Digital Trace Data in the Study of Public Opinion: An Indicator of Attention Toward Politics Rather Than Political Support—Online Appendix

Andreas Jungherr, University Mannheim, Mannheim, Germany
Harald Schoen, University Mannheim, Mannheim, Germany
Oliver Posegga, Bamberg University, Bamberg, Germany
Pascal Jürgens, Johannes Gutenberg-University, Mainz, Germany

Appendix 1: Character strings used to identify politically relevant messages on Gnip

We queried the API of *Gnip's Historical Powertrack* (http://support.gnip.com/apis/historical_api/) for all messages containing the following character substrings irrespective of capitalization. This collection covers mentions of political parties in Germany, prominent candidates, campaign related keywords, and important campaign related media events in various spelling variations: cdu, cducsu, csu, spd, die_linke, dielinke, linke, linkspartei, linken, buendnis90, bündnis90, bündnis90diegrünen, bündnis90grüne, bündnisgrüne, bündnisgrünen, die_gruenen, die_grünen, diegrünen, gruene, grüne, grünen, gruenen, fdp, afd, piraten, piratenpartei, merkel, angie_merkel, angelamerkel, angela_merkel, seehofer, horstseehofer, horst_seehofer, steinbrück, steinbrueck, peer_steinbrück, peer_steinbrueck, gysi, gregorgysi, gregor_gysi, wagenknecht, sahrawagenknecht, sahra_wagenknecht, göring-eckardt, goering-eckardt, göringeckardt, goeringeckardt, katringöring-eckardt, katringöringeckardt, katringoering-eckardt, katringoeringeckardt, katrin_göring-eckardt, katrin_goering-eckardt, katrin_göringeckardt, katrin_goeringeckardt, katrin_göringeckardt, katrin_goeringeckardt, katrin_göring_eckardt, katrin_goering_eckardt, katringoering_eckardt, katringöring_eckardt, göring_eckardt, goering_eckardt, trittin, jürgentrittin, juergentrittin, jürgen_trittin, juergen_trittin, brüderle, bruederle, rainerbrüderle, rainerbruederle, rainer_brüderle, rainer_bruederle, lucke, berndlucke, bernd_lucke, btw13, bundestagswahl, wahlkampf, btw2013, wahl13, tv-duell, wahlarena, dreikampf, kanzlerduell.

Appendix 2: Strings used to identify party mentions

To identify messages referring to political parties we searched the original data set for occurrences of party names in messages be it through keywords or hashtags. The following list covers the spelling variations of party names we counted as mentions of specific parties irrespective of capitalization. CDU: cdu, cducsu; SPD: spd; Die LINKE: die_linke, dielinke, linke, linken, linkspartei; Bündnis 90/Die Grünen: buendnis90, bündnis90, bündnis90diegrünen, bündnis90grüne, bündnisgrüne, bündnisgrünen, die_gruenen, die_grünen, diegrünen, gruene, grüne, grünen, gruenen; CSU: csu; FDP: fdp; AfD: afd; Piratenpartei: piraten, piratenpartei.

Appendix 3: Sentiment Analysis

As described in our paper, we used three approaches to the classification and analysis of sentiment contained in tweets mentioning political parties. First, we had one percent of original tweets—no retweets—mentioning a political party hand coded for negative, neutral, or positive sentiment towards the mentioned party. Second, we used an automated approach for the detection of sentiment (Hopkins & King 2010). Third, we analyzed messages posted by users containing hashtags explicitly supporting or critiquing a party (e.g. #cdu+ or #cdu-). Here we provide background information to the first two approaches.

1) In a first step, we randomly selected one percent of original tweets mentioning a political party by keyword or hashtag. This selection excluded retweets. This resulted in a selection of 6,479 tweets. The distribution of these selected tweets across parties is documented in Table 6. We had two research assistants code these messages by hand on their sentiment towards the mentioned party. We had them differentiate between negative, neutral, and positive sentiment. The results are also documented in Table 6.

   To calculate inter-coder reliability we led both assistants code a random selection of 900 messages. Both agreed on their coding with 714 tweets and disagreed in 186 cases. On this basis we calculated Cohen's Kappa (Cohen, 1960), resulting in a value of 0.671 indicating substantial agreement between the coders.

2) In a second step, we decided to use an automated content analyses approach to assess the sentiment in all tweets mentioning parties. The automated sentiment analysis of tweets is non-trivial (Gayo-Avello, 2012). Even for English—a language most development in the automated detection of sentiment is focused on—, results of dictionary-based approaches have been shown to provide far from stable results (González-Bailón & Paltoglou, 2015). This led us to use an approach for the semi-automated content analysis of text corpora by Hopkins & King (2010). While developed originally for corpora of larger texts than tweets, the approach has been used in a commercial setting for the automated classification of tweets by the firm *Crimson Hexagon*. The approach has also been used in the past by one research team with apparent success in the classification of political mentions on Twitter (Ceron et al 2014; 2015). Given this, the approach seemed a sensible choice as second assessment of the sentiment in tweets mentioning political parties. To train the algorithm, we used the hand coded data set of tweets described before. The results of the approach are documented in Table 6.

Table 6: Tweet Sentiment

Party	**No. observations (hand coded)**	**Share of Neutral mentions (hand coded)**	**Share of Neutral mentions (Hopkins/King)**	**Share of Positive mentions (hand coded)**	**Share of Positive mentions (Hopkins/King)**	**Share of Negative mentions (hand coded)**	**Share of Negative mentions (Hopkins/King)**
CDU/CSU	1048	55.57	50.05	6.72	5.71	37.72	44.24
CDU	836	46.01	40.83	7.49	7.77	46.50	51.41
SPD	920	57.22	48.30	12.47	8.02	30.31	43.68
Die LINKE	477	59.48	50.08	26.78	27.91	13.74	22.01
Die Grünen	389	60.10	61.30	7.61	7.18	32.28	31.52
CSU	355	49.44	41.66	5.65	5.81	44.92	52.52
FDP	794	49.23	59.76	5.61	4.67	45.15	35.57
AfD	617	38.69	33.35	30.82	30.53	30.82	36.12
Piraten	1043	57.49	54.44	31.15	32.50	11.36	13.06

The table reports results of two sentiment analyses of tweets mentioning politics parties. The first column reports the total number of tweets hand coded for mentions of the respective parties by keywords or hashtags. The following columns report the shares of neutral, positive, and negative mentions for each party. For example, of all hand coded mentions of the CDU/CSU 55.57% were coded as containing neutral sentiment. These results are reported both for the hand coded sample and the automated analysis based on Hopkins and King (2010). The results reported for the CDU/CSU are based on a separate sample of mentions of either CDU or CSU. They, therefore, do not necessarily match the aggregation of CDU and CSU sentiment.

Appendix 4: Party shares of various Twitter-based mention metrics

Table 7: User and mention shares between July 1 and September 22, 2013

Party	**Vote share, 2013**	**Vote share, 2009**	**Polling results (median)**	**User share, keywords**	**User share, hashtags**	**Keyword share**	**Hashtag share**	**Positive sentiment share (hand coded)**	**Positive sentiment share (Hopkins/ King)**	**Positive sentiment share (#+)**	**Negative sentiment share (hand coded)**	**Negative sentiment share (Hopkins/ King)**	**Negative sentiment share (#-)**
CDU/CSU			40										
CDU	34.1	28.43		42.69	39.69	16.08	13.65	6.98	6.86	5.08	23.22	22.23	18.73
SPD	25.7	24.01	25	39.99	37.29	17.35	13.84	12.84	7.09	4.53	16.71	18.91	10.73
Die LINKE	8.6	12.39	8	31.71	17.84	11.03	6.27	12.73	13.22	6.55	3.50	5.10	1.05
Die Grünen	8.4	11.16	13	24.28	22.30	7.81	7.36	3.27	3.05	1.12	7.42	6.55	7.23
CSU	7.4	6.8		22.92	21.67	6.98	6.40	2.25	2.30	1.21	9.59	10.18	9.16
FDP	4.8	15.18	5	45.92	46.96	16.65	13.18	4.95	4.00	3.60	21.35	14.91	14.34
AfD	4.7		3	28.75	28.11	10.15	10.23	21.17	25.17	40.32	11.22	14.58	35.53
Piraten	2.2	2.04	3	28.50	34.63	13.96	29.07	35.81	38.31	37.58	7.00	7.53	3.22

The table reports vote shares of the parties included in the analyses in the federal election on September 22, 2015. Vote share reflects the share of votes for each party on the total of all votes collected by all parties included in the analysis. This might lead the vote shares reported here to deviate from the official results.

Appendix 5: Correlations between polls and Twitter-based time series at larger lags

Table 8: Correlations between time series of Twitter-based metrics and polls at larger lags

Party	Metrics	Corr. polls (lag -4)	Corr. polls (lag -3)	Corr. polls (lag -2)	Corr. polls (lag +2)	Corr. polls (lag +3)	Corr. polls (lag +4)
CDU/CSU	Keyword Mentions	.070	-.189	-.042	-.016	.045	.054
	Hashtag Mentions	.026	-.099	.099	.130	.317	.167
	Positive Sentiment (#+)	-.281	.002	.189	-.065	-.037	-.009
SPD	Keyword Mentions	.029	.061	-.189	-.009	.071	.078
	Hashtag Mentions	.071	-.082	.074	.039	.163	.101
	Positive Sentiment (#+)	-.040	.185	-.061	-.035	-.283	.156
Die LINKE	Keyword Mentions	.080	.157	-.134	.082	.177	-.145
	Hashtag Mentions	-.097	.115	-.147	.255	.077	.104
	Positive Sentiment (#+)	-.220	-.214	-.107	-.022	-.173	.008
Die Grünen	Keyword Mentions	-.001	-.218	-.124	.045	-.262	-.156
	Hashtag Mentions	-.366	-.467	-.325	-.254	-.444	-.302
	Positive Sentiment (#+)	.256	.274	.185	.152	.280	.087
FDP	Keyword Mentions	.140	.229	.067	.154	.036	.058
	Hashtag Mentions	.118	.165	-.010	.366	.272	.386
	Positive Sentiment (#+)	-.107	-.146	-.050	-.337	-.187	-.389
AfD	Keyword Mentions	.340	.464	.258	.375	.468	.461
	Hashtag Mentions	.526	.602	.449	.273	.437	.538
	Positive Sentiment (#+)	.115	.272	.117	.047	.010	.068
Piraten	Keyword Mentions	.045	.091	-.012	-.121	-.085	.055
	Hashtag Mentions	.157	-.087	.129	-.103	-.066	.155
	Positive Sentiment (#+)	-.332	-.121	-.128	-.131	.110	-.130

Correlation Polls (lag -2 to lag -4) documents the correlation between opinion polls and each Twitter-based time series with Twitter-based time series of lagged by two, three or four days, respectively. Correlation Polls (lag +2 to lag +4) documents the correlation between opinion polls and each Twitter-based time series with opinion polls lagged by two, three or four days, respectively.

Appendix 6: Plots of the time series of party mentions by keywords or explicitly positive hashtags (#+)

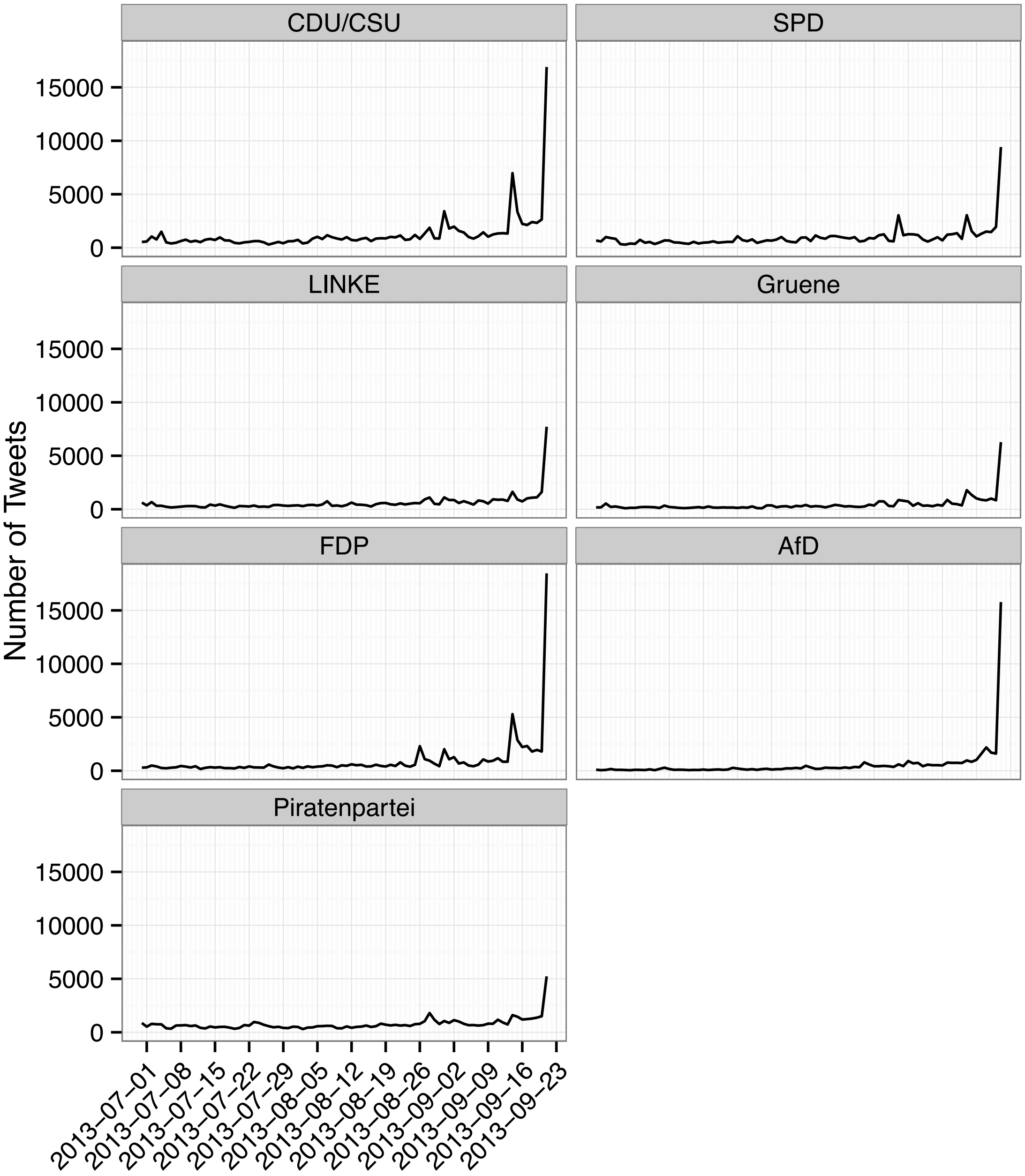

Figure 3: Party mentions, keywords between July 1 and September 22 (6 p.m.), 2013

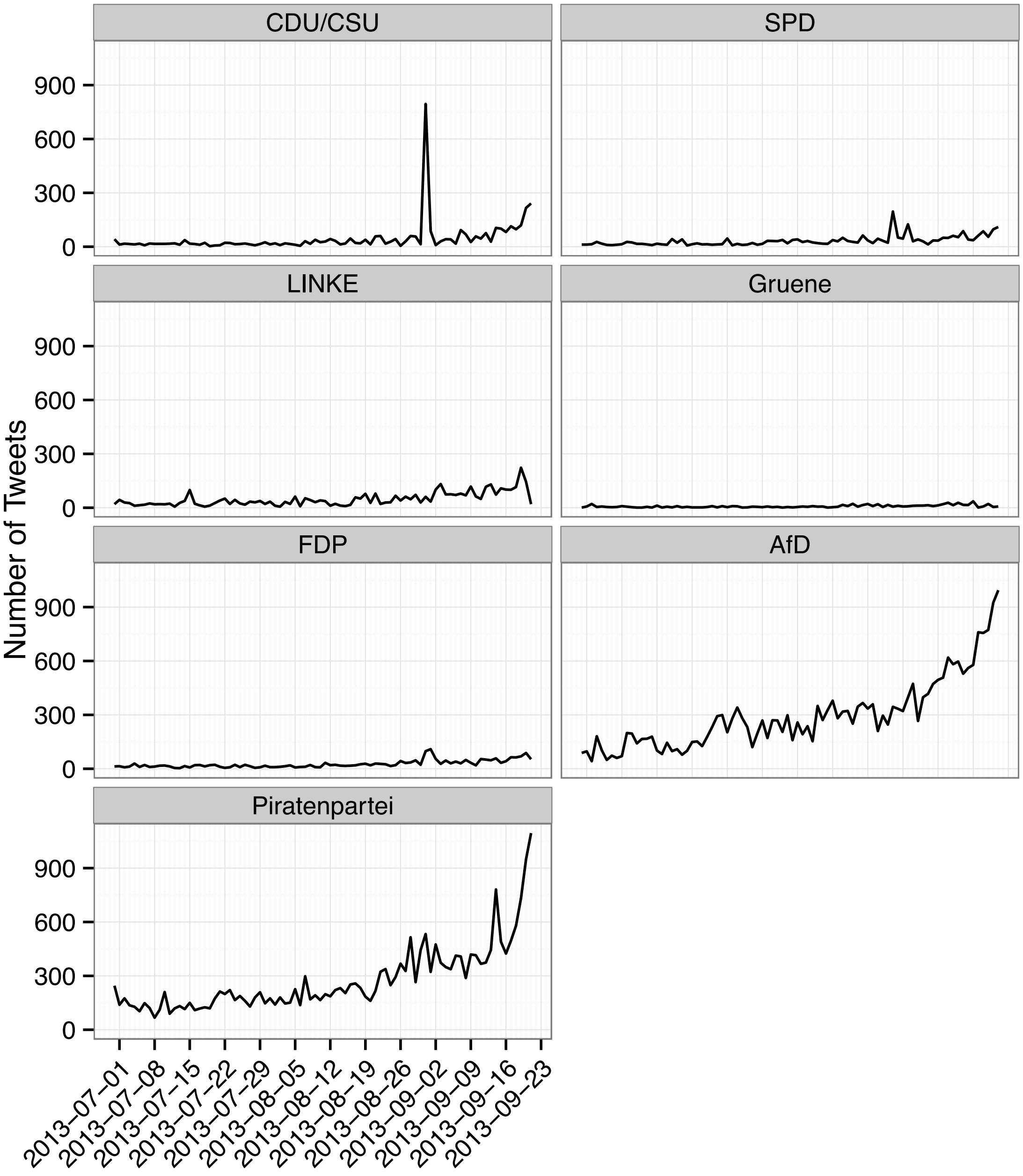

Figure 4: Party mentions, positive sentiment between July 1 and September 22 (6 p.m.), 2013

Appendix Bibliography:

Ceron, A., Curini, L., & Iacus, S.M. (2014). "Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France." *New Media & Society* 16: 340–358.

Ceron, A., Curini, L., & Iacus, S.M. (2015). "Using Sentiment Analysis to Monitor Electoral Campaigns: Method Matters—Evidence From the United States and Italy." *Social Science Computer Review* 33(1):3-20.

Cohen, J. (1960). "A coefficient of agreement for nominal scales". *Educational and Psychological Measurement* 20 (1): 37–46.

Gayo-Avello, D. (2012). "No, you cannot predict elections with Twitter." *IEEE Internet Computing* 16: 91-94.

González-Bailón, S., & Paltoglou, G. (2015). "Signals of Public Opinion in Online Communication: A Comparison of Methods and Data Sources." *The ANNALS of the American Academy of Political and Social Science* 659(1): 95-107.

Hopkins, D., & King, G. (2010). "A Method of Automated Nonparametric Content Analysis for Social Science," *American Journal of Political Science*, 54(1): 229–247.