Inferring the votes in a new political landscape. The case of the 2019 Spanish Presidential elections.

Didier Grimaldi (didier.grimaldi@salle.url.edu)
Universitat Ramon Llull https://orcid.org/0000-0002-1027-1176
Javier Diaz
Universidad Icesi
Hugo Arboleda
Universidad Icesi

Research

Keywords: machine learning, election, prediction, Spain, Twitter

Posted Date: March 10th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-16463/v1

Version of Record: A version of this preprint was published on August 6th, 2020. See the published version at https://doi.org/10.1186/s40537-020-00334-5.

Abstract

Electoral prediction from Twitter data is an appealing research topic. The article aims at inferring the results of 2019 Spanish Presidential elections analysing Tweets. It defines a specific political dictionnary to analyse the sentiment and the opinion of the messages posted during the campaign. Our predicting model compares the performance of 5 multi-linear regression algorithms and our results are compared to the ones delivered by the standard poll systems based on telephone survey. Our methodology correctly ranks the candidates and gives for the winner of the election (Sanchez) a better prediction of voting share than the national polls. This stream of studies is still in the early stage even if our findings look like very promising. Therefore, as a future line of research, we recommend to include more socio- and economic factors like sex, age, location, etc. in the objective to improve our model and results.

Introduction

For more than a decade now, with the emergence of Internet 2.0, users are able to generate their own content and share it publicly more easily. In this boom, social networks have endorsed great popularity, in particular the Twitter microblogging platform which allows its users to share with their family, friends and followers text messages of 140 characters maximum. More than 500 million messages are daily published which are commonly called tweets (Dietrich & Juelich, 2018; Marozzo & Bessi, 2018). Because the main reason for these publications is to express the point of view and the opinion of the users, they result to be of great interest to be analyzed (Cerchiello & Giudici, 2016; Lokers et al., 2016a).

This huge amount of content circulating in them has attracted first the attention of marketing agencies that expect to seize the behavior of clients (current or future) to adjust their online campaigns and even use the social content (number of likes, retweets, etc.) to predict sales intention in the real world (Volkova et al., 2015). Recently this type of analysis has jumped into the field of politics to try to predict the results of campaigns by monitoring the interaction between candidates and voters (Magalhães et al., 2012). Since more and more people are posting on the Internet, it generates the idea to researchers and journalists that a collective feeling is present in the social media, ready to be listened, captured and analyzed (Budiharto & Meiliana, 2018; Cury, 2019). Moreover, most of the social media (Linkedin, Facebook, Twitter, etc.) make available to anyone the use of API's which allow to collect the data published by its users and to take the pulse of the public opinion (Le et al., 2017).

Considering how in Spain the electoral cyberspace is articulated, we can contemplate several possible spaces for the analysis of a political campaign: Facebook, Twitter, Instagram or Pinterest. For our manuscript, we decide to use Twitter and our key contribution is to capture and analyze if we can infer the voting tendencies and predict the 2019 Presidential Spanish elections results with this media. The novelty of our paper is to study if a social media can be a better predicting tool while the political landscape has suffered a big change.

Indeed, in the race for the presidency in 2019 five candidates competed for one of the most important Spanish political elections, including the President Sanchez candidate for the left party and for his re-election. The other candidates were Rivera for the central-right party, Iglesias for the far-left party, Casado for the right party and Abascal for the far-right party (called "*Vox*"). The historical level of participation (75.7%) shows the importance of this scrutiny due to two reasons. On one hand, the context of high political divide due to the question of the Catalan referendum for independency and on the other the emergence of the far-right party proposing very reactionary arguments such as the end of the authorized abortion or a recentralisation of the national governance around Madrid capital of Spain.

"Vox" party is competing for the first time in this type of election and jeopardizes the standard system of polls forecasting, usually based on the scores obtained in the previous elections (Huberty, 2015) and Twitter could be a better predicting tool. The vote participation was about 9 points more than in the last election and the greatest of the 21st century. The presidential election was held under a one-round voting system held on April 28th, 2019.

This paper is organized as follows. Section II reviews the literature about the use of social media to infer elections while the section III explains the methodology used to get a clean dataset to study. In Section IV, the results are described and analysed.

Section V concludes and proposes further lines of research.

Related work

The Twitter microblogging network allows for its users to place their thoughts, feelings and opinions in the form of text and concerning a wide variety of issues. It is a space from which it is possible to extract the public opinion using linguistic programming models (Preoţiuc-Pietro et al., 2017). These methods of analysis are known as sentiment or sentimental analysis. Recent studies (Cerón-Guzmán & León-Guzmán, 2016; Volkova et al., 2015; Cohen & Ruths, 2013) examine if it is possible to infer election results through Tweets analysis. Tumasjan et al. (2011) compare the results of Twitter analysis with the ones of traditional surveys. They look for the margins of error taking as example the latest electoral polls spread in Germany and their conclusions are that the analysis of tweets mentioning a political party can be considered as a plausible reflection of the vote share and its predictive power comes close to traditional election polls.

Shi et al. (2012) during 2010 US legislative elections compare the number of followers achieved by each candidate on Twitter with the electoral predictions. They conclude that in 71% of the cases, the candidate with the highest number of followers is also who occupies the first position in the polls. However, these results are challenged by Cha & Gummadi (2010) who state that the number of followers (they coined them as indegree) is a sign that reveals the popularity of a user but not necessary related to its ability to influence the audience. Their study reinforces the hypothesis of the "fallacy of one million followers" endorsed in a previous research paper of (Avnit, 2009) in which this latter shows that some users follow others simply by education, becoming followers of those who also follow them but who hardly read the tweets published by whom they follow.

O'Connor et al., (2010) collect one billion messages posted on Twitter during 2008 and 2009. They make two studies to compare the "consumer confidence" with the Index of Consumer Sentiment (ICS) of Reuters and with the Economic Confidence Index of the Gallup consulting company for presidential job approval. Their findings are in the case of "consumer confidence", twitter and standard opinion survey have similar results with a correlation of 80% but found divergent results regarding the topics of "political opinion": for presidential job approval in 2009 the data obtained from Twitter replicates those obtained in the polls but for 2008 the correlation is not significant. They however conclude expensive and time-intensive polling can be supplemented or supplanted in certain conditions with the simple-to-gather text data generated from online social networks.

Recently there has been a growing interest in building a proposal of textual or linguistic classifiers of public opinion enable to name as favorable or unfavorable opinions expressed by individuals about a proposal or political issue (Ramteke et al., 2016). The methodological approaches that have been used so far (mainly dictionary of terms used or pre-tagged texts) have not achieved the level of accuracy as those obtained in the journalistic critics on topics such as business or consumer goods. Yu et al. (2008) analyze the speeches of the US Senate during the period 1989-2006 and those of the Congress in 2005 and compare them with the articles on business and critiques about films in 2006. They found the difficulty that exists today to carry out an effective sentiment analysis in the field of political opinion. Gayo-Avello (2013) argues since not all the potential voters express their opinion on Twitter, the "sample" collected is not representative of the population. Moreover, he adds that these networks easily allow the manipulation from spammers and propagandists bringing noise to the real message. He states also voters who live in urban centers as well as young adults are more likely to use Twitter as a form of expression, and these latter show a tendency towards left party opinions in politics. Smith & Rainie (2008) corroborate these results revealing for 2008 US elections that 65% of Obama's online followers have consulted online political information compared to 56% for McCain, having made more campaign contributions, more online petition signatures, more blog comments than the rest of the electorate.

On the other hand, Grimaldi (2019) highlights for the 2019 Spanish Presidential elections that the number of information / conversations / messages circulating on Twitter grows as the election campaign period progresses, showing peaks in line with relevant events that affect or happen in it, and whose volume falls after the election voting day. The sentiment analysis technique needs to adapt to this evolution incorporating for instance new hashtags which were not present at the beginning of the campaign (Grimaldi, 2019). The fact that campaigns are primarily manifested through Twitter media before standard general channels like TV or radio raises the need to continue investigating the best approach to analyze the data available in

these networks and reveal the messages conveyed. Our manuscript aims at answering if we can capture and infer the voting tendencies with Twitter and if this media is more reliable than standard public opinion polls in a new political landscape.

Methodology

Collect of data

With the Twitter Streaming API[1], we collect all the posts containing one or more hashtags related to the presidential election every day which have been clustered as follows:

- Election general information = #28A; #28Abril; #Vota; #Vota28A; #Elecciones Generales; #28 Abril; #EspanaVaciada
- Political party = #ValorSeguro; #Casado; #laEspaña que quieres; #PedroSanchez; #LaHistoriaLaEscribesTu; #PabloIglesias; #VOX; #SantiagoAbascal; #VamosCiudadanos; #AlbertRivera

The tweets are collected real time, 24 hours after their publication. Our aim is to get some statistics related to the popularity of the tweet such as (a) the number of shares or retweets, which shows how many users shared the tweet with their followers; (b) the number of likes, which shows how many users found the tweet relevant (c) the number of retweets which shows how many tweets are broadcasted within the tube. Between April 12th and April 28th 1.170.000 tweets contributed and table 1 shows the amount of collected data in terms of users and tweets per candidate.

Table 1: classification of the keywords by political faction

Candidate	Nº Tweets
Pedro Sanchez	320.760
Pablo Iglesias	226.472
Albert Rivera	217.423
Pablo Casado	217.380
Santiago Abascal	188.828

When working with the Twitter API it was necessary to take into account certain limitations of it. The frequency limit of the interface only allows 450 requests every quarter of hour. Moreover, Twitter company filters tweets published considering them irrelevant for the user.

Data preprocessing

The dataset resulting from data collection is a collection of tweets. In order to apply sentiment analysis techniques on these data, it is necessary that they go through a process comprised of several phases, also known as pipeline, where the input of each phase is the output of the previous one. The final result will be a matrix that represents the relevant information of the collected data. A tweet is a complex object with many properties, but what we are interested in is the message that the user wrote in the form of a text. So, the collection of tweets is transformed into our corpus by extracting the message of each tweet. While doing this step we lose the information about the author of the tweet and the date of publication, but we keep the reference to the object to be able to link the final result with the original tweet. The result of this stage is a list of sentences or documents.

Secondly, each document resulting from the previous stage is transformed into a list of words and symbols called tokens. In general, tokens are strings of characters between blanks or punctuation, but this is not always the case, as for example in the case of abbreviations. The total set of words used, different and unique, is the vocabulary of the corpus. This step also filters the functional words, which do not have a clear referential semantic, such as articles, pronouns, prepositions… what they are usually called as stop words. It is also necessary to delete certain functional words from the Twitter glossary as \ RT ; \ HO and \ HT". The next step is to generate a sparse matrix by transforming each word list in a vector in a Euclidean space, where each column is a feature. The characteristics are mainly words of the vocabulary extracted from tokenization. We can consider each word of the vocabulary in a column, or we can obtain more detailed representations if we consider as possible characteristic sequences of words, called n-grams, that have occurred in the text. The n-grams can be sequences of one word (unigrams), two words (bigramas), three words (trigrams), and so on (Manning & Raghavan, 2009). Each text of each tweet is represented as a vector $d = \{t1, t2,\ldots tn\} \in R^n$ / where n is the size of the vocabulary. The sum of vectors consists on a Document-term matrix.

New polarity political lexicon

Due to the lack of a sentiment lexicon for non-English language, we decide to create a new polarity lexicon for Spanish political event from two different sources. As a starting point, we use the dataset created by (Molina-González et al., 2013) for Spanish tweets. Then, we choose a random sample of 1.000 tweets as the training set. In order to label a tweet as either positive, negative or neutral, eight volunteers in group of 2 (i.e. 4 groups) were asked to assign a label for each tweet according to the sentiment they understood it was conveyed. The evaluators had access to the full tweet, which is tagged according to its global polarity, indicating whether the text expresses a positive, negative or neutral sentiment. For disambiguation, they had to assess the tone of some tweets after screening the other tweets of the account. Moreover, if there was no agreement inside the group, a third independent volunteer from another group was compelled to help to decide. Volunteers agreed in 80% of tweets which support the statement that humans often disagree on the sentiment of a text (Villena et al., 2015) and determine which additional word should be included in the lexicon. In total 187 words were added based on this manual testing and included in the final vocabulary.

Machine learning classification for sentiment analysis

The sentiment analysis system classifies a given vector as either positive, negative or neutral by summing the number of positive words and subtracting the number of negative words. It assigns only a label class to the tweet. This type of analysis makes sense if we assume that each tweet expresses a single opinion. It may seem like a limitation, because in practice it works well since users usually focus on a single topic in each tweet. Surely in other contexts, or if this limitation of message length were not available, it would be good to consider more complex analysis systems that allow more granular analysis. A sentiment labelled dataset of 1.000 tweets is split into two sets: 80% of the tweets are used for the training set and the remaining as the test set. The splitting of the data is performed in a stratified way.

The process of cross validation is repeated during K iterations, with different possible subsets of test data. The arithmetic mean of the results is done to obtain the final result. This method is very precise since we evaluate K combinations of data from training and testing but has the disadvantage to be slow from the computational point of view. In our case, the classifier is trained on the training set via K=10-fold cross-validation using the R library ISLR implementation of the Logistic Regression algorithm. The system achieves a macro-averaged F1-score of 92,65% and an accuracy of 93.08% on the test set.

[1] https://dev.twitter.com/streaming/overview. Last access on 10-11-2019

Results And Discussion

Feature of the model

Following recommendations from (Gayo-Avello, 2012, 2013; Metaxas & Gayo-Avello, 2011), the following features which are the independent variables used by the inference method, are computed in a daily basis:

1) Day tweet volume: the sum of tweets on that day mentioning a candidate.

2) Day unique tweet volume: the sum of tweets on that day that only mentions a candidate.

3) Day Twitter user number: the sum of different Twitter users with at least one tweet mentioning a candidate.

4) Day unique Twitter user number: the sum of different Twitter accounts whose posts only mention a candidate.

5) Positive or negative tweet volume: the sum of positive or negative posts that mentions a candidate.

6) Positive or negative- based Twitter user number: the sum of different Twitter users with at least one positive or negative posts mentioning a candidate.

7) sentiment score per tweet

The features are normalized by applying the moving average smoothing technique over a window of the past seven days, as it is proposed by (O'Connor et al., 2010). Moreover, the polling data is considered as the dependent variable of our model.

Inference method

The voting intention inference was approached as a multiple linear regression analysis. In this way, several regression models were built to infer the vote of each candidate in the electoral round, using the aggregated polling as the output variable of the models. We use the *fit.models* package in R software. In total, 5 models were built i.e. one model per candidate. We evaluate 5 different algorithms: Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART), k-Nearest Neighbors (kNN)., Support Vector Machines (SVM) and Random Forest (RF). Based on the performance of the mean absolute error (MAE), the best one with the lowest MAE was the LDA model. As an illustration, graph 1 shows the results of the different models applied to Sanchez candidate.

Graph 1: Performance of the 5 models applied to Sanchez candidate

Discussion

Table 2 shows our results. In the first column, it shows the 2019 Spanish Presidential official scores and in the third one the voting intention inference based on our method. As far as it concerns the second column, it presents the last polling results corresponding to 5 days before the election occurred as the Spanish law regulating the presidential elections authorizes it. Our results show our method a) correctly ranks the candidates, b) gives for the winner of the election (*Sanchez*) a better prediction of voting share than the last and definitive polls, c) provides a prediction equivalent as the last poll for *Abascal* candidate. However, they are worse than the last poll to estimate the voting share for the rest of the candidates (*Casado*, *Rivera* and *Iglesias*). Therefore, the obtained results show that the inference method based on Twitter data is reliable, but further research will have to improve and tune the model that we propose.

Table 2: Results and voting inference using the method with the lowest absolute error

Candidate	Election results	Polls (difference)		Inference with our method	
		Results	Difference with elections	Results	Difference with elections
Pedro Sanchez	28,7 %	27.3%	-1.4%	29.2%	+0.5 %*
Pablo Casado	16,7 %	19.4%	+2.7%*	19.9%	+3.2%
Albert Rivera	15,9 %	15.8%	-0.1%*	14.9%	-1.0%
Pablo Iglesias	14,3 %	14.4%	+0.1%*	12.8%	-1.5%
Santiago Abascal	10,2 %	10.2%	+0.0%*	10.2%	+0.0%*

*Highlighted in yellow the method delivering best results by candidate

Conclusions And Future Work

While our model alone may not be sufficient to predict the results, we can point out that our method to listen and analyze tweets has partially achieved our objective to be a fast, cheap and reliable tool for predicting public opinion and election results (Tumasjan et al., 2011). Our dataset was created by mining Twitter for the 16 days of the campaign; however, future lines of research can extend it and create an automated framework which collects data for months. Indeed, we believe election result prediction is a continuous process and requires analysis over a longer period. Finally, we believe future studies should analyze and propose weighting factors for our model to include in the statistical balance, the “inaudible” voice i.e. part of the population

who does not express their opinion using social media tools, in order to reduce the response bias and be more representative of the electors' demography.

Declarations

- Availability of data and materials: The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request
- Competing interests: The authors declare that they have no competing interests
- Funding: not applicable
- Authors' contributions: all the authors have made equivalent contribution to the conception, analysis of the work and have drafted the work and substantively revised it.
- Acknowledgements: not applicable

Bibliography

Avnit, A. (2009). The Million Followers Fallacy. *Internet Draft, Pravda Media*. Retrieved from http://tinyurl.com/nshcjg

Budiharto, W., & Meiliana, M. (2018). Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis. *Journal of Big Data*, *5*(1), 1–10. https://doi.org/10.1186/s40537-018-0164-1

Cerchiello, P., & Giudici, P. (2016). Big data analysis for financial risk management. *Journal of Big Data*, *3*(1). https://doi.org/10.1186/s40537-016-0053-4

Cerón-Guzmán, J. A., & León-Guzmán, E. (2016). A sentiment analysis system of Spanish tweets and its application in Colombia 2014 presidential election. *Proceedings - 2016 IEEE International Conferences on Big Data and Cloud Computing, BDCloud 2016, Social Computing and Networking, SocialCom 2016 and Sustainable Computing and Communications, SustainCom 2016*, 250–257. https://doi.org/10.1109/BDCloud-SocialCom-SustainCom.2016.47

Cha, M., & Gummadi, K. P. (2010). Measuring user influence in Twitter: The million follower fallacy. Retrieved from http://en.scientificcommons.org/58470236

Cohen, R., & Ruths, D. (2013). Classifying Political Orientation on Twitter: It's Not Easy! *Seventh International AAAI Conference on Weblogs …*, 91–99. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewFile/6128/6347%5Cnpapers3://publication/uuid/3532F9EA-312A-4F8E-83C3-6369D71D2171

Cury, R. M. (2019). Oscillation of tweet sentiments in the election of João Doria Jr. for Mayor. *Journal of Big Data*, *6*(1), 1–15. https://doi.org/10.1186/s40537-019-0208-1

Dietrich, B. J., & Juelich, C. L. (2018). When presidential candidates voice party issues, does Twitter listen? *Journal of Elections, Public Opinion and Parties*, *28*(2), 208–224. https://doi.org/10.1080/17457289.2018.1441847

Gayo-Avello, D. (2012). No, you cannot predict elections with twitter. *IEEE Internet Computing*, *16*(6), 91–94. https://doi.org/10.1109/MIC.2012.137

Gayo-Avello, D. (2013). *A Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data*. *Social Science Computer Review* (Vol. 31). https://doi.org/10.1177/0894439313493979

Grimaldi, D. (2019). Can we analyse political discourse using Twitter? Evidence from Spanish 2019 presidential election. *Social Network Analysis and Mining*, 1–9. https://doi.org/10.1007/s13278-019-0594-6

Huberty, M. (2015). Can we vote with our tweet? On the perennial difficulty of election forecasting with social media. *International Journal of Forecasting*, *31*(3), 992–1007. https://doi.org/10.1016/j.ijforecast.2014.08.005

Le, H. T., Boynton, G. R., Mejova, Y., Shafiq, Z., & Srinivasan, P. (2017). Revisiting The American Voter on Twitter, 4507–4519. https://doi.org/10.1145/3025453.3025543

Lokers, R., Knapen, R., Janssen, S., van Randen, Y., & Jansen, J. (2016a). Analysis of Big Data technologies for use in agro-environmental science. *Environmental Modelling and Software*, *84*, 494–504. https://doi.org/10.1016/j.envsoft.2016.07.017

Lokers, R., Knapen, R., Janssen, S., van Randen, Y., & Jansen, J. (2016b). Analysis of Big Data technologies for use in agro-environmental science. *Environmental Modelling and Software*. https://doi.org/10.1016/j.envsoft.2016.07.017

Magalhães, P. C., Aguiar-Conraria, L., & Lewis-Beck, M. S. (2012). Forecasting Spanish elections. *International Journal of Forecasting*, *28*(4), 769–776. https://doi.org/10.1016/j.ijforecast.2012.04.007

Manning, C., & Raghavan, P. . (2009). *Introduction to Information Retrieval*. *Computational Linguistics* (Vol. 35). https://doi.org/10.1162/coli.2009.35.2.307

Marozzo, F., & Bessi, A. (2018). Analyzing polarization of social media users and news sites during political campaigns. *Social Network Analysis and Mining*, *8*(1). https://doi.org/10.1007/s13278-017-0479-5

Metaxas, P., & Gayo-Avello, D. (2011). How_(Not)_To_Predict_Elections.pdf. *IEEE International Conference on Privacy, Security and Risk*.

Molina-González, M. D., Martínez-Cámara, E., Martín-Valdivia, M. T., & Perea-Ortega, J. M. (2013). Semantic orientation for polarity classification in Spanish reviews. *Expert Systems with Applications*, *40*(18), 7250–7257. https://doi.org/10.1016/j.eswa.2013.06.076

O'Connor, B., Balasubramanyan, Routledge, B. R., & Smith, N. A. (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*.

Perboli, G., De Marco, A., Perfetti, F., & Marone, M. (2014). A New Taxonomy of Smart City Projects. *Transportation Research Procedia*, *3*(July), 470–478. https://doi.org/10.1016/j.trpro.2014.10.028

Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., & Ungar, L. (2017). Beyond Binary Labels: Political Ideology Prediction of Twitter Users. *Proceedings Ofthe 55th Annual Meeting Ofthe Association for Computational Linguistics,* 729–740. https://doi.org/10.18653/v1/p17-1068

Ramteke, J., Shah, S., Godhia, D., & Shaikh, A. (2016). Election result prediction using twitter sentiment analysis. *2016 International Conference on Inventive Computation Technologies (ICICT)*. https://doi.org/10.1109/inventive.2016.7823280

Shi, L., Agarwal, N., Agrawal, A., Garg, R., & Spoelstra, J. (2012). Predicting US Primary Elections with Twitter, 1–8.

Smith, A., & Rainie, L. (2008). The internet and the 2008 election. *Spring*.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2011). Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social Science Computer Review*, *29*(4), 402–418. https://doi.org/10.1177/0894439310386557

Villena, J., García, J., Martínez, E., & Jiménez, S. (2015). TASS 2014 - The challenge of aspect-based sentiment analysis. *Procesamiento de Lenguaje Natural*, *54*, 61–68.

Volkova, S., Bachrach, Y., Armstrong, M., & Sharma, V. (2015). Inferring Latent User Properties from Texts Published in Social Media. *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI)*, 4296–4297.

Yu, B., Kaufmann, S., & Diermeier, D. (2008). Exploring the characteristics of opinion expressions for political opinion classification. *Proceedings of the 2008 International Conference on Digital Government Research*, (May), 82–91. Retrieved from http://dl.acm.org/citation.cfm?id=1367832.1367848

Figures

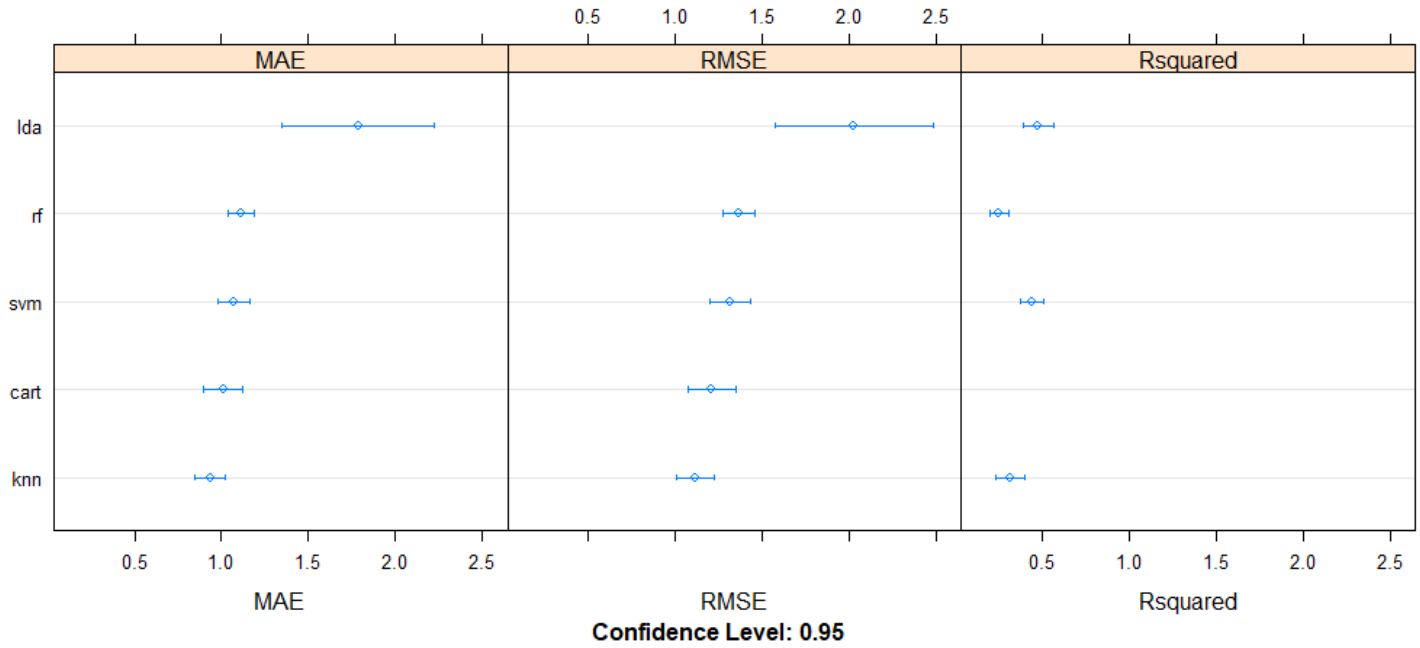

Figure 1

Performance of the 5 models applied to Sanchez candidate