Predicting the Brexit Vote by Tracking and Classifying Public Opinion Using Twitter Data

Julio Amador[*], Sofia Collignon-Delmar[†] and Kenneth Benoit[‡]

January, 2017

Abstract

We use 23M Tweets related to the EU referendum in the UK to predict the Brexit vote. In particular, we use user-generated labels known as hashtags to build training sets related to the Leave/Remain campaign. Next, we train SVMs in order to classify Tweets. Finally, we compare our results to Internet and telephone polls. This approach not only allows to reduce the time of hand-coding data to create a training set, but also achieves high level of correlations with Internet polls. Our results suggest that Twitter data may be a suitable substitute for Internet polls and may be a useful complement for telephone polls. We also discuss the reach and limitations of this method.

1 Social Media and Traditional Polls

Recent events such as the Brexit referendum and the 2016 presidential election in the United States have shown that traditional polling methods face important challenges. Low response rates, low reliability of new polling channels and the time it takes to capture swings in public opinion make it difficult for traditional polling to provide timely information for campaign decision-makers. Consider, for instance, the 2016 presidential election in the United States. Right up to election day, the majority of polls gave Hillary Clinton the victory. Were most polls

[*]Imperial College London
[†]University of Strathclyde/University College London
[‡]London School of Economics and Political Science

wrong? Probably not[1]. However, it is possible that polls were not able to capture swings in public opinion due to fast-breaking events, such as of FBI director James Comey's reopening of the Clinton email investigation less than two weeks before election day (Ackerman, Jacobs, and Siddiqui, 2016). Real-time polling is expensive and rarely done (Beauchamp, 2013). Therefore, decision-makers have to rely on polls that usually reflect a 'lagged' mood in voters' preferences (Silver, 2016). As such, the question of how an electoral campaign can use real-time social media data to obtain timely information to adjust strategic behaviour is of upmost importance.

This article puts forward a simple approach for quickly processing data from the social networking site Twitter. The main premise behind our approach is that in Twitter, user-generated labels for topics, known as hashtags, can be used to train a classifier of favorability toward different outcomes, which can be aggregated to provide information predictive of the election outcome. By eliminating the time it takes to hand-code Twitter data and distributing the processing load, our approach is able to provide timely information using the most up-to-date information available, without the delay and expense of traditional polling. We demonstrate our approach with a dataset of around 23 million Tweets related to the Brexit referendum campaign in the United Kingdom. We show that our approach not only manages to classify millions of Tweets extremely rapidly but also achieves high levels of correlation with polls conducted over the Internet.

A growing body of research has used Twitter data to study or measure public opinion. Scholars have used Twitter data to analyze the way in which people discuss candidates and party leaders during elections in Germany (Tumasjan, Sprenger, Sandner, and Welpe, 2010), the US (McKelvey, DiGrazia, and Rojas, 2014), and the UK (Franch, 2013). These studies have concentrated on comparing the predictive power of Twitter data against information obtained using polls, showing that Twitter data can unveil changes in public opinion as good as opinion polls (Beauchamp, 2013, DiGrazia, McKelvey, Bollen, and Rojas, 2013, Caldarelli, Chessa, Pammolli, Pompa, Puliga, Riccaboni, and Riotta, 2014).

Nevertheless, the optimistic view that claims that Twitter can replace opinion polls is not shared by all scholars and may not be extrapolated to all contexts (Gayo-Avello, 2012, Gayo Avello, Metaxas, and Mustafaraj, 2011). Huberty (2015) concluded that social media does not offer a stable, unbiased, representa-

[1]See: Silver, N. [NateSilver538]. (2016, Nov 09). National polls will wind up being **more accurate** than they were in 2012: 2012: Obama up 1, won by 4 2014: Clinton up 3-4, will win by 1-2 [Tweet]. Retrieved from `https://twitter.com/NateSilver538/status/796411118302302208`

tive picture of the electorate and, therefore, cannot replace polling as a means of assessing the sentiment or intentions of the electorate. Sajuria and Fábrega (2016) analyzed Twitter data in the context of the Chilean 2013 elections, showing while this data could not reliably replace polls, it did provide an informative complement to more traditional methods of tracking public opnion. The same conclusions have been reached by Caldarelli et al. (2014), who showed that the volume of Tweets and their patterns across time could not precisely predict the election outcome in Italy, but that they did provide a very good proxy of the final results. Similarly, Burnap, Gibson, Sloan, Southern, and Williams (2016) incorporated sentiment analysis to show the limitations of using Twitter to forecast the results of elections in multi-party systems.

This disagreement may be due to differences in the types of polls and elections that are compared to the Twitter data; i.e., different samples may relate differently to social media data. Also, as Hopkins and King (2014) point out, there are a number of methodological challenges that emerge from classifying textual data, which may be even more severe when classifying text from social media, especially those using formats enforcing extreme brevity.

While the academic study of social media data has expanded tremendously, the state of the art remains relatively underdeveloped Beauchamp (2013). The majority of researchers use a counting measure of party or candidate mentions. As noted by (Gayo-Avello, 2012, Gayo Avello et al., 2011, Sang and Bos, 2012, Tumasjan et al., 2010), the relevance of including Tweet sentiment into the computation has been overlooked. The latter has been recently included to predict seat share (Burnap et al., 2016) and the popularity of party leaders (Franch, 2013) during the UK 2015 General Election, the popularity of Italian political leaders and candidates in the French election of 2012 (Ceron, Curini, Iacus, and Porro, 2014) and candidate success for elections to the U.S. House of Representatives (DiGrazia et al., 2013). Their success in accurately predicting elections from Twitter data has been mixed.

In what follows, we demonstrate how a large collection of Twitter data about the UK referendum to leave the European Union, known popularly as Brexit, provides an informative source for tracking vote intention. Using machine learning to classify Tweets on the Leave and Remain sides, we show how the relative balance of these classifications, across time, correlates highly with independently conducted opinion polls. In so doing, we contribute to the study of public opinion and electoral campaigning, building on previous research to show that Twitter data can be used to complement polls and provide campaigns with real-time information. Our approach is meant as a complement to more traditional polling, with the purpose of placing timely information in hands of campaign decision-makers obtained through public sources. Our approach distinguishes itself from previous efforts by putting forward the possibility of using user-generated labels and distributing the processing load to speed classification. This approach comes with several limitations concerning the nature of Twitter data and the use of Support Vector Machine (SVM) classifiers. In our conclusion, we discuss such limitations and consider the ways they may affect other cases.

2 Data

2.1 Polling Data

We use polling data from 25 different sources compiled by the poll aggregator at HuffPost Pollster. For a poll to be considered at HuffPost Pollster, it has to follow different criteria that ensure the transparency of the methodology and processing of the data. The criteria for a poll to be included can be found in the following URL: `http://elections.huffingtonpost.com/pollster/faq`. Table 1 presents the pollsters considered.

Polls included were carried out mainly through the Internet (50) and Live Phone (25) and between two populations: Likely Voters (49) and Adults (27). Moreover, polls were dated between April, 1st 2016 and June 22nd, 2016.[2]

2.2 Twitter Data

Twitter provides a continuous stream of public information by allowing its users to broadcast short messages known as "Tweets". Users can "follow" others to receive their messages, forward (or "reTweet", also know as RT) Tweets to their own followers, or mention other users in their Tweets. Tweets may also contain spontaneously created keywords known as "hashtags", that function as hyperlinks to view other Tweets containing the same hashtags. Prefixed with "#", hashtags are used to create and follow discussions or for signalling messages, such as `#strongerin`.

Over a six month period prior to the referendum, we collected publicly available Tweets based on search terms, hash tags, and user names clearly related to Brexit, listed in Table 2.

[2] A complete description of the data can be found in the following URL: `http://elections.huffingtonpost.com/pollster/uk-european-union-referendum`.

Pollsters	
Polling Agency	Number of Polls
ORB - The Daily Telegraph	9
YouGov - The Times	9
ICM	9
ICM - The Guardian	6
Opinium - Observer	6
Ipsos MORI - Evening Standard	4
SurveyMonkey	4
Survation - IG Group	4
TNS	3
ComRes - Daily Mail - ITV News	3
ComRes - Sun	2
ORB - The Independent	2
BMG Research - Herald	2
TNS BMRB	2
YouGov	2
YouGov - GMB	2
YouGov - ITV News	2
YouGov - The Sunday Times	2
Opinium	1
Survation - Mail on Sunday	1
BMG Research	1
Populus - Financial Times	1

Table 1: Pollsters.

Searched terms	
Hashtags	Usernames
#betterdealforbritain	@vote_leave
#betteroffout	@brexitwatch
#brexit	@eureferendum
#euref	@ukandeu
#eureferendum	@notoeu
#eusummit	@leavehq
#getoutnow	@ukineu
#leaveeu	@leaveeuofficial
#no2eu	@ukleave_eu
#notoeu	@strongerin
#strongerin	@yesforeurope
#ukineu	@grassroots_out
#voteleave	@stronger_in
#wewantout	
#yestoeu	
#brevet	

Table 2: Hashtags and usernames used to collect Tweets related to Brexit.

The sample of Tweets consisted on more than 30 million Tweets. However, focus was centered on 23,876,470 Tweets in English published by 35,03,769 users that emerged during the time window. The data contains information such as user ID, date and time the user account was created, the screen name or alias of the user, the number of the user's followers, time when the Tweet was posted, the text of the Tweet, language, the device that was used to post the Tweet, and a user-defined location.

3 Classifying Leave v. Remain Tweets

We use a distributed SVM classifier to categorise around 23 million Brexit-related Tweets. To perform the categorization, we first coded the variables to build a training set to compute the parameters of the SVM and classify the data. Given the size of the data and our aim to speed the process of categorization, we distributed the load across five processing units (servers). The following sub-sections describe each of the steps taken to perform the categorization.

3.1 Preparing the data and selecting relevant features

To analyze the Tweets statistically, we represent their textual content as numerical values. Specifically, we preprocess the text within each Tweet by converting it to lowercase, removing all punctuation and stop-words. To reduce the complexity of the text, we kept only words that appeared at least 10 times in the corpus. We summarize the preprocessed text as a dichotomous variable representing the presence or absence of a single words in every Tweet. This process resulted in 3,274 unique uni-gram terms.

3.2 Training the classifier

On Twitter, users organize themselves around topic-specific interests using hashtags. We made use of Tweets containing hashtags indicating support for Leave/Remain to build a training set. Specifically, we calculated the frequency in which a given hashtag occurred (see Figure 3.2). We found that the ones that indicated the most support for Leave or Remain were `#VoteLeave` and `#VoteRemain` respectively. To make sure that the appearance of those hashtags indicated support for its campaign, we label a Tweet as indicating support for the Leave campaign if it contained hashtags `#VoteLeave` and `#TakeControl`. Moreover, we label a Tweet as indicating support for the Remain campaign if it contained hashtags `#StrongerIn` and any of the hashtags `#Remain`, `#VoteRemain`, `#LabourInForBritain`, or `#Intogether`. This approach produces two training sets of 116,866 Tweets. The first training set contains 99,719 Tweets labeled as *Leave* and 17,147 labeled as *Not Leave*. The second training set contains 17,147 Tweets labeled as supporting *Remain* and 99,719 Tweets as *Not Remain*.[3] With these training sets, we calculated the coefficients of two models using Support Vector Machines. One model used the training set related to Leave/Not Leave and other using the training set related to Remain/Not Remain. The SVMs were fitted with a sigmoid kernel function. We distributed this fitting across five independent servers ni parallel. Training took an average of 50:07 minutes.[4]

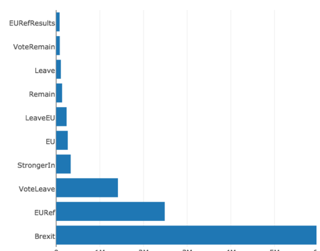


Figure 1: Top Hashtags

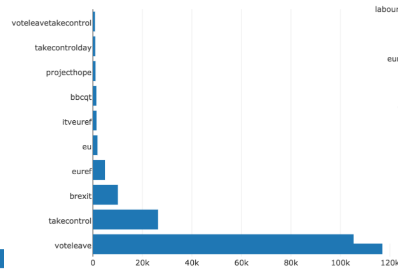


Figure 2: Leave Hashtags

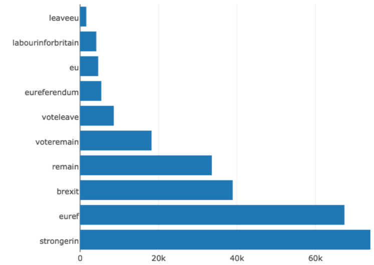


Figure 3: Remain Hashtags

3.3 Precision and recall

Precision and recall for both the Remain and Leave classifiers were calculated over each of the training sets. Doing so may raise questions about overfitting. However, the way in which features were chosen and the size of the dataset should allow sufficient generalization to alleviate this concern. Tables 3.3 and 3.3 present the confusion matrix for both SVM classifiers. Precision and recall for Remain

[3]Even if the number of labels within each training set is unbalanced, this approach allows us, in theory, to have mutually exclusive categories. See Discussion for other benefits and limitations on this approach.

[4]The timing was calculated by taking the average of the time it takes to train the SVM model with this exact training set on five different servers twice.

SVM classifier are 95% and 99% respectively. Precision and recall for the Leave SVM classifier are 99.91% and 100% respectively.

		Predicted	
		Remain	Not Remain
Actual	Remain	16362	785
	Not Remain	64	99655

Table 3: Confusion matrix for Remain/Not Remain

		Predicted	
		Leave	Not Leave
Actual	Leave	99636	83
	Not Leave	0	17147

Table 4: Confusion matrix for Leave/Not Leave

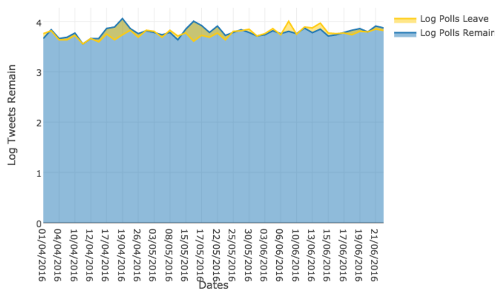
3.4 Classifying the remaining Tweets

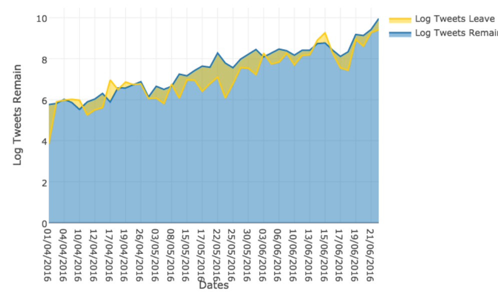
We divided the data into four batches containing 5m Tweets and one batch containing 3,876,470 Tweets. Each batch was assigned to one server for classification. This process produced two scores for every Tweet: one indicating the probability of supporting the Leave campaign and another indicating the probability of supporting the Remain campaign. We say a Tweet supports the Leave/Remain campaign if it scored at least 70% probability for a respective side. This produced 310,932 Tweets supporting the Remain campaign and 182,533 Tweets supporting the Leave campaign. In order to test for the speed of the classification, we took a random sample of 1,000 Tweets and measured the time it took a SVM to classify it. We repeated this experiment 1,000 times and took the mean. On average, it takes our classifiers 27.07 seconds to classify 1,000 Tweets. Given that we distributed the load across 5 different servers, we were able to classify the whole sample in under 35 hours.

4 Results

4.1 Comparing Relative Twitter Predictions to Polling

Given that our interest is centered around using Twitter data when polls are not available, we begin by presenting two time series to gauge how well our classification portraits support for each campaign. Figure 4.1 shows the natural logarithm of the average support for the Leave/Remain campaigns as reported by the polls. Figure 4.1 shows the natural logarithm of the number of Tweets supporting the Leave/Remain campaigns as classified by our approach.





Notice that, despite the difference in scales, both graphics depict similar support trends for each of the campaigns. To explore these patterns more formally, we present the correlation between support for the Leave/Remain campaign as reported by a moving average with five lags of the average of polls available in a given date for every date in which there is polling data, to the moving average with five lags of number of Tweets classified as supporting the Leave/Remain campaign for the same dates. The number of lags was chosen having in mind the lag polls take to capture a trend (Silver, 2016). We calculated the correlations conditioning on polling method and sampled to investigate possible differences. We also present p-values for the correlations calculated. In this case, p-values show the probability of an uncorrelated system producing datasets with correlation at least as extreme as the ones presented. Tables 5 and 6 present our results.

4.2 Internet polls vs. telephone polls

Given the high level of correlation between Internet polls and Twitter data and the low levels of correlation between telephone polls and Twitter data, we present in

Leave		
	Tweets	p-value
Internet Polls	0.71998	0.00000
Telephone Polls	0.59863	0.01112

Remain		
	Tweets	p-value
Internet Polls	0.65858	0.00005
Telephone Polls	-0.81171	0.00005

Table 5: Correlation coefficients between support percentage to *Leave/Remain in the EU* reported by a moving average of polls carried through the internet/telephone and a moving average of the number of Tweets classified as supporting *Leave/Remain in the EU*. p-values show the probability of an uncorrelated system producing datasets with correlation at least as extreme as the ones presented.

Leave		
	Tweets	p-value
Polls - Adults	0.85225	0.00000
Polls - Likely Voters	0.59024	0.00150

Remain		
	Tweets	p-value
Polls - Adults	0.64509	0.00286
Polls - Likely Voters	-0.07381	0.72010

Table 6: Correlation coefficients between support percentage to *Leave/Remain in the EU* reported by a moving average of polls conducted to adults/likely voters and a moving average of the number of Tweets classified as supporting *Leave/Remain in the EU*. p-values show the probability of an uncorrelated system producing datasets with correlation at least as extreme as the ones presented.

Table 4.2 correlations between Internet and telephone polls as a way to interpret the Twitter trends using a benchmark. The p-values show the probability of an uncorrelated system producing datasets with correlation at least as extreme as the ones presented.

Leave		
	Internet polls	p-value
Telephone polls	0.23926	0.41002
Remain		
	Internet polls	p-value
Telephone polls	0.06813	0.81696

Table 7: Correlation coefficients between support percentage to *Leave/Remain in the EU* reported Internet and telephone polls. p-values show the probability of an uncorrelated system producing datasets with correlation at least as extreme as the ones presented.

Tables 4.2 underscore the low correlation between Internet polls and telephone polls. However, these correlations should be taken with a grain of salt for different reasons. First, telephone polls are less frequent than Internet polls. Second, the values used to calculate these correlations are only the simple average of telephone/Internet polls available on each given date. Third, p-values indicate a high probability of an uncorrelated system producing datasets with correlations as least as extreme as the one presented.

5 Discussion

Social media data is notoriously noisy, a pattern also found in our efforts to use this data to measure public changing opinion on Brexit. However, our comparisons also show that even more traditional methods of predicting vote intention, such as telephone polls, are prone to error as well. Our comparisons of polls from Internet data to telephone polls showed a low overall correlation, while correlations between Twitter data and Internet polls were larger than those between Twitter data and Live-Phone polls. Moreover, correlations between Twitter data and polls in which the sampled population was likely voters were smaller than correlations between Twitter data and polls in which the sampled population was adults in general.

Second, the use of hashtags to label Tweets as Leave/Remain implied that the training set was not built out of a random sample. This implies that, most likely, the estimates of the SVM classifiers are biased towards Leave. However, even if the number of Tweets in the training set related to Leave surpassed those of Remain by almost six times, the final classification resulted in 310,932 Tweets related to Remain and only 182,533 related to Leave. Once we limit our data to Tweets gathered before the vote, the number of Tweets related to Remain are 201,078 and those related to Leave number 150,145. Furthermore, our classification appears to be highly correlated to what Internet polls reported.

Third, limitations at the time of coding the variables for the training set imply that the latter does not include all possible information that could be added to accurately predict category Leave/Remain. However, it is important to notice that by design, Twitter limits its users to 140 characters, therefore minimizing the number of words to be included in the training set. Most importantly, the fact that the Leave campaign had a very coordinate set of points they were pushing forward, such as *taking back control of the borders* or *the NHS* greatly helped our ability to correctly classify Tweets as supporting the campaign. This was not the case for the Remain campaign where the points the campaign was pushing forward appear not to be as clear.[5] In fact, the correlations presented above show that, in most cases, correlations related to the Leave campaign are larger than those for the Remain campaign.

Fourth, the process of smoothing trends through moving averages contributed to reduce inter-day biases and fluctuations.

Finally, the effectiveness of our approach may have been affected by the particularities of participation in Twitter for the Brexit referendum. Research looking to reproduce this approach in different contexts should take into account the following considerations. First, there is high level of Twitter users in Britain, some 20% by recent estimates (eMarketer., 2016). The high correlations between Internet polls and Twitter data are most likely due to the relatively high adult level of political participation in social networks in Britain. Applying this approach in countries where the level of participation is lower may lead to different results. Second, the Leave campaign was able to organize their supporters around specific hashtags and topics. Such hashtags allowed us to build a training set without hand-coding the Tweets. Moreover, it is possible that such organization alleviated some of the problems of using SVM classifiers with textual data. This approach may not be as useful in situations in where topics are intrinsically ill-defined. We

[5] A simple count of the hashtags supporting the Leave/Remain campaign supports this point.

believe that individuals looking to replicate these process should bear in mind the methodological limitations discussed above at the time of decision-making.

6 Conclusions

Scholars of public opinion and political behavior have long agreed that information plays an important role in motivating political participation and defining strategic voting (Huckfeldt and Sprague, 1995, Campbell, Converse, Miller, and Donald, 1960, Verba, Schlozman, Brady, and Brady, 1995, Huckfeldt, Carmines, Mondak, and Zeemering, 2007, Settle, Bond, Coviello, Fariss, Fowler, and Jones, 2016). While voters seek information about political affairs, campaign managers consume information *about voters* (Hersh, 2015). The success or failure of these strategies is reflected in changes of public opinion measured using polls and, of course, monitoring social media (Sajuria and Fábrega, 2016). However, recent events such as the Brexit referendums and the latest presidential election in the United States have shown that traditional polling methods face important challenges that derive from low response rates, low reliability of new channels of polling and the time it takes them to capture swings in public opinion. In particular, the time-lag between influential events and results reflect in traditional polls mean that electoral campaigns cannot react quickly to shocks in public opinion. This problem can be addressed by complementing polls with social media data (Settle et al., 2016, DiGrazia et al., 2013, Settle et al., 2016). Our study suggests that Twitter data can provide a valuable source of information for campaign decision-making, as a continuous flow of public information directly posted by individuals who express and share their opinions about politics with a wider network that is broader necessarily than just friends and family (Tumasjan et al., 2010, Fábrega and Sajuria, 2014).

This article built upon previous studies that have used social media data to measure public opinion. We showed that our method of analysis can be used to provide timely information to campaign decision-makers by examining swings in public opinion through Twitter. With the use of hashtags as labels for more than 100,000 Tweets sent during the EU referendum campaign in the UK, we reduced the time required for hand-coding. Moreover, by distributing the computing load across five servers, we were able to train an SVM classifier in less than an hour and classify thousand of Tweets in minutes. Most importantly, by taking moving averages of the time series, we were able to achieve a 71% correlation between our classified data and Internet Polls for those supporting Leave and 65% correlation for those supporting Remain.

It is important to notice there is a low level of correlation between Internet and telephone polls and, conversely, Twitter data and telephone polls. As noted in the section above, the low level of correlation between Internet and telephone polls should be taken with caution. However, these correlations underscore deep differences between polling channels. While we believe such differences deserve more robust exploration, our findings suggest the possibility that Twitter data may be more suited to be a substitute for Internet polling and a complement for telephone polling.

Finally, even if quickly classifying data may not be important for the academic researcher, it is of the highest importance of campaign decision-makers. In our dataset, more than 15 million Tweets were generated in the week before the EU referendum alone. This large amount of information highlights the importance of developing reliable methods to make use this information as a a means of measuring public opinion, and of having methods for doing so that work for such information in massive quantitites. Future research should focus on the conditions under which Twitter data can be a substitute of polling, and when it can be used as a complement. Another future avenue of research will explore the pertinence of using social media data in different type of elections, such as regional elections, as they may present distinctive patterns of political engagement.

References

Ackerman, S., B. Jacobs, and S. Siddiqui (2016): "Newly discovered emails relating to Hillary Clinton case under review by FBI," `https://www.theguardian.com/us-news/2016/oct/28/fbi-reopens-hillary-clinton-emails-investigation`, retrieved on: 2017-01-06.

Beauchamp, N. (2013): "Predicting and interpolating state-level polling using Twitter textual data," in *New directions in analyzing text as data workshop*.

Burnap, P., R. Gibson, L. Sloan, R. Southern, and M. Williams (2016): "140 characters to victory?: Using Twitter to predict the UK 2015 General Election," *Electoral Studies*, 41, 230–233.

Caldarelli, G., A. Chessa, F. Pammolli, G. Pompa, M. Puliga, M. Riccaboni, and

G. Riotta (2014): "A multi-level geographical study of Italian political elections from Twitter data," *PloS one*, 9, e95809.

Campbell, A., P. E. Converse, W. E. Miller, and E. Donald (1960): "Stokes. the American voter," *New Yrok: John Wiley and Sons*, 77.

Ceron, A., L. Curini, S. M. Iacus, and G. Porro (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.

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.

eMarketer. (2016): "Twitter, facebook user growth slowing in the uk," `https://www.emarketer.com/Article/Twitter-Facebook-User-Growth-Slowing-UK/1014326`, retrieved on: 2017-01-31.

Fábrega, J. and J. Sajuria (2014): "The formation of political discourse within online networks: the case of the occupy movement," *International Journal of Organisational Design and Engineering*, 3, 210–222.

Franch, F. (2013): "Wisdom of the Crowds: 2010 UK election prediction with social media," *Journal of Information Technology & Politics*, 10, 57–71.

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

Gayo Avello, D., P. T. Metaxas, and E. Mustafaraj (2011): "Limits of electoral predictions using Twitter," in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, Association for the Advancement of Artificial Intelligence.

Hersh, E. D. (2015): *Hacking the electorate: How campaigns perceive voters*, Cambridge University Press.

Hopkins, D. and G. King (2014): "A method of automated nonparametric content analysis for social science," *American Journal of Political Science*, 54, 229–247.

Huberty, M. (2015): “Can we vote with our tweet? On the perennial difficulty of election forecasting with social media,” *International Journal of Forecasting*, 31, 992–1007.

Huckfeldt, R., E. G. Carmines, J. J. Mondak, and E. Zeemering (2007): “Information, activation, and electoral competition in the 2002 congressional elections,” *Journal of Politics*, 69, 798–812.

Huckfeldt, R. R. and J. Sprague (1995): *Citizens, politics and social communication: Information and influence in an election campaign*, Cambridge University Press.

McKelvey, K., J. DiGrazia, and F. Rojas (2014): “Twitter publics: How online political communities signaled electoral outcomes in the 2010 US House election,” *Information, Communication & Society*, 17, 436–450.

Sajuria, J. and J. Fábrega (2016): “Do we need polls? why Twitter will not replace opinion surveys, but can complement them,” in *Digital Methods for Social Science*, Springer, 87–104.

Sang, E. T. K. and J. Bos (2012): “Predicting the 2011 Dutch senate election results with Twitter,” in *Proceedings of the Workshop on Semantic Analysis in Social Media*, Association for Computational Linguistics, 53–60.

Settle, J. E., R. M. Bond, L. Coviello, C. J. Fariss, J. H. Fowler, and J. J. Jones (2016): “From posting to voting: The effects of political competition on online political engagement,” *Political Science Research and Methods*, 4, 361–378.

Silver, N. (2016): “The myth of the lag,” `http://fivethirtyeight.com/features/myth-of-lag/`, retrieved on: 2017-01-06.

Tumasjan, A., T. O. Sprenger, P. G. Sandner, and I. M. Welpe (2010): “Predicting elections with Twitter: What 140 characters reveal about political sentiment.” *ICWSM*, 10, 178–185.

Verba, S., K. L. Schlozman, H. E. Brady, and H. E. Brady (1995): *Voice and equality: Civic voluntarism in American politics*, volume 4, Cambridge Univ Press.