PoliticAnalytics: Social media analytics for political data science

Luís Gomes
Labs Sapo UP / DEI-FEUP,
University of Porto
Porto, Portugal
Email: lfc.gomes@fe.up.pt

Pedro Saleiro
Labs Sapo UP / DEI-FEUP,
University of Porto
Porto, Portugal
Email: pssc@fe.up.pt

Carlos Soares
INESC TEC / DEI-FEUP,
University of Porto
Porto, Portugal
Email: csoares@fe.up.pt

***Abstract*—In this work we study the problem of predicting the results of political polls based on the combination of several aggregators of buzz and sentiment obtained from Twitter posts, using the Portuguese political scenario as case study. Thus, we built a model to predict the monthly Portuguese polls results, using 2013 as test period. We aimed this problem to be solved as a regression problem. We used tweets from the Portuguese tweetosphere since June 2011. This dataset contains nearly 233 000 tweets, classified according to their polarity (positive, negative or neutral), regarding the five main Portuguese political leaders. Tweets were collected from more than 100 000 users classified as Portuguese. This way, it was possible to process several sentiment and buzz indicators to be used as independent variables in our regression model. Furthermore, we had access to the polls results, since June 2011 to December 2013. We used these polls results, as dependent variable to train and to evaluate our model. We performed some experiments using two regression algorithms (Random Forests and Ordinary Least Squares). Also, we defined a naive baseline to compare our model with. The best model we obtained has an Mean Absolute Error of 0.63 while our reference model (baseline) has an error of 0.91.**

I. Introduction

Nowadays, surveys and polls are widely used to provide information of what people think about parties or political personalities [4]. These methods use the telephone to survey 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 in politics or marketing. However this method is expensive and time consuming [2][4]. Furthermore, over the years it is becoming more difficult to contact people and persuade them to participate in this surveys [8].

On the other hand, the raise of social media, namely Twitter and Facebook, has changed the way people interact with news. This way, people are able to react and comment any news in real time [1]. One challenge that several research works have been trying to solve is to understand how opinions expressed on social media, and their sentiment, can be a leading indicator of public opinion. However, at the same time there might exist simultaneously positive, negative and neutral opinions regarding the same subject. Thus, we need to obtain a value that reflects the general image of each political target in social media, for a given time period. To that end, we use sentiment aggregators. In summary, a sentiment aggregator is a mathematical formula which calculates a global value based on the number of positive, negative, and neutral mentions of each political target, in a given period. We made an exhaustive study and collected and implemented several aggregators from the state of the art [1][2][5].

Thus, the main objective of our work is to study and define a methodology capable of successfully estimating the polls results, based on opinions expressed on social media, represented by sentiment aggregators. Given the monthly periodicity of polls, we needed to monthly aggregate data. This approach allows each aggregator value to represent the monthly value for each political party. We applied this problem to the Portuguese case study, using Portuguese political data. This is a changeling work given that we need to find a relationship between the opinions expressed on Twitter and the real public opinion tendencies. Furthermore, there is no consensus on defining a global methodology to deal with opinions expressed on social media.

In the next section we review related work. In section III we present the methodology we implemented. We describe data in section IV followed by the experimental setup in section V. In Section VI we present and discuss the results we obtained, while section VII is reserved for some important conclusions taken from our study, and for future work.

II. Related Work

Several prior studies examined how social media, namely Twitter, can be used in political scenario. Johnson et al. [4] concluded that more than predicting elections, social media can be used to gauge sentiment about specific events, such as political news or speeches. Defending the same idea, Diakopoulos el al. [9] studied the global sentiment variation based on Twitter messages of an Obama vs McCain political TV debate while it was still happening. Tumasjan et al. [6] used Twitter data to predict the 2009 Federal Election in Germany. They stated that "the mere number of party mentions accurately reflects the election result". Bermingham et al. [1] correctly predicted the 2011 Irish General Elections also using Twitter data. As told before, several approaches use sentiment as a leading indicator of polls. Thus, messages have to be correctly labeled in order to raise the accuracy of our analysis.

Several sentiment analysis methods are used to label messages expressed on social media as positive, negative or neutral [3][7]. However, labeling messages according to their polarity might be meaningless by itself. Thus, one challenge in the political science is to create the right method capable of

aggregating data so the accuracy of our predictions can be raised. Once again, we have two different approaches when aggregating data: using buzz [1][5][6] or sentiment [1][2][5]. Tumasjan et al. [6] used buzz to predict German federal elections in 2009. Bermingham et al. [1] tried to predict the Irish General Elections in 2011. They proposed and included the share of volume for each party in a system of n parties, as predictive measure in the regression model.

Gayo-Avello et al. [5] also tested the share of volume as predictor in the 2010 US Senate special election in Massachusetts. On the other hand, several other studies use sentiment as a polls result indicator. Connor et al. [2] used a sentiment score to study the relationship between the sentiment extracted from Twitter messages and polls result. They defined the sentiment score as the ratio between the positive and negative messages referring an specific political target. They used the sentiment aggregator as predictive feature in the regression model, achieving a correlation of 0.80 between the results and the poll results, capturing the important large-scale trends. Bermingham et al. [1] also included in their regression model sentiment features. They introduced two novel measures of sentiment. For inter-party sentiment, they modified the share of volume formula to represent the share of positive and negative volume. For intra-party sentiment , they used a log ratio between the number of positive and negative mentions of a given party. Moreover, they concluded that the inclusion of sentiment features augmented the effectiveness of their model.

Gayo-Avello et al. [5] introduced a different approach. In a two-party race, all negative messages on party $c2$ are interpreted as positive on party $c1$, and vice-versa. There are several sentiment aggregators that can be combined with different approaches and algorithms. Furthermore, there is no consensus in which prediction algorithm should be used and what aggregators better reflects the image of the political targets. Thus, due to the absence of a general sentiment aggregator with good performance in several case studies, we decided to include all aggregators we collected from the state of the art (buzz and sentiment aggregators) in the regression model. Therefore the learning algorithm is able to adapt to the most informative agrgegators. Furthermore, we also use two regression algorithms.

III. Methodology

Figure 1 represents the global methodology we use in this work.

We collect Twitter messages in order to classify them, and use their polarity to predict polls results. Thus, we use a platform that collects Twitter messages regarding public figures. We focus on the five main Portuguese political leaders. Messages are collected from 100K different users, classified as Portuguese, representing a sample of the Portuguese community on Twitter. We start with a sample of 1000 Portuguese Twitter users and expand it to theirs Portuguese writing *folowees* and followers, until 100K users.

This platform also classifies Twitter messages according to their polarity. The sentiment classification methods use a corpus of 1500 annotated tweets as training set. These 1500 tweets were manually annotated by 3 different users.

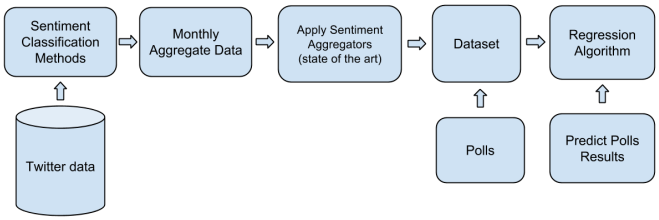

Fig. 1: Solution's general overview

The platform is also capable of returning the daily count of positive, negative and neutral messages regarding each Portuguese political target. Also, we have access to the monthly polls results, that would act as target variable in the regression model. However, if we use daily values, it won't be possible to determine the predictive effectiveness given that we have no daily poll results to compare our predictions with. Thus, we monthly aggregate data so we can calculate the prediction error of our model.

We apply the sentiment aggregators, collected from the state of the art, to the aggregated data. This way, each aggregator represents a monthly value for each political target.

We join sentiment aggregator values and polls results in the same data set. Furthermore, we include in the same data set the value of the polls result of the previous month, for each candidate, which we called y_{t-1} feature. However, there is a small variation in the polls results between two consecutive months. Thus, we also decided to predict that variation. In this approach, instead of using the absolute values of sentiment aggregators and poll results, we use the monthly variations relatively to the previous month. I.e., a given sentiment aggregator i, for a month m, would take the value $Aggreg(i)_m - Aggreg(i)_{m-1}$. Having these two approaches, it allowed us to perform two kind of experiments: (1) using absolute values - to predict poll results absolute value - and (2) using monthly variation - to predict the poll variation with respect to the previous month.

We use all the aggregators along with y_{t-1} feature (or Δy_{t-1} when dealing with monthly variations) as features in our regression model.

IV. Data

A. Twitter Corpus

We apply our work to the Portuguese case study, using Portuguese political data. We use opinions expressed on Twitter. After collect and classify Twitter messages, we built a data set containing the daily count of positive, negative, and neutral mentions referring each leader of the five main Portuguese parties (PS, PSD, CDS, PCP and BE), from August 2011 to December 2013. The data set contains 232 979 classified messages, collected from a network of 100K different users classified as Portuguese.

Table I represents the distribution of positive, negative, and neutral mentions per party, of the data we collected.

TABLE I: Distribution of positive, negative and neutral mentions per political party

	Negative	Positive	Neutral	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

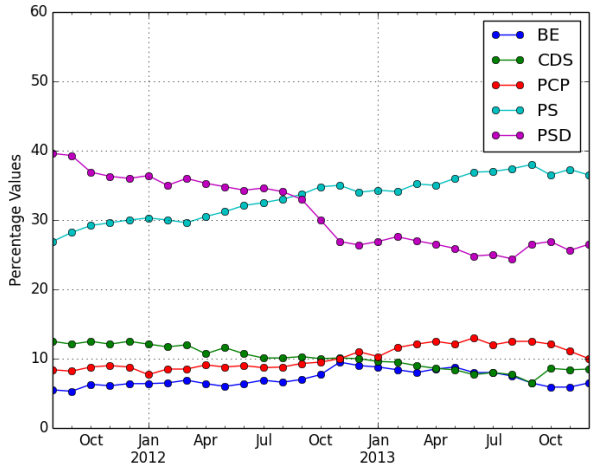

Fig. 2: Representation of the monthly poll results of each political candidate

- 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.
- The positive mentions represent less than 1% of the total mentions of each party, except for BE where they represent 2% of the total mentions.
- The most mentioned parties 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.

B. Public Opinion Polls

The polls results were provided by Instuto de Ciências Sociais da Univerdidade de Lisboa. These surveys were made by Eurosondagem, a Portuguese private company which collects public opinion. This data set contains the monthly polls results of the five main Portuguese parties, from June 2011 to December 2013. In figure 2 is represented the evolution of Portuguese polls results. We can see two main party groups: 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 in general an uptrend. The second group is composed by CDS, PCP and BE. This group has a vote intention range from 5% to 15%. While CDS has a downtrend in public opinion, PCP has an ascendent one. Although the constant tendencies (up- and downtrends), we analyzed data noticed that the maximum variation observed between two consecutive months is 3%. Finding the small variations was an important milestone given that it influenced our approaches to this problem.

TABLE II: List of implemented features

Name	Formula
Sentiment Aggregators	
positive_mentions	$related\_pos$
negative_mentions	$related\_neg$
neutral_mentions	$related\_neu$
total_mentions	$candidate\_buzz$
bermingham [1]	$\log_{10} \frac{related\_pos+1}{related\_neg+1}$
berminghamsovn [1]	$\frac{related\_neg}{total\_neg}$
berminghamsovp [1]	$\frac{related\_pos}{total\_pos}$
connor [2]	$\frac{related\_pos}{related\_neg}$
gayo [5]	$\frac{related\_pos+others\_neg}{total\_pos+total\_neg}$
polarity	$related\_pos - related\_neg$
polarityONeutral	$\frac{related\_pos-related\_neg}{related\_neu}$
polarityOTotal	$\frac{related\_pos-related\_neg}{candidate\_buzz}$
subjOTotal	$\frac{related\_pos+related\_neg}{candidate\_buzz}$
subjNeu	$\frac{related\_pos+related\_neg}{related\_neu}$
subjSoV	$\frac{related\_pos+related\_neg}{total\_pos+total\_neg}$
subjVol	$related\_pos + related\_neg$
share [1]	$\frac{candidate\_buzz}{total}$
shareOfNegDistribution[1]	$\frac{\frac{related\_neg}{candidate\_buzz}}{\sum_{i=0}^{n} \frac{related\_neg_i}{candidate\_buzz_i}}$
pollEuro	y_t
pollEuro-1	y_{t-1}
Normalized Sentiment Aggregators (divided by the candidate buzz)	
normalized_positive	$normalized\_pos$
normalized_negative	$normalized\_neg$
normalized_neutral	$normalized\_neu$
normalized_bermingham	$\log_{10} \frac{normalized\_pos+1}{normalized\_neg+1}$
normalized_connor	$\frac{normalized\_pos}{normalized\_neg}$
normalized_gayo	$\frac{normalized\_pos+normalized\_others\_neg}{normalized\_total\_pos+normalized\_total\_neg}$
normalized_polarity	$normalized\_pos - normalized\_neg$

TABLE III: Explanatory Table

Formula	Description
$total\_pos$	total of positives mentions of all candidates
$total\_neg$	total of negatives mentions of all candidates
$total$	total of mentions of all candidates
$others\_neg$	total of negative mentions, except of the candidate we are calculating)

V. Experimental Setup

As explained in section IV, we had access to the daily count of twitter messages regarding each one of the five main Portuguese political leaders, and theirs respective public opinion polls results. However, given the monthly periodicity of

[1] n is the number of candidates

polls, we need to monthly aggregate the twitter count. The next step is to apply the sentiment aggregators formulas using the monthly count of messages. Thus, each aggregator represents a monthly value for each political target. Also, we transform the dataset in order to allow each sentiment aggregator to carry the variation relatively to the previous month, given that we also intended to predict the polls variations. The last step is to perform the predictions. In our case study we try to predict the monthly polls results of 2013 (both absolute result and monthly variation). To estimate the performance of our regression models, we use a sliding window technique. Thus, we separate our data:

- Training set – containing the monthly values of the sentiment aggregators for 16 months prior the month intended to be predicted.
- Testing set - containing the sentiment aggregators values of the month intended to be predicted.

We use a fixed temporal window of 16 months prior the month we want to predict as training set, and a test period from January 2013 to December 2013. The prediction process is iterative. In the first iteration we predict the poll results for January 2013. (1) Thus, we select the values of the sentiment aggregators for 16 months prior January 2013 (September 2011 to December 2012). (2) We use that data to train our regression model. (3) Then we input data of January 2013 in the the trained model, to obtain a prediction of the poll results. (4) We select the next month of the testing set and repeat the process until all months are predicted. The full list of implemented features can be consulted in table II. Some features are described in table III.

Furthermore, we duplicate each experiment including and excluding the polls result or the polls variation of the previous month as independent variable (y_{t-1} or Δy_{t-1}, respectively) in the regression model.

In the prediction process we use two regression algorithms: a linear regression algorithm (Ordinary Least Squares - OLS) and a non-linear regression algorithm (Random Forests - RF).

After we predict the poll results of 2013, we use Mean Absolute Error (MAE) [3] as evaluation measure, to determine the absolute error of each prediction. Then, we calculate the average of the six MAE's so we could know the global prediction error.

$$MAE = \frac{\sum_{i=1}^{n} |f_i - y_i|}{n} \quad (1)$$

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

We use a naive baseline commonly used in prediction problems as reference model to compare our model with: predict that the polls result of a given month m is equal to the month $m-1$. Thus, there is no variation between two consecutive months. As expected, this baseline has different prediction errors when dealing with absolute or variations values.

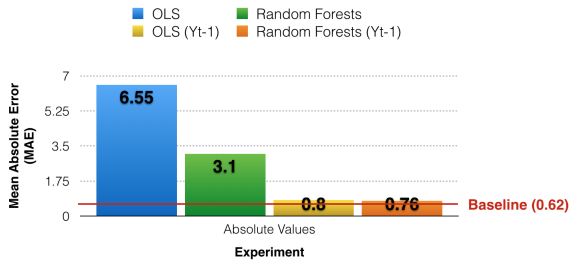

Fig. 3: Model and baseline's errors

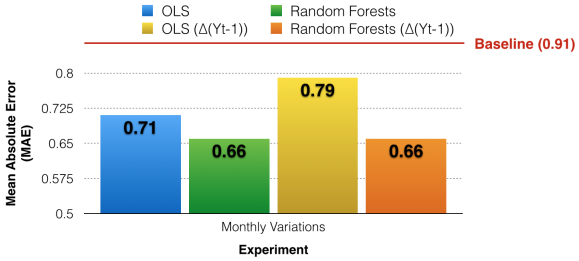

Fig. 4: Model and baseline's errors

VI. Results and Discussion

In this section we explain in detail the experiments we performed and the results we obtained. We performed two different experiments: (1) using absolute values and (2) using monthly variations.

A. Experiment using absolute values

In this experiment, the sentiment aggregators take absolute values in order to predict the absolute values of polls results. Using mathematical notation, this experiment can be seen as: $y \leftarrow \{y_{t-1}, buzzAggregators, sentimentAggregators\}$. In figure 3 we can see the global errors we obtained.

This experiment shows that the inclusion of the polls result of the previous month (y_{t-1} has a determinant role in our regression model. This can be easily explained given that we are including the baseline in the model.

B. Experiment using monthly variations

According to our exploratory data analysis, the polls results have a small variation between two consecutive months. Thus, instead of predicting the absolute value of poll results, we tried to predict that variation. In mathematical notation: $\Delta y \leftarrow \{\Delta(y_{t-1}), \Delta buzzAggregators, \Delta sentimentAggregators\}$

In this particular experiment, the inclusion of the Δy_{t-1} as feature in the regression model has not a determinant role (figure 4). Including that feature we could not obtain lower MAE than excluding it. It means that the real monthly poll variation is not constant over the year. In general, using a non-linear regression algorithm we obtained lower MAE. Also, we achieve better results than the baseline.

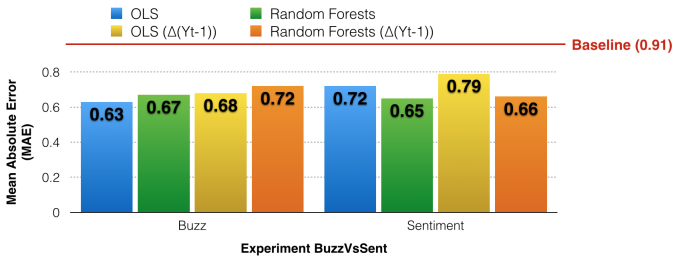

Fig. 5: Model and baseline's errors

1) Buzz and Sentiment: Several studies state the buzz has predictive power and reflects the public opinion on social media. Following this premise, we trained our models with buzz and sentiment aggregators separately to predict polls variations:

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

This experiment allowed us to compare the behavior of buzz and sentiment aggregators.

According to figure 5, buzz and sentiment aggregators have similar results. Although the OLS algorithm combined only with buzz aggregators has a slightly lower error than the other models, it is not a significant improvement. These results also show that Random Forests algorithm performs the best when combined only with sentiment aggregators.

2) Feature Selection: Part of our work is to understand which aggregator (or group of aggregators) better suits our case study. According to the previous experiments, we can achieve lower prediction errors when training our model with buzz and sentiment aggregators separately. However, when training our model with these two kinds of aggregators separately, we are implicitly performing feature selection. As can be confirmed on table II, we only have two buzz features ($share$ and $total\_mentions$). Thus, we decided to apply a feature selection technique to the sentiment aggregators, in order to select the most informative ones to predict the monthly polls results variation. We use univariate feature selection, selecting 10% of the sentiment features (total of 3 features). Using this technique, the Random Forests' global error raise from 0.65 to 0.73. However, OLS presents an MAE drop from 0.72 to 0.67. Another important fact to notice is that if we perform univariate feature selection to all aggregators (buzz and sentiment), we will achieve the same MAE value that when applied only to sentiment aggregators. It means that buzz aggregators are not chosen by the feature selection technique.

We try a different approach and perform a recursive feature elimination technique. In this technique, features are being eliminated recursively according to a initial score given by the external estimator. This method allow us to determine the number of features to select. Thus, also selecting 3 features, the OLS' MAE drop to 0.63. Once again, none of the buzz features were selected. Furthermore, none of the feature selection techniques select the same features for every monthly prediction.

VII. Conclusions and Future Work

The naive baseline proved to be stronger than initially expected. It happens due to the small poll variation between two consecutive months. The lack of polls variation made the baseline a hard model to beat. However, we can drop the baseline's MAE in almost 30%. In general, the models that use absolute values have a bigger absolute error than the ones that use variations as predictive measure. In these cases, the inclusion of the y_{t-1} feature has an important role given that it contributes for a reduction of the absolute error. Also generally speaking, Random Forests has slightly lower errors than OLS. However, when we train our model with buzz and sentiment aggregators separately, Random Forests and OLS have similar results. For sentiment aggregators Random Forests seems to perform the best. On the other hand, for buzz aggregators, Random Forests and OLS are both suitable.

In our study, we build a model where we achieve the lowest MAE using the linear algorithm (OLS), combined only with buzz aggregators variations, using monthly variations. The model has an MAE of 0.63%. We performed two feature selection techniques: (1) Univariate feature selection and (2) recursive feature elimination. Applying the recursive technique to the group of sentiment features, we can achieve and MAE of 0.63, equating our best model. Furthermore, the chosen features were not the same in every prediction. The results show that we can estimate the polls results with low prediction error, using sentiment and buzz aggregators based on the opinions expressed on social media. However, it is important to notice that all improvements are not significant.

The next immediate step is to implement a methodology using time series analysis. Furthermore, it is desirable to test this methodology with difference data sources, such as Facebook messages, blogs or news.

References

[1] Bermingham, A. and Smeaton, A. (2011) On Using Twitter to Monitor Political Sentiment and Predict Election Results. *Workshop at the International Joint Conference for Natural Language Processing (IJCNLP)*, 13th November 2011, Chiang Mai, Thailand.

[2] Connor, B., Balasubramanyan, R., Routledge, B., Smith, N. (2010) From Tweets to Polls : Linking Text Sentiment to Public Opinion Time Series. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 2010

[3] Han, J., Kamber, M. (2006) Data Mining: Concepts and Techniques *Morgan Kaufmann Publishers*, 2006

[4] Johnson, C., Shukla, P., Shukla, S. On Classifying the Political Sentiment of Tweets *cs.utexas.edu.*

[5] Metaxas, P., Mustafaraj, E., Gayo-Avello, D. (2011) 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*, 165—171, October 2011.

[6] Tumasjan, A., Sprenger,TP., Sandner, P., Welpe, I. (2010) Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape *2010 Social Science Computer Review*, 29(4):402-418

[7] Witten, I., Frank, E., Hall, M. (2011) Data Mining. Practical Machine Learning Tools and Techniques, 2011

[8] Kohut, A., Keeter, S., Doherty, C., Dimock, M., Directors, A., Christian, L., (2012) Assessing the Representativeness of Public Opinion Surveys Assessing the Representativeness of Public Opinion Surveys. 2012.

[9] Diakopoulos, N., Shamma, D. (2010) Characterizing debate performance via aggregated twitter sentiment. *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*, 1195-1198