Using the Twitter social network as a predictor in the political decision

Jorge Arroba Rimassa[1], , Fernando Llopis[2,] , Rafael Muñoz Guillena[2], Yoan Gutierrez

1 Facultad de Ingeniería Ciencias Físicas y Matemáticas, Universidad Central del Ecuador, Quito, Ecuador

2 Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Alicante, España

jarroba@uce.edu.ec, {llopis, rafael, yoan}@dlsi.ua.es

Abstract: The use of social networks as a predictive tool to predict the outcome of an election can generate controversy; however if you have a methodology that tries to equate the extracted data as if they were obtained through a conventional survey, that is to say using weighting factors more than what usually should be done, polarity and relevance of each tweet, can make it a very reliable technique in light of the results obtained; the proposed methodology was applied in the presidential election of Ecuador on February 19th, 2017.

Keywords: Twitter, polarity, relevance, weighting factors, Ecuador.

1 Introduction

All the footprint that a man leaves in digital media defines the ideotype of each one.

As a user, as a consumer of what contributes or not to the web. Messaging: through whatsapp, emails; visits and messaging in social networks, the likes; google queries, searches for a certain type of information, ranging from academics, politics, tourism, music, *do it yourself* and other interaction in the virtual world; they are defining people.

The time a man spends and is exposed via computer, TV, entering ATMs, and other interactions with the digital world are also forming the profile of people.

If a detailed analysis of a person's entire fingerprint could be made, one could determine their preferences, their tastes, their friendships, their concerns and, in general, we could predict their behaviors in the face of any future situation.

This imaginary, this kind of Big Brother, still, it is not possible to implement. However, it is possible to partially monitor the transit of a person in the digital world. For example, what are the inquiries of a person in Google; what a person does in social networks; and other follow-ups, which are often constrained by law, by the issue of the privacy of the people.

However, in this privacy, one can analyze one of the tools that today have allowed people of all conditions to "converse" freely. They are Social Networks.

Social Networks have become a mechanism for people to think and interact in a "free" way in front of any THEME given by the same or by an ACTOR.

A thematic is a theoretical construct on aspects that go from the academic to the perceptual. Politics is a topic for example that goes from the academic consideration on the "politics" to the denunciation of acts of corruption. There is partisan, electoral, fiscal, educational policy, etc.

An actor, is an opinion leader or is an official organic entity that presents its position in relation to a particular topic. For example, a government group wants to know the opinion on an action in fiscal policy or in constitutional amendments; a message is sent about such action to be able to know the degree of acceptance that would be had. This is sent through the internet by users to Twitter for example.

As a case study, to evaluate the methodology proposed in the present research, we will carry out the political analysis in the Republic of Ecuador through the Social Networks, specifically Twitter, and be able to determine the messages that the actors issue on the different topics, of the task political, which have greater impact on users, in order to predict the outcome of the election on February 19, 2017.

To the extent that you want to use Twitter messages as a predictation tool, you must perform a conversion process so that the opinion given in the tweets is equivalent to those of the general population, since Twitter users are a subset of these.

The year 2016 can be seen as the year in which the surveys failed. The case of PODEMOS in Spain, the case of the presidential elections in Peru, the BREXIT, the case of the popular consultation in Colombia, the case of the presidential election in the United States to mention the most talked about. In these processes, most of the opinion polls that were conducted, using traditional survey methods; they gave totally opposite results to those that happened.

What the samples were poorly designed, what the questions were biased, what the absenteeism exceeded expectations, what the *"vote of punishment"* and other explanations have been put forward, to try to justify the causes of these errors.

As the journalist Juan Cuvi mentioned in his article [1] “Voto vergonzante” in the newspaper "El Comercio" of Ecuador, on Saturday, November 5, 2016:*"Perhaps out of suspicion, instinct for self-preservation or false complacency, many voters prefer to ignore in the polls a position that will be revealed in the middle of the secret of the polls."*

To indicate sometimes the fear or the indifference of the people in front of the surveys.

Also, a Spanish journalist, Erasmo Quintana, mentions in his article [2] "The shameful vote" in the publication "Norte Gran Canaria": *"The psychology of this town is what it is"* to indicate that respondents prefer to say anything instead of express your opinion.

And surely they are right. It is the citizen, the voter, the common man who is part of a sample; It has its own idiosyncrasies and motivations; that are not reflected in the questions that are asked.

And at this point, the use of different social networks arises with more force, to know the feeling and idiosyncrasy of the common voter.

Is that nothing is more true, when Humberto Eco mentions [3] in "La Stampa" on 06/10/2015 that: *"Social networks give you the right to speak to legions of idiots who first spoke only at the bar after a glass of wine, without harming the community. They were silenced quickly and now they have the same right to speak as a Nobel Prize. It's the invasion of the idiots. "*

All people, who have access to a social network, express their opinion, criticize, warn and make value judgments on any subject, even more so if the topic is political.

And in these media people get naked, they show themselves as they are. Of course there are thousands who are simply passive actors.

According to Statistics Portal mentions [4] that: *"As of the third quarter of 2017, the microblogging service averaged at 330 million monthly active users."*, is the network that provides the best facilities to analyze the opinions of its users and specifically in the Ecuador according to the National Institute of Statistics and Census, INEC [5], through its Survey of Living Conditions - ECV, Sixth Round 2013 - 2014 states that: *"there are approximately one million users"* that belong to all levels of the partner - economic, which are from all geographical regions of the country, which are of all age ranges and which are from all areas of Ecuador; is that it is feasible to carry out the electoral prediction analysis of Ecuador where the President elections will be held on February 19, 2017.

2 State of the art

Using the information given by social networks to determine which candidate will win an electoral process began when they became massive.

We can, however, say that there have been contradictory positions and very different implementation methods.

According to the authors [6] (Gayo-Avello, Metaxas, and Mustafaraj, 2011) they state that it could be better if other techniques were used. Erratic results and values of mean absolute error, MAE, high make the authors do not recommend the use of analysis on Twitter.

In the results delivered by [7] (Tumasjan et al., 2010) in the parliamentary elections in Germany in 2009; in the work of [8] (Fernández Crespo, 2013) in the general elections in Spain in 2011, in the autonomous elections to the Community of Madrid in 2011, in regional elections to the Region of Murcia in 2011 and in the elections to the Parliament of Cataluña in 2010; using simply the count of the number of mentions for an electoral option in the Twitter manage to guess the results. Moreover [9], (Zarrella, 2010) *"randomly chose thirty electoral disputes. In 71% of the cases, the candidate with the greatest number of followers was also the candidate who occupied the first position in the polls"* in the legislative elections of 2010 in the USA. All these authors agree, that in order to have good results in the predictions, there must be a large number of cases, another empirical verification of the Central Limit Theorem, the greater the number of cases considered, these converge towards the true value.

Different authors, in different electoral processes in the world, have used different polarity analysis techniques besides the simple counting of the mentions, like this: in the general elections of 2010 in the United Kingdom, reported [10] in http: // tweetminster .co.uk; the author [11] (Montesinos, 2013) in the primary elections in Chile in 2013; the authors [12] (Ramadhan, Nurhadryani, and Hermadi, 2014) in the legislative elections in Jakarta in 2014 and the authors [13] (Tsakalidis et al., 2014) in the elections in Greece, Holland and Germany in 2014 had very remarkable results.

The authors [14] (Asur, and Huberman, 2010) agree on the use and analysis of Twitter as a predictive tool, however they state that it is not necessary to work with text mining.

Another consideration that must be taken into account is the language in which the predictive analysis will be developed; they are incipient for the Spanish language according to [15] (Martínez Cámara et al., 2011) *"In this work we have made a first approach to the mining of opinions in Spanish"*.

3 Methodological proposal

The methodology used follows a series of steps that range from the extraction of the data to the electoral calculation process as shown (see **¡Error! No se encuentra el origen de la referencia.**).

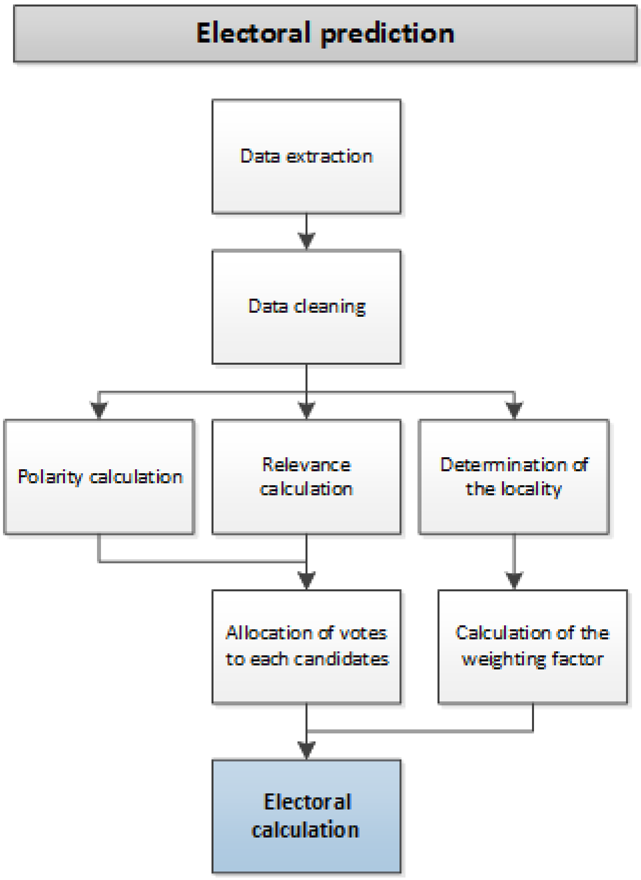

Fig. 1. Methodology used in the Electoral prediction.

This methodology was applied in the presidential elections of Ecuador, a country that is composed of 24 provinces and where all Ecuadorians living abroad can vote, thus becoming another "province". The level in the use of Twitter does not arrive uniformly at all, there are provinces where its reach and penetration is well below the national average. It is clear then that "the voice" of certain voters will not be present in the same proportion as the inhabitants of other areas. Table 1 shows the percentages of the population over twelve years of age that has a Twitter account according to the INEC, through the ECV - Sixth Round 2013 - 2014, a national level of 8.5% has it [5], which represents approximately one million people. It is the inhabitants of the rural areas who have the least, only 2.5% have a Twitter account 3.9% of the inhabitants of the geographical region of eastern Ecuador have a Twitter account; well below the other settlers of the other regions of the country. One way to measure the socio-economic level is through the consumption of people. If quintiles are obtained from this indicator, where the first quintile, will be the people with the lowest consumption and in the fifth quintile, those who have the greatest amount of money to consume; the inhabitants who belong to quintile 3 and lower consumption are below the national average in relation to having an account on Twitter. This asymmetry in the possession of Twitter is one of the reasons why the information downloaded from the Twitter user accounts must be processed differently.

Table 1. Percentage of the population aged 12 and over that has a Twitter account.

Ecuador	**20,4**
ZONE:	
Urban	22,6
Rural	10,5
REGIÓN:	
Sierra	18,7
Costa	22,9
Oriente	11,0
Galápagos	22,0
CONSUMPTION:	
Quintil 1	8,5
Quintil 2	11,5
Quintil 3	14,9
Quintil 4	21,2
Quintil 5	30,6

We will describe each of the processes of the methodology proposed below.

3.1 Data extraction

The objective is to predict the outcome of the presidential elections, in this sense, the first thing that must be done is to obtain the data. In this case, the unit of analysis becomes the diverse opinions that the Ecuadorians have about a certain candidate.
There are many free and paid applications that allow you to download the information of the tweets of the various candidates and the users who follow them. The accounts of the candidates to whom they will be analyzed were first defined. According to a survey conducted by Jorge Arroba Rimassa in October 2016, published in the magazine Vistazo and cited in the article [16] "Who is who? in the duel of the polls "the candidates Moreno, Lasso, Viteri and Moncayo disputed with some option the first places for the presidency, and 10.1% of the electorate would be for a group of four candidates: Bucaram, Espinel, Zuquilanda and Pesántez. To these four candidates, for practical purposes they will be referred to as "others". In this sense, the official Twitter accounts of the candidates and their party were defined to extract the tweets from the users who follow them. Using the tool: Google Tags, the download was started from December 2016 until February 14, 2017; following the observations of [7] (Tumasjan et al., 2010) on the periodicity of data collection. The elections would be February 19, 2017 and a total of 823,135 tweets were downloaded corresponding to users who follow these official accounts.

3.2 Data cleaning

However in the downloads of the various accounts of candidates can filter "bots" or people known as "trolls" who are responsible for inflating the presence in social networks of a certain candidate. According to Carlos Guadián Orta [17], in "How to detect bots on Twitter and their use by digital means" he mentions that: "*the analysis of botnets, which are launched to support candidates and parties ..., to Infact the impact numbers*", have made that there are several ways of detection: for example the difference between unique users and the volume of tweets that they emit or through the publication patterns. The analysis of repeated by the text and the user's identifier was used.

We also eliminated the blank spaces, all the texts were converted to lowercase, the external links were eliminated and we only keep the text that represents the theme of the tweet.

Once this debugging was done, the number of tweets was 132,378, as shown in Table 2 in which are presented the numbers of Twitter messages downloaded-two, the unduplicated ones and the repetition factors; there being a total of almost six repeated for each original tweet.

Table 2. Tweets downloaded, duplicates and repetition factor for each candidate.

Candidate	Downloaded twitter number	Unduplicated twitter number	Repetition factor
Moreno	427.190	40.656	10,5
Lasso	128.976	27.544	4,7

Viteri	84.183	15.005	5,6
Moncayo	71.197	15.700	4,5
Others (Bucaram, Espinel, Zuquilanda y Pesántez)	111.589	33.473	3,3
Total	823.135	132.378	6,2

3.3 Polarity calculation

The next step has to do with the polarity of the messages; as the activity of classification according to the feeling or emotional meaning. Various techniques are used that have to do with the use of natural language. This classification aims to give a quantitative metric, generally in three values, 1 if you have a positive subjective charge text, -1 if you have a negative subjective charge and 0 if you have a neutral subjective charge.

In order to have this polarity metric of a tweet, there are several methods, supervised or computational learning methods, unsupervised methods and hybrid methods.

In this research we used the dictionary method that is based on detecting certain terms that have a positive or negative subjectivity.

For which a dictionary of positive and negative terms was used, extending its power using regular expressions. And then by means of a con-version function the tweet is defined as positive, negative or neutral.

We define if the emitted tweet has a positive, negative or neutral context on the candidate; for which a file of positive terms was used and another of negative terms according to [18] (Ureña López, 2002) "corpora for specific purposes, created in response to a particular purpose".

We proceed to perform a tweet count to identify which ones fall into the positive or negative classification. Then, by means of a difference between positive and negative words, the polarity of each message was determined for each candidate. In Table 3, the results are presented.

Table 3. Polarity for each candidate.

	Polarity			
Candidates	Negative	Neutral	Positive	Total
Moreno	2.520	27.611	10.525	40.656
Lasso	2.370	17.031	8.143	27.544
Viteri	1.095	9.926	3.984	15.005
Moncayo	1.134	12.917	1.649	15.700
Otros	3.368	26.948	3.157	33.473
Total	10.487	94.433	27.458	132.378

3.4 Relevance calculation

A mechanism to detect the importance of Twitter messages is based on the relevance and popularity that a given candidate generates on the voters. There are several metrics to give a measure of the importance of the various tweets; ratio followers / followed, retweets / number of tweets, CTR (Click-through rate) in links; we will evaluate this impact based on the number of retweets and favorites that get the messages.

In this work we have used a variant of what was recommended by Paz Martín in [19] "3 formulas to calculate the engagement rate" to calculate the relevance rate using the weighted interactions; the relevance of a voter's tweet about a certain candidate is a weighted average between the number of retweets that has been assigned a multiplicity factor of 3 and the favorites a factor of 1.

We value the impact of a Twitter emitted by a candidate or on this in the voters, using the equation (1):

$$\text{relevancia}=(\text{favoritos} \times 1 + \text{retwitters}\ \text{x}3)/\text{Number of twitters} \tag{1}$$

3.5. Allocation of votes to each candidates

The number of votes to the candidates is assigned by polarity and relevance, for which the quantiles of the relevance variable were used.

This relevance of each Twitter message must be converted into voices depending on its value; we will determine as a vote in favor of the candidate, if the tweet mentioned has a positive polarity and its relevance will be taken into account in order to scale it; assigning to the vote of the tweet a multiplicity value of the vote; To determine this value, we used the quantiles of the relevance distribution, assigning multiplicity 3 to those that are above the quantile 97.5%, a multiplicity of 2 those that are above the 95% quantile, and the rest a multiplicity of 1.

3.6 Determination of the locality

Since the only "demographic" data available is that of the locality from which the tweets are made and since the location data are not specific in some situations, we will try to identify the locality and parameterize it with the province of origin where was issued, in other cases they will be from abroad and for those who do not have a location in any of these categories they will be defined as others. According to the authors [20] (Peregrino, Tomás, and Llopis, 2013) the "*obtaining a real place from these data is a very complex problem"*.

Due to the low percentage of geolocation, which in general presents Twitter, we have not considered this alternative in its use. According to the INEGI studies [21] "*A high* *percentage of the information generated in social networks does not contain Georeferencing information. The percentage of georeferenced tweets is very low with respect to the total generated, which yields non-representative information on a particular topic. ".*

We will purify certain values of the field of the locality where the Twitter message was emitted, for example there are cases where we find the words: "quito", "uio", "Quito !!!" or another expression associated with the city of Quito, we must find these expressions and associate them with a single name that will be only "Quito", we will carry out a purification for the main 24 provinces of Ecuador, for foreign localities and those that are not localities or that do not fall into the classification. The aforementioned fication or do not possess the Locality field will be assigned to the other locality.

3.7 Calculation of the weighting factor

To the extent that Twitter messages downloaded from users represent a biased sample of the electoral universe, it is that in order to correct this distortion, weighting factors will be used in order to reconstitute the electoral universe.

We calculate the weighting factor by relating the values obtained from the extraction of data, which are only a sample, with the real values that represent the electoral universe. The observations to consider are those that have positive polarity.

The process of **Electoral Calculation**, is the processing of information to obtain the results that are analyzed below.

4 Results

Once for each Twitter message its polarity has been determined, its valuation in votes and its weighting factor calculated, the results have been obtained.

As shown in Table 4, the comparison between the official results given by the National Electoral Council of Ecuador, CNE, the results of the main companies dedicated to conducting political polls in Ecuador, presented on February 8, is presented. 2017; Market and Cedatos linked to the opposition parties and Ciees, Perfiles de Opinion and Opinión Pública close to the officialism [22] and the data processed following the proposed methodology ; using the weighting factor and without using it.

The variability of the results among the polling companies always puts companies in doubt in the first place and then, which is more complex, on the technique of the electoral surveys. In the case of the candidate Moreno, there is a maximum difference of 10.9% between the results of the polling companies.

Table 4. Comparative of official results, given by the polling companies and the data predicted by the methodology.

Candidates	Official results CNE	**Results of POLLING COMPANIES**					**PROPOSAL**	
		Market	CIEES	Perfiles de Opinión	Opinión Pública	CEDATOS	**With** weighting factor	**Without** weighting factor
Moreno	39,4	32,4	43,3	41,7	42,6	38,6	41,7	38,3
Lasso	28,1	20,8	21,3	19,0	22,7	25,7	27,7	29,7
Viteri	16,3	23,0	12,6	16,7	17,1	16,7	14,0	14,5
Moncayo	6,7	13,1	10,8	8,3	9,2	9,2	6,6	6,0
Others	9,5	10,7	12,0	14,3	8,5	9,7	10,1	11,5
TOTAL	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

As shown (see **¡Error! No se encuentra el origen de la referencia.**2), the smallest error that one of the polling companies committed was Cedatos, with an MAE = 1.3 compared to the prediction made analyzing the Twitter messages with the proposed methodology using the factors of weighting with an MAE = 1.1 make this methodology an alternative in the electoral prediction. Also the prediction without using the weighting factors has a good performance.

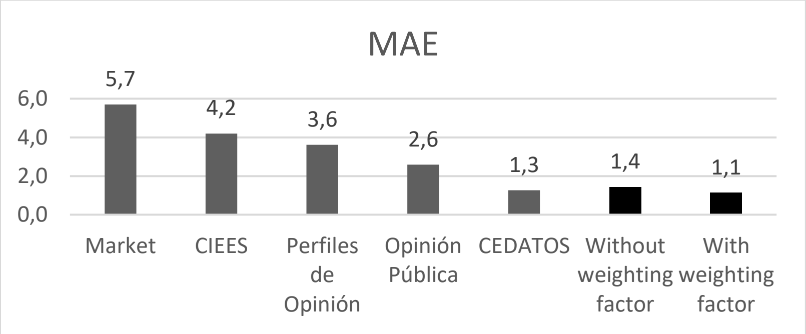

Fig. 2. Comparison of the MAE of the results given by the polling companies and the predicted data.

5 Conclusions

The present study tries to demonstrate that the analysis of the tweets emitted by users about their electoral preferences is as reliable as the results issued by different surveys.

The analysis is based on the debugging of duplicates, which is a way to avoid "bots" whose function is to distort the results.

The contribution of this research is the incorporation of the weighting factors that allow the comparison of the data obtained in a "biased sample", such as the download of Twitter messages, where it can't be controlled in due form that certain regions or provinces have more representation to the detriment of others; with the data of the real electoral universe.

The other contribution of this research is the Electoral Calculation mechanism, which combines weighting factors with relevant measures to enhance the electoral vote.

Additionally, the cost involved in the application of these methodologies, the survey and the analysis of Twitter messages is incomparable; more than the velocity in the delivery of results.

It is clear that this technique has a variety of applications, commercially, in the area of marketing and other areas in which you can have the opinions of users.

References

1. http://www.elcomercio.com/column/juan-cuvi last accessed 2018/02/06.
2. https://nortegrancanaria.es/portal/el-voto-vergonzante last accessed 2018/02/06.
3. http://www.lastampa.it/2015/06/10/cultura/eco-con-i-parola-a-legioni-di-imbecilli-XJrvezBN4XOoyo0h98EfiJ/pagina.html last accessed 2018/02/06.
4. https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/ last accessed 2018/02/06.
5. http://anda.inec.gob.ec/anda/index.php/catalog/358 last accessed 2018/02/06.
6. Gayo-Avello, D., P. Metaxas, y E. Mustafaraj. 2011. Limits of Electoral Predictions Using Twitter. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social, pages 490-493.
7. Tumasjan, A., T. Sprenger, P. Sandner, y I. Welpe. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, pages 178-185.
8. Fernández Crespo M. 2013. Predicción electoral mediante análisis de redes sociales. Memoria para optar al grado de Doctor de la Universidad Complutense de Madrid.
9. https://www.asgroupinc.com/new-data-can-twitter-predict-elections/ last accessed 2018/02/06.
10. http: // tweetminster .co.uk last accessed 2018/02/06.
11. Ramadhan D., Nurhadryani Y., y Hermadi I. 2014. Campaign 2.0: Analysis of Social Media Utilization in 2014 Jakarta Legislative Election. In ICACSIS 2014, pages 102 – 107.
12. Montesinos García L. 2014. Análisis de sentimientos y predicción de eventos en Twitter. Memoria para optar al título de Ingeniero Civil Eléctrico de la Universidad de Chile.
13. Tsakalidis A., Papadopoulos S., Cristea A., y Kompatsiaris Y. 2015. , Predicting Elections for Multiple Countries Using Twitter and Polls. . In Predictive Analytics. IEEE INTELLIGENT SYSTEMS, pages 10-17.
14. Asur S., y Huberman B. 2010. Predicting the Future with Social Media. Technical report. CoRR abs/1003.5699. http://arxiv.org/abs/1003.5699v1.
15. Martínez Cámara, E., M. T. Valdivia Martín, J. M. Perea Ortega, y L. A. Ureña López. 2011. Técnicas de clasificación de opiniones aplicadas a un corpus en español. In Sociedad Española para el Procesamiento del Lenguaje. Volumen 47, pages 163-170.
16. http://www.vistazo.com/search_page?keys=arroba last accessed 2018/02/06.
17. http://ictlogy.net/lo/eadministracion/?cat=1890 last accessed 2018/02/06.

18. Ureña López A. 2002.Resolución de la ambigüedad léxica en tareas de clasificación automática de documentos. Colección de Monografías de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN).Número 1. ISBN 9788484541639
19. http://www.pazmartin.com/calcular-el-engagement-rate/ last accessed 2018/02/06.
20. Peregrino, F. S., D. Tomás, y F. Llopis. 2013. Every move you make i'll be watching you: geographical focus detection on Twitter. In Proceedings of the 7th Workshop on Geographic Information Retrieval, pages 1-8. ACM.
21. http://www.inegi.org.mx/eventos/2015/conacyt/doc/p_Estrada.pdf last accessed 2018/02/06.
22. https://es.wikipedia.org/wiki/Anexo:Sondeos_de_intenci%C3%B3n_de_voto_para_las_elecciones_generales_de_Ecuador_de_2017 last accessed 2018/02/06.

Acknowledgements

This research work has been partially funded by the University of Alicante, Generalitat Valenciana, Spanish Government, Ministerio de Educación, Cultura y Deporte and ASAP - Ayudas Fundación BBVA a equipos de investigación científica 2016(FUNDACIONBBVA2-16PREMIO) through the projects, TIN2015-65100-R, TIN2015-65136-C2-2-R, PROMETEOII/2014/001, GRE16-01: “Plataforma inteligente para recuperación, análisis y representación de la información generada por usuarios en Internet” and PR16_SOC_0013.