Lkjzwm
2 Background

2.1 Twitter Platform

Twitter is a popular social networking and microblogging site extensively explored in recent literature [4, 5]. It has been used to study influence and credibility [6], social structure [7] and to monitor users sentiment [8] among others.

One of Twitter’s defining features is that each tweet is limited to 140 characters. Twitter users can post messages containing text, hyperlinks or hashtags, and interact with one another in a variety of ways. By default, each user’s stream of real-time posts is public. In addition to broadcasting tweets to an audience of followers, Twitter users interact in two public ways: retweets and mentions. Retweets act as a form of endorsement, allowing individuals to rebroadcast content generated by other users, thus raising the content’s visibility [9]. Mentions serve a different function, as they allow someone to address a specific user directly through the public feed, or to refer to an individual in the third person [10].

Hashtags are words prefixed by a # symbol (e.g. #climatechange or #obama), constitute another important feature of the platform, and allow the content produced by many individuals to be aggregated into a custom, topic-specific stream including all tweets containing a given token [11]. Moreover, they make tweets more accessible by hashtag-based search engines such as hashtags.org

These facts, combined with its substantial population of users, renders Twitter an extremely valuable resource for commercial and political data mining and research applications.

2.2 Data Mining on Twitter Data

Much research has focused on detecting significant, unexpected events as they rise in the public feed, since Twitter provides a constant stream of real-time updates from around the globe. Examples of this work include the detection of seismic events [12], influenza outbreaks [13], and the identification of breaking news stories [14].

Another related research area is the application of sentiment analysis techniques to the Twitter corpus. Work by Bollen et al. has shown that indicators derived from measures of ‘mood’ states on Twitter are temporally correlated with events such as presidential elections [15]. Also, Goorha and Ungar used Twitter data to develop sentiment analysis tools for the Dow Jones Company to detect significant emerging trends relating to specific products and companies [16].

Twitter is an ideal platform for monitoring events in real time due to its large scale and streaming nature. However, many of the characteristics that have led to Twitter’s widespread adoption have also made it a prime target for spammers. The detection of spam accounts and content is an active area of research [17,18].

Table 1: Dataset statistics

<table>
<thead>
<tr>
<th>Data</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td># of users</td>
<td>315,862</td>
</tr>
<tr>
<td># of active users</td>
<td>202,790</td>
</tr>
<tr>
<td># of users using hashtags</td>
<td>191,791</td>
</tr>
<tr>
<td># of distinct hashtags</td>
<td>11,746,888</td>
</tr>
<tr>
<td># of original tweets</td>
<td>446,873,576</td>
</tr>
<tr>
<td># of original tweets containing hashtags</td>
<td>103,458,555</td>
</tr>
<tr>
<td># of original tweets containing URLs</td>
<td>197,275,341</td>
</tr>
</tbody>
</table>

Figure 1: Fraction of tweets using P hashtags computed for the set of all original tweets

3 Twitter Data Analysis

Along with all data mining researches done on Twitter, its data has proven valuable for predictions in various domains, including politics. The machine learning approaches paired with sentiment analysis techniques could supplement traditional phone-based opinion surveys by allowing political campaigns to monitor public opinion regarding specific candidates and issues among users in their voting base [19]. As an application of these techniques, here we analyze hashtag and URL usage on a Twitter data set to investigate our research question.

In our study, we collect the Twitter data generated by 315,862 users. To eliminate spammers, only users that posted tweets with climate change and global warming related hashtags are included. We extracted these users from The Carbon Capture Report website [20] which is a free and open service of the University of Illinois devoted to being the preeminent global resource for tracking worldwide perception and developments in Climate Change, Carbon Capture and etc. First, a full list of all Twitter usernames posted a tweet about climate change or global warming were collected. Then, we used the Twitter streaming API [21] to gather the tweets posted by each individual user. 3,200 most recent tweets (which is a limit imposed by the Twitter API) were obtained for each user. Table 1 shows our collected

http://www.carboncapturereport.org
dev.Twitter.com/pages/streaming_api
hashtags they have employed to publish their tweets. First, investigate users political polarization with regard to the suits the analysis.

#Obama) suggesting that our dataset is highly political and are considered as political hashtags (e.g. #tlot, #gop and #obama). In addition, roughly half of the hashtags in this table predominantly by the right- and left-leaning users respectively. The first two hashtags are #tcot ("Top Conservatives on Twitter") and #p2 ("Progressives 2.0"), which are used twice or more hashtags is still significant due to the fact

that the dataset might contain hashtag wars (which we will find later on at this section). Furthermore, the collected users often use hashtags in their tweets while also these hashtags demonstrate interesting facts. Table 2 shows the 30 most frequently used hashtags in our dataset. The first two hashtags are #tcot ("Top Conservatives on Twitter") and #p2 ("Progressives 2.0"), which are used predominantly by the right- and left-leaning users respectively. In addition, roughly half of the hashtags in this table are considered as political hashtags (e.g. #tlot, #gop and #Obama) suggesting that our dataset is highly political and suits the analysis.

**Hashtags categorized by Political Valence.** We investigate users political polarization with regard to the hashtags they have employed to publish their tweets. First,

![Figure 2](https://via.placeholder.com/150)

Figure 2: Percentage of the users with predicted leaning Left and Right, zero political hashtags and undecided based on summation approach.

<table>
<thead>
<tr>
<th>Hashtag</th>
<th>Count</th>
<th>Hashtag</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 #tcot</td>
<td>2,483,853</td>
<td>16 #Syria</td>
<td>318,500</td>
</tr>
<tr>
<td>2 #p2</td>
<td>1,082,667</td>
<td>17 #quote</td>
<td>298,836</td>
</tr>
<tr>
<td>3 #FF</td>
<td>1,018,136</td>
<td>18 #ff</td>
<td>294,757</td>
</tr>
<tr>
<td>4 #fb</td>
<td>805,250</td>
<td>19 #GOP</td>
<td>293,524</td>
</tr>
<tr>
<td>5 #news</td>
<td>749,544</td>
<td>20 #GetGlue</td>
<td>270,986</td>
</tr>
<tr>
<td>6 #auspol</td>
<td>507,224</td>
<td>21 #39</td>
<td>246,327</td>
</tr>
<tr>
<td>7 #leaparty</td>
<td>507,078</td>
<td>22 #News</td>
<td>246,146</td>
</tr>
<tr>
<td>8 #climate</td>
<td>482,740</td>
<td>23 #ObamaCare</td>
<td>242,730</td>
</tr>
<tr>
<td>9 #green</td>
<td>467,972</td>
<td>24 #health</td>
<td>239,733</td>
</tr>
<tr>
<td>10 #tlot</td>
<td>462,158</td>
<td>25 #travel</td>
<td>235,375</td>
</tr>
<tr>
<td>11 #cdnpol</td>
<td>437,534</td>
<td>26 #news</td>
<td>229,852</td>
</tr>
<tr>
<td>12 #Obama</td>
<td>345,824</td>
<td>27 #climatechange</td>
<td>227,510</td>
</tr>
<tr>
<td>13 #jobs</td>
<td>341,933</td>
<td>28 #environment</td>
<td>215,963</td>
</tr>
<tr>
<td>14 #energy</td>
<td>328,902</td>
<td>29 #TCOT</td>
<td>209,745</td>
</tr>
<tr>
<td>15 #1</td>
<td>324,463</td>
<td>30 #DT</td>
<td>207,430</td>
</tr>
</tbody>
</table>

data statistics.

The number of active users is less than the number of all users due to the fact that some of them have deleted their account or set privacy for it, thus its no longer publicly available. The key point about the dataset is the number of distinct users using hashtags and URLs. We separately analyze the usage of hashtags and URLs and their relations to the political polarization in the following subsections. First, we investigate users political polarization with regard to the hashtags they have employed to publish the tweets. Finally, we do a short analysis on the URLs in our dataset.

### 3.1 Hashtag Usage Analysis

There are huge fraction of active users (about 94%) using hashtags in their original tweets, and a substantial fraction of original tweets containing hashtags (<23%). This suggests that many people know how to use hashtags and they frequently tweet using hashtags. As presented in Figure 1, we found out most tweets with hashtag(s) contain only one hashtag. However, the number of tweets containing two or more hashtags is still significant due to the fact that the dataset might contain hashtag wars (which we will demonstrate this later on at this section). Furthermore, the collected users often use hashtags in their tweets while also these hashtags demonstrate interesting facts. Table 2 shows the 30 most frequently used hashtags in our dataset. The first two hashtags are #tcot ("Top Conservatives on Twitter") and #p2 ("Progressives 2.0"), which are used predominantly by the right- and left-leaning users respectively. In addition, roughly half of the hashtags in this table are considered as political hashtags (e.g. #tlot, #gop and #Obama) suggesting that our dataset is highly political and suits the analysis.

**Hashtags categorized by Political Valence.** To identify an appropriate set of political hashtags, we performed a simple co-occurrence discovery procedure. We began with the two most popular political hashtags, #p2 ("Progressives 2.0") and #tcot ("Top Conservatives on Twitter") and we call them key hashtags. For each key hashtag, we identified the set of hashtags with which it co-occurred in at least one tweet, and ranked the results using the Jaccard index. For a set of distinct users U who have used one of a key hashtag, and a set of distinct users H using another hashtag, the Jaccard index between U and H is:

\[
s(U, T) = \frac{|U \cap T|}{|U \cup T|} \tag{2}
\]

Thus, when substantial number of users are using both the key and hashtag, the two are assumed to be related. Using a list of political hashtags is created using the "Political Valence" introduced in [2]. Based on their analysis, a majority of politically active users on Twitter express a political identity in their tweets. Political Valence is a measure that indicates the importance of a hashtag among left- and right-leaning users. Let \(N(h, L)\) and \(N(h, R)\) be the numbers of occurrences of hashtag \(h\) in tweets produced by left- and right-leaning users, respectively. The valence of \(h\) is then defined as:

\[
V(t) = 2 \frac{N(h, R)/N(R)}{N(h, L)/N(L) + N(h, R)/N(R)} - 1 \tag{1}
\]

where \(N(R) = \sum h N(h, R)\) is the total number of occurrences of all hashtags in tweets by right-leaning users and \(N(L)\) is defined analogously for left-leaning users. A table of mostly used hashtags used by left- and right-leaning users have been extracted based on Political Valence in [2]. For each hashtag in this table, we calculated the number of times it has been used distinctively by different users. Our findings suggest that users have used left leaning hashtags more often than hashtags in the opposite groups (94,506 distinct users for left and 71,728 for right).
Table 3: RandomForest performance shown by 6 performance measures based on 10-fold cross validation.

<table>
<thead>
<tr>
<th></th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>ROC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>0.98</td>
<td>0.002</td>
<td>0.998</td>
<td>0.98</td>
<td>0.989</td>
<td>0.999</td>
</tr>
<tr>
<td>R</td>
<td>0.998</td>
<td>0.02</td>
<td>0.98</td>
<td>0.998</td>
<td>0.989</td>
<td>0.999</td>
</tr>
<tr>
<td>Weighted Avg.</td>
<td>0.989</td>
<td>0.011</td>
<td>0.989</td>
<td>0.989</td>
<td>0.989</td>
<td>0.999</td>
</tr>
</tbody>
</table>

Table 4: Hashtags related to #p2 and #tcot

| Just #p2 | #Obama2012 #dems #p21 #p2b #1u #ows #occupy #Romney #oaf #rebellest #union #p1 #libertarian #democrats |
| Just #tcot | #912 #ucot #ccot #tpp #tgpn #Benghazi #liberty #912project #RonPaul #pjnet #NRA #military #rush |

this measure, we are able to eliminate the impact of hashtags that are only used by certain users and might not be related to the desired context. 27 distinct hashtags were identified with exclusion of 8 ambiguous and overly-broad hashtags (e.g. #news, #politics, #economy, etc.). Table 4 shows the list of extracted hashtags. We calculated the number distinct people only used these hashtags to determine the percentage of left- and right-leaning users. 75,492 number of of distinct users posted tweets using the terms mentioned in Table 4. 67% of the these are considered to be left-leaning.

Users Classified by RandomForest. In order to justify our assumption about the collected data, we use a classification approach to detect user’s most probable leaning. The method consist of the followings:

- **Hashtag frequency vector.** Using information collected from previous subsections and the lists of hashtags which are available for left- and right-leaning users deploying Political Valence and Jaccard index, we select two subsets of hashtags named Left and Right. Each set contains specified number of hashtags which are known as popular and highly used by users in our data. We associate two distinct vectors to every user containing hashtags in Left and Right sets weighted by their frequency.

- **Summation.** Next, for each user we do a summation on both Left and Right vectors associated with her. Three cases will occur at this point. If one of the following rules apply to these summations we can predict the user leaning based on the summations:

  \[
  \text{sum}(user.Left) > \text{sum}(user.Right) \times 1.5 \quad (3)
  \]

  \[
  \text{sum}(user.Right) > \text{sum}(user.Left) \times 1.5 \quad (4)
  \]

  The 1.5 ratio is applied to make sure that the user’s predicted leaning will definitely fall in one camp or the other. However, if a specific user does not use any of the hashtags in Left and Right sets or the summations on both vectors are equal to each other for her, she will be called "undecided". Our experiments suggest that the number of users with equal summations is small and can be excluded. Out of 202,790 active users in the data, 62,657 and 17,648 were detected as left and right-leaning respectively. On the other hand, about 58% (118,473) of the users do not fall in left or right camps so the third step is used to make a decision about them. Figure 2 reports the results for this experiment.

- **Classification using RandomForest.** Finally, to do classification on left over users from previous steps, we deploy RandomForest classifier from Weka 3.7 [20] an open source data mining package. RandomForest is trained by the information collected from two groups of users (detected left- and right-leaning) formed in previous step. We extract top 200 frequent hashtags (excluding the hashtags used for the previous step) from each user’s tweets and its label is assigned to it (Left or Right). To determine the performance of our trained classifier we use 10-fold cross validation. It is worth mentioning that since the number of users with left-leaning is about four times more than right-leaning users, we oversampled our right-leaning users (basically copied them four times) to overcome the problem of imbalanced data. Table 5 reports different accuracy measures for this test. Finally, we create hashtag vectors for the users we don’t know their leaning with a similar process. This set of users are considered as test case for our trained classifier. Out of 118,473 users, 106,966 of them were classified as left-leaning and the rest 11,507 as right-leaning.

3.2 URL Usage Analysis

As for the final analysis on the Twitter data, we study popular URLs among left- and right-leaning users. [21] produced a ranked lists of the domains most frequently tweeted by users of each political alignment, based on the predictions of their network classification method. Analyzing URLs in tweets is tricky since many Twitter users rely on URL shortening services to hash hyperlinks into a more compact format. In order to get accurate results, we collected data both on the links and their encoded versions using the popular
bit.ly platform. These results indicate that URLs assigned to left-leaning users (4,335) have appeared more frequently that left-leaning ones (3,427).

4 Conclusions

The widespread adoption of social media for political communication creates unique opportunities to capture the opinions of large numbers of politically active individuals in real time. Using a combination of Random Forest classifier and a rigorously constructed dataset, we demonstrated that self-identified liberals/Democrats are more likely to express personal concern about global warming. This phenomena occurs not only in blogosphere or poll data, but also in social websites such as Twitter. We witnessed that people who were predicted to be left-leaning, post tweets related to Climate Change more often.

References


