ISSN 2007-9737

Sentiment Analysis and Multiple Means Comparison for the 2020 United States Elections

M. Beatriz Bernábe-Loranca[1], Rogelio González-Velázquez[1],
Alberto Carrillo-Canán[2], Erika Granillo-Martínez[3]

[1] Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
Mexico

[2] Benemérita Universidad Autónoma de Puebla,
Facultad de Filosofía y Letras,
Mexico

[3] Benemérita Universidad Autónoma de Puebla,
Facultad de Administración,
Mexico

{beatriz.bernabe, rogelio.gzzvzz, erika.granillo76}@gmail.com,
acarrillo_mx@icloud.com

Abstract. Here comes the abstract. Considering that the presidential elections between Trump and Biden have represented a great impact not only for the United States but also for the world and Mexico, in this work electoral preferences were analyzed using a natural language processing tool known as Sentiment Analysis. The methodology begins with reviewing and categorizing comments related to the 2020 US elections on the social network Twitter. Subsequently, the dictionaries are created to start with the sentiment analysis. In this way, three lines of analysis are established, being reflected in the following way: 1) data collection in the electoral campaign (information retrieval through downloads), 2) creation of dictionaries and 3) sentiment analysis. According to the previous order, first Tweets from different users have been randomly downloaded with the tagging algorithm, considering the comments of the Twitter attendees. The information seen as a corpus led to the definition of dictionaries and consequently, sentiment analysis bifurcates the information into two classes. Such categories have been called praise and name calling for convenience for the purposes of this article. Finally, the frequency of the terms is analyzed with descriptive and inferential statistics using the Fisher mean comparison.

Keywords. Dictionary, Twitter, NLP, Python.

1 Introduction

In August 2020, Democrats Joe Biden and Kamala Harris surpassed Republicans, President Donald Trump and Vice President Mike Pence by 12 percentage points (Biden and Harris with 53% compared to 41% who supported Trump and Pence with respect to the presidential elections of November 3, 2020 [5].

Given the global relevance of these elections and the impact that social networks have on electoral elections, this document presents the analysis of the opinions of Twitter users about the 2020 US presidential candidates. The study seeks to examine the trends from the emotions that users recharge on the candidates.

In market researches on political preferences, Artificial Intelligence (AI) techniques have offered attractive alternatives to respond to electoral problems where different AI efforts have been brought together. Recently, the processing of large amounts of data on political opinions within social networks has been very useful, however, the cost of this task is high due to the manual review of the content in the creation and use of

doi: 10.13053/CyS-26-1-4155

dictionaries that label words. The efforts applied to address this problem are diverse, but it is possible to say that many of them focus on supervised machine learning, even so, little research stands out on this topic. In a pragmatic sense, it has been chosen to resort to predictive models of classification algorithms that many languages already have implemented in their libraries, for example, Naive Bayes.

The purpose of this research is to analyze the influence that Twitter has on the 2020 US political environment, both for ordinary users and for political candidates. The aim is to identify the candidate who has the most followers of the 2020 election period in the US The use of hashtags that represent the problem on twitter has been considered for the extraction of information, in the same way, retweets (RT), followers, likes, etc.

Twitter APIs have been used to download user opinions around Trump [10]. The public information that comes from the Twitter accounts is legally accessible, as long as the use of the platform is for academic purposes. On the other hand, these data are very useful for the sentiment analysis, which is the final part of this work, where the sentiment indicator is obtained by the Naive Bayes classifier of Python [2]. Finally, the results were studied with an Analysis of Variance and Fisher's mean comparison.

The study carried out is reduced to the analysis of political opinions of the US presidential candidates based on Twitter with Natural Language Processing techniques that consists of the following:

1. Define the corpus of the problem according to the names of the presidential candidates expressed in hashtags.
2. Information extraction begins through Tweets downloads. This processed information comprises 2000 random tweets per day for a month.
3. Once the information was freed from intractable characters, 500 comments were selected for each candidate to give rise to the tagging algorithm in order to create dictionaries.
4. Finally, the sentiment analysis is developed that separates the comments of the tweets into two types: praise and insults.

On the other hand, the basic aspects of implementation are briefly described and finally, under the differential statistical technique of means, some conclusions are reached. This work is organized as follows: section 1, Introduction, section 2, the framework of the US elections is presented. Section 3 is responsible for presenting the development of the implementation, in section 4, the data analysis is concentrated, finally, in the last section, the results and future work are discussed.

2 Preliminary Framework of the US Elections

In a poll by The Washington Post newspaper, it was published that Biden and Harris had 53% preferences compared to 41% who supported Trump and Pence. The estimate refers to 2 months prior to the presidential elections on November 3. The sur-vey shows that both President Trump and his vice president are disadvantaged by handling the health crisis caused by the pandemic and the deterioration of the country's economic situation generated by the negative macroeconomic effects of Covid-19.

Among the voters surveyed, 9 out of 10 declared themselves enthusiastic to participate in the elections and to vote for the Democratic duo Biden-Harris, with 65%. In the case of Trump and Pence, 3 out of 4 voters who support the duo opt for the re-election of Republican candidates. Thus, the interest of the second group lies in increasing support for the Republican win instead of concentrating their energies on expressing their rejection of defeating the Democratic Party candidates [5].

Through social networks, both citizens and elected officials debate on issues of political interest, as well as other matters. Thus, with the use of social media platforms, residents choose to organize, meet and communicate. These means allow users to communicate directly with voters, interacting in real time, a situation that would not be possible through postal mail or email. Thanks to Twitter, the interactive process is viable and dynamic [9].

Recently, with the study of political opinions generated by social networks [4], contributions focused on manual classification and / or automated content analysis have been revealed using dictionaries where words are tagged giving a negative or positive prior value for each word [7]. Other considerations made from supervised machine learning or directly derived from artificial intelligence are scarce in research in communication sciences [12, 11]. Although these strategies seem artisanal, they have been very useful for the analysis of opinions in the area of politics [10].

In public, private, academic institutions, sociological and political research centers, Big Data solutions in a broad sense have been very useful. Similarly, various efforts continue to encourage PLN personal learning network tools such as automated sentiment analysis to generate data to help political advisers design, improve and plan both speeches and electoral campaigns on social media.

Since politics reached social networks, the implications have grown in such a way that the objective of political strategies in web 2.0 is to reach and influence those populations that have no interest in politics.

The use of social networks will not be totally decisive in the elections, however, it will be a tool that will contribute votes gradually according to the growth and internet access among voters. It has been observed that the most important thing is the influence of everything that is written on social networks since it directly influences public opinion [8].

The proposal that is exposed in this document allows, based on massive data, that it is possible to analyze opinions and even detect changes in the trend indicators of political parties and their candidates. The study points to the intersection of social science and computer science, particularly artificial intelligence.

3 Methodology

The first stage lies in the collection of data, the criteria used consisted of obtaining information from Twitter to analyze the comments of the US candidates for the 2020 presidency.

To do this, the keywords of the problem must be found, which implies extracting information from Twitter through the downloads and later creating the dictionaries. To download the tweets, the first step is to identify the relevant information in the tweet. When enough tweets are collected, they should be concentrated with the fewest number of words due to the ambiguity of the relationships between the comments and the words of the downloaded tweets. This control allows you to eliminate useless information that is generated in large quantities and that would be impossible to process.

Once a considerable number of tweets have been downloaded, the information is cleaned up with UTF-8 encoding to be used as strings. On the other hand, the Python division function has to be applied, which divides a character string into substrings according to a delimiter to fragment the strings into at least two parts.

3.1 Tweet Downloads and Character Cleaning and Encoding

To start the data processing with the algorithm, it is necessary to start the download of tweets, then the training corpus is created to analyze the success rate. In this step, in addition to the downloads, a conversion to the "UFT-8" encoding is performed [3]. The algorithm is presented in the section of appendix A of the document.

This step consists of the implementation of the algorithm for the substitution of special characters, for example: accent marks.

3.2 Label Classifier

With the document free of special characters, it is possible to build the custom label grouping algorithm. In this case, the tweets are marked by news, opinion or announcement according to the proposed dictionary and that are associated with the hashtags. The first tag has been named "news" since it shows some unbiased comment. This implies that a sentiment is not identified and for our problem it means that the news is irrelevant and discarded from the analysis.

The second label is opinion. An opinion in this study is interpreted as a comment that reflects a positive or negative feeling on a specific topic. In this case, feelings about politics are at the center of opinions.

Table 1. Concentrated negative sentiments

Sentiment	**Trump**	**Biden**	**Kamala**	**Pence**
Excuse	1	0	0	7
Disliked	1	0	0	0
Military	4	9	1	86
Average	3	0	1	0
Criminal	0	0	2	0
Little	5	14	15	2
Guilty.	0	1	0	0
Game	9	2	3	1
Long	7	21	5	3
Creepy	0	4	1	1
Victim	1	0	0	0
Less	5	4	2	3
Little	1	0	0	0
Evil	1	6	15	293
Seriously	1	0	0	0
Long	1	0	1	0
Slow	2	4	1	0
Few.	1	0	0	0
Tries	1	2	2	0
Tense	1	0	0	0
Hardly	2	0	0	0
Loses	0	1	0	0
Thick	1	0	0	0
Broken	4	0	10	6
Lousy	3	0	0	0
Average	1	0	0	0
Close	0	0	1	0
Blind	0	2	0	0
Mere	1	0	0	0
Excuse	1	0	0	7

Once the tag has been assigned, the tweets are divided into two documents: flattery and offense, which correspond to positive and negative feelings respectively.

In general, the steps of the algorithm boil down to the following:

- Start the collection of tweets where the mentioned users are found through the following hashtags: Biden, Trump, Kamala, Pence.
- A list of words is created, performing a previous cleaning of hyperlinks or incomplete words with the ending of "..."
- Two dictionaries of the best words that appear 10 times or more are created, using as a base the 500 collected tweets where the presidential candidates are mentioned.

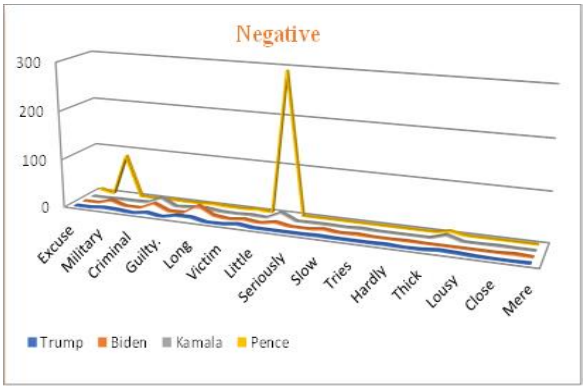

Fig. 1. Offenses for presidential candidates

- A list of tweets where the words inserted in the dictionary from step 2 appear is returned.
- An analysis is carried out where they are divided into two dictionaries: 1 Flattery, 2 Insults.
- A final account is created where the praise and insults appear where the presidential candidates are mentioned.

4 Sentiment Analysis

In this section, a review of the results of the dictionary algorithm and sentiments is presented. Descriptive statistics and comparison of multiple means are the focus of this section. The variability of praise and insults consisted mainly in identifying the number of words that reflected a feeling regarding a compliment or an insult [1]. In this work, the frequency of the qualified terms has been through the main observation of the tweets, therefore, those words that were repeated at least 10 times were included in the dictionaries as shown in Tables 1 and 2 sentiments negative (see Fig. 1 and 2).

4.1 Negative Sentiments

The negative feelings did not produce the expected results according to the diversity and quantity. The Naïve Bayes algorithm identified many neutral words, however, it was possible to associate each of the candidates with the offenses (negative feelings) that are seen in Table 1.

The amount of insults observed in the graph are less so that the information is insufficient to generate a statistical analysis.

In figures 1 and 2, it is observed that the majority of insults or disapprovals were monopolized by Pence with 402 followed by Biden with 70 negative feelings.

It should be noted that the word “game” in table 1 was identified as a negative sentiment, which was surprising because it was expected to be taken as a neutral or positive sentiment.

4.2 Positive Sentiments

Positive feelings are responses that vary according to the individual opinion of each tweeter user. In Table 2, the concentration of positive feelings of the four candidates is shown and in Fig. 2, the graph of praise as positive feelings to the candidates is observed. positive sentiments. Kamala has 939 accolades followed by Trump with 676.

An inconsistent datum was the word "sexual" because the data classifier identified it as a positive feeling, which was associated with a higher score with Kamala.

5 Data Statistical Analysis

According to Table 3, where the averages and standard deviations for positive feelings were concentrated, it is assumed that the means are very similar for the two pairs of candidates and the standard deviation (ST) between Kamala and Pence do not have significant differences.

At this point, to verify this observation, a One-Way ANOVA with a two-tailed hypothesis test confirms whether the means of the four candidates are equal.

The analysis of variance reveals the following results [6]:

- One-Way ANOVA: Positive Feelings vs. Candidates.
- Null hypothesis (Ho), all means of praise are equal.

Table 2. Concentrated of positive sentiments

Sentiment	**Trump**	**Biden**	**Kamala**
Real	205	1	0
More	39	62	29
First	8	18	53
Very	96	27	13
Proud	0	1	165
Love	39	15	83
Secure	0	69	0
Social	3	1	5
Better	15	8	63
Favorite	0	55	60
Right	44	20	21
Many	21	29	39
Good	16	20	23
Sure	5	13	52
Really	12	15	37
Much	4	26	17
Wants	22	22	9
Whole	36	9	6
Most	18	18	7
Cool	1	0	49
Sexual	1	5	37
Proud	0	4	32
Kind	3	8	26
Clearly	35	3	0
Latest	34	0	4
Truth	7	11	12
Great	10	8	9
Sexual	0	0	37
More	1	8	25
Clear	1	2	26

- Alternative hypothesis (Ha) Not all means are equal.
- Significance level α = 0.05
- Equality of variances is assumed for the analysis.
- Factor information.

From table 3, SS is sum of squares that explains the total variance, SMS is the quotient of SS between the degrees of freedom, F-Value means Fisher statistic, P-Value: In statistical hypothesis testing, the P-value or probability value is the probability of obtaining test results at least as extreme as the results actually observed, assuming that the null hypothesis is correct.

In Table 3, the ANOVA p-value is greater than 0.05, therefore, Ho is accepted, which means that

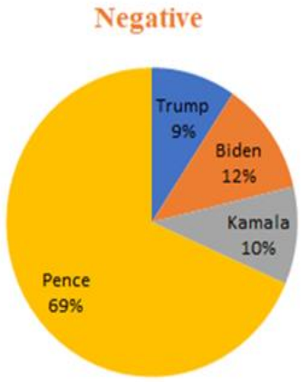

Fig. 2. Offenses percentage graph for presidential candidates

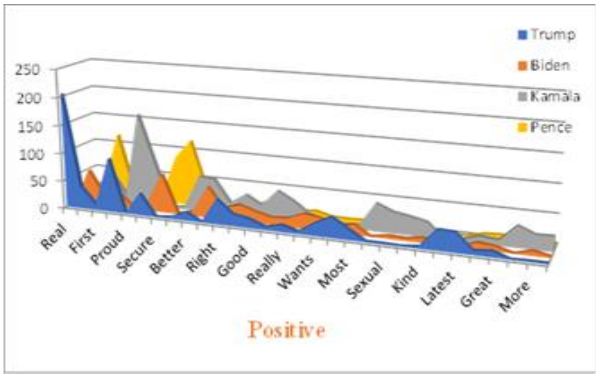

Fig. 3. Concentrate of positive sentiments and compliments

the means for all candidates are equal. To underline this conclusion, the Fisher (F) multiple-mean pair comparison technique was used with the LSD method with 95% confidence. The analysis is reflected in Table 5:

The means that do not share a letter are significantly different, in addition, the value of the p ANOVA is greater than .05, therefore, all the candidates share the same letter according to Fisher, so that Ho is maintained as a valid hypothesis, that is, there is no significant evidence that praise favors a candidate.

For the case of tables 3 and 4, the nomenclature used is the following:

- ANOVA. An F-value is identified for each term in the ANOVA. The F-value is the test statistic used to determine if the term is associated with the response. The F-value is used to calculate the p-value, which allows to make a decision about the statistical significance of the terms and the model.
- FD. Total Freedom Degrees are the amount of information in the data. The total FD is determined by the number of observations in the sample.
- SS. Sequential sum of squares are measures of variation for different components of the model. Unlike adjusted sums of squares, sequential sums of squares depend on the order in which terms are entered into the model.
- SMS. Sequential mean squares quantify how well a variation explains a term or a model. Sequential mean squares depend on the order in which the terms enter the model.
- Unlike sequential sums of squares, sequential mean squares consider degrees of freedom. The sequential mean square of the error (also called MSE or s2) is the variance around the fitted values. Sequential mean squares are used to calculate the p-value of a term.
- F value. The F value is the test statistic used to determine if the term is associated with the obtained result. The F-value is used to calculate the p-value that is used to make a decision about the statistical significance of the terms and the model.
- P Value. The p-value is a probability that measures the evidence against the null hypothesis. Lower probabilities provide stronger evidence against the null hypothesis. To determine if the association between the response and each term included in the model is statistically significant, the p-value of the term is compared with the level of significance to evaluate the null hypothesis.
- The null hypothesis is that there is no association between the term and the answer. Figure 4 shows Fisher's multiple mean pairs comparison technique with the LSD method at 95% confidence.

It also shows that the comparisons between all the candidates the associated interval contains zero, which means that the comparisons of the means are equal [6].

Since the comparison of multiple means indicated that there are no predilection differences between the candidates, an analysis of a box plot was carried out, where it is observed that the greatest variability of compliments is found by Kamala (939) with a single atypical compliment.

For his part, Trump ranks second with 676 compliments and two outliers. Pence has 546 positive sentiments and is the one who concentrates the least praise for variability. Finally, Biden in last place with 478 with 3 outliers.

The interquartile range boxes in a boxplot represent the middle 50% of the data. Whiskers extend to maximum and minimum data points within box heights.

The mean represented by the circle that is above the median, corresponds to Biden with a value of 15.93, which had already been calculated previously, the horizontal line that divides the box indicates quartile 1 (Q1) which is 2.75, the median of 10, the quartile 3 (Q3) that is the end of the box with a value of 20.5, the interquartile range is 17.75, the lower whisker is zero and the upper one is 29, the whiskers signify the variability. It should be remembered that the sample is 30 as seen on the box.

As for Kamala, the mean is 31.3, the Q1 represented by the upper edge of the box, is equal to that of Biden of 8.5, the median that is the horizontal line that divides the box almost in half is 8.5, a median of 25.5, the Q3 is 41.5, previously observed, the means were calculated previously.

Finally, for Trump the mean is 22.53, the Q1 represented by the upper edge of the box is 1, the Q2 is 9, Q3 is 34.25, the interquartile range is 33.25, as for the whiskers, the lower one marks zero and the upper 44, for all cases the sample is 30.

As can be seen, there are no substantial differences in the case of the means, however, in the case of the medians the differences are notable.

Interquartile range of 33, the lower whisker of zero and the upper whisker of 8, in this case if a variability is observed.

For Pence, the mean is 18.2, the median is 7, Q1 is 2, Q3 is 14.5, interquartile range 12.5, the

Table 4. Candidate averages

Candidates	N	Measure	S.D.
Biden	30	15,93	17,93
Kamala	30	31,30	32,98
Pence	30	18,20	31,40
Trump	30	22,53	40,18

Table 5. Fisher's LSD method and 95% confidence

Candidates	N	Measure	Group
Biden	30	31,30	A
Kamala	30	22,53	A
Pence	30	18,20	A
Trump	30	15,93	A

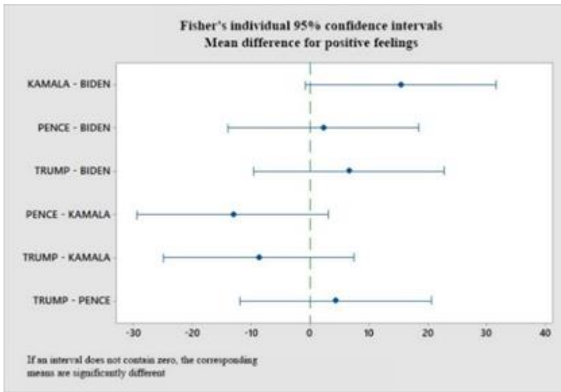

Fig. 4. Fisher multiple comparisons of positive sentiments

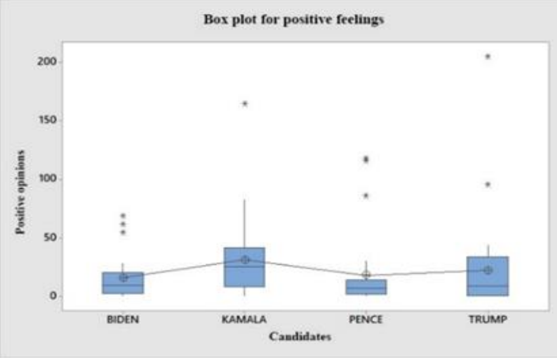

Fig. 5. Box plot for candidates

lower whisker zero, the upper one 31. As Conclusions and Future Works

The corpus handled in this study is made up of random Twitter accounts that are related to the hashtags of the problem.

The accounts were active during the election period, that is, the investigation was conducted months before the presidential elections.

Respecting the ANOVA and the Fisher means comparison, it is stated that with 95% confidence there is no evidence that a candidate has preference on Twitter with respect to compliments.

However, analysis of the box plot reveals that Kamala has an interesting advantage over the other candidates, and although Trump is in second place, Pence helped him with the arithmetic. Between the two they receive 2,161 approvals and the duo Kamala-Biden 1,417. These numbers are not conclusive because individually Kamala has a significant lead, but not as a group.

Therefore, we can say that the ANOVA has shown a correct result, however, it is important to clarify that the study was done 2 months before Trump was diagnosed with Covid-19. Negative feelings need to be examined with other techniques and supplemented with the confusion matrix.

Although the comparison of Fisher means generated that the means are equal in positive sentiments, this does not imply that there is a contradiction, it is even consistent with the scattered numbers that were presented in different social media about the preference of a particular candidate, that is to say, it is encouraging that the results coincide with the closed contest on the days of the voting, since the votes were not counted precisely to know the winner, which is consistent with the results obtained.

In this point, this document leaves a series of questions that will be resolved in future work, for example: What are the political issues that attract the most attention? What makes people want to interact with these public figures? In what way should the candidate express himself on these networks to get users to participate in his social pages?

Appendix A

Algorithm 1. Python code to get the training corpus

```
import sys
import urlib
import re
import codecs
import json
from pattern.web import Twitter
import io
veces=0
 s=open("training.txt","w");
si=open("test.txt","w");
engine=Twitter(Language= “en”)
for j in range(50):
for      tweet             in
Tweetter().search('#ULTIMA
HORA',start=1,count=100):
            if veces%5==0:
            m=tweet.text.e code('utf-8')
            si.write("El tweet: "+m+"\n")
    else:
            m=tweet.text.encode('utf-8')
            s.write("El tweet: "+m+"\n")
    veces+=1
    print veces
```

Algorithm 2. Algorithm for the substitution of special characters

```
import sys
reload(sys)
sys.setdefaultencoding("utf-8")
import csv
import unicodedata
 from pattern.vector import NB, kfoldcv,
count, KNN, Document, Model
#Función para remover acentos
def remove_acents(input_str):

nkfd_form=unicodedata.normalize('NFKD',Unicode
(input_str))
return u"".join([c for c in nkfd_form if
not
unicodedata.combining(c)])
#Para crear un nuevo archivo sin acentos
sinAcentos=open('sin_acentos.txt', 'w')
 #Abre el archivo que se le quitaran
acentos (training)
with open('training.txt') as f:
read = csv.render(f)
for row in read:
    for element in row:

sinAcentos.write(remove_acents(element))
#Cierra el archivo
sinAcentos.close()
```

References

1. **Bermingham, A., Smeaton, A. (2011).** On using Twitter to monitor political sentiment and predict election results. Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), pp. 2–10.
2. **Bernábe, B., Espinoza, E., González, V., Cerón, C. (2020).** Algorithm for collecting and sorting data from Twitter through the use of dictionaries in Python. Computación y Sistemas, Vol. 24, No. 2, pp. 719–724. DOI: 10.13053/CyS-24-2-3408.
3. **Severance, C. (2013).** Python for informatics: Exploring information. Createspace Independent Publishing Platform.
4. **Whitman, W. (2015).** Trending now: Using big data to examine public opinion of space policy. Space Policy, Vol. 32, pp. 119–16. DOI: 10.1016/j.spacepol.2015.02.008.
5. **Esquivel, J.J. (2020).** La dupla Biden-Harris supera a la de Trump-Pence en preferencia electoral, según sondeo. Revista Proceso.
6. **Fallas, J. (2012).** Análisis de Varianza: Comparando tres o más medidas. San José: Universidad de Cooperación Internacional.
7. **Leetaru, K. (2011).** Data mining methods for the content analyst: An introduction to the Computational Analysis of Content. Routledge. DOI: 10.4324/9780203149386.
8. **Oviedo, J.C. (2011).** El uso de las redes sociales en las campañas electorales. (Tesis de Maestría) Escuela de Graduados en Administración Pública y Política Pública, Instituto Tecnológico y de Estudios Superiores de Monterrey.
9. **Ortiz, A.L., Pérez, O.E., Vargas, E. (2015).** Estudio en tendencias diarias en Twitter. Universidad Complutense de Madrid.
10. **Waykar, P., Wadhwani, K., More, P. (2016).** Sentiment analysis in twitter using Natural Language Processing (NLP) and classification algorithm. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Vol. 5, No. 1, pp. 79–81.
11. **Zoonen, W., G.L.A. Van-der-Meer, T. (2016).** Social media research: The application of supervised machine learning in organizational communication research. Computers in human behavior, Vol. 63, pp. 132–141. DOI: 10.1016/ j.chb.2016.05.028.
12. **Vinodhini, G., Chandrasekaran, R.M. (2012).** Sentiment analysis and opinion mining: A survey. International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, No. 6, pp. 282–292.

Article received on 29/07/2021; accepted on 30/09/2021.
Corresponding author is M. Beatriz Bernábe-Lorancai.