Chapter 2
New Methodological Perspectives in Political Communication Research: Machine Learning and Algorithms

Javier García-Marín and Óscar G. Luengo

Abstract Political communication has seen a rise in the use of algorithms and machine learning techniques. Scholars show that these techniques can be used to study new communication routines essential to understanding sophisticated political processes. They state the rise of computer science in the field of political communication is due to the popularity and prevalence of digital media in actual societies. As a result, we firstly draw attention to the fact that our discipline has successfully adopted these new methodological approaches. However, we can detect a slight delay when comparing political communication with other academic fields. Secondly, we draw attention to the fact that it is becoming increasingly common to identify interdisciplinary teams whose scientific output is becoming more and more apparent in reputable publications. In conclusion, social sciences and computer sciences have a common field of work and dialogue to face the new social horizon posed by digital challenges. This chapter describes and identifies some of the different methodological repertoires that have arisen in political communication regarding the employment of algorithms, their fundamental properties, and advantages. This work also identifies potential social, ethical, political, and legal challenges we could face in the forthcoming years due to its explicit usage within social sciences.

Keywords Machine learning · Political communication · Algorithms · Methodology

This research is part of the Project PID2021-128272NB-I00 funded by the Spanish Ministry of Science and Innovation (MCIN/ AEI/10.13039/501100011033/ ERDF, "A way of doing Europe").

J. García-Marín (✉) · Ó. G. Luengo
Department of Political Science and Public Administration, University of Granada, Granada, Spain
e-mail: jgmarin@ugr.es; ogluengo@ugr.es

M. Musiał-Karg, Ó. G. Luengo (eds.), *Digital Communication and Populism in Times of Covid-19*, Studies in Digital Politics and Governance,
https://doi.org/10.1007/978-3-031-33716-1_2

1 Introduction

A new cycle brought about by the popularisation of algorithms has transformed the methodological horizon of political communication. Many contemporary authors, who argue that digital captives have transformed the world into a measurable reality driven by the revolution in computations, explain this pattern. This kind of mysterious truth, boosted by the non-visibility of the algorithms, creates routines of decision orientation, facilitates automatic processes, or even justifies particular policies (Cardon, 2018). However, following the first affirmative phase, digital communications are now being questioned for several reasons, including the production of fake news, the personalisation of messages (creating echo chambers), the radicalisation of views, and the disruption of political processes (Bonneau et al., 2022).

In the last 20 years, the rapid emergence of new methodological approaches within Social Sciences in general, and in political communication in particular, has been configured as one of the visible patterns in the field's evolution. Quantitative techniques based on increasingly complex statistical operations that revolutionised the discipline in the first steps of its institutionalisation after the II World War live together at the moment with proposals offered by other academic areas like genetic studies, neurosciences, or computer science (Luengo, 2016). In this context, these new techniques coming from the latter are compelling and versatile in order to study new communication routines essential to understanding growing sophisticated political processes today. In that sense, for example, we can underline the increasing importance of populism as a current political phenomenon with an essential communication dimension. Moreover, using machine learning techniques, we are developing substantial research quickly, using a massive volume of data that was simply unmanageable with traditional tools. Consequently, we can now unravel complex events like media negativity or affective polarisation, which are still central to many communication science studies, with significant scientific rigour.

By paying attention to a clear indicator—the methodological development of articles in scientific journals—it is simple to confirm the sharp inclusion of this technique (Demšar et al., 2013). The articles published since 2000 in the best-ranked journals within Communication (Social Sciences), according to Scimago Journal Rank (SJR) for 2021,[1] have shown a rising presence of works that include machine learning and algorithms as the chosen techniques for verifying the hypothesis.[2] In

[1] https://www.scimagojr.com/journalrank.php?area=3300&category=3315

[2] We extracted this information using the search tool of Google Scholar. Our sample is the 50 world's best-ranked journals in Communication (Social Science) in 2021, according to Scimago Journal Rank (JRS). The timeline under analysis is 2000–2022. The search keywords are "machine learning" and "algorithm", and we counted their presence in the article title. This measure proposed could sometimes involve a slight misinterpretation of data: some journals published few articles using these techniques but the noted keywords were not present in the title, and consequently were not counted (e.g. Mass Communication and Society). Likewise, we incorporated in the sample a few articles that included the search keywords in the title, but they were not using the cited research

2012 we located the two first publications within these criteria in "Communication Methods and Measures" and "Journalism Practice". Till 2017 we could only find around ten pieces of work using those as the primary methodological strategy. Since then, we have witnessed significant growth in the mentioned techniques, and since 2020 the rise has been prominent, counting 116 articles in 3 years. This basic approach to the data confirms an evident landscape.

This chapter focuses on computer science's contribution to the methodological turnover within Social Sciences. The goal of this work is the systematisation and presentation of these new techniques and tools, which are algorithms for selection, object detection, image classification, supervised, non-supervised, etc. In the following lines, we will provide an overview of the main techniques we have "imported" from other disciplines. Finally, we will illustrate this trend with different research works published by the authors in recent years addressing the scholarly study of increasingly complex social, political, and communication realms.

2 Media Digitisation and Research

We can divide digital techniques in political communication research into three main areas: data collection, cataloguing, and analysis. In all three aspects of research, the discipline has undergone substantial, though not symmetrical, changes.

Perhaps unsurprisingly, it will be in data analysis, which was digitised earlier, where the changes have been smaller. Indeed, traditional statistical analysis tools, usually encompassed in software suites such as SPSS or STATA, are working better and better due to the increased capacity of our computers. Moreover, specific programming languages, such as R or Python (with the Pandas library), and open-source or free programming suites, such as RStudio or Jupyter, have made access to them much cheaper for researchers and students. Because they do not require coding knowledge, open-source visual programming suites like Orange3 are becoming increasingly popular.

Much the same is true of the changes in the data mining and analysis discipline, which, although significant, could not be described as revolutionary. We could speak of evolution in line with the advance in technology. It is true, however, that the data available to researchers has multiplied by several orders of magnitude: from traditional media to the explosion of digital native media, video-sharing platforms, social networks and, more recently, podcasts. Nevertheless, one of the benefits of the digitisation of society is that these materials are more easily accessible than in traditional models. The tools in use are global media databases, such as Lexis Uni or Ingenta Connect, or purely national ones, such as MyNews, in the case of Spain.

techniques, but analysing, for example, the performance of algorithms when offering different options in news consumption, or considering the ethics of algorithms. However, the number of these mentioned cases is not relevant.

These databases offer quick and convenient access to most global or local press, albeit for a fee. Although the instruments that automate this need technical expertise and programming, audiovisual content may also be recorded and stored considerably more conveniently than before. Capturing the comments associated with these media is more complicated without a common platform for all media. Some tools make the task much easier for large sites, such as YouTube. However, data capture is much easier today than in the past. That, of course, has meant that the amount of data available to researchers has multiplied enormously. Moreover, here, we find the challenge that has kept the discipline on tenterhooks in recent years: how to classify and structure so much information for subsequent analysis?

3 Classification and Data Structuring: Machine Learning

Scientists' answer to the above question was machine learning. According to Arthur Samuel, machine learning is the subfield of computer science that gives "computers the ability to learn without being explicitly programmed" (Koza et al., 1996). It searches for the analysis and construction of algorithms that can learn from and make forecasts based on data. In mathematics, computer science, and related disciplines, an algorithm is a well-defined, orderly, and finite operations list that eases a problem's solution findings. It can also be defined as the description of a pattern of behaviour that is displayed using a finite repertoire of actions, and basic information is identified, understood, and obtainable. This repertoire is called a lexicon. An algorithm is a set of serialised operations. Algorithms have been known since ancient times, but the new contribution today relies on the ability to produce very elaborated and complex algorithms thanks to the possibilities offered by the exponential increase in computing capacity. While in the 1950s, the first computers allowed for the use of new research techniques such as survey analysis, nowadays, machine learning is the way to establish a new set of tools that complement the existing ones.

We can therefore argue that an algorithm creates a sequence for categorising data into two or more groups based on criteria established by the algorithm itself (in the case of unsupervised algorithms like the Latent Dirichlet Allocations or LDA) or by the user (in the case of supervised algorithms such as SVM).

ML techniques can be applied to an infinite number of fields. However, our discipline has made the most progress in text analysis. For this purpose, we use natural language processing (NLP) techniques. In its most basic definition, machine learning techniques based on natural language processing break down human language into its constituent parts (whatever they may be) so machines can understand and analyse it. That approach facilitates processing large volumes of data, limited only by the computing power at our disposal. NLP also involves the transformation of messages from audiovisual sources (audio, video, or image) into data. Complementary to NLP is data mining. Text mining, also known as text analytics, examines extensive collections of writings to generate new information. It is used to discover relevant information by transforming text into data that can be

used for further analysis. As we have said, these are two different but complementary techniques.

Finally, there is also a multitude of visual analysis techniques, still little used in our discipline but which, little by little, are gaining ground as they are the best strategy for analysing large quantities of videos.

Below are some examples of the use of these techniques applied by the authors in actual research. In practically all cases, their application has been done without the collaboration of computer science departments and using open-source or free tools, such as R (under RStudio) or Python (using Jupyter or in the form of visual programming with Orange3[3]).

3.1 Lemmatisation and Stemming

The techniques of lemmatisation and stemming are based on word particle location that contains meaning. In the case of stemming, this is done through the elimination of affixes, e.g. the stem of the words eating, eats, eaten is eat (this is the technique used by search engines). In the case of lemmatisation, the root is not the stem but a word, so the result is a word that means the same thing.

These techniques are beneficial for normalising texts that will later be subject to further analysis. Most of the techniques explained below assume that the texts have been normalised. What is achieved, fundamentally, is to reduce the typical complexities of each language by having different words by gender, number, or local variants (slang). The most commonly used algorithms are Porter for stemming and WordNet as Lemmatiser. However, both work exclusively in English (or through translation dictionaries). Others, also prevalent, work in other languages, like the Snowball stemmer or UDpipe and Lemmagen as lemmatisers.

3.2 Sentiment Analysis

Sentiment analysis is one of the most straightforward techniques for extracting digitised text. It consists of simply giving scores to predefined words. Therefore, to perform it, it is necessary to have text in a readable (and, if possible, clean or pre-processed) format and a dictionary file (lexicon) containing a list of words together with a score for each of them. The sentiment analysis will then be the score (mean, median, sum, etc.) obtained by our unit of analysis (sentence, paragraph, article, tweet, post, comment, etc.). It is widely used in the industry to intuit the feelings aroused by a brand or product.

[3] https://orangedatamining.com/

However, in communication, sentiment analysis presents many challenges (see Mohammad, 2017 for a summary of the challenges). The most important of these is that it is difficult to assign average scores to words in isolation from context. Thus, in everyday language, we assign very different meanings to certain words that, in isolation, may be considered highly offensive, but not in these contexts. Naturally, NLP techniques have attempted to overcome this problem by creating lexicons that consider contextual situations, such as nearby words, phrases, etc. Some of these techniques are related to deep learning (Zhang et al., 2018).

Does that mean that sentiment analysis has no use for political communication? No. While it is true that this type of analysis provides us with partial information about the analysed texts, it can be the basis for other, slightly more sophisticated analyses. We can measure the affectivity or polarisation of a group of words, for instance, using typical sentiment analysis. The formula might be:

$$P = |(S - \mathrm{Me})|, \text{where } S\,\epsilon|-1, 1|, p\,\epsilon(0, 2)$$

where P is "polarisation", S is "sentiment", and Me is "mean sentiment" (the median can also be used). Thus, for a lexicon that measures sentiment from 1 to −1, we can perform sentiment analysis on a set of texts (for example, several 1000 tweets) and then calculate the distance of a given text from the mean. The information we obtain, in absolute numbers, would give us the distance between that text and the mean (which can take values between 0 and 2). Note that we do not calculate the negativity of a text but the distance. In this way, we can avoid the main problem of sentiment analysis, the accuracy in measuring negative and positive sentiment.

This type of analysis has been applied quite successfully to analyse the polarisation of large sets of texts (Serrano-Contreras et al., 2020; Luengo et al., 2021; García-Marín, 2021; Serrano-Contreras et al., 2021; García-Marín & Serrano-Contreras, 2023). In the case of Fig. 2.1, for example, it was applied to measure polarisation concerning three topics: the independence of Catalonia, climate change, and political parties. In all three cases, these are comments on YouTube videos (600 videos, $n = 391{,}739$ comments). The average polarisation, in this case, is made with the whole set of comments for each topic. As can be seen, the analysis, although confusing, provides exciting information: in the first case, the political conflict in Catalonia in 2017, we can see how polarisation underwent a sudden increase, coinciding with the climax of the situation. In the third case, we see peaks that coincide with Spain's election periods. The third case is less evident but also shows an increase in polarisation as time progresses on the issue of climate change, which the researchers expect. In any case, the formula gives us a view consistent with the expected reality, which is the aim of any research technique.

Figure 2.2 provides another example of the application of this technique. In this case, it is also research on YouTube, but the polarisation (or affective distance) of both videos and comments is calculated (150 videos, 111,808 comments). The objective is similar to the previous research: verifying user behaviour. However, this time we intend to see if there are differences between Italians, Spaniards, and

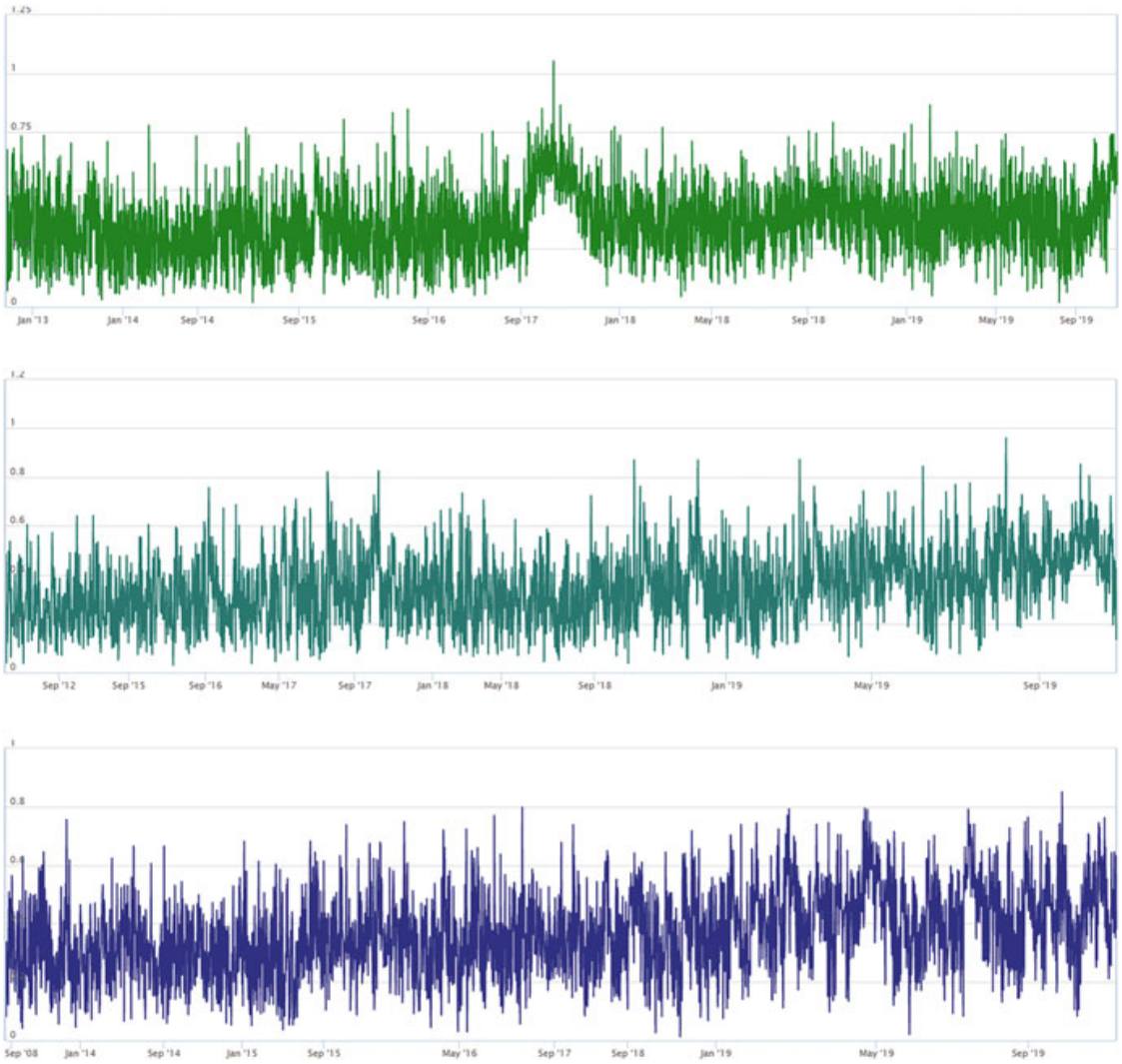

Fig. 2.1 Polarisation through time: Catalonian independence, climate change, and political parties (from top to bottom). (Source: Serrano-Contreras et al., 2020: 68)

British. What Fig. 2.2 shows is how, from the polarisation analysis, it is perceived that Italians and Spaniards behave similarly, but not the British.

Furthermore, Fig. 2.2 shows that the number of likes is an inverse predictor of polarisation for the British, while it is a direct predictor for Spaniards and Italians. That is, the former penalises polarising comments while the latter rewards them. Again, the formula helps us to describe complex realities and is consistent with our expectations using datasets too large to be treated with traditional techniques.

3.3 Summarisation and Topic Modelling

Summarisation and topic modelling techniques attempt to summarise complex texts or reduce them to groups according to word frequencies or other analytical

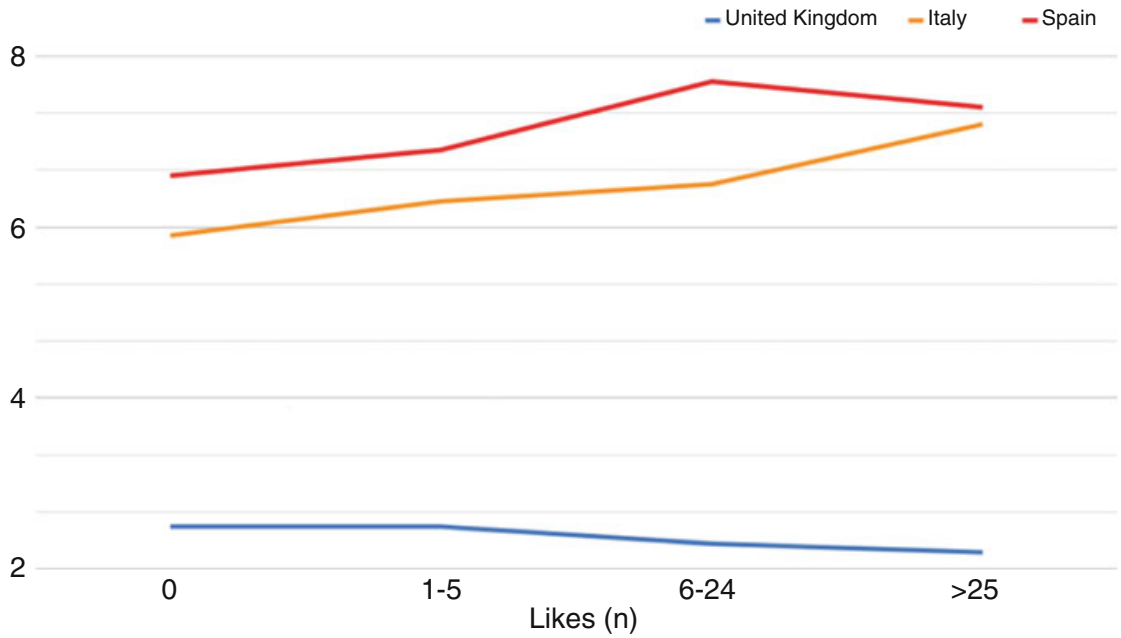

Fig. 2.2 Estimated marginal average of polarisation. (Source: Luengo et al., 2021: 15)

strategies. In both cases, they are unsupervised algorithms, i.e. they do not need to be trained before being applied to the selected sample.

In the case of summarisation, there are algorithms such as SumBasic, LexRank, and TextRank that, using word frequencies, can summarise vast amounts of text and locate redundancies (Shah & Jivani, 2016).

Although text summarisation is a widely used technique in specific disciplines, especially those related to marketing, it has limitations for our field of study (beyond a description of the content of the texts), primarily that it deprives us of the full text we want to analyse. However, topic modelling techniques can be much more helpful. Within these techniques, there are three prevalent algorithms: Latent Dirichlet Allocation (LDA), Latent Semantic Indexing (LSI), and Hierarchical Dirichlet Process (HDA). All three are based on creating clusters of terms according to frequency and return a weighting of each topic per unit of analysis (thus giving us more variables that can potentially be analysed later). LDA is the simplest and only requires the researcher to determine the number of topics he/she wants to extract from the sample. LSI returns both present (representative of the topic) and absent (non-representative) terms. HDP, on the other hand, is more flexible than the previous ones as it allows for greater granularity in its operation, but it is also more computationally demanding. In all three cases, the results obtained may be meaningless, especially if the texts used for the analysis have not been pre-processed.

Such unsupervised algorithms enable us to analyse, at least superficially, large sets of texts. For example, Tables 2.1 and 2.2 describe the coverage of two Spanish media (ABC and El País) on education issues from 1978 to 2018 (Serrano-Contreras et al., 2021).

Table 2.1 Analysis of the coverage of El País (1978–2018)

Topic	Terms present	Terms absent
1	Education, law, government, students, reform, centres, pp	
2	Education, centres, students, teaching	pp, government
3	Law, government, pp, project	Students, years,
4	Millions, pesetas, project	Education
5	Students, pp., government, **castellan, catalonia**, centres, curse	Education, law

Terms have been edited to improve readability. Selected terms in bold
Source: Serrano-Contreras et al., 2021: 507

Table 2.2 Analysis of the coverage of ABC (1978–2018)

Topic	Terms present	Terms absent
1	Education, law, teaching, reform, government, centres, students, educative	
2	Government, pp., Spain	Education, teaching, centre, students, formation
3	Law, education, government, pp., **catalonian**	Reform, formation
4	Years, curse, students	Teaching, law, project, rights, freedom, religion
5	**Parents**, centres, students, teaching, **church**	Reform, education, project, formation

Terms have been edited to improve readability. Selected terms in bold
Source: Serrano-Contreras et al., 2021: 508

Table 2.3 Spearman's Rho El País -selected topics

Part_Gov	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
PP	0.054[a]	−0.006	*−0.196*[a]	0.107[**a**]	*0.201*[a]
PSOE	−0.056[a]	0.003	*0.195*[a]	−0.127[a]	*−0.152*[a]
UCD	0.005	0.008	0.001	0.065[a]	*−0.164*[a]

[a]$p < 0.01$ (bilateral)
[b]$p < 0.05$ (bilateral)
Source: Serrano-Contreras et al., 2021: 509

What is done in this research is to compare the coverage of the two prominent newspapers in Spain, one progressive and the other conservative, during the entire democratic period ($n = 4872$ newspaper articles). The tables help us to divide a large sample into categories that make sense for the research. Thus, for example, we can see how the LSI indicates that religion or the church are terms present in specific articles in ABC, while they are not in El País, which is to be expected. Alternatively, both newspapers treat language use in education differently in Catalonia.

The sample description is essential for the research but does not help us in the theoretical construction. However, one of the benefits of applying these techniques is that it provides us with new variables we can analyse. Thus, Tables 2.2 and 2.3, from the same study, show the correlations in the coverage of both newspapers with the different governments in Spain. That is, they relate the topics covered by each media

Table 2.4 Spearman's Rho ABC-selected topics

Part_Gov	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
PP	−0.074[a]	*0.138*[a]	−0.128[a]	0.118[a]	−0.063[a]
PSOE	0.077[a]	*−0.148*[a]	0.148[a]	−0.123[a]	0.056[a]
UCD	−0.015	0.044	−0.074[a]	0.025	0.018

[a]$p < 0.01$ (bilateral)
[b]$p < 0.05$ (bilateral)
Source: Serrano-Contreras et al., 2021: 509

with the parties that support the governments. It is easy to see how each media outlet applies different policies depending on the government: Tables 2.3 and 2.4 show that specific topics correlate directly or inversely depending on the government. In this way, an unsupervised algorithm has helped us to test hypotheses such as that of Hallin & Mancini (2004) regarding the features of the polarised pluralist model.

3.4 Text Classification

Although many of the techniques we have shown can be included within text classification, such as LDA, we have reserved this denomination for supervised classification algorithms. These algorithms work based on a data set divided into two: one that has been previously labelled by researchers (using the traditional techniques of the discipline, such as content analysis) and one that has not been labelled. Once the distinction has been made, the first group is again divided into two: the training group and the verification group. The training group is fed to the algorithm that will try to find a pattern in the data that justifies the labelling. Once the pattern has been found, the algorithm will try to identify the labelling of the verification group, returning a hit rate that will indicate reliability. Naturally, each algorithm has different strategies for finding the patterns that entail different ways of optimising them and their suitability for different datasets. In general, classification algorithms are divided into linear classifiers, support vector machines (SVMs), quadratic classifiers, kernel estimators (such as KNN), decision trees, neural networks, and learning vector quantisation. The most popular algorithms in non-computer related disciplines are linear (such as logistic regressions or Bayesian classifiers), SVMs (which can behave in a variety of ways, from linear classifiers to neural networks), decision trees (although they are computationally demanding), or neural networks (with all their diversity, from CNNs, DNNs to RNNs).

Generally speaking, we combine these techniques with text analysis in political communication. As the algorithms work with vectors, the first task is to convert them into data and insert them into matrices. This is done using term frequencies in previously processed texts (i.e. from which we have removed words and structures that do not confer meaning or could alter it, usually called stopwords). A more sophisticated strategy is using inverse document frequencies (tf-idf), which assign

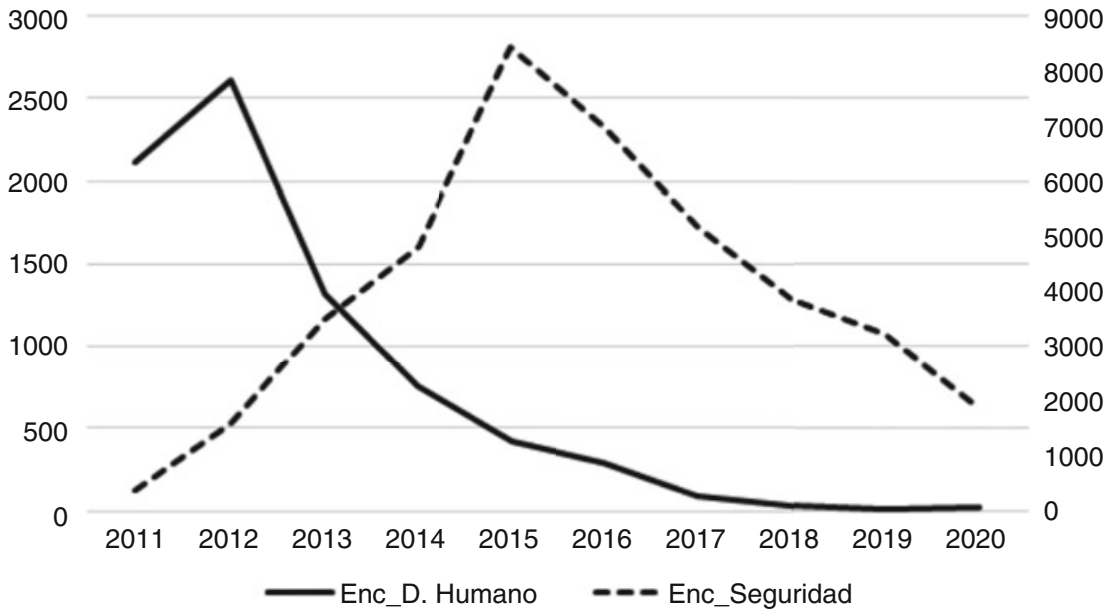

Fig. 2.3 Frames of human drama and security in various Spanish newspapers. (Source: Moreno-Mercado et al., 2021: 8)

decreasing weights for the most common words and increasing weights for those less commonly used in a list of documents. Once we have a spatial representation of each unit of analysis, the algorithms can then proceed to classify them according to the values assigned (a detailed explanation of the use of SVM in political communication can be found in García-Marín et al., 2018). Applying supervised algorithms to obtaining data on large sets of texts has a few limits, as it depends on our ability to analyse the texts differently. Thus, we could train an algorithm for every variable we can think of. Naturally, the simpler the variable, the greater the chance of success.

One of the earliest uses (one of the first in Burscher et al., 2014, although with mixed results) and one that is generating better results from supervised algorithms is the localisation of frames in the media, something that has always been a challenge for researchers.

For example, another case analysed the refugee crisis in the European Union in 2015 (García-Marín & Calatrava, 2018: 185), employing an SVM to locate frames in the coverage of the Spanish press on the event. It concluded that each media has different routines, but they all agree that coverage of the "security" frame increases at the end of the year to the detriment of the "human drama" frame. The research was carried out by coding part of the sample and then training the algorithm, which subsequently coded the remaining articles ($n = 4548$).

Figure 2.3 shows the results of a similar investigation: the use of the frames "human drama" and "security" in Spanish newspapers' coverage of the war in Syria. The research served not only to show the change in the newspapers' coverage but also to support the hypothesis that armed conflicts tend to be "securitised" in the media, which affects all the newspapers analysed, regardless of their ideology (Fig. 2.4).

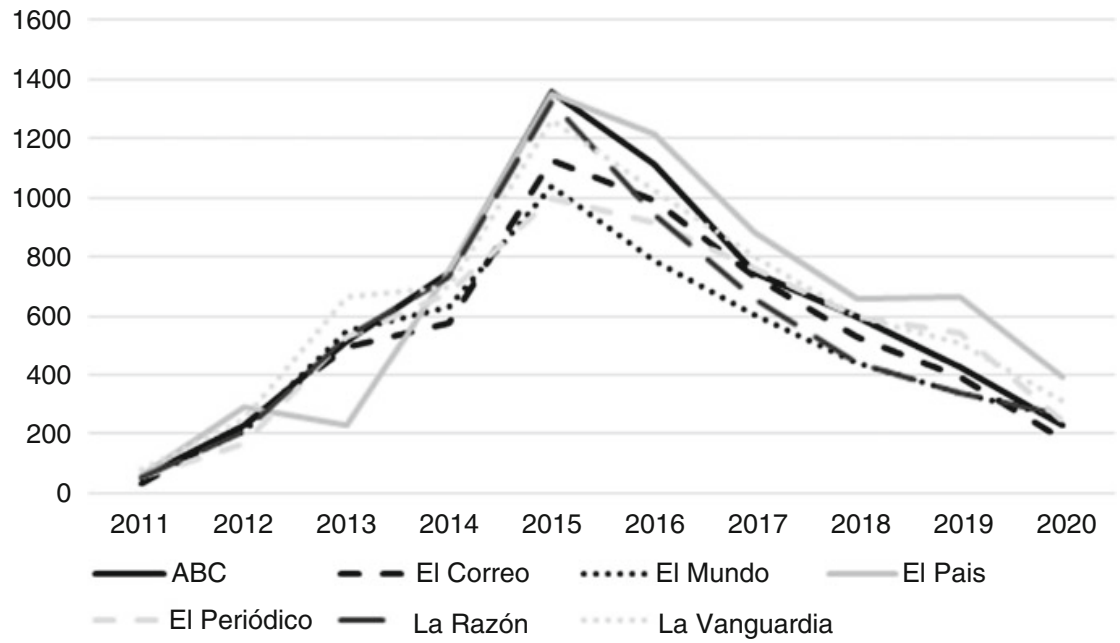

Fig. 2.4 The "security" frame in the analysed newspapers. (Source: Moreno-Mercado et al., 2021: 8)

As can be seen, the results shown were very useful in describing the samples and demonstrating hypotheses, as well as being coherent with the perceived reality. However, there are doubts that research with algorithms is equally effective in locating more complex or non-mutually exclusive frames.

3.5 Keyword Extraction

Keyword extraction is the automatic processing of extracting the most relevant words and expressions from a text. It is a technique that helps narrow down the text content and recognises the topics they deal with. In this sense, it is similar to unsupervised algorithms such as LDA. The difference is that keyword extraction techniques can target specific words or phrases. One can choose to extract names of entities or countries, for example.

The popular word clouds are an example of keyword extraction.

Table 2.5 shows an example of keyword analysis for analysing the content of headlines from various newspapers ($n = 4629$). Although it is simply a descriptive technique, it already provides us with some information, such as the countries most covered by the analysed press (such as Tunisia) and other related concepts, such as the occurrence of religion or the economy.

Table 2.5 Most used terms. By newspaper

ABC			El País			El Mundo			La Vanguardia		
Term	*n*	%	Term	*n*	%	Term	*n*	%	Term	*n*	%
Arab	98	8.91	Tunisia	89	9.05	Syria	80	6.21	Arab	101	8.00
Tunisia	83	7.55	Syria	65	7.27	Arab	53	4.11	Syria	84	6.65
Spring	76	6.91	Arab	56	5.69	Spring	50	3.88	Spring	72	5.70
Syria	66	6.00	Islam	56	5.69	Egypt	45	3.49	Tunisia	71	5.62
New	43	3.91	Spring	44	4.47	Tunisia	45	3.49	Egypt	52	4.12
Spain	41	3.73	Egypt	40	4.06	Deaths	44	3.41	Country	39	3.09
Deaths	39	3.54	New	33	3.35	Egyptian	38	2.95	Tourists	36	2.85
Years	35	3.18	Politics	30	3.05	Years	34	2.64	Years	34	2.69
Egypt	34	3.09	Saudi	30	3.05	Islam	33	2.56	Obama	34	2.69
Morocco	30	2.72	Morocco	29	2.95	Spain	33	2.56	World	33	2.61

Source: Moreno-Mercado & García-Marín, 2021: 358

3.6 Image Analysis

As mentioned above, almost all the analysis we do in communication using machine learning is text-based. Those can be original texts, such as those coming from written media (digital or printed) or extracted from audio or video, such as the research on YouTube that we have seen above. However, ML techniques are very advanced in analysing images, especially for localising faces, objects, image classification, etc. It is easy to say that this is their main field of application. Social sciences can also use these techniques to extract data from images or videos (Moreno Mercado et al., 2021).

When we talk about images or videos, algorithms are usually divided into classification, object localisation, and segmentation algorithms. The first try to classify an entire image. In contrast, the second try to locate specific elements in them, either by an inductive or deductive approach (i.e. if we know what we want to locate, such as a face, or we want to know what is in the image); finally, the third type are dedicated to silhouette localisation (widely used in autonomous driving). It is the same principle in all cases: structuring unstructured information.

There are not many examples of the application of these techniques in communication studies, but the possibilities are very promising. For example, in one of our research studies (García-Marín & Luengo, 2022), we applied this strategy to videos published by jihadist groups on the Internet. The idea was to corroborate whether terrorist groups used the dissemination of audiovisual material to fulfil specific functions such as indoctrination, teaching results or publicity. Furthermore, we hypothesised that the videos could be catalogued if the functions were catalogued. That is, videos dedicated to a function would resemble each other (which also has implications for using professional routines among terrorists).

Figure 2.5 shows the research process and the techniques used. First, 2211 videos of jihadists were used. Next, the videos were converted into images using the FFmpeg library, an open-source standard (one image every 15 s). The ML

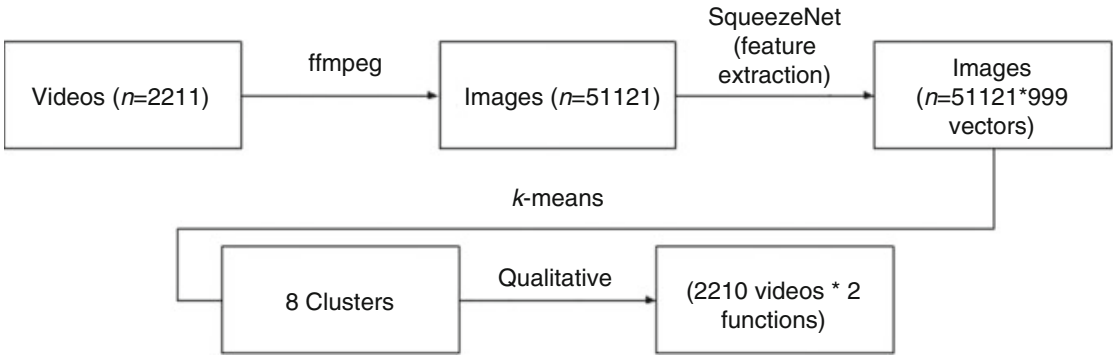

Fig. 2.5 Research Strategy. (Source: García-Marín & Luengo, 2022: 399)

application is the next step, image embedding. Image embedding is a set of techniques used to reduce the dimensionality of data by applying deep neural networks. It is widely used in image classification because it reduces the millions of pixels in a set of variables. In the case of our research, we used SqueezeNet (trained on ImageNet, Iandola et al., 2016) for its simplicity and low computational demand. The result was a matrix with 999 variables, already structured and easily analysable data. Subsequently, using k-means, the number of variables was reduced to eight clusters. With these clusters, a simple qualitative analysis was carried out, where it was possible to extract that there were two differentiated functions: the doctrinal one, with images of people in close-ups, letters of the Arabic alphabet on screen and images of jihadist flags; and another operational one with landscapes, explosions, groups of people in open fields, etc. In other words, by applying an algorithm, it was possible to divide the sample into two functions. However, the actual usefulness came from correlating the functions with the terrorist groups that published the videos (since the videos were labelled by date and organisation). Clearly, groups with greater operational capacity, such as ISIS or Al-Nusra, published longer videos in which they mixed both functions. However, Al-Qaeda, with less presence on the ground in operations in armed conflicts, focuses on shorter and more specialised videos, almost all dedicated to the doctrinal function. In other words, the research technique produced exciting results from quantitative ML and qualitative approaches.

4 Conclusions

This chapter aims to present some relevant studies that are good examples of the growing incorporation of new techniques based on algorithms in political communication research. Of course, we do not support a reductionist viewpoint since we appreciate the appropriate combination of techniques to improve the approach to the object of study, including qualitative tools in the strategies. Instead, we insist on the argument that methodological plans must adapt to the specific research goal,

meaning that algorithms are beneficial to answer some research questions but may be useless in the approach to other questions.

As a result, we firstly draw attention to the fact that our discipline has successfully adopted these new methodological approaches. However, we can detect a slight delay when comparing political communication with other academic fields. Secondly, we draw attention to the fact that it is becoming increasingly common to identify interdisciplinary teams whose scientific output is becoming more and more apparent in reputable publications.

In conclusion, social sciences and engineering have a common field of work and dialogue to face the new social horizon posed by digital challenges. This chapter describes and identifies some of the different methodological repertoires that have arisen in political communication regarding the employment of algorithms, their fundamental properties, and advantages. This work also announces potential social, ethical, political, and legal challenges we could face in the forthcoming years due to its explicit usage within social sciences.

References

Bonneau, J., Grondin-Robillard, L., Ménard, M., & Mondoux, A. (2022). Fighting the "System": A pilot project on the opacity of algorithms in political communication. In A. Hepp, J. Jarke, & L. Kramp (Eds.), *New perspectives in critical data studies. Transforming communications – studies in cross-media research*. Palgrave Macmillan.

Burscher, B., Odijk, D., Vliegenthart, R., De Rijke, M., & De Vreese, C. H. (2014). Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. *Communication Methods and Measures, 8*(3), 190–206.

Cardon, D. (2018). *Con qué sueñan los algoritmos. Nuestra vida en el tiempo de los Big Data*. Dado.

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., et al. (2013). Orange: Data mining toolbox in Python. *The Journal of Machine Learning Research, 14*(1), 2349–2353.

García-Marín, J. (2021). *YouTube and traditional media: Polarization in the Catalan political conflict* (In digitalization of democratic processes in Europe) (pp. 31–41). Springer.

García-Marín, J., & Calatrava, A. (2018). The use of supervised learning algorithms in political communication and media studies: Locating frames in the press. *Communications Society, 31*(3), 175–188.

García-Marín, J., Calatrava, A., & Luengo, O. G. (2018). Debates electorales y conflicto. Un análisis con máquinas de soporte virtual (SVM) de la cobertura mediática de los debates en España desde 2008. *El Profesional de la Información, 27*(3), 624–632.

García-Marín, J., & Luengo, Ó. G. (2022). From image to function: Automated analysis of online jihadi videos. *Pragmatics and Society, 13*(3), 383–403.

García-Marín, J., & Serrano-Contreras, I. J. (2023). (Un)founded fear towards the algorithm: YouTube recommendations and polarisation. Comunicar: Revista Científica de Comunicación y Educación*, 31*(74).

Hallin, D. C., & Mancini, P. (2004). *Comparing media systems: Three models of media and politics*. Cambridge University Press.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and <0.5 MB model size. arXiv preprint arXiv:1602.07360.

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* (In artificial intelligence in design'96) (pp. 151–170). Springer.

Luengo, Ó. (2016). Comunicación política: de la propaganda a las neurociencias. In C. Colino, J. Ferri Durá, J. Olmedo, P. Román-Marugán, & J. Rubio-Lara (Eds.), *(Coords.): Ciencia política: una aventura vital*. Tirant Lo Blanch.

Luengo, Ó., García-Marín, J., & De-Blasio, E. (2021). COVID-19 on YouTube: Debates and polarisation in the digital sphere.[COVID-19 en YouTube: Debates y polarización en la esfera digital]. *Comunicar, 69*, 9–19.

Mohammad, S. M. (2017). *Challenges in sentiment analysis* (In a practical guide to sentiment analysis) (pp. 61–83). Springer.

Moreno Mercado, J. M., García Marín, J., & García Luengo, Ó. (2021). El conflicto de Siria en la prensa española: un análisis sobre la securitización de la guerra.

Moreno-Mercado, J. M., & García-Marín, J. (2021). A quantitative approach to headlines after the Arab Spring. D*oxa. Comunicación, 1*(33).

Moreno-Mercado, J. M., Luengo, Ó. G., & García-Marín, J. (2021). Cyberspace as a global common: Framing the Libyan War in RT, RTVE and La Sexta Television videos. In *Security in the global commons and beyond* (pp. 129–142). Springer.

Serrano-Contreras, I. J., García-Marín, J., & Luengo, Ó. G. (2020). Measuring online political dialogue: Does polarization trigger more deliberation? *Media and Communication, 8*(4), 63–72.

Serrano-Contreras, I. J., García-Marín, J., & Luengo, Ó. G. (2021). Coberturas mediáticas, polarización y reformas educativas en España. *Revista de ciencia política (Santiago), 41*(3), 497–514.

Shah, C., & Jivani, A. (2016, August). Literature study on multi-document text summarization techniques. In *International Conference on Smart Trends for Information Technology and Computer Communications* (pp. 442–451). Springer.

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*(4), e1253.