The PredNews forecasting model

Ana Cristina Bicharra Garcia
Universidade Federal do Estado do Rio de Janeiro
Rio de Janeiro, Brazil
cristina.bicharra@uniriotec.br

William Silva
Universidade Federal Fluminense
Rio de Janeiro, Brazil
william.wws@gmail.com

Luís Correia
BioISI, Universidade de Lisboa
Lisboa, Portugal
Luis.Correia@ciencias.ulisboa.pt

ABSTRACT

The objective is to verify the hypothesis that it is possible to predict election outcomes with similar precision to that of traditional polls. We did sentiment analysis of online news comments about candidates and of their likes and dislikes to predict the two first candidates. Comments from the main Brazilian online newspapers were collected approximately 4 months before the 2016 municipal elections, in 14 municipalities with a total of 70 candidates. Results show that a linear regression model has a high accuracy and robustness, adding to a clear explanatory capability.

CCS CONCEPTS

• Networks → Online social networks; **• Computing methodologies → Supervised learning by regression**; **• Applied computing → Voting / election technologies**;

KEYWORDS

Election forecast, online newspaper, sentiment analysis, social network, collective intelligence

ACM Reference Format:
Ana Cristina Bicharra Garcia, William Silva, and Luís Correia. 2018. The PredNews forecasting model. In *dg.o '18: dg.o 2018: Proceedings of the 19th Annual International Conference on Digital Government Research, May 30-June 1, 2018, Delft, Netherlands.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3209281.3209295

1 INTRODUCTION

Although historically society has relied on predictions using traditional poll methods, these have often been fragile and incorrect, as in for example the USA presidential election of the year 2016, where the victory of candidate Hillary Clinton was taken for granted by the great majority of pollsters using traditional methods. Contrary to those predictions, candidate Donald Trump was elected president, surprising the world. In addition, traditional methods are costly and subject to interlocutor bias.

It is irresistible for researchers to use people's volunteer opinions as basis for creating predictions. Most research with this approach focuses on analysis of tweets [3, 17, 18]. Although Twitter is one of the main sources of analysis of opinion mining, the efficiency of its predictions is questionable mainly due to reproducibility, representativeness and bias of samples since they are generated from keywords chosen by whoever is analyzing [8].

This scenario leads us to seek new sources of information for use in predictive methodology. The alternative adopted strategy was to analyze comments from online newspapers, which are typically the size of tweets. Analyzing these sources of information allows “going back in time" to recover all comments made on political news, avoiding the anticipated definition of keywords which is highly subject to researcher bias. Such sources of information may in the future be explored to obtain a sense of the population's reaction to particular situations where governments may intervene.

The goal of this work is to investigate the hypothesis that is possible to predict election results from media comments and their binary appraisals in likes and dislikes with at least similar precision to that of traditional methods. In this first approach we restricted predictions to the first two candidates. For the construction of the proposed model and its test, we selected as case study the Brazilian municipal elections of 2016. We obtained comments from the main Brazilian online newspapers approximately 4 months before those elections, in 14 municipalities involving a total of 70 candidates. We ran a regression algorithm to define the model for clarity. To check the quality of our model we compared our results to the the estimates generated by IBOPE, the Brazilian major private election polling company, to the simple candidate mention counting and also to a non-parametric technique, artificial neural net (ANN).

In next section we list the related work concerning elections' forecast models, including the ones that use Twitter as the data information source. In section 3, the PredNews approach is presented explaining the data collection process using a web crawler to gather the comments from the online newspapers, the parameter selection and the multivariate regression analysis. In section 4, we present the results of applying PredNews to four different Brazilian city mayor elections. Section 5 summarizes the contributions of our studies and indicates our next steps.

2 RELATED WORK

The original *slogan* from *Twitter - What are you doing ?* - did a great job for it encouraged users to share their thoughts and activities with their friends and followers. Today *Twitter* is considered one of the most important tools for disseminating opinions. On the other hand, it is a fact that the great majority of Internet users, without mentioning people in general, are not using *Twitter* [7].

In their study of election forecasting using social network, Tumasjan et al. [17, 18] collected *tweets* for approximately 5 weeks, using the candidates' names as keyword and verified that the simple mention of these candidates' names can be reflected directly in the

dg.o '18, May 30-June 1, 2018, Delft, Netherlands
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6526-0/18/05…$15.00
https://doi.org/10.1145/3209281.3209295

election result through favorable voting. These researchers claimed the described method was not only possible but also quite simple.

Miranda et al. [12] created a forecast model for the municipal election of 6 Brazilian municipalities. It was considered that if a user commented on a candidate from municipality X, he/she would vote in municipality X. The user is characterized according to demographic attributes such as gender, age, and social class. *Tweets* with editorial content were filtered and sentiment analysis was used to classify them. The percentage of each candidate was estimated taking each positive comment as one vote, and each negative comment distributed in percentages among the other candidates. It predicted 50% of the candidates going through to the second round, without mentioning the order between first and second places, and 66% of the winners of the second round.

Hagar [9] used a linear regression to verify the relation among the independent variables "Twitter Use", "Followees", "Followers", "Candidate Tweets", "Candidate Replies", "Favourites", "Retweets", "Mentions", "Voter Replies", "Sex", "Incumbency", "Age", to examine the relation between votes in candidates called *Incumbents* and *Challengers* in municipal elections in Canada. It was verified that the more the candidate was mentioned the more his/her popularity also grew, despite the relative small sample which corresponded to 1 month before elections. His proposal identified the relationship of each one of these variables with the votes, but did not detail the candidates and the forecast model applying the indicated variables.

Gayo-Avello [7] filtered *tweets* according to geolocation, for USA elections from people from all over the world. Those who share opinions are not necessarily voters, and considering them would distort the forecast considerably. In this work, only users from the country in question are included in the forecast model and the others are ignored. A study that overestimated Obama's victory in 2008 was given as an example of the problem. Gayo-Avello [8] also disputes the predictions based on *Twitter*, arguing that researchers tend not to publish bad results, which can generate a false sense of success. Based on Gayo-Avello's premises, Almeida [2] developed a new methodology to define stratified samples, using demographic attributes for prediction of elections and additionally used sentiment analysis in *tweets* to account for votes, creating samples according to demographic data inferred to forecast the outcome of the election. Despite demonstrating success in his predictions, he reported that to create accurate samples, he would need a fairly large number of *tweets* to match the actual distribution.

Bovet et al. [4], using *tweets*, forecast Hillary Clinton's victory over Donald Trump in the 2016 USA presidential election, through combining natural language processing and machine learning classification. The found result was in accordance with the index *New York Times National Polling Average* which aggregates information from several renowned traditional research institutes.

Burnap et al. [5] used sentiment analysis on *tweets* not to forecast the votes of the candidates, but rather of the parties in the UK general elections in 2015. Although the result of the winning party was close to the forecast, this can not be generalized because it assumed that the voters were equally distributed and this may have been the cause in the distortion of the results.

Saleiro et al. [14] compared the forecast based on over 230,000 tweets generated by 100,000 different users with traditional polls in the 2016 Portuguese parliament elections. The tweets were classified using a sentiment analysis algorithm trained over a *corpus* of 1,500 *tweets* manually annotated by 3 political science students. Their forecasts with a random forests model for a 12 month period reached a mean absolute error of 3.1 when compared with polls.

Using tweets as data source for the prediction of election results has some drawbacks. It is known that computational agents (*robots*) are used to propagate ideas, for instance by simply re-tweeting. Also, tweets can be initiated by any individual, which opens the door for a significant amount of disrespectful messages that may generate a high volume of subsequent tweets. Consequently, tweets are a data source significantly subject to bias and noise.

In an attempt to circumvent these problems, Sápiras et al. [15] described a method for mining the polarity of comments associated with political articles published in online newspapers. They showed to be possible to identify, classify and cluster comments of articles related to issues of health and education, in terms of their polarity. Their results led us to investigate the usefulness of comments associated with political articles published in online newspapers as information source for election predictions. Additionally, we also investigated the likes and dislikes associated to comments as additional information source. They can be considered as an indirect way of showing support for or opposition to candidates.

We are aware that comments on news in public media are usually moderated by staff, which naturally introduces some bias. However, these moderation actions are subject to public scrutiny and therefore they tend to be careful enough to limit the filtering to offensive comments. Also, by selecting high circulation and popular media we believe there is a higher chance of plurality in the data sources.

3 THE PREDNEWS APPROACH

PredNews is an election results predictor derived using a multivariate linear regression method over data collected from four different newspapers, in a four month-period, on election related news of ten different Brazilian cities. The model was then tested with data from four additional Brazilian cities. For assessment of the model results we later compare them to other non-explanatory models such as artificial neural networks (ANN) and poll results. This section explains the studies that were conducted on PredNews and analyses the ensuing model. The main unique features of our approach are:

(1) All comments used as data were collected from those made on political news of public written online media that mention the names of the candidates, segmented by region. Besides the candidates' names no other specific predefined keywords were considered to select comments;
(2) Comments appraisals, in the form of *likes* and *dislikes*, were considered as an additional data source;
(3) The research can be reproduced at any time, and other keywords can be added, since the data is still available on the original websites;
(4) PredNews generates an ordered list of candidates with their estimate percentage of votes.

3.1 Data collection

A comment in an online newspaper article is composed of its content, a date stamp, the identification of the user who made the comment, and the number of *likes* and *dislikes* of the comment. In

Table 1: The amount of political news by newspaper in the 4 months data collection period

Online newspaper	Number of Articles
Extra	688
Folha de Sao Paulo	331
G1	2930
Gazeta do Povo	273

Table 2: Training dataset - Amount of news and comments mentioning candidates, by municipalities

Municipality	News	Comments	Population
Belo Horizonte	289	1789	2.513.451
Campos	168	611	487.186
Fortaleza	154	3480	2.609.716
Guarulhos	93	653	1.337.087
Osasco	228	2696	696.382
Manaus	68	270	2.938.092
Nova Iguacu	145	1920	797.435
Porto Alegre	222	1476	1.481.019
Curitiba	133	1048	1.893.997
Santos	104	5046	434.359

Table 3: Test dataset - Amount of news and comments mentioning candidates, by municipalities

Municipality	News	Comments	Population
Rio de Janeiro	460	10.196	6.498.837
Recife	364	1.890	1.625.583
Salvador	131	914	2.938.092
São Paulo	564	14.374	12.038.175

data collection we retrieved the comments concerning candidates from the municipalities used in the training and test datasets. This was carried out for four months, prior to the 2016 Brazilian mayor elections of 14 major municipalities. We chose as information source the four major Brazilian online newspapers in order to diversify the representativeness of the target audience where the articles were published. The chosen newspapers were *G1*[1], *Folha de São Paulo*[2], *Gazeta do Povo*[3], and *Extra Online*[4], providing different amounts of political news as can be observed in table 1. Municipalities for the training and test data sets were randomly assigned.

For each municipality, up to five of the main candidates were chosen, with the objective of maximizing the chance of each candidate to be the subject of relevant comments. There were cities with less than five candidates, in which case we considered all available candidates. We developed a web crawler to daily navigate on these online newspapers looking for target candidate names.

The newspapers were divided in sections such as General, Sports, Entertainment, Economy and Politics. Each article was classified in one of these sections. The webcrawler was tuned to navigate within articles classified as Politics. News that mentioned any of the targets were selected and all associated comments retrieved for further analysis. The resulting numbers of news and comments are presented in tables 2 and 3. The imbalance between the number of comments in the training and test sets was purposeful to test for the generalisation capability of the model. On average, 26 people have posted comments associated with each article. Each person included, on average, 5.2 comments associated with the same article. The number of posts varied from 1 to 361. As shown in Figure 1, most people post comments only in one article.

The posts have on average 112 characters, which is within the range of Twitter posts [1]. However, there is no restriction on posts' size and messages varied from 1 to nearly 600 characters. Investigating the posts, we noticed that very short messages were agreement or disagreement signs. However, the short messages were automatically eliminated because there was no clear reference to any candidate. We tried to infer in short messages such as "Ele é o cara" that means "He is the man", which candidate the post referred to. However some ambiguities inevitably remained. We noticed a large amount of short messages compared to the smaller amount of long messages.

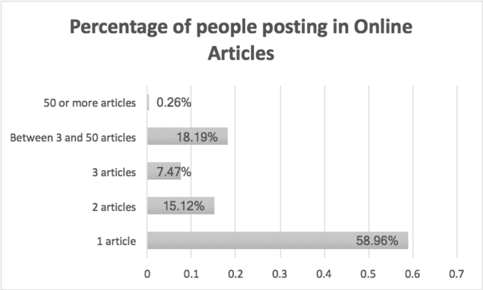

Figure 1: Percentage of people posting comments to 1 or more articles in online newspapers

The amount of comments would be overwhelming for manual processing, consequently the news crawler was key for PredNews. It only collected comments that mentioned the candidates by name or other predefined reference, such as well-known nicknames.

3.2 Data analysis

We used Sentilex algorithm [16] to classify the polarity of each comment related to each candidate. Let x be an element of the set of candidates X of one election process, let c be an element of the set of comments C, and let Sc be the function which evaluates the polarity of sentiment of a comment containing a reference to the candidate, where the result of this function is P for positive, N for negative and Z for neutral. Being K the total of comments we have[5]

$$positive\_comments(x) = \sum_{c=1}^{K} 1\,[Sc(c,x) = \mathrm{P}], \qquad (1)$$

[1] http://g1.globo.com/

[2] http://www.folha.uol.com.br/

[3] http://extra.globo.com/

[4] http://www.gazetadopovo.com.br/

[5] Using Iverson's bracket [11]: value inside brackets produces 0 or 1 as the logical expression evaluates to false or true respectively.

Table 4: Independent variables first tried to build the model

Size of city population
Number of characters in the message
Number of mentions to candidates
Number of candidates in the city
Number of likes on posts
Number of dislikes on posts
Number of likes supporting a candidate
Number of dislikes related to candidates
Number of comments associated with articles
Number of people posting messages associated to an article

$$negative\_comments(x) = \sum_{c=1}^{K} 1\,[Sc(c,x) = \mathrm{N}], \tag{2}$$

$$neutral\_comments(x) = \sum_{c=1}^{K} 1\,[Sc(c,x) = \mathrm{Z}]. \tag{3}$$

Since the same user can express *likes* or *dislikes* in more than one comment, we can not simply quantify data considering it all to be from different people. Thus, to better estimate the opinions of different users, the percentage of returning visitors was taken into account, which indicates the percentage of users returning from previous visits to the newspaper's website. In this way we tend to avoid bias towards opinions of more active users.

We started by considering the independent variables listed in table 4. Our initial analysis did not present any relevant correlations. Consequently, we investigated the derivative parameters. Let Lc be the function that counts the number of *likes* in a comment, Dc the function that counts the number of *dislikes* in a comment and α the returning visitor percentage. We have

$$negative\_likes\_index(x) = \left(\max_{c} Lc(c,x) + (1-\alpha)\sum_{c=1}^{K} Lc(c,x)\right)[Sc(c,x) = \mathrm{N}], \tag{4}$$

$$negative\_dislikes\_index(x) = \left(\max_{c} Dc(c,x) + (1-\alpha)\sum_{c=1}^{K} Dc(c,x)\right)[Sc(c,x) = \mathrm{N}], \tag{5}$$

$$positive\_likes\_index(x) = \left(\max_{c} Lc(c,x) + (1-\alpha)\sum_{c=1}^{K} Lc(c,x)\right)[Sc(c,x) = \mathrm{P}], \tag{6}$$

$$positive\_dislikes\_index(x) = \left(\max_{c} Dc(c,x) + (1-\alpha)\sum_{c=1}^{K} Dc(c,x)\right)[Sc(c,x) = \mathrm{P}]. \tag{7}$$

As in our test dataset, we have information regarding all candidates ($x \in X$) from each municipal election, we should use the aggregated functions for each candidate always in relation to their competitors in the same municipality, computed as percentages. Therefore, all variables are converted into shares as illustrated in equation 8 for *positive comments*.

$$positive\_comments\_share(x) = \frac{positive\_comments(x)}{\sum_{x \in X} positive\_comments(x)} * 100\% \tag{8}$$

The final list of variables available for the training process is presented in table 5. Notice that each of those variables will have as many instances as candidates in the election considered.

Table 5: Set of independent variables available to build the model. All variables have instances for each candidate

Positive comments share
Neutral comments share
Negative comments share
Negative likes index share
Negative dislikes index share
Positive likes index share
Positive dislikes index share

3.3 The PredNews Model

We decided to use a multivariate linear regression technique to obtain a human-readable formula that could be used to justify the predictions. The model was created using the aggregate data of the candidates from the 10 municipalities chosen for training (see table 2) and considering $\alpha = 70\%$ for the "returning visitor". In this process, the tool weka [10] was used with the 10-Fold cross-validation test strategy together with the M5 method for attribute selection, which removes the attributes with the lowest correlation coefficient until there is no further improvement in the error coefficient [13]. As result of this learning process, we obtained the prediction model presented in equation 9.

$$\begin{aligned} vote\_share(x) = 0.5392 * positive\_comments\_share(x) + \\ 0.4942 * neutral\_comments\_share(x) - \\ 0.2854 * negative\_likes\_index\_share(x) + 5.0375 \end{aligned} \tag{9}$$

This produced a correlation coefficient of 0.86 and a mean absolute error (MAE) of 6.69 in the training set. We observe that only three parameters (aggregate functions) were relevant. The more *positive comments share* and the more *neutral comments share*, the higher will be the candidate's percentage of votes. On the other hand, the more *negative likes index share* the smaller will be the percentage of votes. This is consistent with the intuitive perception that candidates with more positive comments have a higher chance of being elected. From this we also get to know that the four variables not used by the regression model have a lower power to predict the election results, relatively to the three selected ones.

To assess the model robustness we repeated the learning process leaving out each one of the 10 municipalities of the training set. The MAE increased slightly with a maximum of 9.8, and the correlation decreased with a minimum of 0.67. In all cases the selected variables did not change. Therefore we may conclude that the model is robust to variations in the dataset.

Table 6: Results on the test set - a checkmark means that the two candidates to go through to the second round were predicted in the correct order. Correct predictions for those candidates were obtained in 7 out of 8 cases (87.5% accuracy)

Election	Candidate	Real	PredNews
Rio de Janeiro	Crivella	32.62	✓44.794
Rio de Janeiro	Freixo	21.44	✓33.026
Rio de Janeiro	Pedro Paulo	18.93	10.869
Rio de Janeiro	Bolsonaro	16.44	4.785
Rio de Janeiro	Índio	10.55	6.488
Recife	Geraldo Júlio	49.72	✓32.064
Recife	João	23.94	✓23.076
Recife	Daniel Coeho	18.73	21.203
Recife	Priscila Krause	5.47	13.722
Recife	Edilson Silva	2.11	9.9
Salvador	ACM Neto	74.24	✓57.471
Salvador	Alice Portugal	14.6	11.543
Salvador	Sargento Isidório	8.64	6.567
Salvador	Cláudio	1.46	13.322
Salvador	Fábio Nogueira	1.04	11.081
São Paulo	João Dória	54.96	✓39.157
São Paulo	Fernando Haddad	17.22	✓22.181
São Paulo	Celso Russomanno	14.06	12.019
São Paulo	Marta	10.45	13.079
São Paulo	Erundina	3.28	13.535

4 RESULTS WITH THE TEST SET

The test dataset was used to verify the generality of the model. The MAE obtained in this dataset is 8.2, a small increase over the training data, confirming the robustness of the model. In table 6, we can observe that 7 out of 8 (87.5%) of the cases in which the candidates were eligible to the second round were hit in the correct order. These results show the feasibility of the hypothesis.

To compare the accuracy with other algorithms, including ANN and the Simple mention of candidates method [18], the result was discretized in terms of FIRST, SECOND and OTHER. Table 7 illustrates the comparison among the algorithms.

The IBOPE, which is a traditional, well-accepted, Brazilian public opinion poll, presented the best results hitting 100% of the cases. The proposed model, with a relatively much smaller financial cost and logistics than IBOPE, obtained nearly 90% of accuracy in relation to the placement of the first two candidates.

As for the ANN, 6 out of 8 (75%) were correctly classified. This model does not meet the premise that we will necessarily have a “FIRST" candidate, a “SECOND" place candidate and three candidates classified as “OTHER". The model of mentioning the candidate’s name, has the smallest accuracy (5 out of 8 well classified).

To verify the efficiency of the proposed model for the second round, with only two candidates, comments from newspapers during the corresponding period up to the election day were collected from the cities of *Rio de Janeiro* and *Curitiba*. The total of the voting percentages may not result in 100% of the votes because they were not normalized. In this example all the candidates’ positions in both elections were correctly predicted showing the feasibility of the hypothesis. The result is illustrated in table 8.

5 CONCLUSIONS AND FUTURE WORK

This paper presented PredNews, a model for predicting the results of elections. It only uses opinions expressed in the form of comments associated to political articles published in online newspapers, and is a faster and cheaper approach than traditional methods of public opinion polls. The results of an experiment carried out in Brazilian municipal elections support the hypothesis that it is possible, only with news’ comments data to predict election results. The performance was compared to a poll result by a renowned Brazilian institution, IBOPE, to an ANN model, and to simple mention of the candidate’s name. PredNews predicted the two candidates with the highest percentage of votes, achieving 90% accuracy, and 100% in the winning candidate on the second round, in the municipalities used to test the model.

It should be emphasized that the PredNews model was developed and tested considering more than one election. In fact, PredNes covered 14 distinct elections, of which ten were used as the training dataset and four were used as the testing dataset, while most related research proposes forecasting election models based on data from just one election. Noteworthy, it is the fact that the performance in the testing dataset is virtually identical to that obtained in the training phase, with a low mean absolute error (MAE), below 10 percent points. Additionally, the model is shown to be robust since it did not significantly change in face of leave-one-out training trials. In summary, PredNews reliably predicted local elections results in Brazil and it therefore seems feasible to apply it to other contexts.

Currently, we are considering whether the obtained model is general enough to predict results of similar elections in other Brazilian municipalities since Brazil is a vast country with thousands of cities. It is also unknown if the model can produce accurate results in other types of elections in Brazil, such as presidential or parliamentary. In addition, it will be interesting to check the validity of PredNews outside Brazilian borders. However, the strongest challenge is to verify the validity of PredNews over time. Given that new technologies are evolving very fast, the type of interaction people have with online media may prove to be significantly different in the near future. What kind of adaptation will be required to provide a good predictor of the Brazilian 2018 elections?

Although PredNews proved to be efficient in predicting eligible candidates for the second round of municipal elections, the existence of important networks of users in social media may be influential. In future work the predictor should use as a source of data not only comments in newspaper but also related *tweets* and statistical information taken from the *Facebook*, such as the amount of followers on the candidate’s page during the election period.

For sentiment analysis, we used a strategy based on the dictionaries made available by [16]. It did not take into account candidates’ nicknames, events’ associations, or even the candidates’ party number (in Brazilian elections parties are numbered) as additional ways to refer to the candidates within the comments. In some comments we noticed indirect references to candidates by either their previous job, some physical defect or even naming pranks. For example “bispo" (*bishop*) was used to negatively address the candidate *Crivella*, in the 2016 *Rio de Janeiro* elections for mayor, since he was a former protestant preacher. Similarly, the term “frouxo" (*irresolute*), in that same election, was used to disqualify candidate *Marcelo*

Table 7: Forecast models comparison

Election	Candidate	Real	IBOPE	ANN	Simple mention	PredNews
Rio de Janeiro	Crivella	FIRST	✓FIRST	✓FIRST	✓FIRST	✓FIRST
Rio de Janeiro	Freixo	SECOND	✓SECOND	FIRST	OTHER	✓SECOND
Rio de Janeiro	Pedro Paulo	OTHER	OTHER	OTHER	SECOND	OTHER
Rio de Janeiro	Bolsonaro	OTHER	OTHER	OTHER	OTHER	OTHER
Rio de Janeiro	Índio	OTHER	OTHER	OTHER	OTHER	OTHER
Recife	Geraldo Júlio	FIRST	✓FIRST	✓FIRST	SECOND	✓FIRST
Recife	João	SECOND	✓SECOND	✓SECOND	FIRST	✓SECOND
Recife	Daniel Coeho	OTHER	OTHER	OTHER	OTHER	OTHER
Recife	Priscila Krause	OTHER	OTHER	OTHER	OTHER	OTHER
Recife	Edilson Silva	OTHER	OTHER	OTHER	OTHER	OTHER
Salvador	ACM Neto	FIRST	✓FIRST	✓FIRST	✓FIRST	✓FIRST
Salvador	Alice Portugal	SECOND	✓SECOND	OTHER	✓SECOND	OTHER
Salvador	Sargento Isidório	OTHER	OTHER	OTHER	OTHER	OTHER
Salvador	Cláudio	OTHER	OTHER	OTHER	OTHER	SECOND
Salvador	Fábio Nogueira	OTHER	OTHER	OTHER	OTHER	OTHER
São Paulo	João Dória	FIRST	✓FIRST	✓FIRST	✓FIRST	✓FIRST
São Paulo	Fernando Haddad	SECOND	✓SECOND	✓SECOND	✓SECOND	✓SECOND
São Paulo	Celso Russomanno	OTHER	OTHER	OTHER	OTHER	OTHER
São Paulo	Marta	OTHER	OTHER	OTHER	OTHER	OTHER
São Paulo	Erundina	OTHER	OTHER	OTHER	OTHER	OTHER
		Accuracy	100%	75%	62.5%	87.5%

Table 8: Second round results

City	Candidate	Real	Forecast	Err	Position
RJ	Crivella	59.36	51.40	-8	1st ✓
RJ	Freixo	40.64	33.47	-7	2nd ✓
Curitiba	Rafael Greca	53.25	56.09	2	1st ✓
Curitiba	Ney Leprevost	46.75	28.77	-18	2nd ✓

Freixo by a pun with his last name. In addition, ironic and double-meaning expressions may have misled automatic interpretations, of current sentiment analysis methods. According to [6], 35% of opinions considered positive are related to irony. Consequently a possible future approach would be to add the regional terms identifying candidates and the corresponding sentiment polarity.

Our study considered each comment referring to only one candidate and with no references to other comments. We are currently adjusting PredNews model to move from this flat representation to a more graph-oriented one in which we can automatically filter opinions not related to the candidates, but concerning the comments' authors, or the comments' content. It is also in our future plans to adjust PredNews to account for users' changes of opinion.

Finally we stress the potential of such an approach to be extended beyond prediction of election results. A tool providing an anytime realistic prediction of popular intention concerning societal issues is definitely useful for local and wider forms of government.

REFERENCES

[1] CM Alis, MT Lim, HS Moat, D Barchiesi, T Preis, and SR Bishop. 2015. Quantifying Regional Differences in the Length of Twitter Messages. *PLoS ONE* 10, 4 (2015), e0122278. DOI: http://dx.doi.org/10.1371/journal.pone.0122278

[2] Jussara M Almeida, Gisele L Pappa, and others. 2015. Twitter Population Sample Bias and its impact on predictive outcomes: a case study on elections. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, 1254–1261.

[3] Marta Arias, Argimiro Arratia, and Ramon Xuriguera. 2013. Forecasting with twitter data. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 1 (2013), 8.

[4] Alexandre Bovet, Flaviano Morone, and Hernan A Makse. 2016. Predicting election trends with Twitter: Hillary Clinton versus Donald Trump. *arXiv preprint arXiv:1610.01587* (2016).

[5] Pete Burnap, Rachel Gibson, Luke Sloan, Rosalynd Southern, and Matthew Williams. 2016. 140 characters to victory?: Using Twitter to predict the UK 2015 General Election. *Electoral Studies* 41 (2016), 230–233.

[6] Paula Carvalho, Luís Sarmento, Mário J Silva, and Eugénio de Oliveira. 2009. Clues for detecting irony in user-generated contents: Oh...!! "it's so easy" ;-). In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion.* ACM, 53–56.

[7] Daniel Gayo-Avello. 2011. Don't turn social media into another'Literary Digest'poll. *Commun. ACM* 54, 10 (2011), 121–128.

[8] Daniel Gayo-Avello. 2012. No, you cannot predict elections with Twitter. *IEEE Internet Computing* 16, 6 (2012), 91–94.

[9] Douglas Hagar. 2015. # vote4me: the impact of Twitter on municipal campaign success. In *Proceedings of the 2015 International Conference on Social Media & Society.* ACM, 19.

[10] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* 11, 1 (2009), 10–18.

[11] Donald Knuth. 1992. Two Notes on Notation. *Amer. Math. Monthly* 99, 5 (1992), 403–422.

[12] Renato Miranda Filho, Jussara M Almeida, and Gisele L Pappa. 2015. Twitter population sample bias and its impact on predictive outcomes: a case study on elections. In *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on.* IEEE, 1254–1261.

[13] John R Quinlan and others. 1992. Learning with continuous classes. In *5th Australian joint conference on artificial intelligence*, Vol. 92. Singapore, 343–348.

[14] Pedro Saleiro, Luís Gomes, and Carlos Soares. 2016. Sentiment Aggregate Functions for Political Opinion Polling using Microblog Streams. In *Proceedings of the Ninth International C* Conference on Computer Science & Software Engineering.* ACM, 44–50.

[15] Leonardo Augusto Sápiras and Karin Becker. 2014. Mineração da opinião sobre aspectos de candidatos a eleições em comentários de notícias. In *Proceedings of Simpósio Brasileiro de Banco de Dados.* 117–126.

[16] Mário J Silva, Paula Carvalho, and Luís Sarmento. 2012. Building a sentiment lexicon for social judgement mining. In *International Conference on Computational Processing of the Portuguese Language.* Springer, 218–228.

[17] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. 2010. Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social Science Computer Review* (2010), 0894439310386557.

[18] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *ICWSM* 10 (2010), 178–185.