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

Alireza Hajibagheri, Gita Sukthankar
Department of EECS, University of Central Florida
alireza@eecs.ucf.edu, gitars@eecs.ucf.edu

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. 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 within the American public over the existence of anthropogenic climate change by analyzing Twitter data from 315,862 users who have participated in climate change discussions.

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 largely dominated by discourse between the Democrats and the Republicans and can be expressed as a left-right political spectrum of viewpoints. 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.

Twitter[1] is a massive social networking site tuned towards rapid communication. Politicians worldwide have realized the power that social media carries for campaigning. Here, Twitter is on the frontline as it provides a forum for users to engage in political debates and mobilize grassroot movements. 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 Spring[2] and the Occupy Wall Street movement[3].

To find relevant content, 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 Republicans[4].

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 described by Conover et al. [2] for Twitter discourse, where users are less likely 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 (1) report scientific beliefs and (2) 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 (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). 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?

[1] http://www.twitter.com
[2] http://bit.ly/N6illb
[3] http://nyti.ms/SwZKVD
[4] See, e.g., http://bit.ly/Lkzjwm

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 a broad range of topics including influence and credibility [6], social structure [7] and user sentiment [8].

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, 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 through hashtag-based search engines such as hashtags.org[5].

These features, combined with its substantial population of users, render Twitter an extremely valuable resource for commercial and political data mining, as well as an asset for social science researchers.

2.2 Data Mining on Twitter Data

Much research has focused on detecting significant, unexpected events as they trend 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].

[5]http://www.hashtags.org

Table 1: Dataset statistics

Data	Count
# of users	315,862
# of active users	202,790
# of users using hashtags	191,791
# of distinct hashtags	11,746,888
# of original tweets	446,873,576
# of original tweets containing hashtags	103,458,555
# of original tweets containing URLs	197,275,341

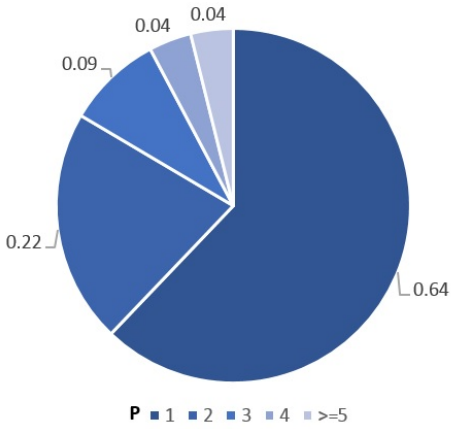

Figure 1: Fraction of tweets using P hashtags computed for the set of all original tweets

3 Twitter Data Analysis

Apart from research applications, Twitter data has proven valuable for predictions in various domains, including politics. 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]. In this paper, we analyze hashtag and URL usage on a Twitter data set to investigate the relationship between political viewpoint and frequency of climate change related discourse.

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[6] which is a free and open service of the University of Illinois devoted to tracking worldwide perception and developments in climate change, carbon capture, and related topics. First, we extracted a full list of all Twitter usernames who posted tweets about climate change or global warming. Then, we used the Twitter streaming API[7] to gather the tweets posted by each individual user. The 3,200 most recent tweets (which is a limit imposed by the Twitter API) were obtained for each user. Table 3 shows the statistics of our dataset.

[6]http://www.carboncapturereport.org
[7]dev.Twitter.com/pages/streaming_api

Table 2: Frequency of hashtag appearance

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

Due to the fact that some users have deleted their accounts or changed the privacy settings, the number of active users is less than the user total. This paper focuses on the subset of distinct users using hashtags and URLs. We separately analyze the usage of hashtags and URLs and their relationship to 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

A large subset of the active users (about 94%) use hashtags in their original tweets, and a substantial fraction of original tweets contain hashtags (<23%). This suggests that many people know how to use hashtags and they frequently tweet using hashtags. As shown in Figure 1, most tweets with hashtag(s) in our dataset contain only a single hashtag. However, the number of tweets containing two or more hashtags is still significant due to the fact that the dataset contains hashtag wars. Table 2 shows the thirty 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, a list of political hashtags is created using the *Political Valence* measure 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:

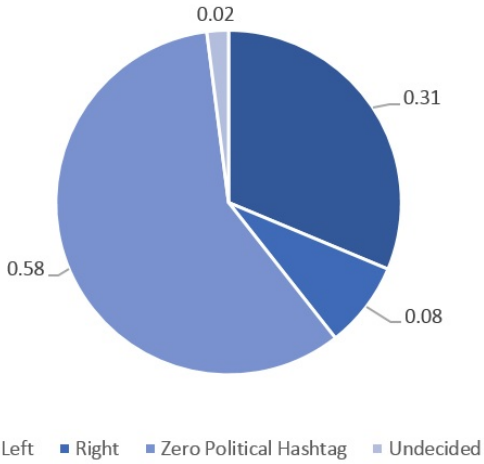

Figure 2: Percentage of the users in our dataset with predicted leaning (Left and Right), zero political hashtags, and undecided based on the summation approach.

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

where $N(R) = \sum_t N(h, R)$ 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 frequently used hashtags employed 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. The 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).

Hashtags categorized by Jaccard Index. 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:

$$\sigma(U, T) = \frac{|U \cap T|}{|U \cup T|} \qquad (2)$$

Thus, when substantial number of users are using both the key and hashtag, the two are assumed to be related. Using 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

Table 3: RandomForest performance based on 10-fold cross validation.

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
L	0.98	0.002	0.998	0.98	0.989	0.999
R	0.998	0.02	0.98	0.998	0.989	0.999
Weighted Avg.	0.989	0.011	0.989	0.989	0.989	0.999

Table 4: Hashtags related to #p2 and #tcot

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

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 who 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 users' most probable leaning. The method consist of the following procedure:

1. **Hashtag frequency vector.** Using Political Valence and Jaccard Index, we select two subsets of hashtags named Left and Right. Each set contains a specified number of hashtags which are known to be popular and highly used by users in our dataset. Two distinct vectors were assigned to every user containing hashtags in the Left and Right sets weighted by their frequency.

2. **Summation.** Next, for each user we do a summation on both Left and Right vectors. 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:

$$sum(user.Left) > sum(user.Right) * 1.5 \quad (3)$$

$$sum(user.Right) > sum(user.Left) * 1.5 \quad (4)$$

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 the Left and Right sets or the summations of both vectors are equal to each other, the user 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.

3. **Classification using RandomForest.** Finally, to classify the remaining users, we deploy the RandomForest classifier from Weka 3.7 [20] an open source data mining package. RandomForest is trained using the information collected from two groups of users (detected left- and right-leaning). We extract the top 200 most frequently used hashtags (excluding the hashtags used for the previous step) from each user's tweets and assigned a label (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 left-leaning users in our dataset is about four times more than right-leaning users, we oversampled our right-leaning users to overcome the problem of imbalanced data. Table 5 reports different accuracy measures for this test. Finally, we create hashtag vectors for the undetermined users with a similar process. This set of users are considered as the 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

For a 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[8] platform. These results indicate that URLs assigned to left-leaning users (4,335) have appeared more frequently than right-leaning ones (3,427) in our dataset.

[8]https://bitly.com/

4 Conclusion

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 the RandomForest classifier and a rigorously constructed dataset, we conclude that self-identified liberals/Democrats are more likely to express personal concern about global warming, as manifested by increased Twitter activity. This phenomena occurs not only in blogosphere or poll data, but also in social websites such as Twitter.

5 Acknowledgments

The authors would like to thank Hossein Rahimi for his helpful insights. This research was funded in part by NSF award IIS-0845159.

References

[1] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 US election: divided they blog," in *Proceedings of the International Workshop on Link Discovery*. ACM, 2005, pp. 36–43.

[2] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini, "Political polarization on Twitter," in *International Conference on Weblogs and Social Media*, 2011.

[3] A. M. McCright and R. E. Dunlap, "The politicization of climate change and polarization in the American public's views of global warming, 2001–2010," *The Sociological Quarterly*, vol. 52, no. 2, pp. 155–194, 2011.

[4] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proceedings of the International Conference on World Wide Web*. ACM, 2010, pp. 591–600.

[5] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts, "Who says what to whom on Twitter," in *Proceedings of the International Conference on World Wide Web*. ACM, 2011, pp. 705–714.

[6] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in *Proceedings of the International Conference on World Wide Web*. ACM, 2011, pp. 675–684.

[7] P. A. Grabowicz, J. J. Ramasco, E. Moro, J. M. Pujol, and V. M. Eguiluz, "Social features of online networks: The strength of intermediary ties in online social media," *PloS One*, vol. 7, no. 1, 2012.

[8] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[9] B. Danah, G. Scott, and L. Gilad, "Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter," *HICSS-43, January*, vol. 6, 2010.

[10] C. Honeycutt and S. C. Herring, "Beyond microblogging: Conversation and collaboration via Twitter," in *Hawaii International Conference on System Sciences*. IEEE, 2009, pp. 1–10.

[11] M. D. Conover, C. Davis, E. Ferrara, K. McKelvey, F. Menczer, and A. Flammini, "The geospatial characteristics of a social movement communication network," *PloS One*, vol. 8, no. 3, p. e55957, 2013.

[12] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: real-time event detection by social sensors," in *Proceedings of the International Conference on World Wide Web*. ACM, 2010, pp. 851–860.

[13] A. Culotta, "Towards detecting influenza epidemics by analyzing Twitter messages," in *Proceedings of the Workshop on Social Media Analytics*. ACM, 2010, pp. 115–122.

[14] S. Petrović, M. Osborne, and V. Lavrenko, "Streaming first story detection with application to Twitter," in *Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 181–189.

[15] J. Bollen, H. Mao, and A. Pepe, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena." in *International Conference on Weblogs and Social Media*, 2011.

[16] S. Goorha and L. Ungar, "Discovery of significant emerging trends," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010, pp. 57–64.

[17] S. Yardi, D. Romero, G. Schoenebeck *et al.*, "Detecting spam in a Twitter network," *First Monday*, vol. 15, no. 1, 2009.

[18] A. H. Wang, "Don't follow me: Spam detection in Twitter," in *Proceedings of the Internatinal Conference on Security and Cryptography (SECRYPT)*. IEEE, 2010, pp. 1–10.

[19] N. A. Diakopoulos and D. A. Shamma, "Characterizing debate performance via aggregated Twitter sentiment," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2010, pp. 1195–1198.

[20] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[21] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer, "Predicting the political alignment of Twitter users," in *International Conference on Social Computing*, 2011, pp. 192–199.