COMMUNICATION & SOCIETY

Special issue
Political communication in Uncertain Times

The Use of Supervised Learning Algorithms in Political Communication and Media Studies: Locating Frames in the Press

Javier García-Marín
jgmarin@ugr.es
Associate Professor. Department of Political Science and Administration. University of Granada, Spain.

Adolfo Calatrava
alatravag@gmail.com
Assistant Professor. Department of International Relations. Antonio Nebrija University, Spain.

Submitted
January 11, 2018
Approved
April 30, 2018

Communication & Society
ISSN 0214-0039
E ISSN 2386-7876
doi: 10.15581/003.31.3.175-188
www.communication-society.com

2018 – Vol. 31(3)
pp. 175-188

How to cite this article:
García-Marín, J. & Calatrava, A. (2018). The Use of Supervised Learning Algorithms in Political Communication and Media Studies: Locating Frames in the Press. *Communication & Society* 31(3), 175-188.

Abstract

To locate media frames is one of the biggest challenges facing academics in Political Communication disciplines. The traditional approach to the problem is the use of different coders and their subsequent comparison, either through statistical analysis, or through agreements between them. In both cases, problems arise due to the difficulty of defining exactly where the frame is as well as its meaning and implications. And, above all, it is a complex process that makes it very difficult to work with large data sets. The authors, however, propose the use of information cataloging algorithms as a way to solve these problems. These algorithms (Support Vector Machines, Random Forest, CNN, etc.) come from disciplines linked to neural networks and have become an industry standard devoted to the treatment of non-numerical information and natural language processing. In addition, when supervised, they can be trained to find the information that the researcher considers pertinent. The authors present one case study, the media framing of the refugee crisis in Europe (in 2015) as an example. In that regard, SVM shows a lot of potential, being able to locate frames successfully albeit with some limitations.

Keywords

Algorithms, Framing, Press, Spain, SVM, Refugees, Refugee crisis.

1. Introduction

Since the 1970s, probably because of its progressive institutionalisation, political communication has been linked to a certain methodological consensus, remaining close to positivist and quantitative approaches, which is particularly visible in the American academic context, with English being the 'lingua franca' in this area. All these elements have denoted a clear ethnocentrism. This stability has provided explanatory knowledge to (at present) such consolidated aspects as *agenda-setting*, *framing*, the formation of public opinion, the interaction between news

coverage during political campaigns and voting behaviour, or even the limits of information manipulation. Therefore, quantitative behaviour studies have been the dominant paradigm in the field.

Traditionally, Political Science, Social Psychology and some other areas under the label Mass Communication Studies, have given coverage to the analysis of the complex interaction between the world of communication and politics. However, the social and political dynamics of the past decade have brought about a context without precedent regarding social relations, political processes, and communication interactions. Due to the sophistication of research and massive data collection techniques, other proposals have acquired special relevance, such as those from neuroscience and genetics (Luengo, 2016). Indeed, research in political communication has been affected by all this.

Some authors, such as Lance Bennett and Shanto Iyengar (2008), suggest that in recent years a divide seems to have emerged within the edifice of political communication, indicating some scepticism regarding the capacity of the process of renovation of old concepts and the regeneration of methodologies to approach new scenarios, where structures, processes and actors seem to act according to different parameters. A new horizon of communicative processes has been defined, characterised by 'de-professionalisation,' with the irruption of mass information transmitters in the field of grass-roots citizen journalism; fragmentation, by the multiplicity of media, new supports and new alignments in the map of the information companies; and unpredictability, by a gradual complication of the modelling of these relationships and the consequent loss of a predictive capacity.

However, this environment of growing sophistication described coincides with a particular historical moment. On the one hand, greater technical resources have been developed in the field of automated/supervised analysis. On the other hand, a greater volume of data, available for analysis, has accumulated in the social sciences. In this new context, it seems that some of the classic tools employed in political communication, i.e. content analysis, have being displaced. Today, it is easy to verify a hypothesis by analysing the entire field of study, without having to restrict the process to a sample. Also, armies of encoders are no longer required, and complex processes for testing the reliability of fieldwork are not necessary anymore. The use of algorithms, big data, neural networks, and the adaptation of the techniques of machine learning or artificial intelligence to new fields of knowledge are all good examples of these new forms of research in political communication. Without a doubt, they involve moving one step forward in the development of the study of, for example, the impact of communication campaigns, the effects of the generation of frames in news coverage, or the way in which citizens are informed.

In addition, the scope of comparative studies is acquiring a new dimension with these innovative research strategies, since some of the traditional barriers, such as language skills or sample size, are significantly minimised. Therefore, the potential is undeniable and this makes possible the implementation of proposals based on comparative studies, such as those pointed out by Hallin and Mancini (2004), which became a constant reference of analytical framework.

This paper aims to explore the projection of the incipient use of machine learning (algorithms) in research in political communication. There is a need, in this field, for relocating objects of study in a context of increasingly fragmented audiences, extremely changeable communicative and political institutions, and actors readapting to them. The use of these types of research techniques may involve a decisive event: crossing a threshold in the improvement of traditional methodological strategies in political communication in general, for example in agenda setting or framing research. To achieve this goal, we will use as a case study the analysis of the Spanish press during the *refugee crisis* of 2015.

ISSN 2386-7876 – © 2018 Communication & Society 31(3), 175-188

2. Machine learning

Machine learning is the subfield of computer science that, according to Arthur Samuel, gives 'computers the ability to learn without being explicitly programmed.' (Koza, Bennett, Andre & Keane, 1996). Machine learning explores the study and construction of algorithms that can learn from and make predictions on data. Such algorithms overcome the need to follow strictly static programme instructions by making data-driven predictions or decisions, through building a model from sample inputs. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms with good performance is difficult or infeasible; sample applications include email filtering, detection of network intruders or malicious insiders working towards a data breach, optical character recognition (OCR), learning to rank, computer vision, etc.[1]

In mathematics, computer science, and related disciplines, an algorithm is a well-defined, orderly and finite operations list that allows us to find the solution to a problem. It can also be defined as the description of a pattern of behaviour that is expressed using a finite repertoire of actions and basic information, identified, well understood and achievable. This repertoire is called lexicon (Scholl & Peyring, 1991). In basic terms, an algorithm is nothing more than a set of serialised operations. What is, then, the novelty? It lies not in the existence of algorithms, known of since ancient times, but in the ability to create very complex algorithms due to the possibilities offered by the exponential increase in computing capacity, i.e. no longer may the lack of having a supercomputer be considered a limit to the ability to apply these tools. And, while in the 1950s the first computers allowed for the use of new investigative techniques such as surveys—and their scientific treatment by statistics—in present day machine learning may be a candidate for establishing a new a set of tools that complement the existing ones.

Hence, in our case, we can say that an algorithm establishes a sequence for classifying data into two or more groups based on conditions laid down within the algorithm itself (in the case of non-supervised algorithms such as the Latent Dirichlet Allocations, LDA) or by the user (in the case of supervised algorithms such as SVM, which we use in this paper).

Among the myriad of existing algorithms, we have chosen to work with Support Vector Machines. SVM is particularly suited to working with texts due to being one of the best tested algorithms in the industry and in the academic world (Joachim, 1998). Thousands of works use it to measure sentiment on Twitter, for example, or on other social media (Facebook, Instagram, etc.). It is also better than neural networks for long texts (such as news stories) and uses much less computing power than other algorithms.[2] It is also a supervised algorithm, which means that we can train it to classify information using our variables.

SVM is based on the idea that any linear model is valid for classification if the classes are linearly separable, sufficing to find a hyperplane that discriminate both sets. In other words, any regression technique can be used for classification if we separate a sample into two groups: one group for training, where values of 1 are assigned for examples of one kind and 0 for the rest; and another group where the value of the regression is calculated and assigned to the corresponding class, comparing it to the value given by the investigators (measuring, thus, effectiveness). Of course, linear models can only obtain linear decision boundaries, and thus there is the problem of how to use them for problems that are not linearly separable, such as those we usually find in the social sciences. In this case we can

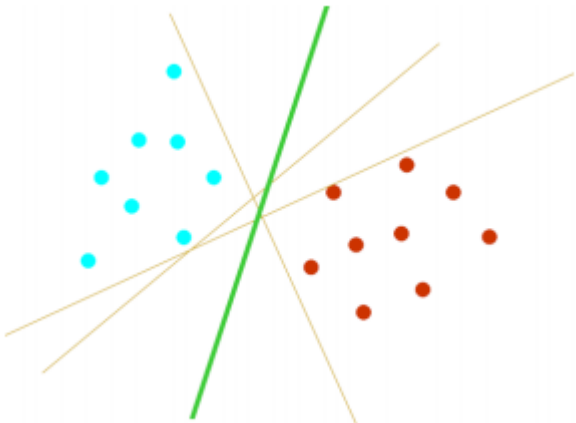
[1] See, for example, the works in fields like genetics (several authors, 2017), to detect plagiarism (VVAA, 2013), the cataloguing of images (Chum, Philbin, & Zisserman, 2008), etc.

[2] Especially less than other popular algorithms like Maximum Entropy, Random Forest and most Convolutional Neural Networks.

ISSN 2386-7876 – © 2018 Communication & Society 31(3), 175-188

apply a non-linear model (increasing the complexity of the model), and also build a linear classifier for a transformed space, where a linear border may represent a nonlinear boundary in the original space. In other words, if we increase the number of space dimensions to sort, we can apply a linear model to separate classes while maintaining the simplicity of the model, which will, at the same time, be applicable to initially non-linear problems.

Chart 1. Optimal separating Hyperplane



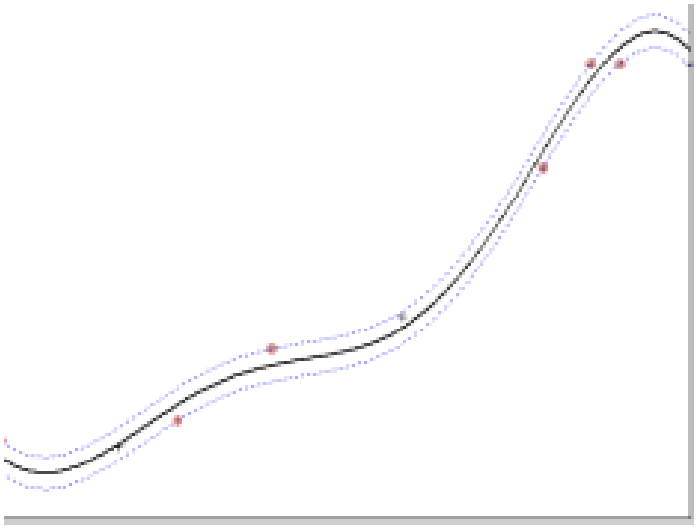
Source: Gunn, 1998: 11

Unfortunately, the universes to study are usually not presented in idyllic two-dimensional cases but, rather, an SVM algorithm should deal with a) more than two predictor variables, b) nonlinear curves of separation, c) cases where the data sets cannot be completely separated, d) classifications in more than two categories. But SVM is an algorithm designed to find linear classifiers in transformed spaces, being highly efficient in nonlinear cases. What is it that it makes? It creates a hyperplane of maximum margin, i.e., a *straight line* (if a two-dimensional space, which it is not) that separates two (or more) classes, and is located at the maximum distance of both. How does it do it? If the data are not linearly separable we can use non-linear applications, which are called *kernel*, of which there are several (polynomial, Gaussian, radial basis, perceptron and sigmoid[3]).

[3] Each one of them have different parameters. See, for example, Cherkassky & Ma (2004).

ISSN 2386-7876 – © 2018 Communication & Society 31(3), 175-188

Chart 2. Radial basis function regression



Source: Gunn, 1998, p. 35

Moreover, mistakes are admitted into the model through a constant (c), limiting the effect of data that is not classifiable in the final model. The value of C is usually given experimentally, (i.e., it tends to be 0 and increases as greater reliability of the model—without harming the classification—is required).

The potential of these research dynamics in the traditional study of the process of framing is beyond question. The development of political communication shows that the attention devoted by scholars in the field not only to the theoretical implications of framing, but also to methodological assumptions, has been significant. A macro approach to framing that examines media frames as outcomes of journalistic norms or organisational constraints assumes that individuals cannot understand the world fully and consequently actively classify and interpret their life experience to make sense of it (Scheufele, 2000, pp. 300-301). The individual´s reaction to sensory information therefore depends on schemes of interpretation called primary frameworks (Goffman, 1974). The most important implication for the field of mass communication research, then, is that there are various ways of looking at and depicting events in news media that depend on the framework employed by journalists (Scheufele, 2000, p. 301)[4]. Hence, there is an enormous potential for a connection to develop between mass media coverage of political events and the framework individuals use to interpret these events.

Most research in this specific realm of political communication has been elaborated over recent decades using content quantitative analysis as the main methodological technique. The limits of these interesting and useful resources are well known and mainly have to do with problems arising from subjectivity in the coding process, as well as limits to its consistency (reliability tests).

How can machine learning techniques be combined with text analysis? Algorithms work with vectors, so the first thing that must take place is the conversion of documents into vectors, i.e., into a matrix of computable data (a news story is, after all, a form of non-structured data).

One of the central problems in the analysis of texts, (text mining or natural language processing) is how to determine what a document is. One of the traditional approaches has

[4] For more information on framing theory, see: Entman (1993 and 2004), Scheufele (1999) or D'Angelo & Kuipers (2010).

ISSN 2386-7876 – © 2018 Communication & Society 31(3), 175-188

been to focus on the *words*. Indeed, on many occasions the frequency of terms (tf) is used to determine the importance of a word in a text. However, there are many words whose usage frequency really tells us little, such as 'the,' 'is,' or 'it,' etc. Normally, in order to limit this problem, there are lists of words that are removed from the corpus (called *stop words*). Another, more sophisticated, way to solve this problem is the *inverse document frequencies* (idf) statistic, which decreases the weight assigned to the most common words and increases it for those that are least used in a series of documents.

Regarding the frequencies of terms, the main step is to create vectors of terms, where the frequencies are the unit of analysis. The interesting thing, of course, is that these matrices are also inverse, calculating not only the presence of terms, but the absence of other terms that are present in other units of analysis (in our sample newspaper articles). According to these values, the matrix is weighted in one direction or in another. The algorithm tries to classify the data sets to homologate it to points on a plane (the matrix displayed as a spatial dataset) and it is then that it establishes the classification.

Tf-idf has limitations, such as not considering synonyms, or forms in the plural or of a different genus (Ramos, 2003). The latter is easy to solve through the elimination of ambiguity in words, with the use of libraries intended for this, referred to as stemming.

3. Design

To test these techniques of data cataloguing, we decided to analyse the existence of two exclusive frames in the Spanish press for the year 2015. The two frames are, from our point of view, quite clear and refer to the sections of the news describing the problem of Syrian refugees in 2015. Over the course of that year, the *European refugee crisis*, as the media called it, took place. According to UNHCR, in 2015, more than one million people—grouping together emigrants and refugees—arrived in Europe by sea (UNCHR, 2016, p. 7).[5] This was a fourfold increase compared to the previous year. Over the first six months of 2015, Greece replaced Italy as the entrance door for migrants to Europe. The corridor was from Turkey to Greece and through the Balkans en route to northern European countries. According to UNCHR, this new flow of people is highly related to the conflicts in Syria and Iraq (UNCHR, 2015).[6]

Hence, the questions we want to answer are: do the media use a unique frame to present the news related to the *refugee crisis*? When do the media use one frame or another? To prove the effectiveness of this tool in order to classify news, we have chosen news and opinion articles from the main Spanish newspapers: *El Pais*, *El Mundo*, *ABC, El Periódico* and *El Correo*. The time frame for the analyses is from 1st January, 2015 to 31st December, 2015, so effectively covering the whole of 2015.

We also used the Lexis-Nexis database, using the word '*refugiado*' (refugee in Spanish) in the title or the body of the article. We got 6,224 units, but after a first review our corpus was composed of 4,595 news stories.

---

[5] The terms *refugee* and *migrant* are not interchangeable. The former refers to persons who flee armed conflicts or persecution (for political, ethnic, religious, sexual, etc., motives). It is a political concept because these people can apply for asylum under international legislation. (1951 Geneva Convention) The second refers to persons who choose to move mainly to improve their lives. Usually, the boundaries of both terms are not well defined and, in many cases, placing an asylum seeker in one group or the other is a political decision (Onghena, 2015).

[6] This can be proved by checking the asylum seekers registered in EU member states in 2015. For the first time, the number of seekers reached 1.2 million. Of them, 29% were Syrians, 14% Afghans and 10% Iraqis (Eurostat, 2016). Two things are significant here: first, there was a two-fold increase in asylum applications and second, more than a third were made in Germany.

ISSN 2386-7876 – © 2018 Communication & Society 31(3), 175-188

Reading news stories from our sample, it was clear that the *refugee crisis* was not covered under a unique frame. Instead of this, we realised that most of the news stories could be classified into two groups (not totally mutually exclusive but almost):

(1) The news story refers to aspects of the humanitarian situation of the refugees and migrants, the danger of the road to Europe or the nature of the persecution in their countries of origin. This is what we called the *human rights frame*.

(2) The news story refers to aspects of the securitisation of the migration flows. This news story draws on the arrival of *these people* as a threat to Europeans. This is what we called the s*ecurity frame*.

Both frames proposed can be described according to Entman's categories[7]. Regarding to the *human rights* frame, the problem definition refers to people fleeing the war and who must face the dangers of a long trip to Europe across the Mediterranean, crossing countries that do not want them where they are mistreated. They are seen as refugees and compared to the European refugees of the former European wars (i.e. in the Spanish news they are usually compared with refugees of the Civil War in the 1930s). The problem causes are violence or wars in their countries (Syria, Iraq, and Libya) and their collapse. Western countries' foreign policy is considered a key factor to explain what is happening in the Middle East. So, this statement is linked with the moral judgement: European countries are mostly responsible for this situation, so they must help this people. Also, if the EU want to be a humanitarian actor, it should behave as such; they must respect the human rights regulations we signed. The proposed solution for the frame is simple: Europeans must help this people opening their borders, securing their trip to Europe, facilitating the asylum processes and giving them the resources to start a new life.

One of the previous assumption was that the *human rights* frame would be the most important in terms of quantity. But, especially in columns and editorials, we could find the other frame we want to study, the *security* frame. This frame is related to the securitization of the *refugee crisis*. This phenomenon has been observed during the last decades: Governments and International Organizations have included the migration topic to their security agendas, being more evident after the 9/11 attacks (Díaz & Abad, 2008; Ferrero & López, 2012; Calvillo, 2017)[8]. We understand this frame as a counterpoint of the human rights frame. Hence, we decided not only to include news stories or opinion pieces that consider the refugees as an existential threat to Europe –which imply urgent measures-[9]; but also, we included this category when the refugee phenomenon was only framed as a collateral element of a geopolitical conflict[10]. The problem definition in this frame is that a lot of refugees (*too many*) could come to Europe. Adding to those who are already there. Countries cannot afford to have so many refugees, mainly for existential issues. They represent a threat for European *civilization*. So, the cause of the problem is not just the high number of refugees, but their different culture and religion since they are Muslims. Problems of integration and the threat of radicalization are placed on the table. There is even another subversive threat: terrorists can move into Europe camouflaged among refugees. The moral judgment would be, then, 'us first': Europeans must protect their societies first and later, to think in welcoming the refugees. The proposed solution is to control the flow of migrants, and to check thoroughly the credentials of asylum seekers. Furthermore, the refugees should adapt to European culture.

---

[7] The four categories are: problem definition, its causes, moral judgment and proposed solution (Entman, 2004).

8 Some evidence of this is the inclusion of human trafficking as a threat for Europe in the European Security Strategy of 2003.

9 This refers to the definition of an act of securitization of A. Collins (2013: 153).

[10] I.e. "Russia backing El Assad Government and attacking Syrian opposition forces leads to more refugees".

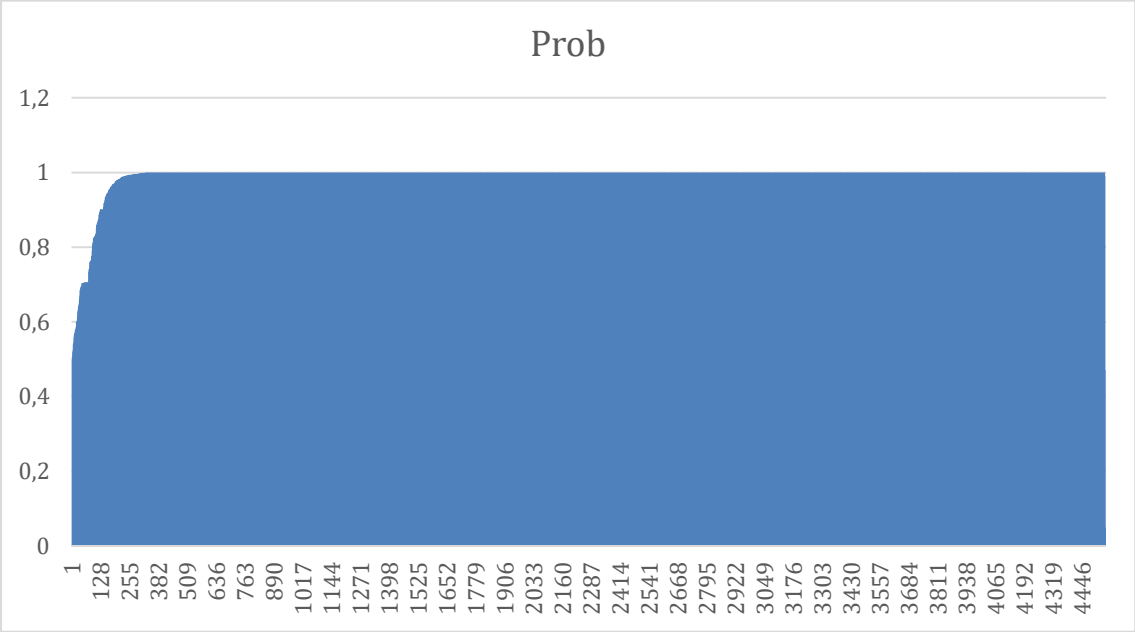
ISSN 2386-7876 – © 2018 Communication & Society 31(3), 175-188

Our hypothesis is that the two frames located in the press (*human rights* and *security*) were not together in the news stories. Here, the context is very important in explaining why one of the frames was more dominant than the other: the tragic death of some refugees in the sea (especially that of children) or the racist behaviour of some police forces' members would lead to the use of the *human rights* frame; while the terrorist attacks in Europe (such as the attack in November 2015 in Paris) would foster the use of the security frame.

The following step was to transform this unstructured data into a more structured form. We decided to employ R because it is easier to use than other languages (like Python, at least for us), and it is also very easy to import libraries to work with data in multiple forms. The implementation of SVM was 'libsvm[11]' (Chang & Lin, 2011) through 'e1071' (Grün & Hornik, 2011; Meyer & Wien, 2001), and the library to transform the data into a tf-idf matrix was 'tm' (Meyer, Hornik & Feinerer, 2008). Both are widely used open-source software. Through this last library, the data set (the text) was transformed, removing any kind of punctuation (and the Spanish 'ñ'), capital letters, numbers, whitespaces, and *stop words*, and remaining words were stemmed.[12]

After that, the corpus was transformed into a term-frequency matrix and, after that, into an inverse term-frequency matrix, which was used to train and apply the SVM model. To do so, the authors analysed and classified 342 news stories according to the frames located before (*human rights* and *security*), and trained the model with 280 of them, testing it with the remaining ones (62). Since SVM was unable to classify correctly the 62 units of the test data using the default linear model, we tried other kernels and configurations. The final configuration was the radial kernel with default values (c=1), and the reliability obtained was 85%, high enough to be useful from our perspective. The last step was to apply the model to the rest of the sample, more than 4,000 news stories. Another useful feature of the model is that it estimates its own effectivity in each case, so it is easy to see in which cases the model was unsure about which frame to apply to a given unit.

Chart 3. Accumulated probability given by the model on its own prediction over the sample



[11] A very good website with information about the library can be accessed via this link: https://www.csie.ntu.edu.tw/~cjlin/libsvm/

[12] Users should consult Meyer et al. (2008) to take full advantage of pre-processing possibilities.

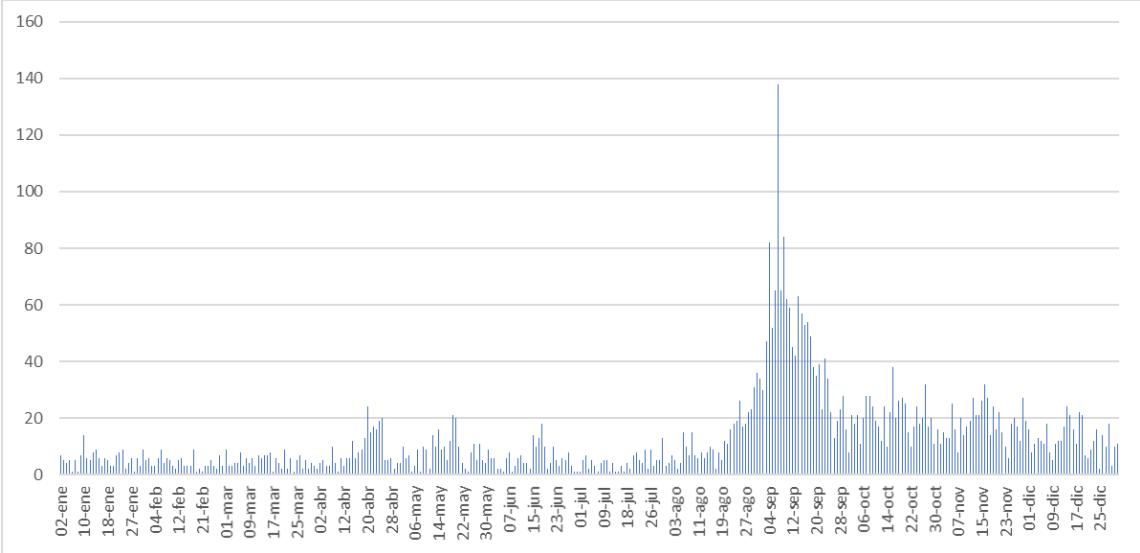
ISSN 2386-7876 – © 2018 Communication & Society 31(3), 175-188

Chart 3 shows that in most cases, the model assigned a very high probability on its own predictions (close to 1). However, in our experience, after analysing another random sample, the reliability of 85% was still present, and some stories were covered under a different frame than that predicted by the model. But, in those cases where the model was unable to offer a clear prediction (close to 0.5), it was clear that the news stories in question were not referring to the refugee crisis. Hence, another useful feature of the algorithm can be to clean a sample with a very high reliability.

4. Analysis

Once the data was classified, a more orthodox analysis was conducted. First, Chart 4 illustrates in greater depth that the interest of the press in themes grows significantly after August, in all the papers and in both frames.[13]

Chart 4. Number of stories per day



If we look at the events, there are three reasons that explain the increase, and they can be checked in the news. From minor to major importance: (1) since August 2015, different European governments suspended the Dublin Accord for regulatory procedures for asylum application. Germany was the first to do so, and later other countries such as the Czech Republic, Austria or Hungary followed. Similarly, some countries decided to close the borders inside the Schengen area. We must understand these actions in a context of civil society's criticism of the European Union's decisions. (2) Since the end of September, the Russian government decided to act more directly in the Syrian conflict. These decisions had some geopolitical implications related to the role of Russia, the United States and the European Union in the Middle East. Many of the news stories during these months are linked to this event. (3) The most important specific event was the publication of the photograph of Alan Kurdi, a three-year-old Syrian child (Kurdish) who died, having

[13] In the first eight months, we can note a brief increase in the numbers of news stories from April. In that month the public and the authorities started to realise that a significant number of people were coming to Europe from Syria and the Middle East. Use of the term 'refugee crisis' was already generalised from 2014 but in April 2015 the number of deaths at sea rose to record levels (UNCHR, 2015).

ISSN 2386-7876 – © 2018 Communication & Society 31(3), 175-188

drowned in the Mediterranean, on 2nd September. The picture and the story behind it (a Syrian family trying to reach Europe and eventually Canada) went viral. It soon emerged as a reference for the *refugee crisis*. The impact was global and criticism of European governments and institutions was generalised.

Table 1. Use of frames by newspaper

	ABC	El Correo	El Mundo	El Pais	El Periodico	Total
Frame: *Human rights*						
Jan	11	12	8	21	3	**55**
Feb	8	15	7	16	7	**53**
Mar	6	25	9	23	2	**65**
Apr	15	16	16	33	8	**88**
May	3	16	12	29	14	**74**
Jun	13	12	15	11	8	**59**
Jul	10	5	5	12	11	**43**
Aug	49	26	20	57	42	**194**
Sep	178	158	111	169	114	**730**
Oct	57	56	41	51	46	**251**
Nov	29	37	35	31	31	**163**
Dec	28	42	25	29	24	**148**
Total *Human rights*	407	420	304	482	310	**1923**
Frame: *Security*						
Jan	17	8	10	58	10	**103**
Feb	5	15	10	24	4	**58**
Mar	18	12	12	32	8	**82**
Apr	35	22	13	58	18	**146**
May	43	23	15	69	25	**175**
Jun	24	19	15	33	21	**112**
Jul	16	13	9	26	21	**85**
Aug	59	30	21	50	33	**193**
Sep	186	119	89	202	86	**682**
Oct	98	67	36	114	53	**368**
Nov	94	70	51	116	51	**382**
Dec	62	46	30	66	35	**239**
Total Security	657	444	311	848	365	**2625**
Total	**1064**	**864**	**615**	**1330**	**675**	**4548**

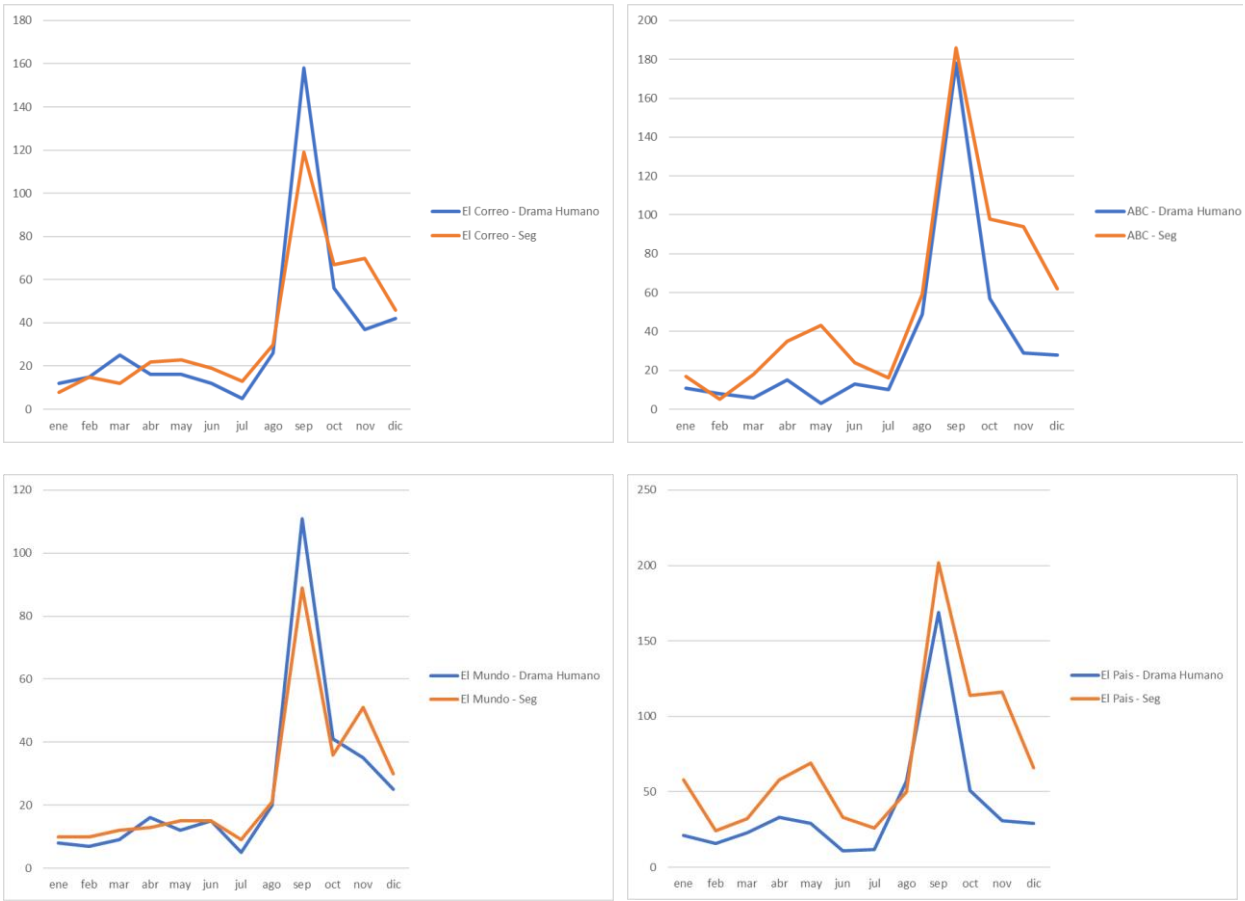
The numbers we can see in Table 1 (or, more visually, in Chart 5) show that one of our initial assumptions is false: there are more news stories using the *security frame* than the

ISSN 2386-7876 – © 2018 Communication & Society 31(3), 175-188

human rights frame, 57.1% of news pieces versus 41.8%.[14] Also, it is significant that the *security frame* is used more regularly throughout the entire year, not only in the second semester of the year when the intervention of the Russian forces in Syria was much more direct. So, even when the *refugee crisis* was a major topic in the Spanish (and European) media, on many occasions the approach to the topic was not humanitarian, but the issue was highlighted as a cause or consequence of a security problem.

If we analyse by paper, *ABC* and *El Pais* stand out for their use of the *security frame*, especially in late December (see Chart 5 for greater clarity).[15] This can be explained by the fact that the news stories related to military intervention by Russia are more numerous in these papers.

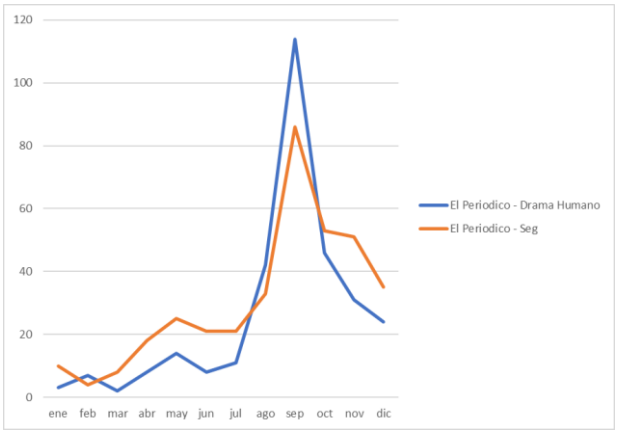
Chart 5. Use of frames per newspaper and month



[14] We must realise that both frames are not completely exclusive and there can be news without any frame or with an inconclusive frame.

15 The total number of news stories coming from each paper is not important because it could depend on the database used to get the news. What is more significant is the relationship between one frame and the other. El Pais has a rate of 1.7 (times greater use of the security frame) and ABC 1.6, while in the other it is close to 1.1.

ISSN 2386-7876 – © 2018 Communication & Society 31(3), 175-188



Another very significant element from the *security frame* is the link between the refugees and the *Islamic State*, in two ways: as a consequence of the conflict in Syria and as a possible way for the terrorists to enter Europe, camouflaged. The latter was usually criticised in most news and opinion articles, which may denote that there could be a frame for this already. The relationship between the news stories on refugees and ISIS can also be noted after the Paris attacks in November.[16] We can see that the number of news stories about refugees increased between October and November and, in the case of *ABC* or *El Periódico*, that the decrease that had begun in September began to slow down.

One quick conclusion we can draw relates to how the relationship between the two frames works. We can see from the charts that the two frames run parallel to each other—even when an event such as the publication of the photo of Alan Kurdi and the subsequent critical reaction against Western politics[17] (the increase of the *human rights frame*) is parallel—except for the peak in some newspapers. Thus, the issue is presented in the press using both frames at the same time. The only exception is after the Paris attacks, when the increase of the *security frame* is not followed by a similar increase in the *human rights frame*.

5. Conclusions

The first conclusion we wish to mention here is obvious: machine learning techniques are useful for researchers in the social sciences. We have demonstrated that creating research on framing is possible without human coders, at a fraction of the time and cost. After all, it took the computer less than an hour to code the whole sample (after the authors coded 342 news stories 'by hand'). Of course, this is not the first time that this has been achieved, as the scientific literature is full of quantitative analysis on framing. However, most of them used only word frequencies or other similar techniques (similarity analysis, word distance, etc.). All of these techniques encountered similar problems when approaching the real meaning of these words (like 'terrorism'). Machine learning techniques, at least those based on supervised algorithms, are somewhat different. First, because they do not rely solely on word frequencies, since each algorithm weighs frequencies and inverse frequencies; and, second, because the researchers *teach* the algorithm qualitatively. This algorithm tries to *learn* how to separate those groups created by the researchers. This implies that the amount of training is correlative to the complexity of the task. That is the reason behind choosing

[16] We are referring to the terrorist attacks in Paris and Saint-Denis on Friday 13th November. The attacks caused 130 deaths (almost 89 in the Bataclan theatre) and almost 400 injures.
[17] The strength and scope of the reaction may indicate that there was a very strong latent criticism to European policies.

ISSN 2386-7876 – © 2018 Communication & Society 31(3), 175-188

such a straightforward issue for this paper. In some cases, we are sure that the complexity of frames will make the technique unfeasible. But, in other cases, it can be an easy technique to use. However, since there are so many algorithms and other similar tools, maybe it is just a matter of choosing a different one, which constitutes a nice source of research for the coming years.

The most important consequence, at least from our perspective, is the *black box* effect. These algorithms are so complex that a lot of knowledge is required in order to know what they really do. This can be a minor problem in supervised algorithms, like SVM, since it is the researcher who tells the tool what to look for. It can, however, be a major problem in non-supervised algorithms (like LDA). However, the danger of using tools without a complete understanding is still there and can be a real issue when dealing with big databases, where a small mistake could lead to a big error. An example could be the parameters of each kernel in SVM. Most kernels have different parameters, according to the mathematics in use. For example, we used a radial basis kernel in this research. Radial basis kernels have two main parameters, C—as already explained—and *gamma*. Both parameters have default values but, if we change them, the effectiveness of the model is altered without our knowing the exact changes that have taken place in it. Furthermore, 'the correct choice of kernel parameters is crucial for obtaining good results. This essentially means that an extensive search must be conducted on the parameter space before results can be trusted, and this often complicates the task' (Meyer, 2007: 8).

To conclude, the authors firmly believe that although machine learning techniques may present challenges, they are useful for the social sciences. Therefore, the academic community should pay more attention to the dynamics mentioned previously, as many other fields of study are doing already.

References

Bennett, W. L. & Iyengar, S. (2008). A new era of minimal effects? The changing foundations of Political Communication. *Journal of Communication 58*, 707–731.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST) 2*(3), 27.

Cherkassky, V., & Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. *Neural networks* 17(1), 113-126.

Chum, O., Philbin, J., & Zisserman, A. (2008, September). Near Duplicate Image Detection: min-Hash and tf-idf Weighting. In *BMVC* 810, 812-815.

Churchland, Patricia (2008) "The impact of neuroscience on Philosophy", *Neuron 60*(3), 409-411.

Calvillo, J. M. (2017). Actualización del régimen jurídico internacional de los refugiados. *Actas III Congreso Internacional do Observare* http://hdl.handle.net/11144/3359

Collins, A. (2007). *Contemporary security studies*. Oxford: Oxford University Press.

Cong, Y., Chan, Y. B., Phillips, C. A., Langston, M. A., & Ragan, M. A. (2017). Robust inference of genetic exchange communities from microbial genomes using TF-IDF. *Frontiers in Microbiology*, *8*.

Díaz, G. & Abad, G. (2008). Migración y seguridad en España. Seguridad humana y el control de las fronteras, el caso de Frontex. *Unisci Discussion Papers* 17(2), 135-150

D'Angelo, P., & Kuypers, J. A. (Eds.). (2010). *Doing news framing analysis: Empirical and theoretical perspectives*. Routledge.

ISSN 2386-7876 – © 2018 Communication & Society 31(3), 175-188

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of communication 43*(4), 51-58.

Entman, R. M. (2004). *Projections of Power: Framing News, Public Opinion, and U.S. Foreign Policy*. Chicago: The University of Chicago Press

Ferrero, R. & López, A. M. (2012). Fronteras y Seguridad en el Mediterráneo. In Zapata, R. and Ferrer, X. (ed.) *Fronteras en movimiento. Migraciones a la Unión Europea en el contexto mediterráneo* (pp. 227-252). Barcelona: Edicions Bellaterra:

Forman, G. (2008, October). BNS feature scaling: an improved representation over tf-idf for svm text classification. In *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 263-270). ACM.

Grün B and Hornik K (2011). Topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, 40(13), pp. 1–30. doi: 10.18637/jss.v040.i13.

Gunn, S. R. (1998) Support Vector Machines for Classification and Regression. *University of Southhampton.*

Karpf, D.A., Kreiss, D., & Nielsen, R.K. (2014). A new era of fieldwork in Political Communication research? In L. Lievrouw (Ed.), Challenging Communication Research (pp. 43-57). New York: Peter Lang,

Scholl, P. & Peyrin, J. P. (1991). *Esquemas algorítmicos fundamentales: secuencias e iteración*. Masson.

Koza, J. R., Bennett, F. H., Andre, D., & Keane, M. A. (1996). Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. Artificial Intelligence in Design '96. Springer, Dordrecht.

Luengo, Ó. G. (2016). Comunicación Política: de la propaganda masiva a las neurociencias. In Colino, César et. al. (comp.) *Ciencia Política: Una aventura Vital* (pp.721-740). Valencia: Tirant Lo Blanch

Meyer, D., & Wien, F. T. (2001). Support vector machines. *R News 1*(3), 23-26.

Meyer, D., Hornik, K., & Feinerer, I. (2008). Text mining infrastructure in R. *Journal of statistical software 25*(5), 1-54.

Nielsen, R. K. (2014). Political Communication Research: New Media, New Challenges, and New Opportunities, *MedieKultur* 56, 5-22.

Onghena, Y. (2015) "Migrant or refugees" *CIDOB Opinion 355*, 2015

Potthast, M., Hagen, M., Gollub, T., Tippmann, M., Kiesel, J., Rosso, P., ... & Stein, B. (2013). Overview of the 5th international competition on plagiarism detection. In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation* (pp. 301-331). CELCT.

Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning.*

Scheufele, D. A. (1999). Framing as a theory of media effects. *Journal of communication 49*(1), 103-122.

Scheufele, D. A. (2000). Agenda-setting, priming, and framing revisited: Another look at cognitive effects of Political Communication. *Mass Communication & Society 3*(2-3), 297-316.

UNCHR (2015, July). Mediterranean Crisis 2015 at six months: refugee and migrant numbers highest on record. Retrieved from http://www.unhcr.org/news/press/2015/7/5592b9b36/mediterranean-crisis-2015-six-months-refugee-migrant-numbers-highest-record.html

UNCHR (2016) *Global Trends 2015. Force Displacement in 2015*, Geneva: UNCHR publications. Retrieved from http://www.unhcr.org/576408cd7.pdf

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF* IDF, LSI and multi-words for text classification. *Expert Systems with Applications 38*(3), 2758-2765.

ISSN 2386-7876 – © 2018 Communication & Society 31(3), 175-188