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

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

4 September 2017

Abstract

We use 23 million 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. In the current dataset, this approach achieves high level of correlations with Internet polls while reducing the time of hand-coding data to create a training set. Our results suggest that Twitter data has a potential to be a suitable substitute for Internet polls and be a useful complement for telephone polls. We also discuss the reach and limitations of this method.

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

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 wrong? Probably not (Silver, 2016b). However, it is possible that polls were not able to capture swings in public opinion due to fast-breaking events, such as 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, 2017). Therefore, decision-makers have to rely on polls that usually reflect a "lagged" mood in voters' preferences (Silver, 2016a). 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 to tracking public opinion through fast processing of 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[1] 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 by 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,

[1] A hashtag is an alpha-numeric string with a prepended hash ("#") in social media texts. Users use this hashtag to emphasize the topic of a Tweet.

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 well as opinion polls (Beauchamp, 2017, DiGrazia, McKelvey, Bollen, and Rojas, 2013, Caldarelli, Chessa, Pammolli, Pompa, Puliga, Riccaboni, and Riotta, 2014). This also makes it possible to use Twitter data to predict electoral outcomes because citizens' opinion made public via Twitter, correlates with their voting history (Barberá, 2014).

Nevertheless, some scholars, who do not share the the optimistic view that Twitter can replace opinion polls, argue that there are only limited contexts where Tweets can be used as a substitute for opinion polls (Gayo-Avello, 2012, Gayo Avello, Metaxas, and Mustafaraj, 2011), and that Twitter data replicates biases observed in other forms of political exchanges (Barberá and Rivero, 2015). Huberty (2015) concluded that social media does not offer a stable, unbiased, representative picture of the electorate and, therefore, cannot replace polling as a means of assessing the sentiment or intentions of the electorate. One reason offered is that social bots can massively influence Twitter trends during campaigns (Howard and Kollanyi, 2016). Sajuria and Fábrega (2016), who analyzed Twitter data in the context of the Chilean 2013 elections, showed that while this data could not reliably replace polls, it did provide an informative complement to more traditional methods of tracking public opinion. 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) illustrated the limitations of using Twitter to forecast the results of elections in multi-party systems by showing that Twitter is useful only when estimates are adjusted with previous party support and sentiment analysis.

While the academic study of social media data for political research has expanded tremendously, the state of the art remains relatively underdeveloped (Beauchamp, 2017). There are still a number of methodological challenges that emerge from classifying textual data, which may be even more severe when classifying text from social media which enforces the extreme brevity, and the extreme sparsity of the document-term matrix would result in the underperformance of classification of individual documents (Hopkins and King, 2010), although the brevity may not always be a disadvantageous for sentiment classification (Bermingham and Smeaton, 2010).

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) and 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, the popularity of party leaders during the UK 2015 General Election (Burnap et al., 2016) and public views of the top election candidates in the USA (Franch, 2013, Chin, Zappone, and Zhao, 2016), 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. Once potential cause of this disagreement on the reliability and accuracy of Twitter to measure public opinion is the differences in the types of polls and elections that are compared to the Twitter data. In other words, different samples may relate differently to social media data.

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 and apply this to the context of the EU referendum in the UK. 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. In particular, polls considered are required to disclose the sponsorship of the survey, fieldwork provider, dates of interviewing, sampling method, population that was sampled, size of the sample, size and description of any subsample, margin of error, survey mode, complete wording and ordering of questions mentioned and percentage results of all questions reported[2]. Table 1 presents the pollsters considered.

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: Polls included in our study.

Polls included were carried out mainly through the Internet (50 in total) and

[2]A more detailed description of the criteria can be found here: `http://elections.huffingtonpost.com/pollster/faq`.

Telephone (25 in total) and between two populations: "likely voters" (49) and adults (27). Moreover, the timing of fielding these polls were spanned across the Brexit referendum campaign, dated between April, 1st 2016 and June 22nd, 2016.[3]

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 known as its abbreviation, RT) Tweets to their 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`.

In order to capture the discussion on the Brexit referendum in Twitter in its entirety, we set up Twitter downloader through an access to the Twitter "firehose," which guarantees the delivery of all Tweets that matched the capture criteria. Another option to capture Tweets is to use Twitter's Streaming API,[4] which can capture Tweets according to search terms up to one percent of all Tweets generated at a given time, a threshold above which it samples randomly. The use of the streaming API is a preferred choice of method in many studies, because it is freely accessible, although it is subject to both rate limits and to a cap of one percent of the volume of all Tweets. While our capture of Brexit-related Tweets did not approach this limit, our capture method was also not subject to these constraints.[5]

Through a careful examination of the terms from a directed search on Brexit topics, we selected our search terms for capturing messages related to the Brexit referendum. These search terms, presented in Table 2, consist of three sets. The first is the general search term "Brexit"; the second contains hashtags related to the topic; and the third consists of Twitter user screen names found in our research to be strongly associated with the Brexit debate. We started with the key hashtag

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

[4] `https://dev.twitter.com/streaming/overview`

[5] For the detailed comparison between Firehose and Streaming API, see Morstatter, Pfeffer, Liu, and Carley (2013). They argue that the sample size is the key to obtain high quality results using sampled Tweets from Streaming API, and given the large volume of Tweets in our data, Streaming API would provide the similar outcomes in this research.

#brexit and conducted a trial.[6] From the Tweets during this trial, we selected frequently used hashtags and user-mentions. While this process involved subjective researcher judgment rather than an automated procedure, the selection effects should not significantly affect our conclusions, since the primary goal is not to estimate the level but to estimate the trend. Any Tweets that contained one of these terms were captured in our data collection, which ran from January 6, 2016 through the day of the referendum.

Search Criteria	Terms
Simple words	brexit
Hashtags	#betterdealforbritain #betteroffout #brexit #euref #eureferendum #eusummit #getoutnow #leaveeu #no2eu #notoeu #strongerin #ukineu #voteleave #wewantout #yes2eu #yestoeu
User screen names	@vote_leave @brexitwatch @eureferendum @ukandeu @notoeu @leavehq @ukineu @leaveeuofficial @uk-leave_eu @strongerin @yesforeurope @grassroots_out @stronger_in

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

The sample consisted of more than 30 million Tweets. However, focus was placed on 23,876,470 Tweets in English published by 3,503,769 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 according to whether they were pro-Leave or pro-Remain. To perform the

[6] #brexit originally indicated the pro-Brexit position, but during the process of Brexit campaign, it used mostly as a term to refer the EU membership referendum. The neutrality of the term is also proven in the empirical analysis (see Section 4.3)

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. We used Python's NTLK library (Bird, Loper, and Klein, 2009) to remove English stopwords and augmented the list by including frequently repeated tokens that are not included in itḟootnoteThe list of stopwords provided by the English dictionary in NLTK is available here: `http://www.nltk.org/book/ch02.html` and the following tokens were added to that dictionary: `"...", "http", "..", "n't", "'s", "‘‘","’’", " ", "'d",",", " ,", "", "ex","https", "rt"`. Moreover, we used the NLTK tokenizer to separate tokens that, because of the nature of Twitter data, may have been written together. To reduce the complexity of the text, we kept only words that appeared more than 10 times in the corpus. Using this simple rule allowed us to reduce the number of features from 11,653 to 3,274 unique unigram terms. Also, doing so allows us to prevent overfitting of the training set. We summarize the preprocessed text with a binary weighting for each term in every Tweet.

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 Table 3). 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`.

Given the nature of Twitter data, we expected many Tweets to contain spam. In view of this, we decided to train two classifiers: one to learn features related to Remain and other to learn features related to Leave. By treating the probability of a Tweet belonging to Remain (Leave) different from the complement of the probability of belonging to Leave (Remain), we allow some Tweets not to belong to any such categories. This approach produces two labelled sets of 116,866 Tweets. The first 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*.[7] We further divided each of the sets into two: one containing 78,300 Tweets which we used to train each of the Leave and Remain models, and another containing 38,566 Tweets which we used to test each model. With the training sets, we calculated the coefficients of two models using the Support Vector Machines class within Sci-Kit Learn library for Python 2.7[8]. 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 ReLU kernel function.[9] We distributed this fitting across five independent servers in parallel. The servers used a Linux Ubuntu 14.0 distribution had 16 cores with 116 Gb in RAM. Training took an average of 50:07 minutes.[10]

3.3 Robustness of the classifiers

3.3.1 Feature selection

A relevant question to validate our approach is: are there features, other than the hashtags, keywords and users we used to select the sample, that are useful to predict a category? To investigate this, we perform a chi-square test to find which features are significant at the 5% level. We found that of the 3,274 features, 2,442 were statistically significant at the 5% level. Of these, there are interesting features that provide insight into the position of those supporting leave such as "national", "believeinbritain", "peoplesmomentum", "independenceday" and

[7]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.

[8]See: Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot, and Duchesnay (2011)

[9]Given that one of the aims was performance, we opted for choosing a ReLU kernel. By doing so, gradient descent algorithm used in SKLearn converges faster to the global optimum.

[10]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.

All		Remain		Leave	
Hashtag	*Count*	*Hashtag*	*Count*	*Hashtag*	*Count*
Brexit	8,635,203	Brexit	10,889	Brexit	110,657
Euref	2,626,167	Euref	10,834	TakeControl	99,719
VoteLeave	1,994,025	VoteLeave	8,620	Euref	33,500
brexit	1,319,070	VoteRemain	8,074	VoteRemain	19,624
EURef	698,355	Remain	6,500	InOrOut	19,421
Eureferendum	685,765	ITVEURef	4,086	LeaveEU	19,378
StrongerIn	568,611	BBCDebate	3,697	bbcqt	15,954
EU	510,146	LabourInForBritain	2,529	bbcdebate	13,060
LeaveEU	477,096	Intogether	2,154	ITVEURef	12,303
BREXIT	183,837	Eureferendum	1,946	BBCDebate	11,349

Table 3: Top-ten hashtags of the Twitter sample and by side.

"immigrant" and also features providing insight into the remain position, such as "scientists4eu", "academicsforeu", "economy", "open" and "sadiqkhan". Even if it is not the final goal of this paper to research important features to predict each category, this exercise shows that there is useful information in every Tweet, apart from hashtags, that can be used to predict support for Remain or Leave.

3.3.2 Cross-validation

To perform cross-validation of our models, we used the test sets mentioned above. Recall that our SVM models were not trained with the test data. As such, cross-validation provides a measure for how well our classifiers generalize to the overall corpus of Tweets. Cross-validation for the Leave model gives a 97.12% accuracy. Cross-validation for the Remain model gives 97.05% accuracy.

3.3.3 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. Table 4 presents the confusion matrix for both SVM classifiers. Precision and recall for Remain 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	Leave	Not Leave
Actual	Remain/Leave	16,362	785	99,636	83
	Not Remain/Not Leave	64	99,655	0	17,147

Table 4: Confusion matrix for Remain/Not Remain

3.4 Classifying the remaining Tweets

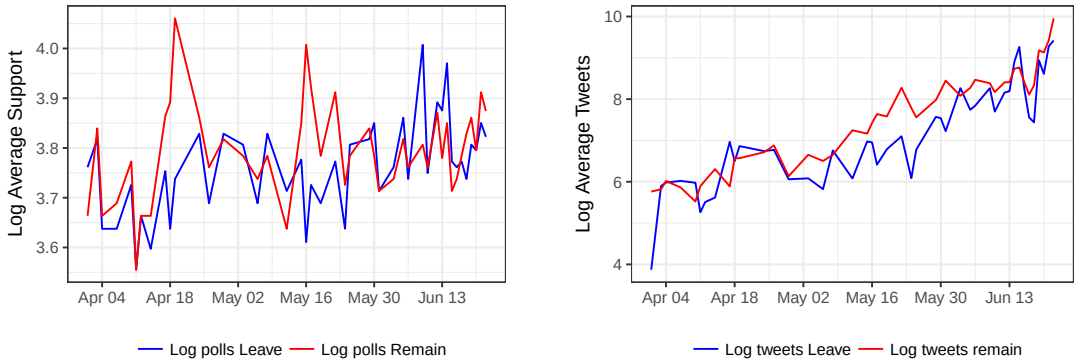
We divided the data into four batches containing 5M Tweets and one batch containing 3,876,470 Tweets. Each batch was assigned to one server for classification. This process produced two scores for every Tweet: one indicating the probability of supporting the Leave campaign and another indicating the 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. The tweets that did not pass this threshold were not assigned any category and, hence, not used for this analysis. 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 1a shows the natural logarithm of the average support for the Leave/Remain campaigns as reported by the polls.

Figure 1b shows the natural logarithm of the number of Tweets supporting the Leave/Remain campaigns as classified by our approach.



(a) Opinion Polls

(b) Twitter

Figure 1: Time-series of Opinion Polls and Tweets.

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, 2016a). 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 as in subsequent tables, the 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.

To further check that our results are not driven by the smoothing of the series, we present correlations of the number of Tweets classifies as supporting the Leave/Remain campaign to the average of polls available in a given date for every date in which there is polling data. Tables 7 and 8 present our results. Notice that, even if correlations are not as strong as with the smoothing, most of our results still hold.

	Leave		**Remain**	
	Tweets	p-value	Tweets	p-value
Internet Polls	0.719	0.00000	0.658	0.00005
Telephone Polls	0.598	0.01112	-0.811	0.00005

Table 5: Internet/telephone poll correlation with Twitter classification. Correlations are the Pearson's *r* between the support percentage to *Leave/Remain in the EU* reported by a moving average of Internet and telephone polls, and a moving average of the number of Tweets classified as supporting *Leave/Remain in the EU*.

	Leave		**Remain**	
	Tweets	p-value	Tweets	p-value
Polls - Adults	0.852	0.00000	0.645	0.00286
Polls - Likely Voters	0.590	0.00150	-0.073	0.72010

Table 6: Adults and "Likely" Voters correlation with Twitter classification. Correlations are the Pearson's *r* 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*.

	Leave		**Remain**	
	Tweets	p-value	Tweets	p-value
Internet Polls	0.448	0.00532	0.550	0.00041
Telephone Polls	0.251	0.27135	-0.178	0.43974

Table 7: Internet/telephone poll correlation with Twitter classification. Correlations are the Pearson's *r* between the support percentage to *Leave/Remain in the EU* reported by a the average of Internet and telephone polls, and the number of Tweets classified as supporting *Leave/Remain in the EU*.

	Leave		Remain	
	Tweets	p-value	Tweets	p-value
Polls - Adults	0.590	0.00301	0.614	0.0018
Polls - Likely Voters	0.214	0.3632	-0.221	0.34800

Table 8: Adults and "Likely" Voters correlation with Twitter classification. Correlations are the Pearson's *r* between support percentage to *Leave/Remain in the EU* reported by a the average of polls conducted to adults/likely voters, and the number of Tweets classified as supporting *Leave/Remain in the EU*.

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 Table 9 correlations between Internet and telephone polls as a way to interpret the Twitter trends using a benchmark.

	Correlation	p-value
Leave	0.239	0.410
Remain	0.068	0.817

Table 9: Correlation coefficients between Internet and telephone polls for support percentage to *Leave/Remain in the EU*.

Table 9 underscores 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.

4.3 Validating the classification using a different classifier

One of the characteristics using SVM classification is that separation of categories is non-linear for most kernels. The obvious advantage of non-linear classification

is that if appropriately tuned, it can exhibit a very high performance in classification. However, there are some drawbacks for using non-linear classifiers. One of these drawbacks is that it is essentially impossible to interpret the effect of features on the classification. The features in the models are words and special entities in Twitter texts and the effects of these features could be substantively interesting. In this subsection, we conduct an additional classification using a multinomial Naive Bayes classifier (Manning, Raghavan, and Schütze, 2008, Ch. 13) in order to complement the issues with SVM classification.

In this classification, we combined all Tweets made by each user, to focus on predicting whether each user was pro-Leave or pro-Remain. We first select 200 accounts which have the largest number of Tweets during the period, excluding obvious bots. For the top 200 most mentioned accounts, we verified for each user (in December 2016) whether the accounts were still active, and noted their position in the Brexit debate. We found that fifty-five accounts clearly supported Leave, and twenty-five clearly supported Remain. From this list of known accounts, we extended the list of users on each side by analyzing the hashtag usage in Tweets from the clearly classifiable accounts. We identified the hashtags disproportionally used by one of two sides by looking at the hashtag use of human-coded accounts.[11] We calculate the log-ratio of the use of Leave and Remain hashtags by users frequently Tweeted on the issue (i.e. more than 50 Tweets in our data, there are about 20,000 such accounts). We generate the training data from the training data by assigning top and bottom ten percent of the frequent Twitter users as Remain and Leave accounts. The features used for the classification are uni-grams used by more than 10 users. We estimate Naive Bayes models of two outcome categories, Leave and Remain, with uniform priors. To check the model performance, we conducted ten-fold cross validation, finding an average predictive accuracy of 0.926.[12]

Since our Naive Bayes classification is conducted at user level as opposed to Tweet level classification using SVM, we create an index from the SVM classifier for each user and compare the results from Naive Bayes classification. The index we use is the rank-order of difference between average probabilities of Leave and

[11] Hashtags by the Remain side are #strongerin, #intogether, #infor, #votein, #libdems, #voting, #incrowd, #bremain, and #greenerin. Hashtags used by the Leave side are #voteleave, #inorout, #takecontrol, #voteout, #takecontrol, #borisjohnson, #projecthope, #independenceday, #ivotedleave, #projectfear, #britain, #boris, #lexit, #go, #takebackcontrol, #labourleave, #no2eu, #betteroffout, #june23, and #democracy.

[12] For the model fitting and prediction, we use the `quanteda` package in `R` (Benoit, Watanabe, Nulty, Obeng, Wang, Lauderdale, and Lowe, 2017).

Remain for Tweets by each users. The correlation coefficient of probability of Remain and Leave from both models is 0.604. Although these two are not perfect match, but the results from different classifiers with different setups seem to yield approximate outputs.

Naive Bayes model results provide the probabilities of each *feature* to belong to each category. Based on these probabilities we categorize features into Leave, Remain and Neutral. Figure 2 is a wordcloud of hashtags which belong to Leave, Remain or Neutral. Many hashtags are in an expected category, such as `#leaveeu`, `#brexitthemovie` for Leave and `#ukineu`, `#britsdontquit` for Remain. Many of the hashtags related to economy and finance (e.g. `#pound` and `#GDP`) are Remain hashtags, and hashtags arguing to get free from the EU (e.g. `#freedom` and `#takecontrolday`) are Leave hashtags. Also, there are hashtags in an unexpected category. For example, `#farage`, the then-leader of the UK Independence Party, is classified as Remain, probably because the use of this hashtag in many occasions was intended to message the criticism against the party and their claims made during the campaign.

5 Discussion

Social media data is notoriously noisy. Our efforts to measure change in public opinion on Brexit through the Twitter data confirms that. 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 of likely voters were smaller than correlations between Twitter data and polls of adults in general. Having said this, we discuss four possible shortcomings that our approach faces.

First, the use of hashtags to label Tweets as Leave/Remain implied that the training set was not built out of a random sample. As such, there exists the possibility that, 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 classi-

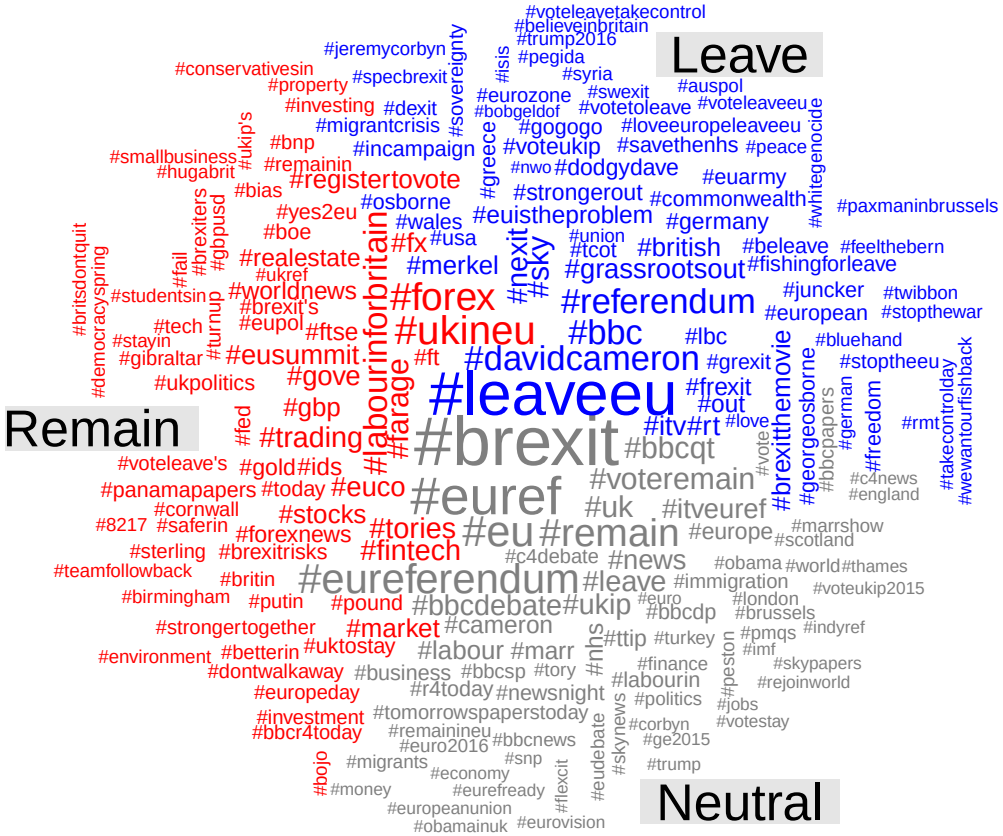


Figure 2: Wordcloud of Influential Hashtags.

Note: The hashtags are classified into three categories with arbitrary cutoffs in probability for a hashtag to be Remain or Leave at 0.25 and 0.75. Hashtags between 0.25 and 0.75 are classified as Neutral. We select 100 most frequent hashtags in each category. The font size is proportional to the frequency of the tag usage in logarithmic scale.

fication appears to be highly correlated to what Internet polls reported.

Second, 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 Twitter forces users to limit the length of message 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.[13] 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.

Third, further processing of post-classification results improves the detection of opinion shifts from Tweets. In particular, 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 are large population of Twitter users in Britain, some 20% by recent estimates (eMarketer., 2016). The high correlations between Internet polls and Twitter data are potentially due to the relatively high level of political participation by adults through social networks in Britain. Applying this approach in countries where the level of online political 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 organized discussion structure of Tweets on Brexit 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 believe that individuals looking to replicate these process should bear in mind the methodological limitations discussed above at the time of decision-making.

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

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 referendum and the 2016 United States Presidential Election 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 (Berinsky, 2017). In particular, the time-lag between influential events and results reflect in traditional polls means that electoral campaigns cannot react quickly to shocks in public opinion. Scholars have discussed the possibility to address this problem 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 than just friends and family (Tumasjan et al., 2010, Fábrega and Sajuria, 2014, Barberá, 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 hundreds of thousands 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 note that the correlation between Internet and telephone polls is low and, conversely, Twitter data and telephone polls is as well. As noted in the discussion, 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 rigorous 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, the fast classification would be of the highest importance to practitioners than to academic researchers. 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 quantities. 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.

Barberá, P. (2014): "Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data," *Political Analysis*, 23, 76–91.

Barberá, P. and G. Rivero (2015): "Understanding the political representativeness of twitter users," *Social Science Computer Review*, 33, 712–729.

Beauchamp, N. (2017): "Predicting and Interpolating State-Level Polls Using Twitter Textual Data," *American Journal of Political Science*, 61, 490–503.

Benoit, K., K. Watanabe, P. Nulty, A. Obeng, H. Wang, B. Lauderdale, and W. Lowe (2017): *quanteda: Quantitative Analysis of Textual Data*, URL `http://quanteda.io`, r package version 0.99.

Berinsky, A. J. (2017): "Measuring Public Opinion with Surveys," *Annual Review of Political Science*, 20, 309–329.

Bermingham, A. and A. F. Smeaton (2010): "Classifying sentiment in microblogs," in *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*, 1833, URL `http://portal.acm.org/citation.cfm?doid=1871437.1871741`.

Bird, S., E. Loper, and E. Klein (2009): *Natural Language Processing with Python*, OReilly Media Inc.

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.

Chin, D., A. Zappone, and J. Zhao (2016): "Analyzing twitter sentiment of the 2016 presidential candidates," .

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 (2010): "A method of automated nonparametric content analysis for social science," *American Journal of Political Science*, 54, 229–247.

Howard, P. N. and B. Kollanyi (2016): "Bots,# strongerin, and# brexit: Computational propaganda during the uk-eu referendum," *Working paper*.

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.

Manning, C. D., P. Raghavan, and H. Schütze (2008): *Introduction to Information Retrieval*, 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.

Morstatter, F., J. Pfeffer, H. Liu, and K. Carley (2013): "Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose," *Proceedings of ICWSM*, 400–408.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011): "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, 12, 2825–2830.

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. (2016a): "The myth of the lag," `http://fivethirtyeight.com/features/myth-of-lag/`, retrieved on: 2017-01-06.

Silver, N. (2016b): "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]," `https://twitter.com/NateSilver538/status/796411118302302208`, 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.