Detecting Suicide Risk Through Twitter

Javier Fabra
Aragón Institute of Engineering Research (I3A)
Department of Computer Science and Systems Engineering
Universidad de Zaragoza, Spain
email: jfabra@unizar.es

Ana B. Martínez-Martínez
IIS Aragón
Faculty of Health Sciences
Universidad de Zaragoza, Spain
email: amarmar@unizar.es

Yolanda López-Del-Hoyo, María C. Pérez-Yus, Bárbara Oliván-Blázquez
IIS Aragón
Department of Psychology and Sociology
Universidad de Zaragoza, Spain
emails:
{yolandal, mcperezy, bolivan}@unizar.es

Abstract— **Mental illness is one of the main causes of illness worldwide. Currently, it is estimated that about 300 million people suffer from depression according to the World Health Organization (WHO). In this context, this work deals with the construction of a platform that allows to detect the risk of suicide using data from Twitter. This platform combines external emotional processing systems, clustering techniques and a system based on machine learning that facilitate the automatic classification of the information obtained. The entire process is articulated around a multidisciplinary team of professionals in Health Sciences and Information Technology, generating as a result a useful prototype for suicide prevention in the population.**

Keywords-mental health; suicide; prevention; Twitter; clustering; automatic classification.

I. INTRODUCTION

Mental illness is one of the main causes of illness worldwide. Currently, it is estimated that about 300 million people suffer from depression according to the World Health Organization (WHO). However, the provision of services for the identification, support and treatment of this type of mental illness globally is considered insufficient [1]. Although 87% of the governments of the different countries offer some type of basic care service for mental health problems, 30% of them do not have specific programs or budgets for mental health [1]. Furthermore, there are no definitive tests for the reliable diagnosis of most mental illnesses. The typical diagnosis is based on the patient's self-reported experiences, behaviors reported by family and friends, and the clinical examination of their mental state.

The data traditionally obtained through survey methodology are not a real-time reflection of the true state of mental and emotional health of individuals, which does not allow to offer a reliable estimate of the population's mental health. In Spain, suicide is the main cause of unnatural death, doubling the number of deaths in traffic accidents. The impact of suicide on families is devastating, when many of the deaths caused by suicide could be prevented [2]. Understanding how people communicate their suicidal tendencies is a cornerstone to preventing such deaths [3].

Social network platforms, such as Twitter, Instagram or Facebook are a source of faithful and real-time data on the emotional state of people [4]. In this work, we focus specifically on suicide-related aspects and propose the development of a platform capable of detecting and analyzing the emotional states of people globally and individually from the information available on social networks, specifically Twitter. The objective is twofold: on the one hand, to understand efficiently demand in the places and times that it occurs; on the other, to develop tools for suicide detection and prevention. In this first approach, we will focus on Tweets written in Spanish, although the methodology is applicable in other languages.

As the literature review points out, Twitter is one of the most widely used social media platforms worldwide, and has been the subject of numerous previous studies. Geographic, daily, weekly, and seasonal patterns of positive and negative affect have been observed in some of these studies [5][6]. The measurement of happiness levels in the populations of certain countries has also been analyzed [7]. Happiness was found to correlate with demographic and general well-being characteristics. The potential of Twitter to detect depression has also been studied [4]. The greater detection of emotional patterns specifically related to mental health variables is of special interest for global health, by helping to understand the places and moments of greatest demand (unmet) and the effective provision of resources that respond to these needs [8][9].

Previous studies have collected and classified the Tweets related to suicide [6]-[11]. However, these databases are still insufficient and the development of models for automatic detection is still rather immature. Although Twitter may provide an unprecedented opportunity to identify those at risk of suicide [10], as well as an intervention mechanism for both at the individual and community level, valid, reliable and acceptable methods for online detection have not been developed yet [12]. The best modus operandi for suicide prevention through social media remains to be clarified, so this work points to address open and existing problems in this area.

The potential of social media as data sources for improving people's health and quality of life is a relatively new phenomenon that society is beginning to value and

 ISBN: 978-1-61208-812-9

understand. This work represents a step within the enormous range of possibilities that the use of social media opens in the area of health. Accessing data and managing the large amount of information that can be collected (millions of daily entries) is a complex task that requires the integration of different technologies in order to integrate the various sources of information with a data processing system that allows us to obtain an adequate analysis of the data collected.

In this work, a deployment based on computer technologies for Tweet collection and analysis, as well as a system based on machine learning techniques that facilitate the automatic classification of the information obtained are depicted. This work deals with a subject and a set of technologies that have been previously considered by other researchers [13]-[16]. However, it represents a full framework, engineered and implemented using various technologies, and structured around a multidisciplinary team of professionals in Health Sciences and professionals in Information Technology. As a result, a useful prototype for suicide prevention and detection of real emotional states in the population has been developed.

The techniques developed in this work are easily adaptable to other contexts and studies in mental health and even in other sectors and institutions. The use of a technological framework based on languages and tools for the processing of information flows (streams) in real time facilitates the reuse of the main concepts and ideas underlying this work in other areas.

The remainder of this paper is as follows. Section II depicts the methodology and steps carried out and the architecture overview of our proposal. Section III presents implementation details, as well as the results from the experimentation conducted. Finally, Section IV concludes the paper and presents some related lines that currently are being addressed

II. METHODOLOGY AND ARCHITECTURE

Figure 1 presents the methodology of the solution that we propose, and that corresponds to the workflow to be carried out.

The first step is to obtain the Twitter entries from some keywords related to suicide. To identify potentially emotional tweets, a large vocabulary of emotional terms has been compiled from different sources, including *The Spanish adaptation of Affective Norms for English words* (ANEW) and the *Spanish dictionary of the Linguistic Inquiry and Word Count* (LIWC) [17][18]. ANEW provides a set of emotional normative scales for a set of words. Furthermore, LIWC is an analysis software that calculates the degree to which people use different categories of words across a wide spectrum of texts. Validation studies reveal that LIWC satisfactorily assesses positive and negative emotions.

One of the hypotheses of our proposal is that adding properties to the text contained in the Tweet facilitates and improves the identification and classification of suicide risk groups. Therefore, a series of properties associated with the text are obtained and added below, which are based both on external natural language processing systems and platforms and on internal algorithms that obtain the information through a text evaluation platform by part of selected reviewers in the area of Health Sciences and Medicine. The emotional vocabulary has been organized by combining the hierarchy of emotions by W. G. Parrott [28] and the *tree of emotions* by Shaver et al. [29]. Each emotional word has been classified into six categories of *primary emotions* of love, joy, surprise, anger, sadness and fear, with 25 subgroups of *secondary emotions*. This task has been carried out by integrating the execution of the Indico affective and emotional text processing tool [19].

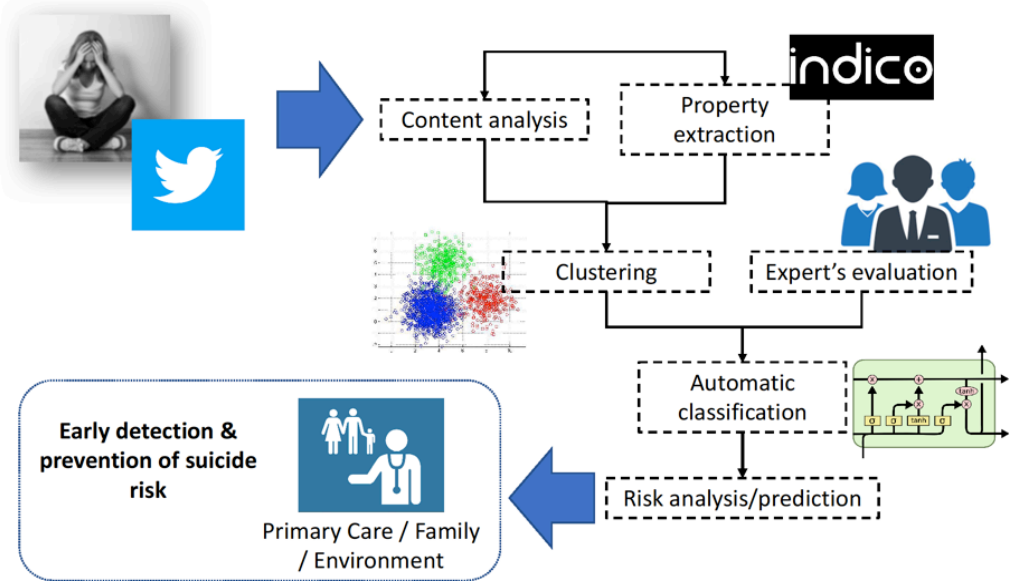

Figure 1. Methodology for early detection and prevention of suicide risk in Twitter.

Once the tweets with the properties have been obtained and annotated using Indico, a clustering is carried out to group the rich tweets. Clustering aims to generate data groups with similar characteristics. In our case, we want to identify suicide risk groups. The selected method for clustering is *k-means* [20]. This method is based on partitioning the data into k well-defined groups. To select this k value, two methods that offer good results have been used: the *elbow* method [21] and the *cross validation* [22].

The clusters that are obtained must be analyzed to know the characteristics of each one. This step must be carried out by a group of expert reviewers on the subject to make an evaluation of the clusters and to know the quality of the groups. The objective of this analysis is to validate that the clusters obtained correspond to suicide risk groups. At this stage, human coding is used to determine the degree of relationship of the classified Tweets, according to the judgment of the coding team made up of researchers from mental health and medicine. These researchers are specialized in suicide prevention and have training in detecting suicide risk.

Finally, in the last step an automatic classifier fed with the clusters and Tweets is created. This classifier is capable of receiving new Tweets and classifying them into one of the groups in order to identify whether there is a risk of suicide or not. In this work, the use of a *Long Short-Term Memory* (LSTM) neural network has been chosen. The result of this process is a tool that is capable of predicting whether a Tweet is in the suicide risk group or not. Since neural networks improve as they feed, the tool will always be in constant advance.

III. EXPERIMENTATION AND RESULTS

In this section, we will detail the implementation and experimentation that have been carried out in a first pilot study to validate the approach presented in this work.

A. Obtaining and adding properties

To obtain Tweets that contain the designed vocabulary of emotional-type words described in Section II, the Amazon Web Service infrastructure has been used. Multiple instances of Elastic Compute Cloud (EC2) [23] have been used to receive the data stream through an application implemented with NodeJS and using the Twitter API [24]. The summary data has been stored in MongoDB for access from the web and API. As a result, 3,051 context-sensitive Tweets have been extracted. The Tweets with emotional information have been then annotated using the Indico API, providing more information about the characteristics of the text.

Let us to briefly describe the properties that have been extracted:

- *Positivity*: value that represents the probability that the Tweet is positive or negative, if the value is greater than or equal to 0.6 the text has a positive feeling and if it is less than 0.6 the feeling is negative.
- *Engagement*: probability that the text will be bookmarked or retweeted by other people.
- *Emotional content*: they represent five values to determine the emotions that the author has expressed in the Tweet. The five emotions that are obtained as a result are: *anger*, *joy*, *fear*, *sadness* and *surprise*.
- *Personality*: set of four values to define the author's personality traits. The four personalities are: *extroversion*, *sincerity*, *sympathy* and *meticulousness*.
- *People*: there are sixteen values that represent the probability that the author adjusts to one of them. People are those described by Myers Briggs [25].

B. Clustering and expert's evaluation

Knime [26] has been used to implement clustering, although there are other equally valid alternatives (R or Weka, among others). Knime is a platform for data analysis that also integrates components for machine learning and data mining. Developed on the Eclipse platform and programmed, mostly in Java, it has a very comfortable interface that allows you to see the work done at all times. This is possible thanks to the *workflows*. A workflow is a graphical representation in which nodes and meta-nodes, a set of encapsulated nodes, are added to read data, do operations with them, or generate output data.

The work with Knime is mostly graphic, but it also allows a high degree of configuration, addition of external code and integration with other tools, making it the most suitable candidate to carry out the clustering phase in this project.

The corresponding workflows have been implemented to both calculate the optimal k and to cluster using the k-means method. As a result of the execution of the workflows, a value of k = 4 was obtained as optimal.

Taking this value as optimal-k, the clusters were obtained. The results of the clustering process with k-means for k = 4 show a well differentiated distribution in the input collection. Table I shows some data about the Tweets that correspond to each of the clusters (C0 to C3), as well as the positivity and the five emotions detected from the contextual information of the Tweet.

TABLE I. CLUSTERS OBTAINED WITH K=4

Cluster	#Tweets	Positivity	Anger	Joy	Fear	Sadness	Surprise
#0	654	0.68	0.15	0.32	0.14	0.27	0.13
#1	884	0.80	0.24	0.23	0.15	0.26	0.12
#2	604	0.29	0.25	0.11	0.16	0.40	0.08
#3	909	0.42	0.24	0.09	0.22	0.39	0.05

As it is shown, cluster 0 groups 21% of the Tweets of the input collection (3051 Tweets), cluster 1 (C1) 29%, cluster 2 20% and cluster 3 30%. Tweets classified in both clusters 0 and 1 show a positivity above the mean (0.56), with values of 121% and 143% with respect to the mean. Clusters 2 and 3, however, group negative Tweets (0.29 and 0.42, respectively).

 ISBN: 978-1-61208-812-9

The analysis of the emotions contained in the context of the Tweets allows us to detail the identified clusters. Tweets in cluster 0 show a low level of *anger* compared to the other clusters (65% of the mean, 0.22, compared to 107%, 113% and 110% with respect to the mean in the other clusters). The same occurs with *joy*, where both cluster 0 and cluster 1 stand out (172% and 125% respectively compared to the mean, 0.18, and 58% and 51% of clusters 2 and 3, respectively). The *fear* analysis shows that the Tweets samples contained in cluster 3 have a value above the mean (129% above the mean, 0.17), while the other clusters remain below it (79% for cluster 0, 89% for cluster 1 and 93% for cluster 2). From the analysis of *sadness*, values are obtained in line with expectations. Cluster 2 and 3 Tweets contain values above the average, 0.33 (122% and 118%, respectively), while the values of clusters 0 and 1 are below (around 80% of the half). Finally, the *surprise* analysis shows that the Tweets of cluster 3 are the ones that show the least surprise (56% with respect to the mean, 0.1), which indicates an apathetic or lazy profile in the context. The analysis of the other properties allowed to refine the results obtained. For example, *engagement* shows similar values for the four clusters (between 95% and 103%), indicating that there is not a predominant trend in any of the groups analyzed.

Subsequently, a human coding was performed to determine the degree of relationship of the classified Tweets. Reviewers were asked to conceptualize the task as the level of concern they would have if they viewed that post on their own online social network and whether they would consider the post to require further investigation from a friend, family member, or a third party. Tweets were individually examined and coded according to a classification system validated by the research team.

The analysis of the results obtained allows us to deduce that there is a direct relationship between the identified clusters and the possible suicide risk groups. Clusters 2 and 3 have a strong emotional content that is reflected in states of anger and sadness, fear and apathy in their content. The population that is at high risk of committing suicide expresses boredom with their life, routine, very negative content (some talk of death, or of being tired of living) and fear of certain scenarios or situations (the negative sensitivity towards simple situations in the day to day increases significantly).

In conclusion, the experts considered that the identified clusters fit perfectly into a classification of potential suicide risk, establishing it as follows: cluster 0 corresponds to tweets and individuals who have a very low risk of committing suicide, cluster 1 represents a low risk, cluster 2 represents a medium risk (to be monitored, since the parameters indicate that it is closer to a high risk than a low risk), and cluster 3 a high risk. Therefore, this suggests that the subsequent phases of analysis will focus on cluster 3, since it is the indicator that this information should be analyzed with priority and work together with the Primary Care services to identify the authors of the Tweet and try to implement prevention procedures as soon as possible.

C. *Automatic classification*

The LSTM neural network has been designed with Tensorflow [27]. Tensorflow is a machine learning framework developed by Google, programmed in Python and C ++. It stands out for its simplicity when it comes to building and training neural networks, but at the same time obtaining great results.

To create and train the neural network, we have developed scripts in Python. The network configuration is based on 10 LSTM hidden layers with 20 neurons in each layer. The output layer is a normal layer that produces a single output. The *loss function* measures the inconsistency between the actual values of the output and the predicted ones, in this case the mean of the absolute error is used as loss. The *optimization function* helps to minimize the loss and sight function. *Adam* (*Adaptive Moment Estimation*) has been used as it offers very low loss values. To evaluate the classifier, the *accuracy* will be used to obtain the percentage of correct answers in the predicted values.

Regarding training, the input data used are the properties of the Tweets together with the cluster to which they belong. These data have been divided into training data (70% of the total) and test data (30% of the total). To improve training, the data that accounted for 70% of the total has been partitioned to have 80% of that set as training data and 20% as validation data.

Each workout has been run 5 times to ensure that the result obtained is reliable. The training is done with 100 iterations, adding a field to end the training before completing them if the loss value does not improve in 3 consecutive iterations. When executing a problem appears, sometimes the best solution is not reached. To avoid this problem and ensure convergence, the 100 iterations have been left for the results shown below.

Training has been considered providing the network with all the properties in the Tweet. With K = 4, the evaluation function returns an accuracy of 93.34%. The confusion matrix obtained can be seen in Table II.

The first row represents the actual values that belong to cluster 0, the second row to cluster 1, the third row to cluster 2, and the last row to cluster 3. The columns appear in the same order. The results obtained in this training are very positive.

It can be observed that 24 data correspond to cluster 3, but belong to 1, and 16 data that belong to cluster 2 have been predicted as belonging to 3. For cluster 0, a success rate of 98.96% is obtained, for cluster 1 the percentage is 99.26%, for cluster 2 87.79% is obtained and for cluster 3 the percentage is 87.23%. In the cases of cluster 0 and cluster 1, the percentages are very close to 100%. The other two clusters do not reach 90% but they get a very high percentage as well.

The success rate that has been obtained is quite high, with an accuracy of 93.34% for the test data. Furthermore, the confusion matrix only presents a few false positives and negatives for cluster 3, assuming a fairly low percentage with respect to the successes achieved.

 ISBN: 978-1-61208-812-9

TABLE II. CONFUSION MATRIX

	C0	C1	C2	C3
C0	**191**	1	1	0
C1	1	**267**	1	0
C2	4	1	**151**	16
C3	0	24	12	**246**

IV. CONCLUSIONS

This work has addressed the problem of automatically identifying suicide risk groups in the context of Twitter. Based on a collection of Tweets obtained by keywords in relation to suicide, a series of properties have been added to enrich them. Then, clustering was applied using k-means with an optimal value of k = 4. The process has been validated by a group of experts in the context of mental health. This validation allowed establishing a relationship between the clusters obtained and the levels of suicide risk.

Finally, an automatic classifier has been built using an LSTM neural network. The neural network has been configured with 10 hidden layers and 20 neurons per layer. After training and evaluating the neural network with the test data, an accuracy of 93.34% has been obtained.

The proposal presented in this work shows very satisfactory and promising results. This approach is currently being extended, deploying the platform on an Amazon AWS infrastructure to automate the entire process and the different phases. As a result, direct connection with Primary Care Services is being worked on, so that the detection of a positive case allows initiating a series of actions to identify and contact the possible author of the content of the tweet. However, this is a very complex process that is being developed.

In addition to the work done, there are several ideas to improve the results obtained. Regarding the clustering process, other distance functions could be used. When implementing k-means the distance function that has been used is the Euclidean distance. Using other types of distance could improve the clusters obtained. A good option would be to use the Tanimoto coefficients to find the similarity and diversity of the sample set.

On the other hand, the properties could be hierarchized, studying the current properties to know which of them are more important in the generated groups. Then, we would have to add a weight to the most relevant properties so that they would have more importance when clustering or look for alternative techniques that allow us to apply these priorities.

We have also considered the possibility of using other classification techniques. Neural networks generate very good results, but there are other techniques, such as Random Forest or Support Vector Machine. It would be interesting to classify with these or other methods and compare the results between them.

ACKNOWLEDGMENT

This work has been supported by the JIUZ-2018-TEC-04 project, granted by Fundación Ibercaja and Universidad de Zaragoza. The authors of this paper want to specially thank David Fustero for his collaboration in the implementation of the platform; Claudia García Martínez for her help and support in conducting the study; as well as the experts in Health Sciences and Medicine for providing us with the revisions of the data used in this study, and for giving feedback on the results: Lara Barahona, Alberto Barceló, María Beltrán, Luis Borao, Roberto Buil, Daniel Campos, Luis Cortés, Irene Delgado, Paola Herrera, Laura Izquierdo, Andrea Lafuente, Andrea Llera, Marta Modrego, Alicia Monreal, Héctor Morillo, Mar Posadas, Marta Puebla, Marta Puértolas, Yaravi Rodríguez, Samara Sáez, Sara Sin, Sol Torres, David Valera, and Francisco Daniel Vinués.

REFERENCES

[1] R. Detels, "Oxford textbook of Public Health", in Oxford medical publications, Oxford University Press, 2009.

[2] J. M. Antón San Martín, "The impact of suicide on the family: the specific process of family grief", in Redes: Revista de Psicoterapia Relacional e Intervenciones Sociales (24), pp. 109-123, 2010.

[3] B. O'Dea, S. Wan, P. J. Batterham, A .L. Calear, C. Paris, and H. Christensen, "Detecting suicidality on Twitter", in Internet Interventions, vol. 2(2), pp. 183-188, 2015.

[4] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting Depression via Social Media", in the 7th International Conference on Web and Social Media (ICWSM), 2013.

[5] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth, "Temporal patterns of happiness and information", in A Global Social Network: Hedonometrics and Twitter. PloS one, vol. 6(12), 2011.

[6] J. Luo, J. Du, C. Tao, H. Xu, and Y. Zhang, "Exploring temporal suicidal behavior patterns on social media: Insight from Twitter analytics", in Health Informatics Journal, vol. 26(2), pp. 738–752, 2020.

[7] L. Mitchell, M. R. Frank, K. D. Harris, P. S. Dodds, and C. M. Danforth, "The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place", in PLoS ONE, vol. 8(5), pp. e64417, 2013.

[8] M. J. Paul and M. Dredze, " Discovering Health Topics in Social Media Using Topic Models", in PLoS ONE, vol. 9(8), pp. e103408, 2014.

[9] S. Alotaibi, R. Mehmood, I. Katib, O. Rana, and A. S. Albeshri, "A Big Data Analytics Tool for Healthcare Symptoms and Diseases Detection Using Twitter, Apache Spark, and Machine Learning", in Appl. Sci., vol. 10, 2020.

[10] J. Jashinsky et al., "Tracking suicide risk factors through Twitter in the US", in Crisis, vol. 35, pp. 51–59, 2013.

[11] M. J. Vioulès, B. Moulahi, J. Azé, and S. Bringay, "Detection of suicide-related posts in Twitter data streams", in IBM Journal of Research and Development, vol. 62, no. 1, pp. 7:1-7:12, 2018.

[12] H. Christensen, P. J. Batterham, and B. O'Dea, "E-health interventions for suicide prevention", in International Journal of Environmental Research and Public Health, vol. 11(8), pp. 8193-8212, 2014.

 ISBN: 978-1-61208-812-9

[13] P. Burnap, G. Colombo, R. Amery, A. Hodorog, and J. Scourfield, "Multi-class machine classification of suicide-related communication on Twitter", in Online Social Networks and Media, vol. 2, pp. 32-44, 2017.

[14] A. Abboute et al., "Mining Twitter for Suicide Prevention", in Natural Language Processing and Information Systems, vol. 8455, pp. 250-253, 2014.

[15] S. Fodeh, J. Goulet, C. Brandt, and A. T. Hamada, "Leveraging Twitter to better identify suicide risk", in Proceedings of The First Workshop Medical Informatics and Healthcare, 23rd SIGKDD Conference on Knowledge Discovery and Data Mining, PMLR 69:1-7, 2017.

[16] K. D. Varathan and N. Talib, "Suicide detection system based on Twitter", in 2014 Science and Information Conference, pp. 785-788, 2014.

[17] J. Redondo, I. Fraga, and I. Padrón, "The Spanish adaptation of ANEW (Affective Norms for English Words)", in Behavior Research Methods vol. 39, pp. 600–605, 2007.

[18] J. Pennebaker, M. Francis, and R. Booth, "Linguistic inquiry and word count (LIWC)", 1999.

[19] "Indico - Intelligent Process Automation for Document Intake, Understanding and Digitization", available at https://indico.io/ [Last access: 2020-09-11]

[20] V. Faber, "Clustering and the continuous K-means algorithm", vol. 22, Los Alamos Science, 1994.

[21] A. Hardy, "An examination of procedures for determining the number of clusters in a data set", in New Approaches in Classification and Data Analysis, pp. 178-185, Springer Berlin Heidelberg, 1994.

[22] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions", in Journal of the Royal Statistical Society. Series B (Methodological), vol. 36(2), pp. 111-147, 1974.

[23] "Amazon Web Services", available at https://aws.amazon.com/es/ [Last access: 2020-09-11]

[24] "Twitter for Developers", available at https://developer.twitter.com/en [Last access: 2020-09-11]

[25] "16 Myer Briggs personalities", available at https://www.16personalities.com/personality-types [Last access: 2020-09-11]

[26] "KNIME", available at https://www.knime.com/ [Last access: 2020-09-11]

[27] "TensorFlow - open-source framework", available at https://www.tensorflow.org/ [Last access: 2020-09-11]

[28] W.G. Parrott, "Emotions in Social Psychology" in Psychology Press, Philadelphia, 2001.

[29] P. Shaver, J. Schwartz, D. Kirson, C. O'Connor, "Emotion knowledge: further exploration of a prototype approach", in J. Pers. Soc. Psychol., vol. 52, no. 6, pp. 1061-1086, 1987.

 ISBN: 978-1-61208-812-9