

LK-Means algorithm for evaluation of behavior of university students in social networks related to cyberbullying and netiquetas topics

Cota Ortiz María de Guadalupe
Departamento de Matemáticas
Universidad de Sonora
Hermosillo, Sonora, México
guadalupe.cota@unison.mx

Ojeda Cota María de Guadalupe
Departamento de Matemáticas
Universidad de Sonora
Hermosillo, Sonora, México
malupitaoc85@gmail.com

Flores Pérez Pedro
Departamento de Matemáticas
Universidad de Sonora
Hermosillo, Sonora, México
pedrofloresperez@gmail.com

Abstract—This paper presents use of the LK-Means algorithm to classify a sample of university students, which was used to analyze the knowledge that they have about cyberbullying and use of Netiquetas in digital social networks. Information was collected in the context of the topics of virtual violence, peaceful coexistence, and knowledge in Netiquetas application. Each participant was assigned by category according to their knowledge level, grouping them into 5 classes to identify those who should receive training in the topics. Results allowed to recognize that 51% corresponds to violent relationships, 65% to peaceful coexistence and 58% have knowledge about the use of Netiquetas and that the algorithm showed high precision and efficiency improving manual procedure and decreasing processing time.

Keywords: *pattern-recognition, cyberbullying, netiquetas, social-networks.*

I. INTRODUCTION

This paper describes the work that was aimed at applying the LK-Means algorithm, which was used to analyze and classify the level of knowledge of Cyberbullying and Netiquetas usage. To apply the algorithm, a sample of university students of the University of Sonora was taken as a training set, to which an instrument was applied to collect data on topics of virtual violence, peaceful coexistence and knowledge of Netiquetas use in digital social networks. The data were used for an exploratory study in order to determine if the relationships are regulated individually (self-regulation) or through educational tools such as Netiquetas. The participants were classified to determine who should receive complementary training in the use of digital social networks.

Currently, Internet has become an indispensable resource for exchanging information regardless of the distance between the connection points. This form of communication

has expanded worldwide since the end of the 1990s, becoming a habitual and powerful tool identified as a web (world wide web or www) (Instituto de Tecnologías Educativas, s. f. [1]).

With the internet, digital social networking platforms are created, which serve as a way for coexistence among the individuals of the world and become popular in 2006 when Facebook and Twitter are created [2]. Due to resources such as synchronous communication between two or more people, lower prices compared to telephony services, and the exchange of experiences, opinions and affinities [3]. This form of communication has become one of the most important in the world. However, it has been proven that these platforms can present interactions with tendency to violence because they don't have legal backing that guarantees physical and psychological safety of Internet users. In Mexico, the National Institute of Statistics and Geography (INEGI) [4] reports that 83.2% of Internet users of 12 years or more have been involved in situations of virtual violence, which is increasing due to the lack of norms that regulate the behavior of the people who use them.

II. CLASSIFICATION: "LK-MEANS" ALGORITHM

Machine-Learning [5] is a scientific discipline of the area of Artificial Intelligence through which you can automate analytical models to generalize behaviors in a set of data called "training set". To later you can search similar characteristics in the data and group them in "classes" or categories [6], using mathematical methods [7].

This approach is used by some algorithms applying mathematical metrics to measure distances that exist between the points that correspond to the records of the data set. In this work we used a modified version that we have named "LK-Means, based on the procedure of the algorithm" K-Means" [5] (Fig. 1).

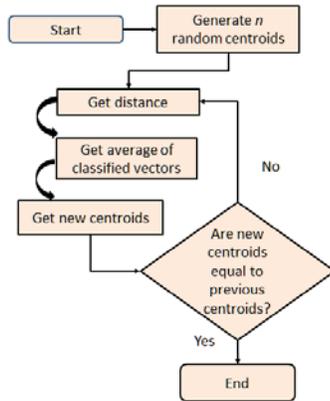


Fig. 1 K-Means procedure

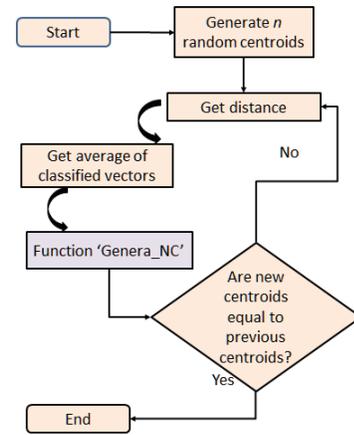


Fig 2. LK-Means procedure

III. LK-MEANS MODEL

In the "LK-Means" version (Fig. 2) of the original algorithm "K-Means" (Fig. 1) [5], as a modification to the original version. It was developed as a function written in programming language 'C', which has been named 'Genera_NC' and is based on the following model:

- C Number of classes that going to be identify (a centroid is created for each class).
- NC Number of attributes or characteristics that identify each instance.
- NE Number of instances of the training set.
- XM Average higher value obtained per class. Initial value = 0.0.
- XA Actual value of PX for each class. Initial value = 0.0.
- ES Auxiliary value = 0.1.
- FX Numeric factor named "XFactor". Initial value = 0.0.
 $FX = XM / (NC * 0.01)$. The calculation of FX is applied in the stage where new centroids are generated, such as shown in Figure 3.
- PX Vector with average values of classes.

For each centroid "XFactor" is estimated according to the value that results from calculating the average of the vectors, selecting higher value, wich is divided between the number of attributes. "XFactor" is added to the value that corresponding to the 'new_centroids'. This procedure modifies efficiently the values obtained with the traditional method, changing the behavior of the original algorithm.

Thus, the "centroid" is a point that identifies the selected instance or "point" to identify the central instance of a "class".

The distance metric used in this work is the "Euclidean distance", which is applied through formula number 1, taking as a basis the points $P = (p_1, p_2, \dots, p_n)$ and $Q = (q_1, q_2, \dots, q_n)$, with $i = (1, 2, \dots, N)$.

$$\sqrt{\sum_{i=1}^n (p_i - q_i)^2} \quad \text{Fórmula (1)}$$

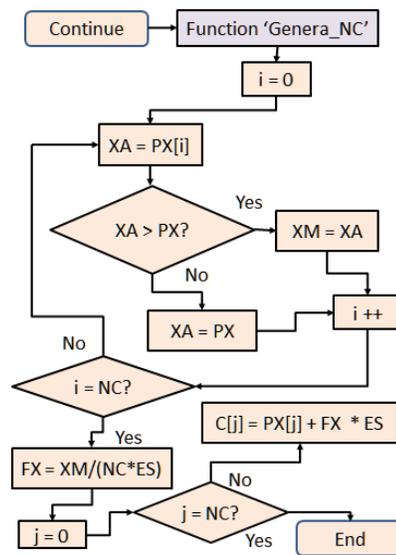


Fig 3. Genera_NC Function

Processing method:

1. LK-Means is started select randomly a number of points equal to number of "classes" to identify them as centroids.
2. The metric used to calculate distances from one point to another in the data set is euclidian distance, but the algorithm "LK-Means" can use any one of existing formulas, like other versions of algorithms based in original method "K-Means".
3. The proximity of each "point" or instance of the "training set" to the "points" identified as centroids is calculated and assigned to the group or subset whose distance is smaller.
4. Once the "points" are grouped, the average of each one is calculated and the "XFactor" is added so that they become the new centroids.
 - a. The values of the new centroids with the previous centroids are compared, evaluating results based on the following conditions:
 - The algorithm converges and ends if new centroids are equal as the previous centroids.

- b. Another difference of "LK-Means" is that converges when the number of iterations is high and doesn't end under condition discussed in the previous point 'a', taking as a criterion that the groups don't vary in the number of elements grouped by "class".
 - c. In case the new centroids are different from the previous centroids, the algorithm will repeat step 4, until the algorithm converges under the conditions specified in the previous points 'a' or 'b'.
5. The results obtained can be evaluated by the analyst to issue his conclusions, identifying the classified groups.

IV. COEXISTENCE OF UNIVERSITY STUDENTS IN VIRTUAL SPACES

Digital social networks (RSD) work through the Internet and provide their users some privacy and control that is difficult to have in physical and natural spaces, which allows diversity in interactions. For example, a young college student can relate in a way and do it different in RSD. In this case, the regulation of behavior is important, especially if the behavior is based on universal values that promote a peaceful coexistence.

UNESCO [8] is an organization that promotes universal values such as respect, freedom, justice, equality, and human rights through the Culture of Peace Program. These values are part of the "tolerance" value which is promoted by UNESCO [8] so well as dialogue and peaceful coexistence among countries:

- a) Tolerance: supports human rights and promotes international standards for responsibility among countries.
- b) Respect: everyone has the right to think and must respect the rights of others.
- c) Freedom: of thought, conscience and religion.
- d) Justice: everyone has the right to be treated with justice, impartiality and equality of conditions.
- e) Equality: in treatment and opportunities.

The coexistence of university students in the RSD is not exempt from conflicts. If they are treated through behaviors regulated by ethical and moral principles, it increases the possibility of avoiding situations of violence and danger among those involved.

Galtung's theory [9] is based on the violence generated by the impossibility of resolving conflicts, which indicates that a conflict involves:

- a) An attitude that may be hatred toward one or more people.
- b) An action that can generate violence in the relationship.
- c) A contradiction that may not be resolved by those involved.

Galtung [9] says that in order to resolve a conflict effectively, one must start by working in contradiction (negation of needs) and then continue with attitude and behavior (see figure 4).

If a conflict is not resolved satisfactorily, it can cause violence, in which case, Galtung [9] mentions three types of violence:

- a) Direct violence (conflict in personal relationships).
- b) Structural violence (own conflict of a social structure).
- c) Cultural violence (conflict between cultures).

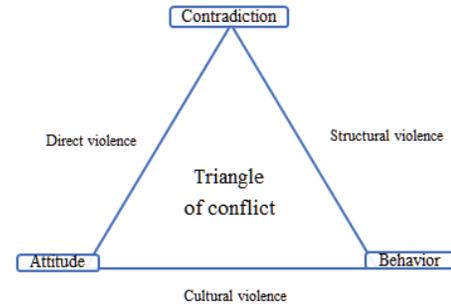


Fig. 4 Conflict triangle Galtung's [9], [10].

In this research we study three elements of conflict (contradiction, attitude and behavior) that can be generated by "direct violence" in university students. This is based on the argument that these three elements must be worked in a functional way to avoid this type of violence.

The test applied to a sample of university students included situations based on universal and network values to analyze them based on the contradiction, attitude and behavior of the Galtung conflict theory [9].

On the other hand, Rodríguez [11] defines a new concept called "Netiquetas", which to be applied as social norms to regulate the behavior of university students in virtual spaces for a peaceful coexistence, such as:

- a) Rule 1: "Never forget that the person reading the message is another human being with feelings that can be hurt".
- b) Rule 2: Adhere to the same standards of online behavior that you follow in real life.
- c) Rule 3: Writing everything in capital letters is like shouting, and makes reading difficult.
- d) Rule 4: Respect the time and bandwidth of other people.
- e) Rule 5: Show the good side of yourself as long as you stay online.
- f) Rule 6: Share your knowledge with the community.
- g) Rule 7: Help keep debates in a healthy and educational environment.
- h) Rule 8: Respect the privacy of third parties.
- i) Rule 9: Don't abuse your power or the advantages that you may have.
- j) Rule 10: Excuse the mistakes of others. Understand the mistakes of others as you expect others to understand yours.

This analysis on netiquettes [11], universal values [8] and the theory of conflicts of Galtung [9] is precisely the basis of this research, evaluating contradictions, attitudes and behaviors of university students regarding to various situations raised based on universal values and "Netiquetas" (see figure 5).

Phase I: in this phase of the investigation, attitudes and behaviors that involve virtual violence and level of knowledge and application of "Netiquetas" are analyzed. These elements allow us to detect the level of education that university students practice in virtual spaces. For this, in the applied test the questions were grouped into two groups: virtual violence and "Netiquetas".

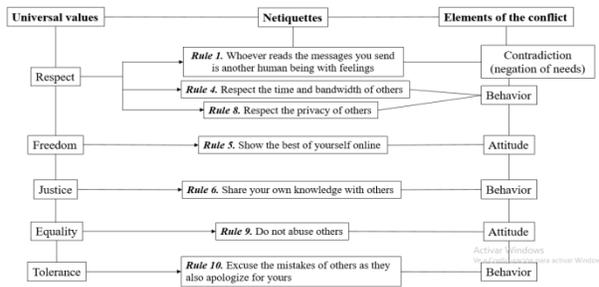


Fig. 5 Forms of coexistence in digital social networks

V. RESULTS

In Figure 6 shows the distribution of the "training set" that corresponds to the general set of 136 people who were grouped into 5 classes. In it you can see, that even for a small sample like the one used in the present work, it presents a high level of difficulty for the manual classification of the "points" in a fast and precise way without being able to avoid human errors. However, using pattern recognition techniques with classification algorithms, it is possible to treat problems with optimal results in a short time.

In this context, the application of algorithms such as "LK-Means", unlike other algorithms, allowed us to verify the results of the sample, identifying the class to which each person participating, and we can fully identify the class or category that corresponds to each person to decide which should participate in training programs on the proper use of virtual social networks.

On the other hand, it is important to comment that the levels established to classify or categorize the data set through the algorithm "LK-Means", are based on numerical values that allow to identify the frequency that is given in response to the questions of the applied test (see table 1).

As an example, to question 40 "I have recorded and / or published embarrassing information about someone, without their approval", the individual chooses one of the frequencies established in the range shown in table 2 as an answer.

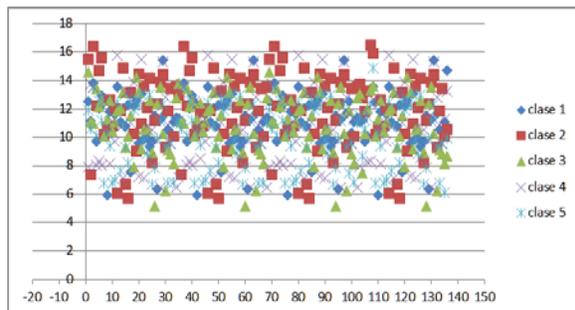


Fig. 6 General data distribution

Results obtained with the algorithm "LK-Means"

For comparison purposes, calculations were made manually, requiring 2 people working 3 hours per day in 1 week and applying "LK-Means" algorithm obtaining results in 10 seconds (see table 2 and figure 7). Next, we compared the effectiveness of both procedures and we found a 90% match.

Table 1
Values established to answer the applied instrument

Category	Value	Frecuency	Education level in topics
Class 1	1	0%	Lower
Class 2	2	25%	Low
Class 3	3	50%	Midle
Class 4	4	75%	High
Class 5	5	100%	Higher

To decide which people need to participate in training programs on the proper use of social networks, we group the classes (5) into two groups:

The people who were classified in classes 1, 2 and 3 correspond to very low, low and medium level (1-3). We consider that they must increase their level of knowledge in the use of social networks.

The people who were classified in classes 4 and 5 correspond to high and very high level of knowledge in the use of social networks (4-5). We consider that their behavior is regulated to an acceptable but not optimal level and that they only need to receive information through talks or printed and / or electronic materials.

The "LK-Means" algorithm showed results that indicate that 56% represents the first group (classes 1-3) and 44% represents the second group (classes 4-5)

Table 2
Grouping

Category	instances	(%)	Education Level
Class 1	19	14	Lower
Class 2	13	10	Low
Class 3	44	32	Midle
Class 4	32	23	High
Class 5	28	21	Higher

Total 136

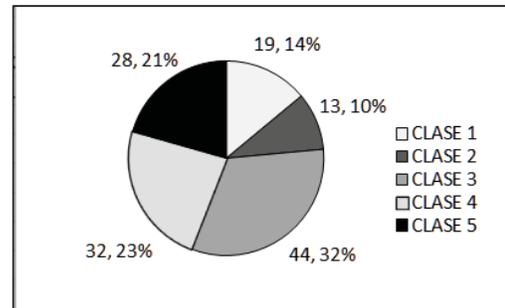


Fig. 7 General data distribution

On the other hand, Table 3 shows the results of the manual count performed to obtain average values by category. For the group of classes 1-3, it is observed that they are higher in the answer, whose value is between 1 and 3, and that for the group of classes 4-5 lower values were found. These results agree with those of the "LK-Means" classification algorithm, as shown below.

Taking the values in Table 3 as a reference, in Figure 8 the following precisions can be observed:

- The greater the range of response values of classes in group 4-5, the lower the range of response values in group of classes 1-3.
- The greater the range of response values for classes in group 1-3, the lower the range of response values for group classes 4-5.

Table 3.

		Average count- Manual procedure				
Frecuency identifier:		1	2	3	4	5
Class 1-3		30	31	29	24	24
Class 4-5		22	21	24	29	29

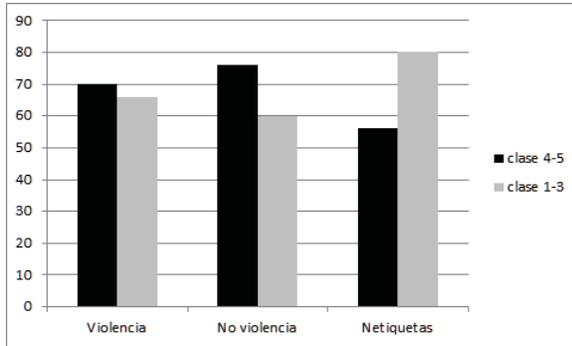


Fig 8 Manual estimation by group of questions

Now we describe the results obtained through the manual counting of the answers, observing, in general, a behavior coinciding with the one thrown by the algorithm "LK-Means".

Table 4 presents the manual calculation performed on the responses of the 136 people to whom the test was applied, based on the grouping of questions of the groups of classes 1-3 and classes 4-5. In Figure 9 it can be seen that there is a balance in the percentages obtained, so it is concluded that the individuals classified in group 1-3, require training in the topics on the use of virtual social networks, and that the classified in class group 4-5 they can receive talks or written or electronic information.

Tabla 4.

		Percent obtained manually			
Topic	No. Question	Class 1-3		Class 4-5	
		No. instances	%	No. Instances	%
Virtual violence	8	66	49	70	51
Opposed to virtual violence	33	60	45	76	55
Knowledge and application of Netiquetas	12	80	58	56	42

Table 5 shows the following differences between the results of the "LK_Means" algorithm and the manual count:

- In virtual violence: 2%.
- In coexistence with virtual violence: 10%.
- In the knowledge of "Netiquetas": 4%.

With these results, it can be concluded that at least 90%, the two ways of evaluating the data present coincidences, but we believe that more accurate results are obtained with computational algorithms and the processing time is shortened. In addition, we find that in the manual process it is easier for human errors to be committed and that more time and human resources are required to carry out this work

Table 5.
Comparative

Topic	Class 1-3		Class 4-5	
	% LK-Means	% Manual count	% LK-Means	% Manual count
Virtual violence	51	49	49	51
Opposed to virtual violence	65	45	35	55
Knowledge and application of Netiquetas	62	58	38	42

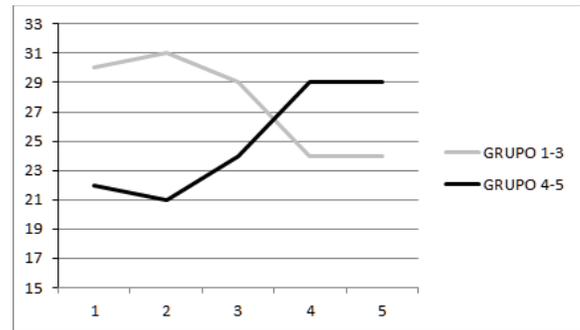


Fig 9 Averages count

Class grouping with "LK-Means" algorithm

In figures 10, 11, 12, 13 and 14, correspond to the graphing of data of the classes (1, 2, 3, 4 and 5), respectively. Here we show the separation of data according to the answers provided by the 136 people to whom the test was applied, and the identification of patterns that allow them to be grouped by category. In these graphs, people are identified through the horizontal axis and the response values are represented on the vertical axis.

- Class 1.- It is considered as the lowest level. 19 people were classified, whose response values of the group of classes 1-3 are higher than the value of the responses of the group of classes 4 -5. The range of responses is between 15 and 36, which allows us to conclude that the people who qualified for this group should participate in a training process in the topics raised in this document (see figure 10).

- Class 2.- It is considered as the low level. 13 people were classified, whose response values of the group of classes 1-3 are higher than the value of the responses of the group of classes 4 -5. The range of responses is between 19 and 32, which allows us to conclude that the people who qualified for this group should participate in a training process in the topics raised in this document. In addition, class 2 presents a difference of 5 points over class 1 in the responses of class 4-5, which is why it is considered to be a higher level than class 1 (see figure 11)..

- Class 3.- It is considered of average level. The algorithm grouped 44 people. In figure 12, it can be seen that the values of class 1-5 increase while those of class 1-3 decrease in a range of 16 to 37. Although the response values of class group 4-5 increase, we consider that the people who stayed in this class should participate in training programs in the use of social networks, as well as the people who stayed in class 1 and class 2.

- Class 4.- It is considered high level. The algorithm grouped 32 people. In this case, we can observe in Figure 13,

that the values that correspond to the classes 4-5 increase, while those of the classes 1-3 decrease. We believe that the people who classified in this group should only receive talks or written and / or electronic information about the use of social networks.

- Class 5.- It is considered at the highest level. The algorithm grouped 28 people. As can be seen in Figure 14, the response values are significantly higher in classes 4-5 compared to group 1-3. We consider that the people grouped in this class, like those of class 4, should receive talks or written and / or electronic information about the use of social networks.

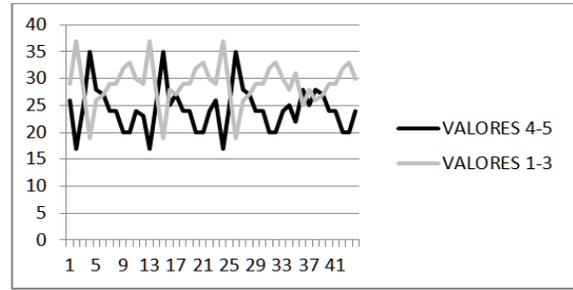


Fig. 12 Class 3. Average level

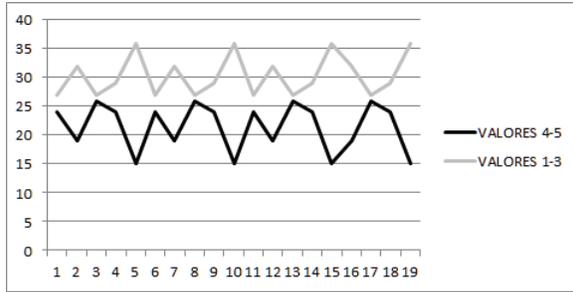


Fig. 10 Class 1. Lowest level

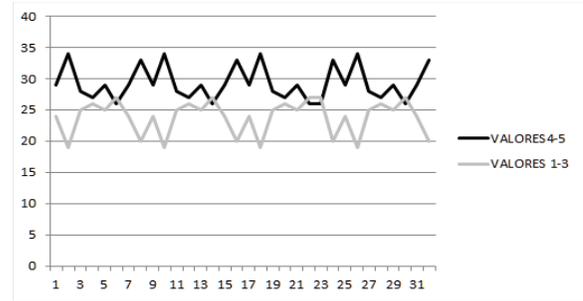


Fig. 13 Class 4. High level

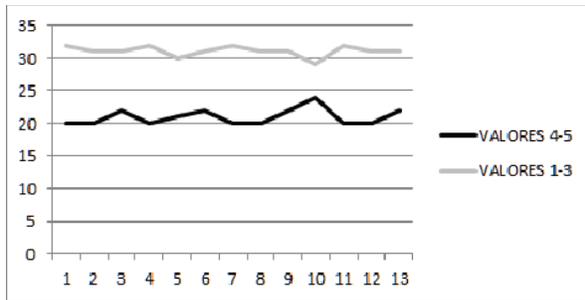


Fig. 11 Class 2. Low level

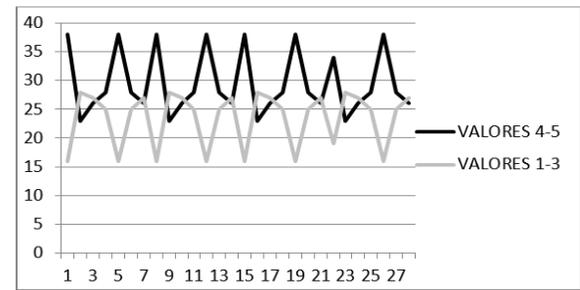


Fig. 14 Class 5. Highest level

Phase II. In this phase, elements of Galtung's conflict theory [9], [12], [13], [14] are presented. The first element "A" refers to the attitude of hatred, the element "B" represents violent behavior and "C" implies contradiction.

In order to know the attitudes of hate and violent behavior referred by university students in social networks, contradictions were raised where the problems of cultural diversity, security, human dignity, expression of freedom, aggression, extortion and invasion affect privacy.

The results show a variety of combinations between attitudes and behaviors referred by university students. To illustrate 8 of the 136 participants were randomly chosen. The evaluation is shown in table 6.

To demonstrate that class grouping works well, table 7 shows 5 examples of participants, one for each class. It can be seen that the largest number of responses of the person classified in class 1 (very low level) corresponds to group 1-3, and that the highest number of responses of the person classified in class 5 are in the range of group of classes 4-5.

With the values in Table 7, figure 15 shows the values of groups of classes 1-3 and 4-5 and in figure 16 the values of responses grouped by each of the 5 classes are shown.

Table 6.
Conflict in social networks

Subject *1	Hate attitude *2	Conduct violent *3
1	El 25% de las veces	El 0% de las veces
3	El 0% de las veces	El 0% de las veces
7	El 75% de las veces	El 0% de las veces
8	El 50% de las veces	El 0% de las veces
19	El 100% de las veces	El 0% de las veces
64	El 50% de las veces	El 100% de las veces
95	El 0% de las veces	El 25% de las veces
124	El 75% de las veces	El 25% de las veces

*1 Contradiction: Freedom expression.

*2 Attitude: "I hate when others want to limit my opinions and/or publications on social networks".

*3 Conduct: "I have threatened others in social networks for wanting to limit my freedom expression".

Table 7.
Number of responses per class of 5 participants

Id	Number of answers					Grouping		
	1	2	3	4	5	Class 1-3	Class 4-5	Class (LK-Means)
1	18	8	1	2	24	27	26	3
19	17	18	3	5	10	38	15	1
3	10	4	1	7	31	15	38	5
7	18	5	1	5	24	24	29	4
8	12	10	11	8	12	33	20	2

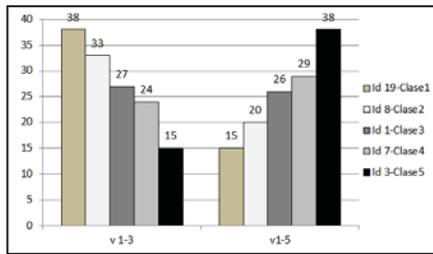


Fig. 15. Classification of users by group (1-3 or 4-5)

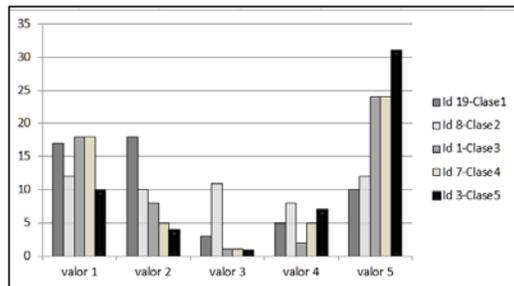


Fig. 16. Classification of users by class (1-5)

In the sample to which the instrument was applied, it is found that there is a low level of education in the appropriate use of social networks. We recommend continuing with research in the area to strengthen the evidence, as well as writing material on the appropriate use of virtual spaces directed to communities of university students.

VI. CONCLUSIONS

The applied test includes 53 questions with 5 possible values for each answer, which doesn't facilitate manual analysis and requires human resources and a long processing time.

The algorithm "LK-Medias" is effective to classify the people who answered the applied test. We found a coincidence of 90% with respect to the manual procedure applied.

We consider that the test can be applied electronically to nearly 30,000 university students from the University of Sonora, and apply the "LK-Means" classification algorithm, avoiding human errors and significantly reducing the time of manual processing.

The problem of cyberbullying is increasing, so it is important, in addition to the care for the victim, detect people with victimizer or observer profiles to provide them information on the proper use of virtual spaces and increase their level of education through programs and actions to reduce virtual violence.

The sample of people to whom the test was applied revealed the following indicators on the use of virtual social networks:

- 51% show violent relationships.
- 65%, show opposite to virtual violence.
- 58% refers knowledge and application of "Netiquetas".

REFERENCIAS

- [1] Instituto de Tecnologías Educativas. (s/f). World Wide Web. (s/n). Recuperado de http://www.ite.educacion.es/formacion/materiales/157/cd/ml_1_conceptos_basicos_de_internet/world_wide_web.html
- [2] Verdejo, M.A. (2015). Ciberacoso y Violencia de género en redes sociales, Análisis y herramientas de prevención. Universidad Internacional de Andalucía. ISBN: 978-84-7993-281-7. pp. 14-18
- [3] Maturana, J. (2009). Historia de Internet 2000-2009. MC. Recuperado de https://www.muycomputer.com/2009/11/17/actualidadesespecialeshistoria-de-internet-2000-2009_we9erk2xxdcs1811r633dmvsuhcb05ih8priucxkk9ushyv2wbfrvp7qk129ybf
- [4] Instituto Nacional de Estadística y Geografía (INEGI, 2017). Módulo sobre Ciberacoso. Recuperado de <http://www.beta.inegi.org.mx/contenidos/proyectos/investigacion/ciberacoso/2015/doc/702825084745.pdf>
- [5] Smola, A. & Vishwanathan, S. (2010). Introduction to Machine Learning. Cambridge University PRESS. ISBN 0 521 82583. pp. 32-34
- [6] Soto, C. y Jiménez, C. (2011). Aprendizaje Supervisado para la discriminación y clasificación difusa. Recuperado de <http://www.redalyc.org/articulo.oa?id=49622390003>
- [7] Navarro, B. (2017). Claves de la Inteligencia Artificial: Machine Learning y Deep Learning. Planeta Chatbot. Recuperado de <https://planetachatbot.com/claves-de-inteligencia-artificial-machine-learning-y-deep-learning-53a2032aaad>
- [8] Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO, 2017). Una nueva campaña mundial con miras a hacer frente al ciberacoso. (s/n). Recuperado de https://www.google.com.mx/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&uact=8&ved=0ahUKEwjg3MWgg8LXAhXjzQKHXAB6YQFggxMAE&url=http%3A%2F%2Fwww.unesco.org%2Fnew%2Fes%2Fmedia-services%2Fsingle-view%2Fnews%2Fa_new_global_campaign_to_address_cyberbullying%2F&usq=AOvYaw1CUsBZoGPOaZVTs1ra5vIU
- [9] Galtung, J. (1958-1973). Theories of conflict. Definitions, Dimensions, Negations, Formations. Recuperado de: http://gawharshad.edu.af/wp-content/uploads/2016/08/Galtung_Book_Theories_Of_Conflict_single.pdf
- [10] Gironés, J. (2013). Algoritmos. Recuperado de http://openaccess.uoc.edu/webapps/o2/bitstream/10609/71345/1/Business%20analytics_M%C3%B3dulo%20Algoritmos.pdf
- [11] Rodríguez, I. (2018). Comunicación virtual: Netiquetas. Boletín científico de la Escuela Superior de Atotonilco de Tula. Recuperado de <https://repository.uaeh.edu.mx/revistas/index.php/atotonilco/article/view/2892/2916>
- [12] Palacios, A. (2018). Tras la violencia, las tres erres de Galtung: reconstrucción, reconciliación y resolución. Global Affairs Strategic Studies. Recuperado de <https://www.unav.edu/web/global-affairs/detalle/-/blogs/tras-la-violencia-las-tres-erres-de-galtung-reconstruccion-reconciliacion-y-resolucion>
- [13] Realidad expuesta.org. (Productor). (2011). Comparte Johan Galtung en Monterrey experiencias sobre el conflicto, la violencia y la cultura de paz. [DVD]. De <http://www.realidadexpuesta.org/2011/10/compartir-johan-galtung-en-monterrey.html>
- [14] Pastor-Satorras, R. y Vespignani, A. (2004). Evolution and Structure of the Internet: A Statistical Physics Approach. Recuperado de https://books.google.com.mx/books?hl=es&lr=&id=EiySN0V4T_0C&oi=fnd&pg=PP1&dq=A+Brief+History+of+NSF+and+the+Internet&ots=JdoC0zMr_i&sig=-g2Uhv1g6jbtEgBeiLi5-x410ug#v=onepage&q=A%20Brief%20History%20of%20NSF%20and%20the%20Internet&f=false