CLASSIFICATION OF TWEETS WITH A MIXED METHOD BASED ON PRAGMATIC CONTENT AND META-INFORMATION

M. ESTEVE, F. MIRÓ & A. RABASA
University Miguel Hernández of Elche, Spain.

ABSTRACT

The sharp rise in social networks in any field of opinion has led to the increasing importance of content analysis. Due to the concretion of the texts published on Twitter from its limitation to 140 characters, this network is the most suitable for the analysis and classification of opinions according to different criteria. Therefore, there are multiple tweet analysis tools oriented from the perspective of semantics for trying to classify content characteristics such as feeling and polarity.

In this paper, the authors present a new approach to classification from a different perspective. The proposed approach addresses a complex mixed model from a perspective of pragmatics, the analysis of opinions in the context of their issuer carried out by a panel of experts, along with the classification of the type of discourse by considering the meta-information of the tweet.

From this new approach, the paper presents a complete and complex analysis process of Big Data, which covers all the characteristic phases of the life cycle: capture, storage, preprocessing and analysis of a tweets database. The aim is to classify the tweets as violent or non-violent in their reference to terrorist acts.

If the classification models based on the metadata of tweets reach acceptable levels of accuracy, this methodology will offer a reliable and semiautomatic alternative for tweet classification.
Keywords: analysis, big data, classification, social networks.

1 INTRODUCTION AND OBJECTIVES

'CRÍMINA' (centre for crime prevention and detection) is working on different fronts related to the prevention and early detection of criminal activities. One of their most current lines of work is oriented towards the classification and search for patterns in the *tweets* posted under different *hashtags* related to terrorist attacks.

Twitter is a huge source of data which can be gathered from opinions freely expressed by users. Five hundred million *tweets* a day are generated [1], which in turn provide 20.5 thousand million data a day to anyone who is able to compile this information. Twitter has a policy to share data freely, and provides APIs which allow a user with developer permits to access its database to compile the *tweets* generated in JSON format. To be exact, it provides API *Streaming*, which returns the *tweets* in real time (*streaming*) according to the consultation being made [2]. For example, the search for three types of *hashtags* (event, humanitarian and *StopIslam*). One of the drawbacks of this API is that the volume of data is 1% of all *tweets* generated in a given time and with the initial bias of the search, which means that the samples obtained from Twitter are representative and biased.

Once a 'clean' enough collection of *tweets* is available, a committee of experts in criminology classifies the content of the *tweets* according to the underlying pragmatics in the main body of the tweet, and supported by a well-defined ontology. Now in the analysis phase, the next step is to generate predictive models of Data Mining which will try to 'learn' the most fitting classification of the *tweets* and replicate it, but this time taking into account the *tweet*'s metadata alone and not its pragmatics. In general lines, this paper aims to present a classification method of *tweets*, starting from learning to classify according to pragmatics and

ISSN: 1755-7437 (paper format), ISSN: 1755-7445 (online), http://www.witpress.com/journals
DOI: 10.2495/DNE-V13-N1-60-70

ending by making classifications on the basis of the metadata alone, or the environmental parameters of the *tweet.*

2 STATE OF THE ART

To be able to make accurate enough classifications, it is necessary for the data sets to undergo adequate pre-processing. Although a lot has been published about pre-processing, in this case, as in many others, pre-processing will be a set of specific tasks aimed at cleaning and formatting the data, taking into account that the aim is to perform a classification task about types of discourse.

With respect to the analysis method used when dealing with Big Data from social networks, one of the analytical techniques most in demand is classification according to different criteria of opinions and content that users express on them. Through Data Mining, specially oriented towards extracting valuable information about very extensive data sets, a wide range of descriptive and predictive methods is proposed [3].

To be exact, classification tasks are considered predictive methods which aim to model the different possible outputs from a consequent variable, also called target class variable or dependent variable, based on a set of antecedent variables, also called attributes or independent variables. In the context of classification problems, the consequent variables are required to be nominal, also called categorical, discrete or non-numerical, while the attributes of the antecedent can be both nominal and numerical.

The results of a classification algorithm are sets of rules under the form:

(Attribute1, value); (Attribute2, value); … → (Consequent, value) (support, confidence)

where support refers to the likelihood of this tuple of the antecedent attribute occurring within the data set, and confidence refers to the conditioned likelihood of the consequent taking this determined value, knowing that the antecedent of this rule has occurred.

Some of the most well-known classification algorithms are ID3 [4] and variations of it, like C4.5 [5] which incorporates a series of improvements to the original algorithm, such as being able to deal directly with numerical attributes (like antecedents) that the algorithm itself segments based on the criteria of gaining information. These types of classification algorithms frequently produce outputs that are not only in the form of rule sets, but also in the form of trees called classification trees, whose interpretation is more immediate.

Nevertheless, the high number of attributes in the antecedent that characterize Big Data problems means that very often classification algorithms are not as accurate and efficient as one would wish. This is because not all the attributes that form the input data set are really important for predicting the consequent, that is to say, not all of them are equally important and many can even become dispensable. In these cases, it becomes necessary to resort to an automatic selection method of characteristics [6]. If a classification algorithm only works with the antecedents that really correlate with the consequent, more accurate and legible models are obtained.

3 INPUT DATA SET

Twitter is a *microblogging* network that permits reading and writing messages on the internet of no more than 140 characters, which are received by anyone who opts to receive them. It is a good source for observing communication because it is used to comment on everything that causes consternation or social interest. Furthermore, the limitation of messages, which on Twitter are no more than 140 characters, allows experts to easily identify the

communicative sense of the sender. Based on these limitations, 'CRÍMINA' has defined a basic taxonomy [7] of violent communication from the most basic messages known on Twitter as *tweets*.

The initial method, in this case, is the observation of the phenomenon of violent communication on Twitter as a consequence of criminal acts of terrorism against the population. In order to obtain a sample that would contain a significant amount of references to reactions towards the terrorist attacks two selection criteria were followed. The first was extraction from the three *hashtags* which at some time had been identified as *trending topic* in Spain in the 6 days following the terrorist attack. The second criterion was to try and balance the sample for its analysis from the point of view of the tendencies of the communicative content. So, within the *trending topic hashtags*, labels were selected: one was for humanitarian and supportive contents, another referred to the description of the event, and a third *hashtag* where negative attitudes towards the attackers and their background could be expressed, *#StopIslam*. With these initial criteria, data files with 41 variables each in JSON format were extracted through Twitter API liberated for this purpose, including both original messages and *retweets*.

Below, each *tweet* field is listed:

text, retweet_count, favorited, truncated, id_str, in_reply_to_screen_name, source, retweeted, created_at, in_reply_to_status_id_str, in_reply_to_user_id_str, lang, listed_count, verified, location, user_id_str, description, geo_enabled, user_created_at, statuses_count, followers_count, favourites_count, protected, user_url, name, time_zone, user_lang, utc_offset, friends_count, screen_name, country_code, country, place_type, full_name, place_name, place_id, place_lat, place_lon, lat, lon, expended_url and *url*.

4 COMPLEX SYSTEM METHODOLOGY

4.1 Overview of the analysis process

According to Hand [8] *"Data mining is the analysis of (often) large observational data sets in order to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner"*. The availability of large volumes has generated the need to convert them into useful information and knowledge [9].

In order to analyse and extract something useful from the data, they first need to be available. In some cases, this may seem trivial, being based on a simple data file to be analysed, but in other cases the diversity and size of the sources means that the data compilation process will be a complex task which requires its own methodology and technology [10]. In the case of Twitter, we start from a file in JSON format, which requires a cleaning process before it is analysed. Section 'Compiling *tweets*' outlines the process for compiling *tweets* by making use of API *Streaming* enabled by Twitter.

Data compilation should be accompanied by a cleaning process so that the data are in a condition to be analysed [10]. The benefits of the analysis and of the extraction of knowledge from data depend to a great extent on the quality of the compiled data. Because of the characteristics specific to Data Mining, it is necessary to carry out a transformation of the data to obtain a 'raw material' which suits the exact purpose (for example, discretize the date the *tweet* was created according to its time frame – early morning, morning, afternoon evening

Figure 1: System methodology.

and night), and the techniques that are to be used (automatic selection of characteristics and classification rules).

Once the data are pre-processed, an algorithm for *automatic selection of characteristics* is applied to them, from which we obtain a list of the attributes that most influence the class variable together with their numerical relevance. The class variable is *Do_Dv*, which has been created by the team of inter-judges from 'CRÍMINA' in order to categorize a *tweet* according to whether it is a *tweet* which expresses violent communication or not. In this way, we obtain the metadata of the *tweet* which most influence violent communication, leaving aside the pragmatics of the *tweet*.

Once the metadata of the *tweet* that most influences violent communication have been obtained, the next step is to generate the predictive models of Data Mining which will try to 'learn' the classification of the *tweets* and replicate it, omitting the pragmatics of the *tweets*. These models are represented in the form of a classification tree (Fig. 1).

4.2 Compiling *tweets*

In order to begin analysing and extracting something useful from data, they first need to be available. Twitter is a very attractive platform for investigators for several reasons, one of which is that Twitter is a huge source of data where users can freely express their opinions following the *hashtag* rules, which makes it easy for the investigator to follow the opinions and conversations that are generated on Twitter. The main function of the *hashtag* is to order a large amount of information which is generated on the social networks, allowing users to observe content related to that particular word.

The first step for extracting knowledge from data is to identify and gather the data to be analysed. As we have explained in this paper, the data that we want to extract from Twitter are related to the issues of terrorist attacks that have taken place in Europe in recent years. Therefore, by taking into account the *hashtag* rules that are applied on Twitter, three types of *hashtags* are then selected: event ('*#CharlieHebdo*'), humanitarian ('*#JeSuisCharlie*') and *#StopIslam.*

Twitter offers its users some REST APIs which through programming provide access for users to be able to read and write data from it. The REST API identifies the applications and the users who use *Twitter OAuth*, an industrial protocol for authorization. Because of this, in order to compile data it is necessary to go to the web https://apps.twitter.com/, log in with a user account and create an application which will have the above-mentioned authentication permits *OAuth.*

The next step is to know what our intentions are, that is to say, we should establish whether what we want is to retrieve past *tweets* (maximum 6 days before the current date) or compile the *tweets* that are being generated at that very moment. In our case, when a terrorist attack took place we were interested in compiling the *tweets* that were being generated at that very moment. For this reason, the API that was most suited to us is *API Streaming*.

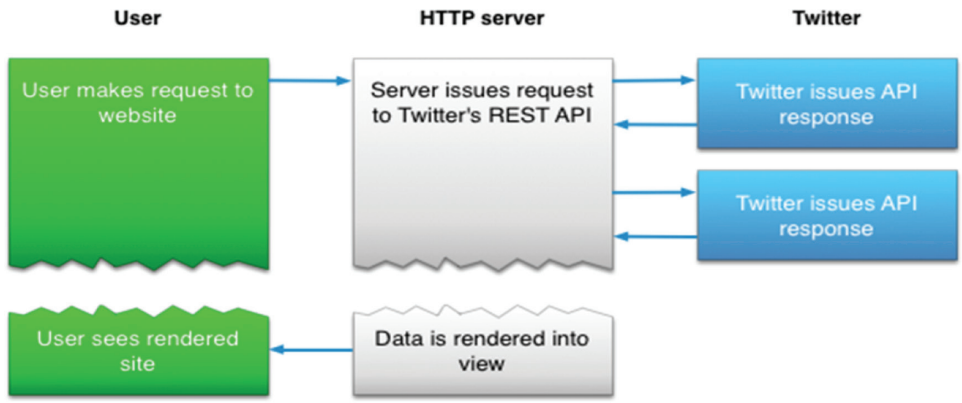

Figure 2: Twitter Apps [11], https://apps.twitter.com/

Figure 2 shows the internal process that is performed when, through programming, a request is made to obtain the data (with the initial bias of *hashtags*) to Twitter. The data that Twitter provides are 41 variables for each *tweet* collected dump, in files in JSON format.

4.3 Preprocessing

Although there is a large bibliography about pre-processing techniques, it is true that this phase is very different for each problem being [12]. It is also only a slightly automatic process where a series of changes are being made iteratively to the original data set. Below we outline some of the frequent steps for data set treatment directed at acting on attributes (columns) and also tuples (rows).

With respect to the data set attributes, the aggregation of characteristics consists of creating new fields to improve the quality, the visualization, or compressibility of the extracted knowledge. It is often necessary to resort to discretization of numerical (or date type) variables to establish ranges or segments that will facilitate the execution of some analytical models that are unable to work with continuous numerical variables. Operations like the elimination of attributes are frequently carried out because they accumulate a dispersion of values that are untreatable in practice. For example, the main body of a *tweet*, or for being primary keys or candidates that identify the registers unequivocally, like a user name. Those attributes, which as a result of some type of selection or bias present for the most part a single value (over 90% of the sample) are also usually eliminated because for practical reasons they are considered uni-valuated. A particular case can be found in the data sets with attributes with a high percentage of null values. Normally, if there is no sense in treating a null value as just another attribute value, and if the concentration of null values is more than 25–30%, this attribute is usually eliminated too. In special cases, where the variables follow a type gradient or known distribution, the null values can be replaced by interpolation. For example, absent values in a series of temperatures.

With regard to the tuples (rows) of the data set, it is necessary to eliminate those that present null values or clearly erroneous values in attributes which are potentially critical for the analysis. For example, a user's date of creation which is post the current date. Also, those registers with a high percentage of null values (above 24–30%) among all their attributes should be eliminated. On the other hand, the values outside the range (outliers) can be treated as null

values following the strategy of being replaced when possible or suitable or by eliminating their corresponding row.

To be exact, in the case that concerns us in this paper, the pre-processing actions carried out are the following:

- Elimination of the following attributes because of the high dispersion of values and text: *text, source, Name, Screen name, description* and *Agent.*
- Elimination of user identification attributes and *tweet*: *IdUser.*
- Elimination of attributes with a high value of null values (higher than 70%): *location, url, Time_zone, Geo_Enabled, Coordinate, Entities, Urls, In_reply_to_screen_name, Geo_Lon* and *Geo_Lat.*
- Elimination of the following attributes after being used for the creation of attributes derived from a greater analytical interest: *create_twt* and *create_user.*
- Creation of the attribute: *hoursAnt_Twt*, *month_usr*, *TextLength* and *Do_Dv.*

The new attributes created (to be included into the original attribute list), and the possible values they can have are as follows: *hoursAt_Twt* (hours that have passed since the terrorist attack and the creation of the *tweet*), month_usr (months that have passed since the creation of the user and the terrorist attack) and *TextLength* (length in characters of the *tweet*).

4.4 Classification method of the *tweets*

4.4.1 Preliminary analysis

The computational experience described below has been carried out on databases of *tweets* previously classified by experts. These databases are related to the attacks of Charlie Hebdo and Brussels.

Both experiments have seen very similar behaviours in three fundamentals of the study:

- The importance of balancing the sample so that input subsamples are chosen with a ratio of approximately 60% neutral *tweets* and 40% *tweets* of violent communication.
- The similarity of the predictions reached and their respective confusion matrices. In both cases, an average precision of 70% was exceeded, with a higher precision in the classification of neutral *tweets* and a higher false-positive rate.
- The three variables that are extracted as more significant in the face of the prediction of the type of discourse are repeated, albeit in different order.

To simplify the writing and to facilitate the understanding of the work, only the Charlie Hebdo set reclassification is shown.

After the *automatic selection of characteristics* was done, the confusion matrices have been obtained on the original database. Very high mean accuracies were also obtained (close to 98%). However, there was a clear example of over fitting produced by the very high proportion of neutral *tweets* (99.2%) against violent communication *tweets* (0.8%). So, the classification system presents a high number of false negatives that would make it inapplicable in real police contexts.

Thus, the sample of *tweets* was filtered, considering a hypothetical scenario of much social tension, which collected two-thirds of it with neutral content *tweets* and the remaining third with *tweets* of violent communication. The following subsection shows the complete analysis process on said sample.

4.4.2 Classification of the *tweets* as neutral communication (0) or violent communication (1). Charlie Hebdo case.

In this section, *tweets* classified as 0 (neutral) or 1 (violent communication) will be analysed, taking into account the following phases:

1. Selection of more influential attributes on the class variable (with *RandomForestClassifier* algorithms)
2. Confusion matrices and accuracy achieved (with *RandomForestClassifier* algorithms)
3. Classification trees (with *DecissionTreeClassifier* algorithms)

Next, the phases are described:

1. Selection of more influential attributes on the class variable.
   The *RandomForestClassifier* method is used first, to extract a ranking of the attributes most highly correlated with the class variable, as shown in Figure 3.
2. Confusion matrices and accuracy achieved.
   Again, *RandomForestClassifier* is used, but this time to extract the general accuracy and confusion matrix of the classifier, as shown in Fig. 4.

	feature	v_importance
8	TextoLength	0.242672
0	hoursAt_Twt	0.201431
5	statuses_count	0.114815
6	month_usr	0.092647
2	friends_count	0.090591
4	favourites_count	0.088896
1	followers_count	0.088064
3	listed_count	0.069195
7	Geo_Enabled	0.011690

duracion: 0:00:05.922788

Figure 3: Most influential attributes.

	precision	recall	f1-score	support
0	0.83	0.86	0.84	865
1	0.82	0.79	0.81	728
avg / total	0.83	0.83	0.83	1593

PREDICCION	0	1
REAL		
0	741	124
1	152	576

Figure 4: Confusion matrix and accuracy (*precision*).

The classifier has an average accuracy of 82%. Analysing the confusion matrix, it can be concluded that over the 1,514 instances used by the model, 804 (694 + 110) were neutral tweets ('0'). Of these, 86% (694) were well classified and only the remaining 14% would correspond with false positives (prediction of violent communication, which they really were not). In the case of violent communication ('1'), 710 cases were counted, of which 77% (548) were correctly classified.

3. Classification trees
   Using different parameters of the *DecissionTreeClassifier* (Python) model, this section shows and interprets one of the classification trees obtained.

The trees shown use the nomenclature of variables, as follows:

X[0] = hoursAt_Twt
X[1] = followers_count
X[2] = friends_count
X[3] = listed_count
X[4] = favorites_count
X[5] = statuses_count
X[6] = month_usr
X[7] = Geo_Enabled
X[8] = TextLength
X[9] = Do_Dv

A split binary, three-depth classification tree has been generated with.

The tree represented in Fig. 5 can be interpreted as follows (see the shaded branch): the variable that discriminates most is *TextLength* (*tweet* length in characters) rather than when it is less than 139 characters (≤139.5, i.e. including extra information recently allowed as URLs or links to images). Next, the model finds that the next most discriminating variable is *hoursAt_Twt* (hours elapsed between attack and *tweet*). Specifically, 45.5 hr is a critical threshold value, such that if the elapsed time is less than said value, the tree ends in a leaf node that collects 525 instances: none neutral (and this is very significant). The 525 instances are of violent communication.

4.5 Interpretation of more relevant results

As Figure 5 shows, and looking for the leaves of minimum entropy, the most relevant rules that can be inferred are the following:

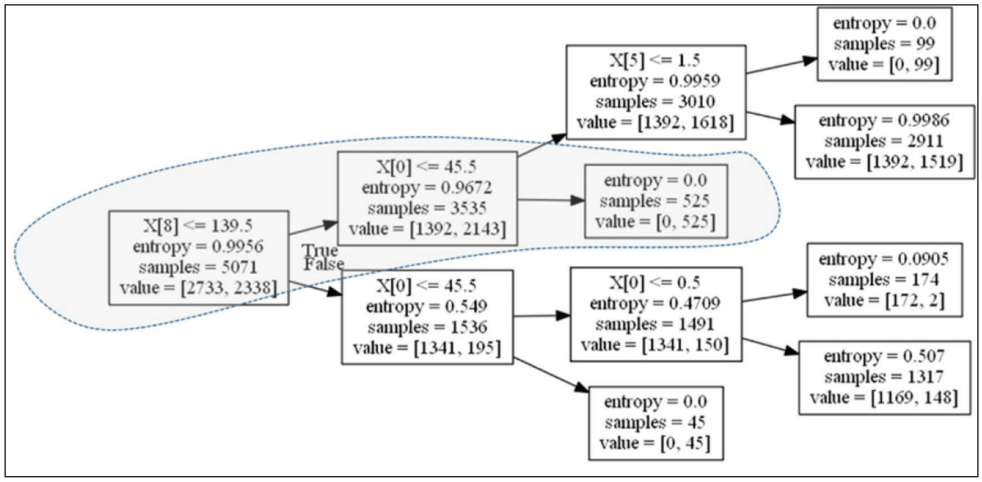

Figure 5: Classification tree.

- If the text does not include URLs or links, the time between the attack and the *tweet* are less than 45 hr and the user has already written a *tweet*, then all samples should be classified as violent communication.
- If the *tweet* incorporates URLs or links and less than an hour has elapsed between the attack and the *tweet*, then only 1.2% of the sample will be classified as violent communication.
- If the *tweet* incorporates URLs or links and it is broadcast more than 45 hr after the attack, then all samples should be classified as violent content.

5 CONCLUSIONS AND FUTURE RESEARCH

5.1 Overcoming the limitations of the language itself

It should be assumed that the pragmatic classification carried out by the experts is based on *tweets* posted in Spanish. This type of classification is dependent on the language the *tweet* is written in and to some extent it is mostly for the expert to judge the ambiguities, turns of speech and context of language. This inevitable point of subjectivity, dependence on language, disappears as soon as the classification of the *tweet* is based only on the environmental issues [13] and not on its semantic and pragmatic content [14].

5.2 Levels of accuracy reached based only on environmental information and not on pragmatics

These levels of accuracy in the classification cannot be interpreted as although the method followed is going to classify any data base of *tweets* with the same accuracy. However, it should be concluded that the model developed is able to classify the *tweets* in almost exactly the same way as the group of experts does and that in some way the algorithm is able to faithfully replicate the criteria of the pragmatic classification in a more objective model of environmental classification based on the meta information of the *tweet*.

On the other hand, as is normal, the accuracy of the classifiers decreases considerably as variables found to have little influence on the selection of characteristics are incorporated. In some scenarios, the accuracy decreases by 40%.

5.3 Possible police uses

The search for hateful or violent discourse potentially generated from criminal situations can become an intractable task to complete within short periods of time, as is evident in the number of *tweets* gathered in the time after a terrorist attack. However, the method proposed facilitates to focus attention on the most relevant variables in each case, that is to say those that have most influence when determining the final nature of the *tweet*. Likewise, and focusing attention on these variables, patterns of potentially conflictive *tweets* can be extracted with high accuracy in much shorter times. In this sense, the proposed system can act both as a filter when searching for opinions which are ‘dangerous’ for a particular group, and as a system of early alarms when recording a *tweet* whose pattern fits some type of undesirable discourse.

5.4 Other fields of application in crime prevention (radicalization)

Hateful or violent discourse is just another manifestation of the many feelings of users. 'CRÍMINA' is currently working on the classification of *tweets* from the perspective of radicalization, with an immediate practical focus on protecting the European population.

5.5 Other technologies to be incorporated: data streaming

The process presented in this paper corresponds to two completely different sequential phases: first collecting the *tweets* and afterwards their pre-processing and classification. A possible improvement could be to study the viability and suitability of incorporating data streaming technologies which would make it possible to carry out classification processes as data are being compiled. This improvement, which would probably entail a considerable reduction in computation times, is not a small matter at all. It would require automatizing pre-processing phase and pragmatic classification by experts and also coordinating it with the dynamic classification of the *tweets* based on their metadata.

5.6 The importance of adequate pre-processing

It is clear how important adequate pre-processing of data is, and how different pre-processing inevitably lead to different accuracies of the classification models. For this reason, another future line of research to be considered, and one that is absolutely critical, is the systematic analysis of the pre-processing phase and how to implement it, so that the highest possible accuracies can be reached in classification tasks of *tweets*.

5.7 Sub-samples selected to avoid over fitting.

In the two databases studied, it was verified that if a very unbalanced sample is considered regarding the amount of possible values in the class variable, the classifiers also provided severe cases of over fitting, with respect to the class variable. In a new research line, this fact must be formally checked on semi-synthetic sub-samples and under different loading conditions.

ACKNOWLEDGEMENTS

The present paper was carried out in the framework of research project DER2014-53449-R entitled "*Incitación a la violencia y discurso del odio en Internet. Alcance real del fenómeno, tipologías, factores ambientales y límites de la intervención jurídica frente al mismo*", from MINECO.

REFERENCES

[1] Internet Live Stats, Online http://www.internetlivestats.com/twitter-statistics/. (accessed March 2017).
[2] Morstatter, F., Pfeffer, J. & Liu, H., *When is it biased? Assessing the Representativeness of Twitter's Streaming API*. Ed. Cornel University Library, 2014.
[3] Rabasa, A., *Método para la reducción de Sistemas de Reglas de Clasificación por dominios de significancia* (doctoral thesis). University Miguel Hernández of Elche, 2009.

[4] Quinlan, J.R., *Discovering rules by induction from large collections of examples. In D. Michie* (Ed.), Expert systems in the micro electronic age. Edinburgh University Press, 1979.
[5] Quinlan, J.R., *Bagging, Boosting, And C4.5*., University of Sydney. Technical Report, 2006.
[6] Dernoncourt, D., Hanczar, B. & Zuckera, J.D., Analysis of feature selection stability on high dimension and small sample data. *Computational Statistics and Data Analysis,* **71**, pp. 681–693, 2014.
https://doi.org/10.1016/j.csda.2013.07.012
[7] Miró, F., Taxonomía de la comunicación violenta y el discurso del odio en Internet. *Journal of law and political science studies*, **22**(I), 2016.
[8] Hand, D., Mannila, H. & Smyth, P., *Principles of Data Mining*, Cambridge, MA: The MIT Press, 2001.
[9] Han, J. & Kamber, M., *Data Mining: Concepts and Techniques* (3th. ed.), San Francisco: Morgan Kaufmann, 2012.
[10] Hernández, J., Ramírez, M.J. & Ferri, C., *Introducción a la minería de Datos*, Pearson. Prentice Hall, pp. 19–45, 2004.
[11] Twitter Apps, online https://apps.twitter.com/. Accessed on March 2017.
[12] Wasilewska, A. & Menasalvas, E., *Data Preprocessing and Data Mining as Generalization. Data Mining: Foundations and Practice, 118 of the series Studies in Computational Intelligence*, pp. 469–484, 2008.
[13] Miró, F. & Johnson, S., Cybercrime and Place: Appliying Environmental Criminology to Crimes in Cyberspace. In G. Bruinsma & S. Johnson (eds), *The Oxford Handbook of Environmental Criminology*. Oxford: Oxford University Press, 2017.
[14] Burnap, P. & Williams, M.L., Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, **7**(2), pp. 223–242, 2015.
https://doi.org/10.1002/poi3.85