Hindawi
Complexity
Volume 2019, Article ID 1238780, 13 pages
https://doi.org/10.1155/2019/1238780

Research Article

# Detection of Jihadism in Social Networks Using Big Data Techniques Supported by Graphs and Fuzzy Clustering

**Cristina Sánchez-Rebollo,[1] Cristina Puente,[1] Rafael Palacios,[1] Claudia Piriz,[2] Juan P. Fuentes,[2] and Javier Jarauta[2]**

[1] Universidad Pontificia Comillas, 28015 Madrid, Spain
[2] Grupo SIA, Alcorcón, 28922 Madrid, Spain

Correspondence should be addressed to Cristina Puente; cristina.puente@comillas.edu

Received 7 December 2018; Accepted 11 February 2019; Published 10 March 2019

Guest Editor: Alicja Krzemień

problems
Goal of terrorist
influencing people
& recruiting new members in Real time.

Goal to solve
Analyze message acc
• level of activity
• Ability to influence other users
• the contents of their messages

Social networks are being used by terrorist organizations to distribute messages with the intention of influencing people and recruiting new members. The research presented in this paper focuses on the analysis of Twitter messages to detect the leaders orchestrating terrorist networks and their followers. A big data architecture is proposed to analyze messages in real time in order to classify users according to different parameters like level of activity, the ability to influence other users, and the contents of their messages. Graphs have been used to analyze how the messages propagate through the network, and this involves a study of the followers based on retweets and general impact on other users. Then, fuzzy clustering techniques were used to classify users in profiles, with the advantage over other classifications techniques of providing a probability for each profile instead of a binary categorization. Algorithms were tested using public database from Kaggle and other Twitter extraction techniques. The resulting profiles detected automatically by the system were manually analyzed, and the parameters that describe each profile correspond to the type of information that any expert may expect. Future applications are not limited to detecting terrorist activism. Human resources departments can apply the power of profile identification to automatically classify candidates, security teams can detect undesirable clients in the financial or insurance sectors, and immigration officers can extract additional insights with these techniques.

Why manually
① influencer
may

Role of Graphs
• Analyze how the msg propagate through the network
• which involves a study of the followers based on retweets & general impact on other us

Roles of fuzzy clustering technique
• use to classify users in profiles. {probability for each profile instead of binary categorization}

ers.
Classification of elements into two groups on the basis classification rule

## 1. Introduction

Social networks are playing a very important role in the way people think. When accurately targeted, repeated messages can reinforce political ideas or even flip the way of thinking of the most indecisive. In this regard, Jihadism has been identified as one of the movements that relies the most on social networks to spread propaganda and try to influence the public opinion. The Madrid bombings of 2004 [1] are used as a case study, where they analyze grassroot jihadist networks and how terrorist organizations use collective action from local level to cause enormous impact.

Database
① Kaggle &
② Twitter data extraction technique

Social networks are also used by terrorist organizations as a tool for recruiting new members. Sentiment analysis to detect radicalization has been applied to social networks in the past as an evolution of previous analysis that were traditionally focused on websites and forums [2]. The problem of nodes that play an important role as influencers or that spread propaganda and the way in which it is propagated is a growing area of research [3–5].

Application future
① Human resource dep
② security Team
③ immigration officials

A very challenging part of the analysis presented in this paper is how to measure the impact of each user in the network, as it depends on the volume of tweets (activity) combined with the number of followers but is also amplified by the number of retweets. For this purpose, a deep analysis is carried out using graphs. General theory of networks and graphs, in particular, have been used for social network analysis (SNA) as one of the most relevant tools [6].

Labelling users as influencers, followers, or neutrals is also very difficult and false positives or false negatives of a standard classifier may yield dramatic consequences. Missing data can be corrected by applying genetic algorithms able to predict the absent text, as in the case of missing answers in questionnaires [7]. However, in the current research the

For this analysis the Python library nltk and Vader Lexicon were used, specialized in natural language processing.

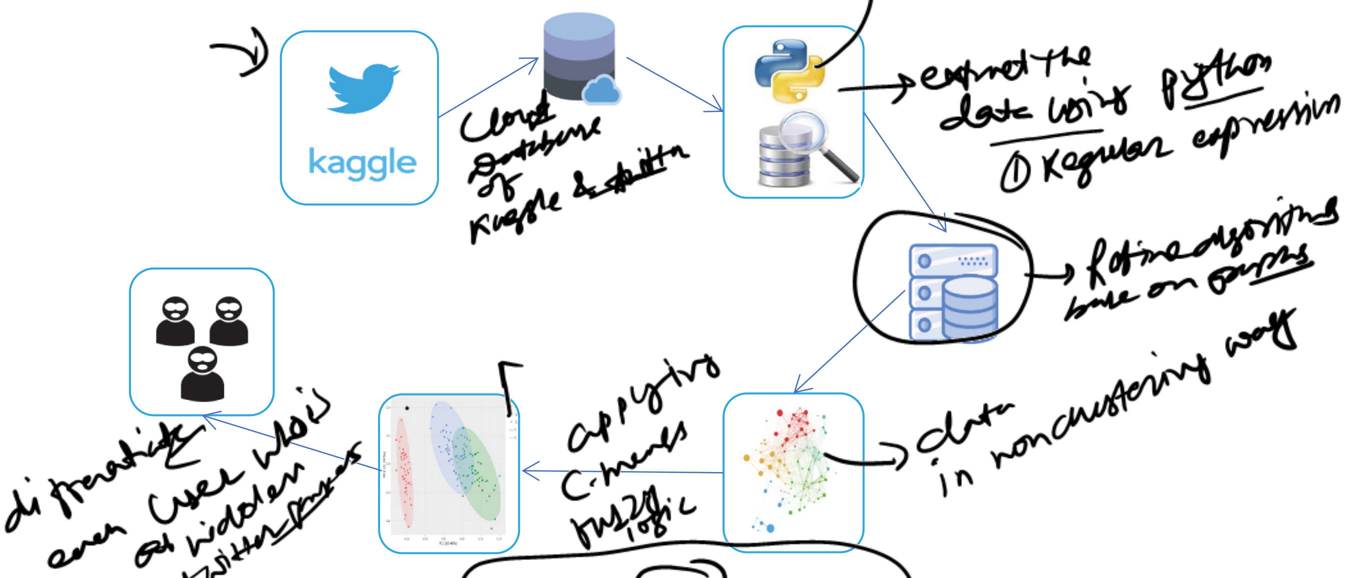
kaggle
Cloud database of Kaggle & twitter
extract the data using python
① Regular expression
Refine algorithms base on graphs
data in nonclustering way
applying C-means fuzzy logic
differentiate even user who's are hidden twitter profiles

Figure 1: Stages of the procedure to select hidden Twitter profiles.

problem is more related to ambiguity as a result of unreliable data or contradictory terms in the messages. In this case advanced text mining or natural language processing techniques would be appropriate [8–10]. Nevertheless, due to the fact that the messages to be analyzed are mostly translations from Arabic language and dialects to English, it was decided not to put a big amount of effort into natural language, because of the risk of ending up modeling the translation process more than the original meaning of the messages.

For the purpose of assigning profiles to the users, the proposed methodology utilizes fuzzy clustering techniques that provide probability of classification for each possible profile. Fuzzy clustering has been successfully applied in semisupervised environments [11] in combination with the classic k-means clustering method [12] and more specifically to detect malicious components [13]. In this paper the fuzzy clustering method takes as an input the results obtained from the graph analysis, along with some characteristics directly extracted from the social network.

## 2. Description of the Methodology: Architecture Based on Graphs and Fuzzy Clustering

*2.1. Big Data Architecture.* A big data architecture is proposed with the goal of monitoring Twitter in real time being able to predict threats either by detecting changes in the profiles or by detecting changes in the level of activity. The system can retrain itself to update profiles and classifications patterns, while maintaining its detection capabilities. A big data approach is very suitable for this kind of real-time analysis, especially on social networks such as Twitter in which messages are generated continuously and the system must collect, analyze, and archive-or-discard them [14, 15].

The proposed implementation was simulated using Kaggle's databases plus Twitter extraction API for demonstration purposes and to refine the algorithms based on graphs and fuzzy clustering (as shown on Figure 1).

*2.2. Fuzzy Architecture to Isolate Suspicious Profiles.* Using the previous works as inspiration and [13], we have designed a system capable of locating those profiles hidden at first sight but prone to modify their behaviour based on the influence received. In order to do so, we have considered a set of tweets as the source of information to measure the impact of an influence user in others.

Our process has followed five steps. First, the information was acquired using Kaggle's database and Twitter user's accounts by extracting their tweets. Then, that information was filtered to select the most and the less active users (understanding active users as those that both generated information or actively retweeted information from others). Next, we established those parameters with the potential to differentiate users, like sentiment analysis of the messages, users followed, and some others. Next the network graph was created, using the relationships among users to apply centrality measures in order to obtain new parameters that will serve as additional inputs to the fuzzy clustering process. Finally, the system will show those hidden users with an undefined profile, susceptible to be traced in the future. Figure 1 represents the interaction of these five stages.

*2.3. Tweet Extraction Techniques.* We have used several sources of information to get as many profiles as possible to perform the study. As primary source, we have used Twitter and Kaggle and, as secondary sources, several forums with terrorist ideologies and the official website of ISIS, known as Wafa Media Foundation, as seen in Figure 2. Primary sources were to collect the base information to be studied, to define the profile of the users actively spreading terrorist content while secondary sources were used to get familiar with the vocabulary used by these users in social propaganda. Many

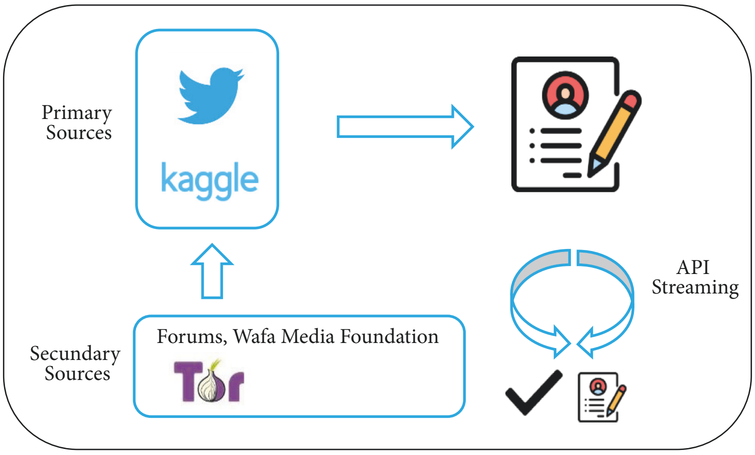

FIGURE 2: Tweet extraction sources.

keywords were obtained in this analysis to perform Twitter searches in a second step.

To download tweets from users speaking of terrorism (in favour and against), a connection between API Rest of Twitter and API Streaming was established. API Rest was used to check if the users whose information is being extracted had active accounts. API Streaming was used to download new users in real time and so expand the knowledge base about terrorists or potential targets. Although previous studies on network evolution show that social networks properties tend to reach an equilibrium [16], studies focused directly on terrorism are able to detect new trends and platform changes [17].

This way, we have expanded the suspicious profiles that we had from the initial databases "Isis_fanboy" and "About_isis" (Kaggle), with profiles that follow and spread the information and users included in these databases. The probability of obtaining users and repeated content is very high, as many times downloads belonged to followers of the downloaded users retweeting the same content (friends of friends sharing the same news, opinions and so). This was understood a clue to consider that we were searching information in the right direction.

95 the data is extracting in right direction.

This set of data was already registered and classified by the level of risk and continued downloading information associated, indicating the existence of active accounts, which allowed us to continue researching about it.

The extraction and analysis were focused on a social circle, in which we defined as "popular users" those with highest number of followers + publications + impact, serving as basis to categorize others that we already had. For this purpose, we retrieved several fields as

(i) Location is one attribute that we retrieved but have not been used, as we have not considered it as relevant information. In Twitter, as in many other social networks, location can be removed, or even faked.

(ii) Tweet_ID is the identification of a user on Twitter, though in this analysis we have not taken it into account as we have the username of the profile.

(iii) Time: date and time of the tweet. We used it to measure the periods of higher activity in an interval of five months. Once those maximums were located, we contrasted the dates to check if those days had any correlation with some political, social, or economic ISIS event.

How to build data base

With this information, we built the database shown in Figure 3.

*2.4. Information Preprocessing.* With the previous fields, we have discarded the fields name (many times is fake) and location (many times undetermined), using the rest to generate new knowledge. In particular the following variables were introduced:

(i) Frequency: using the field date, to calculate the interval within the user sends a tweet.

(ii) Sentiment: analysis of the sentiment of tweets of each user to polarize them in positives and negatives. For this analysis the Python library nltk and Vader Lexicon were used, specialized in natural language processing.

Python library for WLP
nltk &
Vader
Lexicon

(iii) Extraction from each tweet the mentioned or retweeted users, by using regular expressions and Python.

From this new dataset, a new filter was applied in order to locate those active users generating tweets and named or retweeted by other users (most impact users). From this set of users, the set of connections by means of a graph was computed.

*2.5. Graphing the Network.* We have created several graphs based on different metrics that would lead to different interpretations. The objective was to analyze and visualize social relationships among users and communities.

Once the graph was created, as the one displayed in Figure 4, we have applied indicators of centrality to identify the most important vertices based on several criterions, to detect as well as the most influential users those who receive

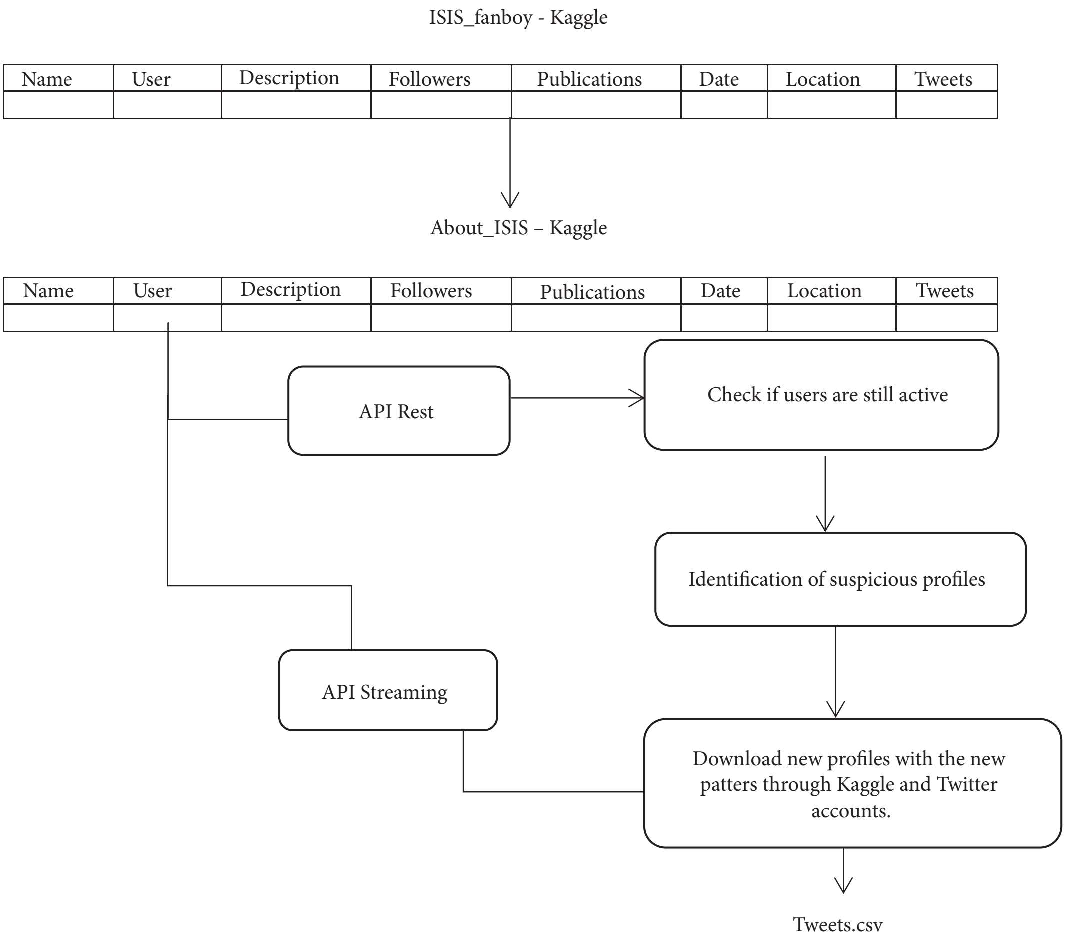

Figure 3: Tweet extraction procedure.

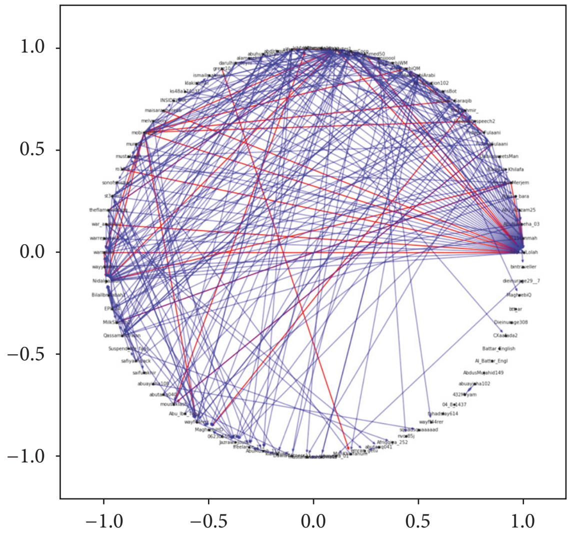

Figure 4: Graph indicating the interaction of users in Twitter related to the topic Jihadism.

information to broadcast it, or those who are following many influential users. To do so, we have weighted every link based on the number of retweets or mentions to a user (inside of the message @user, @user2. . .and so). We have not taken into account the sentiment of the message nor the frequency of shipment; these parameters will be used later in the fuzzy clustering procedure.

Most influential nodes are important in graph analysis, but many times in social communities those users are detected and located. Other criteria can be more important like nodes likely to be the most direct route between two influencer's nodes, or key nodes to reach the rest of nodes. That is why we have used centrality measures, to get different properties of a network and its behaviour. The more a node is centred, the more important it is. In particular, we have used two geometric measures, one being path based and the other one being a spectral measure to evaluate the influence and connections of a node within its community.

(i) Degree Centrality Measure is defined as $C\_deg(V) = deg(V)/|N| - 1$.

This geometric measure is used to find users very connected, those with the highest number of links with other nodes on the net. It takes into account the weighting on edges. We are not focusing on this type of users as usually this measure could serve as a measure of popularity among nodes, though it is important to evaluate their relationship with the rest of nodes.

(ii) Closeness is defined as $C(x) = (|N| - 1)/\sum_y d(y,x)$

This geometric measure represents the importance of each node according to how close is to the rest of nodes. Nodes with a high value tend to be very well connected to the most relevant nodes on their network and are perfect to broadcast information. They do not have to be very influential but are very active followers who broadcast information on the net and are very close to the most influential nodes.

(iii) Betweenness centrality is a path based measure defined as

$$\boxed{g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}} \tag{1}$$

where $\sigma_{st}$ is the total number of shortest paths from node s to node t.

$\sigma_{st}(v)$ is the number of those paths that pass through $v$.

It indicates the nodes included in the shortest path between most of the nodes. For this work, this was a very interesting measure because it could highlight those nodes that serve as bridge for influential nodes. In this case, these nodes might not have many followers but can connect many relevant nodes.

speciality.

(iv) The last measure to be applied is the eigenvector, which takes into account the number of links of a given user, as the number of links of its connections, and so could be considered as a hierarchical measure that computes not only your connections but the connections of whom you follow. Its value for node $v$ is given by the $v_{th}$ element of the eigenvector related to the first eigenvalue of the adjacency matrix of the graph [18],

As seen in Figure 5, the results obtained applying eigenvector and betweenness are quite similar in distribution. In this case measures that are not correlated are used as input for the fuzzy cluster to avoid redundant information and overtraining the cluster giving more relevance to some variables than others. We have chosen eigenvector as input for the clustering part as we are dealing with a structured problem, where there are people that train other users to broadcast information and so in a hierarchical model. The measure that better reflects this way of behavior is the eigenvector.

*2.6. Fuzzy Clustering.* Soft Computing techniques have been used in many different fields to deal with imprecision and uncertainty [19, 20].

In our problem, segmentation techniques are unsupervised methods used to classify information in groups created from similarities among individuals. The potential of the segmentation algorithms to show underlying structures in data can be applied in different fields such as classification, image processing, pattern recognition, modeling, and identification.

Segmentation techniques can be applied to quantitative or qualitative data. In this paper only quantitative ones will be used, to build the data matrix, which will have records as columns and measured variables as rows. By segmenting this data, the users are grouped in base to their similitude, understood from a mathematical point of view and defined as the "distance" among data or according to some prototypes of the group. This group depends, therefore, on the individuals being grouped together and on the definition of distance.

Within these segmentations we found two approaches:

(i) Hard Clustering: Objects belong to just one segmentation. Different groupings are excluding.

(ii) Fuzzy Clustering: This grouping technique applying fuzzy logic [21] allows different objects to belong to different segments simultaneously, but with different membership degree [22]. In many cases, this segmentation is more logical than the previous. In our case, a user can be interested in terrorism, though it has not been catalogued as dangerous, but with an elevated belonging degree in this group.

Therefore, fuzzy segmentation or fuzzy clustering is applied in this work, taking into account the nature of our problem with an objective function to obtain the optimal number of partitions. This optimization will lead to applying some nonlinear optimization algorithms to find a local minimum.

optimised is locally using some non linear optimization algorithm to find local minimum.

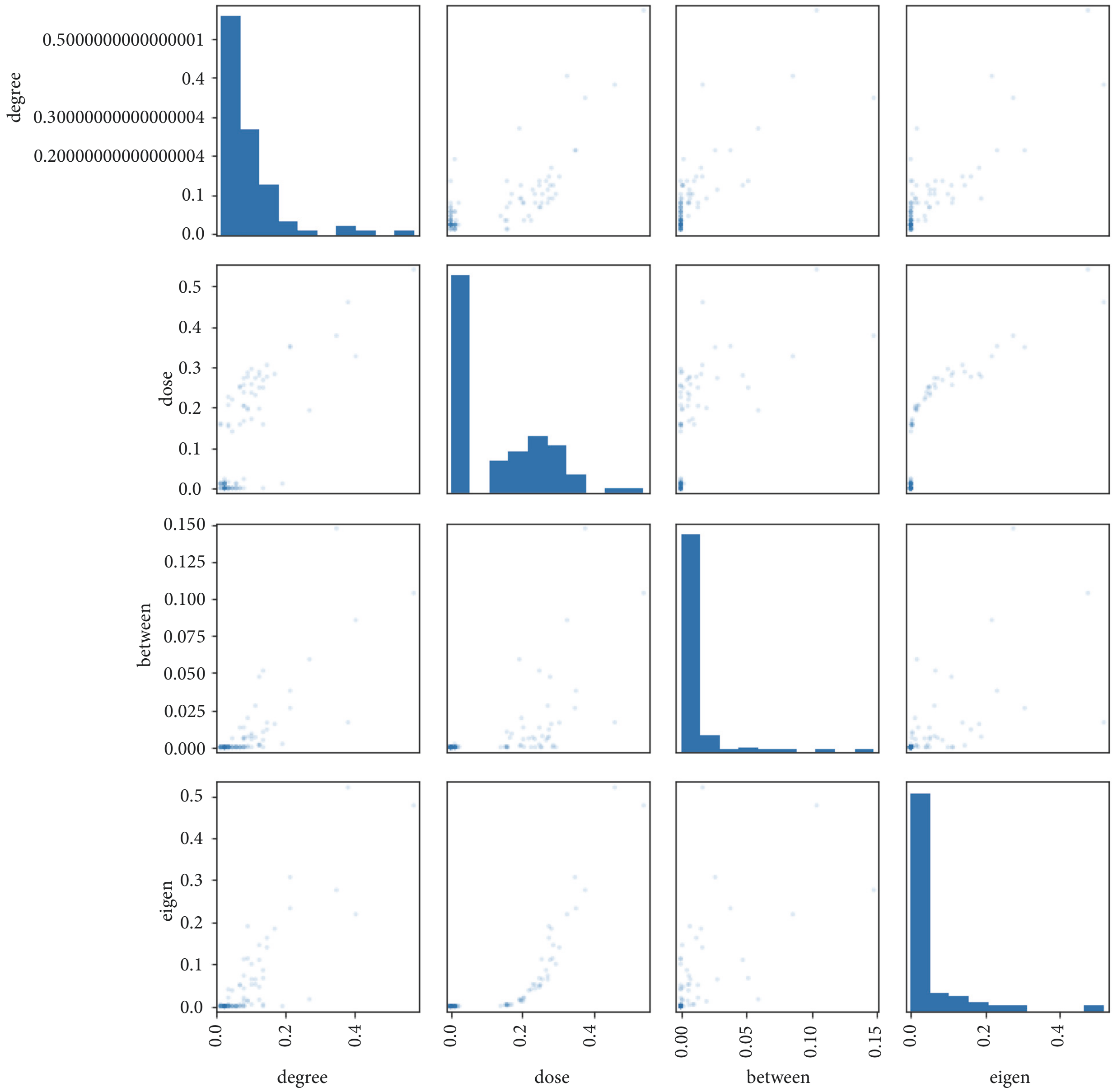

Figure 5: Distribution of users according to the proposed metrics.

*2.6.1. Fuzzy Clustering.* Groupments in this algorithm satisfy the following conditions:

$$\mu_{ij}\epsilon\left[0,1\right], \quad 1\le i\le c, \ 1\le j\le N$$

$$\sum_{i=1}^{c}\mu_{ij}=1, \quad 1\le j\le N \qquad (2)$$

$$0<\sum_{j=1}^{N}\mu_{ij}<N, \quad 1\le i\le c$$

where $c$ is the number of groups and $N$ is the number of records.

Fuzzy space for our data is the set defined by

$$F_c=\left\{U\in\mathbb{R}^{cxN}\mid\mu_{ij}\in\left[0,1\right], \ \forall i,k; \sum_{i=1}^{c}\mu_{ij}=1, \ \forall k; \ 0 < \sum_{i=1}^{c}\mu_{ij}<N, \ \forall i\right\} \qquad (3)$$

*2.6.2. Fuzzy Clustering c-Means.* This algorithm is based on the optimization of Fuzzy partitions [23, 24].

$$J\left(Z,U,V\right)=\sum_{i=1}^{c}\sum_{j=1}^{N}\left(\mu_{ij}\right)^{m}\left\|z_j-v_i\right\|_A^2 \qquad (4)$$

where $U$ is the membership matrix $[\mu_{ij}] \in F_c$ of our data and $V = [v_1, v_2, \ldots, v_c]$ are the vectors characterizing the centers of these groupings for which we want to minimize our functional.

The standard $A$ between our centers and the data is given according to the following:

$$D_{ijA}^2 = \left\| z_j - v_i \right\|_A^2 = \left( z_j - v_i \right)^T A \left( z_j - v_i \right) \qquad (5)$$

The parameter $m \in [1, \infty)$ determines the fuzziness of the segments. The ~~value of the~~ cost function $J(Z, U, V)$ can be interpreted as a measure of the deviation ~~between~~ points $v_i$ and centers $z_j$.

The minimization of this functional leads to a nonlineal optimization problem that can be solved through different methods as genetic algorithms or iterative minimization. However, the most popular for this application in particular is the iterative method of Picard.

The restriction of membership values $\mu_{ij}$ is imposed by Lagrange multipliers.

$$\begin{aligned} J(Z, U, V) = {} & \sum_{i=1}^{c} \sum_{j=1}^{N} \left( \mu_{ij} \right)^m \left\| z_j - v_i \right\|_A^2 \\ & + \sum_{j=1}^{N} \lambda_j \sum_{i=1}^{c} \left( \mu_{ij} - 1 \right) \end{aligned} \qquad (6)$$

We can demonstrate that to minimize the functional it is necessary that

$$\begin{aligned} \mu_{ij} &= \frac{1}{\sum_{k=1}^{c} \left( D_{ijA} / D_{kjA} \right)^{2/(m-1)}}, \quad 1 \le i \le c, \; 1 \le j \le N \\ v_i &= \frac{\sum_{j=1}^{N} \left( \mu_{ij} \right)^m z_j}{\sum_{j=1}^{N} \left( \mu_{ij} \right)^m}, \quad 1 \le i \le c \end{aligned} \qquad (7)$$

Therefore, the parameters to be determined in the algorithm are as follows:

(i) *Number of clusters*: this parameter is the most important and the one with greatest impact in the segmentation. If the number of groups to be divided our data is known, this parameter would be determined. We will determine the number of clusters though the fuzzy partition coefficient (FPC). This validation measure indicates how well our data are explained by this grouping; that is, the membership to each one of the segments of our data is, in general, strong and not fuzzy.

(ii) *Fuzziness parameter*: parameter $m$ affects significantly fuzziness in the segmentation. As it approaches 1, grouping ceases being fuzzy to be *hard*, and if it tends to $\infty$ it will be completely fuzzy. We have chosen the value ($m = 2$), as being the most used in bibliography.

(iii) *Termination criterion*: as it is an iterative algorithm, it is necessary to establish a termination criterion to stop iterations. In our case, we have set 1000 iteration or reach an error lower than 0.005.

(iv) *Distance matrix*: the calculation of distance implies establishing the scalar product matrix. The natural election is the identity matrix ($A = I$) but a distance matrix that is very extended is the inverse of the covariance matrix of the data, leading to the Mahalanobis standard.

$$\begin{gathered} A = R^{-1}, \\ R = \frac{1}{N} \sum_{i=1}^{N} \left( z_i - \bar{z} \right) \left( z_i - \bar{z} \right)^T \end{gathered} \qquad (8)$$

The norm used influences the segmentation criterion changing the measure of dissimilarity. The Euclidean norm leads to hyperspherical groupings in the coordinate axes, while Mahalanobis leads to hyperelipsoidal groupings in the axes given by covariances between variables.

In addition to these parameters, in bibliography we can find several modifications of the algorithm:

(i) Modifications that use an adaptive distance measure, as the algorithm of Gustafson-Kessel [25] and the fuzzy maximum likelihood estimation [26].

(ii) Algorithms relaxing the condition on the probability of belonging to each segment ($\sum_{i=1}^{c} \mu_{ij} = 1, 1 \le j \le N$) indicating a level between each one of the groups.

In this work, we have checked the Euclidean norm, Mahalanobis, and Gustafson-Kessel.

The Gustafson-Kessel algorithm expanded the adaptive distance to detect different groupings with different geometrical forms. Each segment has its own distance given by

$$D_{ijA_i}^2 = \left( z_j - v_i \right)^T A_i \left( z_j - v_i \right) \qquad (9)$$

The matrices $A_i$ become variables that are optimized within the functional $J$ so each group will have the distance that minimize its value. The only restriction imposed is that the determinant has to be positive, ($|A_i| = \rho_i, \rho_i > 0, \forall i$). Optimizing using the Lagrange multipliers method, we obtain that the distance matrices must fulfill this

$$A_i = \left[ \rho_i \det \left( F_i \right) \right]^{1/m} F_i^{-1} \qquad (10)$$

where $F_i$ is the fuzzy covariance matrix of each one of the segments.

$$F_i = \frac{\sum_{j=1}^{N} \left( \mu_{ij} \right)^m \left( z_j - v_i \right) \left( z_j - v_i \right)^T}{\sum_{j=1}^{N} \left( \mu_{ij} \right)^m} \qquad (11)$$

The parameters of this algorithm are, in addition to the general parameters of segmentation, the volumes of the groups $\rho_i$. If we do not have knowledge about this value, we set 1.

We have tested several measures to check which ones fits better to segment our dataset [27].

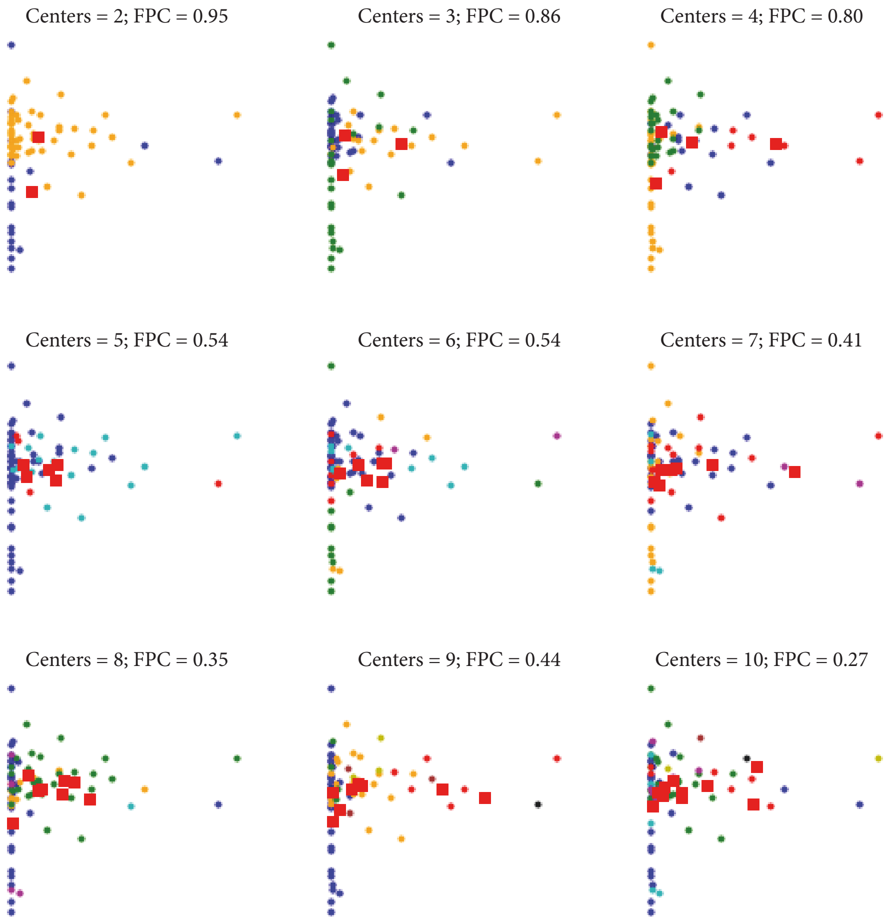

FIGURE 6: Results of the pro-Isis users with different levels of segmentation.

## 3. Results and Discussion

Once the previous methodology was defined and programmed, the Kaggle dataset with the pro-Isis Twitter registered users was firstly used. As criterion to choose which distance matrix and number of distances to be used, the configuration that allows a high membership degree for most of data has been established. For this dataset, the maximum value was obtained using the Gustafson-Kessel algorithm with two segments, as seen in Figure 6.

With this criterion, the division has been performed according to the variables mentioned (frequency, sentiment), eigenvector, and coherently with the rest of variables; it is possible to identify a more dangerous user group (red) and a less active group (blue).

However, one of the main advantages of this methodology is that we can identify users that have been identified within one group more than another, in spite of having a low membership degree. Figure 6 shows the membership degree to both groups, and the marked zone would be a user zone to be analyzed in detail. For example, with this same dataset, establishing a fuzzy membership threshold of 35-65%, we would obtain 2 doubtful users among 74 profiles. These 2 users are the focus of our work and should be studied individually and in time, to check whether they remain in the same place or have turned their behaviour to a more radical one, as seen in Figures 7 and 8.

To verify the consistency of our methodology, we performed a new experiment with an expanded dataset, added to the profiles identified as "fan boys". We have included other profiles considered of interest because of their connections with the users of the first set.

When downloading Twitter information, we have discarded the number of followers (as we have the information of users whom mention/retweet and are mentioned/retweeted) and number of publications (as we have the number of tweets published in the sampling).

After filtering users by those who generate content and those who receive it, we programmed a weighted graph to obtain centrality measures in our network.

The best segmentation value was obtained using Mahalanobis distance in two clusters, obtaining a FPC of 0.85 as we can see in Figure 9, which are represented the FPC obtained in base to the number of fuzzy clusters.

To clarify the results, in Figure 10 we are representing segmentation in base to the variables that we have used to

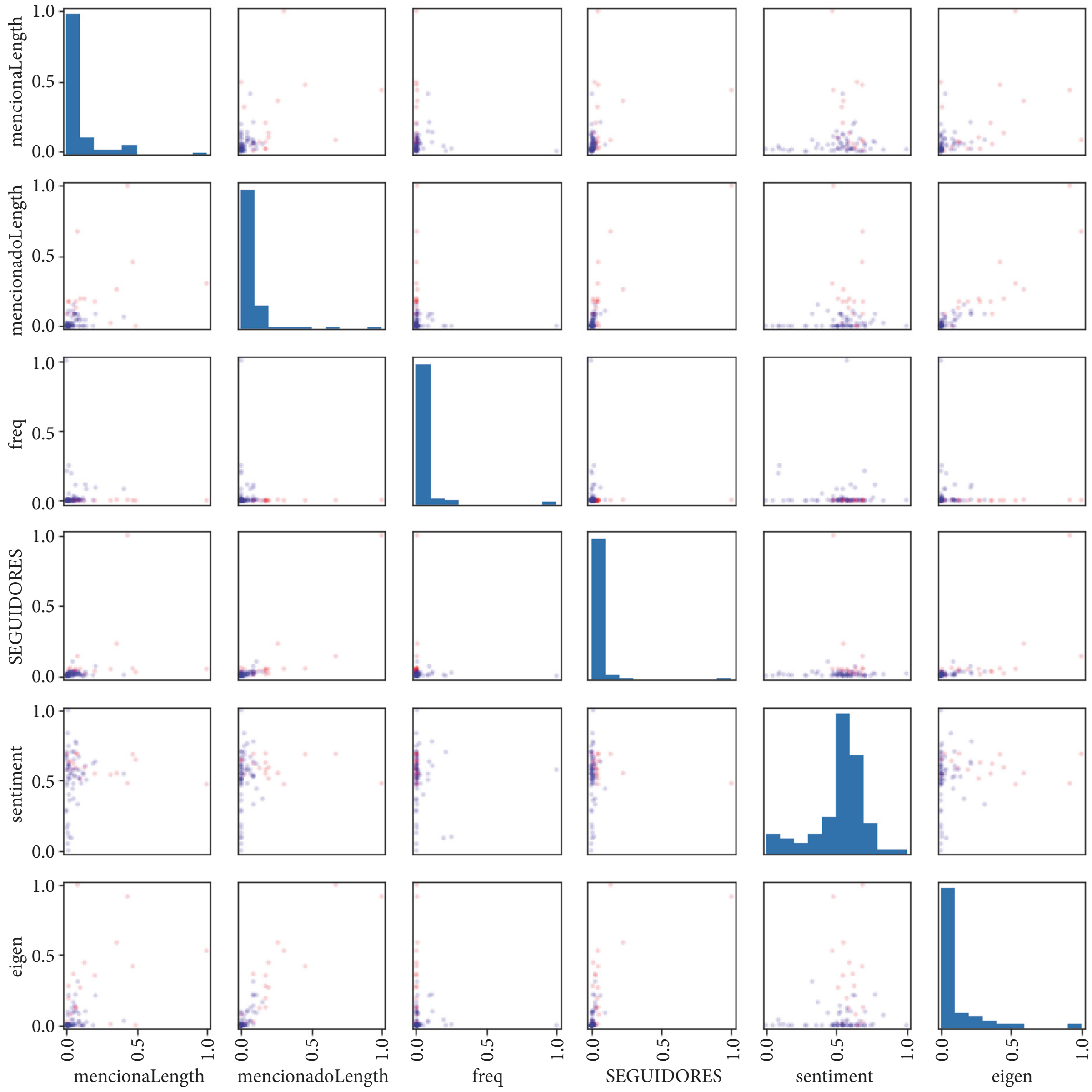

Figure 7: Distribution of the variables used in the fuzzy clustering.

calculate it, as frequency, eigenvector, and number of times that it has been mentioned.

For this experiment, if we set a fuzzy membership threshold of 0.45, we find 123 users to be studied from the 3395 in total. As in the previous case, these users, represented as the blue box of Figure 11, are susceptible to be studied and monitored to trace their behavior along time.

On the other hand, after identifying the most active users, we checked how many of the users categorized as active users in the first dataset were categorized as active as well as in this second set (blue group). From 74 users of the first categorization, 59 were correctly grouped, and the fuzzy membership of 15 left users is represented in the boxplot of Figure 12. The distribution of the membership shows that more than 25% of these false positive and negative users had a very weak membership (between 0.50 and 0.6), which means that these users should be traced to check their behavior using our methodology.

## 4. Conclusions

The use of social networks as a manner to broadcast information has become a popular way to attract new followers to terrorism in general and Jihadism in particular. In this work, we have developed a methodology to identify potentially dangerous users that remain partially hidden, separated from those that are best known for being very active. In fact, the terrorist attack in Barcelona in 2017 was organized by terrorists whose profiles had not been classified as dangerous. Detecting and monitoring those new profiles can be crucial

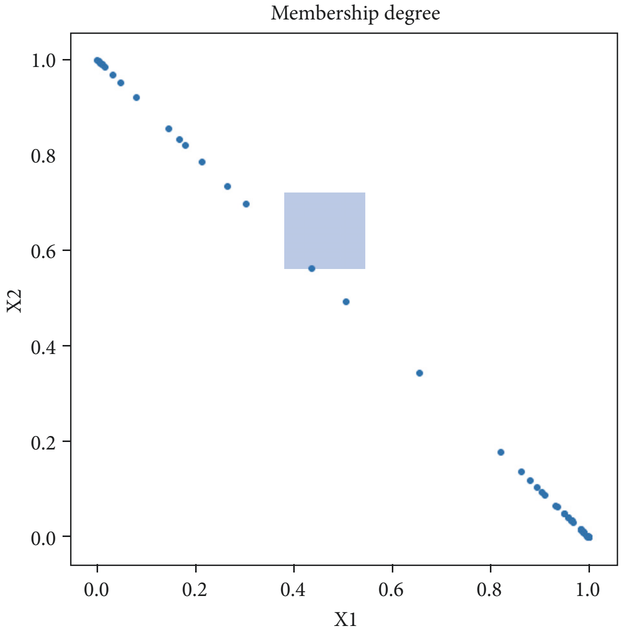

Figure 8: Membership degree function remarking the zone of users that should be analyzed.

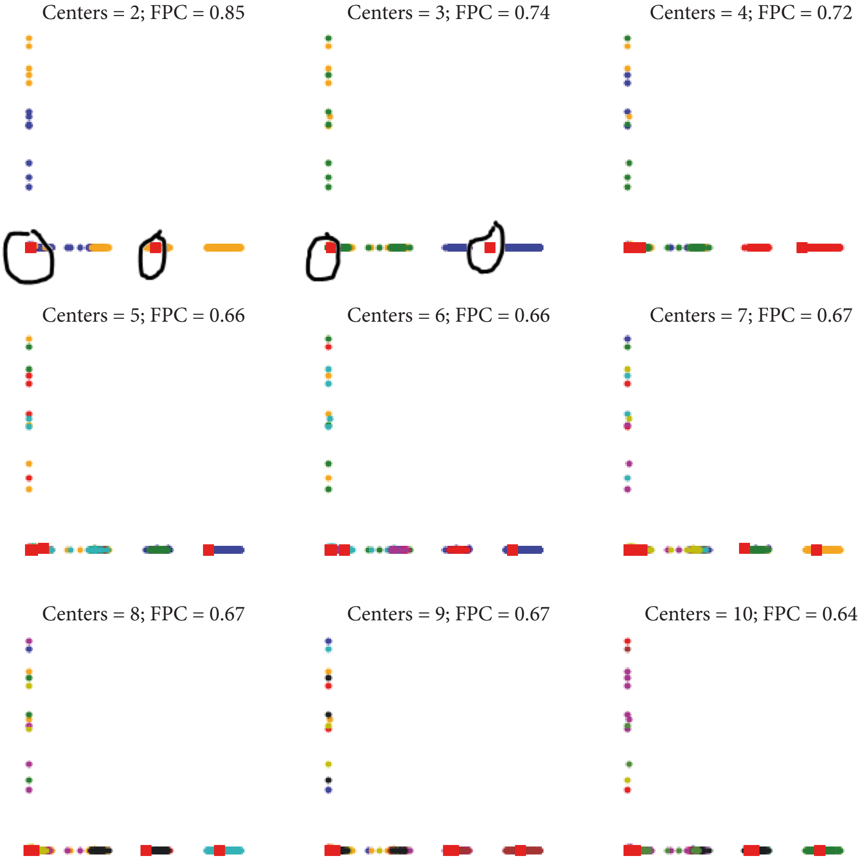

Figure 9: Representation of the FPC in base to the number of clusters.

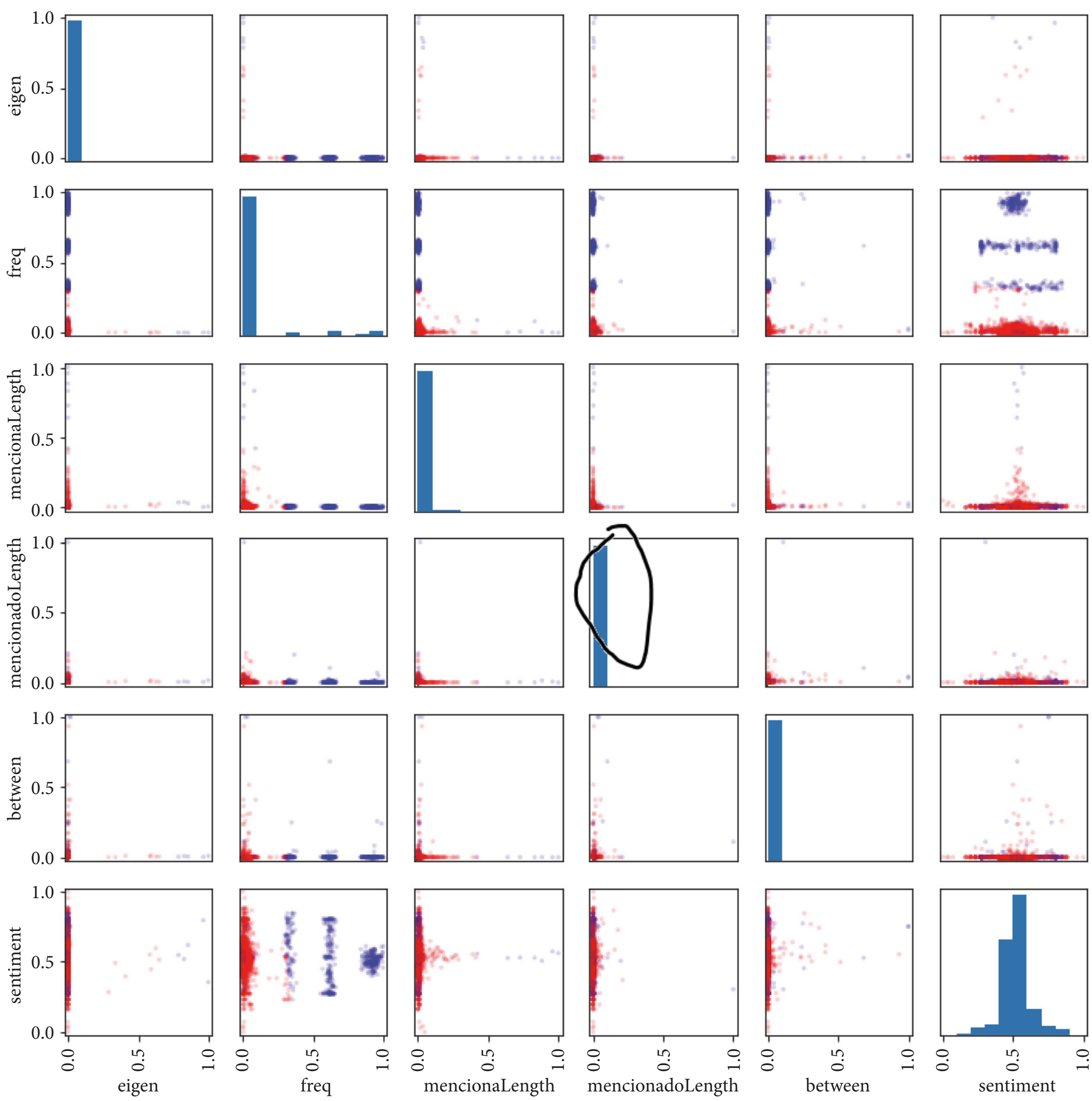

FIGURE 10: Segmentation variables for experiment 2.

to prevent or predict future terrorist actions and terrorist recruiting.

The presented methodology consists of defining a dataset of users plus several metrics to locate influential users. In addition, other metrics are obtained like frequency or the sentiment of their tweets. These metrics are used as vectors to perform segmentation of data. For this type of procedures, it is recommendable to use Soft Computing techniques that deal with the imprecision of the information. In the present problem we have used fuzzy clustering techniques to point out those users that were susceptible to be more active in the future and in consequence to be followed in time to check their behavior. Moreover, the analysis techniques proposed involve unsupervised algorithm, so they can be applied continuously, thereby this same methodology could be used to monitor users marked as fuzzy.

As for future works, we would like to expand this methodology other environments where user profiles could have similar patterns, like pedophilia or fake news. Fake news is known to have an important economical impact if they damage the image of a company and an important political impact if they can manipulate a significant number of voters.

## Data Availability

The Kaggle dataset used to support the findings of this study have been deposited in the Kaggle repository under the name How ISIS Uses Twitter: https://www.kaggle.com/fifthtribe/how-isis-uses-twitter.

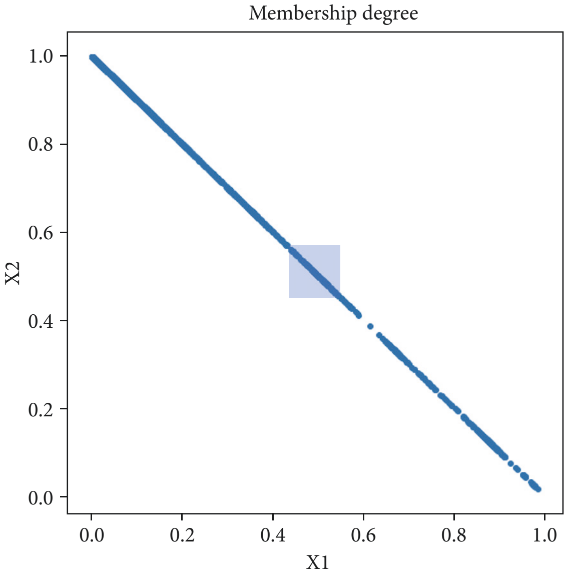

Figure 11: Membership degree function of experiment 2.

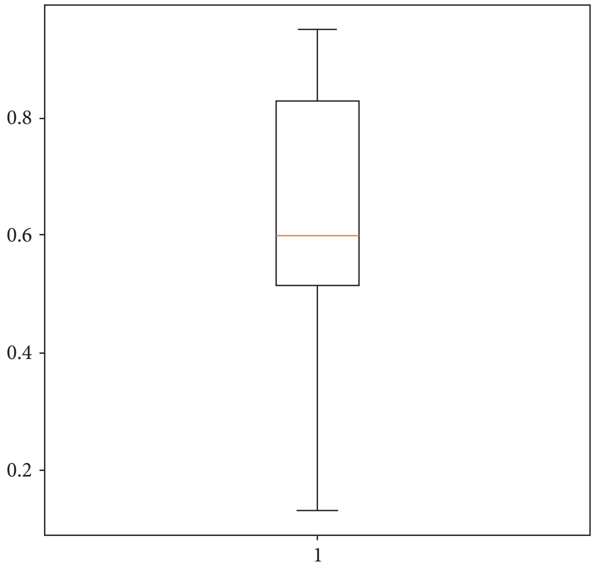

Figure 12: Boxplot representing the categorization of the most active users.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

## References

✓[1] J. Jordan, F. M. Mañas, and N. Horsburgh, "Strengths and weaknesses of grassroot jihadist networks: the madrid bombings," *Studies in Conflict & Terrorism*, vol. 31, no. 1, pp. 17–39, 2008.

✓[2] A. Bermingham, M. Conway, L. McInerney, N. O'Hare, and A. F. Smeaton, "Combining social network analysis and sentiment analysis to explore the potential for online radicalisation," in *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining (ASONAM)*, pp. 231–236, Athens, Greece, July 2009.

[3] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '03)*, pp. 137–146, New York, NY, USA, August 2003.

[4] J. H. Fowler and N. A. Christakis, "Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham heart study," *British Medical Journal*, vol. 337, Article ID a2338, 9 pages, 2008.

[5] C. Mao and W. Xiao, "A comprehensive algorithm for evaluating node influences in social networks based on preference analysis and random walk," *Complexity*, vol. 2018, pp. 1–16, 2018.

~~Not got~~

[6] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca, "Network analysis in the social sciences," *Science*, vol. 323, no. 5916, pp. 892–895, 2009, American Association for the Advancement of Science.

[7] C. Ordóñez Galán, F. Sánchez Lasheras, F. J. de Cos Juez, and A. Bernardo Sánchez, "Missing data imputation of questionnaires by means of genetic algorithms with different fitness functions," *Journal of Computational and Applied Mathematics*, vol. 311, pp. 704–717, 2017.

[8] M. Delgado, M. J. Martín-Bautista, D. Sánchez, and M. A. Vila, "Mining text data: special features and patterns," in *Pattern Detection and Discovery*, vol. 2447 of *Lecture Notes in Computer Science*, pp. 140–153, Springer, Berlin, Germany, 2002.

[9] W. Aziguli, Y. Zhang, Y. Xie et al., "A robust text classifier based on denoising deep neural network in the analysis of big data," *Scientific Programming*, vol. 2017, Article ID 3610378, 10 pages, 2017.

[10] C. J. de la Torre, D. Sánchez, I. Blanco, and M. J. Martín-Bautista, "Text mining: techniques, applications, and challenges," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 26, no. 04, pp. 553–582, 2018, World Scientific Publishing Company.

[11] A. Hamdi, N. Monmarché, M. Slimane, and A. M. Alimi, "Fuzzy rules for ant based clustering algorithm," *Advances in Fuzzy Systems—Applications and Theory*, vol. 2016, Article ID 8198915, 16 pages, 2016.

[12] I. A. Atiyah, A. Mohammadpour, and S. M. Taheri, "K C - means: a fast fuzzy clustering," *Advances in Fuzzy Systems—Applications and Theory*, vol. 2018, Article ID 2634861, 8 pages, 2018.

[13] R. Singh, J. Singh, and R. Singh, "Fuzzy based advanced hybrid intrusion detection system to detect malicious nodes in wireless sensor networks," *Wireless Communications and Mobile Computing*, vol. 2017, Article ID 3548607, 14 pages, 2017.

[14] G. Bello-Orgaz, J. J. Jung, and D. Camacho, "Social big data: Recent achievements and new challenges," *Information Fusion*, vol. 28, pp. 45–59, 2016.

[15] Z. Su, Q. Xu, and Q. Qi, "Big data in mobile social networks: a QoE-oriented framework," *IEEE Network*, vol. 30, no. 1, pp. 52–57, 2016.

[16] G. Kossinets and D. J. Watts, "Empirical analysis of an evolving social network," *Science*, vol. 311, no. 5757, pp. 88–90, 2006.

[17] J. Klausen, E. T. Barbieri, A. Reichlin-melnick, and A. Y. Zelin, "The YouTube jihadists: a social network analysis of al-muhajiroun's propaganda campaign," *Perspective on Terrorism*, vol. 6, no. 1, pp. 1–12, 2012, Terrorism Research Institute, http://www.jstor.org/stable/26298554.

[18] M. E. J. Newman, *Networks: An Introduction*, Oxford University Press, Oxford, UK, 2010.

[19] L. Álvarez Menéndez, F. J. de Cos Juez, F. Sánchez Lasheras, and J. A. Álvarez Riesgo, "Artificial neural networks applied to cancer detection in a breast screening programme," *Mathematical and Computer Modelling*, vol. 52, no. 7-8, pp. 983–991, 2010.

[20] P. J. G. Nieto, J. R. A. Fernández, F. S. Lasheras, F. J. de Cos Juez, and C. D. Muñiz, "A new improved study of cyanotoxins presence from experimental cyanobacteria concentrations in the Trasona reservoir (Northern Spain) using the MARS technique," *Science of the Total Environment*, vol. 430, pp. 88–92, 2012.

[21] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, pp. 338–353, 1965.

[22] R. L. Cannon, J. V. Dave, and J. C. Bezdek, "Efficient implementation of the fuzzy c-means clustering algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 2, pp. 248–255, 1986.

[23] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *Journal of Cybernetics*, vol. 4, no. 1, pp. 95–104, 1974, Taylor & Francis Group.

[24] J. C. Bezdek, "Objective function clustering," in *Advanced Applications in Pattern Recognition*, pp. 43–93, Springer, Boston, MA, USA, 1981.

[25] D. E. Gustafson and W. C. Kessel, "Fuzzy clustering with a fuzzy covariance matrix," in *Proceedings of the 1978 IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes*, pp. 761–766, 1979.

[26] I. Gath and A. B. Geva, "Unsupervised optimal fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 773–780, 1989.

[27] H.-C. Liu, J.-M. Yih, D.-B. Wu, and S.-W. Liu, "Fuzzy C-mean algorithm based on "complete" Mahalanobis distances," in *Proceedings of the 2008 International Conference on Machine Learning and Cybernetics (ICMLC)*, pp. 3569–3574, Kunming, China, July 2008.