**ARTICLE**/ARTÍCULO

# Big Data in Social Sciences. An Introduction to the Automation of Textual Data Analysis Using Natural Language Processing and Machine Learning

## *Big data* en ciencias sociales. Una introducción a la automatización de análisis de datos de texto mediante procesamiento de lenguaje natural y aprendizaje automático

**Alba Taboada Villamarín**
Autonomous University of Madrid, Spain
alba.taboada@uam.es

**Received/**Recibido: 11/1/2023
**Accepted/**Aceptado: 25/7/2023

**ABSTRACT**

Innovations in the field of computer engineering and artificial intelligence provide new methodological opportunities for scientific research, giving rise to the study of emerging social phenomena that are born in and inhabit virtual spaces. The purpose of this paper is to familiarise the social scientist with the widely established processes in massive text analysis using machine learning techniques that give rise to what we know today as natural language processing (NLP). First, a brief overview of the history of NLP and its relation to text analysis in the social sciences is given. Then, in each section of the text, the steps to follow when applying NLP to social research are assessed, providing information on software, tools, data sources and useful links, with the aim of offering an introductory and simplified guide to serve as an initial approach to this discipline. Finally, the main challenges that the social sciences face when implementing NLP techniques are examined and assessed.

**KEYWORDS:** big data; natural language processing; social sciences; machine learning, text mining.

**HOW TO REFERENCE:** Taboada Villamarín, A. (2024). *Big data* en ciencias sociales. Una introducción a la automatización de análisis de datos de texto mediante procesamiento de lenguaje natural y aprendizaje automático. *Revista Centra de Ciencias Sociales*, 3(1), 51–75. https://doi.org/10.54790/rccs.51

The Spanish (original) version can be read at https://doi.org/10.54790/rccs.51

## RESUMEN

Las innovaciones en el campo de la ingeniería computacional y la inteligencia artificial brindan nuevas oportunidades metodológicas para la investigación científica, permitiendo el estudio de fenómenos sociales emergentes que nacen y habitan en los espacios virtuales. El propósito de este trabajo es familiarizar al científico social con los procesos ampliamente establecidos en el análisis masivo de texto mediante técnicas de aprendizaje automático que dan lugar a lo que hoy conocemos como procesamiento de lenguaje natural (PLN). En primer lugar, se lleva a cabo un breve recorrido por la historia del PLN y su relación con el análisis de texto en las ciencias sociales. Luego, en cada sección del texto, se valoran los pasos a seguir cuando se aplica PLN a investigaciones de carácter social, proporcionando información sobre programas informáticos, herramientas, fuentes de datos y enlaces útiles, con el propósito de ofrecer una guía introductoria y simplificada que sirva como acercamiento inicial a esta disciplina. Por último, se examinan y evalúan los principales desafíos que las ciencias sociales enfrentan al implementar técnicas de PLN.

**PALABRAS CLAVE:** datos masivos; procesamiento de lenguaje natural; ciencias sociales; aprendizaje automático, minería de texto.

## 1. Introduction: big data applied to the social sciences. Natural language processing (NLP)

Natural language processing (NLP) refers to the branch of computational sciences that, combined with linguistics, enables certain computer systems to process and “understand” human language (Bird, Klein and Loper, 2009). Language, in the form of written text, constitutes a primary source of human documentation of great importance in social research contexts. Text analysis has developed greatly, incorporating numerous research techniques and methodological tools that have made it possible to refine the use of this information, both as a primary and secondary data source, especially in the field of qualitative approaches.

However, the importance of the text as a unit of analysis is a concept shared by a number of branches of knowledge. Computational science has demonstrated a growing interest in automating and developing machines that are capable of bridging the gap between human language and “machine language”. Parallel efforts have been made in the two disciplines to extract substantive information from text corpora, a process known as text mining (Justicia de la Torre *et al.*, 2018), resulting in certain points of convergence that are decisive for bringing about methodological advances in social research.

The specialised literature suggests that the need to develop machines capable of performing automatic text translations in various languages arose as a result of the outbreak of the Cold War, with Russian to English translation being most prominent. In response to this demand, the first symbolic textual analysis systems using machines started to emerge (Johri *et al.*, 2021). Meanwhile, although text analysis had already consolidated itself in the fields of anthropology and sociology, towards the end of the Second World War, the Chicago School's pioneering works examining the relationship between migrants and soldiers took on a methodological significance (Abbott, 1997). At that time, the systematisation of textual data analysis, both from primary and secondary sources, required considerable effort in terms of labelling, organisation and text management. This manual work led to the development of various methodological branches that to this day determine the different methods of analysis used in qualitative research.

In its early stages, natural language processing grounded its work in Chomsky's theory of syntactic structures, which was strongly criticised by other linguists (see Radick, 2016; Hockett, 2020). Its detractors argued that human language involves complexities that go far beyond the association rules and comparative models that played such a vital role during the nascent years of computational logic. This argument, which continues to this day, shifted in favour of the machines as a result of the advances made in computing following the incorporation of statistical calculations in machine-based human language processing for the very first time (Bitter *et al.*, 2010). These developments made it possible to address the peculiar and variable characteristics of human language, overcoming the limitations of approaches based solely on syntactic rules.

This paradigm shift culminated in the 1990s, coinciding with the growth of telecommunications and the widespread use of personal computers, shaping what we currently know as the information society (Castells, 1997). The "statistical stage" increased the complexity of text analysis, leading to the launch of the first computer programs specialising in qualitative data analysis (CAQDAS), such as ATLAS.Ti (1993) and NVivo (1999). Thanks to these developments, it was now feasible to carry out text tagging tasks in a semi-automated manner for the first time. Word counting and frequency calculations encouraged the development of approaches such as content analysis and techniques that bordered on mixed analysis and triangulation.

From the first rudimentary techniques to the present day, the information society has undergone significant changes that have set a new stage for NLP and, therefore, opened up the possibility for further advances in textual analysis methods and techniques in social research. Developments in the field of ICT—which led to the creation of new social, scientific and technical models—have brought about three main challenges.

Firstly, intense competition in the international market, due to the inclusion of economies from the East and the Global South, has resulted in an unprecedented reduction in material costs in the technology industry. As a consequence, telecommunications structures have expanded and become more interconnected, turning them into the backbone of the virtual society. Secondly, advances in areas such as computing, applied mathematics, statistics and robotics have imbued these interconnections with intelligence, giving rise to what we know as artificial intelligence. This artificial intelligence moves away from the classic approach to computing based on action and reaction, and instead adopts interactive models that are capable of generating multiple responses to a wide variety of inputs. Lastly, the central element that feeds these two infrastructures—and is influenced by them—is what we call big data.

On a global scale, a growing interest in research of this object of study can be seen. From the perspective of social scientists, big data refers to all parts of the digital footprint generated as a result of interactions between humans, between humans and machines, and between machines in the virtual realm. Previous research (Gualda *et al.*, 2023) has highlighted that text analysis has become one of the most popular methodological approaches when combining big data technologies with the social sciences. This is mainly due to the fact that a considerable percentage of digital footprints are stored as text.

The world wide web continuously records thousands of interactions that occur on platforms such as social networks, personal blogs, websites, instant messaging services and digital forums. This information is a reflection of new narratives, discourses, social representations, interactions and relationships that transpire both in online and offline settings, contributing to phenomena that are characteristic of our contemporaneity, such as the spreading of false news, hate speech, viral trends, polarisation of information, distrust in democratic and scientific establishments, virtual relationships and networks of influence, among others.

While the exploration of new forms of socialisation and their structures is of utmost importance to social research, scientists frequently come up against a number of issues when trying to access these new realities. The problems often posed by these types of data include managing large volumes of information, the dizzying speed with which they are generated, the unstructured format in which they are stored and questions related to their extraction and ownership (Gillingham and Graham, 2017; Gualda and Rebollo, 2020). In addition, the lack of interdisciplinary teams and knowledge of the available tools place considerable constraints on this type of research.

In the context of text analysis, the statistical focus has been replaced by neural networks and machine learning, which, upon further inspection, seems to permit more complex analyses while being more straightforward to apply. For this reason, this work strives to reduce the technical deficiencies that are currently evident in the social sciences and encourage us to explore resources that bring us closer to emerging social problems and build bridges with other disciplines and objects of study.

The following pages provide an introduction to the steps required to apply natural language processing (NLP) to research. Practical information is provided on the procedures and resources used to carry out these analyses, including available text data sources, information extraction techniques, and data cleaning and processing, as well as the main types of analyses that can be performed. Finally, the greatest challenges that the social sciences face when implementing NLP techniques will be examined.

## 2. Computer programs for working with natural language or mining digital texts

When seeking to analyse big data or data from digital sources, it is common to resort to software and programming environments that are not traditionally part of a social scientist's training. However, advances in computing and data analysis have simplified the complexity of programming, making it more accessible to all types of users.

New programming tools have taken a leap forward by revolutionising the way in which we analyse data. On the one hand, it is commonplace to open source software with environments and extensions that can be downloaded free of charge, and which often boast a virtual community that constantly shares information and resources. On the other hand, such software is able to process and apply statistical calculations much quicker, providing greater autonomy and control when refining algorithms. It is also able to handle larger volumes of data and can easily connect to various digital sources and resources. In addition, these approaches incorporate innovative predictive statistical techniques which were not previously available in traditional statistical computer programs.

There are currently two predominant approaches to natural language processing that can be chosen, depending on whether the researcher wants to use a programming language or a computer program with a user interface that does not require any code to be implemented. The latter approach is a more accessible alternative for researchers who lack knowledge in computer science, but who still wish to apply these kinds of analyses.

The two most recognised and widely adopted programming languages in the field of data analysis are R (The R Project for Statistical Computing) and Python. Both are high-level programming languages that have a more accessible syntax that is more similar to that of human language rather than machine language. The R programming language is commonly used with the RStudio integrated development environment, and there are several free courses and manuals for beginners available. Social researchers often favour the R programming language because of its greater statistical analysis and data visualisation potential. Different codes that can be easily applied to different types of data have been shared widely on the Internet. With Python, on the other hand, the use of the Jupyter Notebook environment is recommended, which gives users an interactive way to run code. Python offers many possibilities for both statistical and visualisation purposes. If we compare both languages in terms of their data analysis capabilities, R and Python have similar features, although Python is more widely used by data scientists.

For those who choose not to use programming languages, a wide variety of computer programs can be used to run the same algorithms without having to write code. This article proposes two completely free computer programs that are growing in popularity in the academic community: Orange Data Mining and RapidMiner. They can both be downloaded locally onto a personal computer, and their respective websites offer short video tutorials for performing a wide variety of analyses.

## 3. Textual data sources

As already mentioned, natural language processing (NLP) techniques and methods are applied by analysing units of text or words written in human language. These texts are of a considerably wide variety and origin, since they can come from any number of sources where speech or writing is used. However, in order to use text mining techniques, the data must be in digital format. Under this premise, original text data sources and resources involving human language fall under one of three categories, based on their nature and the context in which they were created, each of which involves different procedures and considerations: 1. Analogue text, 2. Transcribed audiovisual text and 3. Digital text.

### 3.1. Text from analogue files

First of all, we have analogue resources, i.e., all classical and historical texts that were written by hand or typewritten and have been printed but are not yet digitised. Examples of these are historical archives on cities, correspondences between royals and officials, manuscripts and certificates. In social sciences these are classified as secondary data sources and are frequently used in fields such as history, philology and anthropology, although they are also highly valuable in any line of research wherein the goal is to study a specific historical moment or recover information pre-dating the use of computers.

These types of resources have enormous potential when NLP methods and techniques are applied. The only drawback, however, is that they must have previously been digitised in order to be used for this purpose. In fact, most of the pioneering projects that laid the foundations of the digital humanities centred around digitising historical sources (Piotrowski, 2012). Thanks to this, nowadays there are numerous open databases that can be used by researchers from all over the world free of charge. What's more, in the event that the text they are searching for is not found in these databases, there are highly effective tools available that allow them to digitise them. These types of resources can also be of great value for machine learning and for understanding specific language types at specific points in time.

"Culturomics" (Michel *et al.*, 2011) was one of the first experiences that achieved widespread success, leveraging the 15 million books digitised by Google—currently available on the Google Books service—to create a database of more than 500 billion words from books written between the years 1500 and 2008, and which has recently been updated with words from up to 2019. The underlying purpose of this project was to generate a kind of "digital collective memory" that could be consulted to find out more about the use of words and linguistic resources throughout history.

In a similar vein, numerous sources of historical texts can be found, such as databases containing newspapers from several decades ago, permitting researchers to inquire about specific events from the past. However, before these texts can be used online, they must first be processed in order to store them in a file format that is compatible with programming languages. One of the most common options is to save them as plain text files, commonly known by their extension ".txt".

## 3.2. Text from audiovisual files

Next we have texts that are transcriptions of audiovisual media. As is often the case in qualitative research, researchers may wish to work with texts acquired from voice or audio notes, songs, television scripts, radio programmes or conversations that take place in offline settings. For these examples, as long as the transcript is stored as a digital version, the NLP processing tools are equally valid.

Of particular use in this regard are resources that automatically convert audiovisual files to digital text, such as the free tools offered by Zoom and Google Docs. Once the text file has been obtained, the next step, as in the case of analogue resources, is to convert it into a format that is compatible with the computer program that is going to be used. The most commonly used formats are plain text files with the ".txt" extension, JavaScript Object Notation with the extension ".json" and Comma Separated Values files with the ".csv" extension. Occasionally, these files may first need to be structured by the researcher.

## 3.3. Text from digital files

Lastly are digital texts that can be found on Web 2.0. These are texts that have been originally written online and are either fully or partially accessible in the virtual realm. In natural language processing, these resources are widely used because of their diversity, quantity and ease of access. These characteristics, combined with the classic attributes of large datasets, make digital resources the most suitable for process automation and computational textual analysis. These types of texts are available as both secondary and primary data sources for research purposes, although in such digital texts there is also an abundance of unsolicited data (Ruelens, 2022).

The digital resources that store information that can be exploited through natural language processing (NLP) techniques are mainly found in the following areas: 1. Social media, 2. Opinion blogs and virtual forums, 3. Websites, 4. Online newspapers, 5. Scientific databases and online encyclopaedias, 6. Search engine tools and 7. Instant messaging apps. In most of these cases, the data are not specifically oriented to a determined research project; it is the responsibility of the researcher to gather the information and make it relevant to their objectives. However, some of these resources may be primary data sources, as they can be used to create texts in particular digital environments. One of the most commonly used methods are structured interviews conducted in instant messaging applications such as Gmail and Outlook. There are also cases of researchers posing a question in a virtual forum and subsequently analysing the answers written by users (see examples in Dahlin, 2021; Holtz *et al.*, 2012; Murthy, 2008).

# 4. Mechanism for selecting and extracting information

Once the source or repository that houses the relevant data for the research has been identified, the next step involves downloading and storing said data. The procedure and means for extracting them largely depend on two key elements: firstly, the type of access granted to these data; and secondly, the platform or computer program used to analyse and exploit the text.

Regarding the type of access, the information to be extracted can be found in different types of domains: 1) private, wherein to access the data you must request permission from the owner, pay a fee or be the owner of the domain or a member of a specific community; 2) semi-private, as is the case when data owners offer application programming interfaces (APIs) that allow partial access to the data; or 3) open, wherein the data are completely public and available for download and use by anybody who is interested.

## 4.1. Download via API

Application programming interfaces (APIs) are a way of accessing software that allows external users to extract specific information (Qiu, 2017). In essence, an API serves as a communication key provided by the owners of the computer program, allowing users to directly access and obtain certain types of information. There are always access limitations depending on the permissions granted by the owner, although, in most cases, APIs grant the necessary permissions to extract the information required for the research.

A major advantage of working with APIs is the direct and established link between the computer program the information is requested from and the computer program that makes the request, as they use specific codes that remain the same. This streamlines the process of querying and extracting information. However, one of the drawbacks of using APIs is the need to obtain access and authentication keys, which can only be provided by the owners, meaning that requests may not necessarily be authorised. Lastly, APIs can either be used free of charge or for a fee. Links to free APIs for the main social media networks are provided in Appendix 1.

## 4.2. Web scraping

The practice known as web scraping refers to the process of extracting information from web pages by using software or programming code (Vilkova, 2020). The information obtained via this method is an exact replica of what is written in the domain that has been accessed, and the process is generally carried out in an automated manner using robots that simulate human behaviour on the website. This technique can be applied to various types of websites, social networks and even search engine results.

A significant advantage of web scraping compared to APIs is that it does not require a prior request for access, which makes it possible to perform on virtually any web page. However, this technique also presents additional challenges and requires more complex code. Although there are applications and websites that offer the ability to perform web scraping without the need to write code, these are usually paid-for computer programs or they offer only a limited free period. In the area of social research, web scraping could prove useful in order to gain access to opinion platforms and extract information from blogs and forums that cannot be accessed via an API.

## 4.3. Direct file download

The most practical and convenient option is to use data sources that allow you to download files directly. In such cases, the domain owner provides a user-friendly interface that allows you to specify search filters and manually download data in various formats. However, it is common to encounter limitations when it comes to downloading information due to the limited capacity of the servers. When researchers need to download large datasets, it is recommended to look for APIs that establish a direct connection with the data, as they are able to generate automated queries.

It is important to also consider the type of format in which the data is made available. In the case of files stored as Excel spreadsheets, the information can be handled easily as long as it does not exceed one million records (Microsoft, 2022). However, when the project at hand involves working with large amounts of data, it is advisable to store the information in formats such as comma-separated values (.csv) or in JavaScript object notation (.json), which contain data in a machine-readable format based on key-value pairs. When the data is only available for download in a digital document format like PDF (portable document format), it is recommended to convert the file to plain text.

# 5. Cleaning and preparing texts for analysis

Once the data has been stored properly, the next step is to clean and prepare the text, which will make it easier to perform analyses later on. In the field of data science, this is a vital and time-consuming step when working with large datasets. Properly cleaning the data helps ensure the satisfactory performance of the algorithms applied and, therefore, the quality and veracity of the results (Bird *et al.*, 2009). In the specific case of NLP, the chosen strategy mostly depends on the degree of complexity of the analysis to be applied. However, there are two unavoidable actions in textual data processing regardless of the tools used and our objectives.

## 5.1. Removing stop words

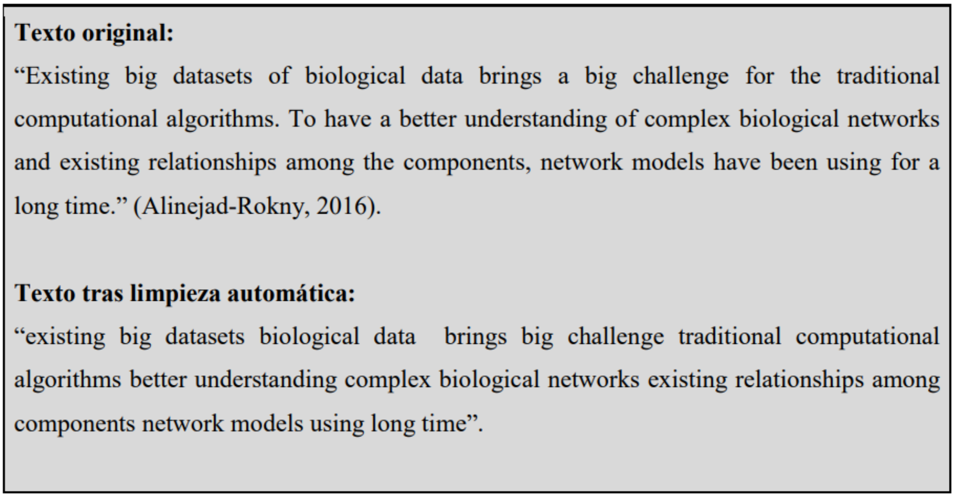
The first step involves excluding what are known as "*stop words*". Human language, in general, is characterised by the presence of connectors and auxiliary words that give coherence and continuity to communication. However, the main ideas and meaning of sentences are found in nouns, adjectives, verbs, and in some cases adverbs, which contain the substance of what we are trying to express or communicate. In text mining, the aim is to simplify the content and keep it homogeneous. To do this, the whole text is converted to lower case and all external links—such as hyperlinks—emoticons and similar characters are removed, if there are any.

Secondly, stop words, such as conjunctions (and, nor, but, same as, because, etc.) and prepositions (to, before, under, near, with, against, of, etc.), are also eliminated. Usually, statistical packages for natural language processing incorporate this function, automatically deleting any stop words. Even so, it is advisable to carry out some kind of test to find out which other words that are not included in the generic dictionaries are of little value to the corpus.

After carrying out the initial cleaning process, a word frequency analysis can be performed in order to identify irrelevant concepts. However, the number of words deleted must be kept to a minimum in order to preserve the integrity of the original information. The exclusion of these words during the subsequent cleaning stage should be clearly documented in the methodology section, providing adequate justification for the decisions made by the researcher. See Figure 1 for an illustrative example of how text is presented after automated cleaning.

**Figure 1**

*Example of automated text cleaning with the Python NLTK library*



**Texto original:**

"Existing big datasets of biological data brings a big challenge for the traditional computational algorithms. To have a better understanding of complex biological networks and existing relationships among the components, network models have been using for a long time." (Alinejad-Rokny, 2016).

**Texto tras limpieza automática:**

"existing big datasets biological data brings big challenge traditional computational algorithms better understanding complex biological networks existing relationships among components network models using long time".

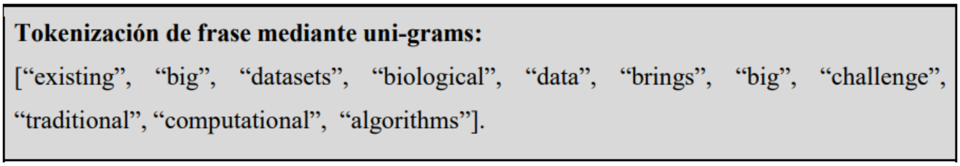
Source: own research.

## 5.2. Tokenisation and dictionary creation

The next step is to tokenise the text. Tokenisation is the process by which text is broken down into unique words or sentences (Saleem *et al.*, 2021). Its purpose is to reduce the text corpus down to the smallest unit size that will be processed as a single piece of data. Continuing with the previous example, the text can be fragmented into unique words known as unigrams so that the first sentence received by the machine will look as shown in Figure 2.

**Figure 2**

*Example of automated sentence tokenisation by unigram*



**Tokenización de frase mediante uni-grams:**

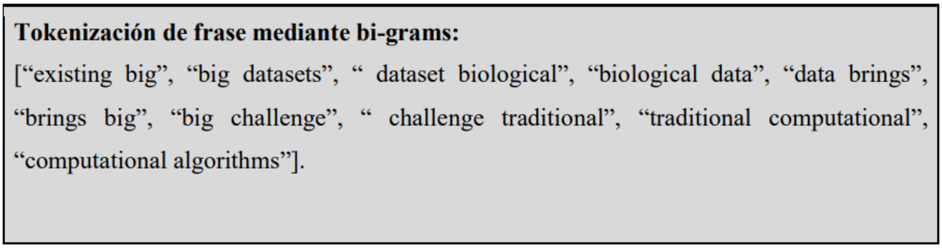
["existing", "big", "datasets", "biological", "data", "brings", "big", "challenge", "traditional", "computational", "algorithms"].

Source: own research using the Python NLTK library.

It may also be an interesting idea to combine two words for the minimum unit, as this offers greater context for the words analysed. Word pair fragments are known as bigrams. Figure 3 shows an example of this type of tokenisation.

## Figure 3

*Example of automated sentence tokenisation by* bigram



**Tokenización de frase mediante bi-grams:**

[“existing big”, “big datasets”, “ dataset biological”, “biological data”, “data brings”, “brings big”, “big challenge”, “ challenge traditional”, “traditional computational”, “computational algorithms”].

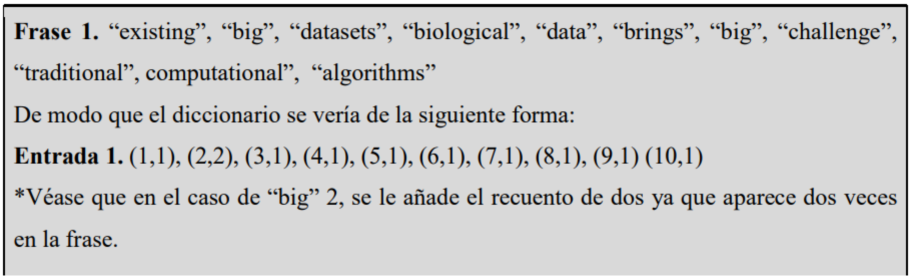
Source: own research using the Python NLTK library.

The most traditional tokenisation methods are by unigrams and bigrams, as shown above in the examples. However, tokenisation is possible by sets of as many words as desired, also known as “n-grams” (Saleem *et al.*, 2021). The final choice of number of words to group together will depend on the type of the corpus we are working with. In addition, performing analyses on different degrees of tokenisation will allow us to compare and give consistency to the results. Tokenisation is an automated process. In general, the software or code allows you to decide the number of grams by which you wish to tokenise the text.

Finally, these “tokens” will be transformed into numerical language, converting them into vectors in order for the machine to be able to count and perform calculations with the words. The vectorisation process can be carried out with multiple libraries and statistical packages, and we must find the transformation that best suits our specific needs. In this step it is important to consider the amount of text, our hard drive’s computational capacity and its speed. This vectorisation will create a dictionary that, to put it simply, assigns each token a unique number, followed by a second number that indicates the number of times that this token is repeated in the sentence. In the example below, the word “existing” is assigned the number 1, the word “big” is assigned the number 2, and so on.

## Figure 4

*Example of dictionary creation*



**Frase 1.** “existing”, “big”, “datasets”, “biological”, “data”, “brings”, “big”, “challenge”, “traditional”, computational”, “algorithms”

De modo que el diccionario se vería de la siguiente forma:

**Entrada 1.** (1,1), (2,2), (3,1), (4,1), (5,1), (6,1), (7,1), (8,1), (9,1) (10,1)

*Véase que en el caso de “big” 2, se le añade el recuento de dos ya que aparece dos veces en la frase.

Source: own research using the NLTK library.

## 6. Application of algorithms using machine learning techniques for text analysis

Machine learning is the branch of artificial intelligence where computational engineering meets mathematical statistics (James *et al.*, 2013). The main difference between this and classical programming is that until now the machine's actions were determined by a computer programmer, who was responsible for establishing a specific rule assigning a determined output to each input. Machine learning, on the other hand, allows the machine to learn the various inputs and outputs by itself, enabling it to generate new output patterns without there necessarily being prior programming (Müller and Guido, 2016). This type of engineering is what has made it possible for computer-based natural language processing to increase the complexity of textual analyses, refining the lexical variation and simplifying its multiple relationships, and bringing machines ever closer to the way humans use language.

As already mentioned, textual data analysis has long been used in the social sciences. This type of data is generally used in qualitative or mixed approaches, which in turn have gone on to form different methodological approaches. The objective of the research, as well as the medium and nature of the data, condition which type of analysis is chosen. However, most authors with experience in this field agree that the qualitative data analysis and, therefore, most text analyses, must necessarily involve two aspects: "data management and interpretation" (Gibbs, 2012, p. 23).

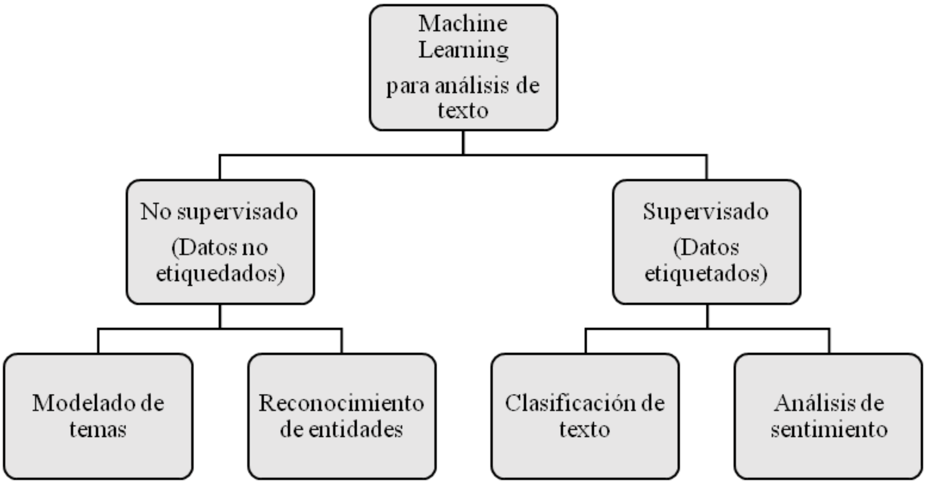
Data management serves to classify, label, order and reorganise the information, a vital step that will help establish the fundamental methodology of its subsequent interpretation. At present, this first "administrative" phase is carried out with the help of private computer-assisted qualitative data analysis software (CAQDAS), such as ATLAS.Ti and NVivo, combined with the researchers' theoretical criteria. Although these programs have undergone great developments in recent years, many of their limitations can be overcome using NLP techniques and machine learning tools.

Machine learning techniques offer a number of possibilities when it comes to managing large text corpora or unstructured datasets, which also opens up new opportunities for using such data when performing statistical analyses, such as correlations and time series. These techniques are not only able to be used for classical content analyses or categorisation, which focus on quantitative approaches, but they also facilitate the development of discursive analyses by uncovering key patterns and concepts in voluminous and complex datasets. Ultimately, the choice of one tool over another will depend on the nature of the research and its objectives, as well as whether or not the researcher wishes to triangulate data and contrast methods and results.

Machine learning, in turn, branches into two types of learning based on the characteristics of the data (inputs) that the researcher uses and the intended actions that the researcher wishes to perform with the obtained results (outputs). In the case of categorising or classifying data lacking associated information, unsupervised learning is used. On the other hand, when a portion of the labelled data is available and we wish to predict the labels of future observations, supervised learning is used. The distinguishing features of each learning approach are described below, as well as the main types of analyses that can be carried out (see Figure 5).

**Figure 5**

*Main types of text analyses using machine learning*


Machine Learning para análisis de texto
No supervisado (Datos no etiquedados)
Supervisado (Datos etiquetados)
Modelado de temas
Reconocimiento de entidades
Clasificación de texto
Análisis de sentimiento

Source: own research.

## 6.1. Unsupervised learning

Unsupervised learning is employed when working with unlabelled data, in other words, data that are not classified or that have previous responses associated with each piece of data or observation. The goal of this approach is to group observations based on their similarity or to perform entity recognition using statistical distance calculations. In doing so, the machine unveils patterns that cannot be seen by the naked eye (Müller and Guido, 2016). This methodology is especially useful for analysing textual data, since it allows us to gain a first look at corpora that are not classified or that lack prior information regarding their content.

Among the many examples we find in the current literature are works in which these approaches are used to discover the discourses and narratives found on social networks, as well as to carry out massive reviews of the literature. One example is the work conducted by Lindstedt (2019) in which unsupervised learning was applied in order to identify the main topics investigated in the literature related to social movements between 2005 and 2017. Another is the research carried out by Pavlova and Berkers (2020), the objective of which was to analyse the discourses on mental health present on the social network X/Twitter.

### *6.1.1. Topic modelling*

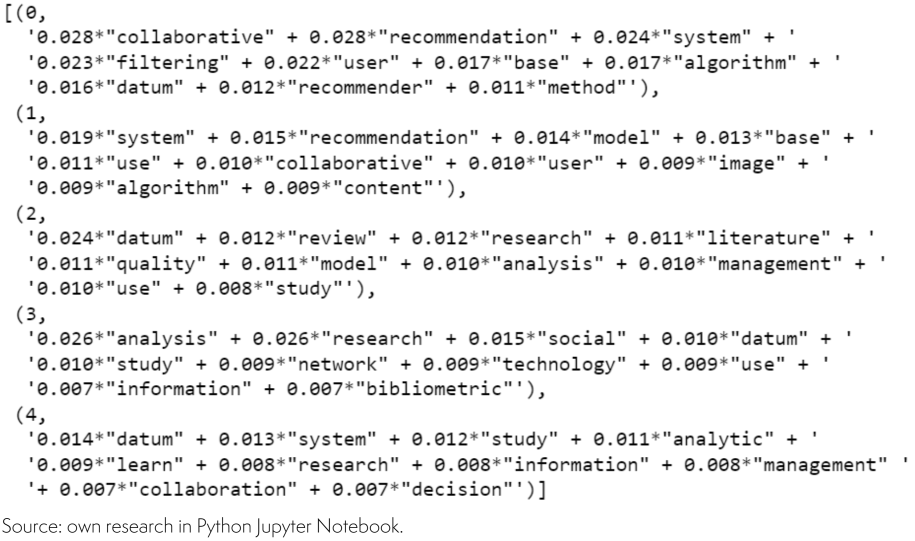
Topic modelling is a text analysis technique based on unsupervised machine learning that is widely used in text mining (Nikolenko *et al.*, 2017), the purpose of which is to discover the statistically significant topics found in the texts analysed in order, as in the examples mentioned above, to be able to obtain substantial generalised information regarding a set of texts.

This technique can be implemented using various algorithms, with latent Dirichlet allocation (LDA) being the most widely used in the scientific literature. LDA groups together the most frequent words based on their similarity. However, recent research comparing the performance of different algorithms has revealed the limitations of LDA. In recent years, the BERTopic algorithm has been proposed as a promising alternative, although it has yet to be widely adopted in social science work (Egger and Yu, 2022).

Regardless of the algorithm used, through topic modelling we can classify texts grouped by similarity and differentiated from one other by distance, while we also obtain a series of keywords from each group which, depending on the researcher's criteria, must be given a descriptive name based on the set of words in question.

## Figure 6

*Example of output from the LDA algorithm applied to a text corpus*



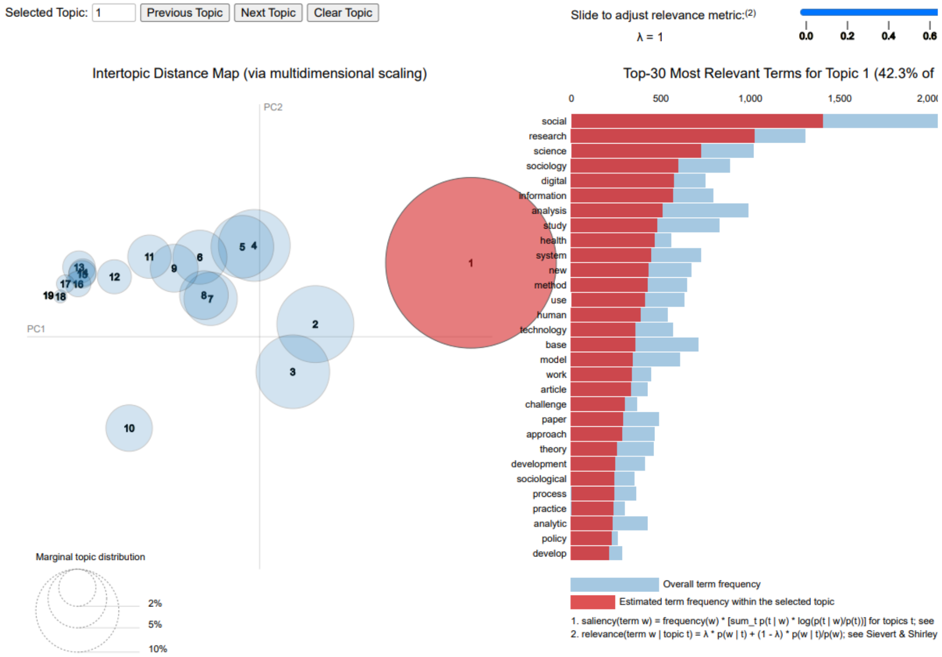
```
[(0,
  '0.028*"collaborative" + 0.028*"recommendation" + 0.024*"system" + '
  '0.023*"filtering" + 0.022*"user" + 0.017*"base" + 0.017*"algorithm" + '
  '0.016*"datum" + 0.012*"recommender" + 0.011*"method"'),
 (1,
  '0.019*"system" + 0.015*"recommendation" + 0.014*"model" + 0.013*"base" + '
  '0.011*"use" + 0.010*"collaborative" + 0.010*"user" + 0.009*"image" + '
  '0.009*"algorithm" + 0.009*"content"'),
 (2,
  '0.024*"datum" + 0.012*"review" + 0.012*"research" + 0.011*"literature" + '
  '0.011*"quality" + 0.011*"model" + 0.010*"analysis" + 0.010*"management" + '
  '0.010*"use" + 0.008*"study"'),
 (3,
  '0.026*"analysis" + 0.026*"research" + 0.015*"social" + 0.010*"datum" + '
  '0.010*"study" + 0.009*"network" + 0.009*"technology" + 0.009*"use" + '
  '0.007*"information" + 0.007*"bibliometric"'),
 (4,
  '0.014*"datum" + 0.013*"system" + 0.012*"study" + 0.011*"analytic" + '
  '0.009*"learn" + 0.008*"research" + 0.008*"information" + 0.008*"management" '
  '+ 0.007*"collaboration" + 0.007*"decision"')]
```

Source: own research in Python Jupyter Notebook.

Figure 6 shows a concrete example illustrating the results of applying latent Dirichlet allocation (LDA) analysis to a text corpus composed of various investigations in the fields of business and technology. This specific case was implemented using the Python programming language in the Jupyter Notebook environment by using the Gensim library. However, as summarised in Table 1, it is feasible to carry out this procedure in a number of environments and even to use the software mentioned in Section 2 without requiring prior programming knowledge.

## Figure 7

*Example of a graphical view of the LDA algorithm*



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley

Source: own research using the Python pyLDAvis library.

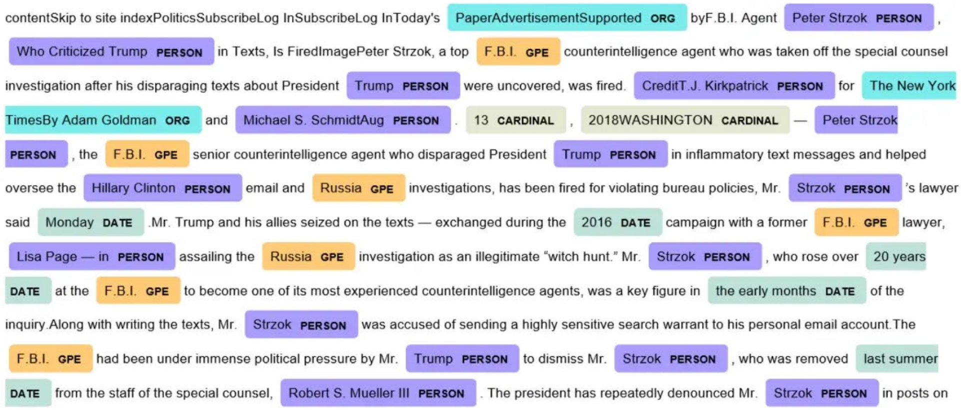
Figure 7 shows a way of viewing this algorithm using the pyLDAvis library. This tool allows us to gain a better understanding of how clusters and topics are distributed, as well as the most relevant words associated with each of them. The next step is to extract the information and, based on a study of a sample of the texts selected for each group, make thematic deductions and assign a name to each cluster.

### *6.1.2. Character or entity recognition*

Another of the most-used techniques in unsupervised learning is entity recognition, or more specifically, named entity recognition (NER). In this process, the machine is able to detect predefined entities in a text corpus. The algorithm can detect where in the text people, places, companies and numbers, among other elements, are mentioned (Calzolari, 2020). This can be an interesting way to label texts and organise them quickly and effectively. These entities are recognised using pre-existing dictionaries that the algorithm consults to detect nouns in the corpus that have been previously labelled in other text corpora. For this reason, although we apply this technique as part of unsupervised learning, the reality is that it is a mixed method that combines aspects of supervised and unsupervised learning.

## Figure 8

*Example output by applying NER*



contentSkip to site indexPoliticsSubscribeLog InSubscribeLog InToday's PaperAdvertisementSupported ORG byF.B.I. Agent Peter Strzok PERSON , Who Criticized Trump PERSON in Texts, Is FiredImagePeter Strzok, a top F.B.I. GPE counterintelligence agent who was taken off the special counsel investigation after his disparaging texts about President Trump PERSON were uncovered, was fired. CreditT.J. Kirkpatrick PERSON for The New York TimesBy Adam Goldman ORG and Michael S. SchmidtAug PERSON . 13 CARDINAL , 2018WASHINGTON CARDINAL — Peter Strzok PERSON , the F.B.I. GPE senior counterintelligence agent who disparaged President Trump PERSON in inflammatory text messages and helped oversee the Hillary Clinton PERSON email and Russia GPE investigations, has been fired for violating bureau policies, Mr. Strzok PERSON 's lawyer said Monday DATE .Mr. Trump and his allies seized on the texts — exchanged during the 2016 DATE campaign with a former F.B.I. GPE lawyer, Lisa Page — in PERSON assailing the Russia GPE investigation as an illegitimate "witch hunt." Mr. Strzok PERSON , who rose over 20 years DATE at the F.B.I. GPE to become one of its most experienced counterintelligence agents, was a key figure in the early months DATE of the inquiry.Along with writing the texts, Mr. Strzok PERSON was accused of sending a highly sensitive search warrant to his personal email account.The F.B.I. GPE had been under immense political pressure by Mr. Trump PERSON to dismiss Mr. Strzok PERSON , who was removed last summer DATE from the staff of the special counsel, Robert S. Mueller III PERSON . The president has repeatedly denounced Mr. Strzok PERSON in posts on

Source: Li (2018).

Figure 8 shows a specific example of the output displayed on screen when applying the spaCy library for visualisation in the Python programming language within the Jupyter Notebook environment. These libraries also offer the ability to store classifications in lists or databases, enabling them to be used and processed in the future.

By applying both topic modelling and entity recognition, we are able to label and classify texts that were originally unstructured. As we have already mentioned, these analyses can be applied both in environments that require programming and in coding-free software. The advantage of this type of technique is that it enables us to quickly create generalised topic and concept maps. However, the researcher plays a much more important role when interpreting the topics and hypothesising.

It is important to note that in the case of topic modelling, it is the researcher who must decide on the number of clusters into which they wish to classify their texts. Although there are evaluation metrics that can be used to statistically compare the results, with the consistency metric being the most popular (Stevens *et al.*, 2012), the current literature argues that these metrics should be considered in an indicative manner. Ultimately, the meaning and significance of the results obtained will come down to the researcher's theoretical criteria.

## 6.2. Supervised learning

Supervised learning is the approach used to process labelled or classified data in order to make predictions about the classification of future observations with the same criteria and characteristics (Shahbaz *et al.*, 2022). Using this type of algorithm, the machine is instructed with a training dataset which provides it with the input characteristics and the corresponding output labels. The main objective of this is to enable the machine to predict what the label will be for a new observation. Supervised learning can reach considerable levels of complexity, and the current possibilities in this field are wide and diverse.

This type of approach is especially useful in identifying specific categories, predicting results and assigning labels to new texts based on previously established patterns. Unlike unsupervised learning, which focuses on general and descriptive analytics, supervised machine learning is suitable for more precise and targeted applications. Supervised machine learning techniques have been used in the scientific literature in a range of contexts. For example, Naseeba *et al.* (2023) employed this approach to classify newspaper articles, while Khanday *et al.* (2022) focused on detecting hate speech on social networks. Mbona and Eloff (2023), on the other hand, applied supervised machine learning techniques to bot identification, and Shevtsov *et al.* (2023) explored its usefulness in electoral analysis.

### *6.2.1. Text classification*

Text classification can be very useful in studies involving large volumes of data. In this scenario, it is first necessary to classify the set of texts and the predicted outputs, and then apply this knowledge to the rest of the corpus to allow the algorithm to perform the classification automatically. In this sense, it is advisable to start based on previous knowledge of possible theoretical categories and typologies in order to generate labels that can either support or refute hypotheses. Therefore, the prior labelling task carried out by the researcher must be solidly grounded and follow a comprehensive coding process, which will allow the algorithms to make predictions with greater accuracy. Before performing this labelling, a clear and detailed coding guide must be created to ensure that coders follow uniform and solid criteria when classifying the text unit to be analysed.

Text classification can be carried out using various algorithms. As with topic modelling, we found multiple papers in the recent literature that evaluate the performance of different classification algorithms (Dogra *et al.*, 2022). In the area of social sciences, one of the most widely used types of algorithms in scientific research are support vector machines (SVMs), which are able to classify text groups with high dimensionality (Joachims, 1998).

It is important to note that, when using classification algorithms, the training data must have the same characteristics as the data to which the algorithm is going to be applied. For example, if the algorithm has been trained to obtain two response classes, such as "extremist" and "moderate", this algorithm's output will always correspond to these two labels. Therefore, if the new sets of texts are expected to contain additional categories that are not provided for in the labelled sample, a new sample in which the additional categories are labelled will be needed and the algorithm will have to be retrained.

### *6.2.2. Sentiment analysis*

Sentiment analysis is a broad analysis type commonly found in the field of business and other social sciences. It has traditionally been carried out through customer satisfaction surveys. However, the increase in the number of digital commerce platforms has led to the development of new methods that, in turn, have enabled these analyses to be applied to new and varied areas of research such as electoral analysis, studies on migration and hate speech and communication studies, among others. Sentiment analysis using machine learning allows us to identify emotions in large text corpora. This works with the same logic as text classification; however, as with the NER technique, there are multiple already trained dictionaries that, without the need to previously label a sample of the database, the algorithm can use to recognise the sentiment in our text corpus.

In both cases the result is a text corpus labelled according to previously established theoretical criteria, unlike in unsupervised analysis. When creating a classification algorithm, the goal is to obtain the highest percentage of accuracy in the prediction model we have generated. Supervised learning models are evaluated by following an established procedure.

First, a sample that requires labelling is extracted from the text corpus. This sample is divided into a training set (70%) and a test set (30%). The model is trained using 70% of the sample data by showing it the observation or data and then the output response. Once the model is trained, its classification ability is evaluated using the other 30% of the sample data, or the test set. Based on the accuracy obtained in this evaluation, a decision is made on whether the model is accurate enough to classify the unseen data. In order to achieve a suitably generalised model, a large-volume text corpus is needed. In addition, the level of precision the model is found to have must be specified in the research methodology.

**Table 1**

*Main libraries for each type of analysis and programming language*

| Type of Analysis | Python programming language | R programming language |
|---|---|---|
| Topic modelling | -Gensim<br>-scikit-learn | -topicmodels<br>-LDA |
| Entity recognition | -SpaCy<br>-NLTK | -spacyr<br>-openPLN |
| Text classification | -NLTK<br>-TensorFlow (Keras) | -TM<br>-Caret |
| Sentiment analysis | -scikit-learn<br>-Keras | -tidytext<br>-quanteda |

Source: own research.

# 7. Conclusion

This work offers a synthesised guide to the basics of applying natural language processing (NLP) in social science research within the framework of machine learning. First, it provides an overview of the history and origin of natural language processing, followed by the practical aspects of applying NLP techniques. Details are given of some of the programming languages and software that can be used to carry out analyses, as well as the various text sources and the methods for extracting and storing them. Next is a section on text cleaning and processing, followed by a description of the various types of analysis techniques that can be applied to natural language processing.

Although in the international scientific literature there is a growing interest in the transdisciplinary integration between the different branches of social sciences and computer techniques, there is a clear need to continue developing research and pedagogical works that encourage and motivate social scientists to adopt and actively use these types of techniques in their studies. The implementation of these methodologies can provide significant benefits, such as deeper and more rigorous data analysis, identification of hidden patterns and trends and a more complete and enriching understanding of social phenomena.

This guide is intended as an introductory resource for researchers interested in delving into the field of machine learning and NLP. Its fundamental purpose is to provide clear and effective guidance that helps researchers choose between the many widely validated approaches, programs and algorithms that are available, since the incredible amount of options can be overwhelming for those who are in the early stages of their training in this field of study.

It is worth pointing out that while more recent work employing natural language processing (NLP) in social science research supports the effectiveness of the techniques discussed here, the current discussion identifies significant challenges that social researchers will face in the near future.

First of all, one of the main challenges encountered by social researchers is the existence of biases both in the models used and in their training process. It is essential to keep in mind that machines tend to reproduce the prejudices inherent in human beings. Therefore, it is crucial to avoid such biases in all stages of the process, from data extraction to training the algorithms (Zwilling, 2023).

Another challenge lies in the current inability of models to "understand" the cultural particularities and jargon expressions found in the data. These particularities and expressions are often of significant value to social researchers, and their correct interpretation is crucial to ensure an accurate and relevant analysis (Sambeek, 2021).

Finally, it is worth highlighting the notable dearth of properly annotated datasets that are suitable for training supervised models in the context of social sciences. In many cases, the lack of labelled data for specific areas or topics of interest in fields such as sociology and political science is rather evident. This leads to the need to resort to transfer learning techniques and devise specific strategies aimed at addressing this limitation, with the primary objective of achieving results that are both reliable and representative.

These challenges underline the need to continue researching and developing NLP applied to the social sciences in order to overcome the current restraints and ensure a rigorous and robust analysis of the texts in this field. In addition, it is essential to highlight the importance of making these analyses replicable and reproducible by indicating in the methodology section the libraries, tools and environments used, as well as the verifications and modifications made, from cleaning the text corpus to evaluating and visualising the results.

# 8. Appendix

*Appendix 1. Links to the most popular social media APIs*

| Platform | API |
|---|---|
| Facebook | https://developers.facebook.com/docs/graph-api |
| Instagram | https://developers.facebook.com/docs/instagram |
| X/Twitter | https://developer.twitter.com/en/docs |
| YouTube | https://developers.google.com/youtube/v3/docs/ |
| Reddit | https://www.reddit.com/dev/api/ |

Source: own research.

# 9. References

Abbott, A. (1997). Of Time and Space: The Contemporary Relevance of the Chicago School. *Social Forces*, 75(4), 1149. https://doi.org/10.2307/2580667.

Ajmal, S., Khan, S., Hossain, M., Lomonaco, V., Cannons, K., Xu, Z. and Cuzzolin, F. (2022). International Workshop on Continual Semi-Supervised Learning: Introduction, Benchmarks and Baselines. *Continual Semi-Supervised Learning*, Vol. 13418 (pp. 1–14). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-17587-9_1

Alinejad-Rokny, H. (2016). Proposing on Optimized Homolographic Motif Mining Strategy Based on Parallel Computing for Complex Biological Networks. *Journal of Medical Imaging and Health Informatics*, *6*(2), 416–424. https://doi.org/10.1166/jmihi.2016.1707

Bird, S., Klein, E. y Loper, E. (2009). *Natural language processing with Python*. O'Reilly.

Bitter, C., Elizondo, D. A. and Yang, Y. (2010). Natural language processing: A prolog perspective. *Artificial Intelligence Review*, *33*(1–2), 151–173. https://doi.org/10.1007/s10462-009-9151-4

Calzolari, N. (2020). *LREC 2020 Marseille Twelfth International Conference on Language Resources and Evaluation$dMay 11–16, 2020, Palais Du Pharo, Marseille, France: Conference Proceedings*. Paris: The European Language Resources Association (ELRA).

Castells, M. (2018). *La era de la información: economía, sociedad y cultura. Vol. 3, Fin de milenio*. 4th ed., 2nd reprint. Madrid: Alianza Editorial.

Dahlin, E. (2021). Email Interviews: A Guide to Research Design and Implementation. *International Journal of Qualitative Methods*, 20 https://doi.org/10.1177/16094069211025453.

Dhiraj, M. (2008). Digital Ethnography: An Examination of the Use of New Technologies for Social Research. *Sociology*, 42(5), 837–855. https://doi.org/10.1177/0038038508094565.

Dogra, V., Verma, S., Kavita, Chatterjee, P., Shafi, J., Choi, J. and Ijaz, M. F. (2022). A Complete Process of Text Classification System Using State-of-the-Art NLP Models. In S. K. Sah Tyagi (Ed.), *Computational Intelligence and Neuroscience* (pp. 1–26). https://doi.org/10.1155/2022/1883698.

Egger, R. and Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, 7. https://doi.org/10.3389/fsoc.2022.886498.

Gibbs, G. (2012). *El análisis de datos cualitativos en investigación cualitativa*. Madrid: Ediciones Morata.

Gillingham, P. y Graham, T. (2017). Big Data in Social Welfare: The Development of a Critical Perspective on Social Work's Latest «Electronic Turn». *Australian Social Work*, 70(2), 135–147. https://doi.org/10.1080/0312407X.2015.1134606

Gualda, E., Taboada Villamarín, A. and Rebollo Díaz, C. (2023). Big data y ciencias sociales: Una mirada comparativa a las publicaciones de antropología, sociología y trabajo social. *Gazeta de Antropología*, *39*(1).

Gualda, E. and Rebollo, C. (2020). Big data y Twitter para el estudio de procesos migratorios: Métodos, técnicas de investigación y software. *Empiria. Revista de metodología de ciencias sociales*, *46*, 147. https://doi.org/10.5944/empiria.46.2020.26970

Hockett, C. F. (2020). The state of the art. *The State of the Art*. De Gruyter.

Holtz, P., Kronberger, N. and Wagner, W. (2012). Analyzing Internet Forums: A Practical Guide. *Journal of Media Psychology*, 24(2), 55–66. https://doi.org/10.1027/1864-1105/a000062

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning* (vol. 103). New York: Springer. https://doi.org/10.1007/978-1-4614-7138-7

Johri, P., Khatri, S. K., Al-Taani, A. T., Sabharwal, M., Suvanov, S. and Kumar, A. (2021). Natural Language Processing: History, Evolution, Application, and Future Work. In A. Abraham, O. Castillo and D. Virmani (Eds.), *Proceedings of 3rd International Conference on Computing Informatics and Networks* (vol. 167, pp. 365–375). Springer Singapore. https://doi.org/10.1007/978-981-15-9712-1_31

Justicia de la Torre, C., Sánchez, D., Blanco, I. and Martín-Bautista, M. J. (2018). Text Mining: Techniques, Applications, and Challenges. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 26(04), 553–582. https://doi.org/10.1142/S0218488518500265

Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R. and Malik, S. H. (2022). Detecting Twitter Hate Speech in COVID-19 Era Using Machine Learning and Ensemble Learning Techniques. *International Journal of Information Management Data Insights*, 2(2), 100120. https://doi.org/10.1016/j.jjimei.2022.100120.

Li, S. (2018). *Named Entity Recognition and Classification with Scikit-Learn.* https://towardsdatascience.com/named-entity-recognition-and-classification-with-scikit-learn-f05372f07ba2

Lindstedt, Nathan C. (2019). Structural Topic Modeling For Social Scientists: A Brief Case Study with Social Movement Studies Literature, 2005-2017. *Social Currents*, *6*(4), 307-318. https://doi.org/10.1177/2329496519846505.

Maud, R. and Blanchard, A. (2022). The Framing of Health Technologies on Social Media by Major Actors: Prominent Health Issues and COVID-Related Public Concerns. *International Journal of Information Management Data Insights*, 2(1), 100068. https://doi.org/10.1016/j.jjimei.2022.100068.

Mbona, I. and Eloff, J. H. P. (2023). Classifying Social Media Bots as Malicious or Benign Using Semi-Supervised Machine Learning. *Journal of Cybersecurity*, *9*(1), tyac015. https://doi.org/10.1093/cybsec/tyac015.

Michel, J.–B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A. and Aiden, E. L. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, *331*(6014), 176–182. https://doi.org/10.1126/science.1199644

Microsoft (2022). *Especificaciones y límites de Excel*. https://support.microsoft.com/es-es/office/especificaciones-y-l%C3%ADmites-de-excel-1672b34d-7043-467e-8e27-269d656771c3

Morimoto, J. and Ponton, F. (2021). Virtual reality in biology: Could we become virtual naturalists? *Evolution: Education and Outreach*, *14*(1), 7. https://doi.org/10.1186/s12052-021-00147-x

Müller, A. C. and Guido, S. (2016). *Introduction to aprendizaje automático with Python: A guide for data scientists*. O'Reilly Media, Inc.

Naseeba, B., Challa, N. P., Doppalapudi, A., Chirag, S. and Nair, N. S. (2023). Machine Learning Models for News Article Classification. *5th International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 1009–1016). Tirunelveli, India: IEEE. https://doi.org/10.1109/ICSSIT55814.2023.10061095

Nikolenko, S. I., Koltcov, S. and Koltsova, O. (2017). Topic modelling for qualitative studies. *Journal of Information Science*, *43*(1), 88–102. https://doi.org/10.1177/0165551515617393

Pavlova, A., and Berkers, P. (2020). Mental Health Discourse and Social Media: Which Mechanisms of Cultural Power Drive Discourse on Twitter. *Social Science & Medicine,* 263, 113250. https://doi.org/10.1016/j.socscimed.2020.113250.

Piotrowski, M. (2012). *Natural Language Processing for Historical Texts*. Cham: Springer. https://doi.org/10.1007/978-3-031-02146-6

Radick, G. (2016). The unmaking of a modern synthesis: Noam Chomsky, Charles Hockett, and the politics of behaviorism, 1955–1965. *Isis*, *107*(1), 49–73. https://doi.org/10.1086/686177

Ruelens, A. (2022). Analyzing user-generated content using natural language processing: A case study of public satisfaction with healthcare systems. *Journal of Computational Social Science*, *5*(1), 731–749. https://doi.org/10.1007/s42001-021-00148-2

Saleem, Z., Alhudhaif, A., Qureshi, K. N. and Jeon, G. (2021). Context-aware text classification system to improve the quality of text: A detailed investigation and techniques. *Concurrency and Computation: Practice and Experience*. https://doi.org/10.1002/cpe.6489

Sambeek, I. (2021). Natural Language Processing & Social Sciences. Towards Data Science. https://towardsdatascience.com/natural-language-processing-social-sciences-94a35a8a7c78

Shevtsov, A., Oikonomidou, M., Antonakaki, D., Pratikakis, P. and Ioannidis, S. (2023). What Tweets and YouTube Comments Have in Common? Sentiment and Graph Analysis on Data Related to US Elections 2020. *PLOS ONE*, *18*(1), e0270542. https://doi.org/10.1371/journal.pone.0270542.

Thorsten, J. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. En C. Nédellec y C. Rouveirol, *Aprendizaje automático: ECML-98*. Vol. 1398, *Lecture Notes in Computer Science* (pp. 137–142). Berlin, Heidelberg: Springer. https://doi.org/10.1007/BFb0026683

Vilkova, O. (2020). Web Scraping as a Method of Data Extraction in Sociological Studies: On Scientific Applicability. *Vestnik Tomskogo gosudarstvennogo universiteta. Filosofiya, sotsiologiya, politologiya,* (54), 163–175. https://doi.org/10.17223/1998863X/54/16.

Yuanbo, Q. (2017). The Openness of Open Application Programming Interfaces. *Information, Communication & Society,* 20(11), 1720–36. https://doi.org/10.1080/1369118X.2016.1254268.

Zwilling, Moti (2023). Big Data Challenges in Social Sciences: An NLP Analysis. *Journal of Computer Information Systems*, *63*(3), 537–554. https://doi.org/10.1080/08874417.2022.2085211.

## Alba Taboada Villamarín

PhD student in Economics and Business at the Autonomous University of Madrid. She was awarded a scholarship as a member of predoctoral research staff in training (FPI) on the CONCERN R&D project (PID2020-115095RB-I00) and she is also part of the NON-CONSPIRA-HATE! R&D project (PID2021-123983OB-I00) team. She graduated in sociology from Complutense University of Madrid (UCM) and completed a master's degree in Big Data Science at the University of Navarra. She was also a fellow at the Centre for Sociological Research (CIS) in 2022. She is currently researching new methodological approaches through big data and machine learning applied to the social sciences.