Predicting the EU 2014 Election Results in Multiple Countries Using Twitter

Adam Tsakalidis[1,2], Symeon Papadopoulos[1], Alexandra Cristea[2], and Yiannis Kompatsiaris[1]

[1]Information Technologies Institute, Center for Research and Technology Hellas, Thessaloniki, Greece
{atsak, papadop, ikom}@iti.gr
[2]Department of Computer Science, University of Warwick, Coventry, UK
{A.Tsakalidis, A.I.Cristea}@warwick.ac.uk

Abstract

During the latest years, the behavior of users in Twitter has been explored for various purposes, one of the most famous being the prediction of election results. Most works so far make their predictions by focusing strictly on Twitter data and are applied on some data after the elections; hence, they are biased towards the actual results.

In the current work we have focused on the 2014 European Union Elections for three countries. We monitored political discussions on Twitter and created time-series of the political parties by extracting various features; at the same time, we aggregated opinion polls to serve as our ground-truth. Based on those features, we tried to predict the election results, publishing our predictions for one country before the elections ended. Our approach achieved low error rates, being better than two prediction websites and suggesting that Twitter can be effectively used for this task.

Keywords: Machine learning; Web mining

1 Introduction

Twitter is a microblogging platform that enables users to share short messages ("tweets") with their "followers". Due to the large volume of user interactions and their frequent updates, Twitter has seen increased overall interest, being the 9th most popular website in July, 2014 (`http://www.alexa.com/siteinfo/twitter.com`). Hence, it is not a surprise that the content produced within it is exploited for various research tasks during the latest years in an attempt to model and predict users' behavior.

The current work focuses on exploiting this content for the task of predicting the 2014 European Union (EU) Election results for Germany, the Netherlands and Greece. While several works have been conducted on the same domain, many of them have relied strictly on Twitter data and have been proven ineffective when tested on different cases. Furthermore, most of the past works have published their results after the elections, while the benefit of using Twitter data for this task is quiestionable in many cases [8].

In this work we treat the users' voting intentions as time-variant features. Instead of trying to predict every user's vote, we treat Twitter political discussions as a general index that varies with time; we extract several Twitter-based features and fit them in time-series models, using opinion polls as our ground-truth. In this way, we combine the twitter-based time-series with the poll-based ones. We test three different forecasting algorithms using three different sets of features; we compare our results with several baselines, achieving lower error rates even than two prediction websites and the polls from the last week before the elections. Furthermore, working on different elections at the same time, we demonstrate our approach's portability; more importantly, we show that there by using our twitter-based features all three algorithms get a significantly important boost in accuracy compared to the one obtained when using only poll-based ones. Last but not least, we are among the first to have published our predictions *before* the announcement of the Exit Polls for one country, preventing any bias towards them, while we follow the exact same methodology for the other two countries.

2 Background

2.1 EU Elections

The EU Parliament elections are held every five years among the EU member states. Political parties from different countries form coalitions that constitute the European political parties; however, people within every country are only allowed to vote for their country's political parties. The 2014 EU elections were held in late May and have been judged as significantly important, mainly because of the economic crisis and the rise of euroskepticism. Due to the nature of these elections, it is difficult to predict the results at a pan-European level without taking into account the important demographic and political differences between the EU members. Thus, we focused on three different countries, transfering the problem to a national level. The elections were held on the May, 22 for the Netherlands and on the May, 25 for Germany and Greece. There were 10 main political parties contesting in the Netherlands, six in Germany and eight in Greece, for which we tried to predict the results.

2.2 Related Work

One of the most popular works on the field of predicting election results was performed by Tumasjan et al. [10], demonstrating that the number of times a political party's name appears on Twitter is a fairly good estimate of its voting share. However, their method was unsuccessfully applied in another context [1]. A naïve counting and a sentiment analysis method did not perform well on the analysis by Metaxas et al. [4] either, predicting the correct result in only half of the cases with two candidates.

Since such approaches cannot be generalized, recent works have started working on Twitter by using opinion polls as their ground-truth. Lei et al. [9] used aggregated poll reports in order to train their Twitter-based models, by also examining the geographical locations of the users in an attempt to predict the results per-location. However, sentiment analysis features, which are considered to be important for this task, were not included in their modelling. Using poll reports as ground-truth, Lampos et al. [3] created time-series by taking into account both user- and keyword-based features for the major parties of two countries. Sang and Bos manually fit their Twitter-based data on polls, achieving however slightly worse results [8]. Even worse though, when they replaced their Twitter-based features with uniform variables, their predictions got better, implying that Twitter did not actually help on the prediciton task. For a more complete review on the field, the reader is prompted to the work of Gayo-Avello [2].

In the current work we follow the idea of poll-based training by using keyword-, user- and sentiment-based features for building our models. We achieve reasonably better results than the polls and we explore the role of different sets of twitter-based features, revealing a statistically significant boost when these are incorporated in the prediction process, contradicting the findings in [8].

3 Methodology

We consider our problem as a multivariate time-series forecasting task. Working on every country separately, we create time-series of eleven twitter- and one poll-based features for every party (section 3.2). An example is the number of tweets mentioning a certain party on a specific day (twitter-based) and the percentage for that party reported on a poll that was conducted on that day (poll-based). After certain normalisation steps, we end up with a single value for each feature for every party on a daily basis. At the final stage, we provide all of these features as an input to different forecasting algorithms, trying to predict the voting share of every party separately (section 3.3).

3.1 Data Aggregation

We started aggregating data published on Twitter and various opinion polls on a per-country basis between April, 6th until two days before the elections (20/5 for the Netherlands and 23/5 for Germany and Greece), leaving one day to conduct our processing. Using the public Twitter Streaming API (`https://dev.twitter.com/`), we aggregated tweets written in the respective language that contained a party's name, its abbreviation, its Twitter account name and some possible mispells (e.g., *grunen* instead of *grünen*). We excluded several ambiguous keywords in an attempt to reduce the noise (e.g., the abbreviation of the Dutch party "GL" may stand for "good luck"). This implies that we have missed some data, making it impossible to replicate accurately naïve mention-counting methods; nevertheless, most of them have been unsuccessfully applied in different cases and we also show that on our aggregated data.

3.2 Modelling

Twitter Features: Working on every country separately, we first assigned equal weights to all parties mentioned in a tweet so that they sum up to one. Let $t_d(p)$ denote the (weighted) number of tweets that mention party p on day d and $tpos_d(p)$ ($tneg_d(p)$) the corresponding number of tweets containing positive (negative) content. Similarly, let $u_d(p)$ denote the number of users mentioning party p on day d, $upos_d(p)$ ($uneg_d(p)$) the number of users that have published a tweet with positive (negative) content about party p on that day. We constructed 10 text- and user-based features on a daily basis:

1. $numTweets_d = \sum_i t_d(i)$
2. $pctTweets_d(p) = \frac{t_d(p)}{\sum_i t_d(i)}$
3. $pctTPos_d(p) = \frac{tpos_d(p)}{t_d(p)}$
4. $pctTNeg_d(p) = \frac{tneg_d(p)}{t_d(p)}$
5. $pctTPosShare_d(p) = \frac{tpos_d(p)}{\sum_i tpos_d(i)}$
6. $pctTNegShare_d(p) = \frac{tneg_d(p)}{\sum_i tneg_d(i)}$
7. $pctUsers_d(p) = \frac{u_d(p)}{\sum_i u_d(i)}$
8. $pctUPos_d(p) = \frac{upos_d(p)}{u_d(p)}$
9. $pctUNeg_d(p) = \frac{uneg_d(p)}{u_d(p)}$

10. $pctTotalUsers_d(p) = \frac{\sum_d u_d(p)}{\sum_d \sum_i u_d(i)}$

Here, $pctTotalUsers_d(p)$ refers to the *distinct* number of the users that have mentioned p divided by the total number of them up to day d. We also added the average sentiment value ($avgSentiment_d$) as a feature (notice that $numTweets_d$ and $avgSentiment_d$ were the same for all parties within a country). Finally, we used a 7-day Moving Averages (MA) filter for all features (except $pctTotalUsers_d(p)$) in order to normalise their values, as suggested by O'Connor et al. [6]. These 11 values for every party were used as our Twitter-based features and were provided as input to our algorithms, along with the opinion poll ones.

Opinion Polls: Opinion polls differ in many aspects with each other. Since there is not a complete polling aggregation service, we had to find different polls manually. Once aggregated, we removed all "small" parties reported polling values and added their voting share into the "Others" parties; then, we distributed proportionally to all parties (including "Others") the voting share of all "undecided" voters. In this way we managed to have consistent polls, adjusting their reports to include only the main political parties of each country, along with the "Others".

While creating time-series of Twitter features without missing values was a straight-forward process, this was not the case for the polls. A poll is usually conducted over two to three days; we treated the adjusted results as the actual voting shares each party would have received if the elections were held on any of these days. If two or more polls were held on the same day, we considered the voting share of each party as the weighted average value, using the sample size of every poll as the weight and making sure that all voting shares sum up to 100. Finally, we filled all days without polling data by using linear interpolation. Thus, we managed to create our poll-based feature for every party, with the only values missing being those corresponding to the days after the last poll.

3.3 Sentiment Analysis

Several Twitter-based features were sentiment-related; hence, we needed to assign a sentimental value on each tweet before proceeding. One of the most popular approaches on sentiment analysis is to train a classifier on a labeled corpus of tweets and apply it on the desired test set. However, past works have revealed the domain-dependent nature of such classifiers [7]. The integration of POS tags is also beneficial but there does not exist a reliable, free-to-use POS tagger for the three languages. Given these constraints, we decided to adopt the lexicon-based approach in order to create a generic method that could be applied in different cases. While such approaches perform only slightly better than a random classifier [4], we were only interested in the daily differences of the expressed sentiment; thus, given that we have

enough data on every day, even a slightly better than the random classifier method fit our goals [6].

Due to the lack of a sentiment lexicon for different languages, we translated three English lexicons using Google Translate (`https://translate.google.com/`). These included SentiWordNet (`http://sentiwordnet.isti.cnr.it/`, about 150,000 synsets with a double value indicating their polarity), Opinion Lexicon (`http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html`, about 6,800 polarized terms) and the Subjectivity Lexicon that serves as part of the Opinion Finder (`http://mpqa.cs.pitt.edu/opinionfinder/`, about 8,000 terms along with their Part-of-Speech (POS), subjectivity – strong/weak – and polarity indication).

We assigned the values of 1 and -1 for the positive and negative terms of Opinion Lexicon respectively; for the case of Subjectivity Lexicon, we used four values (-1, -0.5, 0.5, 1) to represent every subjective word depending on its subjectivity ($|0.5|$ for weak, $|1|$ for strong) and polarity; for SentiWordNet, we kept the values of every synset. We removed all terms that were not a single word, due to the inaccuracy observed in those translations. If the same word appeared in different lexicons, we considered the average as its sentimental value, resulting into 14,060/19,357 German, 13,838/18,993 Dutch and 13,582/18,356 *positive/negative* Greek terms. In order to detect a tweet's sentiment, we used a naïve sum-of-weights method of its keywords according to its language's new lexicon and assigned the majority class label (positive/negative) to it.

3.4 Algorithms

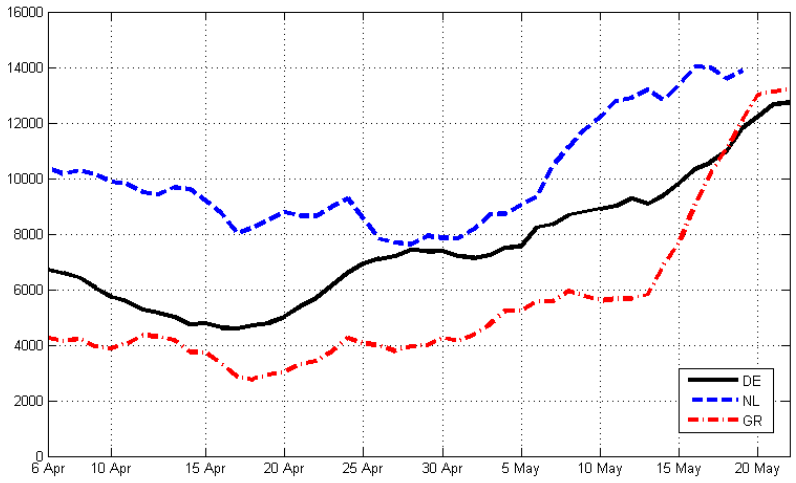
We tested three different algorithms on each political party separately, using only this specific party's features (11 Twitter- and one poll-based) as input. These algorithms were *Linear Regression*, *Gaussian Process* and *Sequential Minimal Optimization for Regression*, all implemented using Weka (`http://www.cs.waikato.ac.nz/ml/weka/`) with the default settings. Since it was difficult to evaluate each algorithm before the elections, we decided to empirically apply a seven-day training window for every algorithm and considered the average predicted percentage for every party as our final estimate (the results we published were based on ϵ-SVR as well; due to its low performance in Greece, it was not applied and is not presented here for any other country). Notice that all three algorithms had to fill the missing values for the poll-based features dated after the last poll. There was only one such point, for the political parties in Germany; in both Greece and the Netherlands the last polls were conducted two days before the elections –our last "training" day. Hence, the predicting horizon of our algorithms was set to 2 for Germany and 1 for the Netherlands and Greece.

4 Data

4.1 Twitter

We aggregated 361, 713 tweets from 74, 776 users in Germany, 452, 348 from 74, 469 users in the Netherlands and 263, 465 from 19, 789 users in Greece. Figure 1 shows the number of tweets per day for every country, after the 7-day MA. Our findings on the average sentiment value reveal that negative opinions dominate in political discussions (-0.54 for Germany, -1.09 for the Netherlands and -0.29 for Greece). As expected, there were far more tweets published in the week before the elections, whereas a slight decrease is noticed in the Easter week ($13 - 20/4$). However, due to the restrictions of the Twitter Streaming API (it returns no more than 1% of all public tweets), it could be the case that we have missed some data. Morstatter et al. have showed that the increase of global awareness over a topic or the decrease of the total number or tweets published on a day could result into a decrease of the coverage of the Streaming API [5]. However, since we have reasonable amounts of data per day and are only interested in time-series modelling, this should not affect our process.

Figure 1: Number of political tweets aggregated per day, after a 7-day MA.



For further analysis, we divided the users into four distinct categories with respect to the number of tweets they published per day. Table 1 reveals that there is a small group of users who have tweeted about half of the tweets, whereas the vast majority of the users (ranging from 76.8% for Greece up to 90.3% for Germany) have tweeted up to one political tweet per week. Finally, as expected, post-processing of our data revealed that none of our features was (consistently) correlated with the results.

Table 1: Frequency of tweets per user category for Germany (DE), the Netherlands (NL) and Greece (GR).

	Users (%)			#Tweets (%)		
Frequency	DE	NL	GR	DE	NL	GR
Once (0, 1]	63.2	58.3	41.3	13.1	9.6	3.1
Weekly (1, 7]	27.1	29.2	35.5	17.9	15.8	9.2
Daily (7, 46]	8.2	10.5	16.9	28.9	29.9	22.9
Higher (46+)	1.5	2.0	6.3	40.1	44.7	64.8
Total	100.0	100.0	100.0	100.0	100.0	100.0

4.2 Opinion Polls

In total, we used 27 different polls from 11 different sources in Greece, 9 from 4 sources in Germany and 13 polls from 3 sources in the Netherlands. More specifically, we used all the polls published in Metapolls (`http://metapolls.net/`); further resources used were `http://www.wahlrecht.de/` for Germany, `http://www.3comma14.gr/` for Greece and polls from Ipsos, TNS Nipo and Peil.nl for the Netherlands.

Table 2: Variance of reported voting shares in the processed polls.

Germany		The Netherlands		Greece	
CDU/CSU	1.15	PVV	1.72	ND	3.20
SPD	0.99	VVD	3.86	SYRIZA	4.15
Grünen	1.00	D66	1.41	XA	2.13
Linke	0.63	CDA	3.52	Potami	5.07
AfD	0.25	PvdA	1.74	KKE	0.52
FDP	0.25	SP	2.20	Elia	1.45
Other	1.00	CU/SGP	0.61	ANEL	0.58
		GL	0.48	DIMAR	0.37
		50+	0.33	Other	1.85
		PvdD	0.37		
Average	0.75	Average	1.62	Average	2.15

Table 2 shows the variance of every party's voting share, after our preprocessing (for the Netherlands the "Others" category was not included in our analysis because of the inconsistency of the polls). In general, the voting shares of the German parties are rather stable; on the contrary, the percentages reported for the Dutch and the Greek parties vary a lot, reflecting the differences of people's voting intentions through time. Intuitively, predicting the results for Germany should be an easier task compared to the other

countries, as long as the polls do not deviate much from the actual results.

5 Results

In the current section we present the results obtained from our method ("Twitter Sensor", "TS"), along with several other methods used as baselines:

CB1 The Count-Based by Tumasjan et al. [10].

CB2 A similar naive method presented in [8]. Working on the last week's tweets, we apply this by keeping the tweets that mention only one party and then the first tweet of every user; at the final stage, voting shares are given to the parties as in CB1. In both CB cases, since we did not have data for the last day before the elections, we worked on the last seven days that we had data for.

S&B This is a replication of Sang & Bos's work [8]. We have used the average of all polls before the last week for training, whereas for sentiment analysis we use our own naive dictionary-based method; for details, see [8].

Polls The average of the polls conducted during the last week; there was one in Germany, two in the Netherlands and seven in Greece.

MP This baseline refers to the predictions of MetaPolls.net. This is the only polling aggregation website that we could find online providing predictions for all EU countries. MetaPolls provide their voting estimates for every party in a range of values; we considered the average value of this range for every party as the predicted percentage, making sure that the values sum up to 100.

PW PollWatch (`http://www.electio2014.eu/`) is the official prediction website that is powered by VoteWatch Europe and Burson-Marsteller/Europe Decides.

PB In order to evaluate the use of our Twitter features, we provide the results by applying our methodology using only polling data as features. Hence, this Poll-Based method is the average of our three algorithms presented in section 3.4, by providing to them the poll data points of the last 7 days.

NS Similarly, the No-Sentiment method was used in order to evaluate the performance of our sentiment analysis features. Hence, its features include the polling data points along with numTweets, pctTweets, pctUsers and pctUsersTotal.

For the case of the Netherlands, we assigned the voting share of the "Other" parties (2.45%) analogously to the remaining parties (see section 4.2). The metrics that we use for evaluation are the Mean Absolute Error (MAE), the Mean Squared Error (MSE) and the Tau Kendall Coefficient. Table 3 presents all the results together, along with the average per-country values of our evaluation metrics.

Table 3: Results and predictions per country.

	Party	Result	CB1	CB2	S&B	Polls	MP	PW	PB	NS	TS
Germany (DE)	CDU	35.30	20.16	19.72	36.58	37.50	37.76	37.70	37.08	38.06	37.04
	SPD	27.30	22.50	19.50	34.80	26.50	26.53	27.00	26.32	26.16	27.20
	Grunen	10.70	10.32	11.24	7.31	10.00	10.07	10.70	9.55	8.68	9.50
	Linke	7.40	11.24	9.48	8.38	7.50	8.28	8.30	7.44	8.00	6.70
	AfD	7.10	25.82	19.64	10.88	7.00	6.58	6.30	7.41	7.37	8.02
	FDP	3.40	9.96	20.43	2.05	3.50	3.39	3.00	4.07	3.50	3.41
	Others	8.80	–	–	–	8.00	7.38	7.00	8.12	8.23	8.13
DE	MAE	0.00	9.13	9.97	2.74	0.69	0.95	0.93	0.80	1.08	0.76
	MSE	0.00	129.24	148.52	10.04	0.95	1.44	1.53	0.92	1.97	0.90
	Tau-a	1.00	0.20	-0.07	0.60	1.00	0.90	1.00	1.00	1.00	0.90
The Netherlands (NL)	D66	15.87	14.79	13.58	24.96	17.49	17.67	18.53	16.49	16.22	15.72
	CDA	15.56	8.73	9.46	13.13	11.84	12.63	11.46	13.33	13.70	13.16
	PVV	13.65	14.94	10.80	15.40	16.28	13.63	14.23	15.66	16.09	16.70
	VVD	12.32	12.25	11.87	14.58	16.26	13.13	13.92	15.51	15.09	15.51
	SP	9.84	17.71	21.30	8.84	12.97	12.32	11.46	13.56	13.28	13.33
	PvdA	9.64	13.10	12.35	6.59	7.64	10.01	10.33	6.88	7.28	7.11
	CU-SGP	7.86	5.57	7.84	4.22	7.21	9.60	9.32	7.90	7.49	7.77
	GL	7.16	7.85	6.96	5.87	4.47	5.25	5.73	4.77	4.91	4.77
	PvdD	4.32	4.20	4.39	4.59	3.24	2.22	1.44	3.32	3.40	3.40
	50plus	3.78	0.86	1.45	1.80	2.60	3.54	3.58	2.58	2.54	2.52
NL	MAE	0.00	2.66	2.85	2.68	2.26	1.44	1.72	1.91	1.80	1.94
	MSE	0.00	13.76	19.50	12.59	6.28	2.99	4.24	4.91	4.22	5.19
	Tau-a	1.00	0.56	0.56	0.82	0.87	0.87	0.85	0.82	0.87	0.78
Greece (GR)	SYRIZA	26.60	26.82	25.17	28.27	28.60	29.00	29.6	27.72	26.48	27.25
	ND	22.71	22.12	18.54	23.52	25.32	25.50	26.00	25.81	23.89	24.67
	XA	9.38	17.74	27.01	16.02	9.45	9.40	8.00	10.02	9.45	9.06
	Elia	8.02	4.75	8.60	7.82	7.13	7.30	6.50	7.40	8.31	8.10
	Potami	6.61	5.22	7.42	4.12	7.73	7.70	8.00	6.56	10.73	8.07
	KKE	6.07	9.71	8.55	11.91	6.50	6.10	6.00	5.97	6.30	6.16
	ANEL	3.47	11.44	3.36	6.40	4.09	4.00	5.10	4.16	3.63	4.26
	DIMAR	1.21	2.19	1.34	1.93	2.32	2.40	3.20	2.61	1.90	2.72
	Others	15.93	–	–	–	8.85	8.60	7.60	9.75	9.31	9.71
GR	MAE	0.00	4.29	4.30	3.08	1.77	1.79	2.51	1.55	1.50	1.45
	MSE	0.00	22.58	46.06	11.46	7.20	7.85	11.33	5.82	6.98	5.34
	Tau-a	1.00	0.57	0.79	0.79	0.89	0.89	0.82	0.94	0.78	1.00
Average	MAE	0.00	5.36	5.71	2.83	1.57	1.39	1.72	1.42	1.45	1.39
	MSE	0.00	55.19	71.36	11.37	4.81	4.09	5.70	3.88	4.39	3.80
	Tau-a	1.00	0.44	0.42	0.74	0.92	0.89	0.89	0.92	0.88	0.89

As expected, naive methods perform the worst in every country in all terms with an average of 5.36 and 5.71 in MAE respectively, whereas they did not manage to predict not even half of the ranking combinations among the parties in every country (Tau Kendall<0.5). Also, while it was shown that CB2 can provide a boost in accuracy compared to CB1 when trying to adjust the data to polls [8], our findings consistenly contradict this statement compared to the actual results for both error rates ($MAE(CB1) < MAE(CB2)$, $MSE(CB1) < MSE(CB2)$ in all three cases).

S&B method fails to perform competitively with the other approaches presented in the table. One the one hand, this might have been caused due to the different sentiment analysis method that we used (in the original paper, manual annotation was performed which is undoubtedly better); on the other hand, it may also highlight the importance of treating people's voting intentions as time-variant features instead of some static values that we could fit some (Twitter) data in.

Last Week's Polls error values also vary a lot among the three different countries. In Germany, Polls were the best predictors for the final result in terms of MAE. On the contrary, in both Greece and the Netherlands, they performed realtively poorly compared to other poll-based methods. This is an interesting point: despite that our models (TS, PB, NS) were based on polls, they manage to outperform the Polls in both error metrics by using knowledge from the past. Given that every poll comes with a standard error (usually around 3%) along with a certain number of undecided voters, treating polls as time-series (along with other features possibly) seems a better practice. Nevertheless, Polls have the highest Tau Kendall value; however, the differences among most models are minor. From the two prediction websites, MetaPolls outperformed PollWatch and achieved the best MAE on average from all models tested here, along with TS.

Overall, our TS algorithm performed the best in both error rate terms; however it failed to perform equally well in terms of correct ranking of the parties, following by a 0.03 the best competing models in Tau Kendall metric. One possible explanation of this effect is that we were not interested in correctly ranking the political parties but instead predict their voting shares individually; in order to do that, only the features related to an individual party were used for a prediction for this party. Enhancing features of different parties in order to predict each party's voting share is a challenging task for our future research.

The comparison between TS and PB shows that our Twitter features were beneficial. However, the differences in both error rates are rather small. Furthermore, in the case of the Netherlands the PB achieved better results than our TS, whereas the comparison between TS and NS yields the same conclusions. So, despite that our approach achieved the best results overall, the question of whether using our Twitter and sentimental features is actually helpful cannot be answered confidently here.

6 Discussion

Recall that all of our models (TS, PB, NS) were based on a 7-day training window and the average of the predictions by Linear Regression (LR), Gaussian Process (GP) and Sequential Minimal Optimisation (SMO) were reported in Table 3. Both of these decisions (window size, averaging) were taken empirically, since we did not know the actual results. In order to better compare these models, we have applied the same algorithms trained on five different window sizes (starting from one-week with weekly increases of training size up to five-week). In this section, we also provide the results obtained from all individual learners (LR, GP, SMO) inside every model.

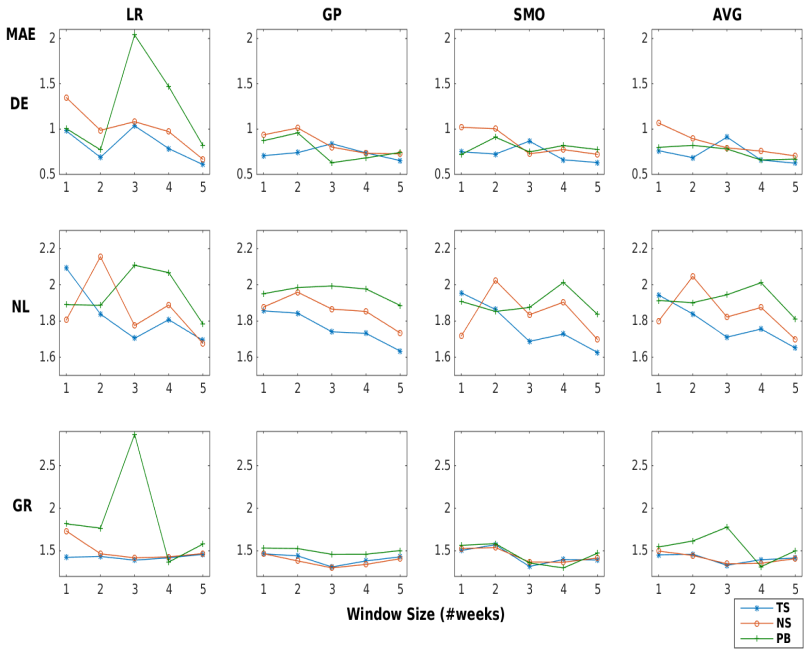
Figure 2 presents the MAE values for all algorithms (including our "averaging" method) when trained on different sets of features (TS, NS, PB) and in different time windows for every country. In the majority of the cases the error drops when we use our full Twitter features; this holds in 64% cases for the individual algorithms (12/15 for LR, 8/15 for GP, 9/15 for SMO) and 67% for the Averaging method(10/15). We also notice that, in most cases, the errors follow a downwards trend as the training window size increases. In fact, had we chosen any such size for our pre-electorial predictions other than the one-week window, we would have achieved the best results among all baselines for the case of the Netherlands as well. However, increasing the training window size does not guarantee that our TS model will perform the best as well.

Despite the domination of TS compared to PB and NS as shown in Figure 2, we still cannot be sure whether there exist *significant* differences between the different models. In order to test our hypothesis (that there do not exist significant differences), we applied a two-step process: first apply the Anderson Darling test to see whether both sets follow a Normal Distribution and then (if so) apply a two-tailed paired t-test or (if not) a Wilcoxon test. We applied this process to all three algorithms using the MAE obtained by every algorithm on every country and training window size (that is, the data points of every algorithm in Figure 2) as our data. Our two pairs of data for every algorithm were the MAE of the pairs (TS, NS) and (TS, PB).

The test between TS and PB revealed that for all three algorithms as well as the "averaging" method, there exist significant differences in MAE for the α level of .05; these results highlight the importance of our Twitter-based features, since by enhancing them into any of our predictive algorithms we can get significantly better results than using only the polls as our features. The similar test in the respective MSEs revealed significant differences for LR and GP but not for SMO, despite that in 9 out of 15 cases the TS performed better than PB in MSE terms for this algorithm as well (see Figure 2, "SMO" column).

The same conclusion does not hold when we compare our TS with the NS model in MAE terms; the differences are statistically significant only when

Figure 2: MAE per training window size for different algorithms and countries.



using LR as our algorithm. This may be due to our naive sentiment analysis methodology; nevertheless, as seen in Figure 2, the TS model achieves better results in the vast majority of cases when compared to NS. Also, the comparison of the respective MSEs revealed that differences are significant for the .025 level for all three algorithms and the averaging method. While our findings indicate that our sentiment analysis method proved statistically important in MSE terms, we plan to test more sophisticated methods for this task in the future, expecting a boost into our TS model.

7 Conclusion

Our work focused on predicting the 2014 EU elections for three countries using Twitter. Working on time-series and using opinion polls as our ground-truth, we extracted several text- and user-based features from political tweets and trained three different algorithms on them. Our results demonstrate the appropriateness of our method in error rate terms, which achieved better results than several baselines, including polls, prediction websites and replication of previous works. Most importantly though, we demonstrated that by enhancing our twitter-based features into poll-based ones we can a statistically significant boost in MAE rates, whereas our methodology was developed before the end of the elections, avoiding any bias towards the

actual results.

Future work includes working on a wider time period in order to find an appropriate training window size, incorporating network-based features for the users, features from different parties in order to predict each party's voting share and a more appropriate method for sentiment analysis, by manually labeling some political tweets for every language and classifying the test data with an in-domain learned model. Furthermore, we have aggregated data from the pages of the political parties on Facebook and we will try to exploit the use of this Social Network for the same task as well. Finally, given that we have enough data for this purpose, we plan to fit a model to the actual results for one country and test it on the others, in an attempt to overcome the need of finding opinion polls for training.

Acknowledgements

This work is supported by the SocialSensor FP7 project, partially funded by the EC under contract number 287975.

References

[1] Jessica Elan Chung and Eni Mustafaraj. Can collective sentiment expressed on twitter predict political elections? In *AAAI*, 2011.

[2] Daniel Gayo-Avello. A meta-analysis of state-of-the-art electoral prediction from twitter data. *Social Science Computer Review*, page 0894439313493979, 2013.

[3] Vasileios Lampos, Daniel Preotiuc-Pietro, and Trevor Cohn. A user-centric model of voting intention from social media. In *Proc 51st Annual Meeting of the Association for Computational Linguistics*, pages 993–1003, 2013.

[4] Panagiotis Takis Metaxas, Eni Mustafaraj, and Daniel Gayo-Avello. How (not) to predict elections. In *Privacy, security, risk and trust (PASSAT), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (SocialCom)*, pages 165–171. IEEE, 2011.

[5] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *ICWSM*, 2013.

[6] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11:122–129, 2010.

[7] Jonathon Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, pages 43–48. Association for Computational Linguistics, 2005.

[8] Erik Tjong Kim Sang and Johan Bos. Predicting the 2011 dutch senate election results with twitter. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 53–60. Association for Computational Linguistics, 2012.

[9] Lei Shi, Neeraj Agarwal, Ankur Agrawal, Rahul Garg, and Jacob Spoelstra. Predicting us primary elections with twitter. In *Proceedings of Social Network and Social Media Analysis: Methods, Models and Applications (NIPS Workshop), Lake Tahoe, NV, December*, volume 7, 2012.

[10] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.