IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 30, 2020, accepted October 12, 2020, date of publication October 19, 2020, date of current version November 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3032015

Known by Who We Follow: A Biclustering Application to Community Detection

JUAN M. COTELO, F. JAVIER ORTEGA, JOSÉ A. TROYANO, FERNANDO ENRÍQUEZ, AND FERMÍN L. CRUZ

Departamento de Lenguajes y Sistemas Informáticos, Universidad de Sevilla, 41012 Sevilla, Spain

Corresponding author: Fernando Enríquez (fenros@us.es)

This work was supported in part by the Spanish Ministry of Economy and Competitiveness through the Project Vision and Crowdsensing Technology for an Optimal Response in Physical-Security (VICTORY) under Grant TIN2017-82113-C2-1-R MINECO/FEDER R&D, UE, and in part by the Spanish Ministry of Science, Innovation and Universities through the Project Energy Efficiency and Performance of Data Centers by Smart Virtualization and Deep Learning Event Detection (REACT) under Grant RTI2018-098062-A-I00 MCIU/AEI/FEDER, UE.

ABSTRACT The detection of communities in social networks is a task with multiple applications both in research and in sectors such as marketing and politics among others. In this paper, we address the task of detecting on-line communities of Twitter users for a given domain. Our main contribution consists in modelling the community detection problem as a biclustering task. We have performed the experimentation with data from the political domain, a very dynamic area with a large number of interested users and a high availability of tweets. We have evaluated our proposal using both extrinsic and intrinsic methods, reaching very good results in both cases. We use the silhouette coefficient as intrinsic metric for clustering evaluation, and a classification task of political leanings of Twitter users as extrinsic evaluation. One of the most interesting conclusions of our experiments is the quality, from the point of view of predictive capacity in the classification task, of the communities identified with the proposed method. The information provided by communities detected through ''follow'' relationships has a predictive capacity comparable to that of the contents of tweets written by users. The results also show how detected communities can give insights about future events related to these communities that arise around social networks.

INDEX TERMS Twitter, community detection, biclustering, politics.

I. INTRODUCTION

Twitter can be seen as a thought shuttle: a channel that allows to launch ideas, in the form of short messages, to anyone who might be interested. At the same time, Twitter is an immense laboratory of contents and connections that allows to analyze opinions on any subject of interest and, in general, to verify sociological hypotheses.

There are a multitude of studies that try to take advantage of this potential in diverse areas such as branding [1], [2] e-commerce and trust management [3], political analysis [4], event detection [5] or market analysis [6].

In this paper we address the task of detecting on-line communities of users in Twitter for a given domain. We have evaluated our proposal with data from the political domain, a very dynamic area with a large number of interested users and a high availability of tweets. Detecting on-line communities can be a good complement to traditional sociological analysis approaches, because the potential dataset size is usually several orders of magnitude greater and is able to target other demographic sectors that surveys typically do not tackle.

Our proposed approach consists in applying an unsupervised process for uncovering ad-hoc communities whose users share similar stances, though no direct connection between them is required. There are many works about community detection in the political domain, most of them using Twitter as data source [7]–[9]. In this sense, our main contribution is the use of the spectral biclustering technique for the detection of communities. To our knowledge, there are no proposals that have used spectral biclustering for this task. Our main goal is to evaluate the usefulness of a biclustering technique to group similar users in a graph. The most important benefit of the use of spectral biclustering is that we have, for each user, a membership degree to each of the detected communities. Thus, we obtain a much richer and more complex profile of users than those resulting from approaches that include users in a single community.

The associate editor coordinating the review of this manuscript and approving it for publication was Hocine Cherifi.

There are many practical applications of the fuzzy characterization of users provided by the spectral biclustering algorithm. We have successfully employed it in the classification of users according to their political orientation, which has served us as an extrinsic evaluation of our proposal as well.

Using the membership degree of the users to each community as input data in order to train a classifier yields similar results (68% accuracy) to those of a classifier based on the texts written by those users. This shows that there is a valuable amount of information in membership degrees produced by spectral biclustering, and also that this information is as useful as the information contained in the texts written by the users.

The rest of the paper is organized as follows. Section II contains a brief review of related work on community detection. In section III we define the task we have used to evaluate our approach. In section IV, we describe our biclustering-based community detection method. We also provide, in section V, an evaluation of our experiments using both extrinsic and intrinsic metrics. Finally, in section VI, we summarize the conclusions and review our main contributions.

II. RELATED WORK

Community detection consists of identifying groups of similar vertices in a network, based on their structural properties. Many papers, books, and surveys, have been published about community detection over the past decades [10]–[12]. This is a sign of the difficulty and variability of the task, to which the saying *one size does not fit all* can certainly be applied. Techniques that detect communities perfectly for a network with certain characteristics (such as size or density), get very bad results when applied to networks of other nature. In addition, the versatility of graphs makes it possible to use very different approaches by applying ideas from disciplines as varied as physics, biology, applied mathematics or social sciences. We briefly describe some relevant works in the area of community detection.

An approach that uses HITS algorithm [13] for link-based analysis is proposed in [14]. It describes a blogosphere community detection system that performs a random-walk algorithm over a network graph in order to detect the communities formed by the most relevant blogs. This graph is built using the blogs as nodes, connecting them according to their citations and applying several iterations of the HITS algorithm to compute scores. In [15], the authors address the task of community detection in Twitter, by assuming that users with similar interests follow the same (topic-dependent) celebrities. They first define an agglomerative method using only topological links to detect communities by using the Clique Percolation Method [16] and the Infomap algorithm [17]. Then, an extension of this method is also proposed by using also implicit links, like mentions among users and retweets.

Another method, OSLOM (Order Statistics Local Optimization Method), is proposed in [18]. They define a degree-based metric intended to evaluate the significance of each node with respect to a given cluster and using the idea that nodes with a high significance score are likely to be part of that community (cluster). This OSLOM algorithm is also used in [19] for detecting topical Twitter communities from user lists. An interesting approach is developed in [20], which is based on LDA [21]. The authors propose TUCM (Topic User Community Model), a method intended to detect communities in social networks. The assumption is that users who interact in a network are likely to belong to a common community. They apply the LDA technique to the content being discussed by users, thus inferring the communities as if they were latent topics.

In [22], the authors propose a community detection method based on what they define as *Ground-truth communities*. These communities are explicitly defined communities whose members share the same role, affiliation or any other attribute. Under this notion, they analyze several structural definitions of what is a network community, categorize them into four groups and find two of these groups that achieve the best performance identifying communities. Another approach to the community detection task is introduced in [23], [24], adapting an algorithm originally intended to model dynamics of chemistry oscillators, to the detection of communities in social networks. Later improvements of this idea have proven their feasibility by adding negative relations between vertices or the addition of a Fourier term to the general equation to better capture the dynamics of the networks [25].

In [26], the authors propose a greedy algorithm that starts from a heuristic initial solution and then it iteratively improves the solution by applying two transformation steps, deconstruction and reconstruction, in order to avoid local solutions and make a wider exploration of the search space. There is also interesting research on topics directly related to the community detection task, such as optimization for very large networks [27], or specific evaluation metrics for the task [28]. Finally, recent approaches like [29]–[31] propose the use of embeddings to characterize graph communities, analyzing the relationship among graph embeddings, community detection and node embeddings.

III. TASK DEFINITION

Our main motivation is to evaluate the effectiveness of our proposal to detect Twitter communities for a given domain. We have chosen the political domain, a very dynamic area in which many users participate and therefore it is easy to obtain a large number of tweets. We will also use this data to illustrate through different graph visualizations the main steps of our proposal.

We built a collection of tweets written in Spanish between 20 December 2013 and 23 December 2013 referencing to one of the two main political parties at that time: *Partido Popular (PP)*, and *Partido Socialista Obrero Español (PSOE)*. It is unfeasible to retrieve and analyze the network composed by the whole subset of Spanish users interested in politics due to technical constraints; the size of this network would be huge and Twitter does limit the amount of information that can

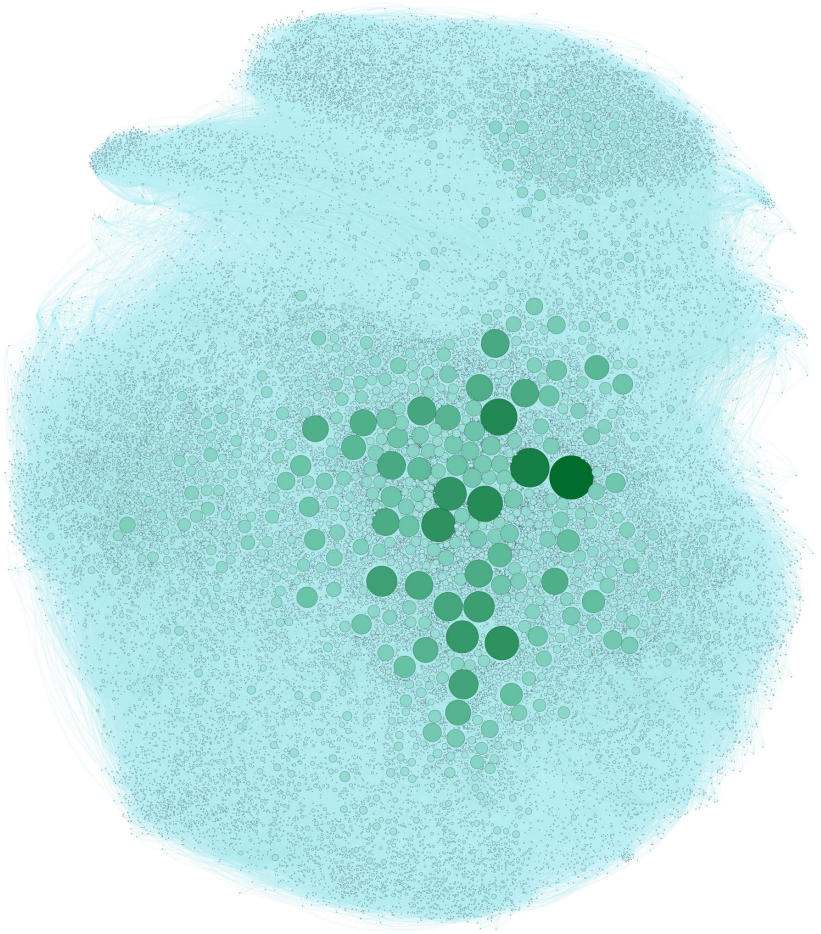


FIGURE 1. Direct friendship graph. Nodes are scaled and colored by their in-degree.

be retrieved. Therefore, we retrieved a sample of those tweets by using an adaptive query generation method that provides much more coverage than simple querying.

The specific method that we used to retrieve the Spanish tweets mentioned was the dynamic retrieval method explained in [32]. This retrieval method starts from a set of terms seeds, to define the domain of interest, and adapts the queries to the texts that it retrieves throughout the search period.

We retrieved a total of 20251 tweets. From this collection of tweets, we extract the users that directly appear within any tweet, being either the author of the tweet or any other user mentioned in the tweet. In addition to appearing users, we retrieve their lists of direct friends (who they follow) thus effectively obtaining their immediate neighborhoods. From the merged list of users and their friends, we compose a direct friendship graph in which nodes are Twitter users, and edges are *follow* relationships.

Figure 1 shows the resulting friendship graph, holding more than 500k users and more than 1,1M relations of friendship. Nodes are visually scaled and colored by in-degree (number of incoming relations of friendship). This friendship graph contains users that are varied in nature: mass media official accounts, politically active people, political party members and common people with no apparent relationship with these political organizations, covering most actors of the political situation.

IV. COMMUNITY DETECTION USING SPECTRAL BICLUSTERING

We propose an approach based on a biclustering to perform an unsupervised community detection. Biclustering is a data mining technique intended for simultaneous clustering of the rows and columns of a given matrix. In figure 2 we can see the general process of our proposal. Starting from the direct friendship graph, we perform several transformations until reaching the final distribution of nodes into communities. In the next subsections, we will detail the aspects of each of the steps of this process.

A. THE BIPARTITE GRAPH

From the perspective of the friendship graph construction process, we can identify two kinds of users: original users in the dataset, and new detected users *followed* by them. From the point of view of the community we can consider these followed users as *content creators*. Especially those with

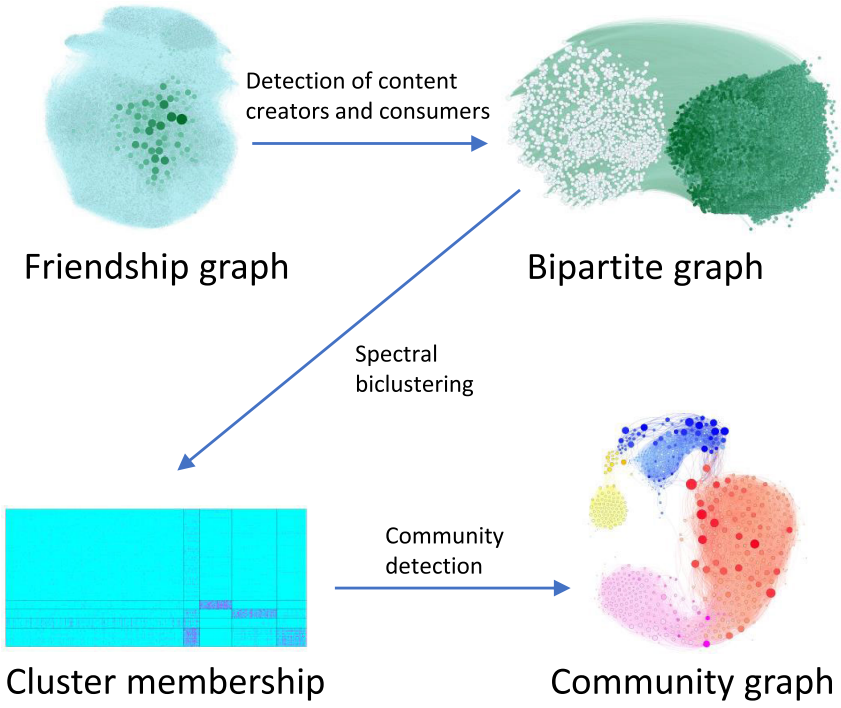


FIGURE 2. General process description.

many followers, who will be referents of the predominant ideology of the community. Similarly, from the community perspective, we can consider original users of the dataset as *content consumers* because through their opinions they have expressed interest in the target domain, and they *follow* the referent users.

If a user is both, content creator and content consumer, its node is split into two nodes in the bipartite graph. One node denotes his role as content creator, and the other one denotes his role as content consumer.

Given a friendship graph defined by *follow* relationships, we build a new bipartite graph to separate content consumers and content creators. Nodes representing users playing both roles have input and output arcs in the original network, but as we said before we split them into two separate nodes, one as a consumer user and one as a creator. The use of a bipartite graph results in a rectangular adjacency matrix, which allows the biclustering algorithm to group follower users (rows) and followed users (columns) differently.

Figure 3 shows the resulting bipartite graph, being the nodes colored by in-degree and laid out by its bipartite class.

Once the bipartite graph is built, we generate a matrix M where $M_{i,j} = 1$ if user i is followed by user j. This matrix is ready to be used as input for a biclustering technique to identify communities of users (both creators and consumers) that exhibit similar behavior.

B. BICLUSTERING TO REARRANGE THE ADJACENCY MATRIX

A bicluster is a subset of the original matrix whose rows exhibit similar behavior across its columns and vice versa. However, the exact definition of "similar behavior" depends on the specific bicluster algorithm used. There are many biclustering techniques [33]. For our experiments we have used spectral biclustering. It computes fuzzy clusters with several degrees of membership, which allows us to detect more flexible communities. Spectral biclustering was first introduced in [34] and its computational complexity is discussed in [35].

Spectral biclustering is a method based on the idea that the adjacency matrix rows and columns can be organized into $n \times m$ biclusters, where n and m are hyperparameters of the algorithm. Each row is assigned to m biclusters and each column is assigned to n biclusters through a membership degree. Seen from the perspective of the community detection, this degree of membership can be interpreted as the degree of affinity of the users to each community.

An intuitive way to show the behavior of the biclustering algorithm is through the concept of reordering. In this sense, the process of obtaining a bicluster partition consists of rearranging the rows and columns of a matrix so that similar values are grouped in the same areas of the matrix. Figure 4 clearly shows this reordering, applied to our community detection problem. In our case, we have an adjacency matrix whose non-zero values correspond to *follow* relationships of the bipartite graph. Before applying biclustering the non-zero values are distributed uniformly. After applying biclustering, very dense areas (groups of rows and columns) can be observed within the matrix. These dense areas correspond to communities where there is a high connection density between content creators and consumers.

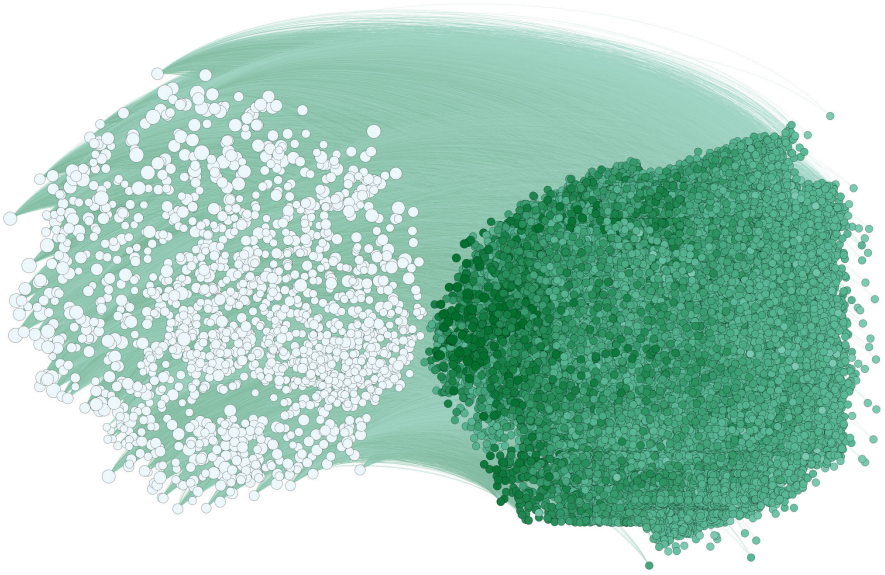


FIGURE 3. Bipartite graph built from the friendship graph. White nodes are content consumers, and green nodes are content creators. Content creators are colored by their in-degree.

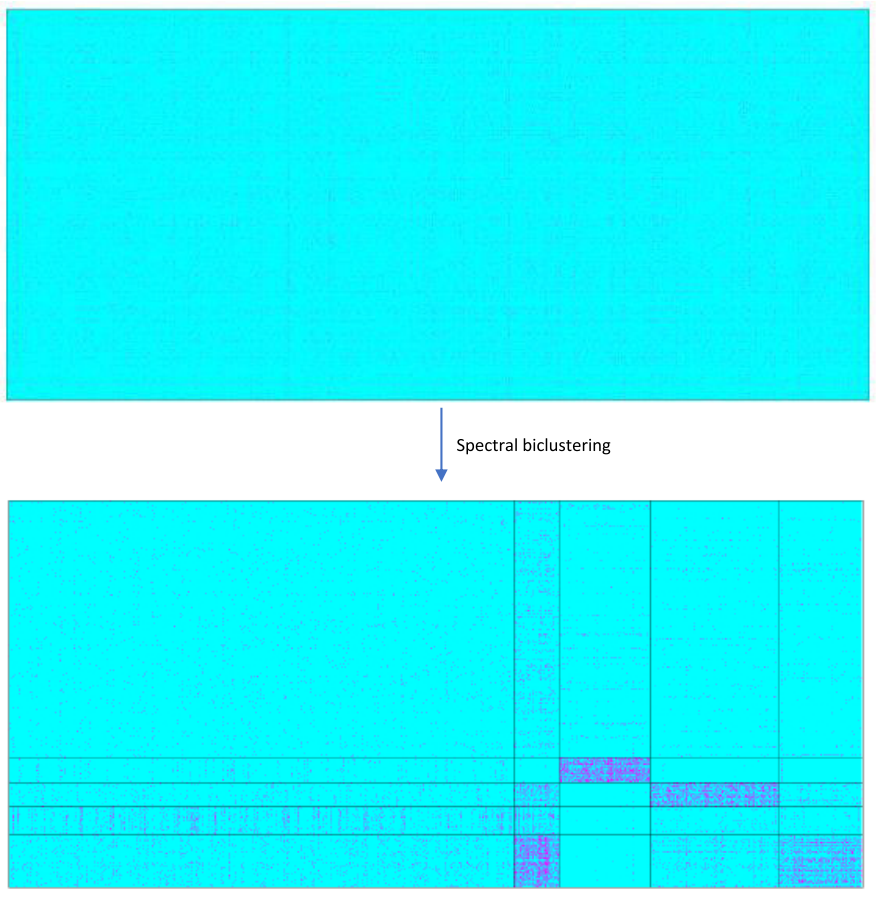


FIGURE 4. Adjacency matrix rearranged after applying spectral biclustering.

C. THE COMMUNITY GRAPH

Figure 5 shows the community graph generated with our biclustering approach for the political dataset. Nodes are scaled by the number of direct followers in the bipartite graph and colored according to the bicluster they belong to. The intra-cluster relevance is represented by the altering

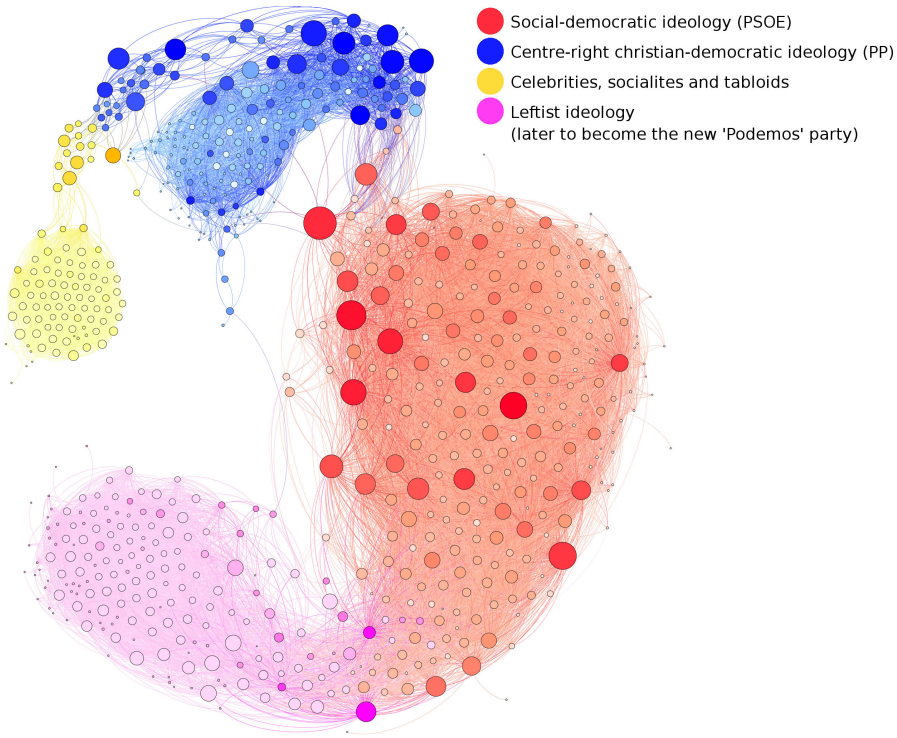


FIGURE 5. Community graph generated using spectral biclustering.

saturation of the color of each node, with more relevant nodes being those with higher saturation values. Edge weights are computed using cosine similarity over the membership vectors of two users. For each user, its membership vector is composed of its membership degrees to the different communities identified by the specral biclustering algorithm.

It is interesting to interpret the graph from the perspective of the political scenario from which the data has been extracted. Red nodes are related to the *PSOE* party or its social-democratic ideology, while purple nodes are related to leftist ideology but do not agree with the Social Democracy ideology. Blue nodes are related to the *PP* party, their centre-right christian-democratic ideology or any other right-wing ideology. Yellow nodes are celebrities, socialites and tabloids with low political relevance but highly followed by the users.

It is worth mentioning big nodes between blue and red communities are the Spanish general-interest daily newspapers *El Pais*, *El Mundo*, *20m* and *ABC*, which reach from socialist to centre-right ideologies, while more right-wing and left-wing papers lie deep within their respective clusters.

Another interesting issue that arises from these results is the interpretation that we can make of the purple nodes knowing what happened months after the dates to which the dataset tweets correspond. After reviewing a sample of the tweets related to that part of the graph, we noticed how they reflect the debate generated around a social protest movement that had recently emerged in Spain, in which a large number of citizens were against traditional parties and which culminated in the founding of a new political party called *Podemos*. It is currently one of the main political parties in Spain, with an ideology located to the left of *PSOE*. Although at the time of our analysis *Podemos* did not yet exist, it can be seen in the results as one of the communities detected along with other more predictable, related to traditional parties. In this sense, we believe that an algorithm such as the one we are proposing can be very useful in different contexts due to its capacity to predict future events related to these communities that arise and grow around social movements.

V. EVALUATION

In this section we present the results of the experiments which have been performed using the dataset described in section III.

Subsections V-B and V-C provide a comparison between our proposed approach based on spectral biclustering and the baseline. The evaluation described in subsection V-B is based on an intrinsic metric, while we use an extrinsic task as an indirect measure in subsection V-C. Finally, in subsection V-D a qualitative analysis is conducted upon the communities extracted by the spectral biclustering approach.

A. THE BASELINE

A well-known and efficient modularity maximization method has been selected a as a baseline for its ability of processing very large networks, and its good performance detecting large-scale communities. *Modularity* [36] is a benefit function designed to measure the quality of a division of a network

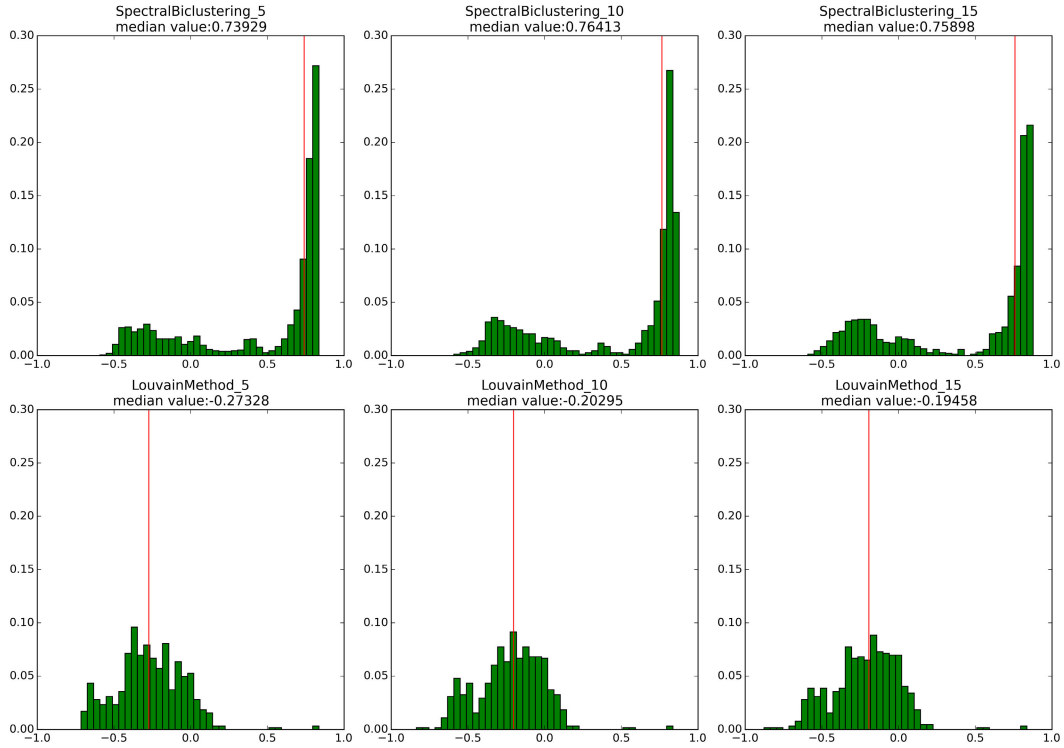


FIGURE 6. Distribution of silhouette cluster performance.

into communities where a high modularity partition exhibits dense intra-community connections and sparse connections between communities.

Nevertheless, brute force search over all possible partitions is intractable in most cases because the search space grows exponentially with the size of the network. Therefore, most algorithms are based on approximate methods such as greedy algorithms, simulated annealing, quasi-Newton methods or spectral optimization, giving approximate solutions to the modularity maximization problem with different degrees of speed and accuracy.

In this work, the popular *Louvain method* [37] is used for approximating the modularity function. The Louvain method yields better results (both in time and modularity categories) than similar methods such as those described in [38] and [39].

However, any modularity-based approach suffers from the resolution limit, failing to resolve communities smaller than some scale, depending on the size of the network. In summary, they accurately detect large-scale communities but they are not appropriate for retrieving microscale communities.

B. SILHOUETTE COEFFICIENT

When there is no labeled information available, evaluation must be performed using the model itself. The *Silhouette Coefficient* [40] is a well known method for evaluating and validating clusters of data, estimating how well each object lies within its assigned cluster.

Given a single sample, its mean intra-cluster distance a and its mean nearest-cluster distance b are computed using a specific metric distance, such as the euclidean distance, the Manhattan distance or the cosine distance. The silhouette coefficient s for that given sample is $s = \frac{b-a}{max(a,b)}$ such that $-1 \leq s \leq 1$. To evaluate the model, we compute the coefficient for every sample of the data set and we apply a central tendency measure such as the mean or the median for scoring the model. Silhouette coefficient ranges between -1 and 1, meaning -1 an incorrect clustering and 1 a tight clustering (dense and well separated). Values around 0 indicate overlapping clusters.

Figure 6 compares the distributions of the silhouette scores (over the content consumer groups) obtained by the Louvain method and by our proposal. We show the values obtained for 5, 10 and 15 clusters. Median values of silhouette scores are shown as red vertical lines.

We observe that silhouette coefficients for the Louvain method are well distributed around the median value and mostly negative. Increasing the number of clusters slightly improves the median value but the deviation increases and the results stay stable. The medians of the coefficients for our biclustering-based method are quite better and most of the coefficients are higher than 0.5. It is interesting that the distribution is clearly bi-modal, negative coefficients are less than 30% of total coefficients and are grouped around the -0.3 value. This indicates mostly dense clustering with a

little of overlapping; varying the number of clusters does not change the results significantly.

Regardless the selected number of clusters, it is clear that the spectral biclustering approach performs remarkably better than the Louvain method approach.

C. EXTRINSIC EVALUATION

Another way of analyzing the performance of an unsupervised task is by measuring how well it contributes to another different task. To do this, we have defined a political opinion classification task on Spanish Tweets. We selected a sample of 3000 high quality tweets from the collection previously mentioned in section III and we manually tagged them to build a dataset for this classification task. Tweets from this collection were labeled as positive, negative or neutral with respect to the opinions they express about PP and PSOE. The classification task is therefore defined on nine categories, corresponding to the Cartesian product of the three stances for the two parties.

TABLE 1. Percentages for each of the nine possible political opinions in the labeled dataset.

	PSOE (+)	PSOE (-)	PSOE (neutral)
PP (+)	0.0%	1.0%	1.3%
PP (-)	1.0%	4.0%	46.1%
PP (neutral)	2.5%	18.8%	25.0%

The percentage of the different political opinions in the labeled dataset can be seen in table 1. Most of the tweets fit into three of the nine possible classes. Due to the low representation of the remaining classes we defined a simplified version of the task along with the original nine-class problem. In this reduced problem we only take in consideration tweets of the three major classes.

We use our community detection approach to calculate a feature vector for each tweet, consisting in the membership degrees of the tweet author to each of the communities detected.

Results are compared to those obtained with a standard Bag-of-Words model. This model is a simplified representation used in natural language processing and information retrieval which represents text documents as the *multiset* of its words. We also included the dummy stratified random classifier for comparison purposes only.

According to this extrinsic evaluation method, table 2 shows that the spectral biclustering approach yields better results than the Louvain method. It is very interesting that the performance of the features extracted from the communities, that is purely based on topological structure, is comparable to the results obtained using a bag of words model for the textual content.

TABLE 2. Cross-validated accuracy for different feature models using a SVM classifier.

Feature Model	Full Problem	Reduced Problem
Dummy Random stratified	31.3%	36.3%
Bag-of-Words	61.9%	68.3%
Louvain Method	50.0%	56.0%
Spectral Biclustering	60.1%	68.7%

D. QUALITATIVE ANALYSIS OF POLITICALLY RELEVANT USERS

In the previous subsections, we have analyzed the performance of our approach and it has been compared with the modularity maximization approach, showing that our approach is significantly better. Conversely, in this subsection, we provide a qualitative analysis intended to measure the ability of our proposal of identifying relevant users within the detected communities.

Given the same community model computed by our spectral biclustering approach in subsection V-B, we have picked those communities containing the official Twitter accounts of PP and PSOE. From both communities, we have selected the top 10 most relevant users that were neither an official media account nor anyone directly affiliated with the corresponding political party. The intra-cluster relevance is given by the biclustering algorithm.

We were specifically searching for individual users that may be related to the political party ideology without being a wide-known member of the party or directly related to any newsgroup. In this way we can measure whether our method is able to identify and cluster ordinary people with the same political ideology, that never interacted between them and do not publicly profess any political affiliation.

For each user, we retrieved the 100 most recent tweets from his timeline whose content were directly related to politics; those tweets were manually tagged the same way that the ones for the task described in section V-C. After that, we computed the affinity of each user to the major political party of their respective cluster group, being the affinity measure defined as follows.

Definition 1: Given the user u, we define the following provisions:

- Let x be one the major political parties. We define $\mathrm{pos}_u(x)$ and $\mathrm{neg}_u(x)$ as the number of times that the user u expresses an opinion regarding the party x in any of his tweets, being positive or negative respectively.
- Being PP and $PSOE$ the major political parties in our study, we define the party "support" measures $\mathrm{sc}_u(PP) = \mathrm{pos}_u(PP) + \mathrm{neg}_u(PSOE)$ and $\mathrm{sc}_u(PSOE) = \mathrm{pos}_u(PSOE) + \mathrm{neg}_u(PP)$.

We define the following affinity scores for user u regarding both political parties as:

$$\underset{u}{\mathrm{aff}}(PP) = \frac{\mathrm{sc}_u(PP)}{\mathrm{sc}_u(PP) + \mathrm{sc}_u(PSOE)}$$

$$\underset{u}{\mathrm{aff}}(PSOE) = \frac{\mathrm{sc}_u(PSOE)}{\mathrm{sc}_u(PP) + \mathrm{sc}_u(PSOE)}$$

Table 3 shows the intra-cluster relevance values and the affinity scores for the selected users. It is worth noticing that

TABLE 3. Affinity scores and intra-cluster relevance values of the top 10 politically relevant users.

User	Affinity	Relevance
PSOE_USER#1	100.0%	68.2%
PSOE_USER#2	100.0%	65.5%
PP_USER#1	97.3%	71.6%
PSOE_USER#3	95.4%	61.0%
PP_USER#2	90.9%	74.3%
PSOE_USER#4	87.5%	54.6%
PSOE_USER#5	85.0%	66.5%
PP_USER#3	83.3%	65.6%
PSOE_USER#6	69.8%	70.6%
PP_USER#4	66.6%	64.8%

users have been anonymized before the manual annotation process for privacy reasons. We observe that users yield high affinity scores (regarding the major party of their cluster) while also exhibiting high intra-cluster relevance values. This means that our approach is able to retrieve highly affine community members despite the fact that the large majority of those users are neither affiliated to any political party nor related to any politically related media.

VI. CONCLUSION

We have presented an original approach for detecting communities of interest in Twitter for a given domain. By modeling the problem as a biclustering task and applying the spectral biclustering technique, we achieve an effective way of addressing the community detection task in large-scale networks. We have applied our approach to the political domain, by extracting underlying communities of Spanish users within the political situation in Spain. In order to test our proposal, we have compared it to the Louvain method, a popular modularity maximization approach. From both intrinsic and extrinsic evaluation methods proposed in this paper, we have observed that the quality of the communities detected with spectral biclustering approach is superior to the ones generated by the baseline. Not only performs better, but it gives richer models that allow overlapping communities, fuzzy membership, smaller communities and provides additional information about the intra-cluster relevance of their members. Another interesting observation is the quality, from the point of view of predictive capacity, of the communities identified with the proposed method. In a task of automatic classification of political leanings of Twitter users, the information provided by communities detected through "follow" relations has a predictive capacity comparable to that of the contents of the tweets written by users. The results include communities related to traditional Spanish political parties, but also a community related to a social movement that led to the creation of a new political party, so it is also worth noting the possibility of using it to predict future events related to these communities that arise around social movements. In addition to the proposed evaluation methods, we performed a qualitative analysis of politically relevant users by making use of the community model provided by our approach. Using the intra-cluster relevance information given by the model, we assessed that our method was able to identify and properly cluster ordinary people sharing the same political ideology even when those users never interacted among them and do not publicly profess any political affiliation.

Our experiments have shown the usefulness of the biclustering technique to detect communities. In the future we plan to continue our research in different directions. We want to evaluate our proposal using data of several domains, and also compare the results with another community detection methods. We are particularly interested in identifying antagonistic communities such as provaccine/antivaccine or proabortion/antiabortion communities. We are also interested in exploring other tasks that may benefit from the membership degrees provided by the spectral biclustering algorithm. The good results obtained in the political stance classification task encourage us to look for other applications where this kind of information is useful.

REFERENCES

[1] A. Culotta and J. Cutler, "Mining brand perceptions from Twitter social networks," *Marketing Sci.*, vol. 35, no. 3, pp. 343–362, May 2016.
[2] D. Zimbra, M. Ghiassi, and S. Lee, "Brand-related Twitter sentiment analysis using feature engineering and the dynamic architecture for artificial neural networks," in *Proc. 49th Hawaii Int. Conf. Syst. Sci. (HICSS)*, Jan. 2016, pp. 1930–1938.
[3] H.-J. Li, Q. Wang, S. Liu, and J. Hu, "Exploring the trust management mechanism in self-organizing complex network based on game theory," *Phys. A, Stat. Mech. Appl.*, vol. 542, Mar. 2020, Art. no. 123514.
[4] P. Barberá and G. Rivero, "Understanding the political representativeness of Twitter users," *Social Sci. Comput. Rev.*, vol. 33, no. 6, pp. 712–729, Dec. 2015.
[5] E. D'Andrea, P. Ducange, B. Lazzerini, and F. Marcelloni, "Real-time detection of traffic from Twitter stream analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 2269–2283, Aug. 2015.
[6] V. S. Pagolu, K. N. Reddy, G. Panda, and B. Majhi, "Sentiment analysis of Twitter data for predicting stock market movements," in *Proc. Int. Conf. Signal Process., Commun., Power Embedded Syst. (SCOPES)*, Oct. 2016, pp. 1345–1350.
[7] M. Ozer, N. Kim, and H. Davulcu, "Community detection in political Twitter networks using nonnegative matrix factorization methods," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2016, pp. 81–88.
[8] O. Fraisier, G. Cabanac, Y. Pitarch, R. Besançon, and M. Boughanem, "Uncovering like-minded political communities on Twitter," in *Proc. ACM SIGIR Int. Conf. Theory Inf. Retr.*, Oct. 2017, pp. 261–264.
[9] L. Belcastro, R. Cantini, F. Marozzo, D. Talia, and P. Trunfio, "Learning political polarization on social media using neural networks," *IEEE Access*, vol. 8, pp. 47177–47187, 2020.
[10] B. S. Khan and M. A. Niazi, "Network community detection: A review and visual survey," 2017, *arXiv:1708.00977*. [Online]. Available: http://arxiv.org/abs/1708.00977
[11] M. T. Schaub, J.-C. Delvenne, M. Rosvall, and R. Lambiotte, "The many facets of community detection in complex networks," *Appl. Netw. Sci.*, vol. 2, no. 1, p. 4, Dec. 2017.
[12] H. Cherifi, G. Palla, B. K. Szymanski, and X. Lu, "On community structure in complex networks: Challenges and opportunities," *Appl. Netw. Sci.*, vol. 4, no. 1, pp. 1–35, Dec. 2019.
[13] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 668–677, 1999.
[14] B. L. Tseng, J. Tatemura, and Y. Wu, "Tomographic clustering to visualize blog communities as mountain views," in *Proc. Workshop Weblogging Ecosystem (WWW)*, 2005, pp. 1–6.

[15] K. H. Lim and A. Datta, "Tweets beget propinquity: Detecting highly interactive communities on Twitter using tweeting links," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, Dec. 2012, pp. 214–221. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6511887

[16] I. Derényi, G. Palla, and T. Vicsek, "Clique percolation in random networks," *Phys. Rev. Lett.*, vol. 94, no. 16, Apr. 2005, Art. no. 160202. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevLett.94.160202

[17] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 4, pp. 1123–1118, Jan. 2008. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2234100&tool=pmcentrez&rendertype=abstract

[18] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, "Finding statistically significant communities in networks," *PLoS ONE*, vol. 6, p. 24, Dec. 2010. [Online]. Available: http://arxiv.org/abs/1012.2363

[19] D. Greene, D. O'Callaghan, and P. Cunningham, "Identifying topical Twitter communities via user list aggregation," in *Proc. 2nd Int. Workshop Mining Communities People Recommenders (COMMPER) ECML*, 2012, pp. 41–48.

[20] M. Sachan, D. Contractor, T. A. Faruquie, and L. V. Subramaniam, "Using content and interactions for discovering communities in social networks," in *Proc. 21st Int. Conf. World Wide Web (WWW)*, New York, NY, USA, 2012, p. 331. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2187836.2187882

[21] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, p. 2003, Jan. 2002.

[22] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," *Knowl. Inf. Syst.*, vol. 42, no. 1, pp. 181–213, Jan. 2015, doi: 10.1007/s10115-013-0693-z.

[23] J. Wu, L. Jiao, C. Jin, F. Liu, M. Gong, R. Shang, and W. Chen, "Overlapping community detection via network dynamics," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 85, no. 1, Jan. 2012, Art. no. 016115. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.85.016115

[24] M.-Y. Zhou, Z. Zhuo, S.-M. Cai, and Z. Fu, "Community structure revealed by phase locking," *Chaos, Interdiscipl. J. Nonlinear Sci.*, vol. 24, no. 3, Sep. 2014, Art. no. 033128, doi: 10.1063/1.4894764.

[25] D. M. N. Maia, J. E. M. de Oliveira, M. G. Quiles, and E. E. N. Macau, "Community detection in complex networks via adapted kuramoto dynamics," *Commun. Nonlinear Sci. Numer. Simul.*, vol. 53, pp. 130–141, Dec. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1007570417301545

[26] J. Sánchez-Oro and A. Duarte, "Iterated Greedy algorithm for performing community detection in social networks," *Future Gener. Comput. Syst.*, vol. 88, pp. 785–791, Nov. 2018.

[27] G. Rossetti, "ANGEL: Efficient, and effective, node-centric community discovery in static and dynamic networks," *Appl. Netw. Sci.*, vol. 5, no. 1, pp. 1–23, Dec. 2020.

[28] M. Jebabli, H. Cherifi, C. Cherifi, and A. Hamouda, "Community detection algorithm evaluation with ground-truth data," *Phys. A, Stat. Mech. Appl.*, vol. 492, pp. 651–706, Feb. 2018.

[29] Y. Zhang, T. Lyu, and Y. Zhang, "Cosine: Community-preserving social network embedding from information diffusion cascades," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[30] C. Tu, X. Zeng, H. Wang, Z. Zhang, Z. Liu, M. Sun, B. Zhang, and L. Lin, "A unified framework for community detection and network representation learning," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 6, pp. 1051–1065, Jun. 2019.

[31] Z. Zhao, X. Zhang, H. Zhou, C. Li, M. Gong, and Y. Wang, "HetNERec: Heterogeneous network embedding based recommendation," *Knowl.-Based Syst.*, vol. 204, Sep. 2020, Art. no. 106218.

[32] J. M. Cotelo, F. L. Cruz, and J. A. Troyano, "Dynamic topic-related tweet retrieval," *J. Assoc. Inf. Sci. Technol.*, vol. 65, no. 3, pp. 513–523, Mar. 2014.

[33] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: A survey," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 1, no. 1, pp. 24–45, Jan. 2004.

[34] Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein, "Spectral biclustering of microarray data: Coclustering genes and conditions," *Genome Res.*, vol. 13, no. 4, pp. 703–716, Apr. 2003.

[35] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler, "A systematic comparison and evaluation of biclustering methods for gene expression data," *Bioinformatics*, vol. 22, no. 9, pp. 1122–1129, May 2006.

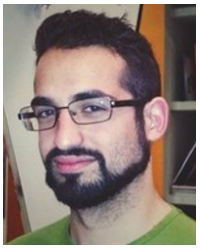
[36] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 2, Feb. 2004, Art. no. 026113. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevE.69.026113

[37] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech., Theory Exp.*, vol. 2008, no. 10, Oct. 2008, Art. no. P10008.

[38] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 70, no. 6, Dec. 2004, Art. no. 066111.

[39] K. Wakita and T. Tsurumi, "Finding community structure in mega-scale social networks," in *Proc. 16th Int. Conf. World Wide Web*, 2007, pp. 1275–1276.

[40] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0377042787901257



JUAN M. COTELO received the Ph.D. degree in computer science from the University of Seville, with which he currently maintains a close collaborative research relationship. He has published several scientific articles about social media analysis, natural language processing, and integration of structured and unstructured information. His research interests include solving natural language processing tasks and language design problems.

F. JAVIER ORTEGA received the Ph.D. degree in computer science. He is currently a Lecturer with the University of Seville. His most recent works are related to opinion mining, social network analysis, web spam detection, and trust and reputation analysis. His main research interests include graph-based algorithms and its application to social network analysis and natural language processing.

JOSÉ A. TROYANO received the Ph.D. degree in computer science from the University of Seville. He is currently a Lecturer with the Department of Languages and Computer Systems, University of Seville. He has published more than 70 scientific articles, 15 of which published in impact journals. He has participated in 18 research projects, being a principal investigator (IP) in four of them. He has directed seven Ph.D. theses, two of which received the Annual Prize for the Best Doctoral Thesis at national level from the Spanish Society for Natural Language Processing. His research interests include natural language processing, machine learning, and social network analysis.

FERNANDO ENRÍQUEZ received the degree in computer engineering from the University of Seville, Spain. He has been an Associate Professor with the University of Seville since February 2012. He was collaborating in research projects on Natural Language Processing in 2003. In 2011, he presented the Ph.D. Dissertation in combination of classifiers for the improvement of different tasks, such as morphological analysis or named entity recognition. He has participated in many research initiatives related to intelligent information processing and machine learning techniques.

FERMÍN L. CRUZ received the Ph.D. degree in computer engineering. He is currently an Associate Professor with the University of Seville. He worked on other issues related to social network analysis and machine learning, such as graph-based labeling and topic-related retrieval in Twitter. His current research interest includes computational linguistics with an emphasis on sentiment analysis.