FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

PoliticAnalytics: social media analytics for political data science

Luís Filipe Castanheira Gomes

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Carlos Soares

Co-Supervisor: Pedro Saleiro

February 7, 2014

PoliticAnalytics: social media analytics for political data science

Luís Filipe Castanheira Gomes

Mestrado Integrado em Engenharia Informática e Computação

February 7, 2014

Abstract

In this project we studied the problem of predicting the results of political polls based on the combination of several aggregators of buzz and sentiment obtained from Twitter posts. We followed a machine learning approach, where the combination model was estimated from data. We used tweets from the portuguese tweetosphere since June 2011. This dataset contains tweets from 100 000 users who were classified as Portuguese. Futhermore, we had access to the polls results, since June 2011, of a private portuguese company that studies portuguese public opinion (Eurosondagens). We performed some experiments using two regression algorithms (Random Forests and Ordinary Least Squares). We compared the real poll values with the predicted values of our regression models. The lower absolute error we could obtain was 0.50 using only buzz aggregators. It means that our prediction model has a small predictive error. These results highlight the potential of using twitter data to complement or substitute traditional surveys.

Resumo

Neste projecto estudámos o problema de prever o resultado de sondagens políticas com base na combinação de vários agregadores de *buzz* e sentimento obtidos de mensagens do *Twitter*. Seguimos uma abordagem de *machine learning* para estimar o modelo de combinação através dos dados recolhidos. Utilizámos *tweets* da tweetosfera portuguesa desde Junho de 2011. Este *dataset* contém *tweets* de 100 000 utilizadores diferentes classificados como Portugueses. Para além disso, tivemos acesso ao resultados das sondagens, desde Junho de 2011, levadas a cabo por uma empresa privada que estuda a opinião publica portuguesa (Eurosondagens). Foram feitas experiências usando dois algoritmos regressão (*Random Forests* e *Ordinary Least Squares*). Comparámos o resultado real das sondagens com os valores previstos pelos nossos modelos regressivos e anotámos os resultados. O menor erro absoluto obtido foi 0.50 usando apenas agregadores de *buzz*. Isto significa que o nosso modelo predictivo tem um pequeno erro predictivo. Estes resultados realçam o potencial existente em usar dados do *Twitter* para complementar ou substituir as sondagens tradicionais.

Acknowledgements

I would like to show my gratitude to Carlos Soares and Pedro Saleiro, for the patience at guiding me, for all the time they spent and for all the knowledge they share with me.
To SAPO Labs, for providing me a workstation and a grant, so I could have all the conditions to develop this project.
To all my friends who were by my side over these years with whom I shared joys and sorrows.
To Vanda, my girlfriend, for her love, support and patience, and for cheering me up in every single moment.
To my family, for their patience and support, even in the nights and days that I was not at home and for providing all the conditions so I could conclude this course.
This work is partially funded by National Funds through the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within projects "REACTION (UTAustin/EST-MAI/0006/2009)" and "POPSTAR (PTDC/CPJ-CPO/116888/2010)" as well as Project "NORTE-07-0124-FEDER-000059", which is funded by the North Portugal Regional Operational Programme (ON.2 – O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF), and the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT).

Luís Filipe Castanheira Gomes

Contents

List of Figures

LIST OF FIGURES

List of Tables

Abbreviations

MAE	Mean Absolute Error
NLP	Natural Language Processing
MSE	Mean Squared Error
OLS	Ordinary Least Squares
RF	Random Forests

Chapter 1

Introduction

Before the internet existance, traditional surveys and polls were the only methods to provide information about what people thought about parties or political personalities [JSS]. These methods use the telephone to collect the opinions about political targets. Surveys randomly select the electorate sample, avoiding selection bias, and are designed to collect the perception of a population regarding some subject, such as public opinion or brand experience. However, this method is expensive and time consuming [CBRS, JSS]. Furthermore, over the years it is becoming more difficult to contact people and persuade them to participate in telephone survey, as is evidenced in figure 1.1 [KKD+12].

Surveys Face Growing Difficulty Reaching, Persuading Potential Respondents

	1997 %	2000 %	2003 %	2006 %	2009 %	2012 %
Contact rate (percent of households in which an adult was reached)	90	77	79	73	72	62
Cooperation rate (percent of households contacted that yielded an interview)	43	40	34	31	21	14
Response rate (percent of households sampled that yielded an interview)	36	28	25	21	15	9

PEW RESEARCH CENTER 2012 Methodology Study. Rates computed according to American Association for Public Opinion Research (AAPOR) standard definitions for CON2, COOP3 and RR3. Rates are typical for surveys conducted in each year.

Figure 1.1: Evolution of contact, cooperation and response rates of traditional surveys

However, social networks, blogs and online forums have turned the internet into an information and opinions repository generating a big amount of data that can be used in scientific research [TBP11]. Thus, researchers started their work in order to understand how social media data can be used in political scenario. People express their opinions or political leanings on social media for free. Thus, why cannot we collect those opinions and try to understand if people are happy with

their government? Moreover, can opinions collected from the social media be used as predictors in the election processes?

With the raise of social media, namely Facebook[1] and Twitter[2], people share online their thoughts and opinions about political targets [BS]. Twitter has roughly 288 million users accessing the site at least once a month [Hol13a] creating 5 787 new tweets every second [Hol13b]. Part of this information is publicly available for free. Additionally, changed the relationship between news and people. Blogs, Facebook and Twitter allow people to consume news and express their opinions about related entities almost in real time. Thus, social media itself has becoming an important source of information both to journalists and politicians.

Furthermore, there is a growing number of studies suggesting that social media messages can be relevant indicators of public opinion and, in some cases, lead to the prediction of election results.

Studying the relationship between social media messages and traditional polls is an active area of research. Due to the amount of data available online, the most conventional way to extract some meaningful information is using machine learning and other data analysis techniques. Approaches in this area usually create predictive models using aggregated social media data as input signal and polls results as target variable. However, there is no consensus on how to aggregate social media data, i.e., how we can predict polls results using raw messages mentioning specific political entities. The most common approaches aggregate the frequency of mentions [TSSW10] or the obtained sentiment of those messages, using a sentiment analysis classifier [MMGA11].

1.1 Objectives

The main objective of this dissertation is to study and define a methodology that allows us to aggregate social media data in order to predict polls results.

We collected buzz data from social media and apply sentiment methods used in literature. However, there is no consensus on how to aggregate the social media data or which aggregation methods should we use. We combined several aggregation methods and used machine learning techniques in order to predict poll results. We propose to evaluate the suitability of the model using data from Portuguese Politics.

This thesis was carried out as part of an FCT funded project, POPSTAR, which aims to study the relationship between online mentions of Portuguese political targets and the results of traditional polls.

1.2 Document Structure

The document is organized as follows:

[1] www.facebook.com
[2] www.twitter.com

Chapter 2 contains the background and the state of the art as well as some related work. We explain in more detail what data mining is as well as regression and classification methods. Furthermore we present a list of opinion aggregation methods. Chapter 3 contains the methodology we used in the portuguese case study. We describe step by step the data transformation so the methodology can be used in different domains. Chapter 4 we present the results of our study and some discussion. In chapter 5 we have some conclusions as well as ideas for future work.

Chapter 2

Background and State of the Art

2.1 Data Mining

Every day the amount of data in the world and in our lives increases in an endless way. Internet provides us a big amount of information on every topic. But what can be done with that data? Usually, the collection of data by itself do not solve any problem or answer our questions. With the increase of information amount all over the world, it became important to create processes able to use information the most meaningful way. One approach is using data mining techniques. Basically, data mining is the technique of extracting knowledge from large amounts of data [HK]. The best practice to extract information correctly from our data set, we need to follow six steps [HK].

1. **Data cleaning**: This step is essential to remove some noisy or inconsistent data;
2. **Data integration**: It might be necessary to use data from different data sources;
3. **Data selection**: One priority of this step is to define •which data is relevant to our study and data we may discard;
4. **Data transformation**: In this step the preparation of the relevant data is performed, so it can be used in the next step;
5. **Data mining**: Application of the intelligent methods in order to extract knowledge;
6. **Evaluation**: After knowledge has been extracted, we need to evaluate it. In this step we need to know what those results really mean, extract some interesting patterns, some interesting behaviours, etc. Without a correct perform of this step, data can be meaningless.

After these steps, and as well as in any research work, we need to present the results and some conclusions we might consider interesting.

Data mining can be used in many research areas. In medicine, for instance, there is the need to analyze patients data to achieve the best diagnosis. In a retail company we need to analyze the

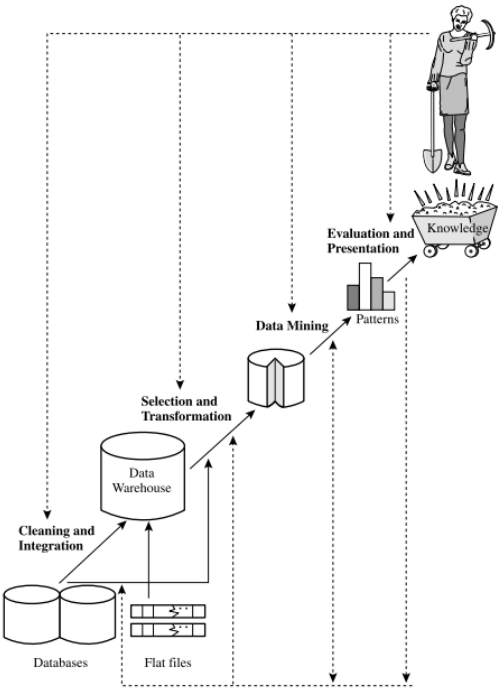

Figure 2.1: Steps to retrieve knowledge from data [HK]

sales to better define marketing strategies. In our case, we need to analyze the social media data on the Portuguese political targets to understand how can it be used to complement or substitute the traditional polls.

2.1.1 Application of Data Mining to Politics

A big effort is being made in order to understand how data mining can be used in politics. Social media, concretely Twitter, can be used extensively for political deliberation. Next, we enumerate some applications where Twitter data was analyzed in politics.

- [JSS] concludes that more than predicting elections, social media can be used to gauge sentiment about specific event (such as political news or political speeches).
- Twitter data can also be used to collect the overall sentiment of a real event (such as political speech, debate, or news) when it is still happening [DS10]. They studied the sentiment variation during a Obama vs McCain TV debate.
- Another interesting study states that "the mere number of party mentions accurately reflects the election result", once it accurately predicted the 2009 Federal Election in Germany [TSSW10].
- In their study, [CBRS], used a sentiment score based on positive and negative twitter messages, showing that text sentiment has a high correlation with polls, which means that text sentiment is a leading indicator of polls.

- The 2011 Irish General Election was correctly predicted using Twitter data. [BS] states tath "both volume-based measures and sentiment analysis are predictive". In summary, they "conclude that Twitter does appear to display a predictive quality".

- Supervised learning techniques were used to learn the public opinion on Obama. They used one billion twitter messages posted over 2008 and 2009, the Gallup's daily tracking poll for presidential job approval and a set tracking polls during the U.S. presidential election cycle as training data [CBRS].

- [BS] also used supervised machine learning techniques once they also collect data from twitter and the results of nine polls which were commissioned during the election.

2.2 Machine Learning

Collecting social media data, by itself does not solve any problem. We need to know how to interpret data we collected. Data mining consists in apply some intelligent methods in order to extract knowledge from data [HK]. Machine learning is all about develop methods for that porpuse. Machine learning is aimed to find and describe "structural patterns in data, as a tool for helping to explain that data and make predictions from it" [WFH11].

We can divide machine learning in two main groups: Supervised and unsupervised learning. Supervised learning, which we will use in this dissertation work, is used when the class label of data set tuples is provided.

The two most common problems within supervised learning are classification and regression problems.

2.2.1 Classification

In their quotidian, companies have the need to make decisions. Banks for instance, have the constant need to know which ones of the loans applicants are "risky" or not. Another emergent market is the betting market. This market consists in predict if a team will win, lose, or draw a given game. In all these cases, classification methods are used.

Basically, classification methodologies are used to predict a categorical label [HK], such as "safe" or "risky", for the bank loans, and "draw", "win" or "lose" for the team result prediction.

One classification method is using **Decision Trees**. "A decision tree is a flowchart-like tree structure, where each internal node (non leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label." [HK]. Figure 2.2 is an example of a classification decision tree. Another method of classification is **classification rules**. In this case, a ruled-based classifier uses a set of IF-THEN rules [HK]:

IF condition THEN conclusion

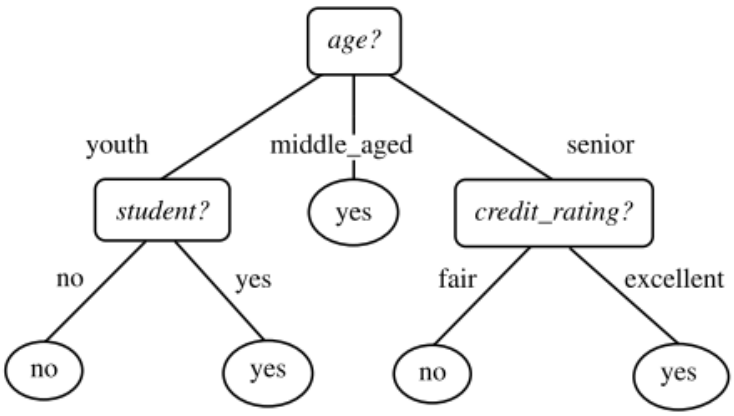

Figure 2.2: Example of a Decision Tree [HK]

Classification problems can also be solved using the **Naïve Bayes Classifier**, which is a classification algorithm based on **Bayes' Theorem**. This method, calculates the probability of an event E to occur, given that H has already happened [HK] and takes the form of the equation 2.1. In other words, **Naïve Bayes Classifier** calculates the probability of an output o to happen, given that an input i. This allows us to extract patterns from the data, which facilitates the classification.

$$P(E|H) = \frac{P(H|E)P(E)}{P(H)} \tag{2.1}$$

Classification methods can be used in several professional activities.

Classification can be used to determine if the overall sentiment within a message is positive, negative or neutral. Concretely, [JSS] tried to classify the political sentiment of tweets related to Obama. One of the used approaches was the Maximum Entropy algorithm. Maximum Entropy, which is a supervised learning algorithm, is based on making as few assumptions as possible, other than the constraints imposed [CN02]. Those constraints are determined based on the the training data [CN02, JSS]. The maximum entropy model used by [JSS] is based on by the following formula:

$$P(c|t,\lambda) = \frac{exp[\sum_i \lambda_i f_i(c,t)]}{\sum_c' exp[\sum_i \lambda_i f_i(c,t)]} \tag{2.2}$$

Where c is a given class where object t fits (*positive* or *negative*), λ is the learned weight vector and $f_i(c,t)$ is the ith feature associated with the tweet. Thus, the aim of the Maximum Entropy algorithm is to learn the λ weights from the training data that maximize the conditional probability [JSS] to better classify a given tweet t.

Another approach in the classification area is using some message components (such as hashtags and URL's) to classify each user according to his political leaning, *i.e.* determine which party each user belongs to [BKY12]. To achieve satisfactory results, [BKY12] used the Bayesian

Classification Algorithm. Bayesian Algorithm is aimed to calculate the probability of X to occur conditioned on H ($P(X|H)$) [HK]. Given that, Boutet [BKY12] calculates the probability $M_k^i(u)$ of a user u to be part of a party i, given that he took action k, $A_k(u)$ (retweet a tweet from a user who belongs to a party y or mention another user on his tweet), given by:

$$P(M_k^i(u)|A_k(u)) = \frac{P(A_k(u)|M_k^i(u))P(M_k^i(u))}{\sum_j P(A_k(u)|M_k^j(u))P(M_k^j(u))} \tag{2.3}$$

Where k is the number of the action.

As told before, each user can retweet a tweet or refer a political party in their tweets. Thus, a clustering analysis was made to characterize the political communication network on Twitter [CR11]. The analysis shows that there are two distinct communication networks. The first one is the retweet network and the second one the mentions network. With this approach, came the results that "users preferentially retweet other users with whom they agree politically, while the mention network appears to form a bridge between users of different ideologies." [CR11]

2.2.2 Regression

In some particular cases, to know if a team will win or lose might not be enough. We might need to predict how may goals will a team score. In a retail company, for instance, it should be possible to predict the sales volume for a given semester, given the past ones. In the political scenario, we might want to predict the vote percentage that a political target will achieve. In these cases, we use regression models. In summary, regression models are "used to predict missing or unavailable numerical data values rather than class labels" [HK]. In other words, after we have a measurements vector x as input, we want the output to be a numeric value y with a concrete significant (such as the number of goals a team will score or the sales volume of a company).

There are some regression models that can be used to the prediction problem. A simple model is **Linear Regression**, which involves a response variable y, and a single predictor variable x, given by the following formula:

$$y = b + wx \tag{2.4}$$

b an w are regression coefficients. The coefficients can also be thought of as weights, that is, the weights of the predictor variables on the model [HK]. Thus, it is necessary to calculate those weights, written by the following:

$$y = w_0 + w_1 x \tag{2.5}$$

So, both w_0 and w_1 can be calculated using the method of **least squares**. This method estimates the best fitting straight line as the one which minimizes the difference between the real data and the predicted one. Given a training data set D containing |D| data points $(x_1, y_1), (x_2, y_2), ..., (x_{|D|}, y_{|D|})$:

$$w1 = \frac{\sum_{i=1}^{|D|}(x_i - x_p)(y_i - y_p)}{\sum_{i=1}^{|D|}(x_i - x_p)^2} \tag{2.6}$$

where x_p and y_p are the predicted values.

However, in some cases we do not have just one predictor variable. It might be necessary to consider more than one predictor variable. In these cases, we have to perform a **Multiple Linear Regression**, which takes the form:

$$y = w_0 + w_1 x_1 + w_2 x_2 \tag{2.7}$$

In this case, we can still use the method of **least squares** to solve this problem.

A regression model has also been used to achieve the election outcome on each party, based on the relevant tweets of the five main Irish parties, on the Irish General Election, 2011. In that study, some sentiment measures were been defined (we will discuss them in detail later) to use as input for the regression model [BS].

Linear least-squares model was used to correlate data collected from twitter related to Obama with two set of polls during the 2008 U.S. presidential election. In this particular case, it was concluded that when consumers confidence changes, it can be noticed first in the text sentiment measure than in polls, which means that text seems to be a leading indicator [CBRS].

However, as told before, regression models can be applied in other professional areas. Actually, there are a some studies on the cinematographic industry that use regression models to predict a movie revenue before its release [AH10, JDGS10, ZS09]. For that purpose, it was built a linear model using the average of the tweet-rate related to a film, in the weekend prior the release, for 24 movies and they conclude that the tweet-rate is a strong indicator of the revenue of a film.

Another work that used regression models is [JDGS10], once they used it to directly predict opening weekend gross earnings, based on some features extracted from the movie metadata (such as whether the film is from, running time, genre, actors or whether the movie opened on a holidays weekend or in summer months.) This information extracted from text was made using three different approaches. (1) n-grams: considering n-grams, bigrams and trigrams. N-grams is a contiguous sequence of n items from a given sequence of text. Bigrams and trigrams were only considered if all the words were included in the 25-word stoplist; (2) part of speech n-grams, using the [TM00] tagger; (3) Dependency relations [KM02].

2.2.3 Evaluation

An important issue in Machine Learning is evaluation. There are several evaluation measures that can be used to measure how good and reality adjusted a model is. We have distinct evaluation measures for classification and regression problems.

One of the measures we can use to evaluate a classifier is **accuracy** [HK]. **Accuracy** represents the percentage of test set that are correctly classified by the classifier. An example of usage of accuracy to measure the performance of their sentiment classification model is given by [JSS].

Other evaluative measures that we can use to evaluate the performance of a classification model are **recall** and **precision**. While **recall** measures the percentage of relevant documents that are correctly classified as relevant by the model, **precision** measures the percentage of documents classified as relevant that actually are relevant [RL94].

On the other hand, we have others fitted to to determine the performance of a regression model.**MAE** (Mean Absolute Error), for instance, is used to know how close the forecast is to the real outcome [BS, HK], and is defined by the following formula:

$$MAE = \frac{\sum_{i=1}^{n} |f_i - y_i|}{n} \tag{2.8}$$

where n is the number of forecasts, f_i is the model forecast and y_i the real forecast.

As the name suggests, the Mean Absolute Error is the average of the absolue errors of each prediction:

$$e_i = |f_i - y_i| \tag{2.9}$$

where f_i is the prediction and y_i is the real outcome.

R^2, also called as coefficient of determination, reflects how strong the variables used as input in the regression model are as predictors [AH10, CBRS]. R^2 varies between 0 and 1. The bigger R^2, the more predictive the measures are [AH10].

2.3 Sentiment Analysis

Sentiment analysis, as well as opinion mining, has enjoyed a huge burst of research activity [PL08]. The rise of machine learning methods in Natural Language Processing (NLP) and information retrieval as well as the availability of data sets for machine learning algorithms to be trained on, contributed to this burst. Sentiment analysis "deals with the computational treatment of opinion, sentiment, and subjectivity in text" [PL08]. In summary, sentiment analysis aims to extract the overall sentiment or opinion from a given text or message. However, with the complexity of the languages syntaxes, this is not a simple task.

In the political scenario, sentiment analysis is a useful tool once there is the constant need of extracting the overall sentiment inherent to a message.

One possible sentiment analysis approach is using lexicons. The lexicon approaches only consider a certain word if it is contained in a pre existing list lexicons. This lexicon list, also contains the information about the positiveness or the negativeness of each word. [CBRS, BS] used the lexicon based approach. [CBRS] classified each tweet as positive or negative according to the existence or not of positive and negative words, respectively. This means that a given message can be both positive and negative if it contains both positive and negative words.

The same approach was used in [MMGA11] with a little modification. In this particular case, each positive and negative word within a twitter message counts as +1 and -1 respectively. Thus,

the message is positive, neutral or negative depending on the sum of all the labelled words it contains.

Another approach to the sentiment classification problem, is using supervised learning algorithms. A classifier can be trained using a data set. [JSS] used a Maximum Entropy classifier to classify the political sentiment of the tweets containing the word "Obama".

2.3.1 Opinion Aggregators

An important decision to make is what to do with labelled data, once labeling messages as positives, negatives or neutrals might be, by itself, meaningless. One challenge on the political science is to create a method to aggregate data the most meaningful way in order to raise the accuracy of our model. Many authors have tried to develop mathematical formulae to quantify the popularity of a given political target or party, based on twitter data, called aggregator. The aggregator, is a numerical value based on twitter messages annotated as positive, negative or neutral by a sentiment analysis method.

We can have two different approaches when aggregating data to include in our predictive model: sentiment [JSS, BS, MMGA11, CBRS] and buzz [MMGA11, TSSW10, BS]. Buzz is merely the frequency that a political target is mentioned in social networks.

The following equation represents a simple aggregation method using sentiment analysis [CBRS].

$$x_t = \frac{count_t(pos.word \wedge topicword)}{count_t(neg.word \wedge topicword)} \tag{2.10}$$

x_t is the sentiment score on day t as the ratio of positive versus negative messages on the topic.

[MMGA11] used an interesting approach based on sentiment analysis. In a two candidate race, $c1$ and $c2$, all negative mentions refered to $c1$ will count as positive for $c2$. Assuming that, the vote share for $c1$ can be calculated as the ratio of the sum of the number of positive messages of $c1$ and negative messages of $c2$, over the total of positive and negative messages.

$$vote\_share(c1) = \frac{pos(c1) + neg(c2)}{pos(c1) + neg(c1) + pos(c2) + neg(c2)} \tag{2.11}$$

However, in a (at least) 3 parties race, we cannot assume that a negative tweet related to $c1$ will automatically represent a positive one for $c2$.

Generally, if the majority of the mentions to a party are negative, we can automatically assume that people are in general negatively disposed towards that party. But what if the negative majority holds true for all parties? To solve that problem, [BS] purpose two models: inter- and intra-party. To better classify each party according to the others, [BS] calculate the share of positive and negative volume, *berminghamsovp* and *berminghamsovn*.

There are also some aggregators that use buzz. [TSSW10] states that "the mere number of party mentions accurately reflects the election result". [MMGA11] also experimented the use of

buzz as predictor to compare the results against using sentiment. Buzz indicators are all the ones that do not need sentiment analysis, such as the mentions total (buzz) and buzz share, of each candidate. Besides using the sentiment analysis to aggregate collected data, [BS] also tries buzz as a predictor (*share*).

Table 2.1 presents the complete list of all aggregators collected from the state of the art, and their formulas. There are buzz and sentiment aggregators. Some sentiment aggregators have been normalized. It means that normalized aggregators instead of using the number of positive, negative and neutral tweets, used the ratio of positive, negative and neutral tweets over the candidate buzz.

Table 2.1: List of implemented aggregators

Name	Description	Formula
bermingham2 [BS]	-	$\log_{10} \frac{related\_pos+1}{related\_neg+1}$
normalized_bermingham2	Normalization of the previous formula	$\log_{10} \frac{\frac{related\_pos}{candidate\_buzz}+1}{\frac{related\_neg}{candidate\_buzz}+1}$
berminghamsovn [BS]	Share of negative buzz	$\frac{related\_neg}{total\_neg}$
berminghamsovp [BS]	Share of positive buzz	$\frac{related\_pos}{total\_pos}$
connor [CBRS]	Ratio between positive and negative tweets	$\frac{related\_pos}{related\_neg}$
normalized_connor	Normalization of the previous formula	$\frac{\frac{related\_pos}{candidate\_buzz}}{\frac{related\_neg}{candidate\_buzz}}$
gayo [MMGA11]	-	$\frac{related\_pos+total\_neg}{total\_pos+total\_neg}$
normalized_gayo	Normalization of the previous formula	$\frac{\frac{related\_pos}{candidate\_buzz}+normalized\_total\_neg}{normalized\_total\_pos+normalized\_total\_neg}$
ind	-	$\frac{related\_neg+related\_pos}{candidate\_buzz}$
polarity	-	$related\_pos - related\_neg$
normalized_polarity	Normalization of the previous formula	$\frac{related\_pos}{candidate\_buzz} - \frac{related\_neg}{candidate\_buzz}$
polarityONeutral	Polarity over neutrality	$\frac{related\_pos-related\_neg}{related\_neutral}$
polarityOTotal	Polarity over total	$\frac{related\_pos-related\_neg}{candidate\_buzz}$
subjNeu	Subjectivity over neutrality	$\frac{related\_pos+related\_neg}{related\_neutral}$
subjSoV	Share of subjectivity	$\frac{related\_pos+related\_neg}{total\_pos+total\_neg}$
subjVol	Subjectivity	$related\_pos + related\_neg$
share [BS]	Buzz Share	$\frac{candidate\_buzz}{total}$
shareOfNegDistribution	Share of negative distribution	$\frac{\frac{related\_neg}{candidate\_buzz}}{\sum_{i=0}^{n} \frac{pos_i}{buzz_i} + \frac{neg_i}{buzz_i}}$
normalized_negative	Normalization of negative buzz	$\frac{related\_neg}{candidate\_buzz}$
normalized_neutral	Normalization of neutral buzz	$\frac{related\_neu}{candidate\_buzz}$
normalized_positive	Normalization of positive buzz	$\frac{related\_pos}{candidate\_buzz}$
positive_mentions	Positive buzz	$related\_pos$
negative_mentions	Negative buzz	$related\_neg$
neutral_mentions	Neutral buzz	$related\_neu$
total_mentions	Candidate Buzz	$candidate\_buzz$

Chapter 3

Methodology and Case Study

3.1 Problem

Nowadays the most used and common way to collect public opinion surveys is through telephone surveys. Surveys randomly select the electorate sample, avoiding selection bias. The questions are designed and tested to measure whatever we want to measure. However, this method is time consuming and expensive [CBRS, JSS]. Another obstacle to telephone surveys is the low response rate. It is becoming more difficult to contact people and persuade them to participate in telephone survey (from 1997 to 2012 the response rate dropped from 36% to 9%) [KKD+12].

The big challenge of this dissertation is to understand how public opinion expressed within twitter messages can be used as a predictor of political target popularity.

3.2 Goals

The main objective of this dissertation is to study and define a methodology that allows us to aggregate social media data in order to predict poll results using a regression model.

Another objective is to understand which type of data aggregator (buzz, sentiment, or both) are more suitable to predict poll results.

To perform the study, we apply two regression algorithms (Random Forests and Ordinary Least Squares) to predict poll results. Its results are validated based on the minimization of the prediction error.

3.3 Data Sources

To study the relationship between portuguese electorate opinions on social media and poll results we use two main distinct data sets: (1) Social media data and (2) Portuguese poll results. The social media data set may contain data collected from three different sources: Twitter, blogs and news. With respect to this case study, the social media data set does not need to contain the

message/news/text itself. The social media data set contains the daily count of positive, negative or neutral message/news/text referring to each political target or party.

The second main data set contains the monthly polls result of the political targets or parties we want to predict.

3.4 Data Preparation

As described in section 3.3, we have access to the political parties daily count of positive, negative and neutral messages/news/text. Futhermore we have access to monthly polls results. Polls results would act as target variable in our regression model.

At this point the need of monthly aggregate the daily count emerged. If we used daily counts as predictive features, we would be trying to predict daily vote intention. Due to the lack of daily poll results to be used as target variable, it would not possible to determine the predictive effectiveness of our model, given that we had no real results to compare our prediction with. On the other hand, monthly aggregated data allowed us to use polls results to compare with our vote intention prediction and determine the prediction error. After aggregating monthly data, we are able to apply the opinion aggregators described in section 2.3.1.

The next step is joining monthly aggregated data and real polls results in the same data set. This allows us to store political parties monthly count of positive, negative and neutral messages, values of the opinion aggregators, and polls results on the same data set. Furthermore, we include in the aggregated data set the value of the previous month polls result, for each candidate, which we called y_{t-1} feature.

Each political target poll results has a small variation from month to month. Thus, we decided to use a different approach. Instead of using absolute values, we transform both polls and aggregators data sets to carry the monthly variation from the previous month. In a mathematical notation, an aggregator i in month m would take the value $agg(i)_m - agg(i)_{m-1}$. It allows the model to predict the poll monthly variation instead of predicting the absolute polls result.

Having these two data sets allows us to perform two kind of experiments: (1) Using absolute values and (2) using monthly variations.

3.5 Architecture

Figure 3.1 represents the architecture of our vote intention prediction model prototype.

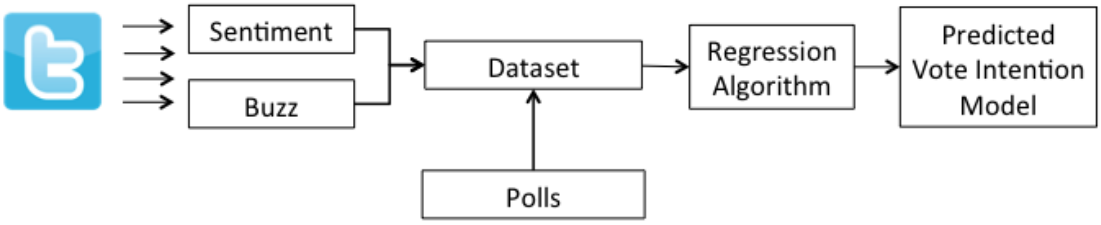

Figure 3.1: PoliticAnalytics' Solution Architecture

This methodology can support multiple data sources (twitter, blogs and news). With respect to this particular case study, we only used one of the three possible data sources: Twitter. After the correct preparation of this data set, we use the second main data set: Polls results. This second data set is defined as our target variable in the regression model. After data preparation, we can use the dataset as input in our prediction model. We apply a regression algorithm and predict the monthly vote intention for each candidate.

Chapter 4

Results

4.1 Overview

This chapter is reserved to describe in detail the data sources we used as well as to explain the data preparation. Ultimately, we describe in detail all the experiments we performed as well as their results.

We performed multiple experiments in order to predict poll results from January 2013 to June 2013 with low MAE. Thus, we used different two different regression algorithms combined with different data sets.

4.2 Experimental Setup

4.2.1 Data

Although the methodology defined in chapter 3 can support multiple data sources, in this project we center our efforts on Twitter. POPSTAR project, which collect and label twitter messages as positive, negative or neutral using sentiment analysis methods, provided us a data set containing the daily count of labeled twitter messages with respect to national and international public figures, since June 2011. In our case study, it was not necessary to access the messages themselves. The second main data set contains the mothly polls result. We had access to the polls results from Eurosondagens, provided by the Instituto de Ciências Sociais da Universidade de Lisboa, also as part of the POPSTAR project. Eurosondagens is a private portuguese company that studies public opinion and performs monthly polls on public political preference. This data set contains the traditional monthly polls results of the five main portuguese political parties (PS, PSD, CDS, BE and CDS), from June 2011 until June 2013.

Both twitter and polls data are stored in a MongoDB database. MongoDB is an open-source document database, and the leading NoSQL databases. The programming language we used was Python, and its scientific libraries scipy, numpy and scikit-learn.

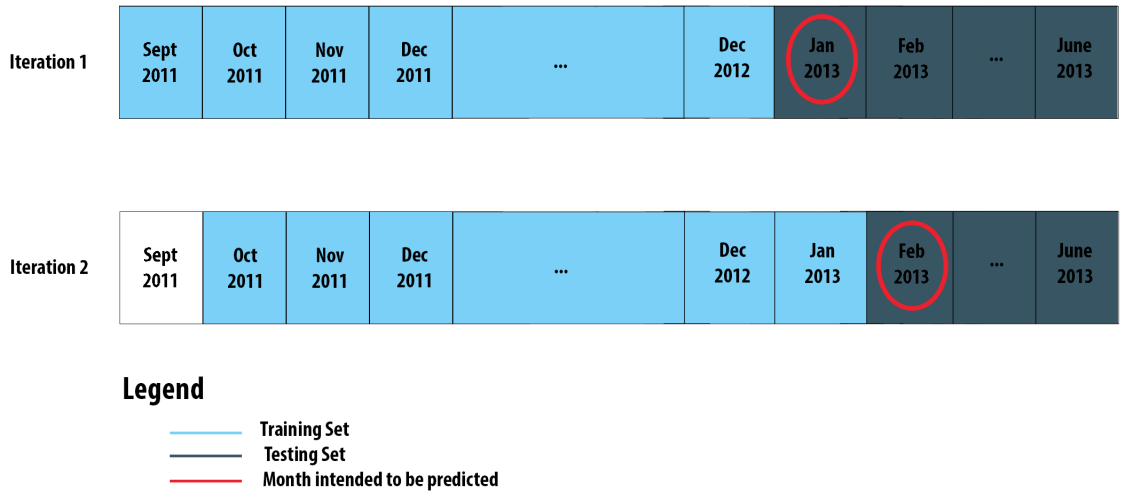

Figure 4.1: Two iterations of the sliding window technique method.

As described in section 3.4, we address the problem using two different approaches: (1) a data set containing the monthly absolute values and (2) another one containing monthly variation values.

To estimate the performance of the models, we use a sliding window technique method:

- Training set, which contains the values of the aggregators for 16 months prior the month we want to predict.
- Testing set, containing the aggregators values from January 2013 to June 2013.

The prediction process is systematic. In the first iteration of the predictive process we try to predict the poll results for January 2013. (1) Thus, we train the predictive model with the training set, containing the 16 months data prior the month we want to predict. (2) We provide the values of the independent variables of the month we want to predict as input to the trained model, to obtain a prediction poll results. (3) We select the next month of the testing set and repeat the process until all months are predicted.

In other words, the training set can be seen as a time window that has always the same size (16 months). In the first iteration of the predictive proccess, we are predicting poll results of January 2013. In this case, the training set contains data from August 2011 to December 2012. Following the same reasoning, when predicting the poll results of February 2013, the training set contains data from September 2011 to January 2013. In figure 4.1 are represented two iterations of the prediction process.

After predicting the poll value for each month, we calculate the average of Mean Absolute Error (MAE) and Mean Squared Error (MSE) between the predicted values and the real poll values. However, we only used the MAE to measure the predictive effectiveness of the models. Thus, in the experiments section we only present tables with the MAE.

4.2.2 Experiments

As described in section 3.4 we create a dataset containing the monthly variation values. It allows us to perform two different kind of experiments: (1) using the aggregators absolute monthly values and (2) using the monthly variation of those values.

We performed a total of four experiments, using different target variables:

- In the first experiment we used the absolute monthly values: $y \leftarrow \{y_{t-1}, buzzAggregators, sentimentAggregators\}$.
- In the second experiment we used the monthly variations: $\Delta y \leftarrow \{\Delta(y_{t-1}), \Delta buzzAggregators, \Delta sentimentAggregators\}$
- To perform experiment *Buzz vs Sent* we used buzz and sentiment data separately:

  $\Delta y \leftarrow \{\Delta(y_{t-1}), \Delta buzzAggregators\}$

  $\Delta y \leftarrow \{\Delta(y_{t-1}), \Delta sentimentAggregators\}$
- In experiment *ExpBuzzSent All* we included the previous poll variation of all candidates:

  $\Delta y \leftarrow \{\Delta(all\_y_{t-1}), \Delta buzzAggregators\}$

  $\Delta y \leftarrow \{\Delta(all\_y_{t-1}), \Delta sentimentAggregators\}$

All the aggregators values are used given that we did not perform any feature selection technique, except in the experiments *Buzz vs Sentiment All* and *Buzz vs Sentiment All*, where we use buzz and sentiment aggregators separately.

All the models are obtained using regression algorithms. Furthermore, in all experiments we performed we use both Ordinary Least Squares and Random Forests algorithms.

To validate the effectiveness of our regression model, we use a naive baseline: the polls result for month m_i is equal to m_{i-1}. As explained in section 4.4, there is a small polls results variation along the time. Thus we believe this premise is a strong baseline, with small prediction error.

Different experiments allow us to compare which aggregator (or list of aggregators) performs the best.

4.3 Twitter Data

The twitter data sample provided by the POPSTAR project contains the daily count of positive, negative and neutral messages of national public figures. However, we filter the data set and use only the data referred to the leaders of the five main political targets (PS, PSD, CDS, CDU and BE). It contains 232 979 annotated tweets, from 100 000 different twitter users. The complete data set information can be seen on table 4.1.

Table 4.1 helps us to take some important notes:

- The negative mentions represent the majority of the total mentions, except for CDU where the number of negative mentions is smaller than the neutral ones.

Table 4.1: Distribution of positive, negative and neutral mentions per political party, provided by POPSTAR

	Negative Mentions	Positive Mentions	Neutral Mentions	Total Mentions
PS	28 660	225	15 326	44 211
PSD	69 723	121	37 133	106 977
CDS	41 935	51	17 554	59 540
CDU	2 445	79	5 604	8 128
BE	9 603	306	4 214	14 123

- The positive mentions represent less than 1% of the total mentions of each party, except for BE where the negative mentions represent 2% of the total mentions. This supports the idea that people use the social media mainly to express their negative opinions.
- The parties most often mentioned are PS, PSD and CDS. The total mentions to these three parties represent 90% of the data sample total mentions. PSD and CDS are the ruling parties while PS is the main opposition party in the time frame the data is from.

4.4 Polls

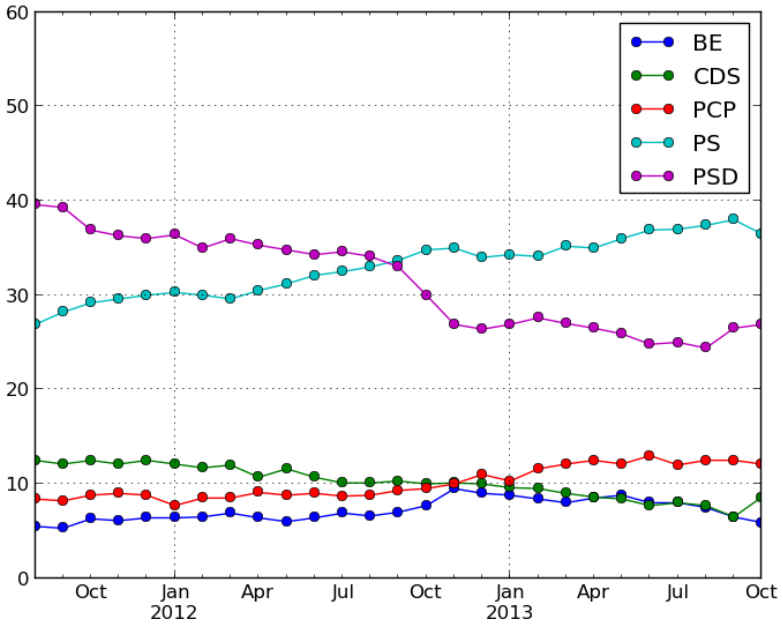

Figure 4.2: Representation of the monthly poll values of each political candidate

As described in section 4.2, the target variable of the regressive model is the poll result. Eurosondagens polls values from June 2011 to October 2013 are represented in figure 4.2. Figure 4.2

shows us two main party groups: (1) The first group, where both PSD and PS are included, has a higher value of vote intention (above 23%). PSD despite starting as the preferred party in vote intention poll, has a downtrend along the time, losing the leadership for PS in September 2012. On the other hand, PS has an uptrend, except in the last month. (2) The second group is composed by CDS, PCP and BE. This group has a vote intention range from 5% to 15%. While PP has a downtrend in public opinion, PCP has an ascendent one.

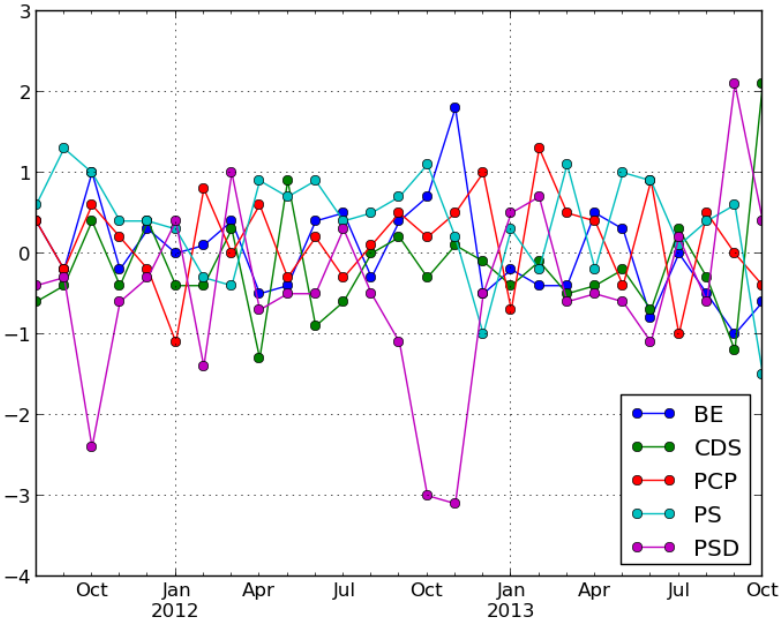

Figure 4.3: Percentage points difference between the current and the previous month of the poll values from August 2011 to October 2013

Figure 4.3 represents the monthly public opinion variation from August 2011 to October 2013. The variation is the difference between the poll value of the actual month and the previous one. In mathematical notation, the variation is given by $new_m = m_i - m_{i-1}$. An obstacle in defining a predictive model capable of correctly reflecting the public opinion is the small monthly variation of the vote intention poll. Generally, the monthly variation is within a range from -2% to 2%. However, the greatest variation verified is -3%.

4.5 Experiment using absolute values

Table 4.3 show the MAE comparison between OLS and Random Forests results (with and without y_{t-1} feature), as well as the baseline results.

All graphical representations can be consulted on the appendix section B.1.

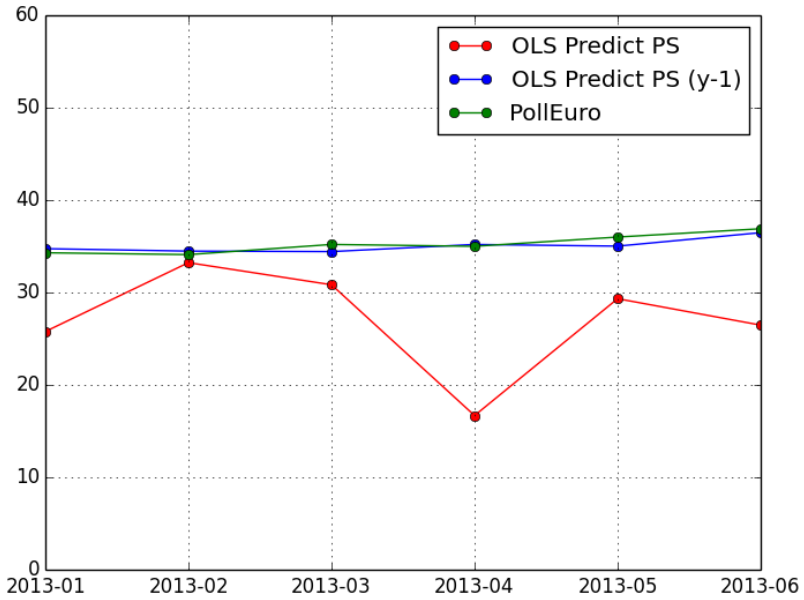

Figure 4.4: Predicted and real values of vote intention, from January 2013 to June 2013, for António José Seguro (PS), including and excluding the y_{t-1} feature

Table 4.2: Mean Absolute Errors of predictive models and baseline

Prediction Method	PS	CDS	PSD	PCP	BE	Overall
Baseline	0.62	0.38	0.67	0.70	0.43	0.56
Random Forests	4.08	6.34	2.93	2.13	0.58	3.19
Random Forests (y_{t-1})	0.92	0.28	0.74	0.80	0.40	0.64
OLS	7.78	7.08	4.74	5.02	3.33	5.59
OLS (y_{t-1})	0.54	0.41	1.03	0.57	0.70	0.65

Random Forests is the algorithm with lower mean absolute error (MAE). However, those results are not as good as the naive baseline. It means that some modifications have to be performed in order to improve the predictive effectiveness of these models. The inclusion of the y_{t-1} feature in the training and testing sets represents a significant improvement. In the Random Forests and the OLS algorithms, the inclusion of y_{t-1} features made the MAE to drop from 3.19 to 0.64 and from 5.59 to 0.65, respectively.

Figure 4.4 represents the poll predictions using OLS, excluding and including the y_{t-1} feature, respectively. Including the latter feature the absolute error (MAE) dropped from 5.59 to 0.65. This means that y_{t-1} has a big predictive power. However, the prediction error is still greater than the baseline's.

Random Forests is the second algorithm we used to build our predictive model. Figure 4.5

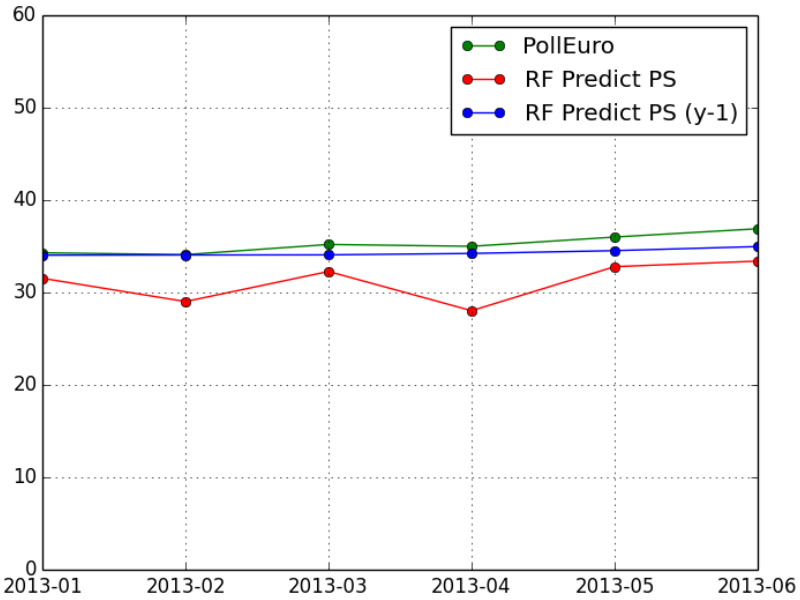

Figure 4.5: Predicted and real values of vote intention, from January 2013 to June 2013, for António José Seguro (PS), including and excluding the y_{t-1} feature.

represents the graphical representation of the experience we performed using Random Forests algorithm.

In this case, without including the y_{t-1} feature in our regressive model, the absolute error is smaller than using OLS (3.19). According to our experiment, when y_{t-1} feature is included, the predictions absolute error is 0.64.

These results allow us to conclude that when using absolute values, our model performs the best with Random Forests algorithm.

4.6 Experiment using monthly variations

As explained in section 4.2, to perform this experiment, instead of training and testing our model with absolute values, we decided to use monthly variations.

Table 4.4: MAE and MSE comparision when the variation from the previous month is greater or smaller than 1

	MSE		MAE	
	<1	>=1	<1	>=1
Baseline	0.27	1.28	0.47	1.13
Random Forests	0.37	1.56	0.51	1.15
Random Forests ($\Delta(y_{t-1})$)	0.36	1.52	0.52	1.15
OLS	0.41	1.67	0.53	1.27
OLS ($\Delta(y_{t-1})$)	0.57	1.49	0.58	1.19

Table 4.3: Mean Absolute Errors of predictive models and baseline

Prediction Method	PS	CDS	PSD	PCP	BE	Overall
Baseline	0.62	0.38	0.67	0.70	0.43	0.56
Random Forests	0.84	0.41	0.57	0.70	0.46	0.60
Random Forests ($\Delta(y_{t-1})$)	0.85	0.40	0.57	0.74	0.48	0.60
OLS	0.78	0.30	0.60	0.86	0.62	0.63
OLS ($\Delta(y_{t-1})$)	0.84	0.37	0.53	0.86	0.70	0.66

This experiment has lower error values than the experiment described in section 4.5. Also, it presents some different behaviours. If we include the Δy_{t-1} feature as input in the Random Forests algorithm we will obtain lower error value than excluding it.

However, the MAE difference between using Random Forests including and excluding the $\Delta(y_{t-1})$ feature is smaller than it was in section 4.5. On the other hand, OLS presents greater absolute error values when including the $\Delta(y_{t-1})$ feature.

Another important fact is that the absolute error of Random Forests (excluding the $\Delta(y_{t-1})$ feature) is smaller than the one obtained in experiment 4.5, which is an important improvement. The MAE dropped from 3.19 to 0.60. At this point, we can conclude that $\Delta(y_{t-1})$ feature has not the determinant role that y_{t-1} had in the first experiment.

In this experiment we also analyzed the behavior of the model when the poll variation is greater than or equal to 1 (1%) and when is smaller than 1, presented in table 4.4. We can see that the baseline has lower error values than our predictions.

In this experiment, we verified that almost all results of both algorithms present lower absolute error than in experiment 4.5. Including the $\Delta(y_{t-1})$ feature in the OLS algorithm, we obtained greater absolute error values than in the previous experiment. However, table 4.3 shows that baseline still has lower absolute error value that our experiments. The smaller error value we could obtain is 0.60 using the Random Forests algorithm. This is closer to the baseline MAE (0.56) than in any other experiments.

The error of the predictions, as discussed above, presents one perspetive on the performance of the models. However, in this domain it also makes sense to analyse if the model is able to

predict the trends correctly. In other words, if the model predicts a positive or negative variation in the vote intentions and the observed value is consistent with that prediction or not. Thus, we build a confusion matrix for each party composed by the times that our model (Random Forests, without Δy_{t-1} feature) correctly and mistakenly predicted the signal of the poll variation (tables 4.5 to 4.10), and a global one. Our model has some difficulty to predict true positives, specially when predicting PS and PSD poll results. It means that in the majority of the months where the poll results rised, the model predicted a fall. On the other hand, in 70% of the cases where the poll results actually dropped, our model correctly predicted the drop.

Table 4.5: Global confusion matrix

		Predicted Value	
		Positive	Negative
Real Value	Positive	5	7
	Negative	5	13

Table 4.6: Confusion matrix PS

		Predicted Value	
		Positive	Negative
Real Value	Positive	1	3
	Negative	0	2

Table 4.7: Confusion matrix PSD

		Predicted Value	
		Positive	Negative
Real Value	Positive	0	2
	Negative	0	4

Table 4.8: Confusion matrix CDS

		Predicted Value	
		Positive	Negative
Real Value	Positive	0	0
	Negative	3	3

Table 4.9: Confusion matrix CDU

		Predicted Value	
		Positive	Negative
Real Value	Positive	3	1
	Negative	1	1

Table 4.10: Confusion matrix BE

		Predicted Value	
		Positive	Negative
Real Value	Positive	1	1
	Negative	1	3

All graphical representations can be consulted on the appendix section B.2.

4.7 Experiment Sentiment vs Buzz

So far, the lower error value we could obtain is 0.60 performing experiment 4.5, using the Random Forests algorithm. As described in section 3.4, in this experience we divided the aggregator list into two groups: buzz and sentiment aggregators. We repeated the process used in previous experiments for each group. The next chapters, will present the results for OLS and Random Forests, using sentiment and buzz aggregators, separately. A detailed representation of results is given in appendix section B.3.

4.7.1 Sentiment

In this experiment, Random Forests has smaller MAE than in experiment 4.5. OLS on the other hand, has lower absolute error when including $\Delta(y_{t-1})$ feature, and a greater one when including it, as can be seen in table 4.15.

Table 4.12 shows that Random Forests is the algorithm with the lowest MAE when the poll variation is greater or equal to 1. This results means that when the poll variation is bigger or equal to 1, Random Forests has better preditive power than the baseline itself. However, the global MAE is greater than the baseline due to the lack of predictiveness when there is a small poll variation between two consecutive months.

Table 4.11: Mean Absolute Errors of predictive models and baseline

Prediction Method	PS	CDS	PSD	PCP	BE	Overall
Baseline	0.62	0.38	0.67	0.70	0.43	0.56
Random Forests	0.81	0.37	0.51	0.72	0.48	0.58
Random Forests ($\Delta(y_{t-1})$)	0.81	0.40	0.48	0.69	0.48	0.57
OLS	0.85	0.38	0.48	0.90	0.72	0.66
OLS ($\Delta(y_{t-1})$)	0.76	0.32	0.57	0.89	0.63	0.63

Table 4.12: MAE and MSE when the variation from the previous month is greater or smaller than 1

	MSE		MAE	
	<1	>=1	<1	>=1
Baseline	0.27	1.28	0.47	1.13
Random Forests	0.33	1.44	0.49	1.12
Random Forests ($\Delta(y_{t-1})$)	0.33	1.44	0.49	1.10
OLS	0.57	1.54	0.58	1.21
OLS ($\Delta(y_{t-1})$)	0.42	1.73	0.53	1.30

At the moment, the smaller absolute error we could obtain is using sentiment aggregators with Random Forests algorithm. Despite that fact, and despite the fact that that Random Forests algorithm can predict poll values with lower error than OLS, it still has greater absolute error than the baseline.

4.7.2 Buzz

This section is dedicated to study the relationship between polls results and buzz aggregators. In other words, we ignored all the aggregators from the aggregators list that used sentiment anal-

ysis. Thus, to perform this experiment we used the aggregators *share* (buzz share) and the *total_mentions* (number of tweets that refers to a given candidate), together with the party id.

Table 4.13: Mean Absolute Errors of predictive models and baseline

Prediction Method	PS	CDS	PSD	PCP	BE	Overall
Baseline	0.62	0.38	0.67	0.70	0.43	0.56
Random Forests	0.63	0.33	0.54	0.60	0.45	0.52
Random Forests ($\Delta(y_{t-1})$)	0.71	0.38	0.68	0.60	0.42	0.56
OLS	0.57	0.35	0.70	0.71	0.42	0.55
OLS ($\Delta(y_{t-1})$)	0.68	0.32	0.68	0.75	0.38	0.56

Table 4.14: MAE and MSE comparision when the variation from the previous month is greater or smaller than 1

	MSE		MAE	
	<1	>=1	<1	>=1
Baseline	0.27	1.28	0.47	1.13
Random Forests	0.29	1.33	0.42	1.14
Random Forests ($\Delta(y_{t-1})$)	0.29	1.44	0.46	1.19
OLS	0.27	1.36	0.46	1.16
OLS ($\Delta(y_{t-1})$)	0.30	1.49	0.46	1.20

In table 4.13 we can see that buzz has an important role in predicting the poll results, given that it presents lower MAE than other experiments. Using buzz aggregators, the bigger absolute error we obtained was 0.56. This means that the MAE of the worst performance of this experiment is equal to the baseline. At this point we can conclude that $\Delta(y_{t-1})$ feature has not the determinant role it had in other experiments. According to table 4.13, excluding the $\Delta(y_{t-1})$ feature, OLS or Random Forests obtained lower error value than the baseline itself.

So far, these results are the best results of the three experiments we made. Without using the $\Delta(y_{t-1})$ feature we can obtain a better prediction (with lower MAE) than the baseline. Including the $\Delta(y_{t-1})$ feature, the MAE error is equal to the baseline.

These results show that when using buzz aggregators, Random Forests algorithm performs better than OLS and the baseline.

4.8 Experiment Sentiment vs Buzz All

At this point, we are able to conclude that using Random Forests algorithm with buzz aggregators, without including the $\Delta(y_{t-1})$ feature, we can obtain better predictive power than the one provided by the baseline. However we decided to perform one last experiment. The variation of the surveys

results of a candidate is not an independent event by itself. Probably, it is correlated with the variation of the popularity of another political target or party. Based on this premise, we decided also to include the previous monthly variation of polls of all parties as predictive features in our regression model.

We perform this experiment using sentiment and buzz aggregators separately.

All graphical representations can be consulted on the appendix section B.4.

4.8.1 Sentiment

In this section we present the results of the experiment using only sentiment aggregators.

Table 4.15: Mean Absolute Errors of predictive models and baseline

Prediction Method	PS	CDS	PSD	PCP	BE	Overall
Baseline	0.62	0.38	0.67	0.70	0.43	0.56
Random Forests ($\Delta(y_{t-1})$)	0.81	0.39	0.46	0.70	0.47	0.57
OLS ($\Delta(y_{t-1})$)	0.74	0.32	0.58	0.93	0.58	0.63

Table 4.16: MAE and MSE comparision when the variation from the previous month is greater or smaller than 1

	MSE		MAE	
	<1	>=1	<1	>=1
Baseline	0.27	1.28	0.47	1.13
Random Forests ($\Delta(y_{t-1})$)	0.32	1.46	0.48	1.11
OLS ($\Delta(y_{t-1})$)	0.44	1.50	0.54	1.20

This experiment showed that including the previous monthly poll results of all political target in every prediction worsens the model, given that there is an increase of the MAE. This means a decrease in the predictive power of our model.

All graphical representations can be consulted on the appendix section B.4.

4.8.2 Buzz

In this section we present the results of the experiment using only buzz aggregators.

All graphical representations can be consulted on the appendix section B.4.

Table 4.17: Mean Absolute Errors of predictive models and baseline

Prediction Method	PS	CDS	PSD	PCP	BE	Overall
Baseline	0.62	0.38	0.67	0.70	0.43	0.56
Random Forests ($\Delta(y_{t-1})$)	0.61	0.30	0.56	0.58	0.43	0.50
OLS ($\Delta(y_{t-1})$)	0.50	0.43	0.66	0.70	0.45	0.55

Table 4.18: MAE and MSE comparision when the variation from the previous month is greater or smaller than 1

	MSE		MAE	
	<1	>=1	<1	>=1
Baseline	0.27	1.28	0.47	1.13
Random Forests ($\Delta(y_{t-1})$)	0.26	1.16	0.41	1.06
OLS ($\Delta(y_{t-1})$)	0.30	1.43	0.45	1.19

According to this study, when using Random Forests algorithm we can minimize the absolute error to 0.50, which is an important improvement. Table 4.17 also shows that OLS has lower MAE than the baseline.

4.9 Experiment with different data set (predicting July to December)

In the previous experiments, we used the same set of data for multiple experiments, namely varying the features, the algorithm and its parameters. Although an adequate methodology was used to ensure that models are tested on data that was not used for training, multiple experiments increase the probability that a good result is obtained by chance. Therefore, it is a good practice to develop a model using a data set, and test it with a different one. Thus, in order to evaluate our model with different data from the one used to develop it, we used the best model obtained in experiment 4.5, to predict poll results from July to December 2013. We could use the model with the lowest MAE we obtained, however given the proximity of the MAEs of all experiments that use variations, we used a model that includes all the aggregators.

Table 4.19: MAE comparision

Prediction Method	PS	CDS	PSD	PCP	BE	Overall
Baseline (Jan to Jun)	0.62	0.38	0.67	0.70	0.43	0.56
Random Forests (Jan to Jun)	0.84	0.41	0.57	0.70	0.46	0.60
Baseline (Jul to Dec)	0.70	0.70	0.92	0.67	0.45	0.69
Random Forests (Jul to Dec)	0.47	0.94	0.84	0.80	0.54	0.72

Table 4.21: Average of the poll variation module in the first and second semester

	PS	CDS	PSD	PCP	BE
First Semester	0.6	0.4	0.7	0.7	0.4
Second Semester	0.7	0.7	0.9	0.7	0.5

Table 4.20: MAE and MSE when the monthly variation is greater or smaller than 1

	MSE		MAE	
	<1	>=1	<1	>=1
Baseline (Jan to Jun)	0.27	1.28	0.47	1.13
Random Forests (Jan to Jun)	0.37	1.56	0.51	1.15
Baseline (Jul to Dec)	0.23	2.05	0.40	1.37
Random Forests (Jul to Dec)	0.24	3.67	0.40	1.80

Table 4.19 presents the comparision between the MAE of poll results predictions from Jan to June 2013 and from July to December 2013. The baseline MAE is different because we aimed to predict the poll results of the second semester of the year 2013. Table 4.19 shows that the baseline MAE had a slight increase. It happened because in the second semester of the year, the poll results were not so stable as they were in the first semester, as table 4.21 suggests.

Table 4.19 shows that our model has a predictive power slightly weaker than the baseline, except when predicting PS poll results.

Also, we build confusion matrices the same way we did in section 4.6 (tables 4.22 to 4.27). In table 4.26 and 4.27 are only represented 5 and 4 predictions, respectively. It happens because there are one and two months where the CDU and BE poll results, respectively, did not variate. However, the 3 times the poll results variation is 0, the model predicted a rise. According to table 4.22, only in 50% of the cases that the poll results actually dropped were correctly predicted by our model.

Table 4.22: Global confusion matrix

		Predicted Value	
		Positive	Negative
Real Value	Positive	8	5
	Negative	7	7

Table 4.23: Confusion matrix PS

		Predicted Value	
		Positive	Negative
Real Value	Positive	4	0
	Negative	1	1

Table 4.24: Confusion matrix PSD

		Predicted Value	
		Positive	Negative
Real Value	Positive	3	1
	Negative	1	1

Table 4.25: Confusion matrix CDS

		Predicted Value	
		Positive	Negative
Real Value	Positive	1	1
	Negative	1	2

Table 4.26: Confusion matrix CDU

		Predicted Value	
		Positive	Negative
Real Value	Positive	0	1
	Negative	3	1

Table 4.27: Confusion matrix BE

		Predicted Value	
		Positive	Negative
Real Value	Positive	0	1
	Negative	1	2

Chapter 5

Conclusions

Traditional telephone surveys are the most common method used to collect public opinion. However, this method is expensive and time consuming. On the other hand, people freely express their opinions on the web. With this project we proposed to define a methodology capable of predicting the poll results using social media data. First, we performed an exhaustive study of the sentiment aggregation methods present in the literature. The second step was to apply these aggregators in our data set (containing the daily count of positive, negative and neutral messages relative to each candidate). Furthermore we performed multiple experiments, namely varying the features and the algorithm and studied its results.

This dissertation project allows us to take some conclusions with respect to the relationship between the Portuguese tweetoshphere and the polls results of the Portuguese private company Eurosondagem.

The first conclusion is that we could estimate the polls results using the sentiment inherent to twitter messages of our case study with a small prediction error (0.50, in a scale from 0 to 100). Furthermore, due to the small poll variation from month to month, we used a naive baseline, where we predict for month i the same real poll result of month $i-1$. We showed that this baseline has a strong predictive power, given the small MAE obtained (0.56). It happens due to the small variation of the poll results of the Portuguese scenario.

Using absolute aggregators values in the regression models has a worst predictive performance, given that the experiment obtained the highest prediction error of all experiments. Furthermore, the y_{t-1} feature has a determinant role in the regression models predictive effectiveness of that experiment. This can be easily explainable once we are including the baseline in the regression model.

Another important conclusion we can take is that experiments using only buzz aggregators have lower prediction errors, which highlights the strong predictive power of buzz aggregators. The smaller MAE we obtained (0.50, from 0 to 100) was using the Random Forests algorithm, together with buzz aggregators, *share* and *total_mentions*, and the previous poll variation of all candidates.

The experiments using only sentiment aggregators have higher prediction error than the baseline, However, we can assume two different positions: we can simply conclude that (1) the implemented sentiment aggregators have a weak predictive power or (2) these results can be justified by a weak performance of the sentiment analysis methods.

We repeated the experiment where we used the monthly variation of all aggregators, with different data. We tried to use that model to predict the polls results from July 2013 to December 2013. The first fact we noticed was the increase of the baseline MAE for the second semester of the year. It happens because in the first semester of the year, the poll results are more stable, having small monthly variations. In the second semester, there are months where the monthly variation of polls is bigger than the average of the first half of the year. The second conclusion that we took from this experiment is that this model has a bigger MAE when predicting the poll results of the months belonging to the second half of the year.

5.1 Future Work

The next immediate step is to build an online representation widget, where we could see in real time how Portuguese general opinion is changing according to twitter activity. Enable the user to choose which indicators he desires to use in the representation widget is an added value.

Furthermore, testing the methodology defined on chapter 3 with different data sources, such as text from blogs or news would be part of future work.

It is also desirable to implement a methodology using Time Series analysis.

References

[AH10] Sitaram Asur and Bernardo a. Huberman. Predicting the Future with Social Media. *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 492–499, August 2010.

[BKY12] Antoine Boutet, Hyoungshick Kim, and Eiko Yoneki. What's in Your Tweets? I Know Who You Supported in the UK 2010 General Election. *. . . AAAI Conference on Weblogs and Social . . .* , pages 411–414, 2012.

[BS] Adam Bermingham and Alan F Smeaton. On Using Twitter to Monitor Political Sentiment and Predict Election Results.

[CBRS] Brendan O Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From Tweets to Polls : Linking Text Sentiment to Public Opinion Time Series.

[CN02] HL Chieu and HT Ng. Named entity recognition: a maximum entropy approach using global information. *Proceedings of the 19th international conference on . . .* , 2002.

[CR11] MD Conover and J Ratkiewicz. Political polarization on twitter. *Proc. 5th Intl. . . .* , pages 89–96, 2011.

[DS10] Nicholas a. Diakopoulos and David a. Shamma. Characterizing debate performance via aggregated twitter sentiment. *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*, page 1195, 2010.

[GE03] I Guyon and A Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 2003.

[HK] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, second edition edition.

[Hol13a] Richard Holt. Half a billion people sign up for Twitter. Available at http://www.telegraph.co.uk/technology/9837525/Half-a-billion-people-sign-up-for-Twitter.html, Jan 2013.

[Hol13b] Richard Holt. Twitter In Numbers. Available at http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-innumbers.html, Mars 2013.

[JDGS10] Mahesh Joshi, Dipanjan Das, Kevin Gimpel, and Noah A. Smith. Movie reviews and revenues: an experiment in text regression. pages 293–296, June 2010.

REFERENCES

[JSS] Christopher Johnson, P Shukla, and S Shukla. On Classifying the Political Sentiment of Tweets. *cs.utexas.edu*.

[KKD+12] Andrew Kohut, Scott Keeter, Carroll Doherty, Michael Dimock, Associate Directors, Leah Christian, and Senior Researcher. Assessing the Representativeness of Public Opinion Surveys Assessing the Representativeness of Public Opinion Surveys. (202), 2012.

[KM02] D Klein and CD Manning. Fast exact inference with a factored model for natural language parsing. ... *in neural information processing* ..., 2002.

[MMGA11] Panagiotis T. Metaxas, Eni Mustafaraj, and Dani Gayo-Avello. How (Not) to Predict Elections. *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing*, pages 165–171, October 2011.

[PL08] Bo Pang and Lillian Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

[RL94] E Riloff and W Lehnert. Information extraction as a basis for high-precision text classification. *ACM Transactions on Information Systems (TOIS)*, 1994.

[TBP11] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment in Twitter Events. 62(2):406–418, 2011.

[TM00] K Toutanova and CD Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. *Proceedings of the 2000 Joint SIGDAT* ..., 2000.

[TSSW10] a. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape. *Social Science Computer Review*, 29(4):402–418, December 2010.

[WFH11] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining. Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, third edit edition, 2011.

[ZS09] W Zhang and S Skiena. Improving movie gross prediction through news analysis. *Proceedings of the 2009 IEEE/WIC/ACM* ..., 2009.

Appendix A

Correlation Tables

Table A.1: Aggregators correlation for António José Seguro (PS)

	bermingham2	normalized bermingham2	bermingham (sovn)	bermingham (sovp)	connor	normalized connor	gayo	normalized gayo	ind	polarity
bermingham2	1.00									
normalized bermingham2	0.43	1.00								
berminghamsovn	-0.23	-0.09	1.00							
berminghamsovp	0.39	0.14	0.00	1.00						
connor	0.71	0.33	-0.23	0.41	1.00					
normalized connor	0.71	0.33	-0.23	0.41	1.00	1.00				
gayo	0.24	0.08	-0.98	0.02	0.24	0.24	1.00			
normalized gayo	0.11	0.14	-0.28	0.18	0.14	0.14	0.30	1.00		
ind	-0.38	-0.91	0.04	-0.11	-0.28	-0.28	-0.03	-0.16	1.00	
polarity	0.27	0.40	-0.34	0.05	0.11	0.11	0.32	0.02	-0.42	1.00
normalized polarity	0.43	0.95	-0.09	0.14	0.33	0.33	0.08	0.17	-0.95	0.42
polarityONeutral	0.29	0.67	-0.23	0.01	0.23	0.23	0.22	0.13	-0.71	0.55
polarityOTotal	0.43	0.95	-0.09	0.14	0.33	0.33	0.08	0.17	-0.95	0.42
subjNeu	-0.24	-0.67	0.19	0.02	-0.18	-0.18	-0.18	-0.14	0.74	-0.54
subjSoV	-0.22	-0.08	0.99	0.00	-0.22	-0.22	-0.98	-0.29	0.03	-0.32
subjVol	-0.26	-0.40	0.34	-0.04	-0.10	-0.10	-0.32	-0.03	0.43	-0.99
share	-0.16	0.02	0.85	0.02	-0.20	-0.20	-0.83	-0.16	-0.06	-0.30
shareOfNegDistribution	-0.21	-0.27	0.30	-0.09	-0.20	-0.20	-0.32	-0.77	0.24	-0.07
normalized negative	-0.39	-0.91	0.05	-0.11	-0.30	-0.30	-0.05	-0.17	0.99	-0.42
normalized neutral	0.38	0.91	-0.04	0.11	0.28	0.28	0.03	0.16	-1.00	0.42
normalized positive	0.70	0.20	-0.30	0.38	0.84	0.84	0.30	0.09	-0.15	0.12
pollEuro	0.10	0.07	-0.19	0.05	0.08	0.08	0.19	0.06	-0.08	0.27
pollEuro-1	-0.18	-0.29	0.01	-0.19	-0.07	-0.07	0.00	-0.05	0.30	-0.35
positivementions	0.34	-0.02	-0.05	0.25	0.49	0.49	0.07	0.00	0.08	-0.28
negativementions	-0.26	-0.40	0.34	-0.04	-0.10	-0.10	-0.32	-0.03	0.43	-0.99
neutralmentions	-0.13	-0.14	0.17	0.10	-0.02	-0.02	-0.15	0.11	0.14	-0.60
total mentions	-0.24	-0.32	0.33	-0.01	-0.09	-0.09	-0.31	0.00	0.34	-0.90

Table A.2: Aggregators correlation for António José Seguro (PS)

	normalized polarity	polarity Over Neutral	polarity Over Total	subjNeu	subjSoV	subjVol	share	share Of Neg Distribution	normalized negative	normalized neutral
bermingham2										
normalized bermingham2										
bermingham (sovn)										
bermingham (sovp)										
connor										
normalized connor										
gayo										
normalized gayo										
ind										
polarity										
normalized polarity	1.00									
polarityONeutral	0.72	1.00								
polarityOTotal	1.00	0.72	1.00							
subjNeu	-0.72	-0.94	-0.72	1.00						
subjSoV	-0.08	-0.23	-0.08	0.19	1.00					
subjVol	-0.43	-0.56	-0.43	0.55	0.33	1.00				
share	0.02	-0.11	0.02	0.08	0.84	0.30	1.00			
shareOfNegDistribution	-0.26	-0.22	-0.26	0.23	0.32	0.08	0.16	1.00		
normalized negative	-0.96	-0.72	-0.96	0.74	0.04	0.43	-0.06	0.26	1.00	
normalized neutral	0.95	0.71	0.95	-0.74	-0.03	-0.43	0.06	-0.24	-0.99	1.00
normalized positive	0.20	0.14	0.20	-0.09	-0.28	-0.12	-0.30	-0.14	-0.16	0.15
pollEuro	0.07	0.19	0.07	-0.15	-0.20	-0.27	-0.19	-0.03	-0.08	0.08
pollEuro-1	-0.31	-0.13	-0.31	0.17	0.00	0.34	0.00	0.04	0.30	-0.30
positive mentions	-0.03	-0.06	-0.03	0.12	-0.06	0.29	-0.02	-0.08	0.07	-0.08
negative mentions	-0.43	-0.56	-0.43	0.55	0.33	1.00	0.30	0.08	0.43	-0.43
neutral mentions	-0.14	-0.16	-0.14	0.15	0.15	0.59	0.22	0.00	0.14	-0.14
total mentions	-0.34	-0.45	-0.34	0.44	0.32	0.89	0.32	0.05	0.34	-0.34

Table A.3: Aggregators correlation for António José Seguro (PS)

	normalized positive	pollEuro	pollEuro-1	positive mentions	negative mentions	neutral mentions	total mentions
bermingham2							
normalized bermingham2							
bermingham (sovn)							
bermingham (sovp)							
connor							
normalized connor							
gayo							
normalized gayo							
ind							
polarity							
normalized polarity							
polarityONeutral							
polarityOTotal							
subjNeu							
subjSoV							
subjVol							
share							
shareOfNegDistribution							
normalized negative							
normalized neutral							
normalized positive	1.00						
pollEuro	0.08	1.00					
pollEuro-1	-0.09	0.08	1.00				
positive mentions	0.50	-0.15	0.21	1.00			
negative mentions	-0.12	-0.27	0.34	0.29	1.00		
neutral mentions	-0.07	-0.16	0.26	0.33	0.59	1.00	
total mentions	-0.12	-0.25	0.34	0.34	0.89	0.70	1.00

Table A.4: Aggregators correlation for Pedro Passos Coelho (PSD)

	bermingham2	normalized bermingham2	bermingham (sovn)	bermingham (sovp)	connor	normalized connor	gayo	normalized gayo	ind	polarity
bermingham2	1.00									
normalized bermingham2	0.21	1.00								
bermingham (sovn)	-0.18	-0.21	1.00							
bermingham (sovp)	0.58	-0.05	0.07	1.00						
connor	0.64	-0.03	0.04	0.73	1.00					
normalized connor	0.64	-0.03	0.04	0.73	1.00	1.00				
gayo	0.17	0.21	-0.99	-0.07	-0.05	-0.05	1.00			
normalized gayo	-0.11	0.23	-0.50	-0.20	-0.30	-0.30	0.51	1.00		
ind	-0.20	-0.96	0.21	0.08	0.07	0.07	-0.21	-0.23	1.00	
polarity	0.19	0.35	-0.18	0.02	0.05	0.05	0.18	0.03	-0.35	1.00
normalized polarity	0.22	0.99	-0.20	-0.04	-0.03	-0.03	0.19	0.22	-0.96	0.37
polarityONeutral	0.20	0.84	-0.14	-0.05	-0.04	-0.04	0.13	0.12	-0.85	0.39
polarityOTotal	0.22	0.99	-0.20	-0.04	-0.03	-0.03	0.19	0.22	-0.96	0.37
subjNeu	-0.19	-0.83	0.13	0.05	0.05	0.05	-0.13	-0.11	0.85	-0.38
subjSoV	-0.17	-0.22	0.99	0.07	0.05	0.05	-0.99	-0.51	0.22	-0.19
subjVol	-0.19	-0.35	0.18	-0.02	-0.04	-0.04	-0.17	-0.02	0.35	-0.99
share	-0.26	-0.09	0.64	-0.03	-0.05	-0.05	-0.64	-0.19	0.10	-0.21
shareOfNegDistribution	0.04	-0.22	0.44	0.16	0.22	0.22	-0.45	-0.85	0.22	0.03
normalized negative	-0.21	-0.97	0.22	0.07	0.06	0.06	-0.21	-0.24	0.99	-0.36
normalized neutral	0.20	0.96	-0.21	-0.08	-0.07	-0.07	0.21	0.23	-1.00	0.35
normalized positive	0.64	-0.03	0.03	0.75	0.99	0.99	-0.03	-0.28	0.07	0.05
pollEuro	-0.04	-0.11	0.02	0.08	-0.04	-0.04	-0.02	0.00	0.11	0.00
pollEuro-1	-0.06	0.07	-0.13	-0.12	-0.22	-0.22	0.13	0.29	-0.08	0.03
positive mentions	0.58	-0.06	0.00	0.65	0.75	0.75	-0.01	-0.13	0.10	-0.09
negative mentions	-0.19	-0.35	0.18	-0.02	-0.05	-0.05	-0.18	-0.03	0.35	-1.00
neutral mentions	-0.18	-0.14	0.14	-0.01	-0.08	-0.08	-0.13	0.02	0.14	-0.78
total mentions	-0.18	-0.27	0.14	-0.03	-0.07	-0.07	-0.14	0.02	0.27	-0.91

Table A.5: Aggregators correlation for Pedro Passos Coelho (PSD)

	normalized polarity	polarity Over Neutral	polarity Over Total	subjNeu	subjSoV	subjVol	share	share Of Neg Distribution	normalized negative	normalized neutral
bermingham2										
normalized bermingham2										
bermingham (sovn)										
bermingham (sovp)										
connor										
normalized connor										
gayo										
normalized gayo										
ind										
polarity										
normalized polarity	1.00									
polarityONeutral	0.85	1.00								
polarityOTotal	1.00	0.85	1.00							
subjNeu	-0.85	-0.99	-0.85	1.00						
subjSoV	-0.21	-0.14	-0.21	0.14	1.00					
subjVol	-0.36	-0.40	-0.36	0.39	0.18	1.00				
share	-0.10	-0.11	-0.10	0.11	0.64	0.20	1.00			
shareOfNegDistribution	-0.21	-0.10	-0.21	0.09	0.45	-0.04	0.12	1.00		
normalized negative	-0.97	-0.85	-0.97	0.84	0.22	0.35	0.11	0.22	1.00	
normalized neutral	0.96	0.85	0.96	-0.85	-0.22	-0.35	-0.10	-0.22	-0.99	1.00
normalized positive	-0.03	-0.04	-0.03	0.05	0.03	-0.04	-0.05	0.21	0.06	-0.07
pollEuro	-0.11	-0.11	-0.11	0.10	0.02	-0.01	-0.07	0.05	0.12	-0.11
pollEuro-1	0.06	0.06	0.06	-0.05	-0.13	-0.03	-0.01	-0.28	-0.09	0.08
positive mentions	-0.07	-0.11	-0.07	0.11	0.01	0.10	-0.03	0.06	0.09	-0.10
negative mentions	-0.37	-0.39	-0.37	0.38	0.19	0.99	0.21	-0.03	0.36	-0.35
neutral mentions	-0.15	-0.17	-0.15	0.17	0.14	0.77	0.23	-0.09	0.14	-0.14
total mentions	-0.28	-0.30	-0.28	0.30	0.15	0.91	0.19	-0.08	0.27	-0.27

Table A.6: Aggregators correlation for Pedro Passos Coelho (PSD)

	normalized positive	pollEuro	pollEuro-1	positive mentions	negative mentions	neutral mentions	total mentions
bermingham2							
normalized bermingham2							
bermingham (sovn)							
bermingham (sovp)							
connor							
normalized connor							
gayo							
normalized gayo							
ind							
polarity							
normalized polarity							
polarityONeutral							
polarityOTotal							
subjNeu							
subjSoV							
subjVol							
share							
shareOfNegDistribution							
normalized negative							
normalized neutral							
normalized positive	1.00						
pollEuro	-0.04	1.00					
pollEuro-1	-0.20	0.04	1.00				
positive mentions	0.77	-0.09	-0.13	1.00			
negative mentions	-0.05	0.00	-0.03	0.09	1.00		
neutral mentions	-0.07	-0.01	-0.03	0.08	0.78	1.00	
total mentions	-0.05	0.00	-0.01	0.10	0.91	0.86	1.00

Table A.7: Aggregators correlation for Paulo Portas (CDS)

	bermingham2	normalized bermingham2	bermingham (sovn)	bermingham (sovp)	connor	normalized connor	gayo	normalized gayo	ind	polarity
bermingham2	1.00									
normalized bermingham2	0.32	1.00								
bermingham (sovn)	0.06	-0.13	1.00							
bermingham (sovp)	0.49	-0.07	0.13	1.00						
connor	0.55	-0.02	0.07	0.85	1.00					
normalized connor	0.55	-0.02	0.07	0.85	1.00	1.00				
gayo	-0.08	0.11	-0.98	-0.13	-0.07	-0.07	1.00			
normalized gayo	0.00	0.03	-0.22	0.09	0.03	0.03	0.22	1.00		
ind	-0.26	-0.92	0.13	0.13	0.09	0.09	-0.10	0.00	1.00	
polarity	0.22	0.36	-0.18	-0.08	-0.03	-0.03	0.17	-0.07	-0.37	1.00
normalized polarity	0.30	0.95	-0.14	-0.10	-0.05	-0.05	0.12	0.03	-0.96	0.36
polarityONeutral	0.18	0.77	-0.16	-0.20	-0.14	-0.14	0.13	-0.08	-0.85	0.40
polarityOTotal	0.30	0.95	-0.14	-0.10	-0.05	-0.05	0.12	0.03	-0.96	0.36
subjNeu	-0.16	-0.76	0.15	0.20	0.16	0.16	-0.13	0.08	0.84	-0.39
subjSoV	0.08	-0.14	0.98	0.16	0.10	0.10	-0.95	-0.21	0.15	-0.19
subjVol	-0.22	-0.36	0.18	0.08	0.03	0.03	-0.17	0.07	0.36	-1.00
share	0.06	-0.03	0.67	0.16	0.10	0.10	-0.69	0.05	0.02	-0.14
shareOfNegDistribution	-0.05	-0.10	0.31	-0.07	-0.01	-0.01	-0.31	-0.81	0.06	0.00
normalized negative	-0.29	-0.94	0.14	0.10	0.06	0.06	-0.12	-0.03	0.97	-0.36
normalized neutral	0.26	0.92	-0.13	-0.13	-0.09	-0.09	0.10	0.00	-1.00	0.37
normalized positive	0.58	-0.02	0.08	0.83	0.97	0.97	-0.08	0.04	0.10	-0.04
pollEuro	0.03	0.02	-0.07	0.13	0.14	0.14	0.07	0.29	-0.01	-0.14
pollEuro-1	0.00	-0.14	0.02	-0.12	-0.09	-0.09	-0.01	-0.39	0.15	0.13
positive mentions	0.59	-0.07	0.17	0.79	0.86	0.86	-0.17	0.03	0.14	-0.12
negative mentions	-0.22	-0.36	0.18	0.08	0.03	0.03	-0.17	0.07	0.37	-1.00
neutral mentions	-0.27	-0.22	0.10	0.08	0.02	0.02	-0.10	0.11	0.20	-0.75
total mentions	-0.21	-0.34	0.15	0.11	0.05	0.05	-0.14	0.08	0.34	-0.94

Table A.8: Aggregators correlation for Paulo Portas (CDS)

	normalized polarity	polarity Over Neutral	polarity Over Total	subjNeu	subjSoV	subjVol	share	share Of Neg Distribution	normalized negative	normalized neutral
bermingham2										
normalized bermingham2										
bermingham (sovn)										
bermingham (sovp)										
connor										
normalized connor										
gayo										
normalized gayo										
ind										
polarity										
normalized polarity	1.00									
polarityONeutral	0.82	1.00								
polarityOTotal	1.00	0.82	1.00							
subjNeu	-0.81	-0.98	-0.81	1.00						
subjSoV	-0.17	-0.18	-0.17	0.18	1.00					
subjVol	-0.36	-0.39	-0.36	0.39	0.20	1.00				
share	-0.03	-0.03	-0.03	0.02	0.69	0.14	1.00			
shareOfNegDistribution	-0.10	0.00	-0.10	0.00	0.31	0.00	0.02	1.00		
normalized negative	-0.99	-0.83	-0.99	0.82	0.17	0.36	0.03	0.10	1.00	
normalized neutral	0.96	0.85	0.96	-0.84	-0.15	-0.36	-0.02	-0.06	-0.97	1.00
normalized positive	-0.05	-0.15	-0.05	0.17	0.11	0.04	0.11	-0.02	0.07	-0.10
pollEuro	0.00	-0.03	0.00	0.02	-0.05	0.14	0.15	-0.31	0.00	0.01
pollEuro-1	-0.14	-0.12	-0.14	0.12	0.00	-0.14	-0.18	0.32	0.15	-0.15
positive mentions	-0.10	-0.19	-0.10	0.21	0.19	0.12	0.19	-0.02	0.11	-0.14
negative mentions	-0.36	-0.40	-0.36	0.39	0.19	1.00	0.14	0.00	0.36	-0.37
neutral mentions	-0.21	-0.21	-0.21	0.20	0.11	0.75	0.16	-0.05	0.19	-0.20
total mentions	-0.34	-0.37	-0.34	0.36	0.16	0.95	0.12	-0.01	0.33	-0.34

Table A.9: Aggregators correlation for Paulo Portas (CDS)

	normalized positive	pollEuro	pollEuro-1	positive mentions	negative mentions	neutral mentions	total mentions
bermingham2							
normalized bermingham2							
bermingham (sovn)							
bermingham (sovp)							
connor							
normalized connor							
gayo							
normalized gayo							
ind							
polarity							
normalized polarity							
polarityONeutral							
polarityOTotal							
subjNeu							
subjSoV							
subjVol							
share							
shareOfNegDistribution							
normalized negative							
normalized neutral							
normalized positive	1.00						
pollEuro	0.12	1.00					
pollEuro-1	-0.07	-0.35	1.00				
positive mentions	0.89	0.11	-0.07	1.00			
negative mentions	0.04	0.14	-0.13	0.12	1.00		
neutral mentions	0.02	0.13	-0.25	0.09	0.75	1.00	
total mentions	0.06	0.16	-0.18	0.14	0.94	0.80	1.00

Table A.10: Aggregators correlation for Jerónimo de Sousa (PCP)

	bermingham2	normalized bermingham2	bermingham (sovn)	bermingham (sovp)	connor	normalized connor	gayo	normalized gayo	ind	polarity
bermingham2	1.00									
normalized bermingham2	0.70	1.00								
bermingham (sovn)	-0.30	-0.32	1.00							
bermingham (sovp)	0.32	0.22	0.26	1.00						
connor	0.57	0.54	-0.02	0.59	1.00					
normalized connor	0.57	0.54	-0.02	0.59	1.00	1.00				
gayo	0.24	0.29	-0.88	-0.23	0.02	0.02	1.00			
normalized gayo	0.53	0.69	-0.49	0.04	0.31	0.31	0.47	1.00		
ind	-0.48	-0.74	0.41	0.00	-0.31	-0.31	-0.40	-0.78	1.00	
polarity	0.55	0.49	-0.46	0.10	0.35	0.35	0.46	0.54	-0.46	1.00
normalized polarity	0.67	0.97	-0.29	0.22	0.54	0.54	0.26	0.69	-0.75	0.49
polarityONeutral	0.56	0.86	-0.29	0.19	0.52	0.52	0.32	0.71	-0.79	0.55
polarityOTotal	0.67	0.97	-0.29	0.22	0.54	0.54	0.26	0.69	-0.75	0.49
subjNeu	-0.45	-0.74	0.34	-0.07	-0.39	-0.39	-0.38	-0.74	0.90	-0.53
subjSoV	-0.28	-0.30	0.96	0.30	0.02	0.02	-0.88	-0.48	0.41	-0.44
subjVol	-0.50	-0.44	0.48	-0.06	-0.30	-0.30	-0.49	-0.53	0.45	-0.95
share	-0.10	-0.11	0.75	0.36	0.10	0.10	-0.74	-0.26	0.20	-0.31
shareOfNegDistribution	-0.54	-0.74	0.47	-0.05	-0.36	-0.36	-0.43	-0.87	0.79	-0.49
normalized negative	-0.57	-0.85	0.37	-0.10	-0.43	-0.43	-0.36	-0.78	0.88	-0.51
normalized neutral	0.48	0.74	-0.41	0.00	0.31	0.31	0.40	0.78	-1.00	0.46
normalized positive	0.56	0.44	0.09	0.62	0.87	0.87	-0.10	0.20	-0.20	0.28
pollEuro	0.01	-0.03	-0.02	-0.19	-0.04	-0.04	-0.01	-0.03	-0.07	-0.01
pollEuro-1	0.07	0.18	-0.07	0.03	0.07	0.07	0.09	0.01	-0.01	0.02
positive mentions	0.41	0.35	0.24	0.54	0.64	0.64	-0.26	0.08	-0.09	0.04
negative mentions	-0.53	-0.46	0.46	-0.08	-0.32	-0.32	-0.47	-0.54	0.46	-0.98
neutral mentions	-0.35	-0.29	0.45	0.05	-0.15	-0.15	-0.44	-0.32	0.25	-0.68
total mentions	-0.38	-0.32	0.45	0.02	-0.18	-0.18	-0.43	-0.37	0.30	-0.78

Table A.11: Aggregators correlation for Jerónimo de Sousa (PCP)

	normalized polarity	polarity Over Neutral	polarity Over Total	subjNeu	subjSoV	subjVol	share	share Of Neg Distribution	normalized negative	normalized neutral
bermingham2										
normalized bermingham2										
bermingham (sovn)										
bermingham (sovp)										
connor										
normalized connor										
gayo										
normalized gayo										
ind										
polarity										
normalized polarity	1.00									
polarityONeutral	0.90	1.00								
polarityOTotal	1.00	0.90	1.00							
subjNeu	-0.77	-0.87	-0.77	1.00						
subjSoV	-0.27	-0.26	-0.27	0.34	1.00					
subjVol	-0.45	-0.50	-0.45	0.52	0.48	1.00				
share	-0.08	-0.10	-0.08	0.16	0.75	0.35	1.00			
shareOfNegDistribution	-0.74	-0.77	-0.74	0.74	0.46	0.47	0.24	1.00		
normalized negative	-0.86	-0.90	-0.86	0.85	0.35	0.47	0.16	0.82	1.00	
normalized neutral	0.75	0.79	0.75	-0.90	-0.41	-0.45	-0.20	-0.79	-0.88	1.00
normalized positive	0.45	0.41	0.45	-0.28	0.13	-0.23	0.19	-0.26	-0.31	0.20
pollEuro	-0.02	-0.03	-0.02	-0.04	-0.03	-0.03	0.03	-0.04	0.02	0.07
pollEuro-1	0.17	0.11	0.17	-0.04	-0.06	0.02	-0.03	-0.02	-0.07	0.01
positive mentions	0.34	0.28	0.34	-0.14	0.28	0.02	0.30	-0.12	-0.21	0.09
negative mentions	-0.47	-0.52	-0.47	0.52	0.46	0.98	0.33	0.48	0.49	-0.46
neutral mentions	-0.26	-0.27	-0.26	0.27	0.46	0.71	0.41	0.25	0.27	-0.25
total mentions	-0.31	-0.34	-0.31	0.35	0.45	0.82	0.42	0.29	0.32	-0.30

Table A.12: Aggregators correlation for Jerónimo de Sousa (PCP)

	normalized positive	pollEuro	pollEuro-1	positive mentions	negative mentions	neutral mentions	total mentions
bermingham2							
normalized bermingham2							
bermingham (sovn)							
bermingham (sovp)							
connor							
normalized connor							
gayo							
normalized gayo							
ind							
polarity							
normalized polarity							
polarityONeutral							
polarityOTotal							
subjNeu							
subjSoV							
subjVol							
share							
shareOfNegDistribution							
normalized negative							
normalized neutral							
normalized positive	1.00						
pollEuro	0.00	1.00					
pollEuro-1	0.11	-0.31	1.00				
positive mentions	0.74	-0.10	0.18	1.00			
negative mentions	-0.26	-0.01	0.00	-0.01	1.00		
neutral mentions	-0.08	0.02	-0.02	0.12	0.69	1.00	
total mentions	-0.13	-0.02	0.02	0.10	0.80	0.90	1.00

Table A.13: Aggregators correlation for João Semedo and Catarina Martins (BE)

	bermingham2	normalized bermingham2	bermingham (sovn)	bermingham (sovp)	connor	normalized connor	gayo	normalized gayo	ind	polarity
bermingham2	1.00									
normalized bermingham2	0.62	1.00								
bermingham (sovn)	-0.21	-0.12	1.00							
bermingham (sovp)	0.21	0.09	0.27	1.00						
connor	0.46	0.34	-0.04	0.38	1.00					
normalized connor	0.46	0.34	-0.04	0.38	1.00	1.00				
gayo	0.18	0.10	-0.97	-0.25	0.03	0.03	1.00			
normalized gayo	0.19	0.16	-0.52	-0.05	0.19	0.19	0.54	1.00		
ind	-0.35	-0.66	0.11	0.14	0.00	0.00	-0.10	-0.10	1.00	
polarity	0.44	0.33	-0.37	-0.19	0.03	0.03	0.35	0.05	-0.33	1.00
normalized polarity	0.58	0.93	-0.10	0.06	0.27	0.27	0.08	0.18	-0.73	0.34
polarityONeutral	0.44	0.67	-0.05	0.02	0.12	0.12	0.03	0.03	-0.69	0.33
polarityOTotal	0.58	0.93	-0.10	0.06	0.27	0.27	0.08	0.18	-0.73	0.34
subjNeu	-0.37	-0.61	0.07	0.03	-0.02	-0.02	-0.05	-0.02	0.74	-0.32
subjSoV	-0.21	-0.12	0.99	0.27	-0.04	-0.04	-0.96	-0.53	0.11	-0.37
subjVol	-0.38	-0.32	0.34	0.22	0.00	0.00	-0.32	-0.05	0.36	-0.94
share	-0.15	-0.11	0.82	0.33	0.00	0.00	-0.84	-0.42	0.10	-0.44
shareOfNegDistribution	-0.18	-0.18	0.53	0.15	-0.13	-0.13	-0.51	-0.80	0.25	-0.11
normalized negative	-0.43	-0.77	0.11	0.07	-0.11	-0.11	-0.10	-0.11	0.89	-0.35
normalized neutral	0.35	0.66	-0.11	-0.14	0.00	0.00	0.10	0.10	-1.00	0.33
normalized positive	0.40	0.20	0.02	0.35	0.83	0.83	-0.03	0.05	0.14	-0.04
pollEuro	-0.07	-0.13	-0.14	-0.09	0.08	0.08	0.16	0.14	0.06	-0.19
pollEuro-1	0.28	0.23	-0.14	-0.12	0.10	0.10	0.15	0.02	-0.24	0.34
positive mentions	0.14	-0.04	0.26	0.53	0.37	0.37	-0.25	-0.09	0.32	-0.38
negative mentions	-0.42	-0.33	0.35	0.19	-0.01	-0.01	-0.33	-0.03	0.35	-0.98
neutral mentions	-0.29	-0.12	0.40	0.24	0.00	0.00	-0.37	-0.04	0.03	-0.68
total mentions	-0.36	-0.23	0.38	0.25	0.04	0.04	-0.36	-0.06	0.24	-0.89

Table A.14: Aggregators correlation for João Semedo and Catarina Martins (BE)

	normalized polarity	polarity Over Neutral	polarity Over Total	subjNeu	subjSoV	subjVol	share	share Of Neg Distribution	normalized negative	normalized neutral
bermingham2										
normalized bermingham2										
bermingham (sovn)										
bermingham 8sovp)										
connor										
normalized connor										
gayo										
normalized gayo										
ind										
polarity										
normalized polarity	1.00									
polarityONeutral	0.73	1.00								
polarityOTotal	1.00	0.73	1.00							
subjNeu	-0.67	-0.91	-0.67	1.00						
subjSoV	-0.10	-0.05	-0.10	0.07	1.00					
subjVol	-0.36	-0.36	-0.36	0.35	0.34	1.00				
share	-0.10	-0.06	-0.10	0.06	0.82	0.40	1.00			
shareOfNegDistribution	-0.21	-0.10	-0.21	0.14	0.54	0.13	0.39	1.00		
normalized negative	-0.84	-0.74	-0.84	0.74	0.11	0.38	0.11	0.24	1.00	
normalized neutral	0.73	0.69	0.73	-0.74	-0.11	-0.36	-0.10	-0.25	-0.89	1.00
normalized positive	0.13	0.00	0.13	0.09	0.02	0.08	0.08	0.01	0.03	-0.14
pollEuro	-0.12	-0.03	-0.12	-0.01	-0.16	0.18	-0.12	-0.20	0.09	-0.06
pollEuro-1	0.24	0.22	0.24	-0.20	-0.14	-0.31	-0.19	-0.03	-0.27	0.24
positive mentions	-0.11	-0.12	-0.11	0.19	0.26	0.44	0.32	0.20	0.24	-0.32
negative mentions	-0.35	-0.35	-0.35	0.34	0.35	0.96	0.42	0.09	0.37	-0.35
neutral mentions	-0.08	-0.03	-0.08	0.02	0.40	0.64	0.47	0.03	0.07	-0.03
total mentions	-0.25	-0.24	-0.25	0.23	0.38	0.87	0.44	0.10	0.27	-0.24

Table A.15: Aggregators correlation for João Semedo and Catarina Martins (BE)

	normalized positive	pollEuro	pollEuro-1	positive mentions	negative mentions	neutral mentions	total mentions
bermingham2							
normalized bermingham2							
bermingham (sovn)							
bermingham (sovp)							
connor							
normalized connor							
gayo							
normalized gayo							
ind							
polarity							
normalized polarity							
polarityONeutral							
polarityOTotal							
subjNeu							
subjSoV							
subjVol							
share							
shareOfNegDistribution							
normalized negative							
normalized neutral							
normalized positive	1.00						
pollEuro	0.02	1.00					
pollEuro-1	0.10	0.09	1.00				
positive mentions	0.50	0.07	-0.29	1.00			
negative mentions	0.05	0.19	-0.35	0.41	1.00		
neutral mentions	-0.06	0.20	-0.32	0.32	0.65	1.00	
total mentions	0.06	0.21	-0.39	0.44	0.88	0.78	1.00

Appendix B

Graphical Representations

B.1 Experiment using absolute values

B.1.1 Ordinary Least Squares

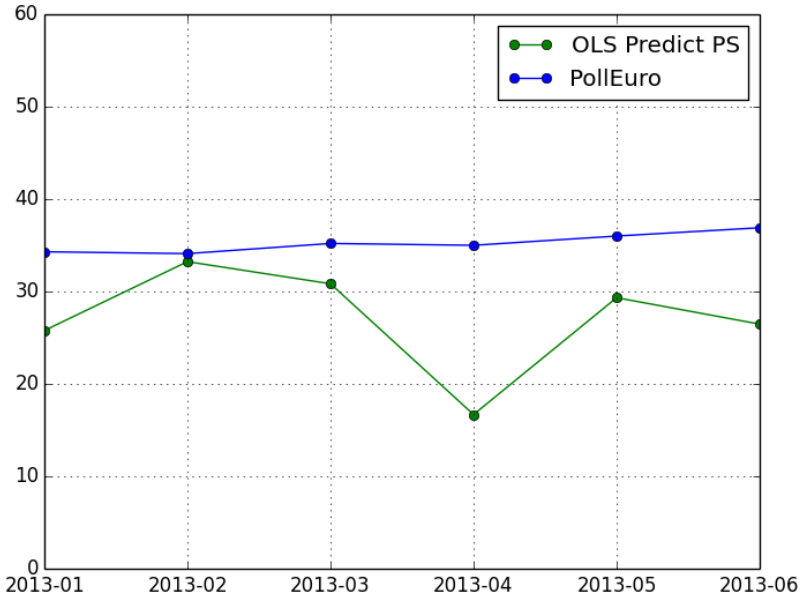

Figure B.1: Predicted and real values of vote intention, from January 2013 to June, for António José Seguro (PS), excluding the y_{t-1} feature

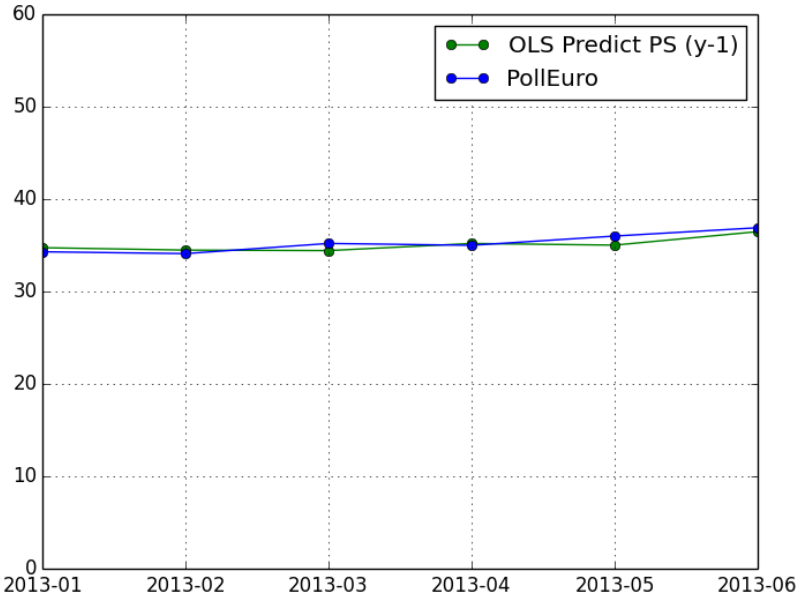

Figure B.2: Predicted and real values of vote intention, from January 2013 to June 2013, for António José Seguro (PS)

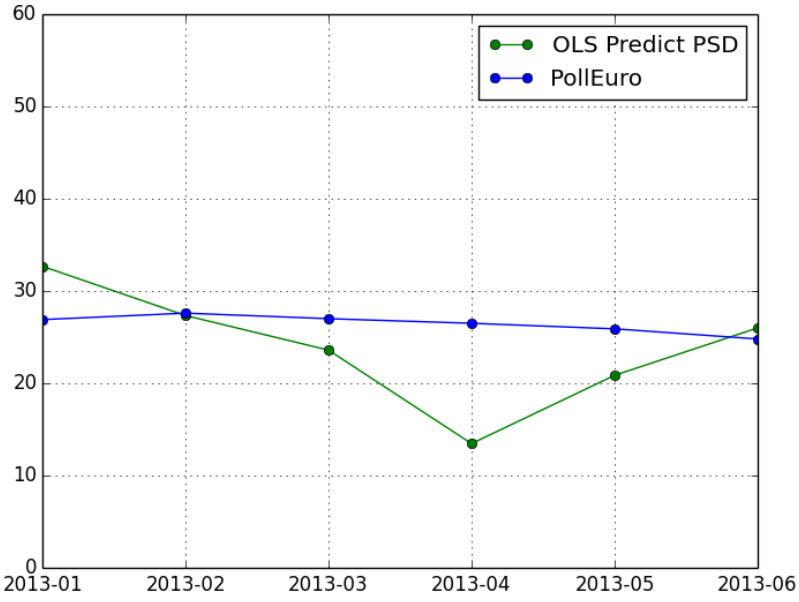

Figure B.3: Predicted and real values of vote intention, from January 2013 to June, for Pedro Passos Coelho (PSD), excluding the y_{t-1} feature

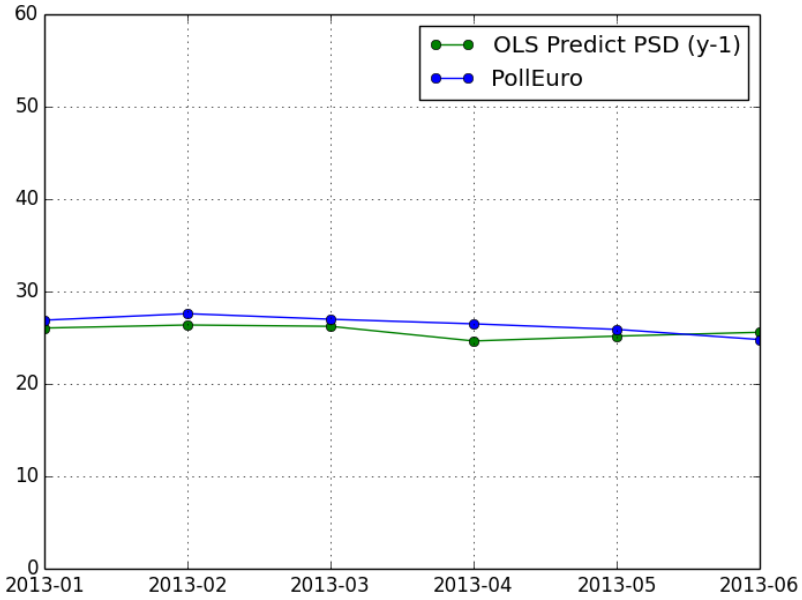

Figure B.4: Predicted and real values of vote intention, from January 2013 to June, for Pedro Passos Coelho (PSD)

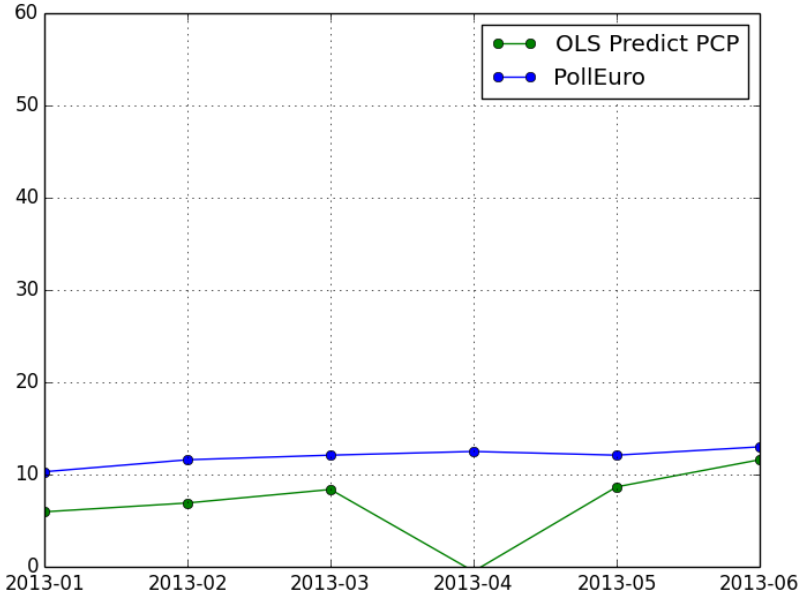

Figure B.5: Predicted and real values of vote intention, from January 2013 to June, for Jerónimo de Sousa (JS), excluding the y_{t-1} feature

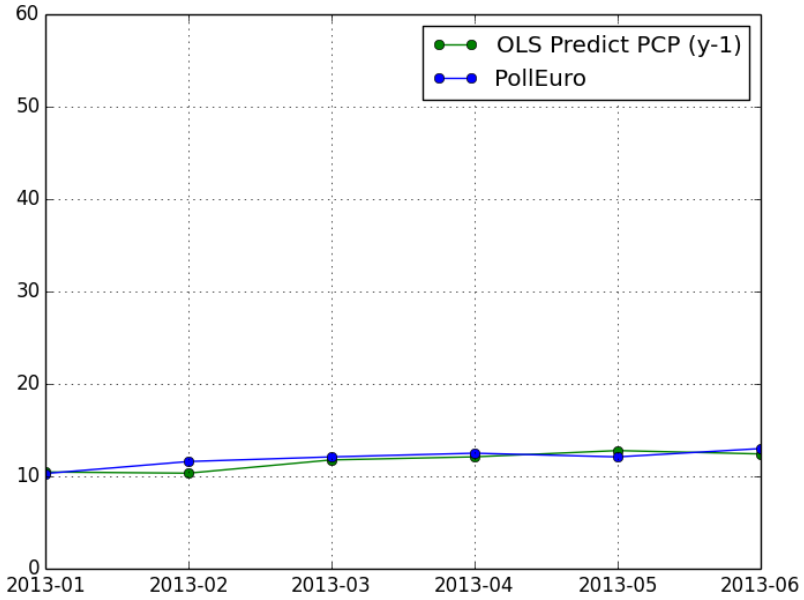

Figure B.6: Predicted and real values of vote intention, from January 2013 to June, for Jerónimo de Sousa (PCP)

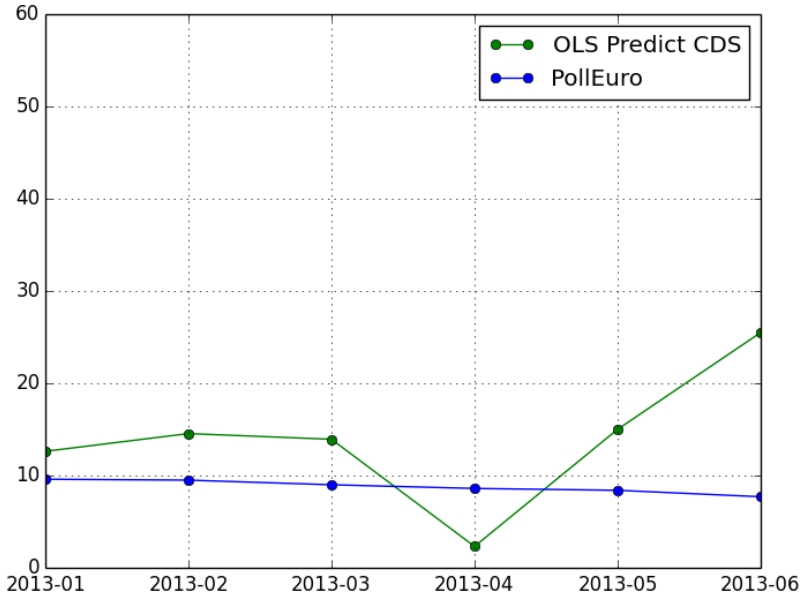

Figure B.7: Predicted and real values of vote intention, from January 2013 to June, for Paulo Portas (PP), excluding the y_{t-1} feature

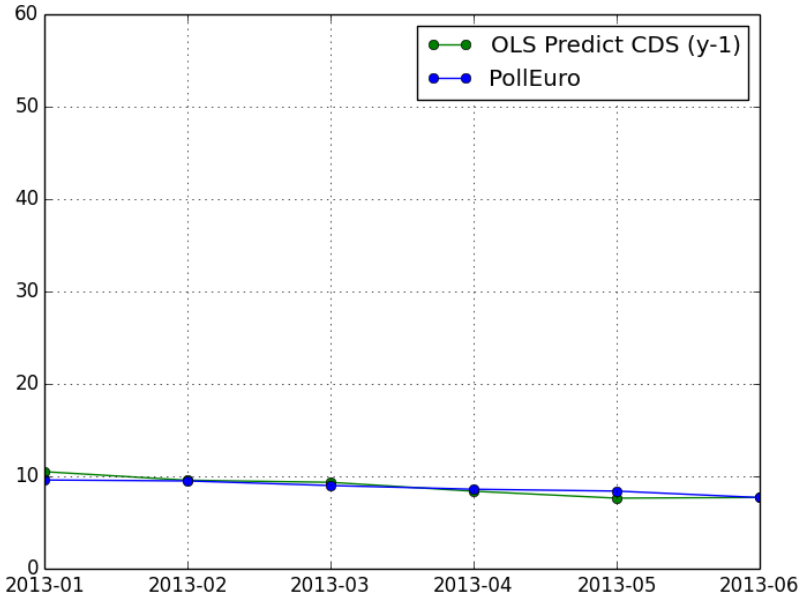

Figure B.8: Predicted and real values of vote intention, from January 2013 to June, for Paulo Portas (PP)

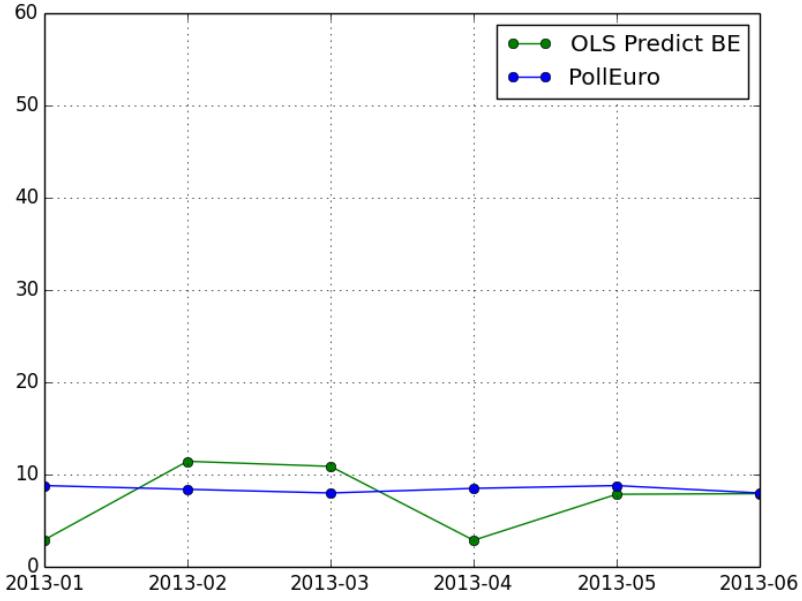

Figure B.9: Predicted and real values of vote intention, from January 2013 to June, for Catarina Martins e João Semedo (BE), excluding the y_{t-1} feature

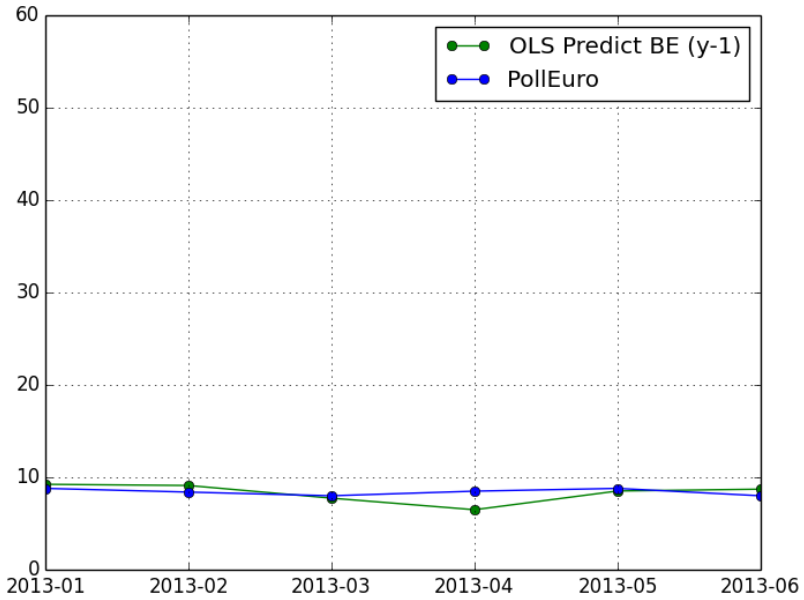

Figure B.10: Predicted and real values of vote intention, from January 2013 to June, for Catarina Martins e João Semedo (BE)

B.1.2 Random Forest

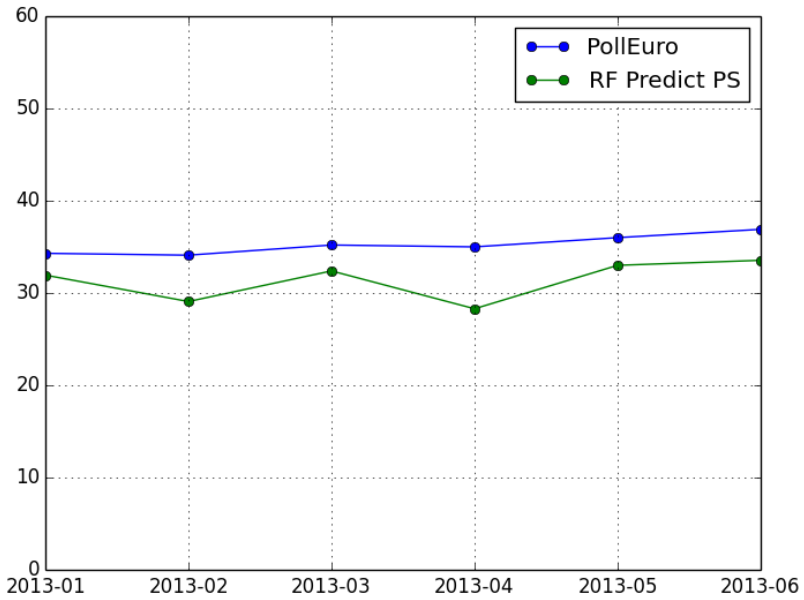

Figure B.11: Predicted and real values of vote intention, from January 2013 to June, for António José Seguro (PS), excluding the y_{t-1} feature.

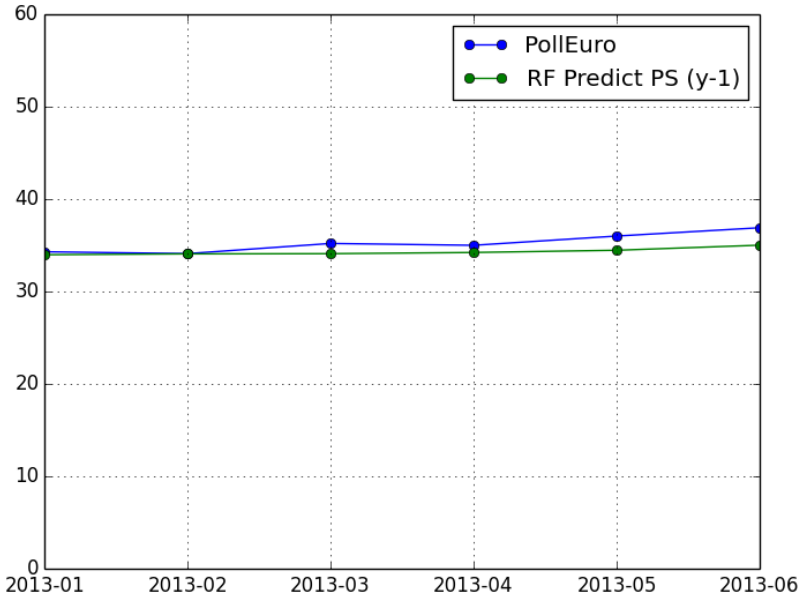

Figure B.12: Predicted and real values of vote intention, from January 2013 to June 2013, for António José Seguro (PS)

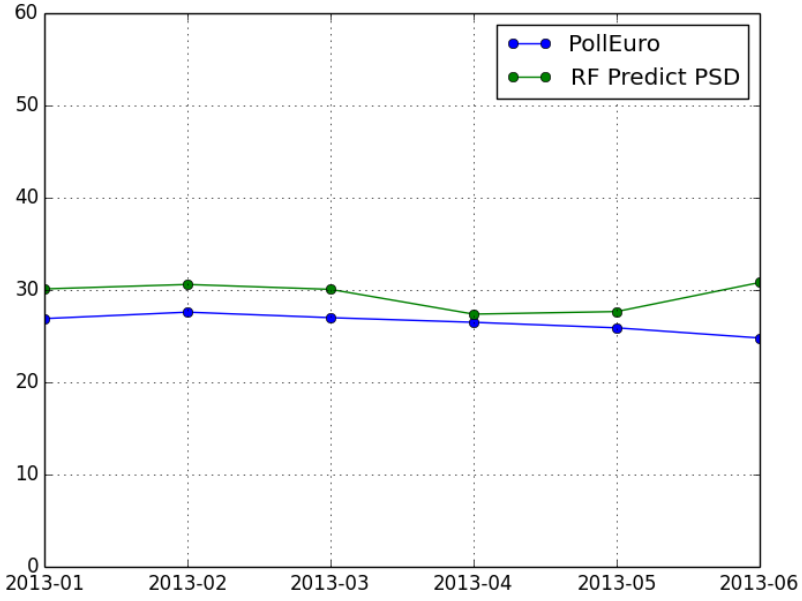

Figure B.13: Predicted and real values of vote intention, from January 2013 to June, for Pedro Passos Coelho (PSD), excluding the y_{t-1} feature.

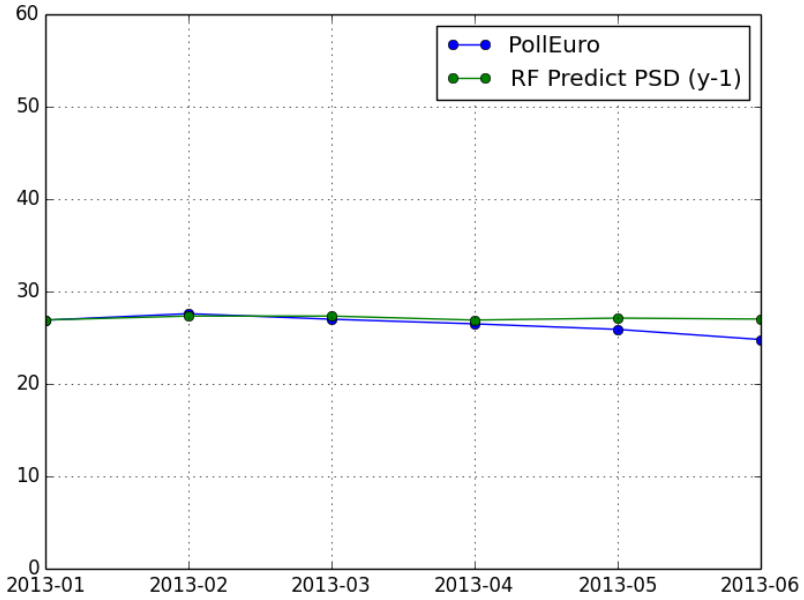

Figure B.14: Predicted and real values of vote intention, from January 2013 to June 2013, for Pedro Passos Coelho (PSD)

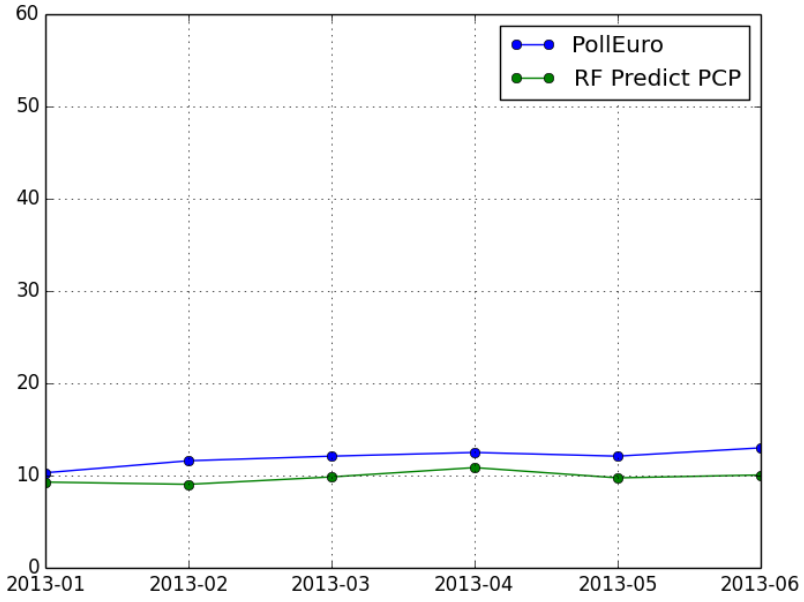

Figure B.15: Predicted and real values of vote intention, from January 2013 to June, for Jerónimo de Sousa (PCP), excluding the y_{t-1} feature.

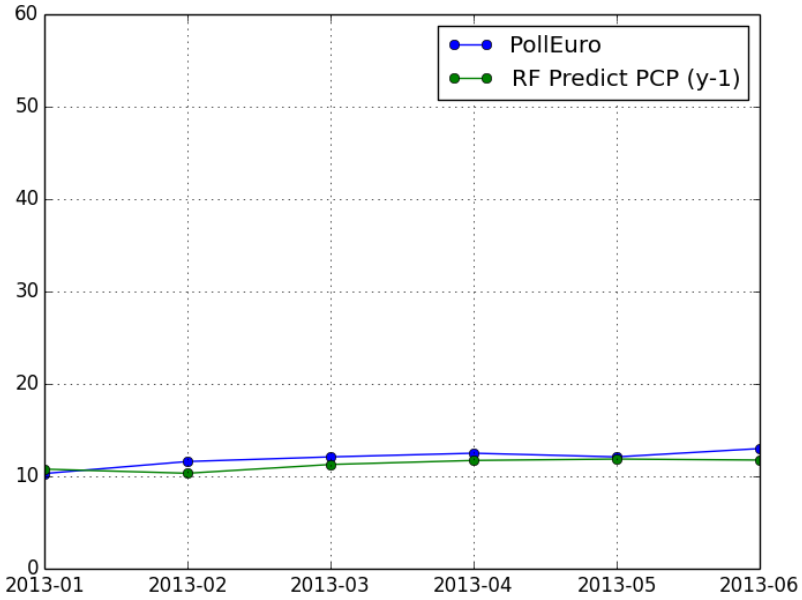

Figure B.16: Predicted and real values of vote intention, from January 2013 to June 2013, for Jerónimo de Sousa (PCP)

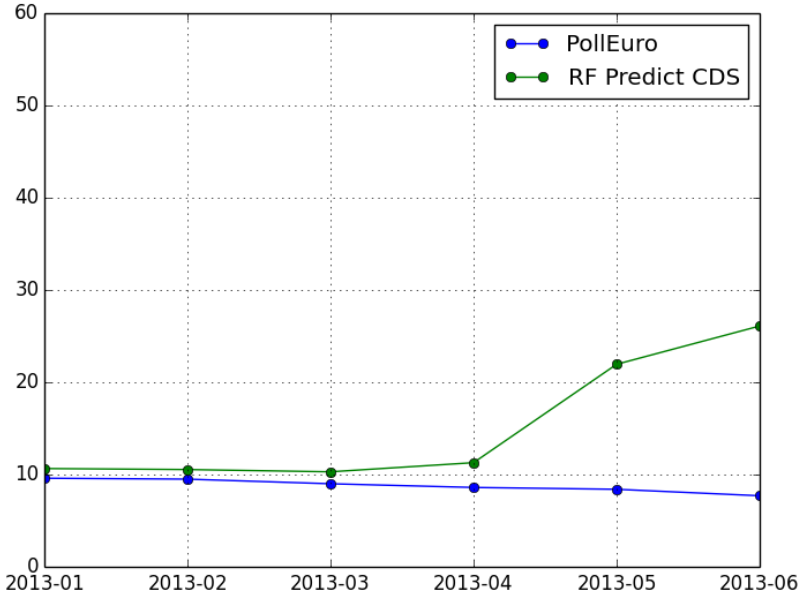

Figure B.17: Predicted and real values of vote intention, from January 2013 to June, for Paulo Portas (PP), excluding the y_{t-1} feature.

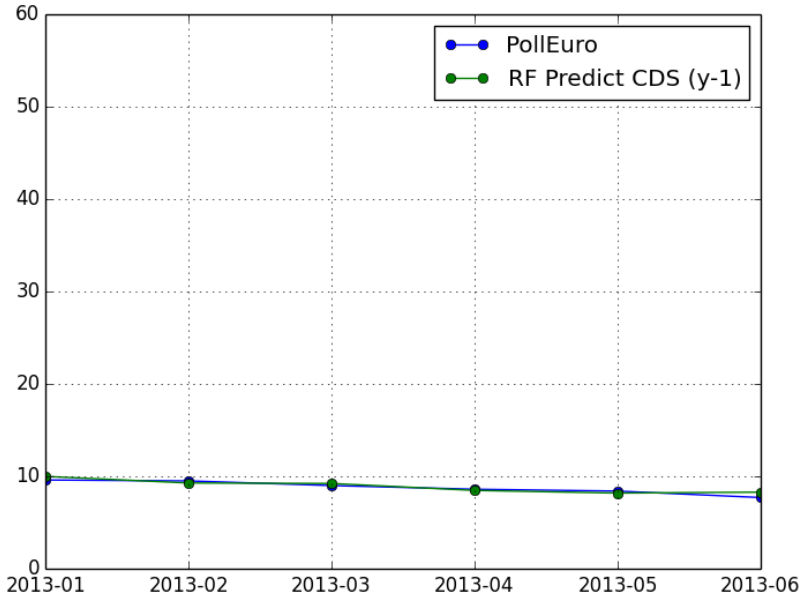

Figure B.18: Predicted and real values of vote intention, from January 2013 to June 2013, for Paulo Portas (PP)

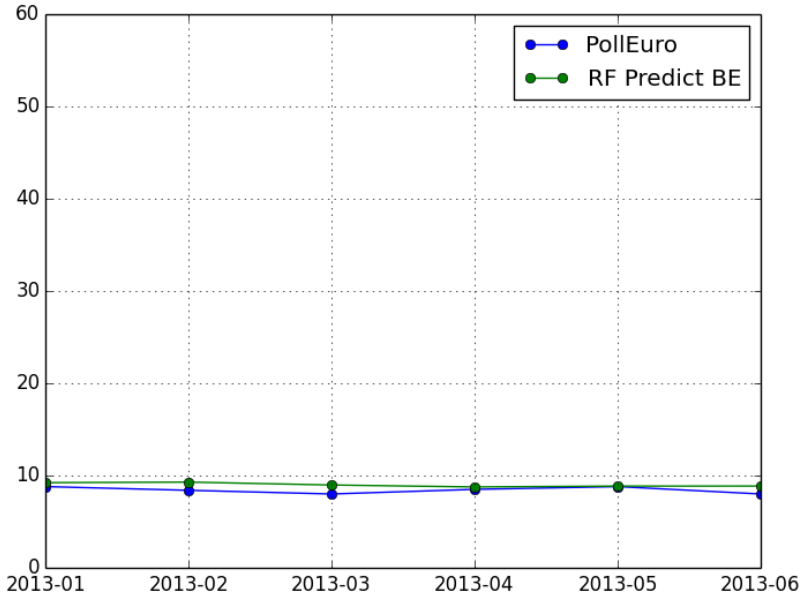

Figure B.19: Predicted and real values of vote intention, from January 2013 to June, for Catarina Martins e João Semedo (BE), excluding the y_{t-1} feature.

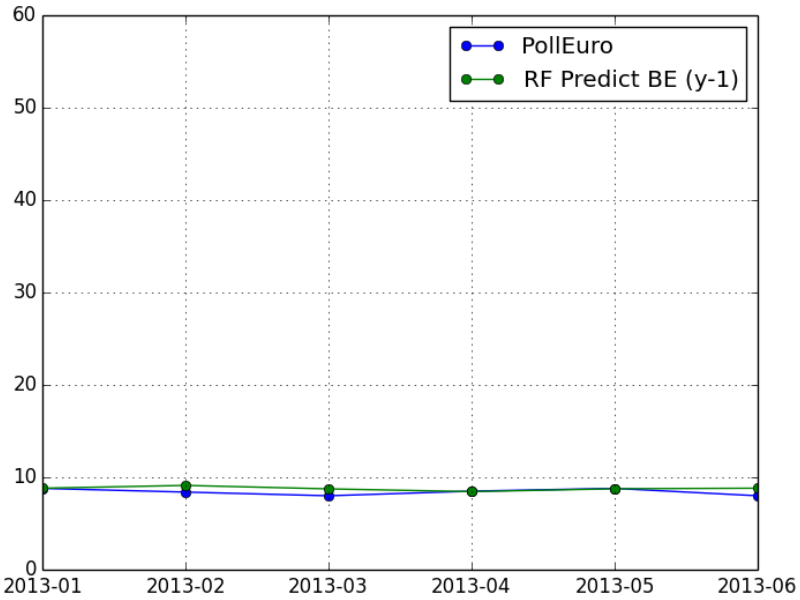

Figure B.20: Predicted and real values of vote intention, from January 2013 to June 2013, for Catarina Martins e João Semedo (BE)

B.2 Experiment using monthly variation

B.2.1 Ordinary Least Squares

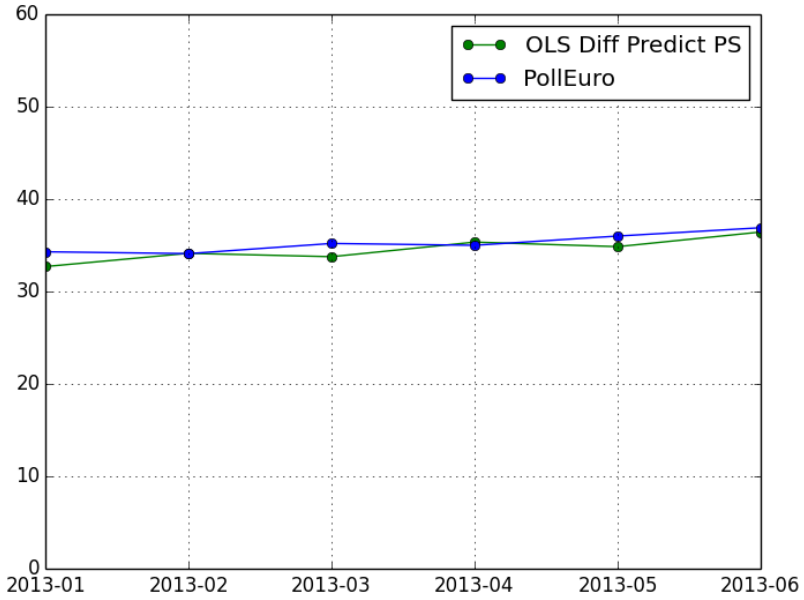

Figure B.21: Predicted and real values of vote intention, from January 2013 to June, for António José Seguro (PS), excluding the $\Delta(y_{t-1})$ feature

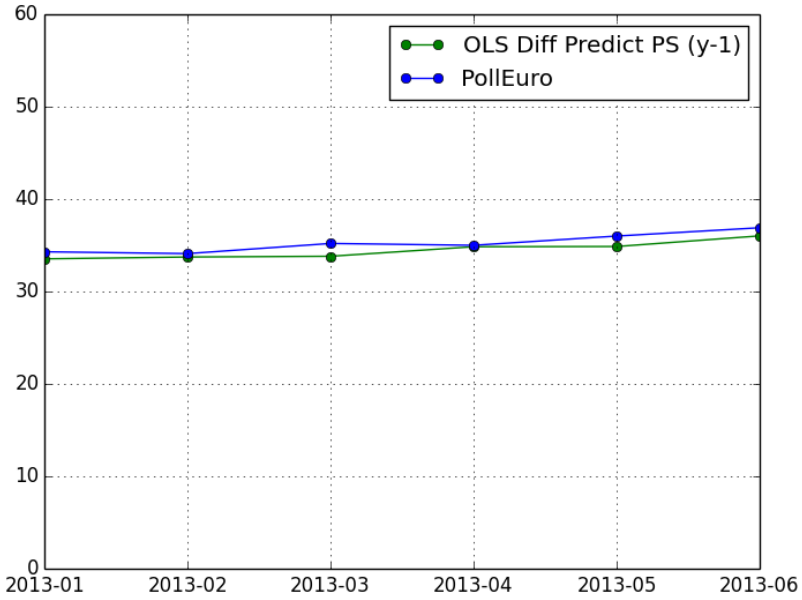

Figure B.22: Predicted and real values of vote intention, from January 2013 to June 2013, for António José Seguro (PS)

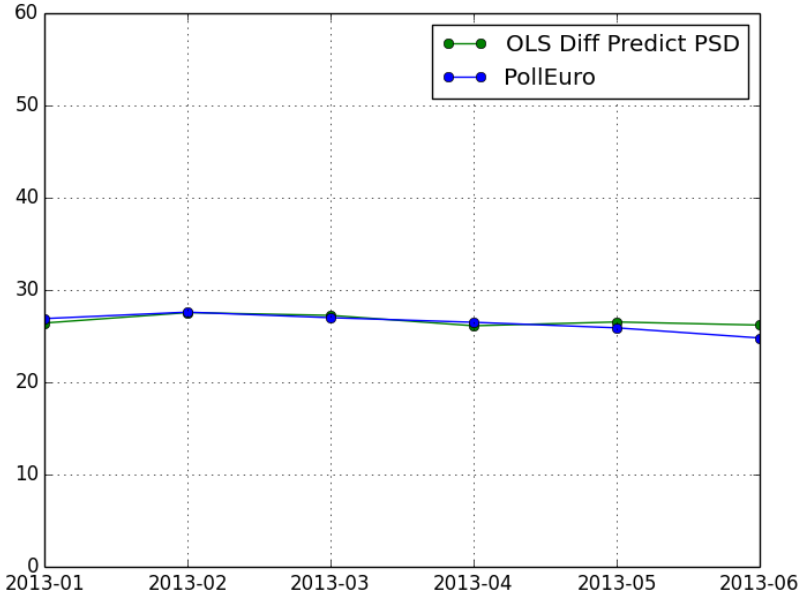

Figure B.23: Predicted and real values of vote intention, from January 2013 to June, for Pedro Passos Coelho (PSD), excluding the $\Delta(y_{t-1})$ feature

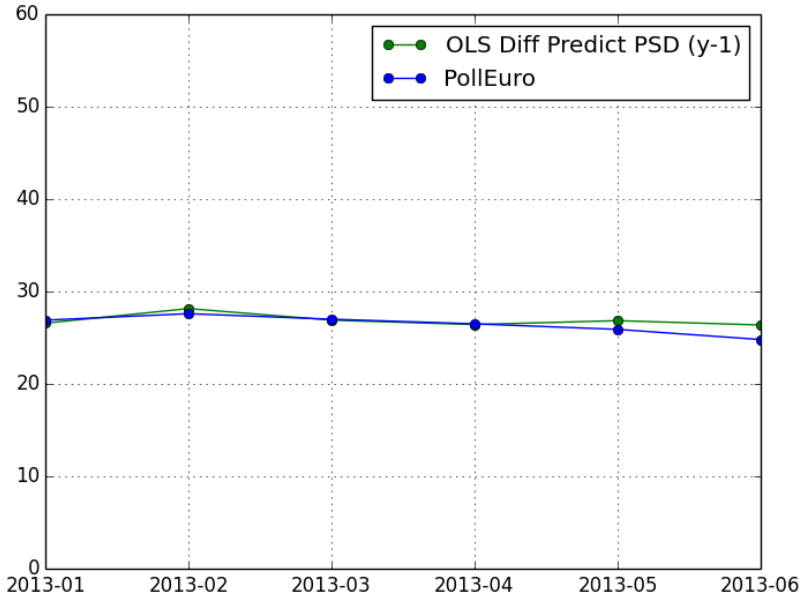

Figure B.24: Predicted and real values of vote intention, from January 2013 to June, for Pedro Passos Coelho (PSD)

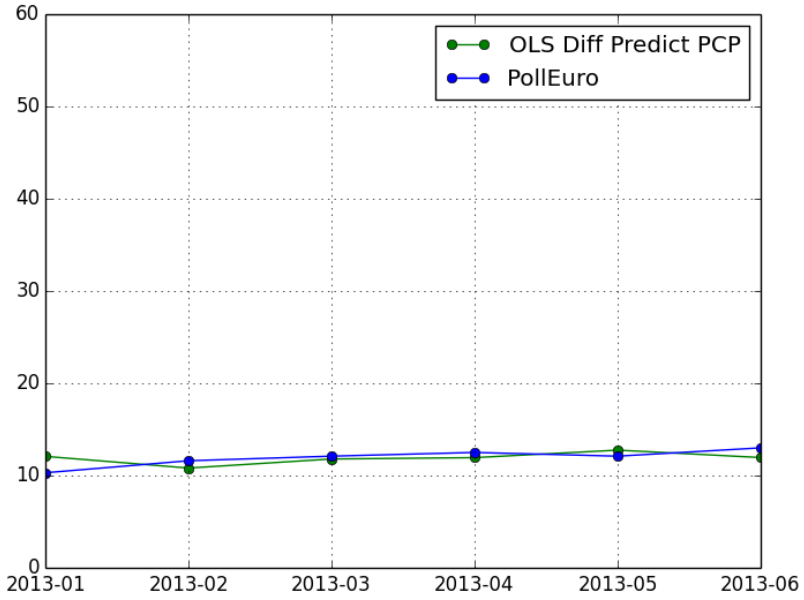

Figure B.25: Predicted and real values of vote intention, from January 2013 to June, forJerónimo de Sousa (JS), excluding the $\Delta(y_{t-1})$ feature

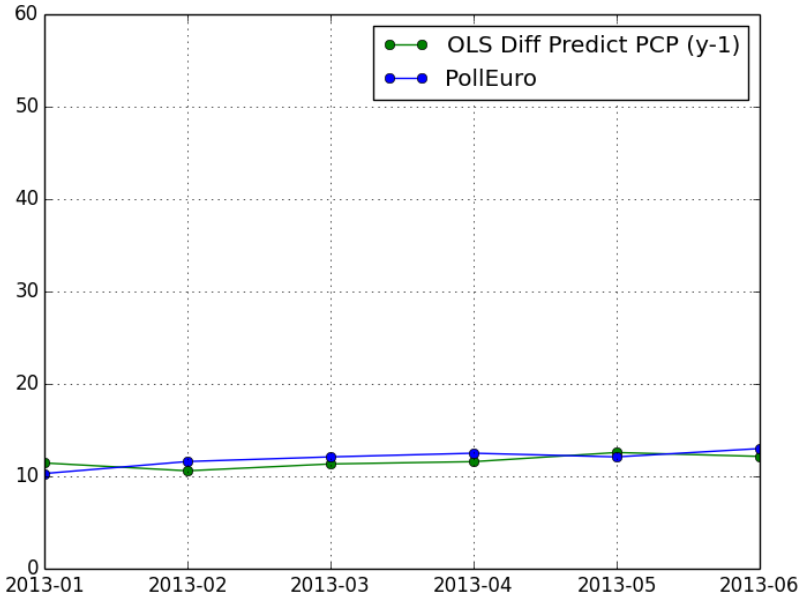

Figure B.26: Predicted and real values of vote intention, from January 2013 to June, for Jerónimo de Sousa (PCP)

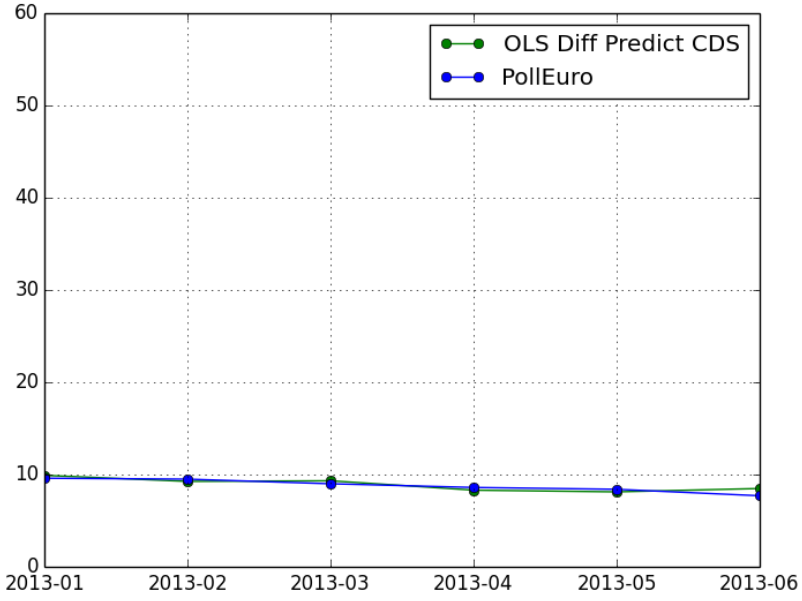

Figure B.27: Predicted and real values of vote intention, from January 2013 to June, for Paulo Portas (PP), excluding the $\Delta(y_{t-1})$ feature

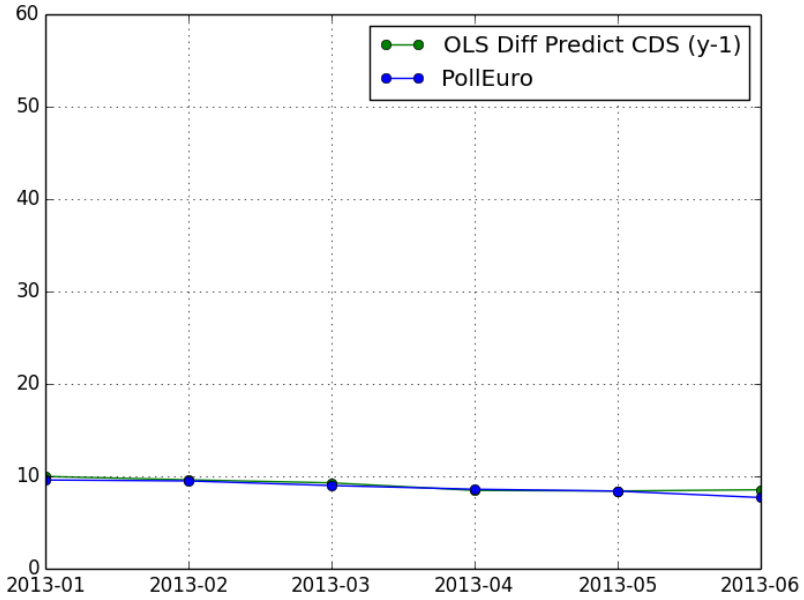

Figure B.28: Predicted and real values of vote intention, from January 2013 to June, for Paulo Portas (PP)

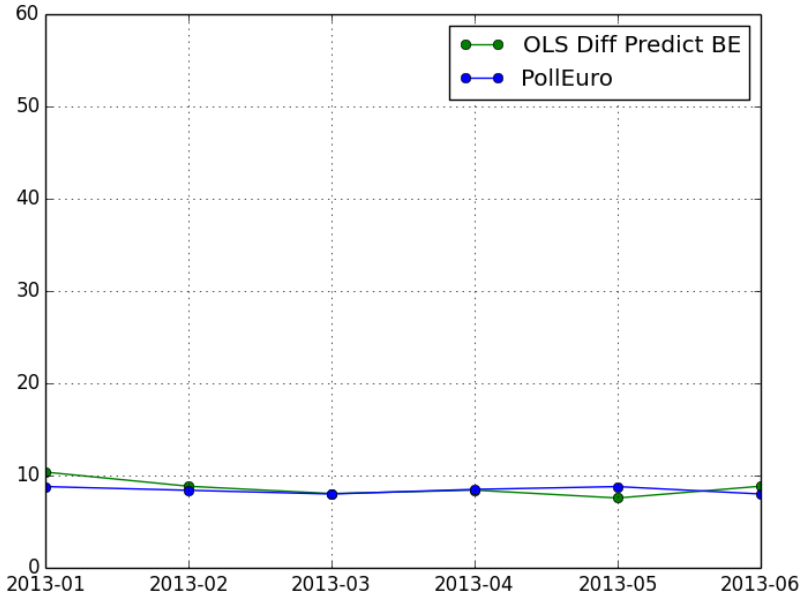

Figure B.29: Predicted and real values of vote intention, from January 2013 to June, for Catarina Martins e João Semedo (BE), excluding the $\Delta(y_{t-1})$ feature

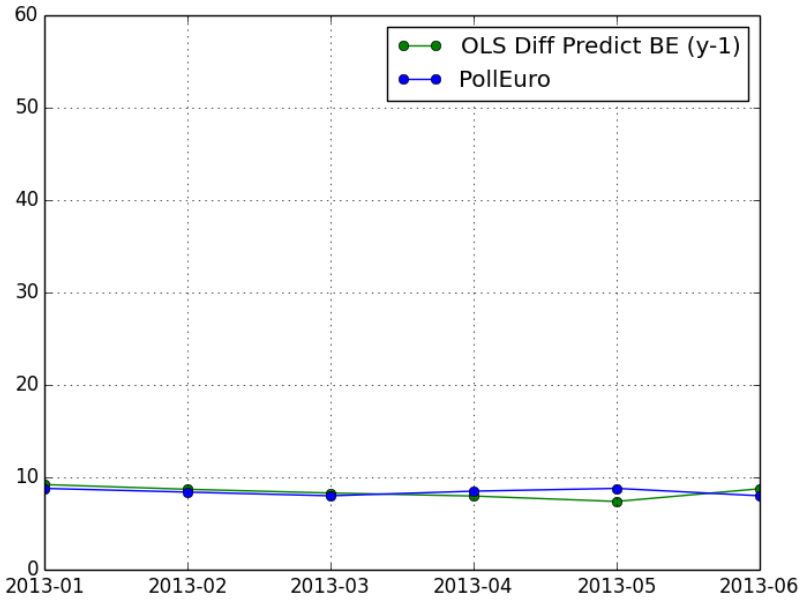

Figure B.30: Predicted and real values of vote intention, from January 2013 to June, forCatarina Martins e João Semedo (BE)

B.2.2 Random Forest

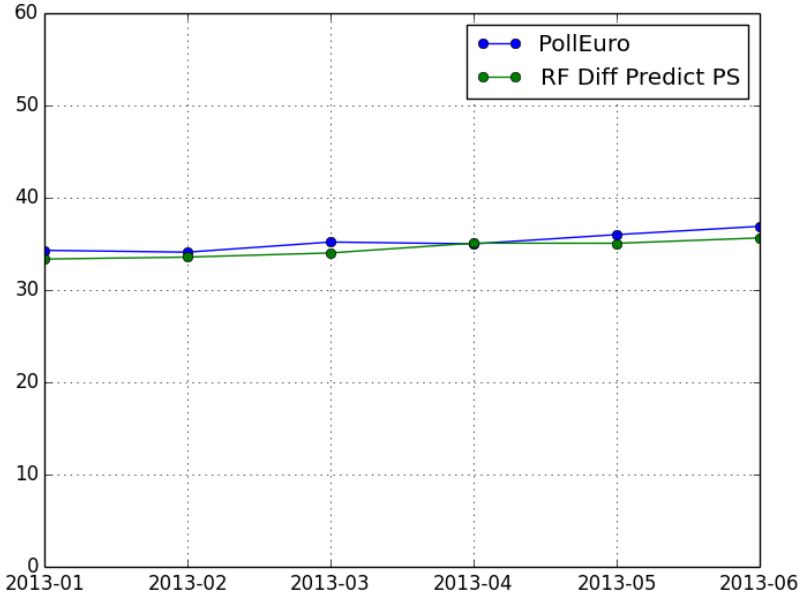

Figure B.31: Predicted and real values of vote intention, from January 2013 to June, for António José Seguro (PS), excluding the $\Delta(y_{t-1})$ feature.

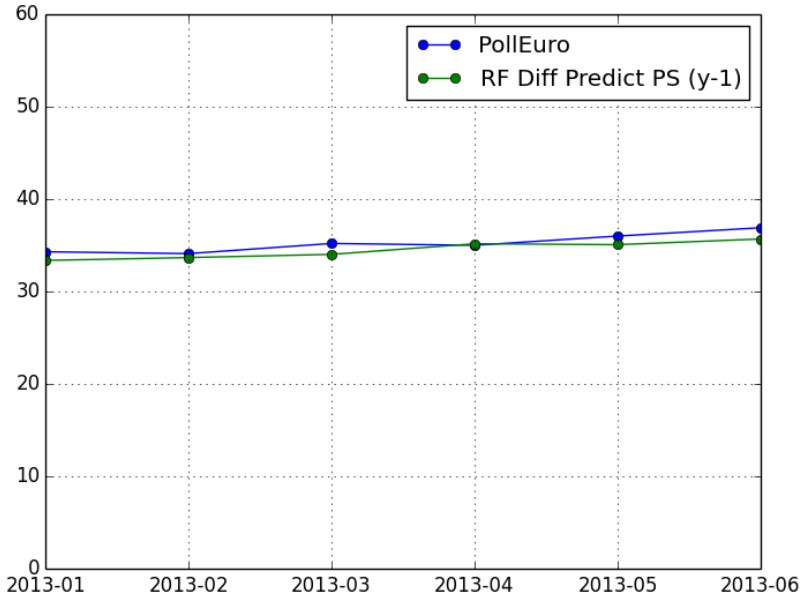

Figure B.32: Predicted and real values of vote intention, from January 2013 to June 2013, for António José Seguro (PS)

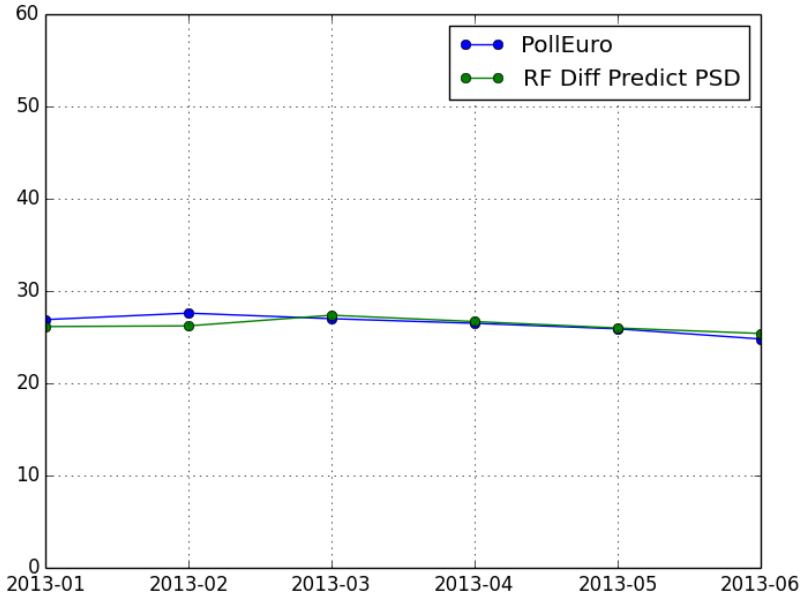

Figure B.33: Predicted and real values of vote intention, from January 2013 to June, for Pedro Passos Coelho (PSD), excluding the $\Delta(y_{t-1})$ feature.

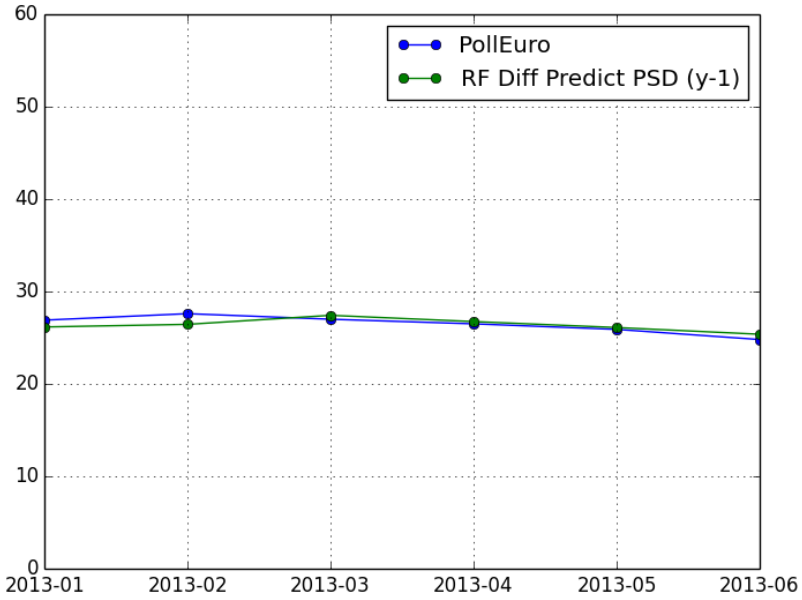

Figure B.34: Predicted and real values of vote intention, from January 2013 to June 2013, for Pedro Passos Coelho (PSD)

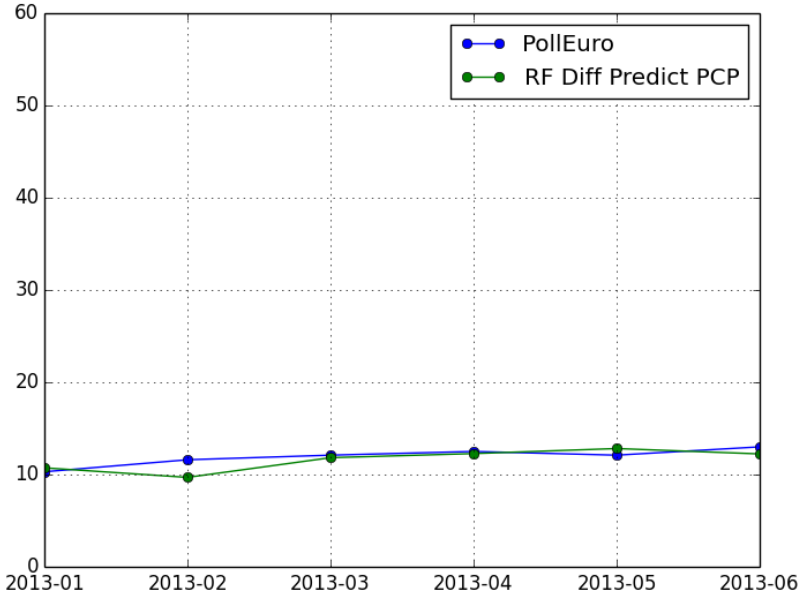

Figure B.35: Predicted and real values of vote intention, from January 2013 to June, for Jerónimo de Sousa (PCP), excluding the $\Delta(y_{t-1})$ feature.

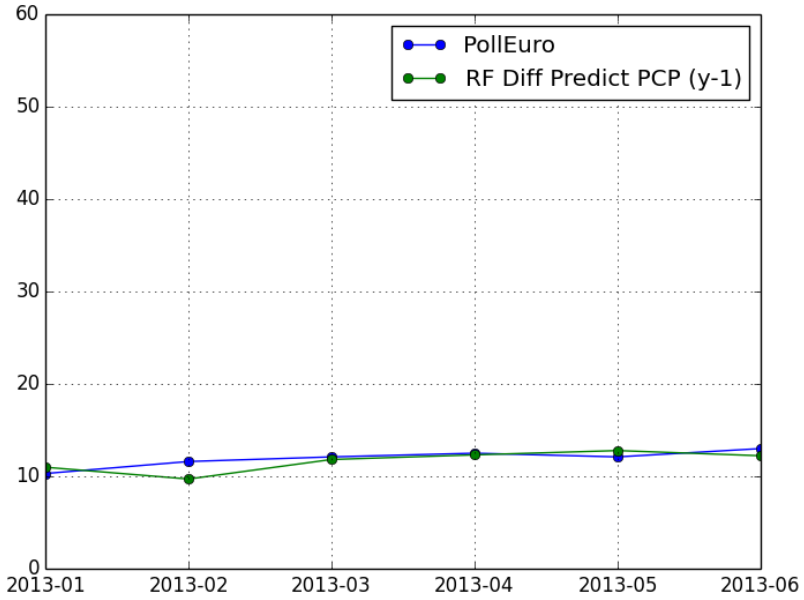

Figure B.36: Predicted and real values of vote intention, from January 2013 to June 2013, for Jerónimo de Sousa (PCP)

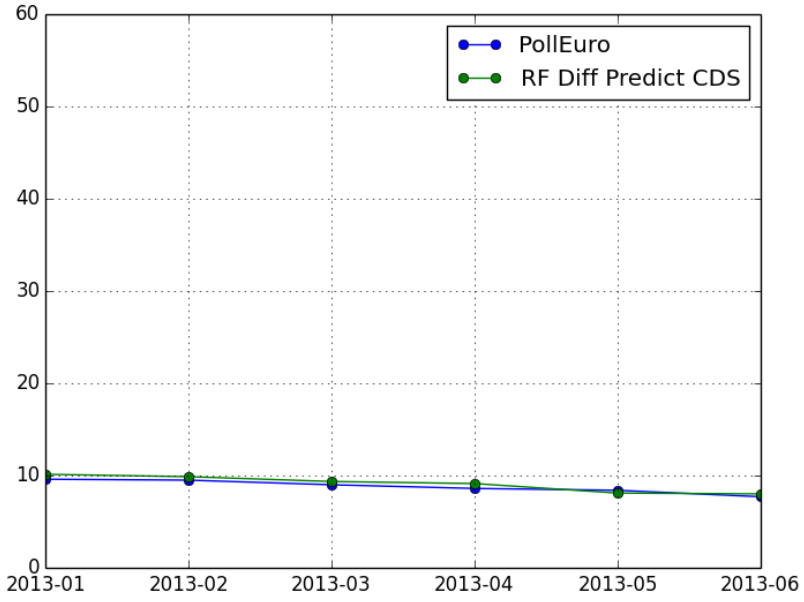

Figure B.37: Predicted and real values of vote intention, from January 2013 to June, for Paulo Portas (PP), excluding the $\Delta(y_{t-1})$ feature.

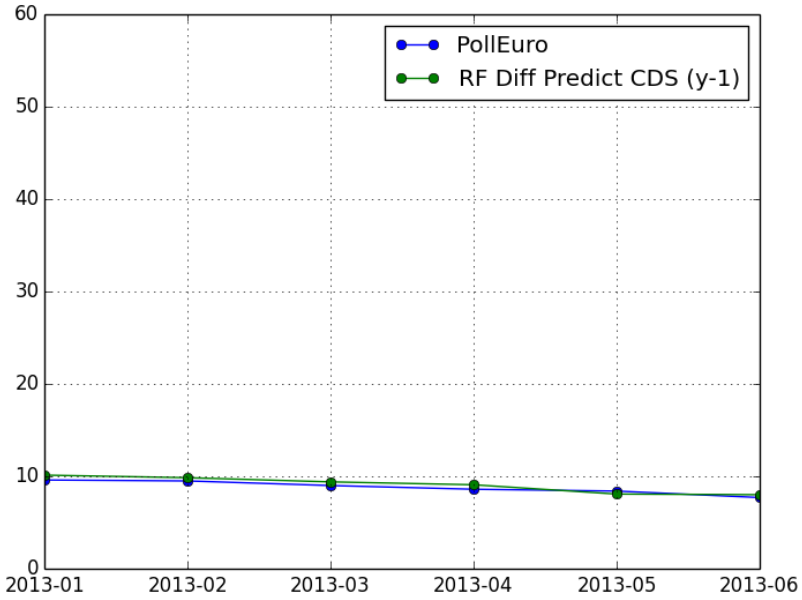

Figure B.38: Predicted and real values of vote int[!ht]ention, from January 2013 to June 2013, for Paulo Portas (PP)

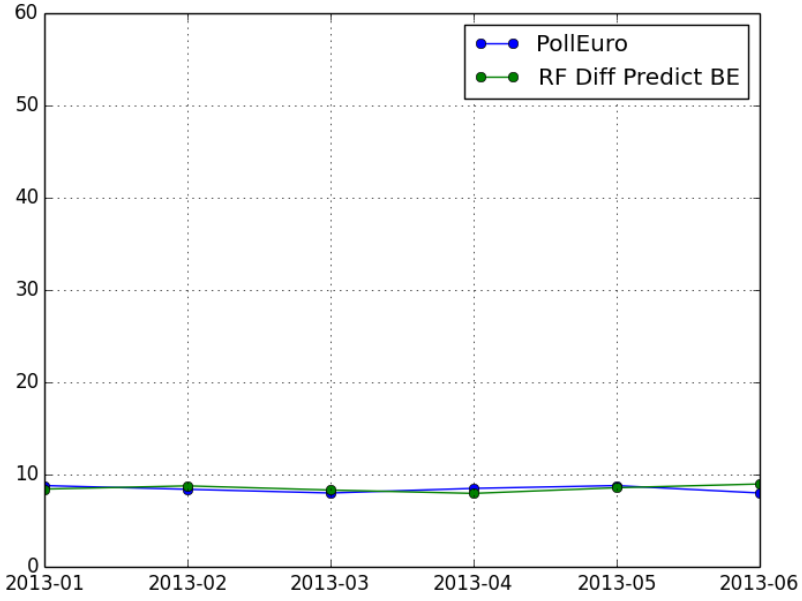

Figure B.39: Predicted and real values of vote intention, from January 2013 to June, for Catarina Martins e João Semedo (BE), excluding the $\Delta(y_{t-1})$ feature.

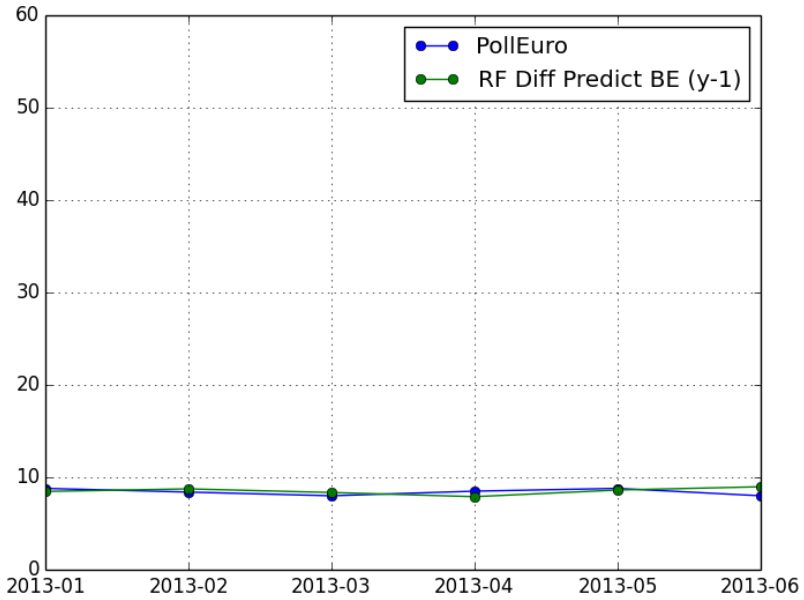

Figure B.40: Predicted and real values of vote intention, from January 2013 to June 2013, for Catarina Martins e João Semedo (BE)

B.3 Experiment Sentiment vs Buzz

B.3.1 Sentiment

B.3.1.1 Ordinary Least Squares

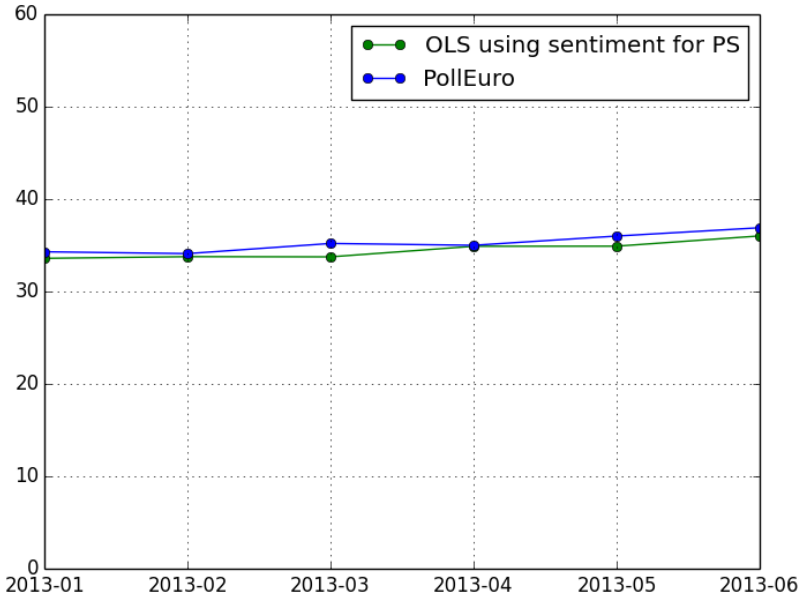

Figure B.41: Predicted and real values of vote intention, from January 2013 to June, for António José Seguro (PS), excluding the $\Delta(y_{t-1})$ feature.

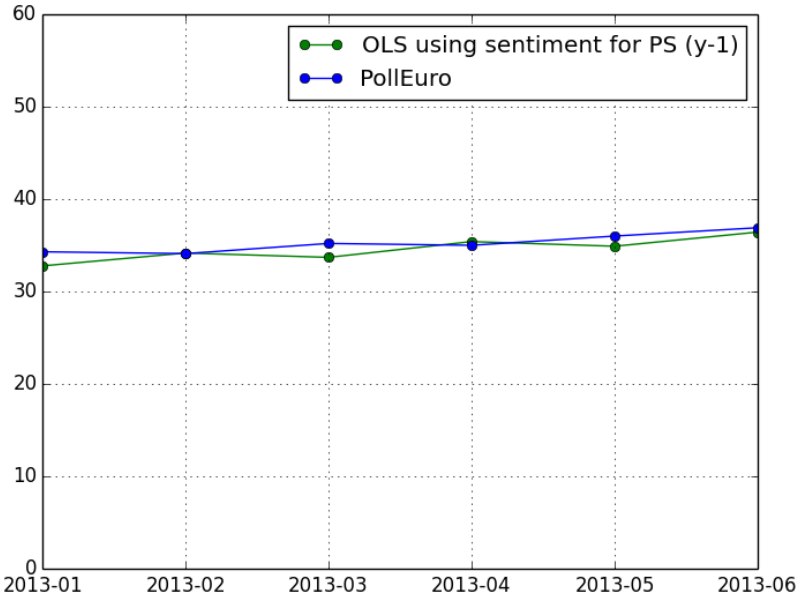

Figure B.42: Predicted and real values of vote intention, from January 2013 to June 2013, for António José Seguro (PS)

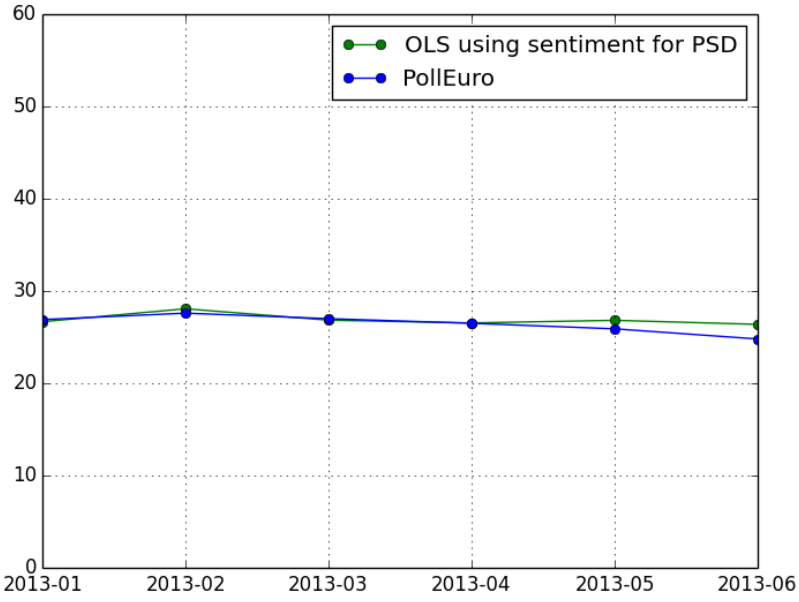

Figure B.43: Predicted and real values of vote intention, from January 2013 to June, for Pedro Passos Coelho (PSD), excluding the $\Delta(y_{t-1})$ feature.

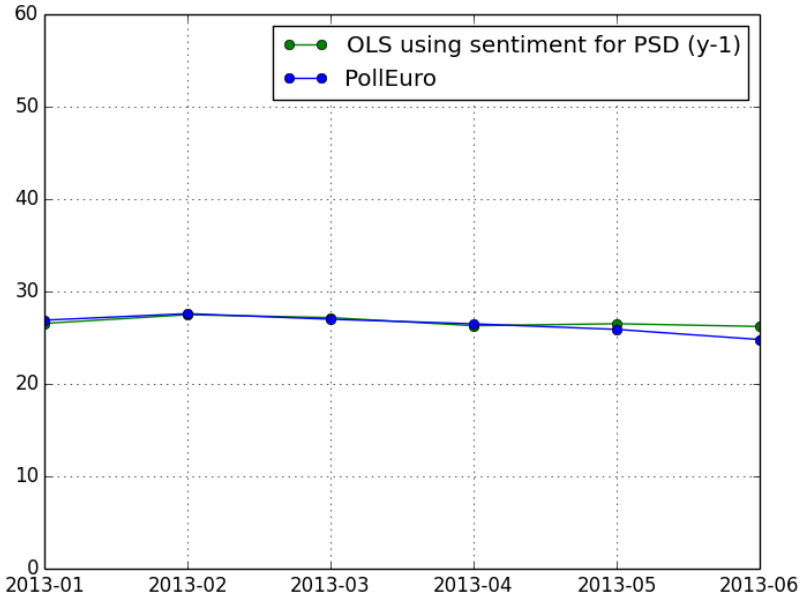

Figure B.44: Predicted and real values of vote intention, from January 2013 to June 2013, for Pedro Passos Coelho (PSD)

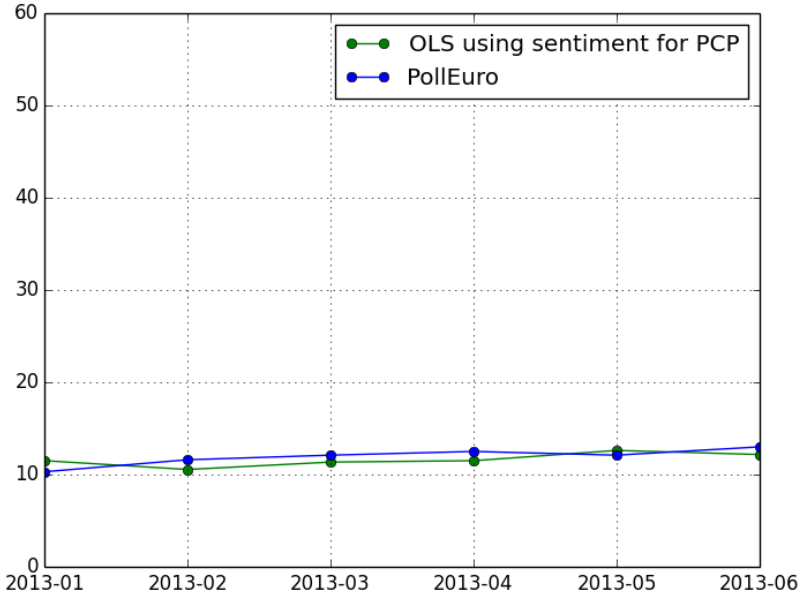

Figure B.45: Predicted and real values of vote intention, from January 2013 to June, for Jerónimo de Sousa (PCP), excluding the $\Delta(y_{t-1})$ feature.

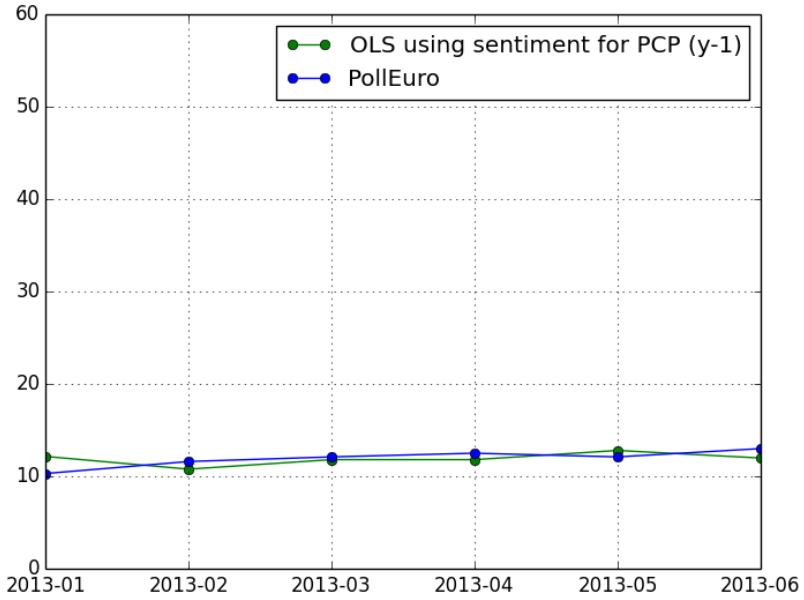

Figure B.46: Predicted and real values of vote intention, from January 2013 to June 2013, for Jerónimo de Sousa (PCP)

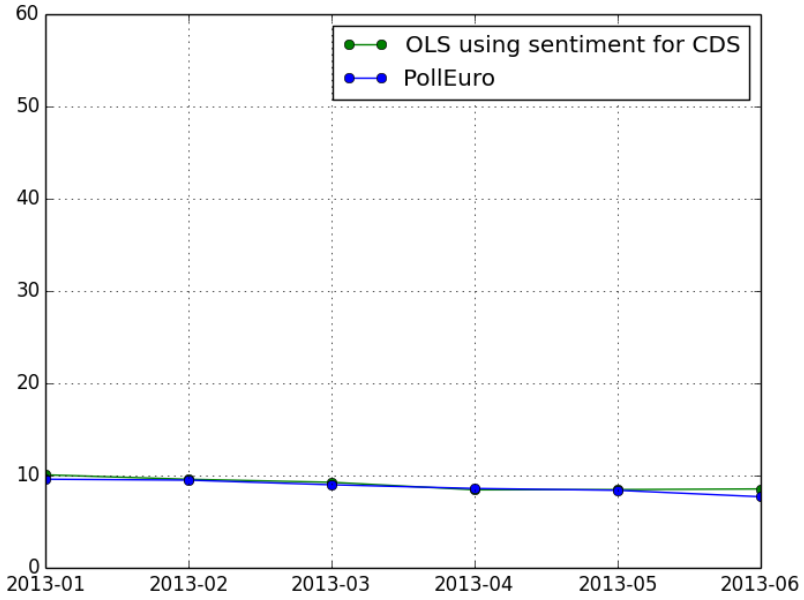

Figure B.47: Predicted and real values of vote intention, from January 2013 to June, for Paulo Portas (PP), excluding the $\Delta(y_{t-1})$ feature.

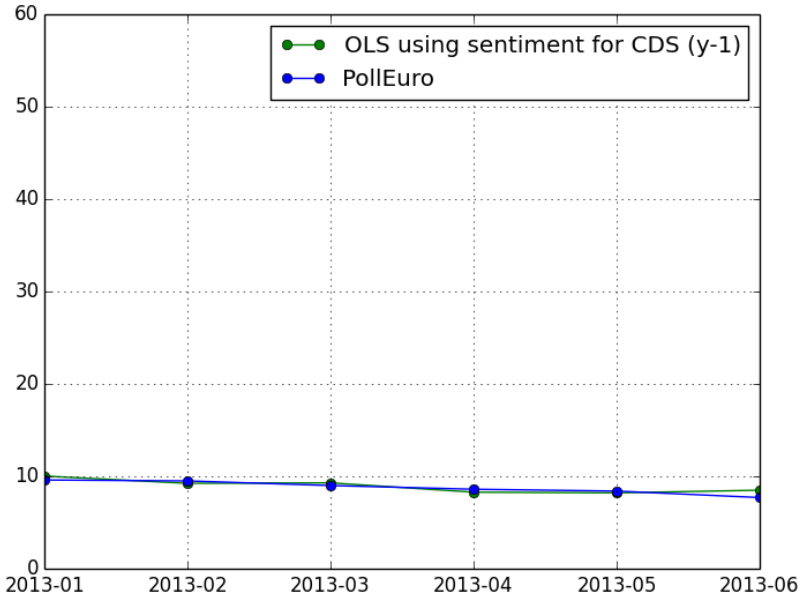

Figure B.48: Predicted and real values of vote int[!ht]ention, from January 2013 to June 2013, for Paulo Portas (PP)

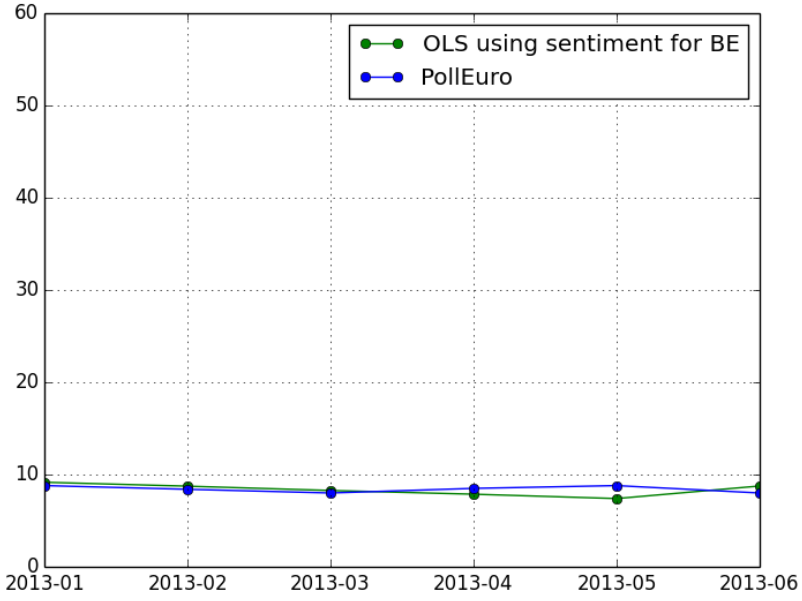

Figure B.49: Predicted and real values of vote intention, from January 2013 to June, for Catarina Martins e João Semedo (BE), excluding the $\Delta(y_{t-1})$ feature.

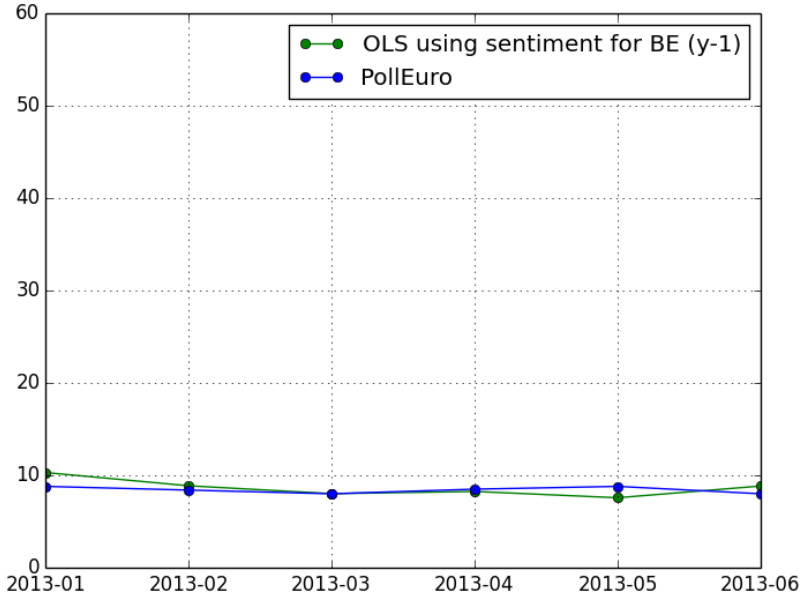

Figure B.50: Predicted and real values of vote intention, from January 2013 to June 2013, for Catarina Martins e João Semedo (BE)

B.3.1.2 Random Forest

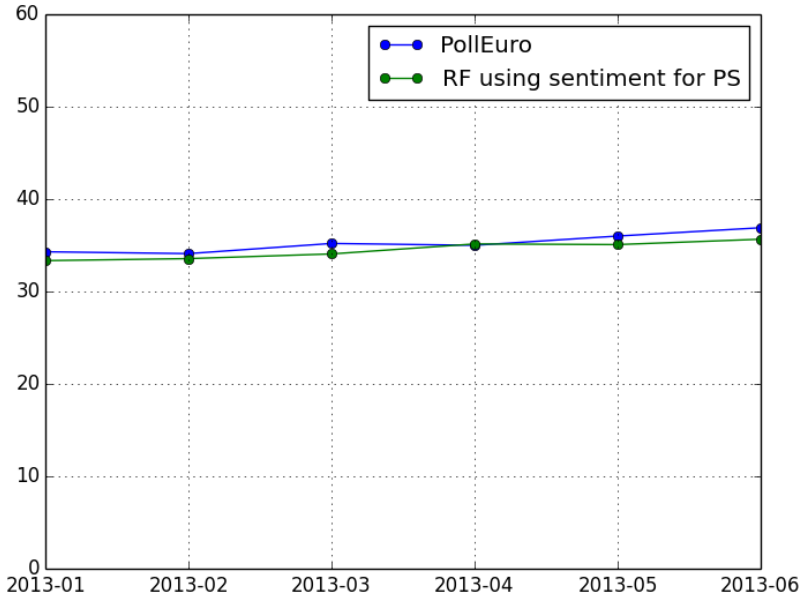

Figure B.51: Predicted and real values of vote intention, from January 2013 to June, for António José Seguro (PS), excluding the $\Delta(y_{t-1})$ feature.

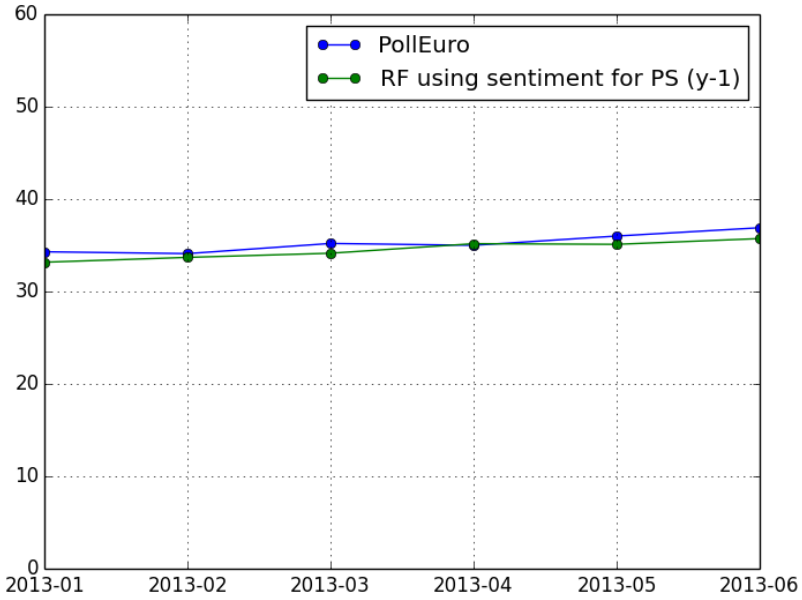

Figure B.52: Predicted and real values of vote intention, from January 2013 to June 2013, for António José Seguro (PS)

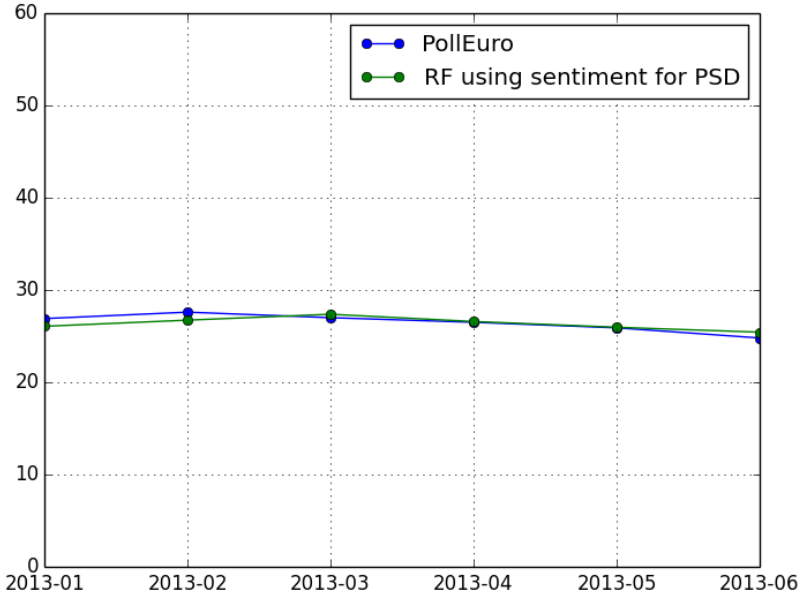

Figure B.53: Predicted and real values of vote intention, from January 2013 to June, for Pedro Passos Coelho (PSD), excluding the $\Delta(y_{t-1})$ feature.

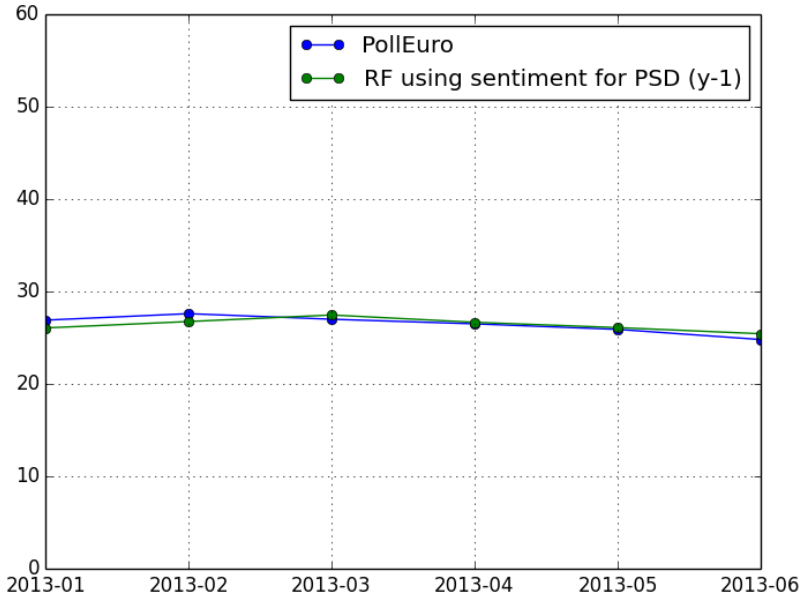

Figure B.54: Predicted and real values of vote intention, from January 2013 to June 2013, for Pedro Passos Coelho (PSD)

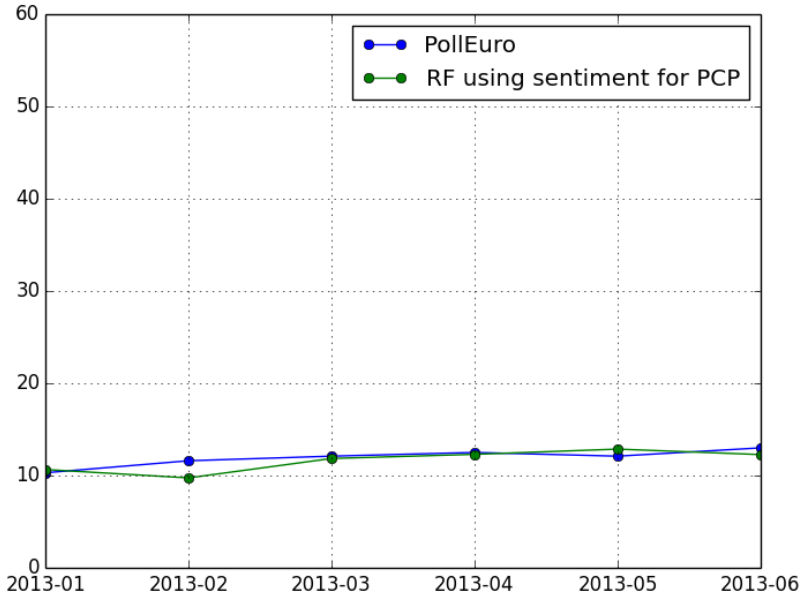

Figure B.55: Predicted and real values of vote intention, from January 2013 to June, for Jerónimo de Sousa (PCP), excluding the $\Delta(y_{t-1})$ feature.

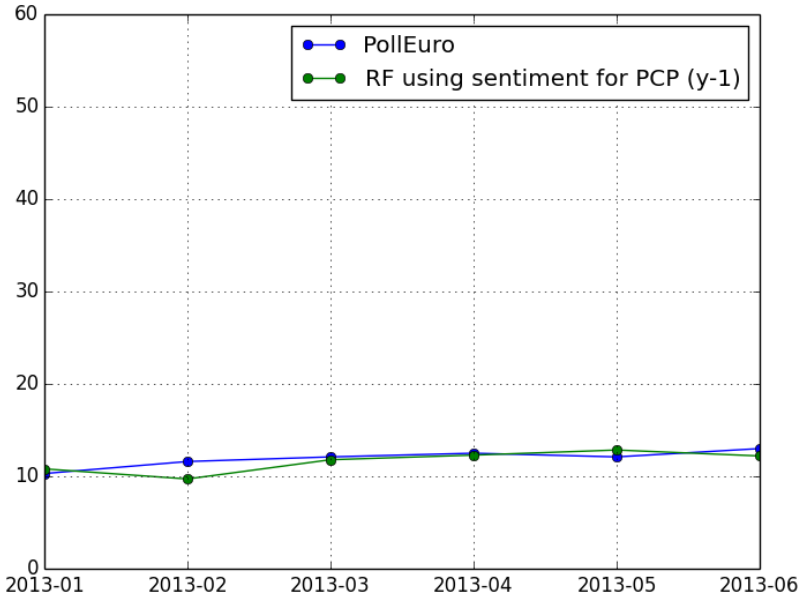

Figure B.56: Predicted and real values of vote intention, from January 2013 to June 2013, for Jerónimo de Sousa (PCP)

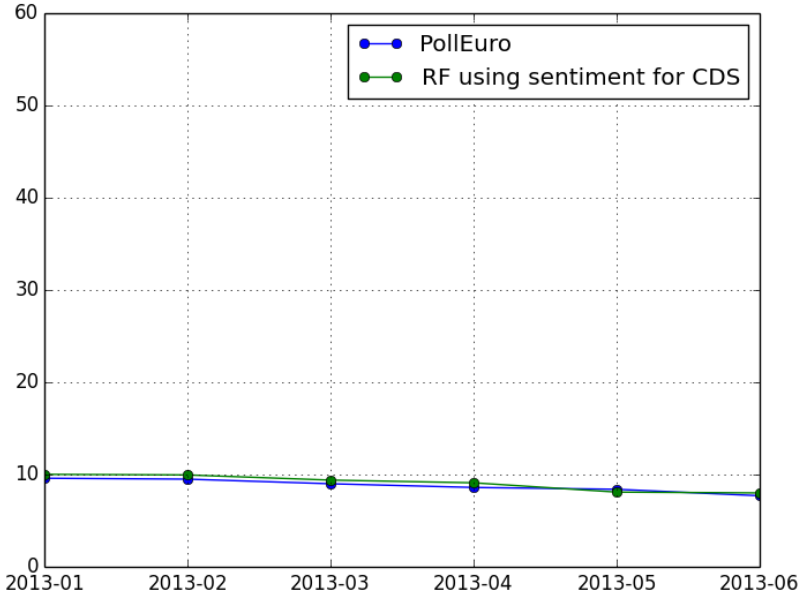

Figure B.57: Predicted and real values of vote intention, from January 2013 to June, for Paulo Portas (PP), excluding the $\Delta(y_{t-1})$ feature.

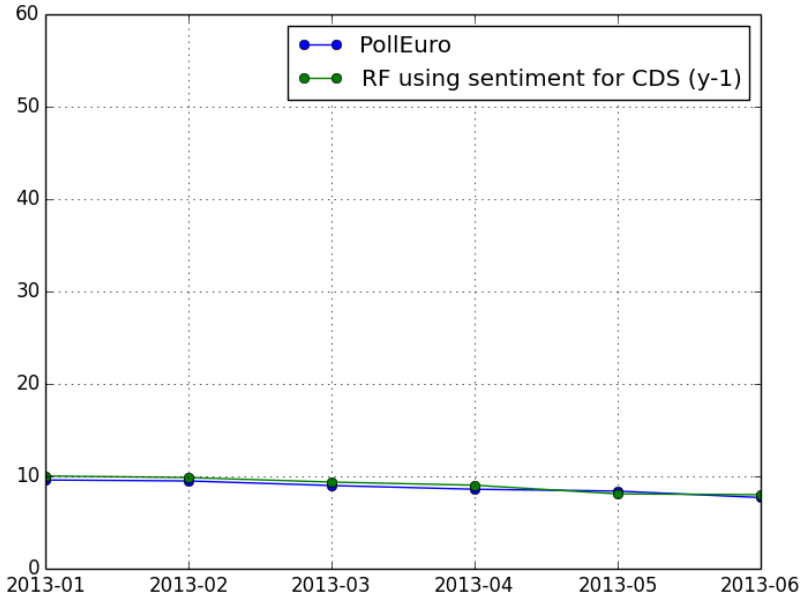

Figure B.58: Predicted and real values of vote int[!ht]ention, from January 2013 to June 2013, for Paulo Portas (PP)

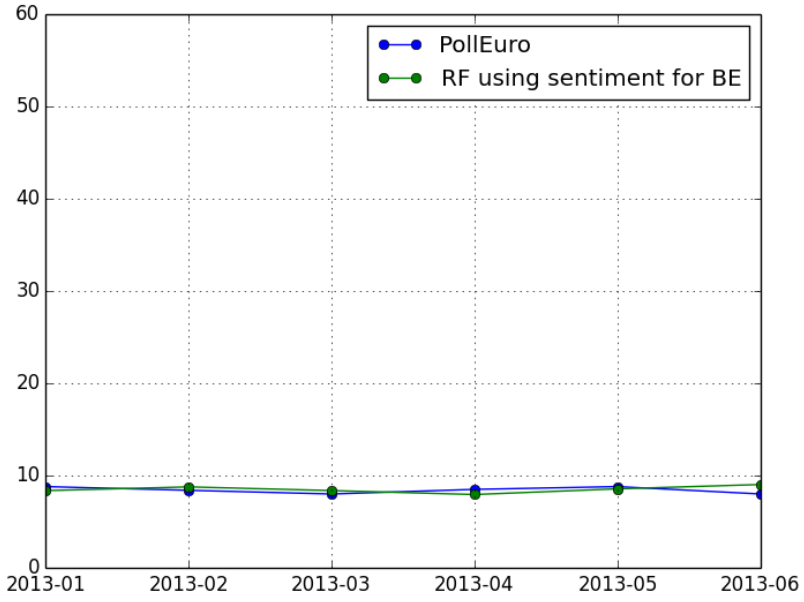

Figure B.59: Predicted and real values of vote intention, from January 2013 to June, for Catarina Martins e João Semedo (BE), excluding the $\Delta(y_{t-1})$ feature.

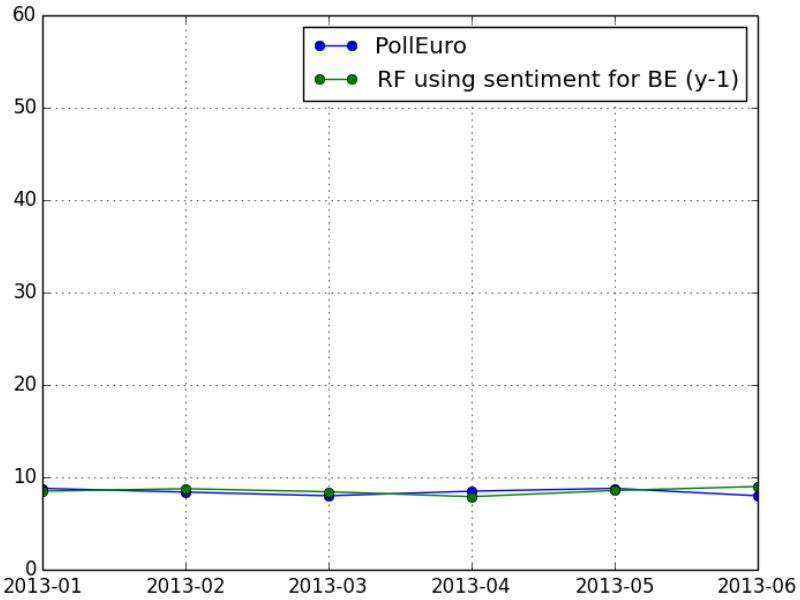

Figure B.60: Predicted and real values of vote intention, from January 2013 to June 2013, for Catarina Martins e João Semedo (BE)

B.3.2 Buzz

B.3.2.1 Ordinary Least Squares

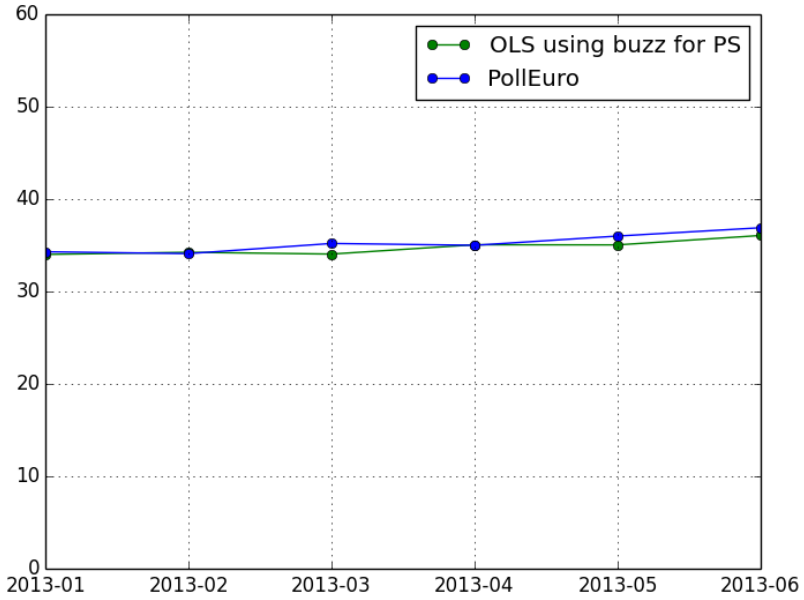

Figure B.61: Predicted and real values of vote intention, from January 2013 to June, for António José Seguro (PS), excluding the $\Delta(y_{t-1})$ feature.

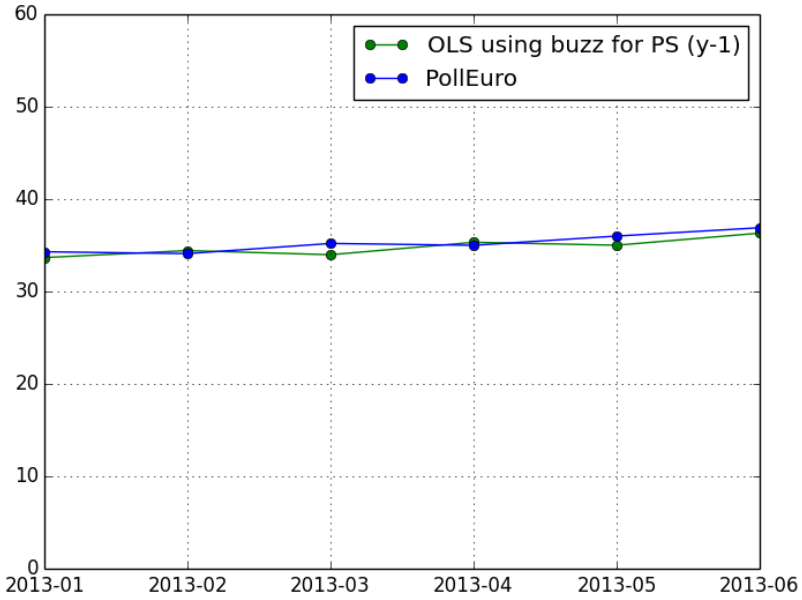

Figure B.62: Predicted and real values of vote intention, from January 2013 to June 2013, for António José Seguro (PS)

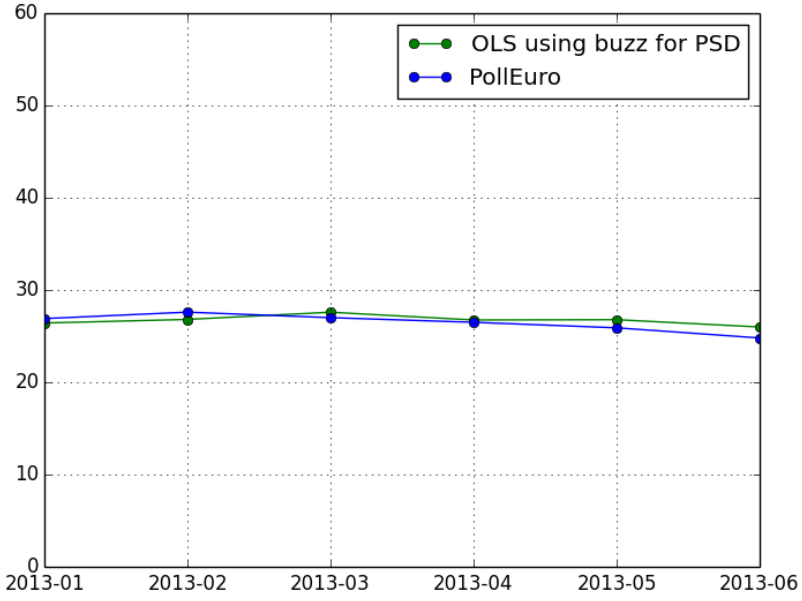

Figure B.63: Predicted and real values of vote intention, from January 2013 to June, for Pedro Passos Coelho (PSD), excluding the $\Delta(y_{t-1})$ feature.

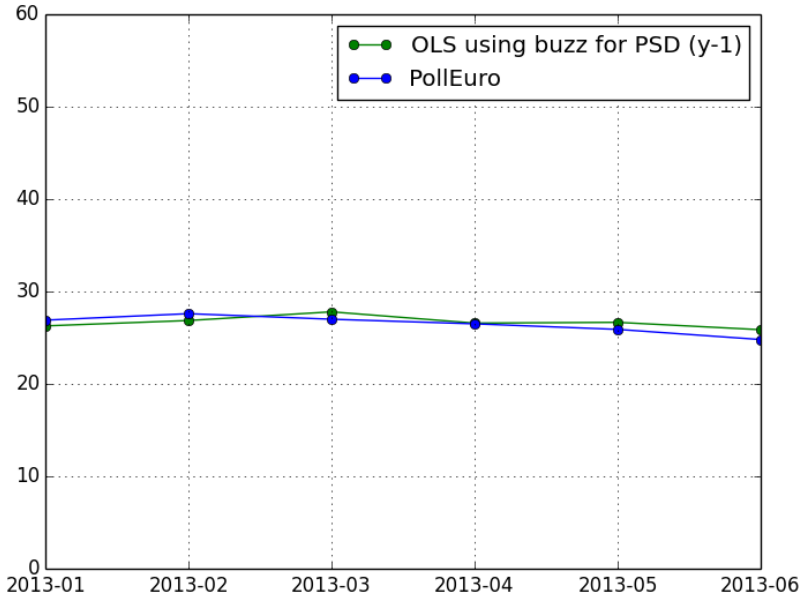

Figure B.64: Predicted and real values of vote intention, from January 2013 to June 2013, for Pedro Passos Coelho (PSD)

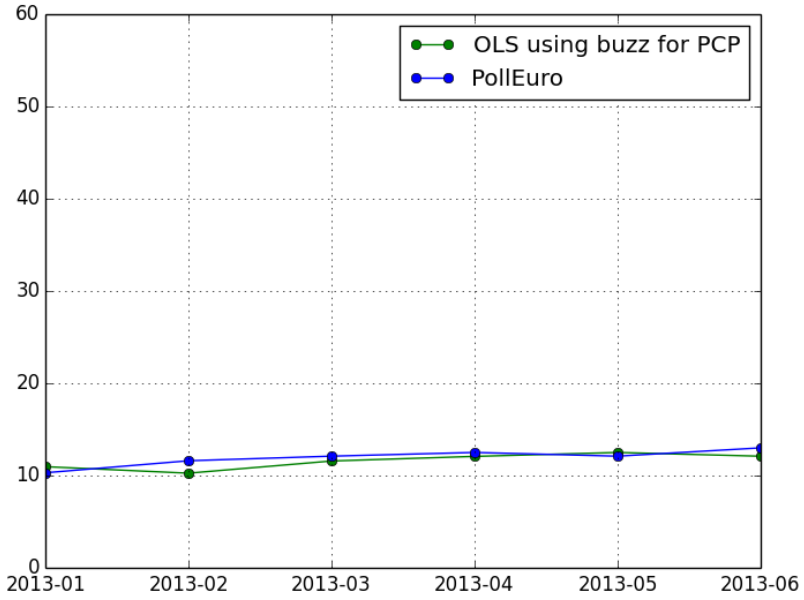

Figure B.65: Predicted and real values of vote intention, from January 2013 to June, for Jerónimo de Sousa (PCP), excluding the $\Delta(y_{t-1})$ feature.

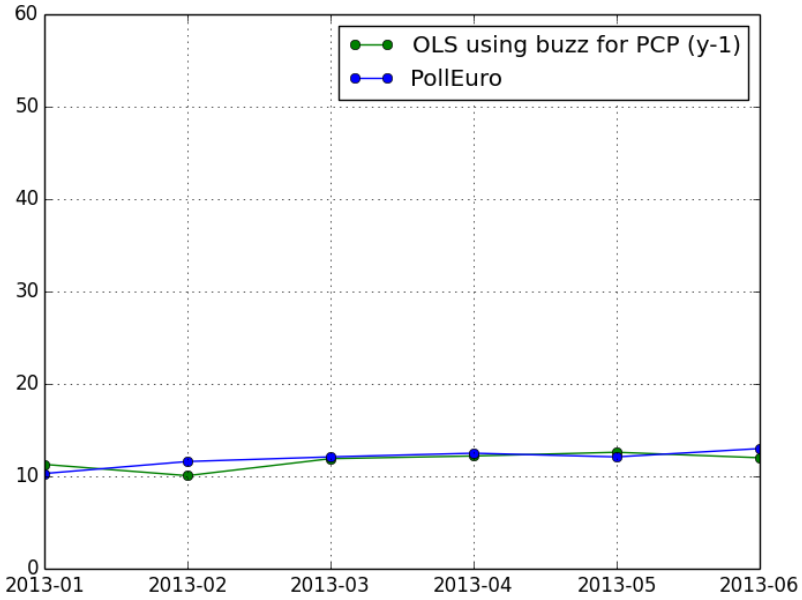

Figure B.66: Predicted and real values of vote intention, from January 2013 to June 2013, for Jerónimo de Sousa (PCP)

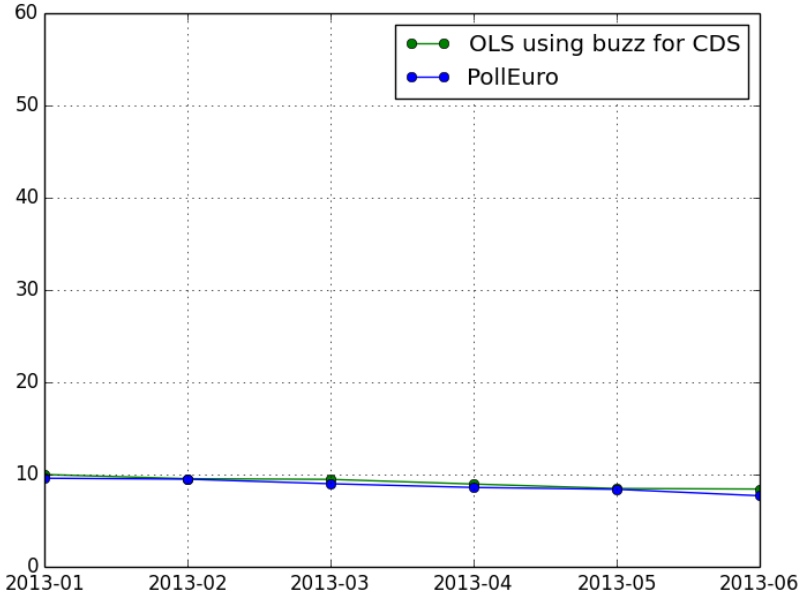

Figure B.67: Predicted and real values of vote intention, from January 2013 to June, for Paulo Portas (PP), excluding the $\Delta(y_{t-1})$ feature.

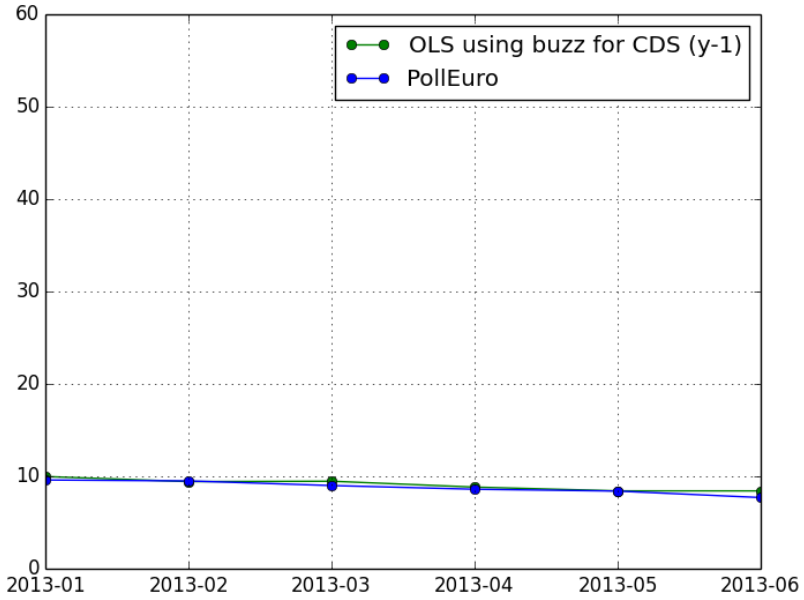

Figure B.68: Predicted and real values of vote int[!ht]ention, from January 2013 to June 2013, for Paulo Portas (PP)

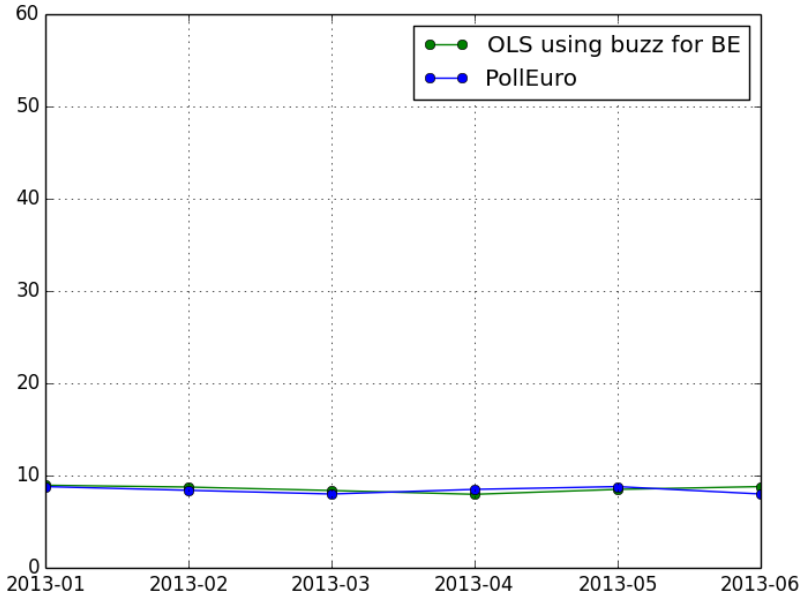

Figure B.69: Predicted and real values of vote intention, from January 2013 to June, for Catarina Martins e João Semedo (BE), excluding the $\Delta(y_{t-1})$ feature.

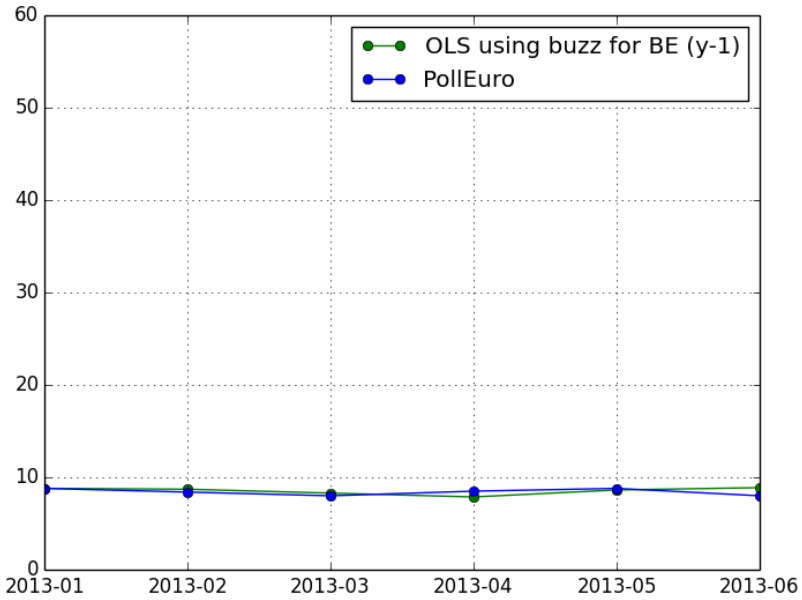

Figure B.70: Predicted and real values of vote intention, from January 2013 to June 2013, for Catarina Martins e João Semedo (BE)

B.3.2.2 Random Forest

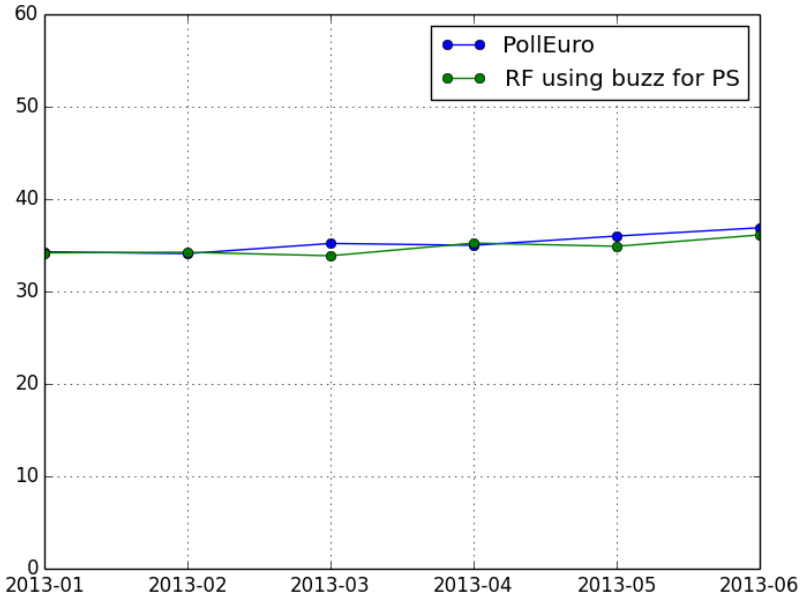

Figure B.71: Predicted and real values of vote intention, from January 2013 to June, for António José Seguro (PS), excluding the $\Delta(y_{t-1})$ feature.

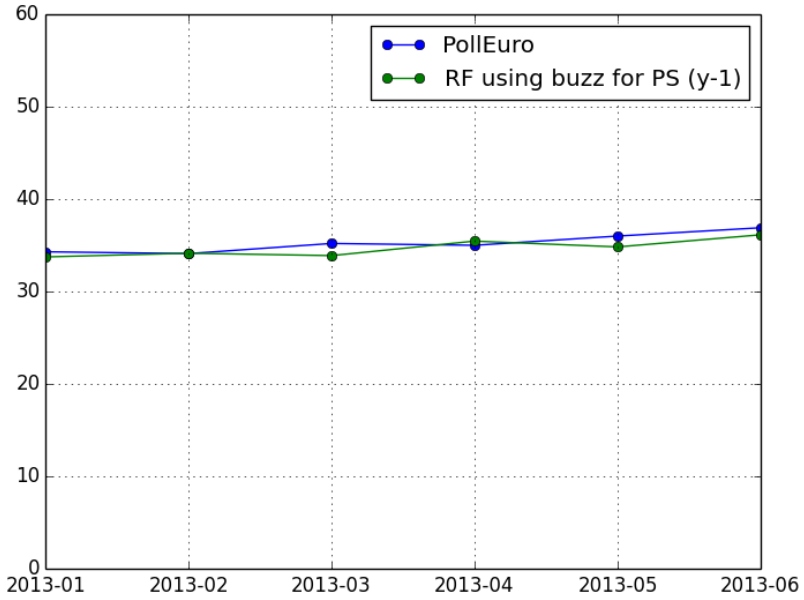

Figure B.72: Predicted and real values of vote intention, from January 2013 to June 2013, for António José Seguro (PS)

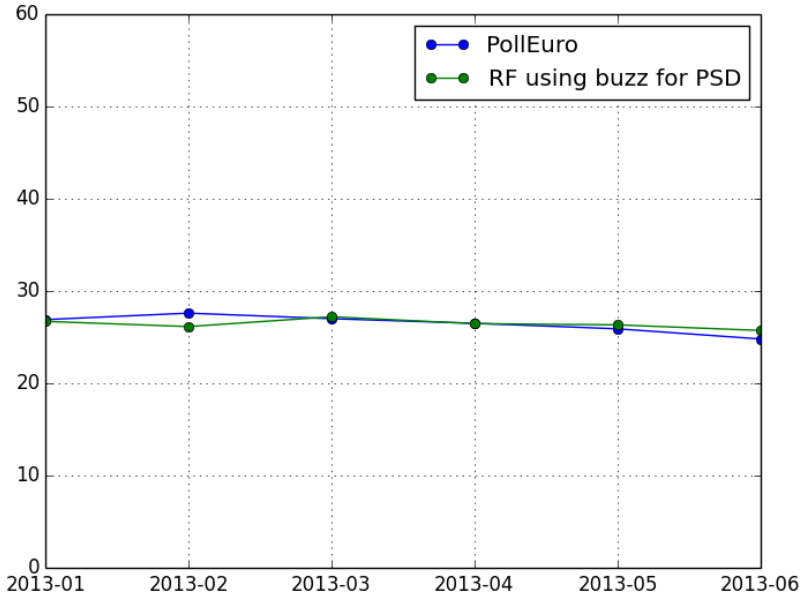

Figure B.73: Predicted and real values of vote intention, from January 2013 to June, for Pedro Passos Coelho (PSD), excluding the $\Delta(y_{t-1})$ feature.

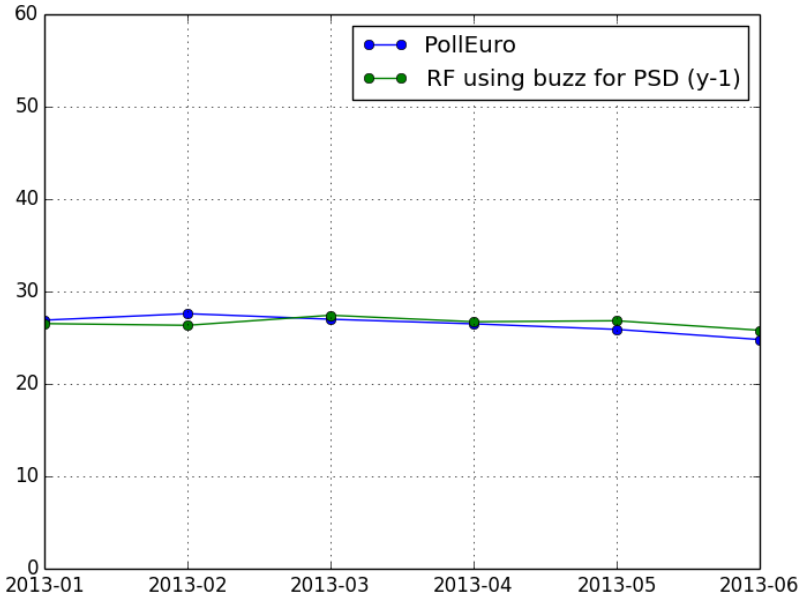

Figure B.74: Predicted and real values of vote intention, from January 2013 to June 2013, for Pedro Passos Coelho (PSD)

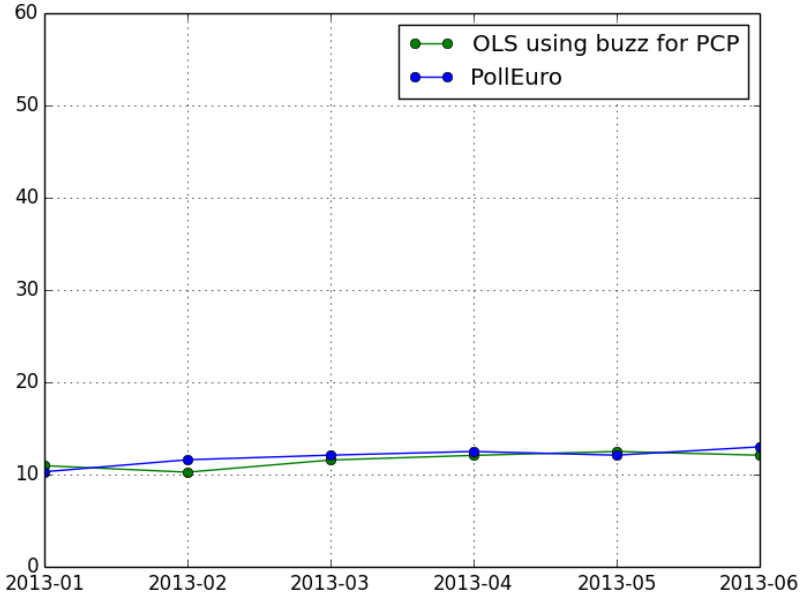

Figure B.75: Predicted and real values of vote intention, from January 2013 to June, for Jerónimo de Sousa (PCP), excluding the $\Delta(y_{t-1})$ feature.

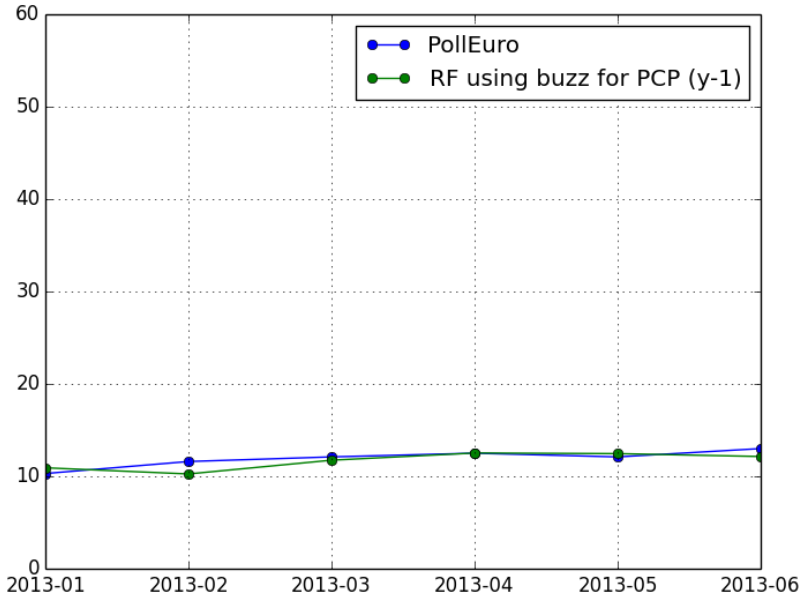

Figure B.76: Predicted and real values of vote intention, from January 2013 to June 2013, for Jerónimo de Sousa (PCP)

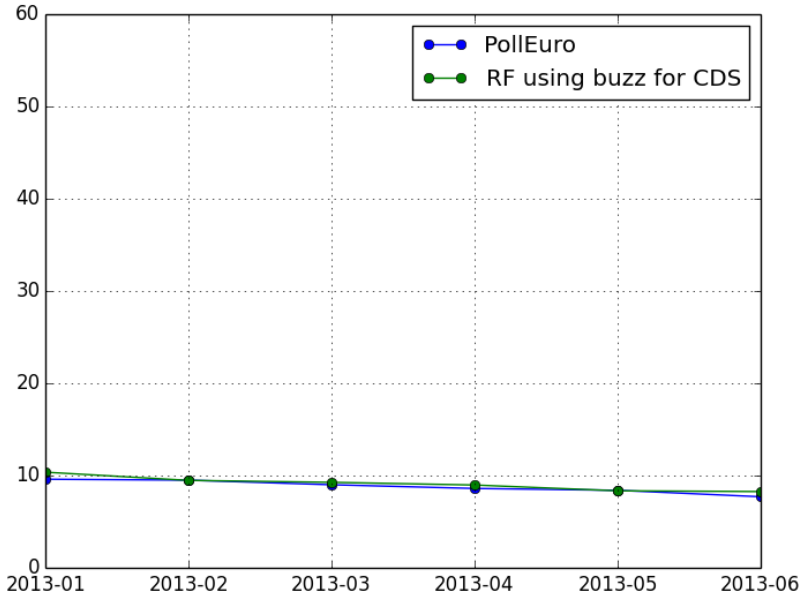

Figure B.77: Predicted and real values of vote intention, from January 2013 to June, for Paulo Portas (PP), excluding the $\Delta(y_{t-1})$ feature.

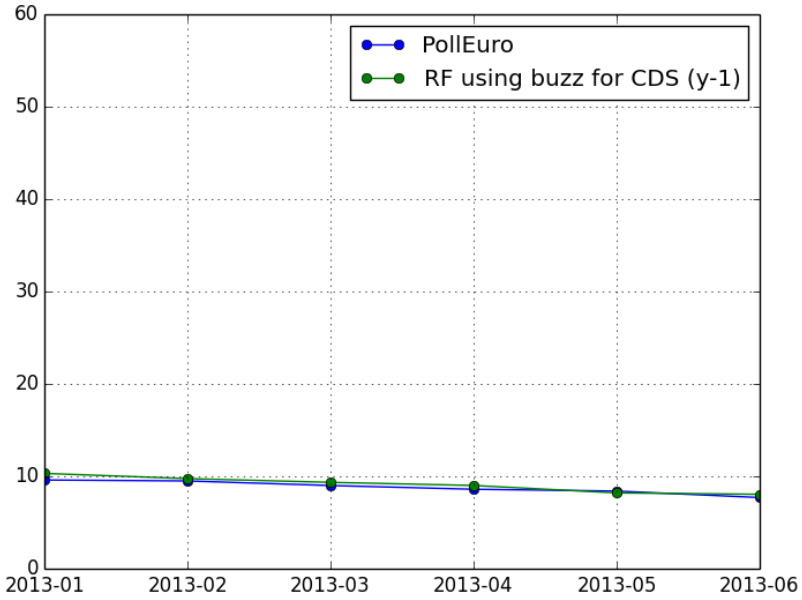

Figure B.78: Predicted and real values of vote int[!ht]ention, from January 2013 to June 2013, for Paulo Portas (PP)

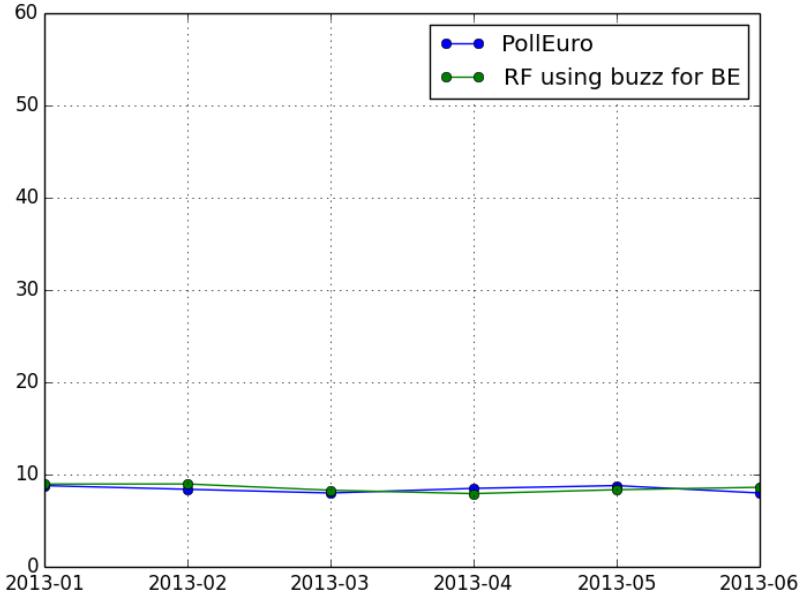

Figure B.79: Predicted and real values of vote intention, from January 2013 to June, for Catarina Martins e João Semedo (BE), excluding the $\Delta(y_{t-1})$ feature.

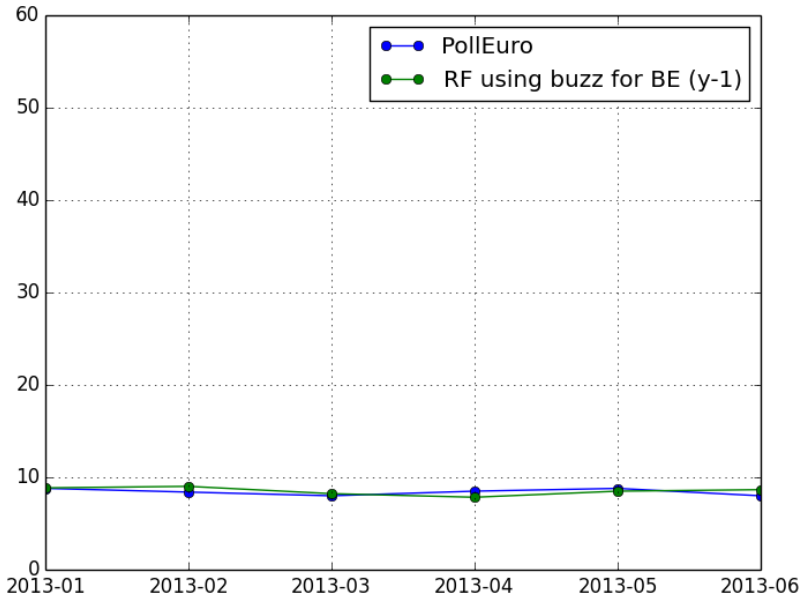

Figure B.80: Predicted and real values of vote intention, from January 2013 to June 2013, for Catarina Martins e João Semedo (BE)

B.4 Experiment Sentiment vs Buzz All

B.4.1 Sentiment

B.4.1.1 Ordinary Least Squares

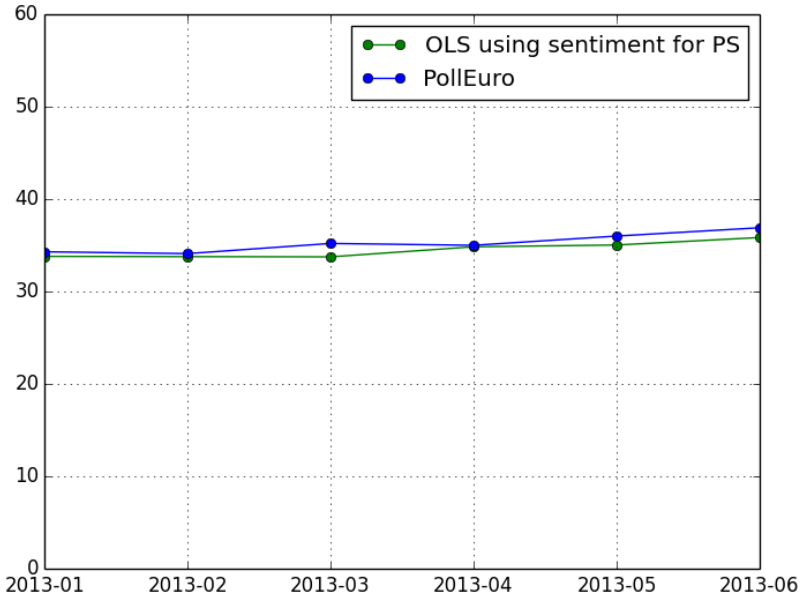

Figure B.81: Predicted and real values of vote intention, from January 2013 to June 2013, for António José Seguro (PS), including the previous poll variation of all political targets

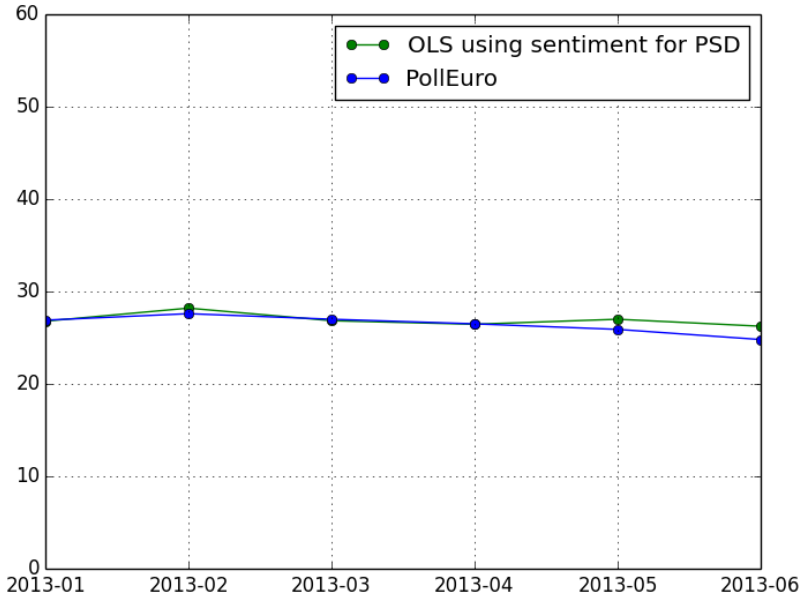

Figure B.82: Predicted and real values of vote intention, from January 2013 to June 2013, for Pedro Passos Coelho (PSD), including the previous poll variation of all political targets

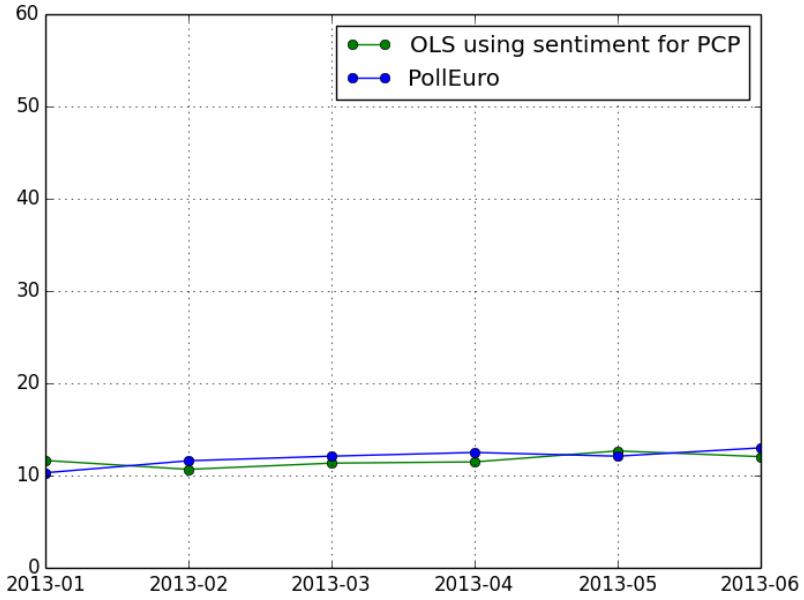

Figure B.83: Predicted and real values of vote intention, from January 2013 to June 2013, for Jerónimo de Sousa (PCP), including the previous poll variation of all political targets

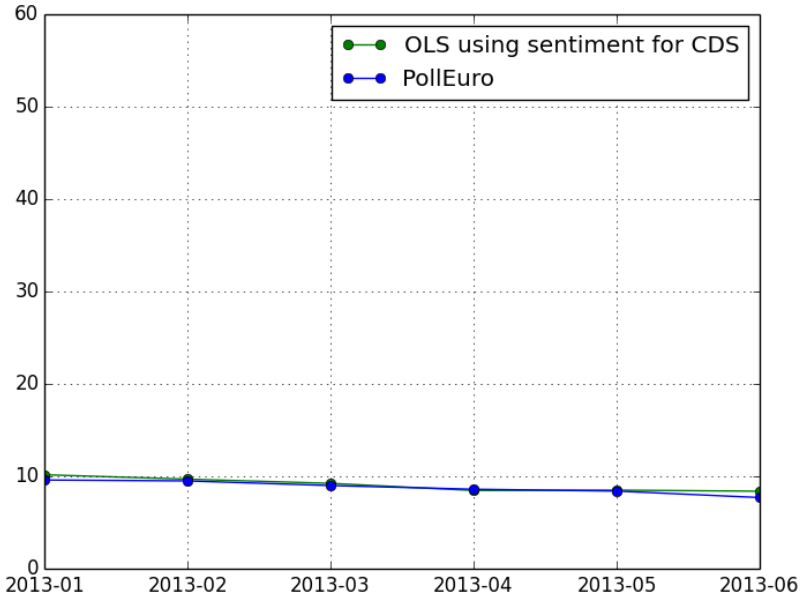

Figure B.84: Predicted and real values of vote intention, from January 2013 to June 2013, for Paulo Portas (PP), including the previous poll variation of all political targets

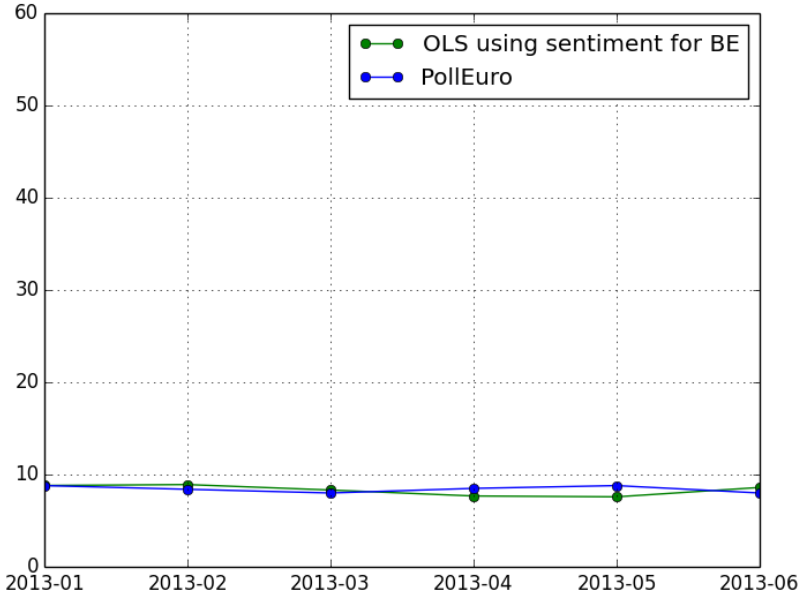

Figure B.85: Predicted and real values of vote intention, from January 2013 to June 2013, for Catarina Martins e João Semedo (BE), including the previous poll variation of all political targets

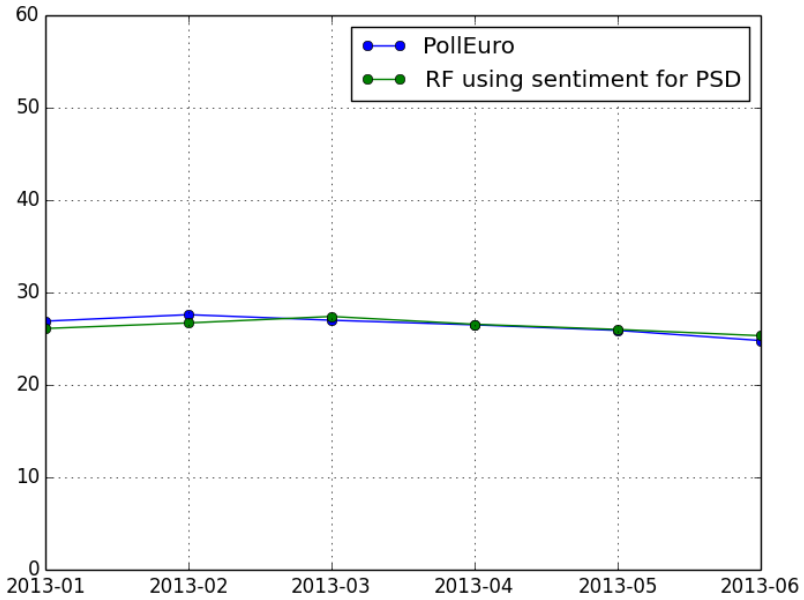

Figure B.87: Predicted and real values of vote intention, from January 2013 to June 2013, for Pedro Passos Coelho (PSD), including the previous poll variation of all political targets

B.4.1.2 Ranfom Forest

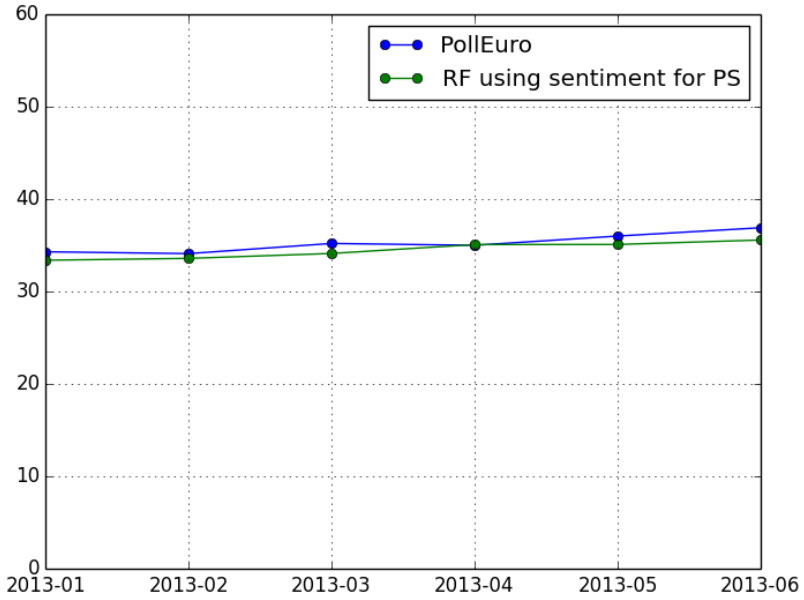

Figure B.86: Predicted and real values of vote intention, from January 2013 to June 2013, for António José Seguro (PS), including the previous poll variation of all political targets

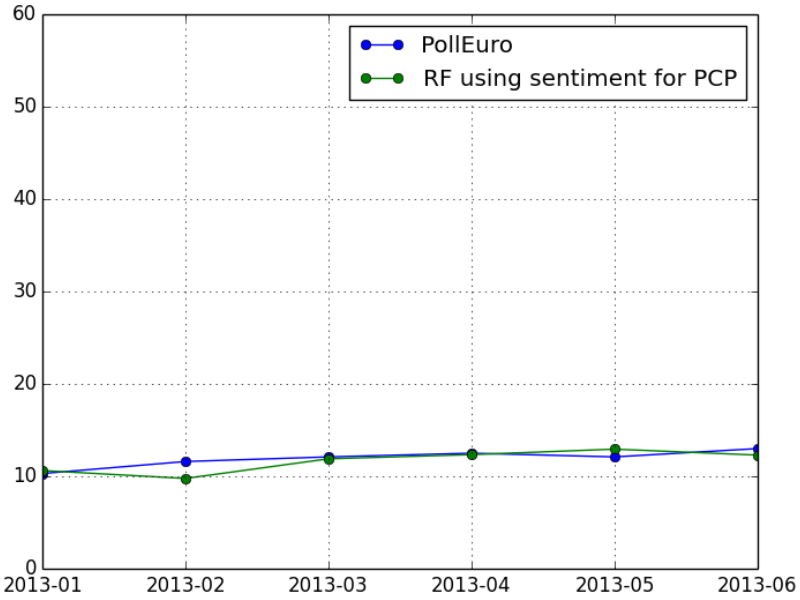

Figure B.88: Predicted and real values of vote intention, from January 2013 to June 2013, for Jerónimo de Sousa (PCP), including the previous poll variation of all political targets

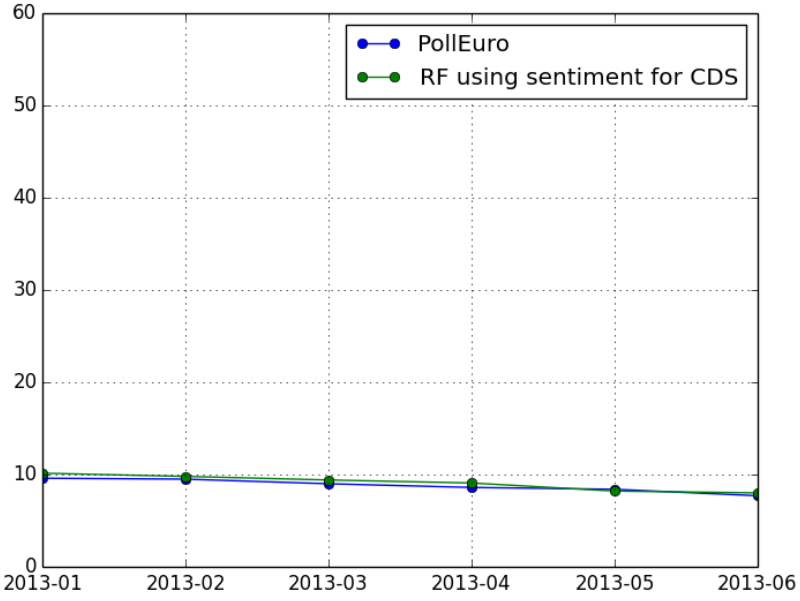

Figure B.89: Predicted and real values of vote intention, from January 2013 to June 2013, for Paulo Portas (PP), including the previous poll variation of all political targets

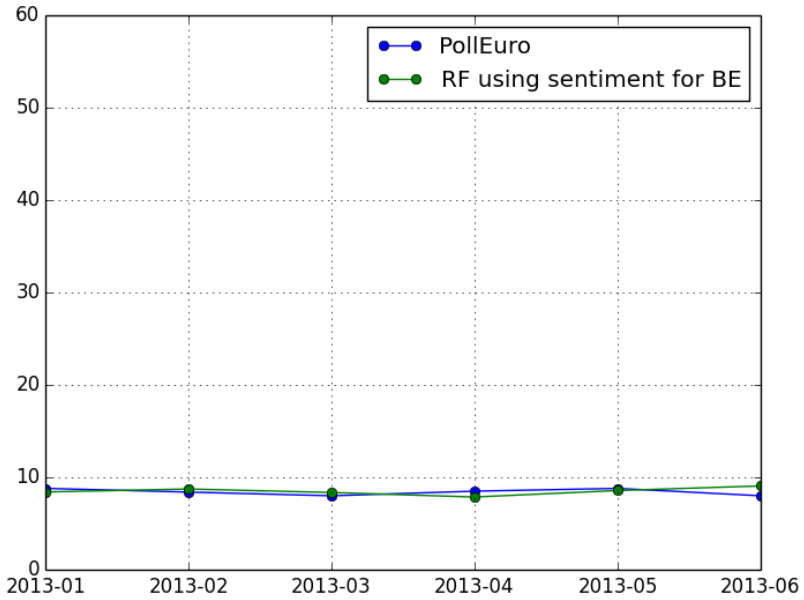

Figure B.90: Predicted and real values of vote intention, from January 2013 to June 2013, for Catarina Martins e João Semedo (BE), including the previous poll variation of all political targets

B.4.2 Buzz

B.4.2.1 Ordinary Least Squares

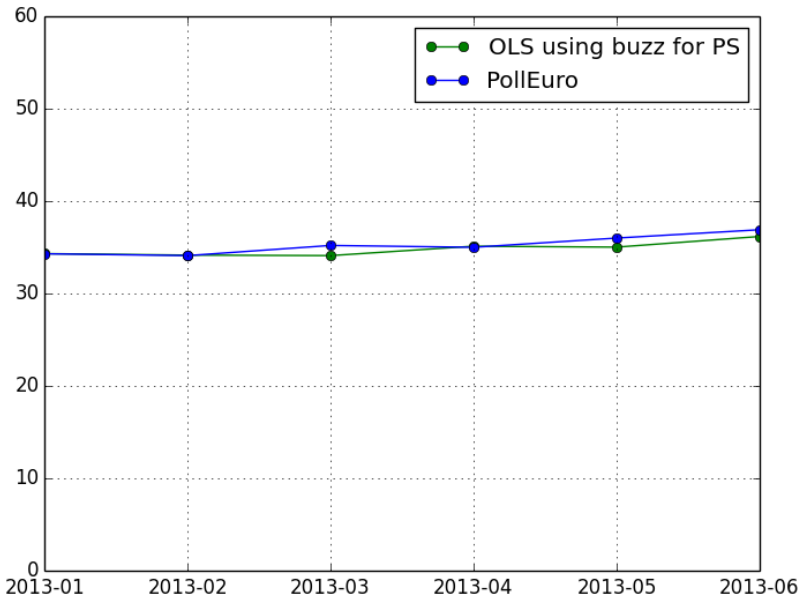

Figure B.91: Predicted and real values of vote intention, from January 2013 to June 2013, for António José Seguro (PS), including the previous poll variation of all political targets

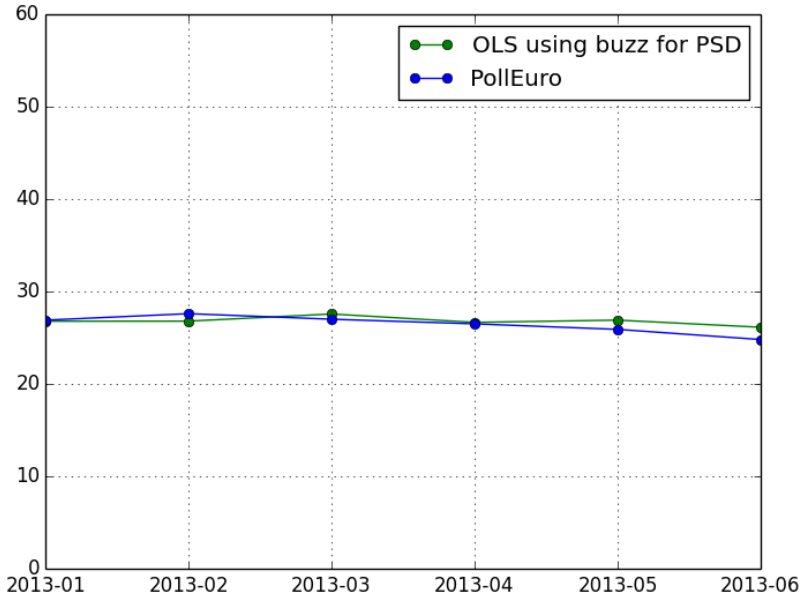

Figure B.92: Predicted and real values of vote intention, from January 2013 to June 2013, for Pedro Passos Coelho (PSD), including the previous poll variation of all political targets

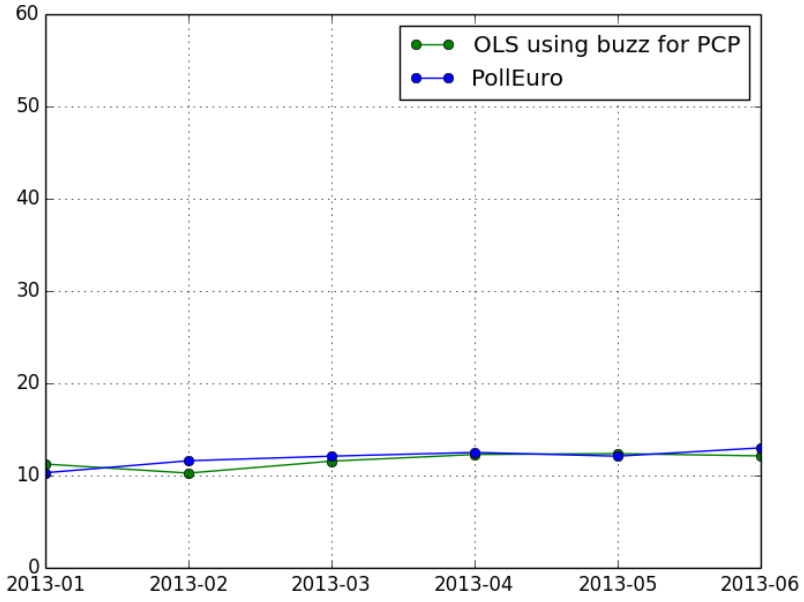

Figure B.93: Predicted and real values of vote intention, from January 2013 to June 2013, for Jerónimo de Sousa (PCP), including the previous poll variation of all political targets

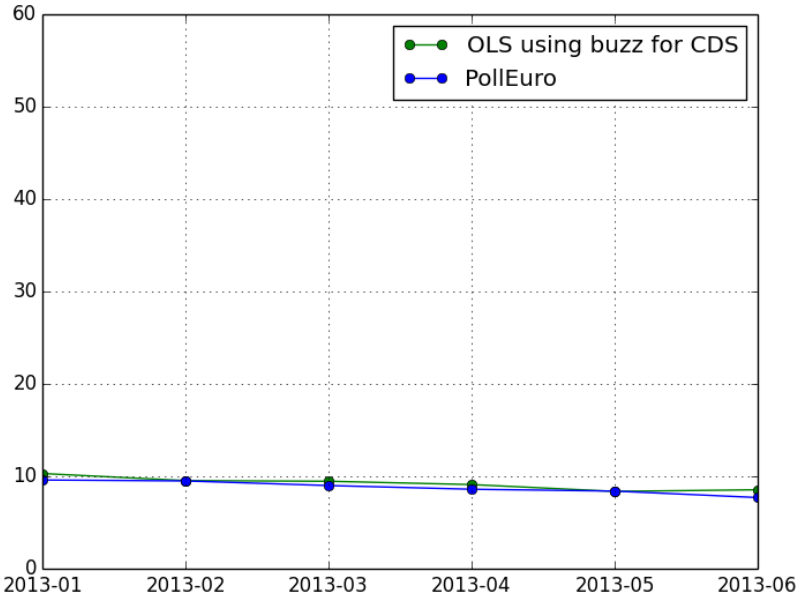

Figure B.94: Predicted and real values of vote intention, from January 2013 to June 2013, for Paulo Portas (PP), including the previous poll variation of all political targets

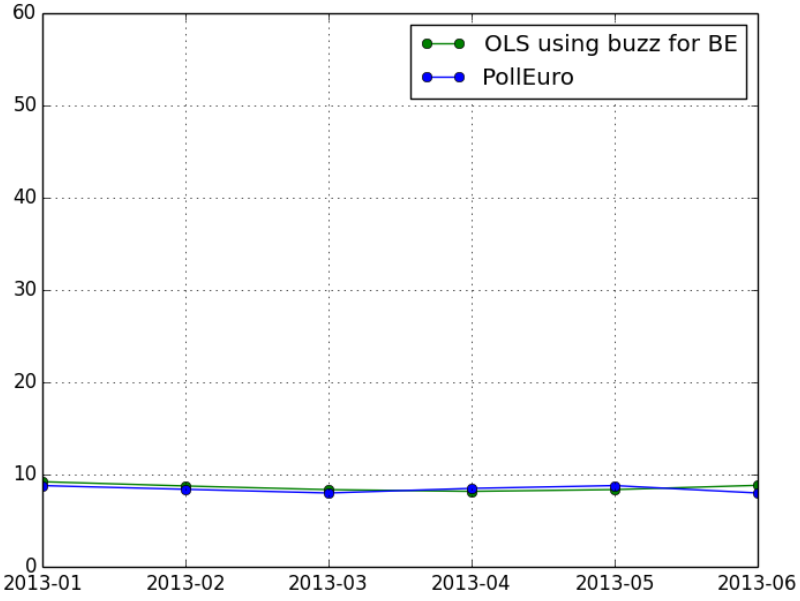

Figure B.95: Predicted and real values of vote intention, from January 2013 to June 2013, for Catarina Martins e João Semedo (BE), including the previous poll variation of all political targets