Quality & Quantity (2020) 54:903–923
https://doi.org/10.1007/s11135-020-00965-z

Analyzing social media, analyzing the social? A methodological discussion about the demoscopic and predictive potential of social media

Pedro Santander[1] · **Rodrigo Alfaro**[1] · **Héctor Allende-Cid**[1] · **Claudio Elórtegui**[1] · **Cristian González**[1]

Published online: 10 January 2020

Abstract
The impact of computational technologies and the worldwide use of Internet entails a theoretical and methodological challenge for social scientists, considering the purpose of observing, interpreting and explaining human and social behaviour. Today, the digital environment seems to be an adequate space for this exploration and the emergence of the Web 2.0 offers common people the possibility of expressing and sharing their opinions on a daily basis. Due to the ubiquity of technology, Internet and social media in people's lives, socialization and its expressiveness have changed. If this is the case, the means to measure the perceptions, opinions and judgements of citizens should also change. The immense quantity of data available to be analysed today poses a challenge for the traditional scientific model. In this sense, it could be necessary for social research to move towards the analysis of the web and consider the potential predictive capacity of digital demoscopy. A new field of study has opened, with interest in exploring the predictive capacity of social media in electoral contexts. As a research group comprised by linguists, communication experts and engineers we explored the predictive potential of social media in three national elections that took place in Chile during 2017. Our objective was to explore a methodological design that allows predicting the result of political elections through the use of inductive algorithms and the automatic processing of messages with political opinion in social media. Through computational intelligence, we were able to follow, collect and analyse millions of tweets, and to improve our forecast each time. Our learning based on empirical research was fundamental to improve our procedures and to refine our variables and, thus, improve our prediction.

Keywords Social media · Digital demoscopy · Political communication · Election forecasting · Chilean's presidential election

✉ Pedro Santander
pedro.santander@pucv.cl

Extended author information available on the last page of the article

1 Introduction

The impact of computational technologies, Internet and mobile telecommunications is creating one of the greatest reconfigurations in humanity in the last 300 years. Changes occur exponentially, which have enabled the development of a new age of technology, based on the Internet of Things, robotics, nanotechnology and Artificial Intelligence (Schwab 2016). There have been transformations in political management and state governance models, in the formation of social movements, etc.

One of the consequences of this technological revolution (also known as the Fourth Industrial Revolution) is the increasing and unstoppable use of Internet and digital technologies in all areas of human life, both public and private. All indicators show that this use is more widespread and intensive every day, i.e., more people of all ages, genders, and social strata from different places around the world access Internet and spend more time staring at their digital devices to carry out multiple activities.

In just a few years, we moved from a situation in which many people that grew up with analog technology and had to adapt and migrate to digital technology, to a reality in which it is increasingly frequent to be digital natives and heavy users, i.e., people for whom digital devices are normal and part of their everyday lives. Not only people, but also various institutions—financial, banking, mass media, educational, corporate, etc.—had to adapt to change and incorporate ideas coming from digital environment.

In this regard, the study of digital technologies is increasingly important for social research, considering its purpose of observing human behavior, both institutional and social, in order to interpret it correctly and to develop methods and theories that enable the progress of knowledge on subjects, organizations and social entities. Consequently, this new revolution entails a challenge for scientists, both on a theoretical and methodological level. Social research is, per se, sensitive to historical and cultural changes, and is commonly faced by a recurring question: how and where should we "read" society in order to interpret, explain and hopefully foresee tendencies, dynamics, attitudes, behaviors, and opinions of human collectives? Today, the digital environment seems to be an adequate and even essential space for this exploration (Burrows and Savage 2014; Ceron et al. 2014; McCormick et al. 2015; Sloan et al. 2015). If this is true, it means that we must know how to work with new types of data, with both their quality (automated, digital, binary, instantaneous) and quantity (huge volumes); then, the challenge for social research, besides being theoretical and methodological, is also interdisciplinary.

Accordingly, many authors took notice of the potential of these new types of data that social scientists may use. Jungherr et al. (2017), for example, argue that the "digital trace data"—data produced by people while interacting with digital services—may have some potential for the study of public opinion, if methodological and conceptual measures are taken. Following that line we wanted to explore a methodological design that allows predicting the result of political elections through the automatic processing of political opinion in social media.

Beauchamp (2017: 490) argues that "the social media data "can track representative measures of public opinion (…) and can provide a method for extrapolating vote intentions" to do social media prediction. The nature of this type of data differs from both traditional data and survey data, which has been used in Social Sciences to analyze societies and make predictions. Unlike survey data, digital trace data enables a "social media analysis" whose role in social research is still an open question (Schober et al.

2016: 182), and within its context researchers are attempting to unlock the potential of digital trace data in the study of public opinion" (Jungherr et al. 2017: 2).

In this sense, computational intelligence, digital traceability and Big Data defy social scientists to redefine the nature of knowledge and the descriptive power of Social Sciences (Burrows and Savage 2014).

Due to its digital nature digital data are radically different from typical data. This explains that, for example, "data scientists" are taking over statistician jobs as a new type of professional that claims expertise on social analysis. As it was demonstrated (Burrows and Savage 2014: 3), "Google Trends shows how in mid-2013 searches for 'data scientist' surpassed those for 'statistician' for the first time". Not only that, but also many professionals unrelated to Social Sciences are producing social knowledge based on the analysis of digital traceability.

All these changes influence the ways in which knowledge is being generated. A methodological and theoretical discussion has started; some suggest that a new way of making science is being created, one in which the *modus operandi* is purely inductive by nature (Kitchin 2017), and in which the traditional deductive method that infers observed facts based on general laws is replaced by the inductive method, in which laws are proposed from observed facts, in an emerging logic. Other researchers such as Anderson (2008) asserted that, in the era of petabyte information, the traditional, hypothesis-driven scientific method would become obsolete. Other authors point out the methodological differences concerning sample representativeness, "from the perspective of data scientists, representativeness is not an issue when one has access to all of the data (Schober et al. 2016).

Mazzocchi (2015) claims that the analysis of vast volumes of data will yield novel and often surprising correlations, patterns and rules. This unprecedented volume of available data requires a maximum inclusion analysis, without needing to focus on limited portions of data, concerns about randomization techniques and sample size are minimized, for size is no longer an issue. With the abundance of data, multiple aspects of the same problem can be investigated, rather than focusing on a random portion of it. Macrodata reduce measurement errors and better reflect the complexity of natural phenomena. Finally, macrodata require a strong emphasis on the correlations among data as a heuristic tool to find unexpected associations created by chance alone.

1.1 The demoscopic tension: Where and how to analyses the social in a digital era?

The emergence of the Web 2.0 and its evolution towards the Web 3.0 (Sheth and Thirunarayan 2012) offers people the possibility of expressing and sharing their opinions on a daily basis and through multiple applications and, since every device is connected to the web, the services can be used anywhere at any time. The evolution of social networks has contributed enormously to this proliferation of discursive flows, providing a global, accessible and instantaneous platform, like social media (SM), to share points of view on products, services and, of course, also current politics. Millions of people, in SM like Twitter or Facebook, give their opinions in continuous and voluntary flow on the most diverse topics, making their opinions, likes, attitudes, and preferences known. Beauchamp (2017: 490), for example, who proposes a method for extrapolating vote intention in states that are poorly polled, by extracting "from the social media data stream the textual features that are the best predictive of the polling data", points out that approximately 100 million tweets were produced on any given day during his collection period in the United States.

Besides being massive, all these online expressions that we can conceptualize as social media data (SMD) can be traced; they represent, in this sense, a real register of points of view of a large part of humanity regarding their most diverse interests.

Consequently, the scientific community is discussing new research methods to analyze public opinion and to improve our understanding of the social world, for example, through opinion mining (Ravi and Ravi 2015), and by comparing SMD with traditional survey data. An inevitable tension arises between new and traditional procedures and methods for social measurement and social analysis, that is what we call the *demoscopic tension*. Multiple attempts at using digital trace data as a sensor of off-line phenomena in various areas are being made worldwide, for example, to test the predictions power of social media analysis.

In this context, as a research group comprised by linguists, communication experts and engineers we wonder if taking into account the development of data science, it is possible, through the use of inductive algorithms and the automatic processing of messages with political opinion in social media, to generate electoral predictions in a context of three national elections that took place in Chile during 2017.

2 Social media analysis and survey data

Nowadays we witness the crisis of two traditional devices used in social research to carry out social analysis: media and opinion polls. During most of the twentieth century, common sense dictated that surveys were key to help us understand, guide, and coordinate the actions of individuals in complex environments (Lippmann 1922; McCombs and Shaw 1972; McQuail 1991; Thompson 1998). For decades, their efficacy has been relied upon to conduct macrosocial readings and to operate as a social thermometer. Polls, for example, would offer the possibility to understand what many authors define as public opinion (Habermas 1986; Price 1994). However, it is a fact that his predictions have not always been successful in recent years.

Moreover, members of the public are increasingly unwilling to participate in surveys (Schober et al. 2016). At the same time, while SM grows as channels used by people to express their opinion, so does the research of the potential of SMD for social analysis. Consequently, the importance of traditional survey data for exploring society is being challenged today by the research of social media data, generating a kind of demoscopic tension.

This dispute becomes even more intense if we consider that a traditional ally of public opinion polls, the media—which have been the main devices of diffusion of these polls—are also facing a difficult and complex situation. Media is another institution originated during the twentieth century and often used in social research as a historical and discursive document. Nowadays, media are facing its own crisis. The appearance of the Web 2.0 caused their loss of privileges: they are not, as they were during the whole of the twentieth century, the main source of mass information for people and the central nodes of information transmission (Dubois et al. 2018; Meraz 2011). We can observe an increasing migration of diverse activity towards digital platforms, e.g. the political one, especially during electoral campaigns (Issenberg 2012). The use of SM has become widespread during elections, becoming increasingly important for users, political parties and the campaign designs of the candidates, and for their everyday direct communication with their electors (Beauchamp 2017; Gulati and Williams 2013; Kreiss 2016; Lobo 2017). Consequently, social media analysis has become research object for social sciences. However, how trustworthy

such measurements are and knowing exactly if SMD is a sensor of off-line phenomena are still open questions being debated by the scientific community.

At present, predictive power of social media analysis is being tested in differents areas, also in the field of electoral forecasting, with the idea that the analysis of digital trace data can be useful to anticipate dynamics and predict trends. This demoscopic exploration of the digital environment implies methodological challenges. There are differences in how participants understand the activity they are engaged in; taking a survey is not the same as posting in SM. There are also differences between the collection of survey data based on intrusive, face-to-face methods, in which respondents offer a discursive and retrospective account of their actions and preferences to an unknown person, and the non-intrusive digital data collection methodology, which is continuous and does not alter the natural context in which data is produced, and in which digital traceability and the use of Big Data allow access to new repertoires and circuits. The nature of the data to be analyzed is also different (digital social media data versus analogic survey data). Schober et al. (2016) make a comprehensive comparison between survey data and SMD, comparing the differences in terms of ethics, collection, and analysis.

2.1 *Sampling criticism* and methodological challenges

Those who distrust the predictive capability of SM often focus of their arguments on *sampling issues*. It is argued that online users do not represent the population and that social media analysis is always a form of nonprobability sampling which may introduce new kinds of bias and measurement error. In other words, there is no equivalence between the demographic data of online users and the general population, particularly in relation to variables such as age, social condition and place of residence; therefore, the sampling of predictive studies based on the analysis of social media users is a priori biased and invalid.

This population representativeness problem is the core issue that often separates data scientists from the traditional statisticians. Social media analysis, unlike traditional surveys, always uses non-probabilistic samples and, in that sense, sample quality is challenged since they were not designed to represent the population. However, for data scientists, sample representativeness is less important if they can access all data. "Reaching for a random sample in the age of big data is like clutching at a horse whip in the era of the motor car" (Mayer-Schönberg and Cukier 2013: 31). Even though it is true that Internet users do not necessarily represent all populations, social media have shown a surprising potential for predicting electoral results in many real cases. Barberá (2015), for example, points out that Twitter data is widely acknowledged to hold great promise for the study of social and political behavior; Beauchamp (2017: 491) points out that "there exists something of a minor industry dedicated to measuring public opinion using social media", and Schober et al. (2016: 186) point out that "data-driven methods have the potential to uncover patterns that researchers have not pre-identified".

The predictive validity of social media analysis does not necessarily rely on how representative the users are of the general population (Ceron et al. 2014; McCormick et al. 2015; Jaidka et al. 2018). We find ourselves before a "new data collection paradigm", in which the challenge is knowing how to select the most appropriate methods for the new digital reality (McCormick et al. 2015). At the same time, it is also mentioned that also the traditional survey methods make electoral predictions with non-representative polls, making suitable statistical adjustments, post-stratifying responses to emulate a representative sample (Wang et al. 2015). The sampling critic that is used to question

the predictive potential of social media analysis bases its criticism on the no use of the classical poll patterns that employs statistical and demoscopic methods. This criticism disregards, however, the fact that current empiric evidence shows us the limitations of the traditional statistical methods for demoscopic purpose, even if those patterns are used and all sampling precautions are taken. As we mentioned before, in recent years, these predictions have been wrong repeatedly in multiple elections and different places all around the world. This sampling criticism that demands the use of traditional statistical procedures in digital demoscopy, does not consider that the number of samples, time intervals and spatial boundaries do not always match between the analog context and the digital environment (Aradau and Blanke 2016). Science is facing a new situation: the immense and unprecedented volume of data available for social analysis today. And as in many other fields of social life, there has been a transition from analog to digital, and the same changes should occur in the study of public opinion, were the use of SMD is being used increasingly.

Likewise, theoretical and methodological deficiencies of social media analysis are pointed out based on the proposed similarities between the political preferences shown by digital environment users and its equivalence in the political support of the candidates. In other words, the *equivalence hypothesis* (Aparaschivei 2011; Deltell et al. 2012; Tumasjan et al. 2010) that suggests that a high level of on-line interest could be indicative of political or electoral support is being questioned. This criticism is aimed at the conceptual basis that supports the equivalence between SMD (for instance, a *like*) and *off line* behaviors (for example, a vote), and points out that the link between SMD and political support is far from stable, in a manner that digital trace data cannot automatically be considered as a sensor of off-line phenomena (Jungherr et al. 2017). In this regard, several *on line* predictive studies are criticized based on their hypothetical interpretation of favorable opinions of a candidate in social networks as equivalent to a vote, based on a weak *equivalence hypothesis*. This criticism is no longer focused on demographic partiality criteria, but rather on a conceptual issue, that is, the equivalence between digital attitudes—for instance, positive comments or *on line search traffic* (Granka 2013)—and the effective electoral behavior of citizens in ballot boxes.

In this sense, Jungherr et al. (2017) note a recurring conceptual weakness of many investigations that analyze SMD for the study of public opinion, i.e., either to infer current levels of support towards political actors or to predict their support in upcoming elections. They note a "classic fallacy" (Jungherr et al. 2017: 2): using a quantitative indicator to draw inferences on a latent target concept, but instead measuring another concept. This often occurs, for example, when Twitter-based metrics are linked with metrics of political support in selected cases. Measuring the activity of Twitter users is an indicator of "digital attention", but not necessarily of "political support". In this sense, the equivalence hypothesis supports many investigations based on Twitter metrics that use confusing indicators and are supposed to measure political support, though they actually measure digital attention towards politics. As we will see further down, in our research we considered these conceptual precautions and created a methodological design that allows making inferences regarding political support based on Twitter metrics.

The context has changed, and in a radical manner, some already talk about the "algorithmic turn" (Klinger and Svensson 2018). Our challenge is knowing how to extract significant signals from these unconventional samples, i.e., knowing how to read them adequately for social analysis in the context of this new complexity. This tension requires methodologies for this new context.

3 Our research, our method, our electoral forecast

Due to the widespread and intensive use of SM, as well as the importance acquired by SM in electoral campaigns and in political communications, a growing body of research examining the relationship between SM and politics has appeared. In this regard, according to Ceron et al. (2017), seven main streams have emerged to analyze this relationship: (1) the political impact of the Web; (2) the agenda setting power of SM; (3) SM, collective action and public policy (digitally networked action); (4) the phenomenon of E-campaigning; (5) estimating of policy positions by analyzing social media data; (6) analysis of perceptions expressed on SM as a source of information to measure the attitudes of the on-line public opinion; and finally (7) the potential predictions power of social media analysis, under the idea that information available online can be useful to anticipate dynamics and forecast electoral results. Our research has been focused on the latter.

Any social media analysis that seeks to generate this kind of knowledge must take conceptual and methodological precautions.

1. Firstly, not adopting weak correspondence hypotheses that, as Jungherr et al. (2017) point out, make us use quantitative indicators like Twitter-based metrics to draw inferences on a target concept like "political support", although we are measuring another concept instead, for instance, political attention.
2. Secondly, not making false equivalencies between a post and a survey response.
3. Thirdly, Beauchamp (2017) mentions four requirements for using social media analysis for extrapolating vote intention:

    (a) Statistical testing
    (b) Benchmarks
    (c) Training
    (d) Out of sample

Our research generated a validation strategy that took into consideration the points mentioned above. Likewise, it is worth noting that the methodological approaches most used to explore the potential prediction power SM are three: the Volumetric Analysis (AV), Social Network Analysis (Net) and Sentiment Analysis (SA). The first is focused on the volume and frequency of mentions and/or followers (for instance, for candidates and political parties). Meanwhile, Net is focused on estimating the strength of an on-line community that supports a candidate or political party, identifying central positions (nodes), extent and expansion potential in user networks (Jaidka et al. 2018). Finally, the Sentiment Analysis is focused on exploring positive and negative sentiments, as well as the emotions expressed by users towards the candidates. We can distinguish mainly two types of classification of emotions (Turner 2000), binary classification (coarse-grained classification of sentiment polarity) and multi-class classification (fine-grained classification multiple classes).

In 2017, six scholars formed a research group called DeepPUCV (https://deeppucv.cl/), comprised by linguists, communication experts and engineers. Our main objective was to predict the result of presidential elections through the use of inductive algorithms and the automatic processing of messages with political opinion in the social network Twitter, using SA. During 2017, three national elections took place in Chile: primary elections to choose presidential candidates from two political coalitions (July 2), the first round of the presidential election (November 19) and the second round (December 17). It was an

intense electoral year and, therefore, a period of intense use of both traditional media and SM among Chilean users and presidential candidates to talk about politics.

In this context, using SA, our research was aimed at exploring Twitter's predictive capability, the most social network used during presidential campaigns to express electoral and political preferences (Ceron, et al. 2014; Jaidka et al. 2018; Kreiss 2016). For this, we carried out a collection of all messages and metadata of Chilean users that mentioned any of the candidates at least once during the campaign period. For this purpose, through a social media monitoring company we automatically extracted the tweets of all these users.

In this way, we collected 9,367,127 tweets, from 372,665 users, through three search criteria:

(a) Mention of a candidate's account;
(b) Mention of the name of the candidate;
(c) Mention of a hashtag related to an event of the candidates.

After the collection and storage procedure, the tweet classification process was carried following the SA criterion that categorizes messages as positive/negative/neutral according to what is expressed regarding the candidates (Aparaschivei 2011; Deltell et al. 2012). In a first stage, this classification process was carried out manually by 6 professionals that tagged 640,224 tweets in the categories mentioned above. The human cataloging work was the basis for training the supervised learning algorithms, i.e., for the machines to correctly define the positive/negative/neutral value of a new message; this stage, called machine learning, is necessary to develop predictive models (see Figs. 1, 2).

Aware of the criticisms made on the equivalence hypothesis, which questions, for example, the lienality that matches one positive mention with one vote, or holding false equivalencies between a post and a survey response we guided the human classification in a conceptually purposeful manner (characteristic of inductive and non-randomized methods). Thus, the professionals who carried out the manual tagging of data only classified tweets generated by Chilean users during the broadcasting of live media events (radio or television) of high political-electoral importance.

This way, we designed our research with a very specific selection of our analysis unit: tweets generated by users during the television or radio broadcasting of debate programs. We distinguish two types of media events:

(a) presidential debates broadcasted live by Chilean radios and television;

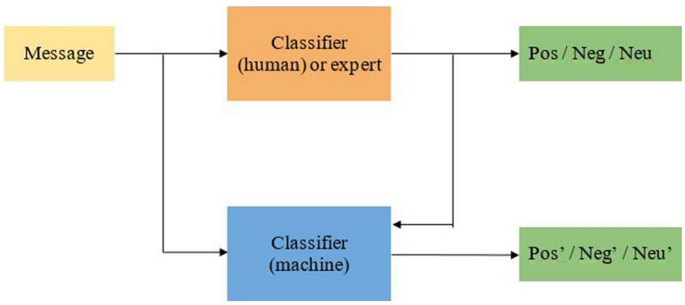

Fig. 1 Supervised learning. Own elaboration

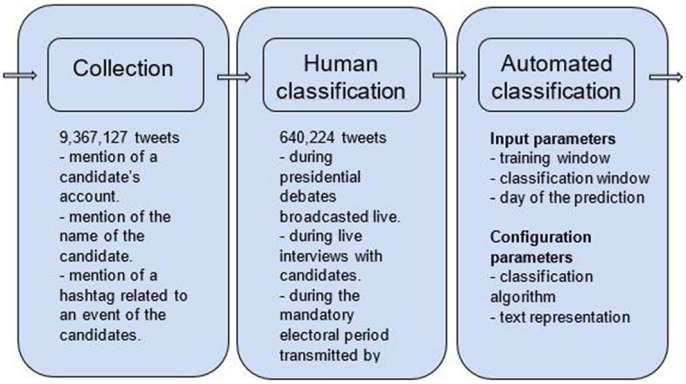

Fig. 2 External parameters for the second round of the presidential election. Own elaboration

(b) live interviews with candidates in the most important political TV shows in the country;

In other words, although the tweet collection process was continuous during 2017 (over 9 million tweets), the human classification of these tweets in positive/negative/neutral was focused only on the tweets generated during media events, with the collection starting 2 h before the media event and finishing 2 h after the program. In total, in 2017, there were three radio debates, four television debates and four political programs with individual candidate interviews, instances in which humans classified 640,224 tweets.

Multiple conceptual criteria guided us to take the decision of carrying out human classification only with the tweets collected during these media events, and not any tweet. From a conceptual standpoint, we wanted to develop a strong correspondence hypothesis that allowed us to claim that the opinions expressed by online users regarding one candidate resemble their political support. We are assuming that the people that watch these political shows do so voluntarily, they choose to watch them instead of many other programs offered by other channels at the same time, and then voluntarily choose to express their political opinions regarding what is said in those shows in the social networks. In our opinion, when we analyze the tweets generated by users during the occurrence of live political programs in which candidates debate, the equivalence hypothesis is also strengthened, even more so in countries where, as in Chile, voting is voluntary, since these are moments when thousands of users voluntarily express their opinions and preferences regarding the candidates, generating a national public debate in the digital context. Although it is true that the socio-economic characteristics of those who use social media do not exactly match the demographic characteristics of populations, this difference is reduced, and, therefore, sampling similarity increases when the sample includes only those who express voluntary their political opinion on the Web (Ceron et al. 2014), in countries where voting is voluntary. Thus, we compare three voluntary actions: watching the political program, expressing political opinions about it in the social networks, and voting. The volumetric weight analyzed should also be considered, namely, millions of tweets by hundreds of thousands of users, as shown in Fig. 2.

We took methodological and conceptual precautions to avoid the error observed by Jungherr et al. (2017), that is "using quantitative indicators and draw inferences on a target concept like *political support*, but instead measuring another concept, for instance,

political attention". With our qualifying technique, we propose a methodology in which we match digital attention (Zhang et al. 2010), political attention and political support.

It is known that there is an important connection between television and social media (Deltell et al. 2012; Gallego 2013; Lahey 2016), between tweeting and television ratings (Fábrega and Paredes 2012; Wang et al. 2015), and between the audience and Twitter users that has created the phenomenon of double or multiple screens. Tweeter is a medium that catalyzes audience discussions and interactions about television; topics related to television frequently appear in Twitter's trending topics, this is the case especially for live television (Deltell et al. 2012; Harrington et al. 2013). In spite of the discredit of politics, presidential debates still attract the attention of citizens, who actively comment and discuss regarding the candidates and generate a parallel debate in the digital context that is triggered by the live debate in traditional media (Benoit and Sheafer 2006; McKinney and Carlin 2004; Ruiz del Olmo and Bustos 2017). This "double debate" occurs in a context of "multi screens", in which the audiences of traditional media, such as radio and television, are also users of digital platforms in which they interact with other people (Claes and Deltell 2015; Gallego 2013; Harrington et al. 2013; Wang 2016). Thus, the experience of consuming television can become a social exercise in which the audience virtually exchange opinions and ideas with strangers—supporters and opposition—candidates and their teams, as well as journalists and the producers of television shows, etc. In this sense, a social media like Twitter can become a thermometer that minute by minute provides information and metrics regarding the reactions, interactions and opinions of viewers (Harrington et al. 2013; Lahey 2016).

All these factors were considered by our team in order to take the classification decision mentioned above, strengthening, in this sense, data quality regarding the analytical objective.

3.1 Automated classification

Concerning the classification procedure for estimating vote intention, we took into account the four requirements proposed by Beauchamp (2017): Statistical testing, benchmarks, training and out of sample.

Given the nature of our data and analysis, statistical tests were used to compare the performance of our models, establishing statistical differences among their different predictions. In other words, given our methodology and its application, we made no statistical comparisons between real and predicted data, but between different data predicted by our algorithms.

In terms of benchmarks, our points of comparison were the real results of the three elections, provided by the Electoral Service of Chile. This way, we measured the success of our model against the official and real electoral results.

Regarding the Training set, again, considering the nature of our investigation, we trained our models with data categorized by human experts during political debate events, in a way that allowed us to relate comments with the intention to vote for one candidate or the other. This way, we have addressed the so-called "lack of representativeness of Twitter users". Subsequently, we have trained our models with a set of categorized messages, which allows us to learn a function of voting intention categorization based on sets of examples.

Finally, in terms of *out of sample*, we tested the performance of the model with unclassified data but based on the previous training that was carried out with human data categorization, which allowed us to generate learning and imitation algorithms. In

this way, we applied a classification model with which we generated electoral predictions a few days in advance. This prediction was determined in a prediction window called "Testing Window", which was designed by us as an input parameter of the model (Figs. 2, 3). Thus, the testing set measures the performance of the model and the prediction window predicts the election, extrapolating the vote intention.

Regarding automated classification, we define three "input parameters" and two "configuration parameters" to control the supervised and automated processing of data in the best way possible. We understood these parameters as independent variables that were refined as our research progressed. In other words, we adjusted the prediction variance according to five independent variables, three external and two internal, namely:

$$P = f(a, b, c, d, e),$$

where a: Training window, b: prediction testing window, c: prediction day, d: Representation, e: Classification algorithm.

Input parameters were also called "external parameters", since they are variables used before automated data processing, in other words, are outside the algorithmic procedure yet condition it. We considered three external variables, also called *input parameters* (see Fig. 2).

Input parameters (external):

(a) Training window

The training window is defined by the set of manually classified messages that allows us to adjust the internal parameters of the learning algorithm. In our case, as we explained before, humans classified these messages during relevant media events. These tweets were used to train the algorithm with the purpose of enabling a prediction based on this learning.

(b) Prediction Testing window

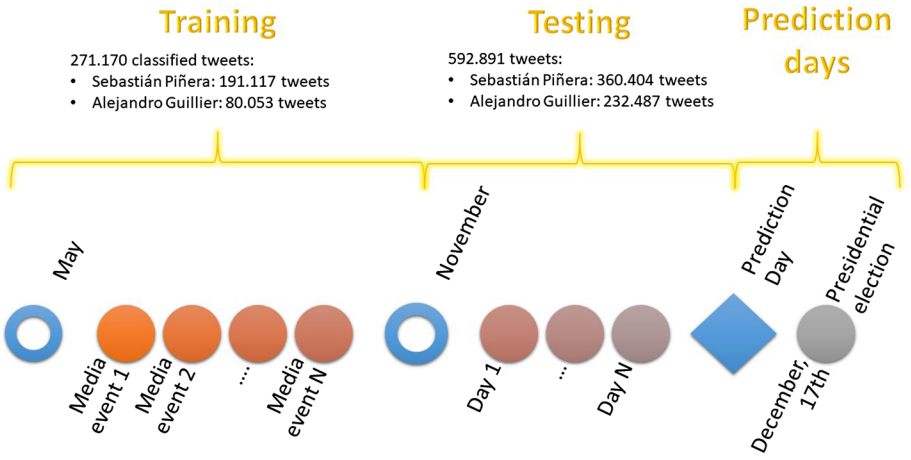

Fig. 3 Stages and variables of the classification process. Own elaboration

This is understood as a specific set of days in which data collected in these days are classified based on the previous training. These are used to make predictions, meaning, this is where we define how many days of "history of opinions" are required to make predictions. After testing different time windows during our research, the 10 days window provided the best performance, with predictions closer to the real values and a smaller margin of error. It is worth mentioning that the size of the classification window is particularly important when some events that could change the citizens' voting intention occur a few days before voting.

(c) Day of the prediction

This third external parameter is related to the decision of which day should the prediction be made. It implies, therefore, deciding how many days pass between the end of the classification window (which used to be 10 days) and the election day. In our case, this parameter usually ranged from the election day to 3 days before it.

In case of the presidential second round, with the two candidates (Piñera and Guillier), 271,170 tweets that mentioned them were manually classified. This learning base, the automated classification (Prediction Testing) was carried out from November onwards, and for that, 592,891 tweets will be analyzed.

Configuration (internal) parameters:

On the other hand, configuration (or internal) parameters are those related to automated learning; therefore, they are related to the selection of textual representations for classifier tasks, such as the very determination of classifiers. We considered two internal variables (see Fig. 3).

(d) Representation (*n*-gram)

The automatic classification of natural language has been studied for decades. For that purpose, a series of textual representation methods have been developed, which are usually based in term weighing, including, for example, the Boolean model, Tf-idf and its variants. Overall, to weigh terms in the vector space model, the term frequency or the frequency of documents that contain a term can be used. Some elements used to represent a text are *n*-grams, the ones that we used. A *n*-gram is a subsequence of "n" elements of a given sequence, a word, phrases, logical terms and statement or any other semantic and/or syntactic unit that can be used to describe text content (Schütze et al. 2008).

(e) Classification algorithm

The second configuration variable is related to the automated classification of tweets using algorithms that receive the messages and its class defined by a human, meaning, the representation of the message and voting intention that a human assigned to that message. With it, the algorithm is trained to separate, identify and relate the messages to the voting intention in the best way possible. It is expected for the algorithm to classify new messages with the same competence than during its training (first external variable). In our case, we used the following algorithms: (1) Decision tree, (2) Random Forest, (3) Adaboost and (4) Linear Support Vector Machines (LSVM).

Figure 3 represents these procedural stages and their considerations within a diagram.

The manipulation and adjustment of external and internal parameters were aimed at decreasing the mean absolute error (MAE) as much as possible and, thus, improving the predictive capability of the classifiers. The margin of error that occurs with the prediction is called mean absolute error (MAE), a metric used to evaluate the performance in the predictions and that accounts for the average difference between the estimates obtained by the classifiers used and the actual results (Willmott and Matsuura 2005). The closer the MAE is to 0, the closer to the actual result of the elections. This metric is calculated with the following formula:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}.$$

In it, x_i corresponds to the real value of an i event, y_i corresponds to the value of the estimate made for an i event, and finally n corresponds to the number of existing events to be estimated. A MAE below five qualifies as acceptable, as occurs in the metrics of traditional surveys (Bermingham and Smeaton 2011).

During our research, our MAE improved progressively in each election, due to the processing of great data volumes and different algorithms that allowed us to improve our capacity to find relevant patterns; thus, we refined the automated mode of processing data during several months. In each of the three elections, we learned from our mistakes and perfected our methodological procedures in connection with variable management and adjustment. Santander et al. (2017) and Rodríguez et al. (2018) have published part of the results.

4 DEEP PUCV forecast

As we mentioned before, there were three important elections in our country during 2017: primary elections, the first round of the presidential election (November 19) and the second round (December 17). In the case of this investigation, we will show mainly the results of the two presidential elections.

The first was on July 2; it was the primary election, voluntary and legal, in which the left wing alliance Frente Amplio (two candidates) and right wing alliance Chile Vamos (three candidates) chose their presidential candidates, respectively. Due to problems and internal conflicts, the ruling party did not participate (political center). A total of 1,811,411 people went to the polls that day. Our prediction (Fig. 4) was far from forecast the real votes for every candidate (Fig. 5); but we went from our initial pessimism to a cautious optimism after seeing that we succeeded at predict the order of preference of the candidates.

This election was our first test and allowed us to analyze our mistakes and successes. Regarding the configuration parameters, for instance, we could identify the two algorithms that best suited the prediction and that produced the lowest MAE. Regarding the external parameters, we identified the best windows and days for prediction, both for the Frente Amplio (Fig. 6), and the right Chile Vamos (Fig. 7).

What we learned from this first test was used to face the first presidential round (November 19) in which 8 candidates participated. The favorites, according to all polls, were three: Sebastián Piñera (right wing; "Chile Vamos"); Beatriz Sánchez (left wing, "Frente Amplio"), and Alejandro Guillier (ruling party candidate, political center; "Nueva Mayoría"). The latter, as we explained before, did not participate in the primary elections, which, as we will see, had consequences for our learning and the prediction we made (Fig. 8).

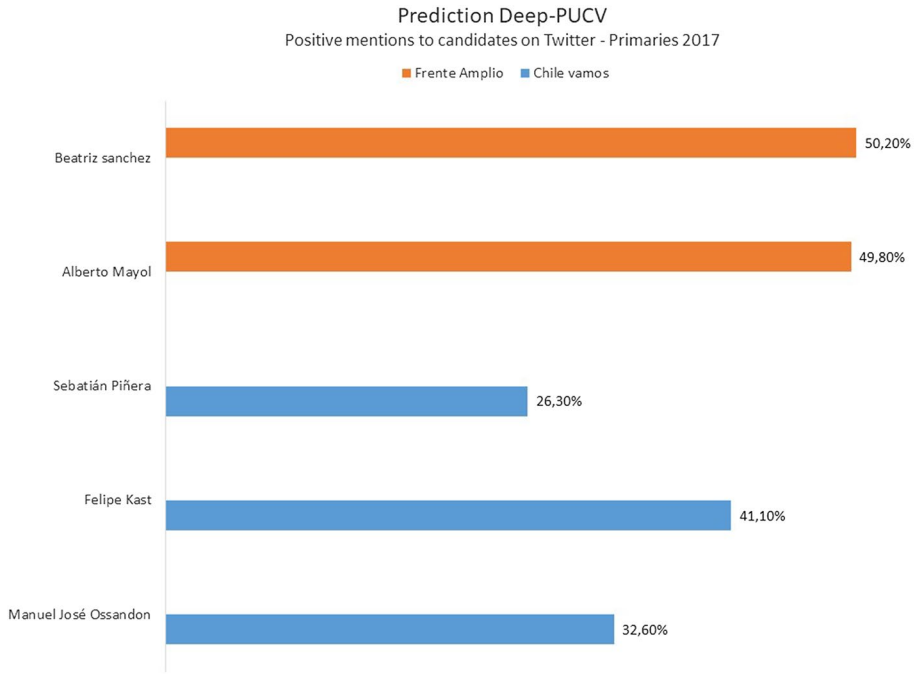

Fig. 4 Electoral prediction made by deep-PUCV, primary elections 2017, Chile. Own elaboration

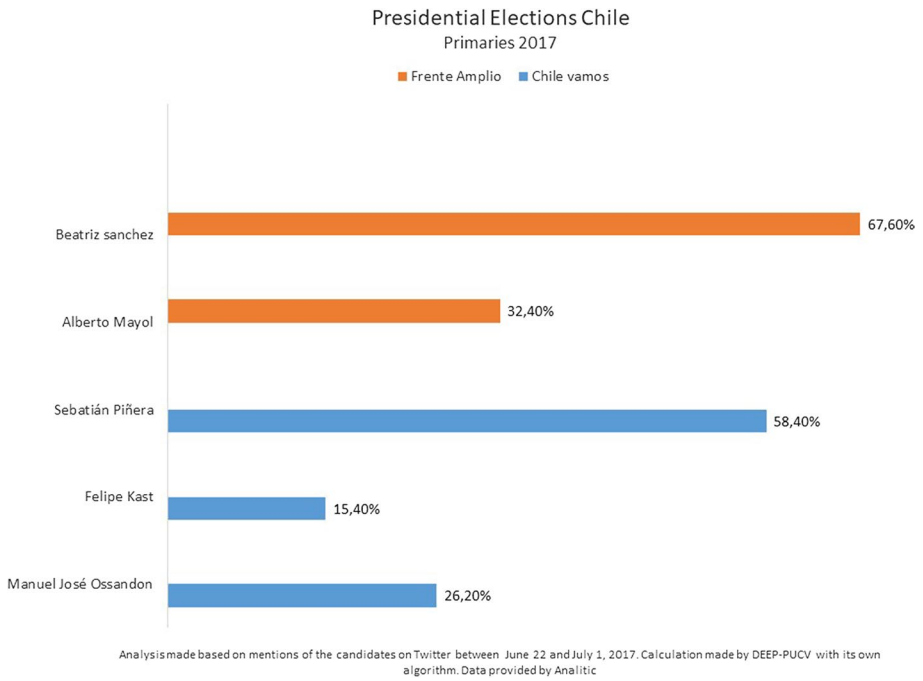

Fig. 5 Official electoral results; primary elections 2017, Chile; own elaboration. *Source*: Chilean electoral service

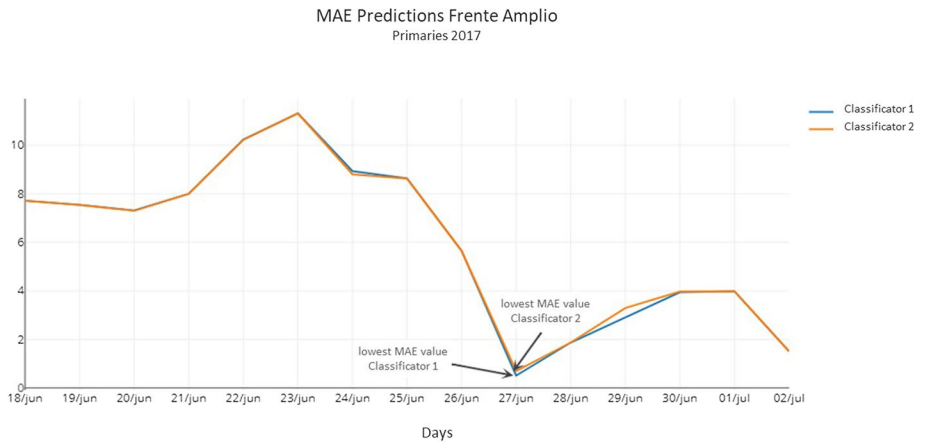

Fig. 6 Lowest MAE for Frente Amplio. *Source*: Own elaboration

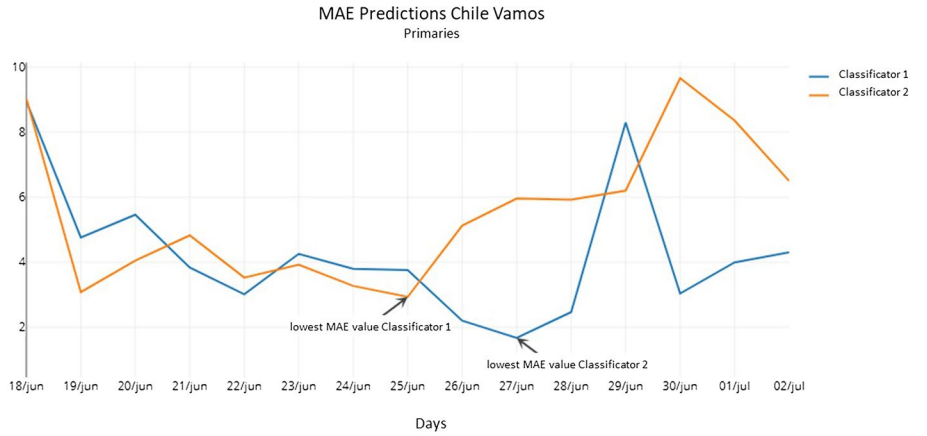

Fig. 7 Lowest MAE for Chile Vamos. *Source*: Own elaboration

As we see in the comparison between both graphs, our figures showed a very low MAE for two of the three strongest candidates: Sebastián Piñera (32.2% versus 36.6%) and Beatriz Sánchez (19.6% versus 20.3%). However, with the other candidate, Alejandro Guillier, our prediction was completely wrong according to the official results, 2.3% vs. 22.7% (Fig. 9).

Upon analyzing this tremendous deviation, we realized that the cause of the problem was data recollection, since, unlike the surnames of the other two strong candidates (Piñera and Sánchez), the French surname "Guillier" is very rare in Chile, and most users wrote it wrong; in fact, it was misspelled even in the headquarters of his campaign, as we checked. Therefore, our error was to algorithmically collect mentions with the right spelling of the surname only, since it significantly undermined the sample and the posterior analysis, since Guillier, along with Piñera, continued to the second round. However, realizing this error in the first round allowed us to collect mentions adequately, in order to make our predictions for the second round that took place a month after, on December 17 (see Figs. 10, 11), in which Sebastián

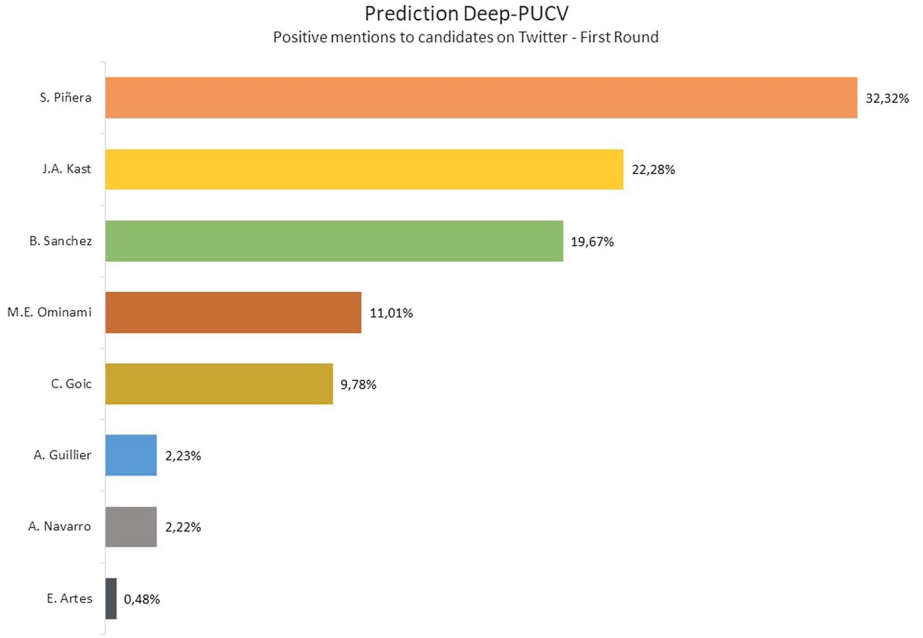

Fig. 8 Deep-PUCV prediction for the first electoral round. *Source*: Own elaboration

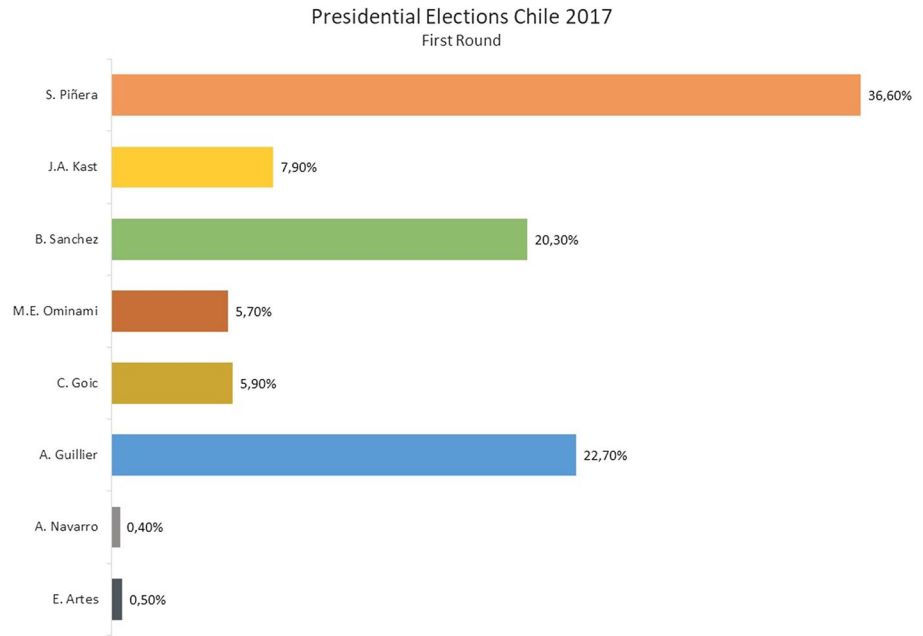

Fig. 9 Official results for the first electoral round. *Source*: Chilean electoral service

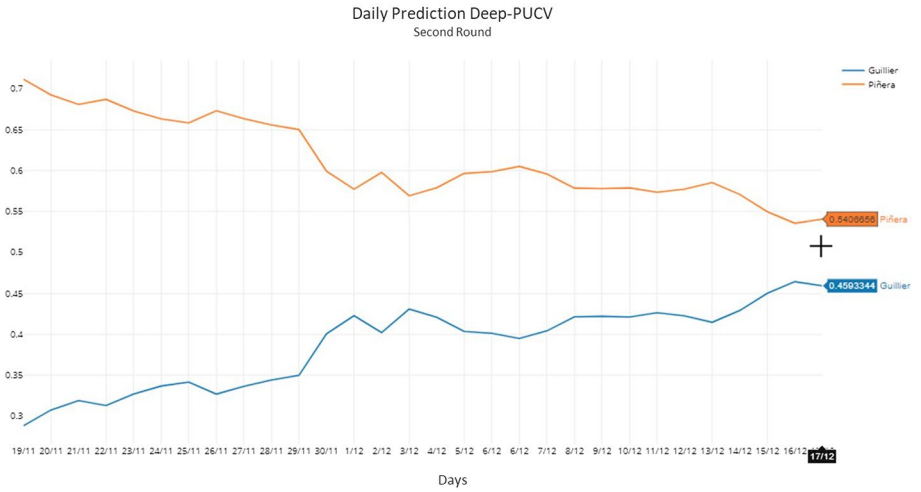

Fig. 10 Daily prediction by deep-PUCV, second round of the presidential election, Chile. *Source*: Own elaboration

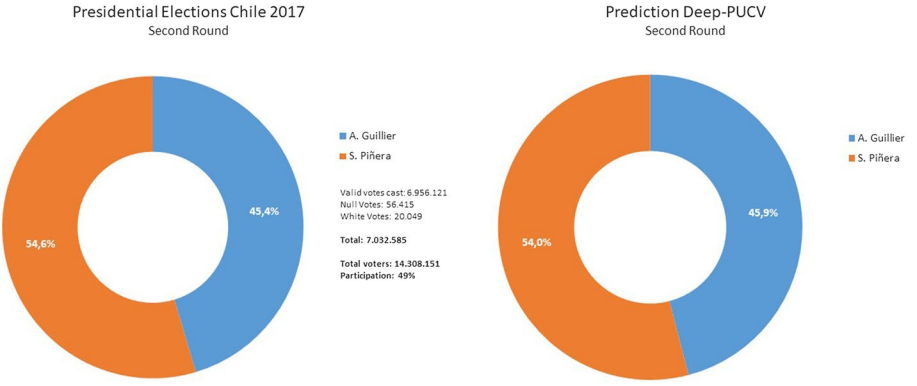

Fig. 11 Comparison: Deep-PUCV prediction and official results for the second round of the presidential election, Chile; own elaboration. *Source*: Deep-PUCV and Chilean electoral service

Piñera and Alejandro Guillier ran for presidency. This time, the MAE related to the real results was minimum.

As we can see, this time our results were practically identical to the official electoral results, disclosed by the Electoral Service of Chile. This was the third time that we tried to predict the results of an election. Our learning based on trial and error in the primary election and the first presidential round was fundamental to improve our procedures and to refine the independent variables and, thus, improve the prediction.

5 Conclusions

As it could be observed in the preceding pages, our knowledge production mode was methodologically inductive and non-randomized. It was a data-driven research where we used inductive algorithms, i.e., algorithms which refine their way of processing not structured data according to the most appropriate use that can be made. During the three electoral moments we operated under the trial and error logic, essential to the inductive research logic, in which knowledge emerges as the research progresses. We proved the importance of external and internal variables, their progressive adjustment and the resulting reduction of MAE. Through computational intelligence, we were able to follow, collect and analyze the presidential campaign by processing over 9 million tweets, that represented all the universe of messages writes naming the terms examinates. This allowed us to possess a great volume of data, and meant that we collected natural and not experimentally produced data, thus avoiding the bias of the so-called social desirability, very typical in cases in which it is necessary to know the political opinion of the respondents.

We achieve our objective, this was to predict the result of presidential elections through the use of inductive algorithms and the automatic processing of messages with political opinion in the social network Twitter, using Sentiment Analysis approach.

It is clear that a great volume of data does not generate knowledge by itself; being aware of this, we did not carry out a randomized collection of data, but a data collection process guided by precise conceptual, methodological, and instrumental frameworks. In turn, we strengthened the equivalence hypothesis with conceptual density. This was carried out through the human classification of messages produced during media events, which are moments characterized by triggering an intense activity in SM, generating a massive and voluntary political debate of users that follow the debates of presidential candidates.

To this we must add the following hypothesis that emerges as a result of this research: in a country with voluntary vote as Chile, Twitter users are very similar to those who actually go voluntarily to the polls on the day of the elections. It is likely that in a country with mandatory voting (for example, Argentina or Brazil) the similarity may be lower.

These and other considerations should continue to be examined since what is clear is that the impact and use of SM in people's daily lives is increasing, therefore, there are increasingly more opinions of citizens about the most various topics in the networks. At the same time, the algorithms are being refined more and more, which suggests that in the future many of them will be used to measure the moods of the users in relation to different topics, for example, popularity of candidates, evaluation of policies public, adherence to brands, etc. This could redefine participation in public policies and democracy.

As a future challenge, it is still pending to conceptually understand the performance differences of the classifiers, as well as verifying the scalability and applicability of the model in other political-electoral contexts in order to develop a predictive methodology that may be used in other elections and even other kinds of contexts, such as opinions on brands or public policies. For that, we could explore transfer learning and other approaches from machine learning.

References

Anderson, Ch.: The end of theory: the data deluge makes the scientific method obsolete. The Wired, June 23 (2008)

Aparaschivei, P.A.: The use of new media in electoral campaigns: analysis on the use of blogs, Facebook, Twitter and YouTube in the 2009 Romanian presidential campaign. J. Media Res. **2**(10), 39–60 (2011)

Aradau, C., Blanke, T.: Politics of prediction: security and the time/space of governmentality in the age of big data. Eur. J. Soc. Theory **20**(3), 373–391 (2016)

Barberá, P.: Less is more? How demographic sample weights can improve public opinion estimates based on Twitter data. Center for Data Science New York University (2015). http://pablobarbera.com/static/lessis-more.pdf

Beauchamp, N.: Predicting and interpolating state- level polls using Twitter textual data. Am. J. Polit. Sci. **61**(2), 490–503 (2017)

Benoit, W., Sheafer, T.: Functional theory and political discourse: televised debates in Israel and the United States. Journal. Mass Commun. Q. **83**(2), 281–297 (2006)

Bermingham, A., Smeaton, A.: On using Twitter to monitor political sentiment and predict election results. In: Proceedings of the Workshop on Sentiment Analysis where AI Meets Psychology (SAAIP 2011). Chiang Mai, Thailand: Workshop at the International Joint Conference for Natural Language Processing, pp. 2–10 (2011). Retrieved from http://doras.dcu.ie/16670/

Burrows, R., Savage, M.: After the crisis? Big data and the methodological challenges of empirical sociology. Big Data Soc. (2014). https://doi.org/10.1177/2053951714540280

Ceron, A., Curini, L., Iacus, S.: Politics and Big Data. Routledge, London (2017)

Ceron, A., Curini, L., Iacus, S.M., Porro, G.: Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. New Media Soc. **16**(2), 340–358 (2014). https://doi.org/10.1177/1461444813480466

Claes, F., Deltell, L.: Audiencia social en Twitter: hacia un nuevo modelo de consumo televisivo. Trípodos **36**, 111–132 (2015)

Deltell, F., Claes, F., Osteso, J.: Predicción de tendencia política por Twitter, Elecciones Andaluzas 2012. Ambitos: Revista Internacional de Comunicación **22**(1), 91–100 (2012)

Dubois, E., Gruzd, A., Jacobson, J.: Journalists' use of social media to infer public opinion: the citizens' perspective. Soc. Sci. Comput. Rev. (2018). https://doi.org/10.1177/0894439318791527

Fábrega, J., Paredes, P.: La política chilena en 140 caracteres [Chilean Politics in 140 Characters]. In: Arriagada, A., Navia, P. (eds.) Intermedios Medios de comunicación y democracia en Chile [Intermedia. Media and democracy in Chile]. Universidad Diego Portales, Santiago (2012)

Gallego, F.: Social TV analytics: nuevas métricas para una nueva forma de ver televisión. index.comunicación **3**(1), 13–39 (2013)

Granka, L.: Using online search traffic to predict US presidential elections. PS Polit. Sci. Polit. **46**(2), 271–279 (2013)

Gulati, G.J., Williams, C.B.: Social media and campaign 2012: developments and trends for Facebook adoption. Soc. Sci. Comput. Rev. **31**(5), 577–588 (2013). https://doi.org/10.1177/0894439313489258

Habermas, J.: L'espace public. Archéologie de la publicité comme dimension constitutive de la société bourgeoise. Payot, Paris (1986)

Harrington, S., Highfield, T., Bruns, A.: More than a backchannel: Twitter and television. J. Audience Recept. Stud. **10**(1), 405–409 (2013)

Issenberg, S.: The victory lab: the secret science of winning campaigns. Public Opin. Q. **78**(1), 363–364 (2012). https://doi.org/10.1093/poq/nft048

Jaidka, K., Ahmed, S., Skoric, M., Hilbert, M.: Predicting elections from social media: a three-country, three-method comparative study. Asian J. Commun. (2018). https://doi.org/10.1080/01292986.2018.1453849

Jungherr, A., Schoen, H., Posegga, O., Jürgens, P.: Digital trace data in the study of public opinion: an indicator of attention toward politics rather than political support. Soc. Sci. Comput. Rev. **35**(3), 336–356 (2017)

Kitchin, R.: Thinking critically about and researching algorithms. Inf. Commun. Soc. **20**(1), 14–29 (2017)

Klinger, U., Svensson, J.: The end of media logics? On algorithms and agency. New Media Soc. **10**(12), 4653–4670 (2018). https://doi.org/10.1177/1461444818779750

Kreiss, D.: Seizing the moment: the presidential campaigns' use of Twitter during the 2012 electoral cycle. New Media Soc. **18**(8), 1473–1490 (2016). https://doi.org/10.1177/1461444814562445

Lahey, M.: Everyday life as a text: soft control, television, and Twitter. SAGE Open (2016). https://doi.org/10.1177/2158244016633738

Lippmann, W.: Public Opinion. Hartcourt Brace, New York (1922)

Lobo, S.: Cómo influyen las redes sociales en las elecciones. Nueva Sociedad (269) (2017). http://nuso.org/autor/sascha-lobo/. Accessed 24 Mar 2019
Mayer-Schönberg, V., Cukier, K.: Big Data: A Revolution That Will Transform How We Live, Work, and Think (2013)
Mazzocchi, F.: Could big data be the end of theory in science? A few remarks on the epistemology of data-driven science. EMBO Rep. **16**(10), 1250–1255 (2015). https://doi.org/10.15252/embr.201541001
McCombs, M.E., Shaw, D.L.: The agenda-setting function of mass media. Public Opin. Q. **36**(2), 176–187 (1972). https://doi.org/10.1086/267990
McCormick, T.H., Lee, H., Cesare, N., Shojaie, A., Spiro, E.S.: Using Twitter for demographic and social science research: tools for data collection and processing. Sociol. Methods Res. **46**(3), 390–421 (2015)
McKinney, M., Carlin, D.: Political campaign debates, en Lynda Kaid [comp.], Handbook of Political Communication Research. Lawrence Erlbaum Associates, USA (2004)
McQuail, D.: Introducción a la teoría de la comunicación de masas [Introduction to the mass communication theory]. Paidós, Barcelona (1991)
Meraz, S.: The fight for 'how to think': traditional media, social networks, and issue interpretation. Journalism **12**(1), 107–127 (2011)
Price, V.: La Opinión Pública [Public opinion]. Gedisa, Madrid (1994)
Ravi, K., Ravi, V.: A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowl. Based Syst. **89**, 14–46 (2015). https://doi.org/10.1016/j.knosys.2015.06.015
Rodríguez, S., Allende-Cid, H., Palma, W., Alfaro, R., González, C., Elortegui, C., Santander, P.: Forecasting the Chilean electoral year: using Twitter to predict the presidential elections of 2017. In: International Conference on Social Computing and Social Media, pp. 298–314 (2018)
Ruiz del Olmo, F., Bustos, J.: La evolución del debate televisivo como herramienta de comunicación política. Informação Sociedade: Estudos **27**(2), 235–252 (2017)
Santander, P., Elórtegui, C., González, C., Allende, H., Palma, W.: Social networks, computational intelligence and electoral prediction: the case of the presidential primaries of Chile 2017. Cuad. Info **41**, 41–56 (2017). https://doi.org/10.7764/cdi.41.1218
Schütze, H., Manning, C.D., Raghavan, P.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
Schober, M., Pasek, J., Guggenheim, L., Lampe, C., Conrad, F.: Social media analyses for social measurement. Public Opin. Q. **80**(1), 180–211 (2016)
Schwab, K.: La cuarta revolución industrial. Debate, España (2016)
Sheth, A., Thirunarayan, K.: Semantics Empowered Web 3.0: Managing Enterprise, Social, Sensor, and Cloud-based Data and Services for Advanced Applications (Synthesis Lectures on Data Management), 1st edn. Morgan & Claypool Publishers, San Rafael (2012)
Sloan, L., Morgan, J., Burnap, P., Williams, M.: Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. PLoS ONE **10**(3), e0115545 (2015). https://doi.org/10.1371/journal.pone.0115545
Thompson, J.: Los media y la modernidad: una teoría de los medios de comunicación. Paidós, Barcelona (1998)
Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with twitter: what 140 characters reveal about political sentiment. In: International AAAI Conference on Weblogs and Social Media, vol. 10, no. 1, pp. 178–185 (2010). http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1441
Turner, J.: On the Origins of Human Emotions: A Sociological Inquiry into the Evolution of Human Affect. Stanford University Press, Stanford (2000)
Wang, Y.: How do televisión networks use Twitter? Exploring the relationship between Twitter use and televisión ratings. South. Commun. J. **81**(3), 125–135 (2016)
Wang, W., Rothschild, D., Goel, S., Gelman, A.: Forecasting elections with non-representative polls. Int. J. Forecast. **31**(3), 980–991 (2015)
Willmott, C.J., Matsuura, K.: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. **30**, 79–82 (2005)
Zhang, Y., Jin, R., Zhou, Z.-H.: Understanding bag-of-words model: a statistical framework. Int. J. Mach. Learn. Cyber. 1(1), 43–52 (2010)

Affiliations

Pedro Santander[1] · Rodrigo Alfaro[1] · Héctor Allende-Cid[1] · Claudio Elórtegui[1] · Cristian González[1]

Rodrigo Alfaro
rodrigo.alfaro@pucv.cl

Héctor Allende-Cid
hector.allende@pucv.cl

Claudio Elórtegui
claudio.elortegui@pucv.cl

Cristian González
cristian.gonzalez@pucv.cl

[1] Pontificia Universidad Católica de Valparaíso, Valparaiso, Chile