Predicting Brexit: Classifying Agreement is Better than Sentiment and Pollsters

Fabio Celli[1], **Evgeny A. Stepanov**[1], **Massimo Poesio**[2], **Giuseppe Riccardi**[1]
[1]Signals and Interactive Systems Lab, University of Trento, Italy
{fabio.celli,evgeny.stepanov,giuseppe.riccardi}@unitn.it
[2]School for Computer Science and Electronic Engineering, University of Essex, UK
poesio.essex.ac.uk

Abstract

On June 23rd 2016, UK held the referendum which ratified the exit from the EU. While most of the traditional pollsters failed to forecast the final vote, there were online systems that hit the result with high accuracy using opinion mining techniques and big data. Starting one month before, we collected and monitored millions of posts about the referendum from social media conversations, and exploited Natural Language Processing techniques to predict the referendum outcome. In this paper we discuss the methods used by traditional pollsters and compare it to the predictions based on different opinion mining techniques. We find that opinion mining based on agreement/disagreement classification works better than opinion mining based on polarity classification in the forecast of the referendum outcome.

1 Introduction

The outcome of the 2016 EU referendum did not only spell disaster for the UK government and the Remain campaign. It also amounted to a Press Release disaster for commercial pollsters. YouGov, Populus, ComRes, ORB, Ipsos-Mori and Survation, all failed to correctly predict the outcome. Of the larger pollsters, only TNS and Opinium correctly called the outcome, although still underestimating the Leave vote. This general failure, moreover, follows hard on the heels of similar failures in both the 2010 and 2015 General Elections, and public faith in commercial polling has taken another serious blow. By contrast, predictions using Natural Language Processing (NLP) and Computational Linguistics (CL) techniques, such as opinion mining, proved to be much more reliable. In what follows we will refer to opinion mining as the automatic task of assigning a polarity to a topic in context (Wiebe et al., 2005); to polarity classification and sentiment analysis as the tasks for the extraction of emotive polarity or scores from text; to agreement/disagreement classification as the task of recognizing the opinion of a message towards others in a thread (Wang and Cardie, 2014) or pairs of replying posts (Celli et al., 2016); and to stance classification as the task of recognizing the overall opinion of an author from text.

In this paper we discuss the methods used by traditional pollsters and compare them to the predictions based on different opinion mining techniques, in particular polarity classification and agreement/disagreement classification. We describe a system that predicted the outcome of the referendum correctly to within one-tenth of a percentage point. Unlike many political prediction papers that provide post-hoc analyses (Gayo-Avello, 2012), our final prediction was publicly

available the day before the referendum (i.e. 22nd of June) on our referendum monitoring web site[1]; thus, it is indeed the prediction of the future result.

The rest of the paper is structured as follows: in Section 2 we report the techniques used by commercial pollsters to forecast the votes, in Section 3 we report related work on forecasts from social media using opinion mining. Section 4 describes the methodology we have used for data collection and the system for the prediction of the referendum outcome. In the same section we provide analyses of the representativeness of our data sources and the methods for topic labeling and automatic annotation of opinions. Finally, in Section 5 we analyze and compare polling and NLP-based predictions. We hope that the results and discussions presented in this paper will contribute to pushing NLP further in the exploitation of para-semantic analysis techniques to forecast and understand collective decisions.

2 Traditional Opinion Polling

Traditional commercial polling for the EU referendum in UK started in the months between the announcement of the referendum (January 22, 2013) and the referendum day (June 23, 2016). Although the UK government started a pro-Remain campaign in April 2016, opinion polls of voters in general tended to show roughly equal proportions in favor of remaining and leaving. Polls done in the weeks preceding the referendum showed majority being in favor of remaining, and the outcome of the referendum showed that there is a bias in the methods used by traditional opinion pollsters to sample and collect the data.

Known issues with traditional opinion polling techniques are related to demographic bias in the way voters are polled. Demographics-wise, post-referendum analyses showed that younger voters tended to support remaining in the EU, but are generally less likely to vote; whereas older people tended to support leaving, and they are less likely to use social media or reply to online polling. According to two out of three pollsters, managerial, professional and administrative workers were most likely to favor staying in the EU, while semi-skilled and unskilled workers, plus those reliant on benefits, were the largest demographics supporting leaving. University graduates were generally more likely to vote Remain compared to those with no qualifications. White voters were evenly split, and all ethnic minority groups leaned towards backing Remain. Support for remaining in the EU was known to be significantly higher in Scotland than it is in Great Britain as a whole, with Scottish voters saying they are generally more likely to vote.

The way voters are polled is known to affect the outcome. Traditional methods consisting of telephone polls usually based on small samples ranging from 1000 to 1500, and online polls are usually based on larger samples (from 2000 to 5000). Telephone polls have consistently found more support for remaining in the EU than online polls. Ipsos-Mori and ComRes argued that telephone polls are more reliable, but YouGov, which uses online polling, has criticized telephone polls because they have a high percentage of graduates, thus skewing the results towards Remain. A study by Populus[2] concluded that telephone polls were likely to better reflect the state of public opinion. However, overall for the EU referendum, online polls seem to have had a better prediction than phone polls .

Table 1 reports the results of the major commercial pollsters, with details on the sample size

[1] http://www.sense-eu.info

[2] http://www.populus.co.uk/2016/03/polls-apart/

time window	Remain	Leave	undecided	sample	pollster	method
22 June	55%	45%	0%	4700	Populus	Online
20-22 June	51%	49%	0%	3766	YouGov	Online
20-22 June	49%	46%	5%	1592	Ipsos Mori	Phone
20-22 June	44%	**45%**	11%	3011	Opinium	Online
17-22 June	48%	42%	10%	1032	ComRes	Phone
16-22 June	41%	**43%**	16%	2320	TNS	Online
20 June	45%	44%	11%	1003	Survation	Phone
18-19 June	42%	**44%**	14%	1652	YouGov	Online
16-19 June	53%	46%	1%	800	ORB	Phone

Table 1: Overview of the results obtained and methods adopted by traditional pollsters for the referendum.

and the methods adopted. In the days before the referendum, only TNS and Opinium predicted the outcome correctly, both using online polling and a three day time window, or larger. But results are contradictory: Populus used online polling, and with a larger sample, but they focused on a one-day time window and their prediction failed. Moreover, YouGov gave a first correct prediction with online polls from June 18 to 19, and then failed using the same method with a larger sample collected between June 20 to 22.

Other pollsters based their predictions on various aggregations of the polls from different companies, adjusting for biases and gaps they have perceived in their methodology, such as the one between telephone and online polling. However, no pollster utilizing this methodology was able to predict the referendum outcome correctly.

3 Opinion Mining and Forecasting

Opinion mining based on sentiment analysis has become one of the most popular tasks in the last decade (Li and Wu, 2010) and many works have demonstrated how much it can be useful for recommendation systems (Zhang and Pennacchiotti, 2013) among other tasks. Opinion mining is traditionally performed by means of sentiment lexica or dictionaries (Cambria et al., 2012), althought other methods based on semantics (Agarwal et al., 2015) or stylometry (Anchiêta et al., 2015) have been tested in recent years. One of the most popular applications of sentiment analysis to event forecasting is perhaps the works initiated by Bollen and colleagues on the prediction of the stock market from Twitter: they found strong correlations between collective mood states extracted from large-scale Twitter feeds and the value of the Dow Jones on a 3-days time window (Bollen et al., 2011). They also attempted to detect the public's response to the US presidential election and Thanksgiving day in 2008, successfully predicting the outcome with an accuracy of 86.7%.

More recent studies analyze political opinions and make political forecasts through sentiment analysis of social media: for example, O'Connor connected measures of public opinion from polls with sentiment measured from text and found strong correlations between public opinion and tweet texts (O'Connor et al., 2010). This highlights the potential of text streams as a substitute fot or supplement to traditional polling. Other studies showed that the mere number of political party mentions accurately reflects the election results (Tumasjan et al., 2010). In

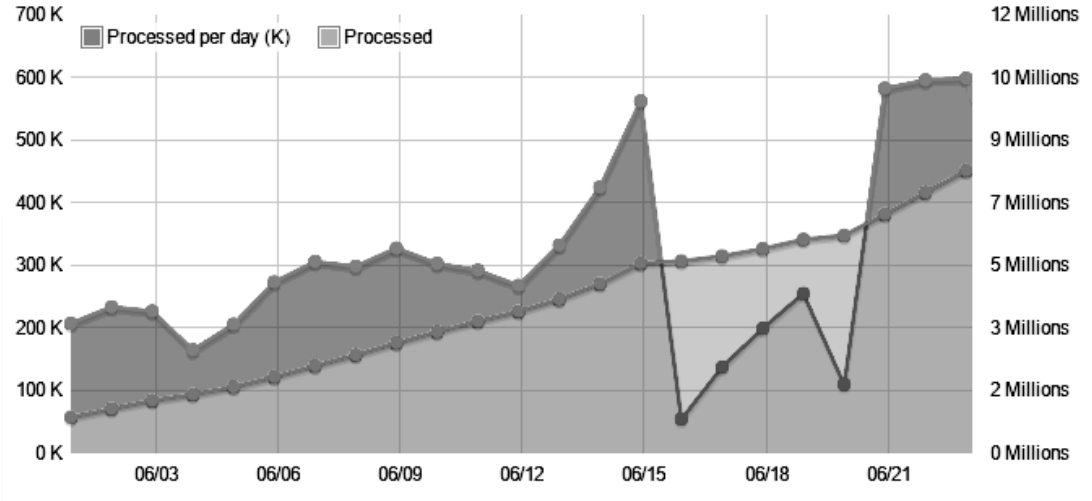

Figure 1: Data collected per day and overall from June 1, 2016 to the referendum date. The lower peaks correspond to the aftermath of the Cox's murder, followed by a news breakout on Brexit.

a recent study, Burnap and colleagues used Twitter data to forecast the outcome of the 2015 UK General Election: they exploited sentiment analysis and prior party support to generate a forecast of parliament seat allocation that turned out to hit the final result with high accuracy (Burnap et al., 2016).

4 Prediction Methodology

Similar to other papers using social media for political predictions, such as (O'Connor et al., 2010), our methodology consists of collecting social media data and applying opinion mining techniques to predict the distribution of votes.

4.1 Data Collection for Referendum Monitoring

Starting from May 19, 2016 we crawled the web for conversations about Brexit using hand-crafted lists of keywords, hashtags and mentions (e.g. *EUreferendum*, *#Brexit*, and *@ukleave-eu*); and created daily data dumps. The conversations were collected from more than 4000 sources such as newspaper blogs, social network sites and other types of social media in 20 languages and from 14 countries of the European Union. The collected data was automatically processed with cross-language algorithms for extracting topics (Leave/Remain) and opinions towards them. Within the referendum monitoring time frame, we have collected and processed more than 8 million posts (see Figure 1), about 80% of which comes from Twitter. The collected data will be made available on http://sisl.disi.unitn.it.

4.2 Representativeness of the Social Media Data for Political Predictions

How representative is the social media data of the voter demographics is a debated topic. As it is illustrated in Figure 2 that shows the distribution of social network and online news readers

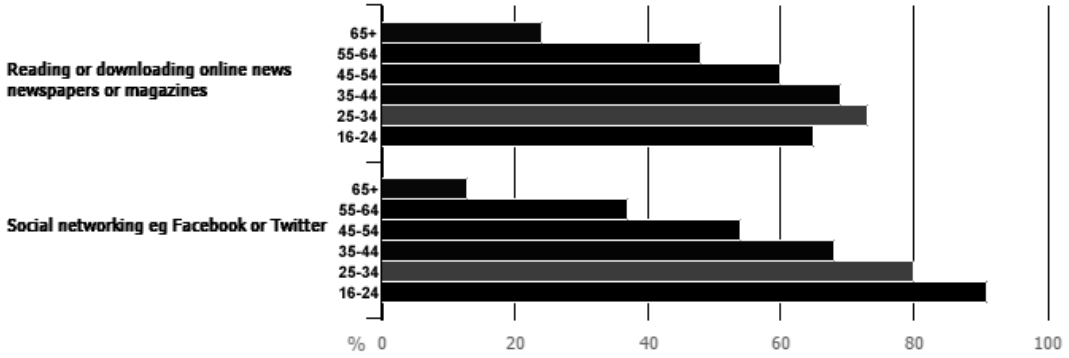

Figure 2: Percentage of online news readers and social network users by age. Source: ONS, year 2014.

Leave	**Remain**
euroscepticism,	SayYes2Europe, Remain,
#beLeave, #betteroffout, #britainout,	#bremain, #betteroffin , #leadnotleave,
#LeaveEU, #noTTIP, #TakeControl,	#Remain, #Stay, #strongerin, #ukineu,
#VoteLeave, #VoteNO, #voteout,	#votein, #voteremain, #VoteYES,
@end-of-europe, @leaveeuofficial,	#yes2eu, #yestoeu, #SayYes2Europe,
@NoThanksEU, @nothankseu,	
@ukleave-eu, @vote-leave	

Table 2: Sets of keywords, hashtags and mentions for assigning posts to Leave and Remain categories.

reported by the national UK statistic agency[3] in 2014, not every age group is equally represented in social media. The same is also true for other demographic factors such as gender, race, social class, etc. An extensive work on 70 million tweets collected between 2011 and 2012 during Spanish and US presidential elections, showed that Twitter users who write about politics tend to be male, to live in urban areas, and to have extreme ideological preferences (Barberá and Rivero, 2014). Moreover, since there is usually no demographic information available in Twitter or other sources as meta data, sample representativeness is not easy to verify. Thus, it is inevitable that it will be biased. For predicting the outcome of Brexit referendum, we did not apply any techniques to account for UK voter demographics. We plan to address this in future work.

4.3 Leave/Remain Topic Labeling

As the first step, the posts in the collected data are automatically assigned Leave or Remain topics. The task is performed by means of simple hand-crafted rules that use keywords, hashtags and mentions to map the posts to classes. If a post contains keywords, hashtags, or mentions for Leave and not for Remain, it is mapped to Leave; and if it contains keywords, hashtags or mentions for Remain and not for Leave, it is mapped to Remain. Unclassified posts were not

[3]http://www.ons.gov.uk/

used for the prediction. The sets of keywords, hashtags and mentions used for each class were selected such that they yield balanced probabilities. The sets for each class are given in Table 2.

4.4 Automatics Classification of Author's Opinions

Just assignment of a topic to a post is not enough for the prediction of authors' opinions. The authors' opinions towards topics could be predicted either as a sentiment polarity expressed in a post, or as an agreement or disagreement with the topic expressed in a post. In this section we describe the sentiment polarity prediction and the agreement/disagreement prediction systems that are used for posts classification.

4.4.1 Agreement/Disagreement

The system for the automatic labeling of posts with agreement/disagreement makes use of language independent stylometric features such as: character-based ratios of upper and lowercase letters, numbers, various punctuation marks and special characters; word-based ratios of URLs, Twitter mentions and hashtags, negative and positive emoticons. Additionally, the model considers ratios of character and word ngrams (bigrams to tetragrams). All the features have their numerical values between 0 and 1.

The model is trained and evaluated on the Italian CorEA corpus (Celli et al., 2014) using 66% of the data for training and 33% for evaluation. The corpus consists of about 2900 posts to online news articles that were manually annotated with respect to agreement, disagreement and neutrality/not applicability to the parent posts. The system was trained only on agreement and disagreement labels, removing neutral and not applicable cases. The inter-annotator agreement on two classes is k=0.85 and the manually annotated posts used for training and testing are approximately 2000. The task is cast as a regression with the well balanced bimodal distribution. The performance of the Support Vector Regressor (Shevade et al., 2000) on the CorEA test set has a Mean Absolute Error (MAE) of 0.32.

Even though the model is trained and tested on Italian data, the features are language independent: semantics of features such as emoticons and punctuation are similar at least across European languages; thus, we believe that the model is applicable to other languages as well.

4.4.2 Sentiment Polarity

The sentiment polarity prediction system is lexicon-based. We used OpenNER polarity lexicon[4] to label each post as either negative, positive, or neutral. The posts classified as neutral were removed for the prediction. The performance of the system on the Movie Reviews 2.0 data set (Pang and Lee, 2004) has accuracy of 68.7%. Even though the system has moderate performance, it is in line to the state-of-the-art lexicon-based approaches to sentiment analysis.

5 Brexit Prediction, Analysis and Evaluation

We have predicted the outcome of the referendum from a subset of approximately 178 thousand posts in a time window of 2 days (June 20 and June 21). In this paper we compare two opinion mining systems – the one based on sentiment polarity and the one based agreement/disagreement classification. The baseline is the volume of posts about Leave and Remain topics, obtained by

[4]http://www.opener-project.eu

System	Leave		Remain	
	Counts	Percentages	Counts	Percentages
Baseline	178,722	(60.97%)	114,403	(39.03%)
Sentiment Polarity	63,788	(51.26%)	60,657	(48.74%)
Agreement/Disagreement	90,847	(51.79%)	84,560	(48.21%)
Referendum Outcome		51.9%		48.1%

Table 3: Counts and percentages for Leave and Remain as predicted by sentiment polarity prediction system (*polarity*) and agreement/disagreement prediction system (*agreement/disagreement*). Baseline is the counts of posts selected by hand-crafted rules. The referendum outcome is provided for the reference.

topic labeling with the hand-crafted rules described in Section 4.3. For the final prediction of each system we compute the percentage of posts that are positive towards one class and negative towards the other. For example, the predicted percentage for Leave counts posts in agreement with Leave and in disagreement with Remain, and vice versa. Neutral posts are ignored (this is why the posts used by the NLP systems are fewer than the posts used for the baseline).

Predictions using each system are reported in Table 3. While sentiment polarity and agreement/disagreement systems yield correct predictions, we found that the baseline is significantly offset and overestimates Leave (60.97%). This suggests that people tend to write a lot about Leave, but mainly to criticize. The agreement/disagreement based prediction is more accurate than the sentiment polarity based prediction. One reason can be that the agreement/disagreement based system considers significantly more posts ($\approx$ 50K more) than the sentiment polarity system.

Our findings support the claim that NLP techniques such as opinion mining can be very useful to opinion polling and social media analytics, and that events such as a referendum can be predicted with high accuracy. However, a correct prediction is the result of a combination of many factors, where the time period is important as well as the analysis method. As literature reports, a time-window of 2 or 3 days is the best for a prediction, and we used a 2 days time-window like many pollsters. However, in this specific case, we were able to capture the moment when undecided people (estimated between 7% to 11% of voters) changed their minds towards Leave areas, while traditional pollsters were not. In the aftermath of the referendum, YouGov attributed the error in their predictions to this higher turnout in Leave-oriented areas not captured by their polls. In our opinion there are three reasons for the NLP techniques being able to produce more accurate predictions than traditional polling:

- NLP techniques can analyze much larger sample sizes. Traditional polls typically interview on average 1000 to 4000 individuals. By contrast, with NLP techniques we processed a minimum of 80K to 100K posts per day, and aggregation of this order produces compelling evidence.

- Traditional polling asks for the peoples behavioral intentions or opinions, whereas analyses carried out with NLP techniques try to infer opinions that motivate behavior. Modern cognitive science has established that direct questions about opinions and behavioral intentions may produce unreliable and invalid responses (Hufnagel and Conca, 1994). Asking subjects to fill questionnaires is only used when more indirect methods cannot be applied,

such as measuring the time it takes to perform a task, or eye-tracking. NLP in this case represents such an indirect method, since it focuses on opinions that are some distance from the behavior.

- Data collected from social media and processed with NLP techniques may well cover posts coming from a wider range of geographical locations and demographic variety than pollster's surveys.

6 Conclusions

We have predicted the outcome of the Brexit referendum with high accuracy exploiting NLP techniques and outperforming a baseline based on the volume of posts. We analyzed some possible causes of this result, comparing our prediction to pollsters' surveys. Our findings are based just on one event, and require further study to be consolidated. To date, however, neither polling organizations nor the media have paid much attention to NLP methods for election and referendum forecasting, but the results of this work suggest that these methods, with all their limitations, can produce reliable forecasts.

At the very least, campaigners and the media alike should consider using NLP methods to compare with or complement the polls. While every new methodology is rightly treated with a degree of suspicion and while it is premature to expect traditional polling to disappear, there are grounds for both campaigners and the media to take NLP techniques seriously in the future.

Acknowledgements

The research leading to these results has received funding from the European Union - Seventh Framework Programme (FP7/2007-2013) under grant agreement n. 610916 - SENSEI. We would like to thank Websays[5] for providing the data for the baseline.

References

Basant Agarwal, Soujanya Poria, Namita Mittal, Alexander Gelbukh, and Amir Hussain. 2015. Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach. *Cognitive Computation*, 7(4):487–499.

Rafael T Anchiêta, Francisco Assis Ricarte Neto, Rogério Figueiredo de Sousa, and Raimundo Santos Moura. 2015. Using stylometric features for sentiment classification. In *Proc. of CICLing 2015*, pages 189–200.

Pablo Barberá and Gonzalo Rivero. 2014. Understanding the political representativeness of twitter users. *Social Science Computer Review*, pages 1–29.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1 – 8.

Pete Burnap, Rachel Gibson, Luke Sloan, Rosalynd Southern, and Matthew Williams. 2016. 140 characters to victory? using twitter to predict the uk 2015 general election. *Electoral Studies*, 41:230–233.

[5] http://websays.com

Erik Cambria, Catherine Havasi, and Amir Hussain. 2012. Senticnet 2: A semantic and affective resource for opinion mining and sentiment analysis. In *Proc. of FLAIRS*, pages 202–207.

Fabio Celli, Giuseppe Riccardi, and Arindam Ghosh. 2014. Corea: Italian news corpus with emotions and agreement. In *Proc. of CLIC-it 2014*, pages 98–102.

Fabio Celli, Evgeny Stepanov, and Giuseppe Riccardi. 2016. Tell me who you are, i'll tell whether you agree or disagree: Prediction of agreement/disagreement in news blog. In *Proc. of NLPMJ*.

Daniel Gayo-Avello. 2012. "i wanted to predict elections with twitter and all i got was this lousy paper" – a balanced survey on election prediction using twitter data. *CoRR*, abs/1204.6441.

Ellen M Hufnagel and Christopher Conca. 1994. User response data: The potential for errors and biases. *Information Systems Research*, 5(1):48–73.

Nan Li and Desheng Dash Wu. 2010. Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems*, 48(2):354 – 368.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. pages 122–129.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proc. of the ACL*.

Shirish Krishnaj Shevade, S Sathiya Keerthi, Chiranjib Bhattacharyya, and Karaturi Radha Krishna Murthy. 2000. Improvements to the smo algorithm for svm regression. *Neural Networks, IEEE Transactions on*, 11(5):1188–1193.

Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. 2010. Election forecasts with twitter: How 140 characters reflect the political landscape. *Social Science Computer Review*, pages 1–17.

Lu Wang and Claire Cardie. 2014. Improving agreement and disagreement identification in online discussions with a socially-tuned sentiment lexicon. *Proc. of the ACL*, pages 97–102.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

Yongzheng Zhang and Marco Pennacchiotti. 2013. Recommending branded products from social media. In *Proc. of ACM conference on Recommender Systems*, pages 77–84.