v. 2, n. 22, 2022

Journal of
Engineering
Research

ANALYSIS OF TWEETS OF SPANISH POLITICIANS AND STUDY THROUGH TEXT MINING TECHNIQUES

Raquel Enrique Guillen
CaixaBank Business Intelligence

Ursula Torres Parejo
University of Granada
ORCID: 0000-0003-0496-7609

Maria Dolores Ruiz Jimenez
University of Granada
ORCID: 0000-0003-1077-3173

Abstract: The objective is to extract information from the social network Twitter, specifically from four users who represent political figures in Spain, for further study applying different statistical and mining techniques. The users studied are: Pedro Sánchez, the current president of the government and of the Spanish Socialist Workers Party (PSOE), Pablo Iglesias, former vice president of the government and of United We Can, Inés Arrimadas, president of Ciudadanos and Pablo Casado, president of the Popular Party (PP).). The methodology used consists of, starting from an average of 1,900 tweets per user and a total of more than 7,600 extracted tweets, applying hierarchical and non-hierarchical clustering techniques, association rules, classification of tweets based on their authorship using machines of support vector and, finally, an analysis of sentiments using a dictionary of words. For the classification, a mathematical model is trained and the politicians are divided into two groups for its application: on the one hand, Sánchez together with Iglesias, and on the other, Arrimadas with Casado, due to the, a priori, similarity between them. The results show the heterogeneity of the data collected, the classification of tweets based on authorship fails in 8.72% for the case of the president and former president of the government, and 10.44% of erroneous classifications are obtained for the other two politicians. Sentiment analysis shows how users' tweets stand out for having a high degree of "trust". The results of the classification show how, despite having formed a government coalition in the past, Sánchez and Iglesias publish more different tweets than those published by Arrimadas and Casado, which did not have to show a more even content, but the model failed more times when classifying their tweets.

Keywords: Text mining, discourse analysis, statistical techniques, politics.

INTRODUCTION

In recent years, the use of social networks has increased exponentially. Similarly, social networks have become important platforms within different topics such as political discourse [Garimella, K. et al. 2016; Hsu, CL et al. 2013], whose impact depends on the way in which the message is transmitted and received.

There are numerous articles that study political discourse and its importance [Bayram, F. 2010; Chilton, P. 2004; Grimaldi, D. 2019; Johnson, K. and Goldwasser, D. 2016; Van Dijk, TA 2002], but this work intends to make a study through the activity offered by the social network Twitter, one of the most important and used that allows direct contact between users and have verified accounts that represent figures or official entities. recognized as such [Garimella, K et al. 2016]. The interest of a study based on this social network lies in the fact that many figures, people or entities have abandoned traditional communication tools to *microblogging services* [1] such as Twitter to express their political opinion [Grimaldi, D. 2019], beliefs or react to recent events [Johnson, K. and Goldwasser, D. 2016], as well as its use in key political moments such as electoral campaigns or critical situations in countries.

The main objective of this work is to study the political discourse of four Spanish political figures using different and diverse text mining techniques. This speech is immersed in the activity that each user offers in the social network Twitter and the politicians that will be the object of study of this work are:

- Pedro Sánchez, current president of the government and of the Spanish Socialist Workers' Party (PSOE).
- Pablo Iglesias, former vice president of the government and United We Can.

- Inés Arrimadas, president of Ciudadanos.
- Pablo Casado, president of the Popular Party (PP).

METHODOLOGY

DATA COLLECTION AND SOFTWARE USED

Twitter only allows us to collect a maximum of 3,200 tweets per user, but it will not always be possible to collect that maximum for each account.

Table 1 shows the number of tweets extracted for each user together with the start and end date of data collection.

STATISTICAL METHODS USED

After all the data collection in textual format, such as tweets, it is necessary to pass the extracted information through a cleaning and tokenization process, to later carry out an exploratory analysis and apply the following text mining techniques:

HIERARCHICAL AND NON-HIERARCHICAL CLUSTERING

Hierarchical clustering aims to group data into clusters, but just forming a hierarchy. Normally the results of this technique are shown in a dendrogram, a scheme where the similarity of the terms will be given by the height of the closest common node. Sometimes this can help us visually determine the number of clusters that best fits our data, but this is not an easy task [Berzal, F. 2017; Zaki, MJ et al. 2014].

To apply non- hierarchical clustering, the algorithm known as k-means will be used. The name of this method is given by the representation of each of the clusters by the mean (or weighted mean) of its points, which characterizes each group and is normally found in the center or in the middle of the elements that compose it. [García Cambronero, C. and Gómez Moreno, I. 2006; University of Cincinnati, nd; Zaki, MJ et al. 2014].

Since the number of clusters must be given when calling the method, it would be necessary to find the optimal value based on our data. Three methods are used for this: Elbow Criterion [Bholowalia, P. and Kumar, A. 2014], Silhouette Method [Shi, C. et al. 2021] and Gap Method [Tibshirani, R. et al. 2001]

ASSOCIATION RULES

Association rules are used to obtain new patterns of objects/attributes that usually appear together [Martínez, CG 2021].

They are evaluated using the support, confidence, and litf metrics [McNicholas, PD et al. 2008].

In this analysis, the Apriori algorithm is used to obtain association rules from sets of items that are considered frequent. [Agrawal, R. et al. 1993].

SORTING WITH SUPPORT VECTOR MACHINES

Using a statistical learning model based on support vector machines, the aim is to classify tweets based on their authorship.

Suppose we have a data set that is completely separable by a hyperplane. A problem arises directly, and that is that there are infinite. This situation is solved from what is known as the optimal separation hyperplane: the most distant hyperplane of all the observations, that is, the one with the maximum margin [Rodrigo, JA 2017]. The margin is defined as the smallest of the distances of each observation to the hyperplane.

This model will be used to classify the tweets according to their authorship, on the one hand Inés Arrimadas and Pablo Casado, and on the other hand those of Pedro Sánchez and Inés Arrimadas. The grouping of politicians this

way is due to the a priori similarity between them.

SENTIMENT ANALYSIS

The last technique that is applied is sentiment analysis, which, based on a dictionary, classifies words into eight emotions, in addition to their polarity. These emotions are anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. All the words of all the politicians' tweets will then be collected and classified, being able to study which terms have been classified more times as a certain feeling or which is the predominant one in the study. In addition, we will analyze the evolution in the polarity of the tweets over time.

RESULTS

Within the exploratory analysis, the activity of the different users is collected over time (Figure 1) and this same superimposed (Figure 2).

We can see how the activity of Inés Arrimadas ceases in a period of time that coincides with the maternity of politics. Omitting this time interval, we can see that their activity together with that of Sánchez and Casado is similar. Three moments stand out in which the activity of Pablo Iglesias is increased; the earliest coincides in time with the country's general elections in which the politician ran, the second could be related to the controversial Dina case that was related, and the last coincides with the Madrid elections in which he also ran as candidate.

In the same way we can obtain the total number of words and the total number of words used by each user, shown in Table 2.

After removing the *stops words,* we can represent in a graph the words most used by each politician, shown in Figure 3.

It is observed how users share the most used words such as "Spain", "government" and "today". We can contextualize this coincidence since we study political figures of the country and they are words related to this theme. The graph also shows that "Sánchez" is one of the words most used by Arrimadas and Casado, politicians who preside over opposition parties.

Similarly, these results can be represented in what is known as Word Cloud or word cloud, where the size of each word is determined by its frequency in the text (see Figure 4).

Likewise, the most frequent n-grams with n=2 and n=3 (bigrams and trigrams, respectively) are collected and represented using a word cloud. These can be seen in Figures 5 and 6.

Being in a Spanish political context, we are not surprised by bigrams such as "Sánchez must", "European funds", "all of Spain", "public services", etc. Also noteworthy are the two bigrams that have appeared the most: "years ago" and "thank you very much". The latter offers us information about a use that politicians could give to this social network and it is, in effect, to publicly thank certain people or actions. In the same way, we can observe that a clear reference to the past is constantly made, since the expression "years ago" is frequently repeated by users.

As for the trigrams, results continue to appear that support all the information obtained so far. An example of this is the appearance of the trigram "years ago today", which tells us that this reference to the past is usually due to anniversaries of certain events or mentions of the former terrorist organization ETA: "years ago ETA murdered". We see a set of trigrams such as "we will continue working together", "we must work together" or "we must continue working" that provide us with information about the use of the social network by politicians: calling on the people to act together. Lastly, trigrams such as "so difficult year" and "doing massive tests"

User	Number of tweets extracted	Start and end date of data collected
@InésArrimadas	1678	011/05/2019 – 0/17/2021
@pablocasado_	2201	01/14/2020 – 08/16/2021
@PabloIglesias	2049	08/22/2019 – 05/04/2021
@sanchezcastejon	1734	02/04/2020 – 08/15/2021

Table 1. Users, number of tweets extracted along with the start and end date of the data collected.

Figure 1. Twitter activity of the different politicians.

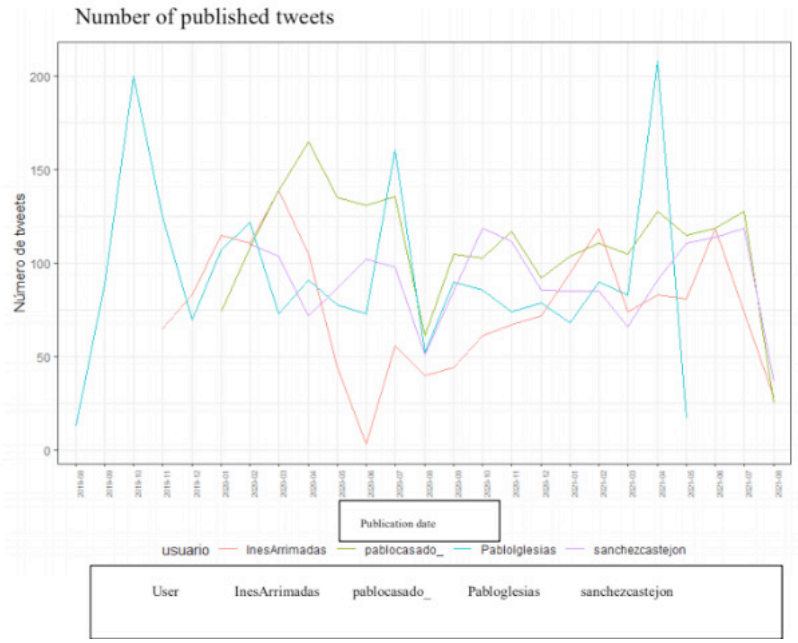


Figure 2. User activity overlaid.

User	*total words*	*Total different words*
@InesArrimadas	60499	7999
@pablocasado_	87365	10092
@PabloIglesias	59271	9498
@sanchezcastejon	67830	8949

Table 2. Total words and total words used by each user.

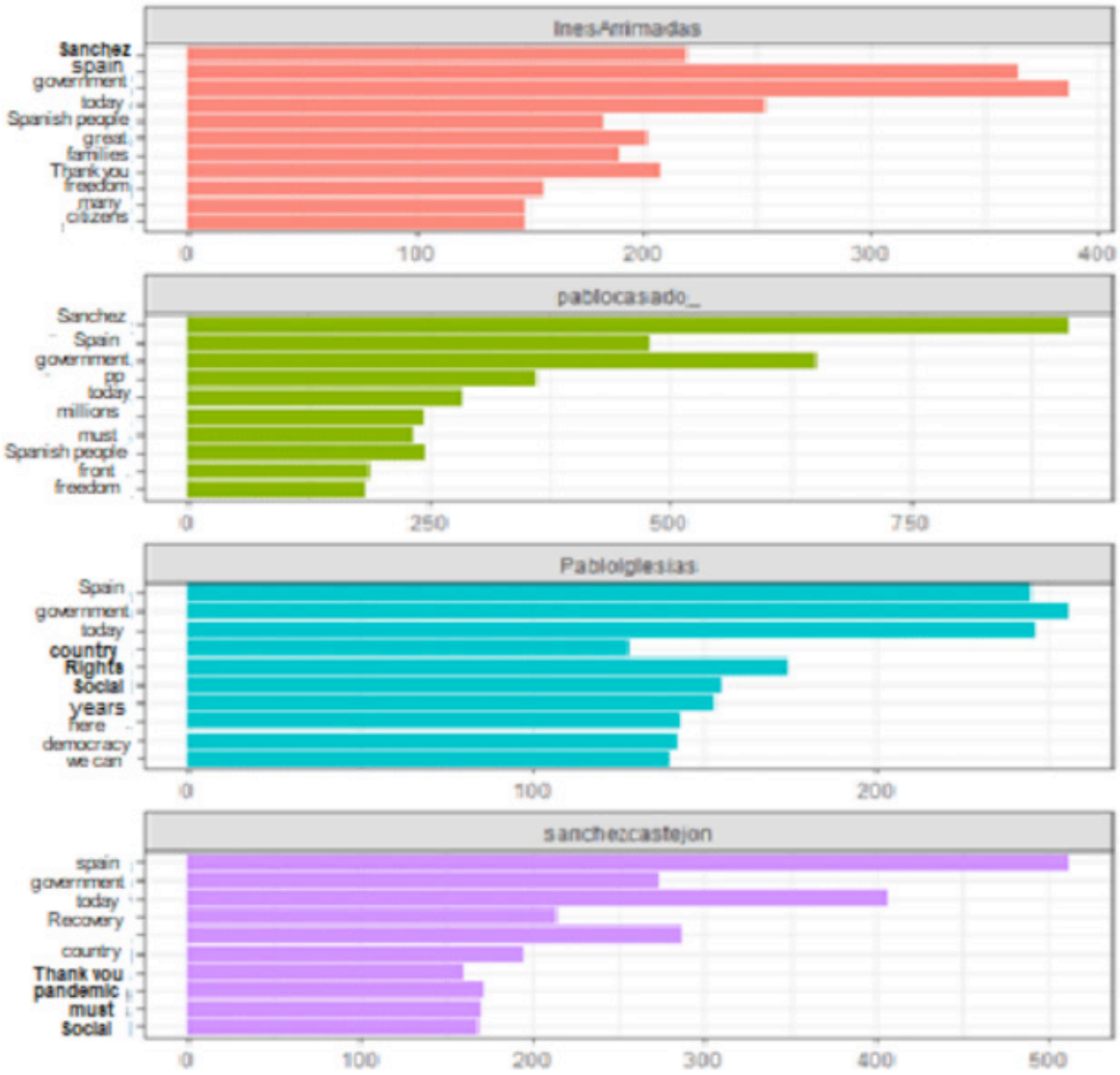


Figure 3. Words most used by users.

Word cloud de Inés Arrimadas

Word cloud de Pablo Casado

Word cloud de Pablo Iglesias

Word cloud de Pedro Sánchez

Figure 4. Word cloud of the different politicians.

Figure 5. Most frequent bigrams.

Figure 6. Most frequent trigrams.

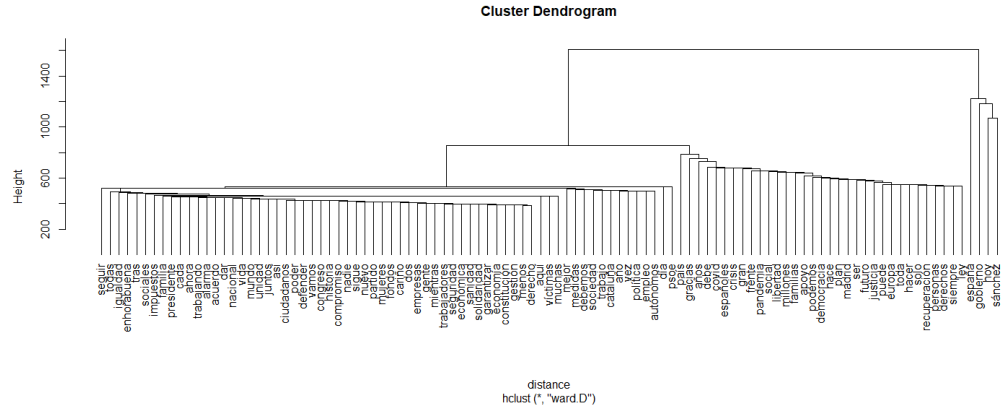


Figure 7. Dendrogram.

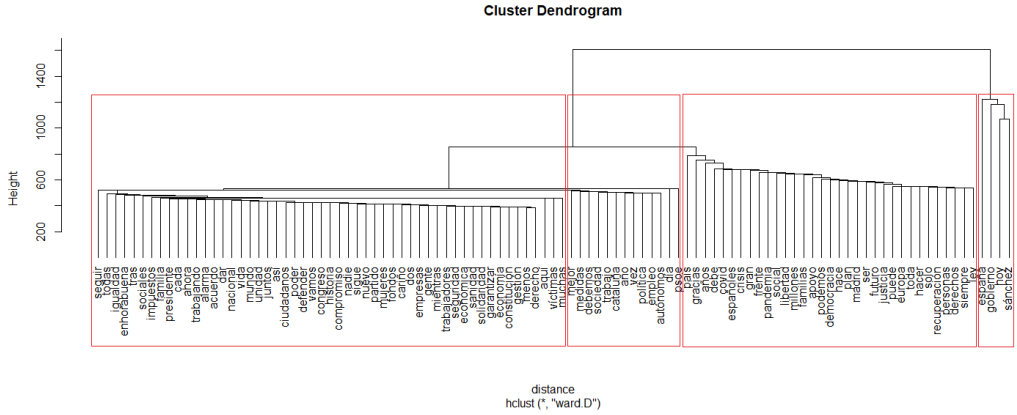


Figure 8. Grouping of terms in the dendrogram.

could be related given the pandemic situation experienced, along with other international issues.

CLUSTERING

After applying the Gap, Silhouette and Elbow Criterion methods, we deduce that the optimal number of clusters is 4, obtaining the dendrogram in Figure 7.

If we wanted to group these terms according to themes, a possible grouping would be the one shown in Figure 8, where the first cluster starting from the left includes words with uses of a political nature. It has been seen in previous sections that users use Twitter to make references to the past, to congratulate, etc., so terms of these topics are collected together with others of an economic nature, among others.

Secondly, a group of words used by politicians that revolve around the current government and its actions are collected, hence the appearance of words such as " psoe ", " politics", "employment ", "measures", etc.

The third cluster includes words whose relationship (in our context) is that of certain problems or fronts of the country, current and old. This explains why words like " covid ", "pandemic" or "crisis" are found in this cluster. Lastly, we see words like " sánchez ", "government", "today" and " spain ". Terms that appeared throughout the exploratory analysis with characteristics in common, such as words widely used by the politicians studied.

ASSOCIATION RULES

To generate the association rules in this study, minimum support and confidence values must be established. In the first place, using the a priori algorithm, the sets of frequent items are obtained. They will be considered frequent when they appear at least 22 times, which translates into a minimum support of 0.003, obtaining 2638 sets of frequent items. From these, the association rules will be generated whose confidence exceeds 85%, obtaining 51 association rules.

We can visualize the rules obtained in numerous ways. In Figure 9, the rulers are represented in a matrix by a dot, whose color depends on the lift (measure of a ruler's quality), and its size from the support. The elements on the horizontal axis represent the antecedent of the rule, and those on the vertical axis the consequent of it.

Another possible visualization is the one shown in Figure 10, where the rules are represented by vertices connected to elements by means of arrows. These vertices have a size depending on the support and a color depending on the lift. Antecedent items are connected with arrows pointing to the vertex representing the rule, and consequent items have an arrow pointing to itself.

SORTING WITH SUPPORT VECTOR MACHINES

Like any mathematical model, this one has to be trained. To do this, for each pair of politicians, 80% of the tweets are randomly chosen, and the remaining 20% will be used to train the model.

A cross-validation algorithm is executed to find the optimal cost value in each case and we apply the model with the corresponding number, obtaining the results shown in Table 3.

The table shows how the model, after having been trained and used with the optimal cost value in each case, has failed more times when classifying tweets according to authorship in the case of Arrimadas and Casado, which It may mean that the tweets of these two politicians are more similar than those of Sánchez and Iglesias.

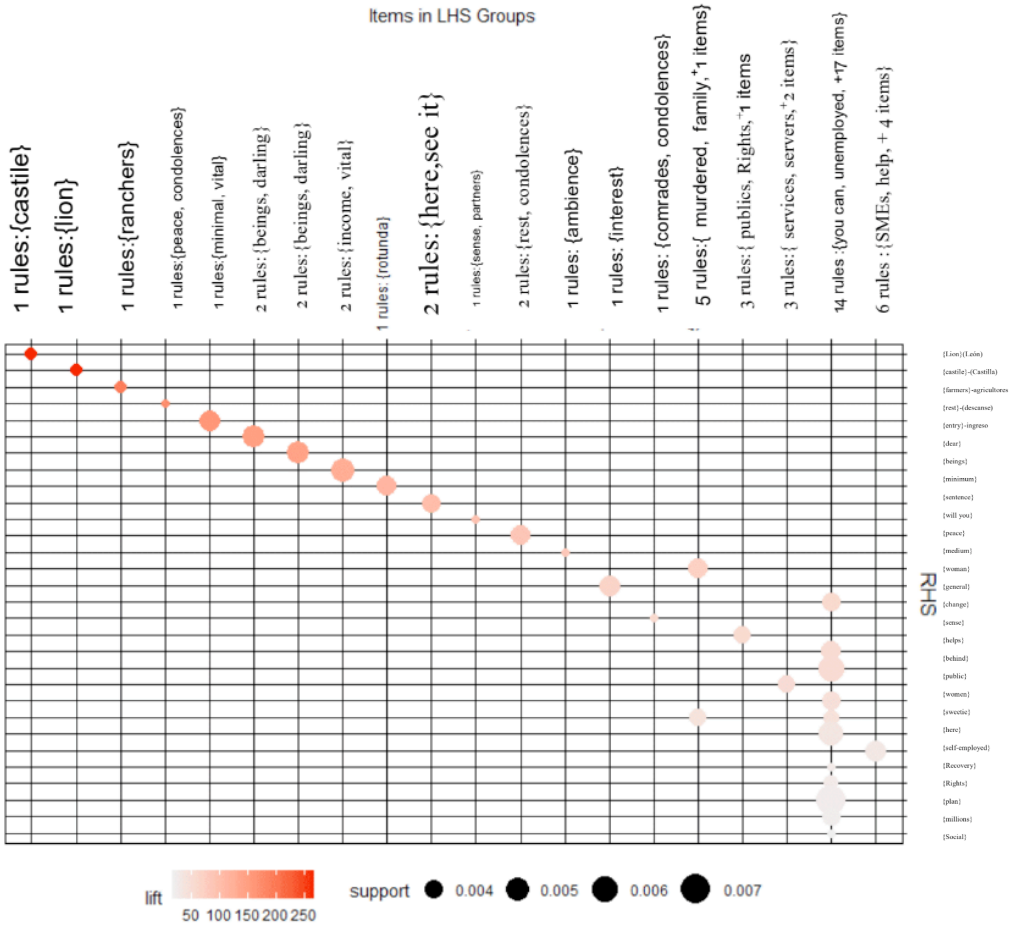


Figure 9. Visualization of the rules by means of a matrix.

FIGURE 10. Visualization of the rules by means of a graph.

Users	*Cost*	*% misclassifications*
Arrimadas and Married	7	10.44
Sanchez and Iglesias	10	8.72

TABLE 3. Results obtained after applying the model based on support vector machines to the data with the value of the cost obtained through cross-validation.

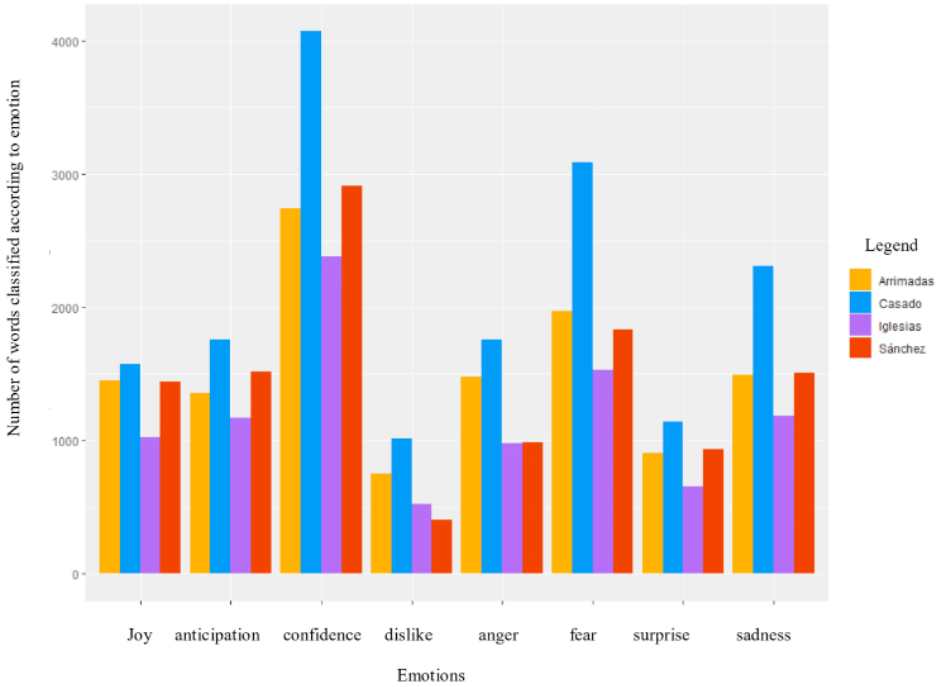


Figure 11. Number of words classified according to emotion for each politician.

	Sadness		Happiness	
	word	count	word	count
Pablo Casado	crisis	159	freedom	184
	pandemic	149	independence	47
	worse	64	To create	38
	let	53	special	36
	avoid	42	equality	36
Inés Arrimadas	crisis	110	freedom	156
	pandemic	86	equality	82
	struggle	47	worked	76
	violence	29	pride	44
	terrible	28	hug	36
Pablo Iglesias	crisis	104	worked	68
	commitment	54	Agreement	61
	violence	34	freedom	42
	pandemic	33	equality	33
	struggle	26	defending	26

Pedro Sánchez	pandemic	172	equality	90
	commitment	102	worked	84
	crisis	97	agreement	80
	emergency	70	advance	80
	struggle	55	progress	50

Figure 12. Words that have been classified the most times as happiness and sadness for each politician.

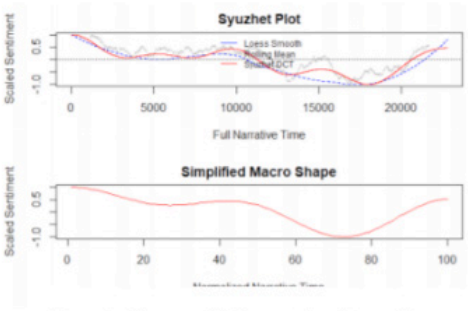
Cloud of emotions by Inés Arrimadas

Cloud of emotions of Pablo Casado

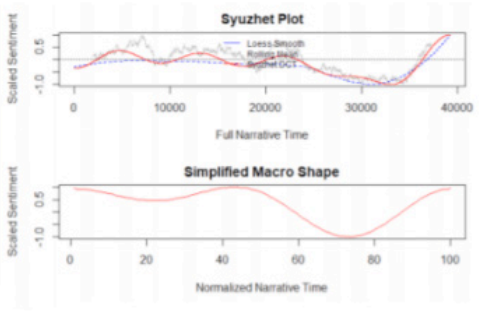
Cloud of emotions by Pablo Iglesias

Cloud of emotions by Pedro Sánchez

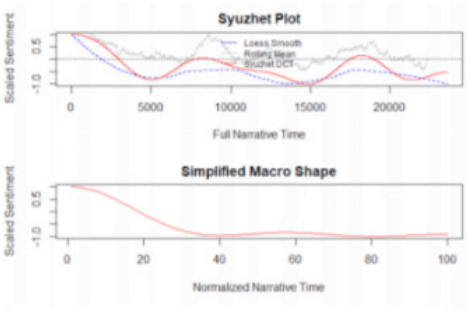
Figure 13. Cloud of emotions of politicians.



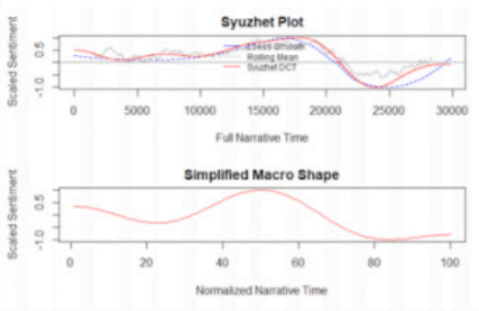
Evolution of Ines Arrimadas



Evolution of Pablo Casado



Evolution of Pablo Iglesias



Evolution of Pedro Sanchez

Figure 14. Evolution of the polarity of the tweets of all politicians.

SENTIMENT ANALYSIS

We can classify the words of the collected tweets according to eight sentiments, and visualize the results obtained for each politician in Figure 11. In the same way, we can collect which words have been classified more times as a certain emotion. Figure 12 collects these results according to each user for *happiness* and *sadness.*

These results can be represented for the eight emotions using the word cloud again, where the size of the word is determined by the number of times it has been classified as a certain emotion, and its color and position within it depends on the number of times it has been classified as a certain emotion. of the feeling on which it has been classified. We can see the clouds of emotions obtained in Figure 13.

It is interesting to see how the word pandemic appears in all the clouds of emotions, as well as government. The appearance of these terms makes sense because we study Spanish political figures and we have suffered the consequences of the well-known SARS-CoV-2 in recent years.

Finally, a study of the polarity of the tweets over time is made. In order to make a joint comparison, the tweets in the same time interval are studied, since we saw before that not all the tweets that were collected coincide in time. The results are shown in Figure 14. A value of polarity close to -1 means that the tweets were classified as negative, according to the feelings identified, and a value close to 1 the opposite, so that values close to 0 represent *neutrality.*

We can see that the polarity of Arrimadas, Casado and Sánchez is similar, while Iglesias begins with a positive value that falls in the first third of the interval studied, to remain negative.

DISCUSSION AND CONCLUSIONS

Within the results obtained in the exploratory analysis, it is interesting to see how, in the temporal distribution of tweets, Pablo Iglesias stands out for having peaks of activity on the social network. These coincide in time with the country's general elections, the Madrid elections and the "Dina" case with which the politician was related. In addition, when collecting the total number of words and the total number of different words used, Pablo Casado is the user that stands out from the rest, however, when collecting the words most used by politicians after eliminating the *stop words,* he has a word that clearly used with a much higher frequency than the rest of the words and users: Sánchez. It is an interesting fact because this politician presides over a party that is currently in opposition, and Pedro Sánchez is the current president of the government.

The results of applying hierarchical and non-hierarchical clustering show the heterogeneity of the data, since different results were obtained when finding an optimal number of clusters to group the data. Furthermore, the dendrogram shows many branches and there is no clear hierarchical grouping of the data.

The association rules show the co-occurrence of sets of terms related to politics, such as minimum vital income, environment, among others. They also show terms in which politicians refer to the past, or offer condolences in dramatic situations.

When classifying using support vector machines, we have seen how the mathematical model failed more times when classifying the tweets of Inés Arrimadas and Pablo Casado than those of Pedro Sánchez and Pablo Iglesias, this means that the tweets of the latter two politicians are more different from each other than those of Casado and Arrimadas.

This is an interesting result, since Sánchez and Iglesias came to form a coalition to form a government, while the other pair of politicians have no other relevant relationship.

The results obtained after applying sentiment analysis are interesting. It was reflected how trust is the predominant feeling among the eight in which the words were classified. This makes sense, since the users we studied represent political figures in the country, and normally they intend to give an image of trust to the people. If the object of study were other types of users, we would probably obtain very different results. In addition, when making an evolutionary study of the polarity of politicians, it is interesting to see how the behavior of this in Sánchez, Casado and Arrimadas is similar, while Pablo Iglesias's tweets start out being somewhat positive, but decline in the first third of the time studied, remaining negative until the end of the period collected.

ACKNOWLEDGMENTS/ SUPPORTS

This work has been partially funded by the projects:

- "FEDER University of Granada" (Spain). B-TIC-145-UGR18
- "I+D+I Project of the Junta de Andalucía" (Spain). P18-RT-1765
- COPKIT Project, through the European Union's Horizon 2020 Research and Innovation Program, under Grant 786687.

REFERENCES

Agrawal, R., Imieliński, T., & Swami, A. (1993). **Mining association rules between sets of items in large databases.** In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data* (pp. 207-216).

Bayram, F. (2010). **Ideology and political discourse analysis of Erdogan's political speech.** Annual review of education, communication & language sciences, 7.

Berzal, F. (2017). **Hierarchical clustering.** Available at: http://elvex.ugr.es/idbis/dm/slides/42% 20Clustering, last accessed on December 13, 2021.

Bholowalia, P. & Kumar, A. (2014). **EBK-means: A clustering technique based on elbow method and k-means in WSN.** *International Journal of Computer Applications, 105* (9).

Chilton, P. (2004). **Analyzing political discourse Theory and practice.** Routledge.

García Cambronero, C. and Gómez Moreno, I. (2006). **Learning Algorithms: KNN & KMEANS.** *Carlos III University of Madrid.*

Garimella, K., Weber, I., & De Choudhury, M. (2016). **Quote rts on twitter usage of the new feature for political discourse.** In Proceedings of the 8th ACM Conference on Web Science (pp. 200-204).

Grimaldi, D. (2019). **Can we analyze political discourse using Twitter Evidence from Spanish 2019 presidential election.** Social Network Analysis and Mining, 9(1), 1-9.

Hsu, C. L., Park, S. J., & Park, H. W. (2013). **Political discourse among key Twitter users: The case of Sejong city in South Korea.** Journal of Contemporary Eastern Asia, 12(1), 65-79.

Johnson, K. & Goldwasser, D. (2016). **Identifying stance by analyzing political discourse on twitter.** In Proceedings of the First Workshop on NLP and Computational Social Science (pp. 66-75).

Martinez, C.G. (2021). **Association rules.** Rpubs. Available at: https://rpubs.com/Cristina_Gil/Reglas_Asociacion last accessed on December 13, 2021

McNicholas, P.D., Murphy, T.B., & O'Regan, M. (2008). **Standardizing the lift of an association rule.** *Computational Statistics & Data Analysis, 52* (10), 4712-4721.

University of Cincinnati. **K-means Cluster Analysis. Github.** Available at: https://uc-r.github.io/kmeans_clustering, last accessed December 13, 2021

Rodrigo, J.A. (2017). **Support** Vector Machines (SVMs).

Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). **A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm.** *EURASIP Journal on Wireless Communications and Networking, 2021* (1), 1-16.

Tibshirani, R., Walther, G., & Hastie, T. (2001). **Estimating the number of clusters in a data set via the gap statistic.** *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63* (2), 411-423.

Van Dijk, T.A. (2002). **Political discourse and political cognition.** Politics as text and talk Analytic approaches to political discourse, 203, 203-237.

Zaki, MJ, Meira Jr, W., & Meira, W. (2014). **Data mining and analysis: fundamental concepts and algorithms.** Cambridge University Press.