Election Forecasting

GRAD-E1234

Digital Trace Data Models

Simon Munzert

Spring Semester 2017
Hertie School of Governance

Session outline

The online data revolution
A thought experiment
The total survey error framework revisited
Gathering data from the web with R
Applications

The online data revolution

- data abundance online
- social interaction online
- services track social behavior
- behavioral data often available for free
- data often available in (near) real-time
- classical data collection techniques transferred to online platforms (online survey panels, crowdsourcing tools, ...)

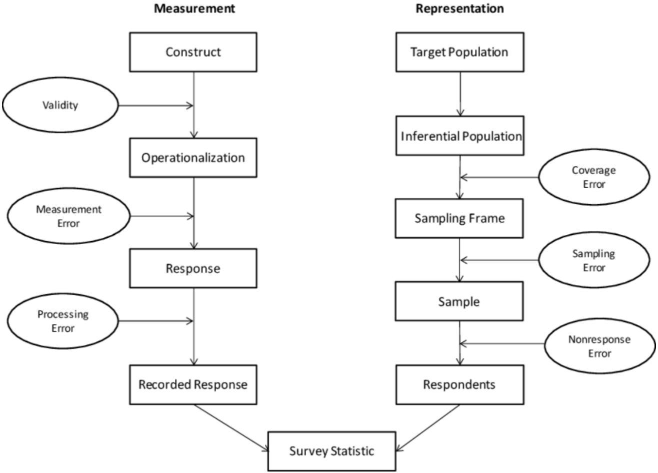
A thought experiment

Imagine you had the technical expertise and opportunity to access and process behavioral data on the web at a large scale.

1. how would you use it to forecast an upcoming election?
2. how would you guard against potential biases?
3. how would you assess the power of your approach (also relative to other, more conventional techniques) before the election takes place?

You have 15 minutes to prepare a 2-minute pitch of your approach with your neighbor!

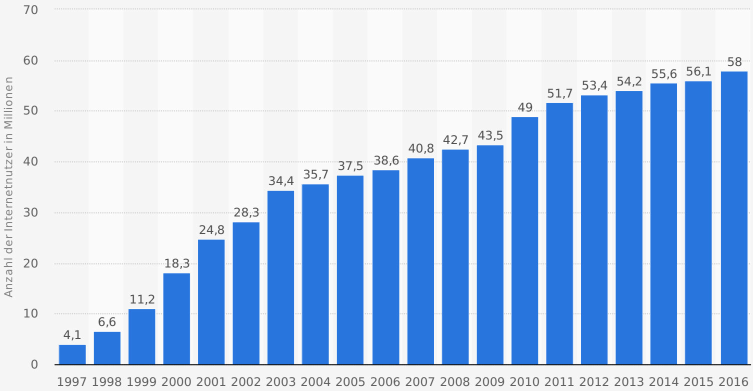
Sources of survey error



- in analogy to the above framework, can you identify sources of error in social media-based forecasts?

WELCOME TO THE INTERNET
PLEASE FOLLOW ME

Anzahl der Internetnutzer in Deutschland in den Jahren 1997 bis 2016 (in Millionen)

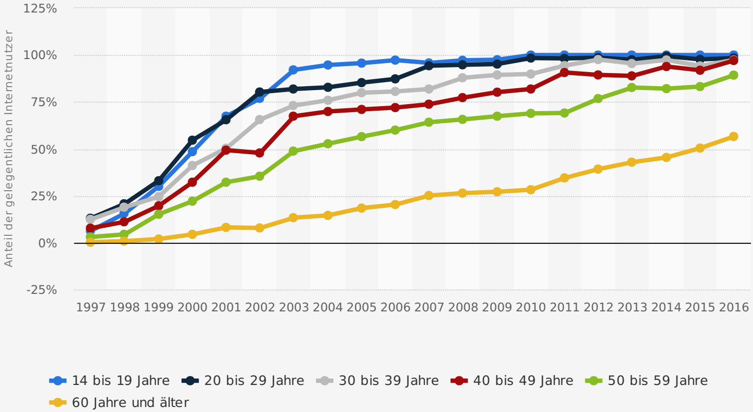


Quelle:
ARD; ZDF
© Statista 2016

Weitere Informationen:
Deutschland; GfK Media and Communication Research; n=1.508 (2016); ab 14 Jahre; Zumindest selten genutzt

statista

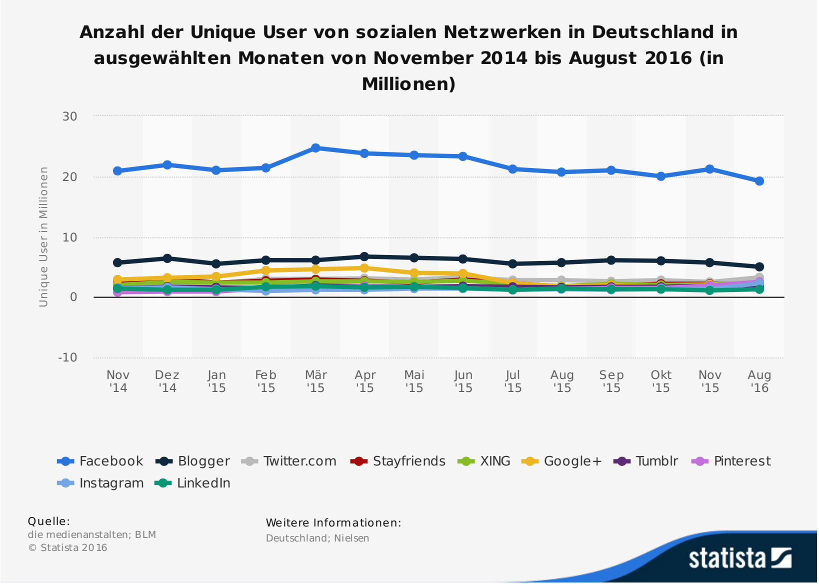
Anteil der Internetnutzer nach Altersgruppen in Deutschland in den Jahren 1997 bis 2016



Quelle:
ARD; ZDF
© Statista 2016

Weitere Informationen:
Deutschland; GfK Media and Communication Research; n=1.508 (2016); ab 14 Jahre; Zumindest selten genutzt

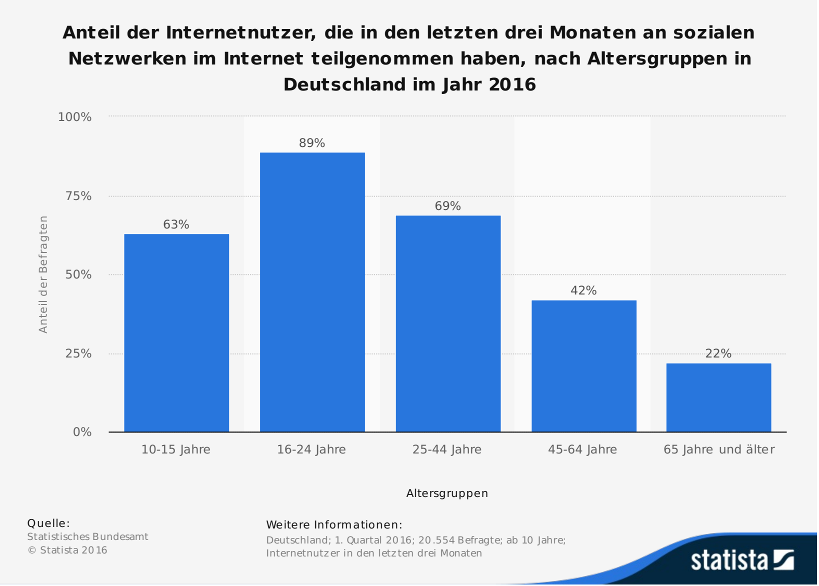
statista



Anzahl der Unique User von sozialen Netzwerken in Deutschland in ausgewählten Monaten von November 2014 bis August 2016 (in Millionen)

Quelle:
die medienanstalten; BLM
© Statista 2016

Weitere Informationen:
Deutschland; Nielsen

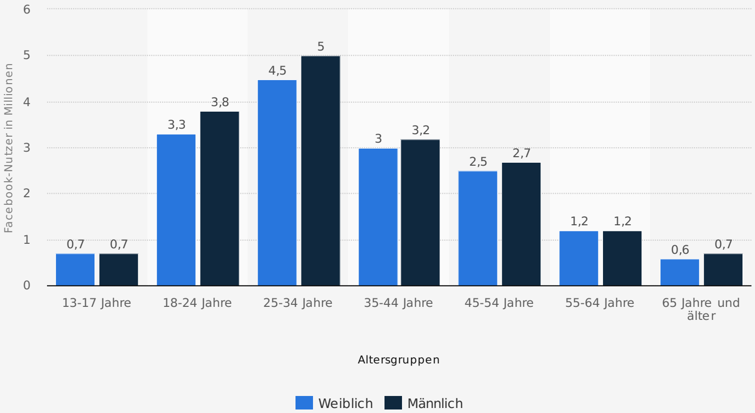


Anteil der Internetnutzer, die in den letzten drei Monaten an sozialen Netzwerken im Internet teilgenommen haben, nach Altersgruppen in Deutschland im Jahr 2016

Quelle:
Statistisches Bundesamt
© Statista 2016

Weitere Informationen:
Deutschland; 1. Quartal 2016; 20.554 Befragte; ab 10 Jahre; Internetnutzer in den letzten drei Monaten

Anzahl der Facebook-Nutzer nach Altersgruppen und Geschlecht in Deutschland im Januar 2017 (in Millionen)

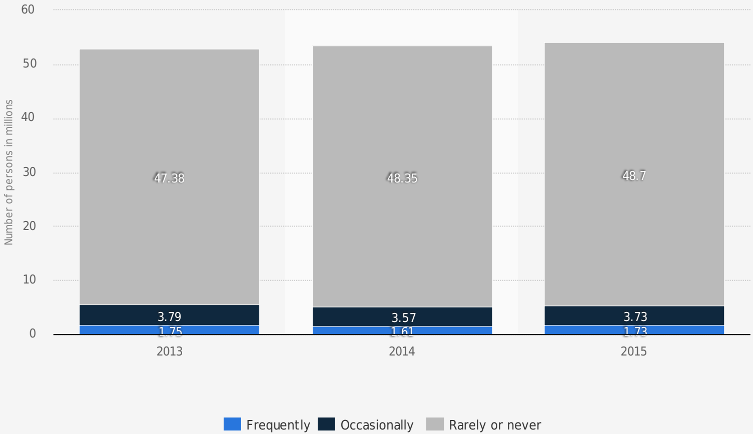


Quelle:
We Are Social; Hootsuite
© Statista 2017

Weitere Informationen:
Deutschland; We Are Social; Facebook; Hootsuite

statista

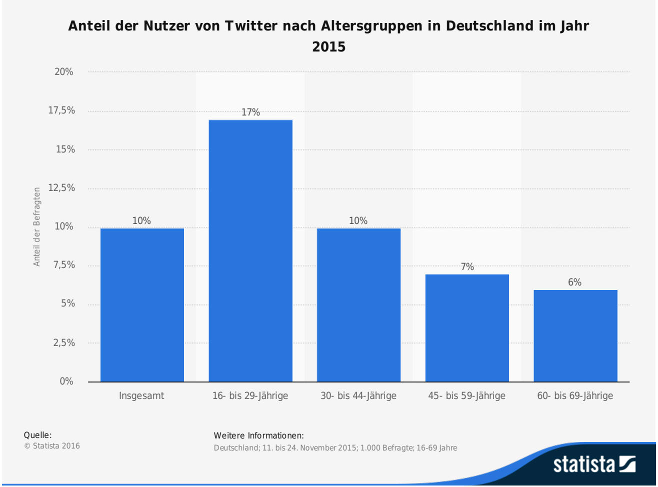
Number of internet users who read posts on Twitter in Germany from 2013 to 2015, by frequency (in millions)



Source:
IfD Allensbach (ACTA 2015)
© Statista 2015

Additional Information:
Germany; 14 years and older; German-speaking internet users

statista

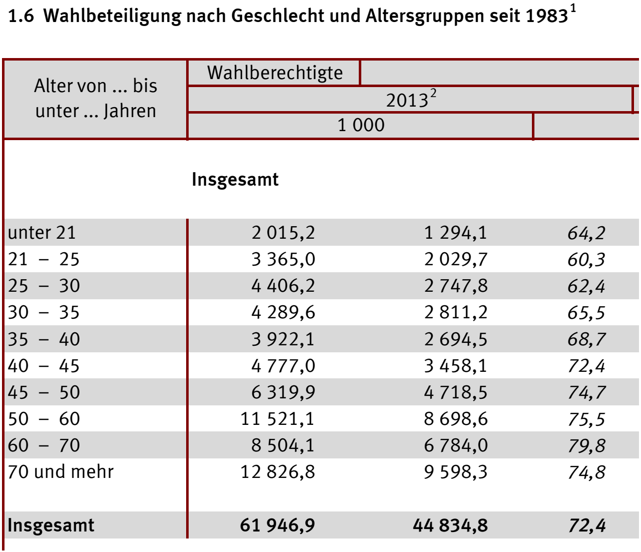


Anteil der Nutzer von Twitter nach Altersgruppen in Deutschland im Jahr 2015

Quelle:
© Statista 2016

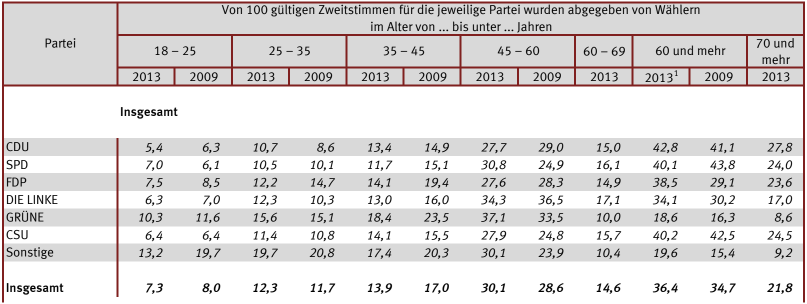
Weitere Informationen:
Deutschland; 11. bis 24. November 2015; 1.000 Befragte; 16-69 Jahre

Why is this important?



1.6 Wahlbeteiligung nach Geschlecht und Altersgruppen seit 1983[1]

Alter von ... bis unter ... Jahren	Wahlberechtigte		
	2013[2]		
	1 000		
	Insgesamt		
unter 21	2 015,2	1 294,1	*64,2*
21 – 25	3 365,0	2 029,7	*60,3*
25 – 30	4 406,2	2 747,8	*62,4*
30 – 35	4 289,6	2 811,2	*65,5*
35 – 40	3 922,1	2 694,5	*68,7*
40 – 45	4 777,0	3 458,1	*72,4*
45 – 50	6 319,9	4 718,5	*74,7*
50 – 60	11 521,1	8 698,6	*75,5*
60 – 70	8 504,1	6 784,0	*79,8*
70 und mehr	12 826,8	9 598,3	*74,8*
Insgesamt	**61 946,9**	**44 834,8**	***72,4***



Partei	Von 100 gültigen Zweitstimmen für die jeweilige Partei wurden abgegeben von Wählern im Alter von ... bis unter ... Jahren											
	18 – 25		25 – 35		35 – 45		45 – 60		60 – 69	60 und mehr		70 und mehr
	2013	2009	2013	2009	2013	2009	2013	2009	2013	2013[1]	2009	2013
	Insgesamt											
CDU	*5,4*	*6,3*	*10,7*	*8,6*	*13,4*	*14,9*	*27,7*	*29,0*	*15,0*	*42,8*	*41,1*	*27,8*
SPD	*7,0*	*6,1*	*10,5*	*10,1*	*11,7*	*15,1*	*30,8*	*24,9*	*16,1*	*40,1*	*43,8*	*24,0*
FDP	*7,5*	*8,5*	*12,2*	*14,7*	*14,1*	*19,4*	*27,6*	*28,3*	*14,9*	*38,5*	*29,1*	*23,6*
DIE LINKE	*6,3*	*7,0*	*12,3*	*10,3*	*13,0*	*16,0*	*34,3*	*36,5*	*17,1*	*34,1*	*30,2*	*17,0*
GRÜNE	*10,3*	*11,6*	*15,6*	*15,1*	*18,4*	*23,5*	*37,1*	*33,5*	*10,0*	*18,6*	*16,3*	*8,6*
CSU	*6,4*	*6,4*	*11,4*	*10,8*	*14,1*	*15,5*	*27,9*	*24,8*	*15,7*	*40,2*	*42,5*	*24,5*
Sonstige	*13,2*	*19,7*	*19,7*	*20,8*	*17,4*	*20,3*	*30,1*	*23,9*	*10,4*	*19,6*	*15,4*	*9,2*
Insgesamt	***7,3***	***8,0***	***12,3***	***11,7***	***13,9***	***17,0***	***30,1***	***28,6***	***14,6***	***36,4***	***34,7***	***21,8***

Gathering data from the web with R

Gathering data from the web with R

See R script!

Applications

Tumasjan et al. 2010

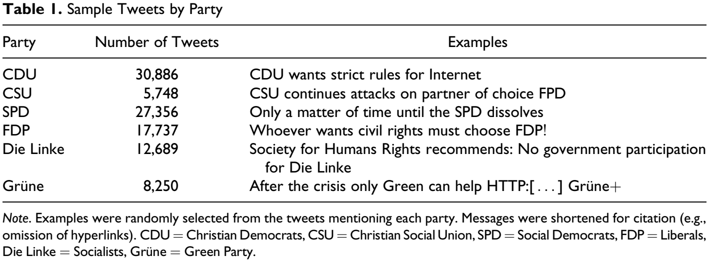


Table 1. Sample Tweets by Party

Party	Number of Tweets	Examples
CDU	30,886	CDU wants strict rules for Internet
CSU	5,748	CSU continues attacks on partner of choice FPD
SPD	27,356	Only a matter of time until the SPD dissolves
FDP	17,737	Whoever wants civil rights must choose FDP!
Die Linke	12,689	Society for Humans Rights recommends: No government participation for Die Linke
Grüne	8,250	After the crisis only Green can help HTTP:[. . .] Grüne+

Note. Examples were randomly selected from the tweets mentioning each party. Messages were shortened for citation (e.g., omission of hyperlinks). CDU = Christian Democrats, CSU = Christian Social Union, SPD = Social Democrats, FDP = Liberals, Die Linke = Socialists, Grüne = Green Party.

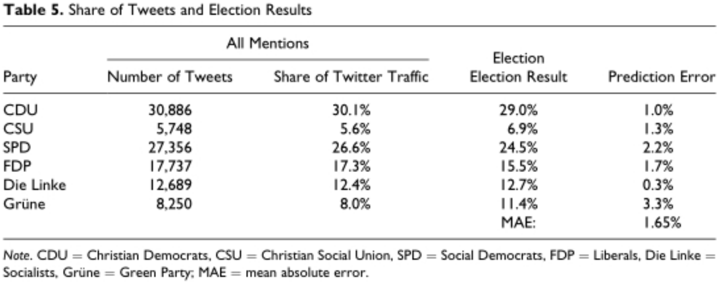


Table 5. Share of Tweets and Election Results

	All Mentions		Election	
Party	Number of Tweets	Share of Twitter Traffic	Election Result	Prediction Error
CDU	30,886	30.1%	29.0%	1.0%
CSU	5,748	5.6%	6.9%	1.3%
SPD	27,356	26.6%	24.5%	2.2%
FDP	17,737	17.3%	15.5%	1.7%
Die Linke	12,689	12.4%	12.7%	0.3%
Grüne	8,250	8.0%	11.4%	3.3%
			MAE:	1.65%

Note. CDU = Christian Democrats, CSU = Christian Social Union, SPD = Social Democrats, FDP = Liberals, Die Linke = Socialists, Grüne = Green Party; MAE = mean absolute error.

Response: Jungherr et al. 2012

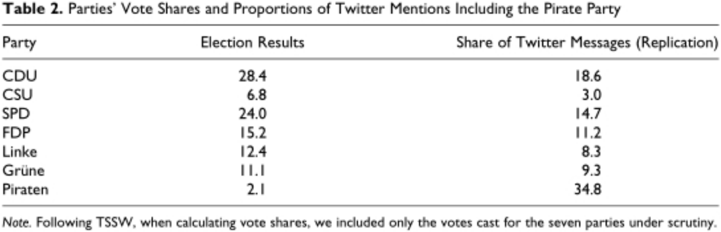


Table 2. Parties' Vote Shares and Proportions of Twitter Mentions Including the Pirate Party

Party	Election Results	Share of Twitter Messages (Replication)
CDU	28.4	18.6
CSU	6.8	3.0
SPD	24.0	14.7
FDP	15.2	11.2
Linke	12.4	8.3
Grüne	11.1	9.3
Piraten	2.1	34.8

Note. Following TSSW, when calculating vote shares, we included only the votes cast for the seven parties under scrutiny.

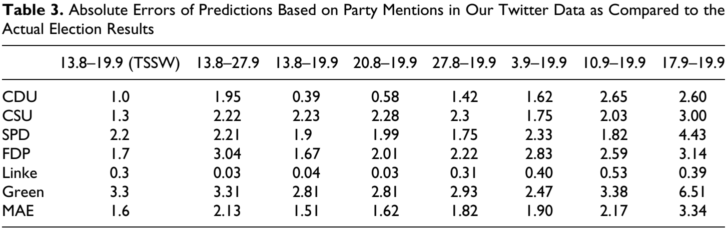
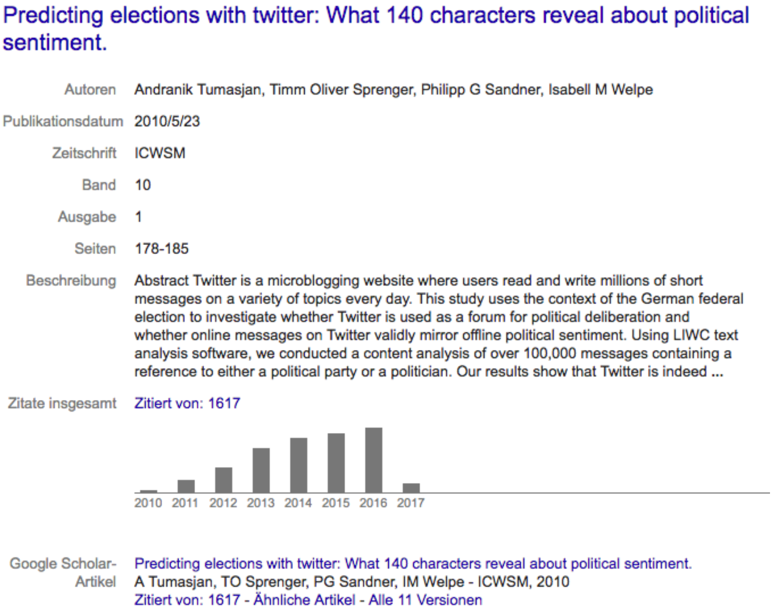


Table 3. Absolute Errors of Predictions Based on Party Mentions in Our Twitter Data as Compared to the Actual Election Results

	13.8–19.9 (TSSW)	13.8–27.9	13.8–19.9	20.8–19.9	27.8–19.9	3.9–19.9	10.9–19.9	17.9–19.9
CDU	1.0	1.95	0.39	0.58	1.42	1.62	2.65	2.60
CSU	1.3	2.22	2.23	2.28	2.3	1.75	2.03	3.00
SPD	2.2	2.21	1.9	1.99	1.75	2.33	1.82	4.43
FDP	1.7	3.04	1.67	2.01	2.22	2.83	2.59	3.14
Linke	0.3	0.03	0.04	0.03	0.31	0.40	0.53	0.39
Green	3.3	3.31	2.81	2.81	2.93	2.47	3.38	6.51
MAE	1.6	2.13	1.51	1.62	1.82	1.90	2.17	3.34

Tumasjan et al. 2010: reception

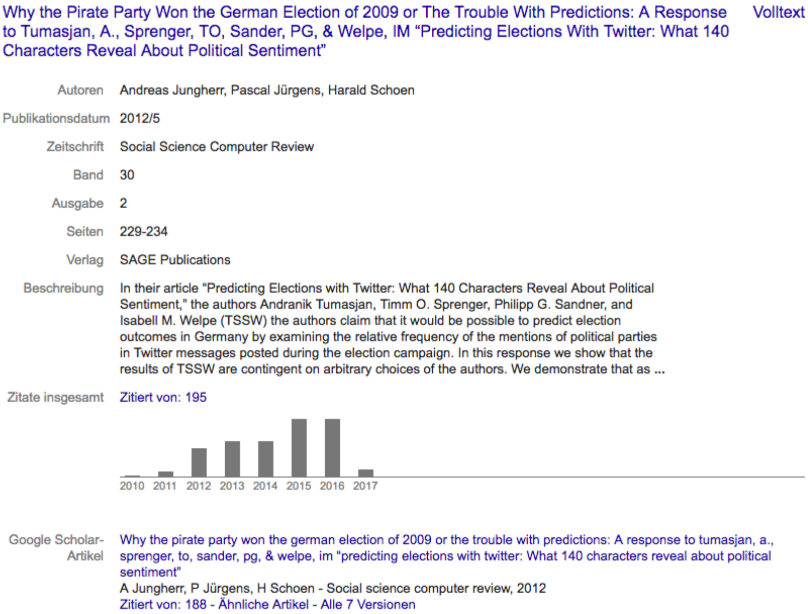


Predicting elections with twitter: What 140 characters reveal about political sentiment.

Autoren	Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, Isabell M Welpe
Publikationsdatum	2010/5/23
Zeitschrift	ICWSM
Band	10
Ausgabe	1
Seiten	178-185
Beschreibung	Abstract Twitter is a microblogging website where users read and write millions of short messages on a variety of topics every day. This study uses the context of the German federal election to investigate whether Twitter is used as a forum for political deliberation and whether online messages on Twitter validly mirror offline political sentiment. Using LIWC text analysis software, we conducted a content analysis of over 100,000 messages containing a reference to either a political party or a politician. Our results show that Twitter is indeed ...
Zitate insgesamt	Zitiert von: 1617

Google Scholar-Artikel	Predicting elections with twitter: What 140 characters reveal about political sentiment. A Tumasjan, TO Sprenger, PG Sandner, IM Welpe - ICWSM, 2010 Zitiert von: 1617 - Ähnliche Artikel - Alle 11 Versionen

Jungherr et al. 2012: reception



Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, TO, Sander, PG, & Welpe, IM “Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment” Volltext

Autoren	Andreas Jungherr, Pascal Jürgens, Harald Schoen
Publikationsdatum	2012/5
Zeitschrift	Social Science Computer Review
Band	30
Ausgabe	2
Seiten	229-234
Verlag	SAGE Publications
Beschreibung	In their article “Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment," the authors Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe (TSSW) the authors claim that it would be possible to predict election outcomes in Germany by examining the relative frequency of the mentions of political parties in Twitter messages posted during the election campaign. In this response we show that the results of TSSW are contingent on arbitrary choices of the authors. We demonstrate that as ...
Zitate insgesamt	Zitiert von: 195
Google Scholar-Artikel	Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, to, sander, pg, & welpe, im "predicting elections with twitter: What 140 characters reveal about political sentiment" A Jungherr, P Jürgens, H Schoen - Social science computer review, 2012 Zitiert von: 188 - Ähnliche Artikel - Alle 7 Versionen

Yasseri and Bright 2016

- use of Wikipedia traffic data to predict electoral outcomes
- information research hypothesis: data reflect need for information, not opinion
- underlying hypotheses:
    1. people more likely to seek information on new political parties
    2. undecided people / people considering changing their vote more likely to seek information
    3. coverage of parties in the mainstream news media correlated with voting behavior, substituting information seeking on Wikipedia

Yasseri and Bright 2016

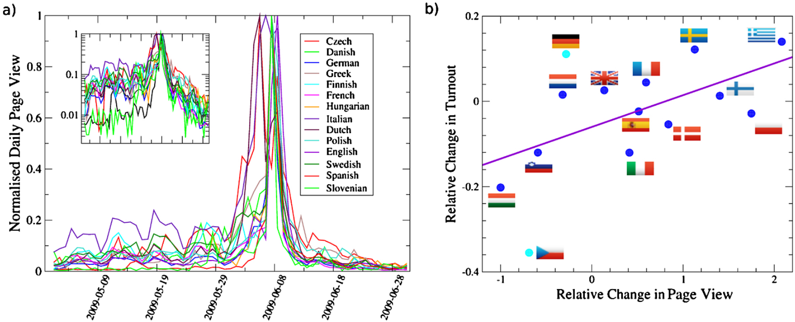


Figure 1 EU elections and Wikipedia page views. (a) Page views of the general Wikipedia article on the European Parliament elections over time and **(b)** relative change in these page views compared to change relative change in turnout. The two outliers, Germany and Czech Republic are excluded from the trend line in (b).

Yasseri and Bright 2016

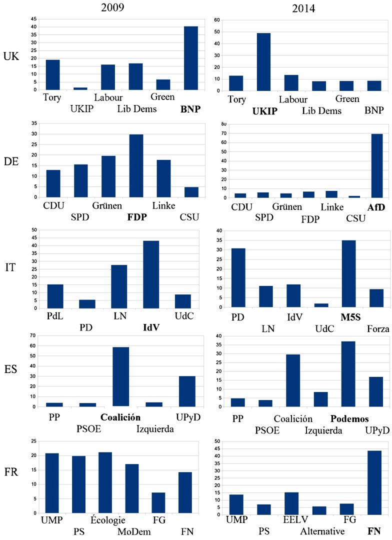


Figure 2 Wikipedia page view statistics for different parties. Relative share of Wikipedia traffic for major parties in five European countries the week before the 2009 and 2014 European parliament elections (for the party names, see the list of abbreviations below).

Yasseri and Bright 2016

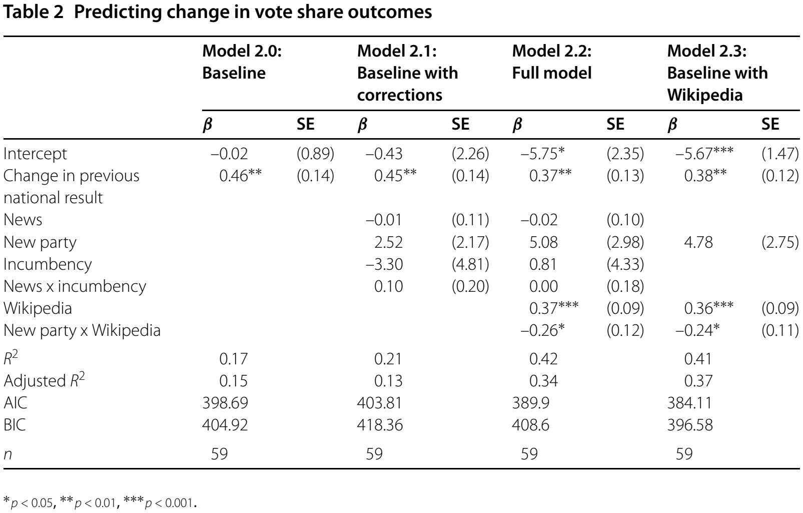


Table 2 Predicting change in vote share outcomes

	Model 2.0: Baseline		**Model 2.1: Baseline with corrections**		**Model 2.2: Full model**		**Model 2.3: Baseline with Wikipedia**	
	β	**SE**	***β***	**SE**	***β***	**SE**	***β***	**SE**
Intercept	–0.02	(0.89)	–0.43	(2.26)	–5.75*	(2.35)	–5.67***	(1.47)
Change in previous national result	0.46**	(0.14)	0.45**	(0.14)	0.37**	(0.13)	0.38**	(0.12)
News			–0.01	(0.11)	–0.02	(0.10)		
New party			2.52	(2.17)	5.08	(2.98)	4.78	(2.75)
Incumbency			–3.30	(4.81)	0.81	(4.33)		
News x incumbency			0.10	(0.20)	0.00	(0.18)		
Wikipedia					0.37***	(0.09)	0.36***	(0.09)
New party x Wikipedia					–0.26*	(0.12)	–0.24*	(0.11)
R^2	0.17		0.21		0.42		0.41	
Adjusted R^2	0.15		0.13		0.34		0.37	
AIC	398.69		403.81		389.9		384.11	
BIC	404.92		418.36		408.6		396.58	
n	59		59		59		59	

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

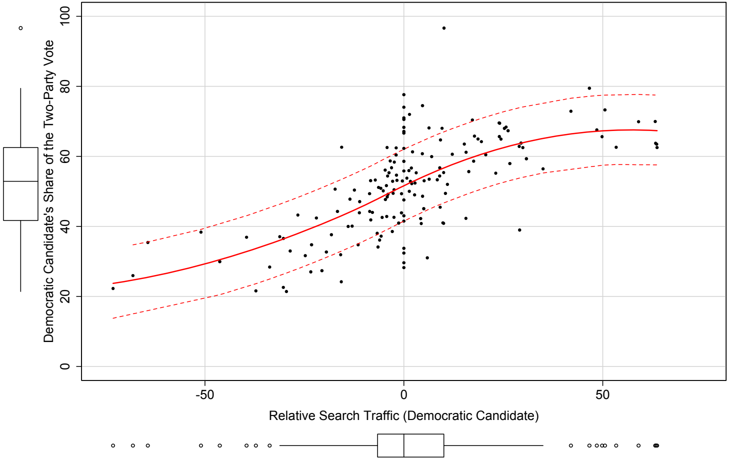
Swearingen and Ripberger 2014

1. What is the core objective of the article?
2. What data do they use to arrive at their goal?
3. How do they evaluate the criterion validity of their measure? Do you think it's a good criterion?
4. What are their findings?
5. How could these findings be used in the context of election forecasting? Do you see room for improvement of the measure?

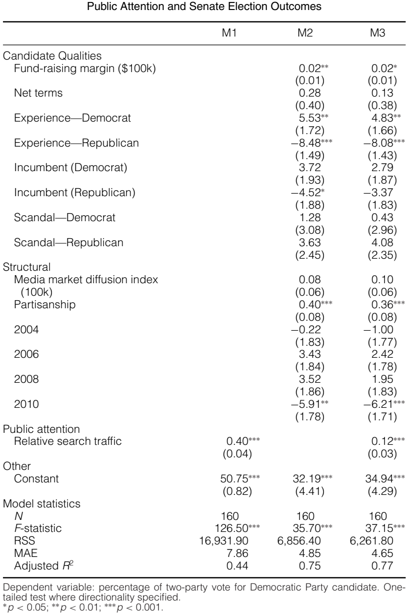
Swearingen and Ripberger 2014

FIGURE 1

Relative Search Traffic vs. Democratic Candidate's Share of the Two-Party Vote



Swearingen and Ripberger 2014



Public Attention and Senate Election Outcomes

	M1	M2	M3
Candidate Qualities			
Fund-raising margin ($100k)		0.02**	0.02*
		(0.01)	(0.01)
Net terms		0.28	0.13
		(0.40)	(0.38)
Experience—Democrat		5.53**	4.83**
		(1.72)	(1.66)
Experience—Republican		−8.48***	−8.08***
		(1.49)	(1.43)
Incumbent (Democrat)		3.72	2.79
		(1.93)	(1.87)
Incumbent (Republican)		−4.52*	−3.37
		(1.88)	(1.83)
Scandal—Democrat		1.28	0.43
		(3.08)	(2.96)
Scandal—Republican		3.63	4.08
		(2.45)	(2.35)
Structural			
Media market diffusion index		0.08	0.10
(100k)		(0.06)	(0.06)
Partisanship		0.40***	0.36***
		(0.08)	(0.08)
2004		−0.22	−1.00
		(1.83)	(1.77)
2006		3.43	2.42
		(1.84)	(1.78)
2008		3.52	1.95
		(1.86)	(1.83)
2010		−5.91**	−6.21***
		(1.78)	(1.71)
Public attention			
Relative search traffic	0.40***		0.12***
	(0.04)		(0.03)
Other			
Constant	50.75***	32.19***	34.94***
	(0.82)	(4.41)	(4.29)
Model statistics			
N	160	160	160
F-statistic	126.50***	35.70***	37.15***
RSS	16,931.90	6,856.40	6,261.80
MAE	7.86	4.85	4.65
Adjusted R^2	0.44	0.75	0.77

Dependent variable: percentage of two-party vote for Democratic Party candidate. One-tailed test where directionality specified.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

See you next week!