GOBIERNO DE ESPAÑA
MINISTERIO DE INDUSTRIA, ENERGÍA Y TURISMO

red.es

ontsi observatorio nacional de las telecomunicaciones y de la SI

PILOT PROJECT ON THE VIABILITY OF USING THE INTERNET AS A DATA SOURCE

Final Report 2014

Universidad Carlos III de Madrid

MANAGEMENT AND COORDINATION:

Luis Muñoz López
Pedro Antón Martínez
Red.es

Jesús Cid Sueiro
Miguel Lázaro Gredilla
Sergio Muñoz Rodríguez
Carlos III University of Madrid (UC3M)

EDITING:

Jesús Cid-Sueiro
Ángel Navia Vázquez
Vanessa Gómez Verdejo
Jerónimo Arenas García
Emilio Parrado Hernández
Sergio Muñoz Romero
Jesús Fernández Bes
Miguel Lázaro Gredilla
Carlos III University of Madrid (UC3M)

Contents

1 Forward

The Government of Spain has determined to formulate a Digital Agenda for Spain as a frame of reference for a road map in the area of Information and Communication Technology (ICT) and electronic administration; establish Spain's strategy for achieving the objectives of the Digital Agenda for Europe; maximise the impact of public ICT policies for improving productivity and competitiveness; transform and modernise the economy and Spanish society through the effective and intensive use of ICT by the population, companies and government administrations.

The Digital Agenda for Spain describes how enhancing digital skills in the enterprise and labour spheres leads to tangible benefits. Finding a job or accessing training resources is easier with computer qualifications than without them. The same applies to developing innovative services for companies or meeting new demands of people and partner entities, customers and users. Improving our educational system to meet the demands of the new ICT professions and linking training measures to the generation of quality jobs are indispensable imperatives. It is vital to adapt our educational systems to respond to recent demand for new ICT skills and professions. Systems must be adapted at both professional and university training levels. In this regard, the new professions will be related to e-commerce, digital marketing, with the digital content industry, cloud computing, intensive computation, smart cities, the Internet of things and with the trusted digital products and services industry. Periodically aligning ICT training with market needs, encouraging collaboration between businesses and educational centres and obtaining more multipurpose ICT skills in the business and management spheres are essential for training new ICT professionals. In order to exceed these ambitious objectives, we perceive a need to adequately characterise the development of the demand for information, communication and digital content professionals as well as the training available at universities and in professional training centres. This need emerges, on the one hand, from the demand for accurate and up to date information from the sector itself and from the population; and, on the other hand, from the various public administrations responsible for developing policies that promote the Information Society in Spain.

The objectives and action lines of the Digital Agenda for Europe, the Digital Agenda for Spain and the specific plans that can be implemented underline the desirability of undertaking activities aimed at expanding and improving systems for measuring Information Society development indicators. Consequently, the Spanish Observatory for Telecommunications and the Information Society (ONTSI), part of the Public Corporate Entity Red.es, reporting to the State Secretariat for Telecommunications and the Information Society, has been working on the collection, production, publication and analysis of data, indicators and studies on the development of the Information society in Spain and their comparison with international sources. ONTSI prepares, collects, synthesises and systematises indicators, performs studies, and offers news and updates on the Information Society, which can be accessed on its website (http://www.ontsi.es). ONTSI uses traditional research methods and techniques, proven and harmonised in a comparative context with the countries of the OECD and the European Union, for producing its studies and reports. ONTSI uses models that integrate and channel the information produced with its own research resources and with those derived from all the initiatives and projects sourced by third party national and international, public and private, initiatives and projects in order to achieve a common objective, guaranteeing the alignment of individual objectives of each project and initiative with the Digital Agenda for Spain and as modulated by the objectives set and the results of the impact on the whole of Society.

In this sense, the State Secretariat for Telecommunications and the Information Society instructed Red.es to implement a pilot project to analyse the viability of using the Internet as a Data Source (IaD) for monitoring availability of and demand for ICT professionals in Spain.

In parallel with this instruction, given the importance and the upswing of e-commerce in the Spanish economy, Red.es and ONTSI started another pilot project to analyse e-commerce in Spanish businesses. One of the specific plans of the Digital Agenda for Spain refers to the use of ICT by SMEs and e-commerce. With respect to e-commerce, point III of this Plan sought to achieve the objectives set in the Digital Agenda for 2015 in the field of e-commerce through financial support measures for increasing what is on offer on the Internet and promoting the availability of professionals trained in the new disciplines and technologies required by electronic sales. The objective of the pilot was to monitor the products and services on offer via e-commerce by Spanish companies, using the information available on the Internet.

Both businesses and citizens leave a large number of 'digital tracks' on the Internet. By collecting and using this information it is possible to discover various socio-economic phenomena in almost real time. IaD enables the identification of data and indicators that can be obtained directly from the Internet, describing new habits and uses that are not covered by traditional methods or that demand an enormous amount of economic and human resources, making them nonviable. Using IaD can provide a quick view on new phenomena, which traditional techniques would find difficult to measure. It can improve the quality of statistics, especially when combined with traditional methods. Furthermore, it can be a way of reducing the workload on the informing units, whether businesses or individuals. In terms of defining future Information Society policies, IaD may be a possible alternative for obtaining data on the uses of the Internet and other phenomena related to the Information Society.

This report outlines both pilot projects on using the Internet as a Data Source, one for monitoring availability of and demand for ICT professionals in Spain, and the other for monitoring e-commerce in Spanish businesses.

It was possible to carry out both projects thanks to the interest and resources provided by the State Secretariat for Telecommunications and the Information Society and by the Public Corporate Entity Red.es, and to the joint work of ONTSI with the Information Management and Processing Group of the Department of Signal Theory and Communications of Carlos III University of Madrid.

2 Introduction

This report is the result of the activities of the service contract "Development of a Pilot Project for Analysing the Viability of Using the Internet as a Data Source, Simplified Procedure, File 107/113-OT". The expression, "Internet as a Data Source" (IaD) refers to the use of advanced data analysis techniques that use the Internet as an additional or substitute source for traditional sources of statistical data. What differentiates these methods from other Internet orientated techniques such as online surveys is their completely automated and nonintrusive nature.

This study describes procedures, results and conclusions regarding two major milestones or sub-projects of this project. The first was aimed at applying automated classification techniques for detecting and characterising the use of e-commerce services (B2C, Business to Consumer) in Spain. The aim of the second was the application of automatic classification techniques for detecting and characterising the availability and demand for ICCT professionals (Information, Communication and Content Technologies) in Spain. In both cases, automation attempts to avoid and minimise the tasks of browsing and manually recording websites (of businesses, job portals and official qualification catalogues).

The sub-project for automatic detection of B2C commerce is described in chapters 2 to 4. The sub-project for detection and characterisation of job offers is described in chapters 5 to 10. The description of each sub-project has been structured in three main parts: the first describes the main features of the software application developed to meet the objectives of each sub-project (chapters 2 and 5). The second describes the results obtained in each sub-projects through the use of the software applications (chapter 3 describes the results of detecting B2C commerce, and chapters 6-9 describe the various results obtained in the second sub-project). Finally, chapters 4 and 10 establish conclusions on the viability of using the Internet as a Data source in each sub-project in the light of the results obtained and described in the previous chapters. We have tried to avoid using excessively technical language so that the results are accessible to readers who are not experts in the technical aspects of the project. More detailed information on the software architecture and the automatic learning and data manipulation algorithms used are described in a second document, internal to the project.

3

MATERIALS AND METHODS FOR AUTOMATED DETECTION OF B2C COMMERCE

3 Materials and Methods for the Automated Detection of B2C Commerce

OBJECTIVE

Analyse the viability of AUTOMATED DETECTION OF B2C COMMERCE (business to consumer e-commerce) on the websites of Spanish businesses.

The starting point of the study was a listing of URLs of Spanish businesses, provided by Red.es, with additional information described in detail in section 4.1. The final objective of this sub-project was to analyse the viability of machine learning (ML) algorithms for automatically detecting the possibility of making electronic purchases (B2C, Business to Consumer) on these websites. To do this, it was necessary to complete three main phases:

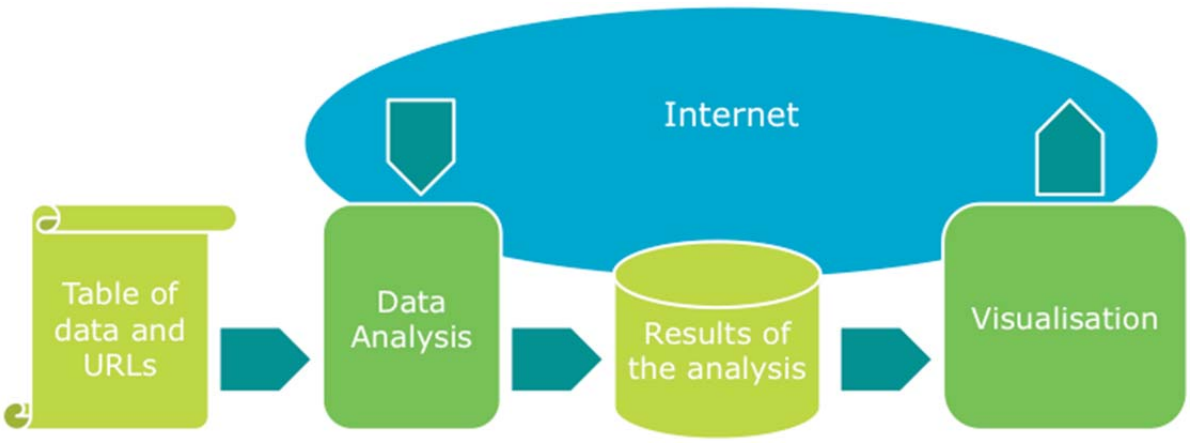
1. Develop software for capturing, analysing and visualising the data.
2. Apply the developed software to detecting B2C e-commerce in Spanish businesses.
3. Analyse the results obtained with the purpose of drawing conclusions on the viability of ML for detecting B2C activity.

These three phases were highly interrelated: the analysis software was modified in successive iterations as a result of the conclusions of phases 2 and 3 in a process that was cyclical rather than sequential. These iterations in design were fundamental for selecting the ML algorithms and configuring them to optimise performance in detecting B2C e-commerce. Although we will discuss part of this process in the following sections, we will focus here on the final application, configured for detecting B2C e-commerce. We start with a user orientated description of the application that seeks to detect B2C e-commerce in a new data base.

3.1 Process of Detecting B2C E-commerce

As indicated above, the software development phase for automated detection of B2C e-commerce requires a first stage of data capture and analysis (automated) and a second phase of visualisation and analysis (by users) of the results of the automated process. This process is shown in Figure 1.

Figure 1. Process of Detecting B2C E-commerce



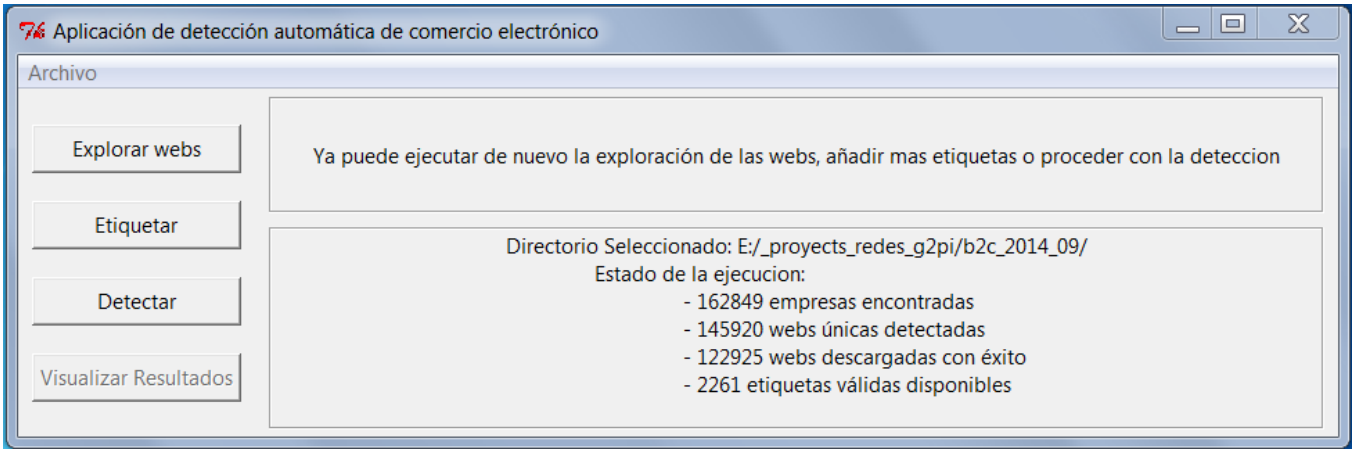
Although data analysis and visualisation can be performed from a single application that integrates both processes, the visualisation module is independent and the results of the analysis can be viewed using a conventional web browser, without the need to launch the integrated application.

The data analysis process is described first.

3.2 Process for Data Analysis

When launching the data capture and analysis application (details on its installation and operation can be found in the appendices document of this report), a window appears that is shown in the following Figure.

Figure 2. Main Window of the Data Capture and Analysis Software



The **Archivo** (File) menu allows the user to select the folder containing all the Excel files that contain lists of businesses with the URLs of their sites.

The application allows maintaining a number of different active projects, each with its corresponding listing of businesses.

The process of data capture, analysis and visualisation requires 4 main steps, which correspond to the buttons shown in the window.

3.2.1 PHASE 1: Web Browsing ("Crawling")

CRAWLER

An **INTELLIGENT CRAWLER** only browses the **LINKS** with **INDICATIONS OF B2C E-COMMERCE**

After clicking the button **Explorar Webs** (Browse Websites), the application starts a process of browsing the Internet, data scraping from the websites of businesses and capturing any information that may be relevant for detecting B2C e-commerce. This process is often called crawling, and the software that performs it is a crawler (other equivalent names are "web spider" or simply "robot"). The result of this process is a record of terms that are relevant for the analysis of each of the visited websites.

This is the phase that takes the longest processing time. To reduce it, the crawler uses an intelligent link selection mechanism, which avoids the need to browse all the pages of a company's website, selecting only those links that are likely to contain relevant information for detecting B2C e-commerce.

Despite this, the browsing time for a large database can be in the order of several hours or even days. Fortunately, it is generally necessary to perform the data scraping operation only once for each project.

3.2.2 PHASE 2: Labelling

When the browsing stage is complete, the next step is to label the websites, which is performed by clicking the button entitled **Etiquetar** (Labelling). The ML algorithms are based on the principle of learning by example, and therefore need some websites to be previously labelled as containing or not containing evidence of B2C e-commerce. After clicking on the Labelling button, the application shows the user a company website in the default browser and asks about the presence of B2C on the site.

Figure 3. Illustration of Manual Labelling on Mac OS X

The user can freely browse the website in order to determine whether or not it has B2C activity. The user must finally choose between the three following options:

1. **YES**: The website contains B2C activity
2. **NO**: The website does not contain B2C activity
3. **Error**: The user could not determine whether the website does or does not contain B2C activity.

The third option may be useful in certain circumstances: for example, when the browser is unable to access the web page, when the company has temporarily shut down the website for maintenance or upgrades, or when it could not be determined whether there is or is not B2C activity because the content is only accessible to registered users.

LABELLING

An ***ACTIVE LEARNING*** algorithm enables the minimisation of manual labelling of websites.

The labelling session is terminated when the user closes the window. It is now that the application saves the labels entered and is ready to apply the ML algorithms.

Labelling can be performed over several sessions. Each session can be interrupted at any time and can be restarted afterwards.

The labelling process is the only process that requires user intervention before the classification of websites and must therefore be performed efficiently. The application uses an intelligent Active Learning mechanism so that the choice of pages to be labelled is not random; only those websites that are likely to be the most informative for automated learning are shown to the user.

3.2.3 PHASE 3: Detection of B2C activity

When the labelling is complete, the ML algorithm can be activated for classification by clicking on the button **Detectar** [Detection]. This algorithm analyses all the labelled websites to identify patterns or features that differentiate sites that have B2C activity from those that do not.

Following the automated learning process, the B2C detector processes all the relevant information captured by the crawling process and determines a **B2C score** for each business. This value is the number of points assigned by the B2C detector to each business before taking the final decision on whether it has or does not have B2C activity. A high score, close to 1, indicates high evidence that the business has B2C activity, while a low score, close to -1, indicates the reverse: evidence that the website does not have B2C activity.

The final decision of the B2C detector is the result of applying a threshold to all the scores:

- Enterprises with B2C score higher than the threshold are assigned to the class, "**with B2C activity**"
- Enterprises with a B2C score lower than the threshold are assigned to the class, "**without B2C activity**"

DETECTOR

For each business, determines a DECISION on the presence or absence of B2C activity, and a RELIABILITY value for this decision

a. Measurement of Detector Performance

Although the aim of the B2C detector is a perfect classification, the presence of classification errors is inevitable. The efficiency of the detector is measured by two key parameters:

- **TPR** (True Positives Rate): the proportion of businesses with B2C activity that have been correctly detected.
- **FPR** (False Positives Rate): the proportion of businesses without B2C activity that have been incorrectly assigned to the class "with B2C".

These parameters can only be estimated by comparing the detector decision with the labels obtained manually, but these are often reliable indicators of the expected performance on the whole set of businesses.

An ideal classifier would be characterised by the values of TPR=1 and FPR=0 but this result is not achievable in practice. A good detector would try to maximise TPR at the same time as minimising FPR.

The selection of the detection threshold for B2C activity detection has a direct impact on TPR and FPR. In general:

- High values for the threshold tend to reduce FPR, but at the cost of also reducing TPR.
- Low values for the threshold tend to increase TPR, but at the cost of also increasing FPR.

The choice of threshold value establishes a compromise between a high proportion of businesses with detected B2C activity and a low rate of false positives. Given that this compromise can depend on the aims of the detector, the B2C detector pays special attention to four characteristic threshold values:

- **BEP** (Break Even Point): the threshold value for which the rates of false positives (FPR) and false negatives (businesses with B2C activity classified as "without B2C") are equal.
- **TPR**=0.95. Threshold value that guarantees 95% true positives.
- **FPR**=0.05. Threshold value for which the rate of false negatives is 5%.
- **FP**=**FN**. Threshold value for which the total (absolute) number of false positives and false negatives are equal.

The difference between BEP and the FP=FN thresholds is important and especially significant when the proportion of positive examples is much lower than that of negative examples. BEP is the point at which the proportion of false positives (evaluated with respect to the total of businesses WITHOUT B2C activity) coincides with the proportion of false negatives (evaluated with respect to the total of all businesses WITH B2C activity). However, given that the set of businesses WITHOUT B2C activity (in the database analysed) is much higher than the number of businesses WITH B2C activity, the number of false positives at the BEP is very much higher than that of false negatives.

By contrast, the FP=FN threshold balances the total numbers of false positives and false negatives. The search for this equilibrium is important when trying to obtain aggregate measures of the detector. For example,

when trying to estimate the number of Spanish businesses that engage in B2C commerce, the estimate based on B2C activity detections with the FP=FN threshold will be more reliable than an estimate based on the BEP threshold (because the errors of both types mutually compensate in the calculation).

Finally, the detector determines a confidence measure on the decision taken, which is calculated as follows:

Confidence = 1 - |B2C decision - B2C estimate|/2

Thus for example:

- When the B2C activity estimate = -0.998 and B2C decision = -1, Confidence = 0.999: the detector is very certain that it is not mistaken.

- When B2C activity estimate = 0.1 and B2C decision = 1, Confidence = 0.55: the detector takes a decision but it is not very reliable.

b. Results of Detection

As a result of classifying all businesses, two files of results are produced:

1. **File of classification results**: this file contains a table in CSV format (semicolon separated variables) combining the classification results with the information on the businesses contained in the data file input to the application. Specifically, the table contains the fields indicated below for each company:

    - Data taken from the input files:

        - **Company name**
        - **NIF code (company tax ID)**
        - **Primary CNAE code (activity classification)**
        - **Net sales** (EUR '000)
        - **Materials** (EUR '000)
        - **Staff costs** (EUR '000)
        - **Other operating costs** (EUR '000)
        - **Property depreciation allowance**. (EUR '000)
        - **Tangible assets** (EUR '000)
        - **Intangible assets** (EUR '000)
        - **Staff numbers**
        - **Web address**

    - Data obtained after B2C activity detection

        - **Manual label** (-1 = "no", 1 = "yes", 0 = "no label"). Indicator of the presence or absence of B2C activity obtained during the labelling process.
        - **Website available** (1 = "website downloaded", 0 = "not downloaded"). Indicates businesses excluded from the analysis because their website could not be downloaded.
        - **B2C estimate** (value between -1 and 1).

- **B2C in BEP** (-1 = "no", 1 = "yes"). Detector decision when the BEP threshold is applied.
- **BEP Confidence** (value between 0 and 1)
- **B2C in TPR**=0.95 (-1= "no", 1 = "yes"). Detector decision when TPR=0.95 threshold is applied.
- **TPR confidence**=0.95 (value between 0 and 1)
- **B2C in FPR**=0.05 (-1 = "no", 1 = "yes"). Detector decision when TPR=0.05 threshold is applied.
- **FPR confidence**=0.05 (value between 0 and 1)
- **B2C in FP=FN** (-1= "no", 1 = "yes"). Detector decision when the FP=TP threshold is applied.
- **FP=FN confidence** (value between 0 and 1)

2. **File of descriptors**: in addition to B2C activity detection, the classifier explores the website for the presence of terms or words that can be useful for subsequent analysis regarding the use of trust certificates, secure websites, accepted means of payment, etc.

   Each line of this file has the following format:

   NIF code; Descriptor

   where "NIF code" is the company tax identifier and "Descriptor" is the description that was found on the company website, and is one of the following:

   - Cestacompra [shopping cart]
   - Enviodomicilio [home delivery]
   - Aenor [Spanish Association for Standardisation and Certification]
   - Agace
   - Americanexpress
   - Aptice [National Association of Internet businesses]
   - Confianzaonline [online security certificate]
   - Devoluciones [repayments]
   - Email
   - Facebook
   - Geotrust
   - Iqua [Internet Quality Agency]
   - Mastercard
   - Mcafee
   - Norton
   - Optimaweb
   - PayPal
   - Qweb
   - Reembolso [refund]
   - RSS
   - Safetypay
   - Thawte [global SSL certification authority]
   - Truste
   - Trustedshops
   - Twitter
   - Verificada [verified]
   - Verisign
   - Visa

If several of these descriptors appear on the company website, the file will contain several records with the same NIF, one for every descriptor found.

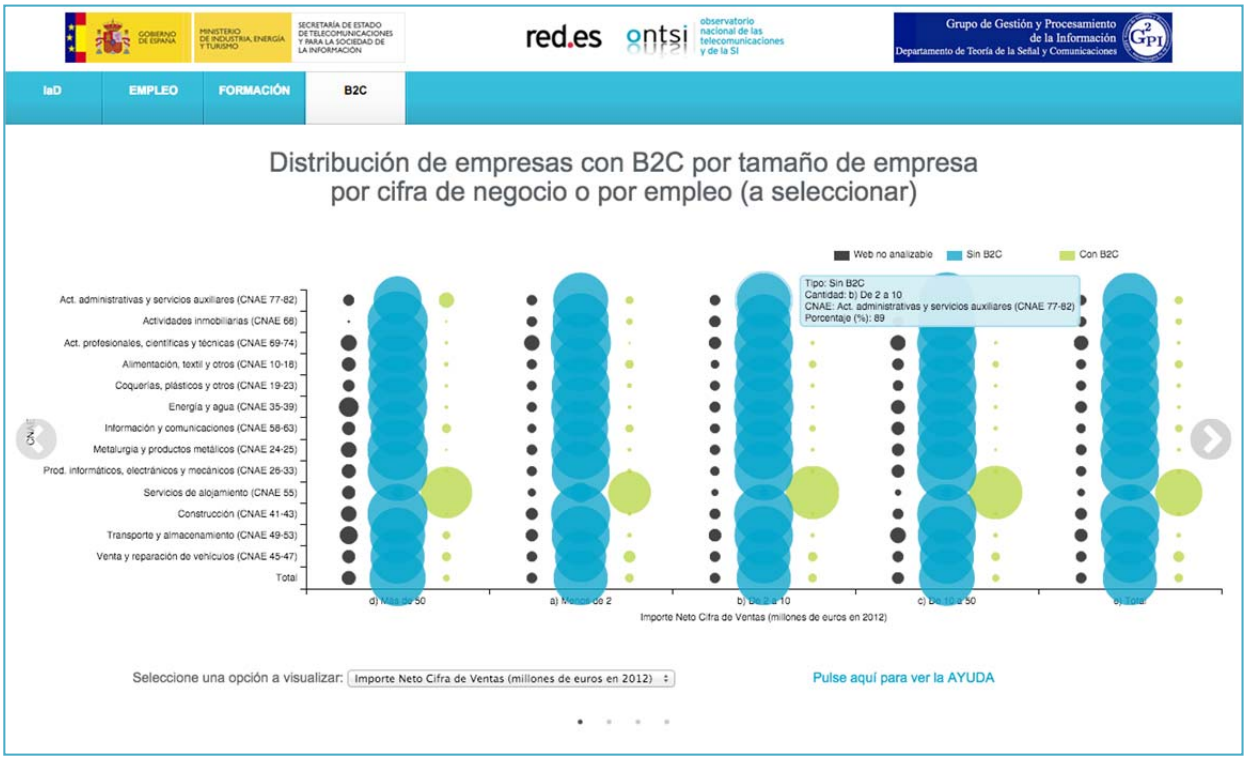
3.2.4 PHASE 4: Visualisation

Clicking on the button **Visualizar** (View) in the application leads to a view of the results of the analysis in the web browser. This is independent of the application, so that the visualisation can be accessed from any browser by entering the URL.

The visualisation website consists of four pages that show different views of the results of the analysis. These pages shown the classification results obtained at the operating point **FP**=**FN** explained in the previous section.

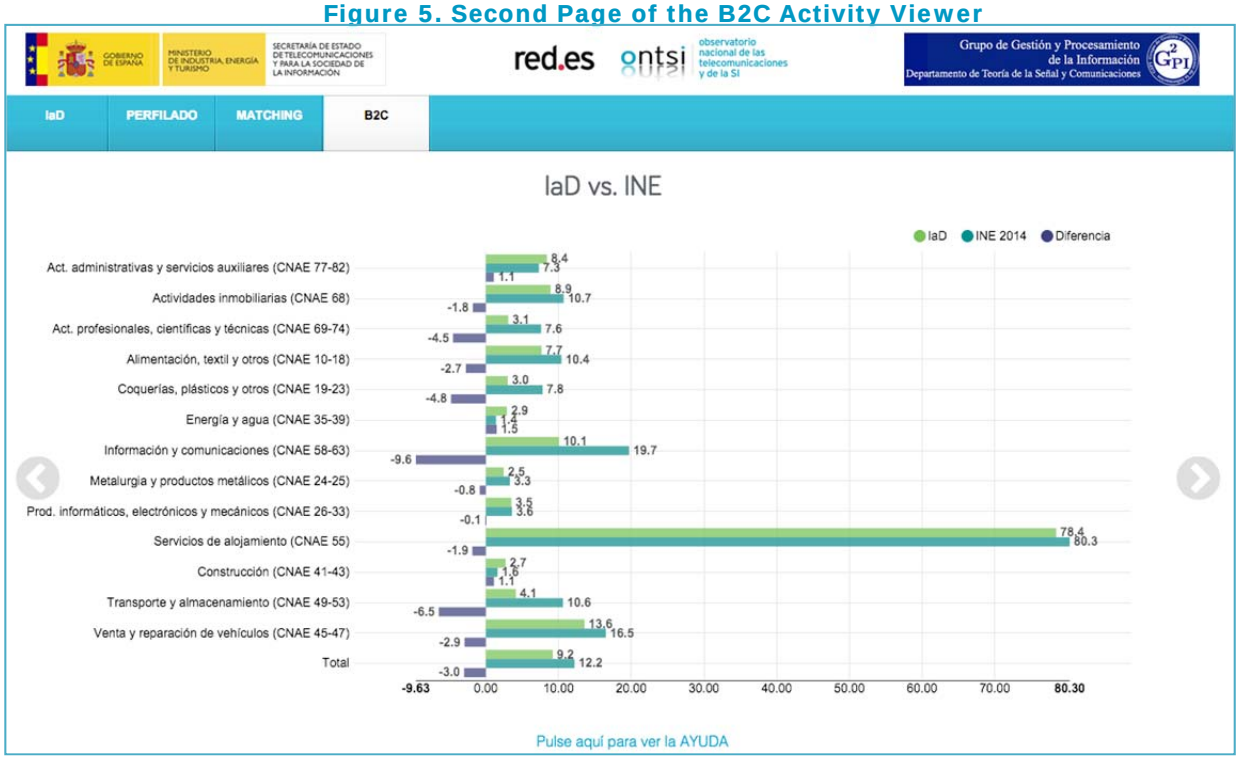
The first page shows the results of the B2C activity detection, broken down by sectors on the vertical axis and broken down on the horizontal axis by a variable that can be selected via the menu (e.g., the size of the company), as shown in the following Figure. Each circle has a size proportional to the number of corresponding websites, in blue if they do not have B2C activity, in green if they have B2C activity and in black if they have been excluded from the analysis.

Figure 4. First Page of the B2C Activity Viewer



The second page compares the results of the B2C detector with those obtained from other sources from the Spanish National Institute for Statistics (INE). Each row shows, for each CNAE sector, the percentage of businesses with B2C activity as reported by the B2C detector compared to the INE data. The bars in dark blue show the difference.

Figure 5. Second Page of the B2C Activity Viewer

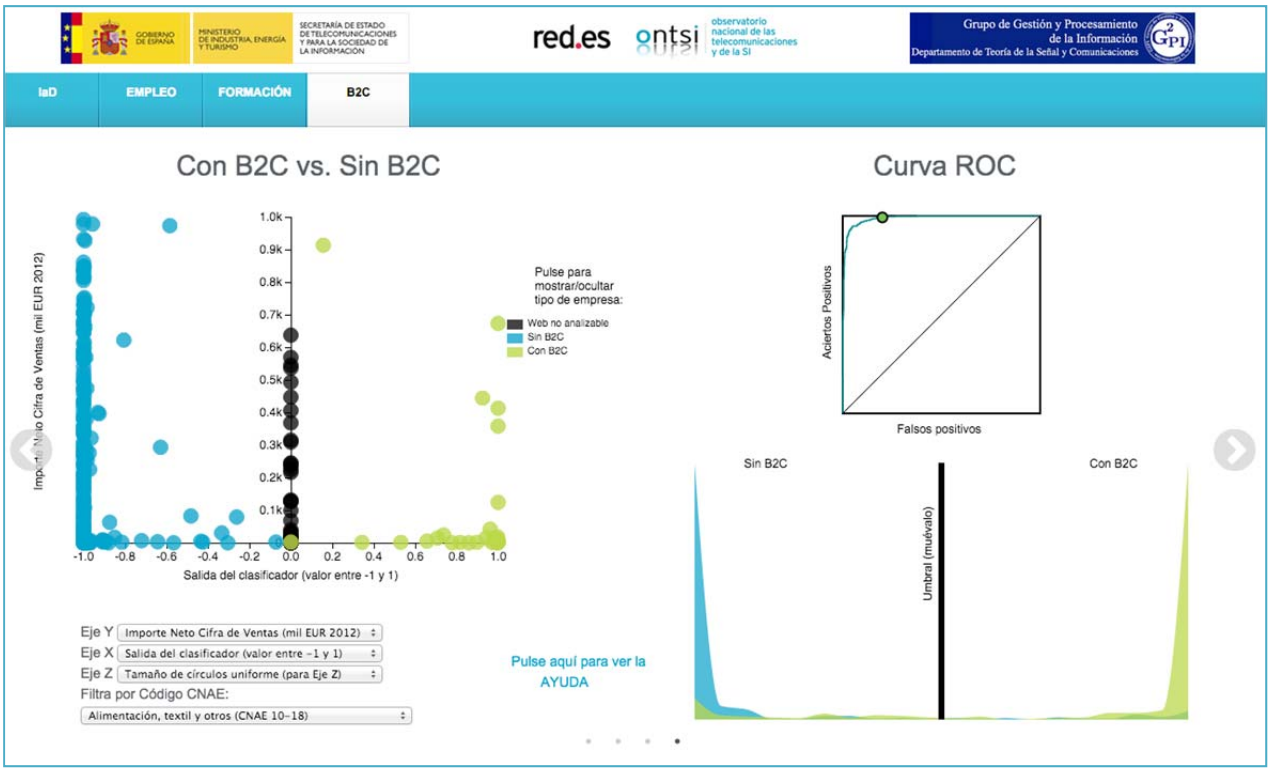


The third page is a visual representation of the indicators obtained in the file of descriptors. The size of each circle is proportional to the number of pages in which the presence of each indicator was detected.

Finally, the fourth page shows a view of a random sample of all websites analysed and the results of their classification. Each point on the figure on the left represents a website, in blue if it has been classified "without B2C activity", in green if it has been classified "with B2C activity" and in black if it has been labelled as "error". The horizontal axis, the vertical axis and the size of each point is associated with three magnitudes that can be selected by the user via a menu.

Figure 6. Third Page of the B2C Activity Viewer

Figure 7. Fourth Page of the B2C Activity Viewer



The left half of the page entitled "With B2C activity vs. Without B2C activity".- This view graphically illustrates the use of e-commerce (B2C commerce) in a specific sector - according to the Spanish national classification of economic activities (CNAE)-. This use of B2C commerce can be analysed interactively according to different characteristics (provided by Red.es) that describe businesses registered for each sector. These characteristics are:

- Net sales
- Fixed assets
- Number of employees in 2012
- B2C estimate:
- Confidence

The right half of the page entitled "ROC Curve", shows two different figures: the ROC (Receiver Operating Characteristic) curve (above), which enables evaluation of the quality of the best classifier used, and the distributions of businesses that engage or do not engage in B2C commerce as a function of the confidence in the decision taken by this classifier for each of the businesses (below). The ROC curve is nothing other than a representation of the variation of TPR (vertical axis) and FPR (horizontal axis) by the effect of the variation of the detector threshold. The vertical bar in black represents the threshold, which can be moved with the mouse pointer, thereby displacing the point of operation reflected in the ROC. Depending on the position of the threshold (moveable) between these two distributions, the decision taken for the various companies is reflected in the cloud of dots on the left of the page.

In order to ensure streamlined interaction with the cloud of dots, only a representative portion of businesses are shown, but they always respect the proportions of businesses with B2C activity compared to those without. The proportion of websites that were not accessed by the crawler for different reasons are also shown.

The process of labelling, detection and visualisation can be activated any number of times. After a first analysis, if it is considered useful to add more labels, click on the labelling button, generate new labels and restart the detection process with the expanded set of labels.

4

ANALYSIS OF ONLINE SALES IN SPANISH BUSINESSES

4 Analysis of Online Sales in Spanish Businesses

Having described the operation of a tool for detecting B2C commerce, an analysis performed on the data provided by Red.es is described below.

4.1 Data Source

In order to start the content analysis of Spanish company websites, a first listing of company data provided by Red.es was obtained. These data were contained in the following Excel files, segmented by CNAE codes:

- CNAEs 10 - 18.xls (food, textile and others)
- CNAEs 19 - 23.xls (Coal and oil refining, plastics and others)
- CNAEs 24 -25.xls (metallurgy metallic products)
- CNAEs 26 - 33.xls (computer, electronic and mechanical products)
- CNAEs 35 -39.xls (energy and water)
- CNAEs 41-43.xls (construction)
- CNAEs 45-47.xls (vehicle sales and repairing)
- CNAEs 49-53 1.xls (transport and storage)
- CNAEs 55-1.xls (accommodation services)
- CNAEs 58-63.xls (information and communications)
- CNAEs 68.xls (estate agencies)
- CNAEs 69-74.xls (professional, scientific and technical activities)
- CNAEs 77-82-2.xls (administrative activities and auxiliary services)

These files contain financial data of the businesses to be used in the analysis, the company NIF (tax code), used as a unique identifier for each company (we observed that businesses with different NIFs could have the same web address) and the website addresses themselves. The table of data contained in these files has the following column structure (this structure must be respected by data files to be used in the future to perform the same type of analysis):

1: Sequential number

2: Company name

4: NIF code

5: Primary CNAE code 2009

6: Net sales

7: Materials

8: Staff costs

9: Other operating costs

10: Provisions for depreciation of fixed assets

11: Tangible fixed assets

12: Intangible fixed assets

13: Number of employees

14: Web address

As a first step, these files were joined into a single listing in CSV ("Comma-Separated Values") format called b2c_2014_05.csv, which was the starting point for all subsequent processes.

Following an initial analysis of this file, a listing of non-repeating website addresses was obtained. This was the starting point for the website exploration module (crawling) and was called b2c_2014_05_urls.txt. A dictionary indexed by NIF was also obtained of other relevant data of the businesses to be analysed; this was to be used later in the results visualisation phase.

4.2 Purpose of the Study

The purpose of this study was to characterise the use of e-commerce on Spanish company websites using the automated detector based on ML algorithms. The results of the study were compared with other data sources.

As explained in the previous chapter, the automated detection process consisted of various stages:

1. Crawling and data scraping of the content of enterprise web sites, extracting from them the features that are relevant for detecting B2C activity.
2. Manual labelling of a fraction of the pages
3. Automatically detecting B2C activity, based on the application of ML algorithms
4. Analysing performance and visualising the results

DATA SOURCE

162,849

COMPANIES,

145,920

different WEB DOMAINS

4.3 Browsing Websites

4.3.1 Crawling

The data files contained a total of 170,620 entries, each corresponding to a company and its CNAE code. The following must be taken into account:

- Some businesses appeared repeated in different listings, associated with different CNAE codes. Excluding these repetitions, the set was reduced to 162,849 businesses.
- Some businesses shared the same web domain. As an example, there are three extreme cases of this situation:
  - 'mediamarkt.es': 56 companies
  - 'ac-hotels.com': 50 companies
  - 'renault.es': 46 companies
- Excluding repetitions, 145,920 unique websites were identified to be browsed by the crawler.
- This figure reduced to 122,925 websites (corresponding to 136,884 businesses) excluding cases in which the crawler could not download the content.

The last point implies that 25,965 businesses were discarded from classification because the crawler could not download the content. This could have been due to various reasons:

- There was no content on the website (in which case these websites could be classified as "without B2C activity")
- The website automatically redirected the crawler to another website outside the domain of the company. The crawler does not browse pages outside the domain given that it is not possible to identify (automatically) the company that the B2C activity, if any, belongs to.
- The website content is in flash or javascript format, that cannot be analysed by the crawler.

Given that the absence of downloaded content does not necessarily imply absence of B2C activity, these businesses were discarded from the analysis.

BROWSING WEBSITES

The **CRAWLER** obtained a representation of each website based on

8,763,024

TERMS, that were reduced to

343,780

4.3.2 BoW analysis

It should be highlighted that after the process of browsing, the content of each website was not downloaded and stored permanently on the computer. The crawler coded the information of each website in the form of Bags of Words (BoW): a list of terms (words and other character strings) that appear in the website, accompanied by a measure of relevance. This measure of relevance is proportional to the number of times that the term appears in the document and is inversely proportional to its abundance (so that terms that are abundant in the language, such as propositions and articles, are generally not very relevant). The abundance of the term in the language was estimated from the frequency of appearance in all web pages analysed.

After browsing the database, a total of 8,763,024 different terms were identified. Obviously, this number far exceeds the lengths of common dictionaries in any language. The proliferation of terms could be due to several reasons: the existence of content in other languages, terms that do not belong to the language but were indexed as words, etc. Therefore it was necessary to perform a prior cleaning of the list. A first selection was based on the frequency of appearance in the web pages. An initial criterion was established for removing words that did not appear in at least 10 websites. A word that appears in very few websites is not informative with respect to the problem of classifying the other pages. After carrying out this process, the number of features retained was 343,780, a much more manageable figure, although still too high. It was therefore necessary to proceed to an additional stage of feature selection.

FEATURE EXTRACTION

Identifies the

10,000

most important **TERMS** for detecting **B2C** activity.

4.3.3 Feature Extraction

Note that the number of features was higher than the number of documents available, and much higher than the number of documents labelled (some 2,260). In these conditions, most of the classifiers were difficult to handle because of what is known as "the curse of dimensionality".

The feature extraction algorithm must select from the whole subset of 343,780 those terms that are most relevant for detecting B2C activity.

After a comparative analysis of different algorithms for extracting features (which are described in the following chapter), an algorithm entitled WeightedTfIdf was applied to select 10,000 features relevant for detecting B2C activity.

4.4 Labelling

4.4.1 Labelling Criteria

LABELLING

2,540 **WEB PAGES LABELLED.**

Labelling is a key aspect in the classification system based on ML. Automated learning is based on the information provided by the labelled samples, so that errors or inaccuracies in the labelling process directly affect detector performance.

Firstly, a clear definition of the category to be detected is required. The study considered that a website offered e-commerce when it met the two following conditions:

- **Condition 1**: It offered a variety of products, products could be added to a shopping basket and then it was possible to pay for and complete the order.

- **Condition 2**: It offered the option of booking hotel rooms, theatre and travel tickets and it was possible to pay or otherwise signal the operation.

This definition implies that, during annotation, websites such as the following were labelled as "without B2C activity":

- Businesses that enable browsing a product catalogue but do not offer the possibility of selecting products for purchasing; the website may offer only a contact address for sending messages.
- Businesses that enable browsing a catalogue and requesting one or several products over the web, but it is the company that later, possibly by other methods, gets in touch with the customer.
- Businesses that enable selecting products in a shopping basket, but payment is not electronic and needs to be made by cheque or bank transfer outside the website domain.

Some cases were also identified of businesses that had a catalogue, products, shopping basket and even information about online payment, but after selecting a product, the customer received a message indicating that the product was not available for electronic purchase. Cases such as this were also considered in the category "without B2C activity".

Manual labelling was also limited by access to the content. On some websites, access to part of the content requires prior registration, which generally requires personal and contact information to be entered. The restricted access part usually allows the formalisation of the purchase, and

therefore it is not possible to determine the means of payment, which prevents the website from being categorised.

During the annotation process, personal data were not entered to access restricted areas. The labelling strategy for websites with limited access areas was based on indicators: if the freely accessible area contained sufficient indicators of presence or absence of B2C activity, it was labelled as such. When access limitations to the website prevented a reliable evaluation, the website was labelled as erroneous or was omitted from labelling.

4.4.2 Labelling Errors

LABELLING

There were

1,606

websites labelled **WITHOUT B2C activity**, and

776

websites labelled **WITH B2C activity**

A precise definition of B2C activity does not guarantee perfect labelling. Various sources of error can give rise to samples assigned to an incorrect category:

- The indicators of presence or absence of B2C activity on websites with limited access can be wrongly interpreted.
- Time delay: the moment of time when the crawler analyses the website and the time when the page is labelled may differ by days or weeks. During this time, changes may have occurred on the analysed website. An example of this is when a website is temporarily out of service for reasons of maintenance or update. If the website did not offer B2C commerce at the time of labelling, it was labelled as "without B2C activity". However, it is possible that the website did offer B2C commerce at the time when the crawling processes occurred.
- Discrepancies between individuals performing the labelling: labelling was carried out by 5 different people. For the more difficult to discern cases, it is possible that the criteria followed by the labellers was not completely homogeneous.

The number of websites labelled for this project was 2,540. Of the 2,540 labels, there were 776 positives and 1,606 negatives, totalling 2,382. The remaining 158 labels were not used, either because the label was an "error" or because the website did not download (not available at the time of crawling) or had few words.

The statistics were calculated counting businesses with B2C activity compared to the total that could be classified.

4.5 Classification

The classification process requires two decisions to be taken: (1) choosing a classification algorithm, and (2) determining a point of operation of the classifier. After a comparative analysis of different classification algorithms, which are detailed in the following chapter, we chose a classification method called logistic regression.

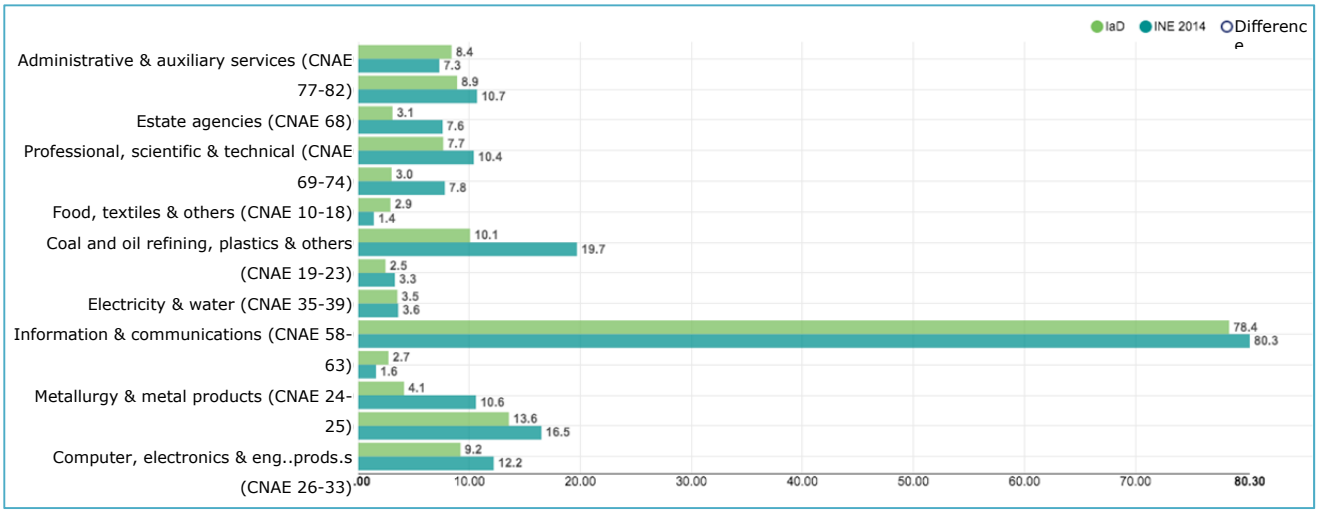
The second important aspect is the choice of the point of operation of the classifier. This point of operation is determined by the classifier threshold. When the value of the B2C activity estimation of a website is higher than this threshold, the website is considered to have B2C activity, otherwise it is considered not to have B2C activity.

Given that one of the main aims of the study consisted in obtaining reliable measures of the level of implementation of e-commerce in Spanish businesses, the threshold was chosen so that the number of false positives was equal to the number of false negatives, so that when obtaining aggregated measures, both types of error tended to balance out.

4.6 Results of the Study

In order to analyse the results provided by the B2C activity detection module, we firstly compared the percentages of the presence of e-commerce by CNAE sectors as reported by the module against other data available from the INE. Only those sectors are shown where data are available from the INE. Where such data were not available, only the results of B2C activity detection are shown. The percentages of estimated B2C activity by sector are shown in the figure below.

Figure 8. Comparison of ML Results Against INE Data. Percentage of businesses selling through their websites



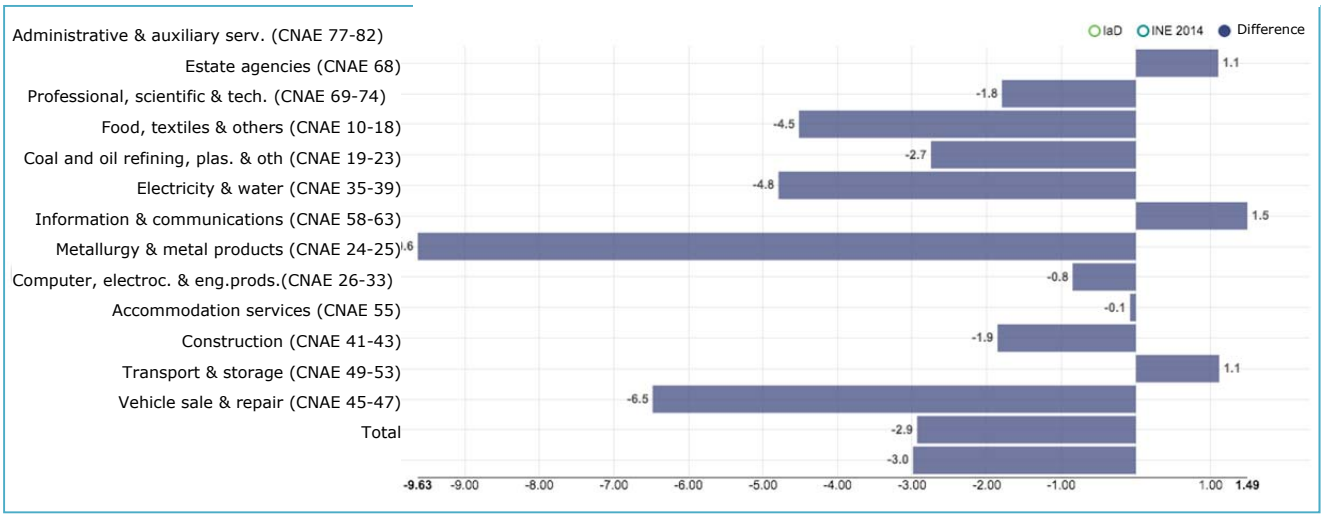
In general, there is a good match between both estimates. The following figure shows the differences between them. The largest discrepancies are seen in "*Information and Communications*", "*Transport and Storage*" and "*Coal and oil refining, plastics and others*", and these could be due to multiple factors:

- Classification errors
- Differences between the ML labelling criteria and those used by INE
- Sample differences (ML errors were calculated on the proportion of websites that could be downloaded)
- INE estimate errors (which follows a process of a questionnaire to a subset of the population of businesses)

4.6.1 Classification Errors

One of the advantages in the use of an automatic detector is that is enables automated identification of problems in labelling or parameterisation; it is possible to easily isolate the training patterns that show a large discrepancy between labelling and to identify two sources of problems:

Figure 9. Difference Between ML Results and INE Data



- In the case of poor manual labelling of a website, it would be possible to correct the label, which could be the result of human labelling error.
- Bad parameterisation of the content of a website: in this case it would be possible to improve the information extraction mechanism so that a new parameterisation would correctly estimate the presence or absence of e-commerce.

As an example of the above, the current classifier shows the following list of "outliers" or abnormal points at both ends:

Label	Score	URL
-1	0.9999999	grandholidaysclub_com
-1	0.9999999	pasitoblanco.com
-1	0.9999996	schmidt-cocinas.com
-1	0.9999583	seguridadmar.com
-1	0.9999342	unidadeditorial.com
-1	0.9998394	sanfrio.com

-1	0.9997789	contrasena.com
-1	0.9996426	azulenapiscinas.com
-1	0.9995962	fourllopis.es
-1	0.9981084	diariojaen.es
-1	0.9968564	grupoi.es
-1	0.9942898	kines.es
-1	0.9942494	gtvcomunicaciones.com
-1	0.9921998	jamonesdonfinardo.com
-1	0.9898554	tecniclima.com
-1	0.9896447	grupoifg.com
-1	0.9889636	destileriasaguilar.com
-1	0.9881324	lanmovil.es
-1	0.9839426	cristaleria-bailon.com
-1	0.9807287	rucecan.com
-1	0.9801564	increasearch.com
-1	0.9747041	aedgency.com
-1	0.9670714	intsa.com
1	-0.9192017	lexnova.es
1	-0.9528232	alared.com
1	-0.9629765	pitneybowes.es
1	-0.9777928	ni2tek.com
1	-0.9813653	supercompdigital.com
1	-0.9864429	iustel.com
1	-0.9887289	duramas.com
1	-0.9902629	dchoteles.net
1	-0.9907376	tginformatica.com
1	-0.9943648	estuseguridad.com
1	-0.9972894	camelibrosinfantiles.es
1	-0.9978397	blauhotels.com
1	-0.9982150	nominalia.com
1	-0.9989692	manolihotels.com

1	-0.9993627	artek.es
1	-0.9995814	wmega.es
1	-0.9999721	puntdoc.es

The first value in this list indicates the label assigned by the human labeller, the second value is the estimation by the e-commerce detector (score) and the third value is the corresponding URL. Brief comments on some of these cases are given below:

- Websites that do not contain content in Spanish, so it is likely that the detector will fail (grandholidaysclub.com)
- Websites that have an e-shop but have an unusual path to get to it (puntdoc.es, lexnova.es)
- Websites that have an e-shop but it is located on another domain (wmega.es)
- Websites that were perhaps incorrectly labelled (wmega.es, puntdoc.es, artek.es)
- Websites that are heavily javascript-based and the crawler cannot analyse them correctly (manolihotels.com, blauhotels.com)
- Websites with an unusual profile, for example the sale of domains (nominalia.com, alared.com)
- Websites that changed their content or disappeared between downloading and labelling, and therefore show an inconsistency (tginformatica.com, ni2tek.com, pitneybowes.es)
- Websites that ask for the username and password at some point, and so the crawler cannot access them (supercompdigital.com, duramas.com)

Instances of doubtful labelling were also identified, which could be the result of erroneous labelling or operations, specifically some websites:

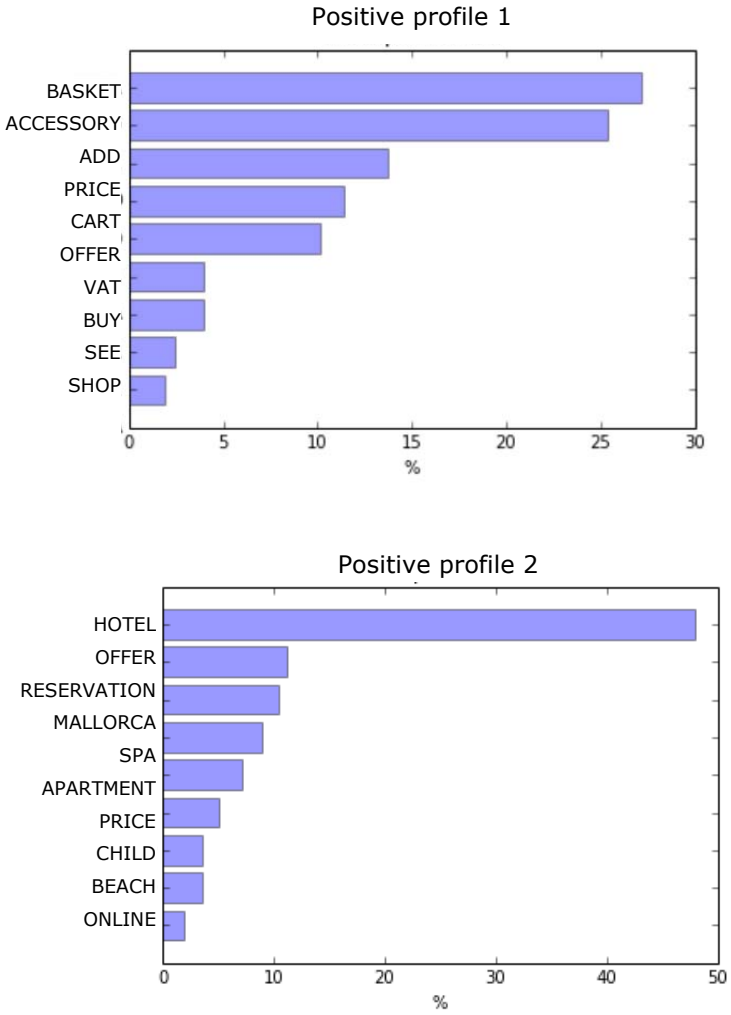
- Sites that indicate the presence of key words, so their code seems to indicate availability of e-commerce, but which in the end did not have it (www.clevisa.com)
- Sites that only allow access to customers, so that a web crawler can never access these contents (www.megasur.es)
- Sites that display e-shop advertising, but the shop either does not belong to the same domain or corresponds to an offer by third parties (www.diariodeibiza.es)
- Sites that contain links to e-shops that redirect the user to other e-shops:
    - www.cinesa.es: redirects to ticketmaster
    - www.desktop.com: redirects to Amazon

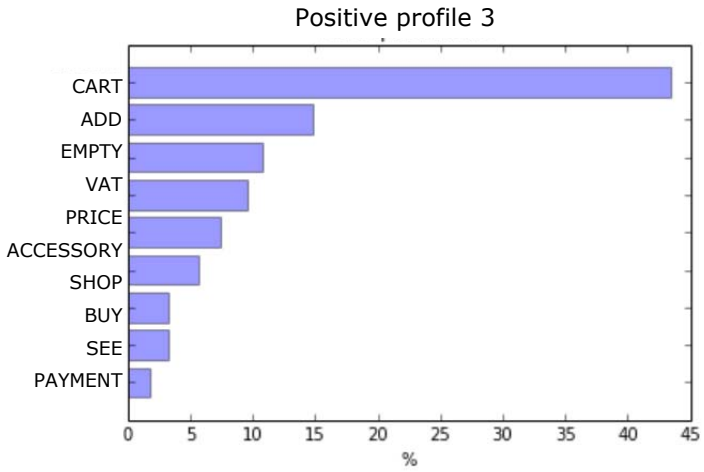
- The order cannot be completed directly on the website, but it is necessary to place an order via e-mail and to complete the payment (www.azuritasystem.com)

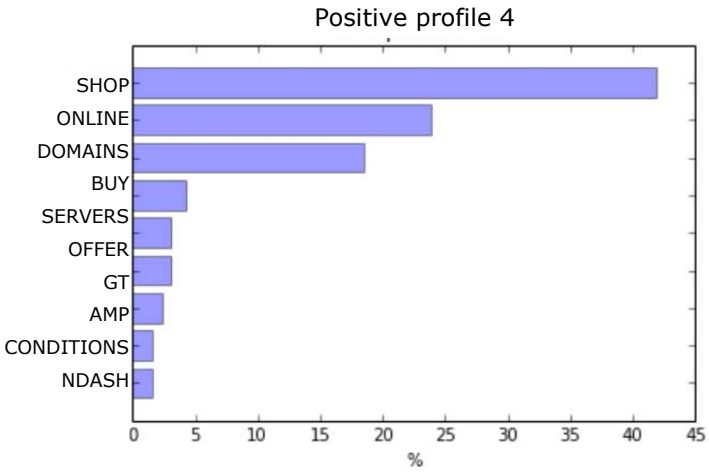
4.6.2 Profiles

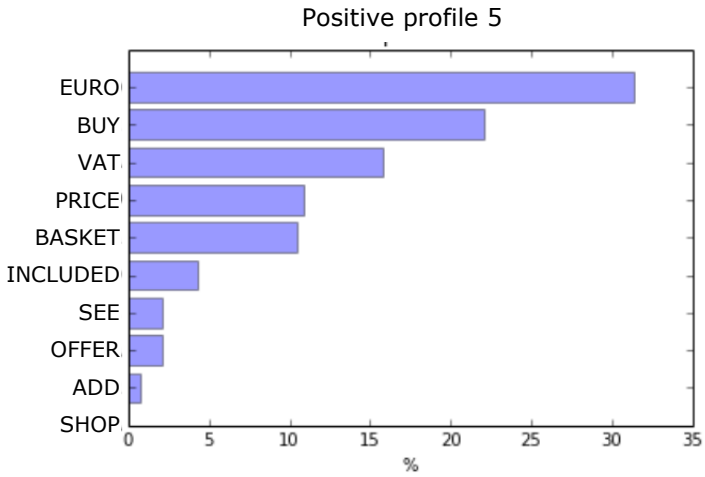
It is interesting to analyse characteristic patterns of websites with and without B2C activity. Figure 10 shows the 5 most characteristic profiles of company websites where B2C activity was detected. Each profile includes a list of terms or words most relevant to each profile, together with a bar where the length in logarithmic scale reflects the importance of each term in the profile.

Figure 10. Characteristic Profiles of Websites with B2C Activity




Positive profile 3
CART
ADD
EMPTY
VAT
PRICE
ACCESSORY
SHOP
BUY
SEE
PAYMENT
0 5 10 15 20 25 30 35 40 45
%


Positive profile 4
SHOP
ONLINE
DOMAINS
BUY
SERVERS
OFFER
GT
AMP
CONDITIONS
NDASH
0 5 10 15 20 25 30 35 40 45
%


Positive profile 5
EURO
BUY
VAT
PRICE
BASKET
INCLUDED
SEE
OFFER
ADD
SHOP
0 5 10 15 20 25 30 35
%

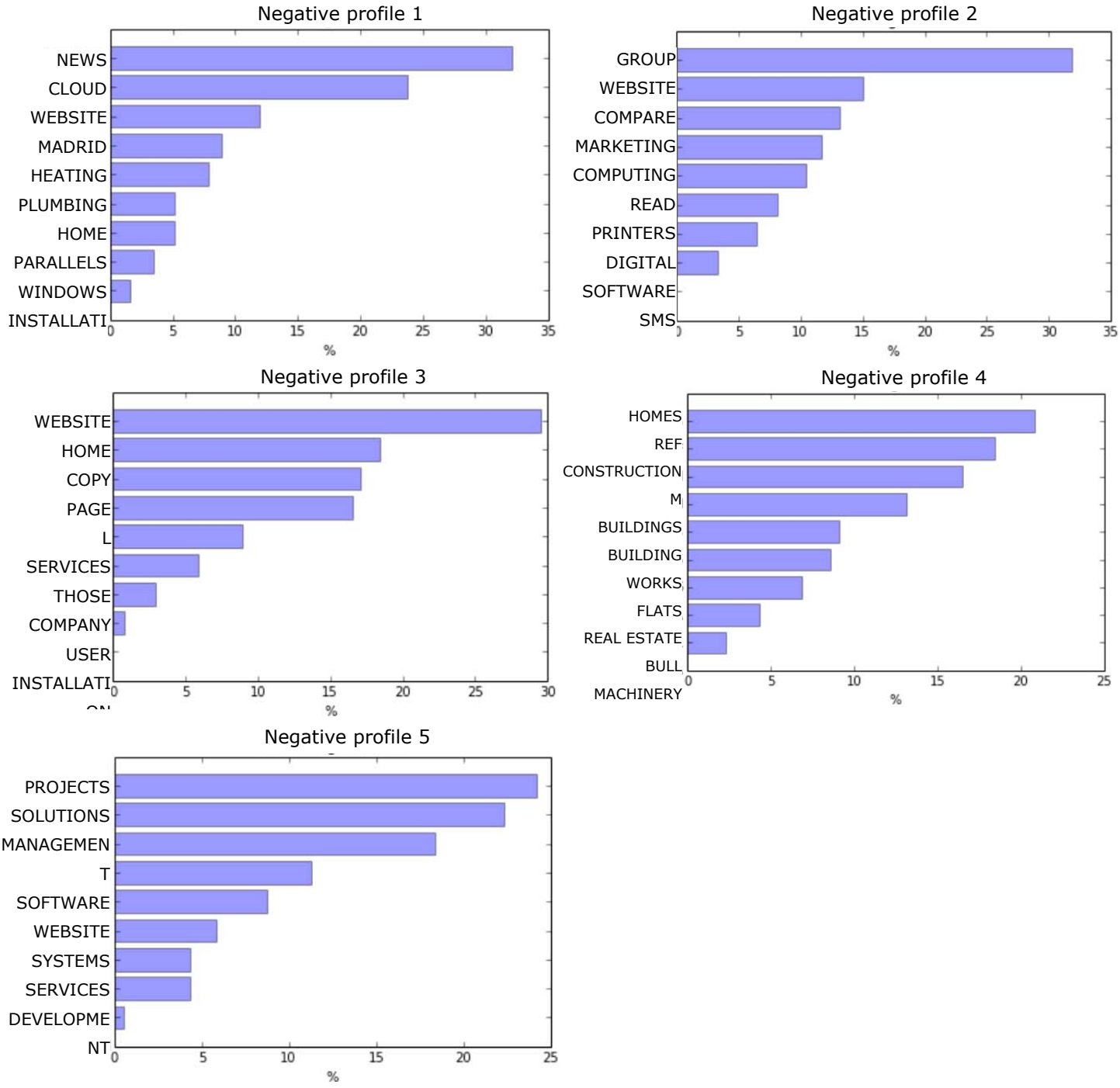
Profile 1 shows the presence of terms such as "basket", "add", "buy", "shop", very often in the buttons of electronic purchasing.

Profile 2 shows terms typical of websites related to the reservation of accommodation: "hotel", "reservation", "apartment", "price".

The other profiles show terms for purchasing over the Internet.

By contrast, Figure 11 shows characteristic profiles of websites without B2C activity. The terms associated with e-commerce have disappeared, in their place are terms associated with different economic sectors. The dispersion of "without B2C activity" profiles is greater.

Figure 11. Characteristic Profiles of Websites without B2C Activity



5

ANALYSIS OF THE VIABILITY OF ML FOR DETECTING B2C ACTIVITY

5 Analysis of the Results of Applying ML Techniques

The overall objective of the first milestone or subproject of this project was the evaluation of the viability of ML for automatically detecting B2C activity. Until now, such analyses were performed by conducting surveys. In addition to their high cost in time and money, these do not easily enable frequent periodic repetition in order to monitor change in e-commerce activity (for example, monthly, half-yearly, etc.) The alternative of manually visiting each and every link is also nonviable.

5.1 Efficiency of the Automated Classifier

CLASSIFIER

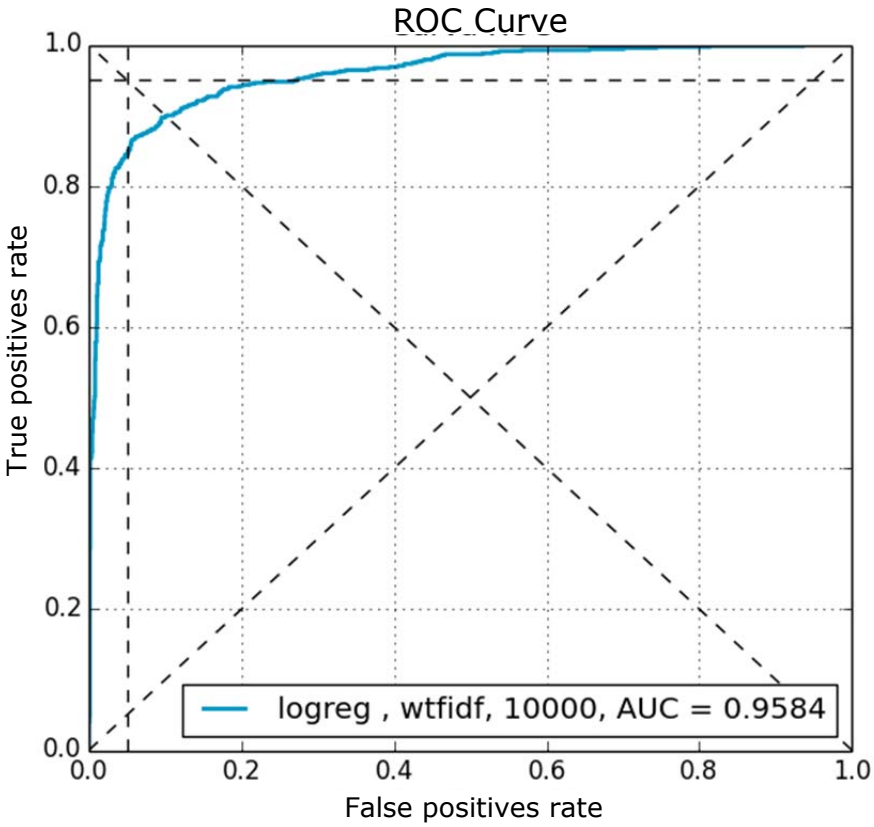
Uses **LOGISTIC REGRESSION** algorithms with the selection of features based on the **WEIGHTED TFIDF**

As noted above, we calculated the ROC curve for each of the possible and available combinations of classification models, feature selection techniques and number of features retained. The ROC curve is a graphical representation of the expected performance of a classifier, irrespective of the adopted decision threshold. The abscissa (X axis) shows the false positives rate (FPR); that is the fraction of training examples in which the classifier was in error, reporting an output as positive when it is really negative (obviously this value should be as close to zero as possible). The ordinate (Y axis) shows the true positives rate (TPR); that is the fraction of training examples in which the classifier was correct, reporting a positive output (obviously this value should be as close to 1 as possible). As the FPR and TPR values depend on the threshold chosen for taking the decision on the classifier output, the ROC curve is obtained by sweeping through the range of all possible thresholds; the ROC curve completely characterises the performance of the model.

The ROC curve of Figure 13 corresponds to the performance of the B2C activity detection method finally selected for the project and it has the following characteristics:

- Classification algorithm: logistic regression
- Algorithm for extracting features: WeightedTfIdf (based on selection by weights of the classifier).
- Number of features: 10,000.

Figure 12. Classifier Performance ROC Curve



From the representation of the curve, the threshold value that provides useful working points can be obtained, such as for example, the point of equilibrium or Break-Even Point (BEP). This point is where the values of 1-TPR and FPR are equal (that is, the false positives and false negatives rates are equal), and in the figure is the point where the diagonal line cuts the ROC curve. This working point is the most balanced, and was used for the rest of the analysis.

Other points that can serve as a reference for classifier consistency are the indicators shown on the previous figure as TPR=0.95 (where the horizontal discontinuous line cuts the ROC curve), that is, when 95% of the positive values are correct; and the indicator FPR=0.05 (where the vertical discontinuous line cuts the ROC curve), that is, when there are 5% false positives.

The visualization tool shows the results obtained at the operating point **FP**=**FN**, which has demonstrated to be the most adequate to the estimation of aggregated results (average presence of B2C in the whole set of companies, or per CNAE sector). This operating point corresponds to TPR = 86.8 % and FPR = 6.7 %, which implies an error rate around 7.3 % and an hit rate around 92.7 %, which is very close to the performance at the BEP.

In order to represent the effectiveness of a classifier using a single number (to order a collection of classifiers from good to poor, for example), the area under the ROC curve can be calculated (AUC, "Area Under Curve").

PRECISION

The B2C detector has a precision of

92%

of **SUCCESSES** (at the BEP) that is

8%

of **ERRORS**

The selected detection algorithm was the result of a process of exploration of different classification and feature extraction algorithms. The selection made was based on four criteria:

1. Detector performance
2. Processing time
3. Ease of installation
4. Interpretability of the results

The following section details other alternatives explored.

5.1.1 Other Methods for Detecting B2C Activity

In order to estimate the ROC curve for different detection methods, it is not acceptable to use the data that was used for training. In such a case, because the number of features used is very much higher than the number of examples labelled, there is a risk of obtaining ROC curves with 100% correctness in all such cases. But some cases were clearly estimated incorrectly and the resulting classifier would not operate correctly in cases not used in the training. In order to verify correct operation of the classifier, the ROC curve must be estimated using what is known as "Leave-one-out" (LOO) validation. This validation consists of using *N*-1 data (of the *N* available for the training) to set up the model, and to estimate the output on the example not used for the training. This procedure would provide one of the desired output values, so it must be repeated *N* times, leaving one different sample out from the training each time. When all the output estimates have been obtained as per the LOO procedure, the ROC curve can be calculated and displayed. This was performed to obtain the ROC curve of Figure 13 and also for analysing the goodness of fit of each of the possible combinations of parameters, the results of which are shown in Figure 14.

The curves of Figure 14 were obtained with a reduced set of some 1,500 labels, less than those used for obtaining Figure 13. It shows that the performances differed notably according to the method of estimating and selecting the features used. However, there are many curves with good performance where the value of AUC is greater than 0.96 (the ideal is AUC=1).

Three feature extraction methods were evaluated:

- SVM, known as "*fastfs*", because it is quite fast.
- Method based on "bagging" techniques, which combine the decisions of many classifiers in order to obtain the characteristics of high robustness, known as "baggedfs".
- Methods that use mutual information between inputs and labels: two of these were tried: "Joint Mutual Information" JMI (jmi) and Conditional Mutual Information Maximisation (CMIM), cmim.

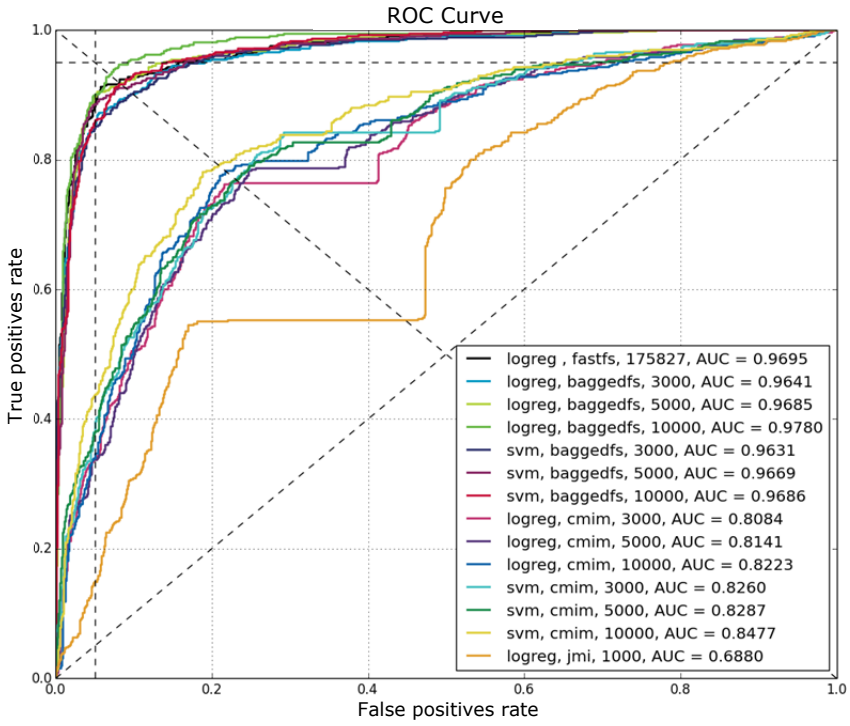
When the reduced sets of features were available, the chosen classifiers (logreg, svm and gp) were trained using 3,000, 5,000 and 10,000 features in the cases where this was computationally possible.

The best combination was obtained for the combination of a logistic regression classifier and the Bagging selection method when using 10,000 features (AUC = 0.978). It was found that the selection methods based on mutual information (CMIM, JMI) produced substantially poorer results on this problem.

CLASSIFIER

Other methods were also tested such as **SVM**, **GAUSSIAN PROCESSES** and **DECISION TREES**.

Figure 13. ROC Curves for Different Classifiers and Feature Extraction Methods



The classification module also included a "Gaussian process" (gp) classifier, as mention in the proposal, but the execution of this method was very costly (up to 5 days of execution time) for rather poor performance, so it was replaced by an alternative method, that of decision tree regression (dtr).

Observe that although from the point of view of classification performance, the best combination was obtained by a classifier with logistic regression and the bagging selection method, the final choice used a different selection method. The reason for this was that bagging selection was computationally very laborious (longer processing time), made the installation of the application software more complicated (because it requires the use of libraries that must be first compiled) and gives rise to a detector, the operation of which was difficult to interpret. The final selection, shown in Figure 13, avoided all these drawbacks with very similar performance.

5.2 Need for Labelling

Correct operation of the e-commerce detector was dependent on the quality of the data extracted from each website (features) and by the accuracy and number of labels used during the training process. It seems trivial to say that if it is not possible to extract suitable features from web pages, any method of automated classification will be unable to provide the expected results. This is, for example, the case of web pages based on flash or javascript, to which the crawling process has no access, and therefore the result is uncertain in these cases. Also, the performance of the classifier is intimately linked to the number of labels available at the time of training. It seems obvious that if 100% of the labels of the web pages were available, the success rate would be 100% (and any other processing would be unnecessary), but this implies an effort that is not directly feasible. The question to be answered here is, what performance can be obtained with a small fraction of websites labelled?

NEED FOR LABELLING

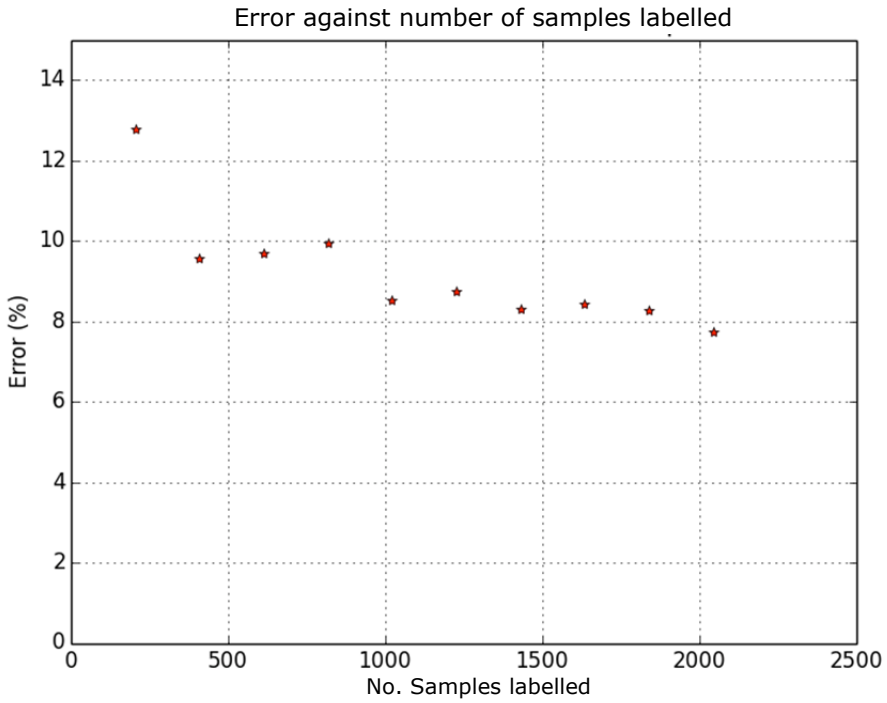
Some

1,500

LABELS are sufficient. Above this value, the error rate decreases slowly.

The following figure shows the change in the detection error rate as the number of labels increases. The error rate tends to stabilise after 1400 samples; this suggests that very significant gains are unlikely for an increased number of samples.

Figure 14. Error Rate as a Function of the Number of Labels



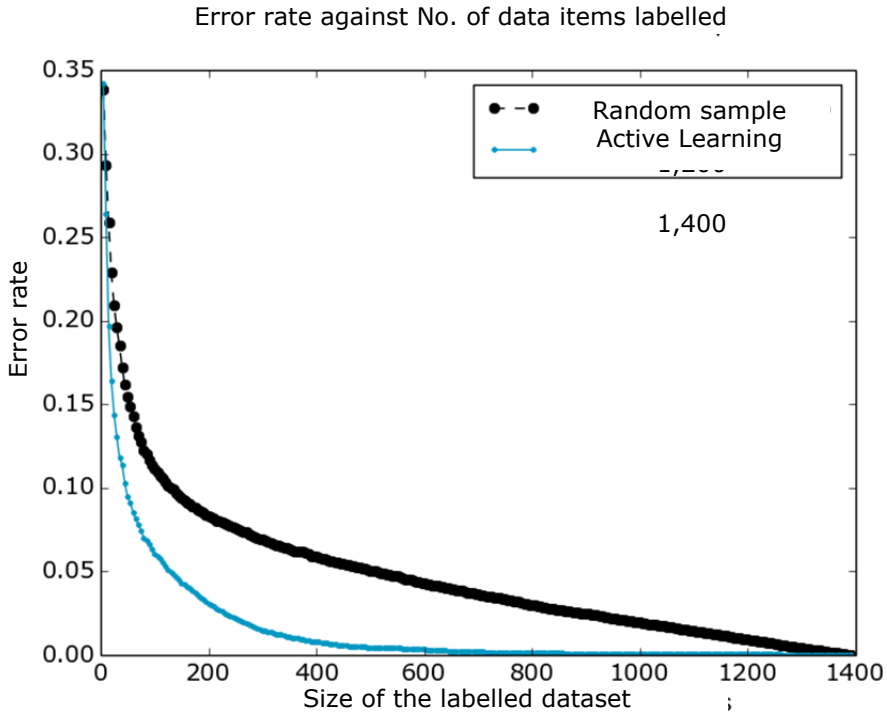
Another interesting question is whether it is possible to apply some intelligent selection method of the websites to be labelled (active learning), that would enable a further reduction in the number of web pages to be labelled and still obtain similar performance. This second question will be addressed in more detail in the next section.

5.3 Gain from Active Learning

In order to start the process, the active learning (**AL**) strategy was very useful, initially performed manually and then performed with the help of an AL tool developed for the purpose. Starting from a few initial labels, in both cases an indication was obtained of possible new websites that would be useful to visit and label for the purpose of improving model performance.

The following figure shows how the error rate decreased with the number of data items labelled. The tests were performed using logistic regression as the classifier and with vectors of 1,000 features. The curve in blue shows the change in error rate when applying an *AL* strategy, which falls toward a minimum much more quickly than when performing labelling on a random sample. Thus, almost the same performance was obtained with AL on 500 labels as was obtained with random sampling with 1,400 labels.

Figure 15. Error Rate With and Without Active Learning



This effect becomes more evident if this figure is represented directly, comparing labelling a random sample with labelling using AL. The curve in blue shows, for each value of the number of labels used with random sampling, the number of labels required by the AL algorithm to achieve the same error rate. It is clearly shown that the AL algorithm can reduce the need for labelling by some 35%.

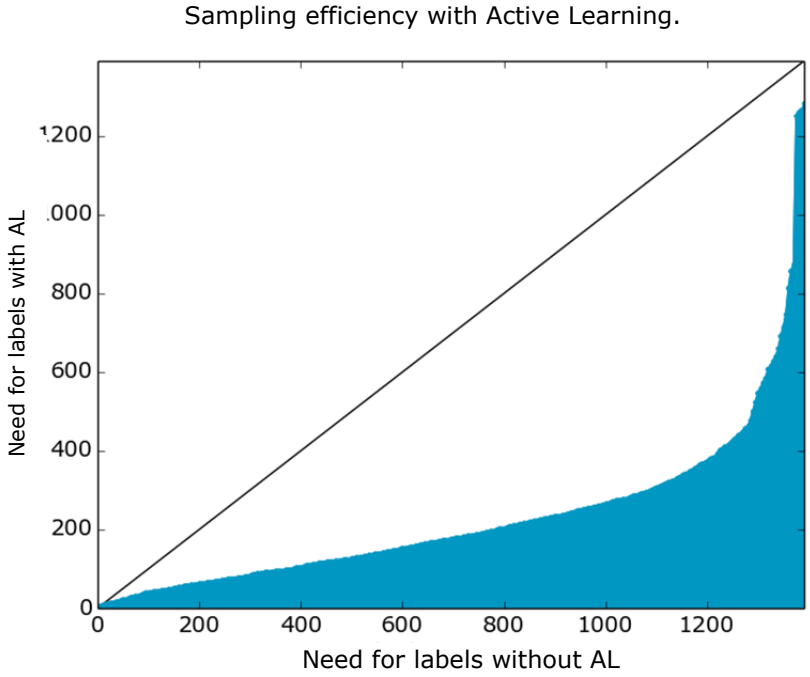
ACTIVE LEARNING

Can **REDUCE** the number of samples that must be labelled by

35%

compared to the number required by random sampling.

Figure 16. Sampling Efficiency with Active Learning



5.3.1 Conclusions on Labelling

This study showed that it is possible to apply ML techniques to automate the detection of the presence of e-commerce on the websites of Spanish businesses. Manual labelling of some 2,300 of a total of around 105,000 businesses (around 2.1% of all businesses) was sufficient to perform automated classification with a high level of precision. The use of active learning techniques can further significantly reduce the needs for labelling. Labelling 2,300 samples without AL could have been done with only 800 labels and still be sufficient (that is around 0.8% of all websites). Nevertheless, it should be taken into account that a large part of the labelling performed (approximately 50%) had already been performed using AL, so that in this case the figure should be somewhat higher; values around 1,000-1,200 are more realistic.

In general, the recommendation for labelling is as follows:

- Perform the labelling in several sessions, evaluating the performance every 200-300 samples.
- Observe the decrease in the error rate. When the error rate stabilises, the labelling can be stopped.

The results of the classification indicate the potential of ML techniques for classifying data (in this case websites) that were not used during the design

phase. Provided that the database used is of the same type (technically, that it comes from the same statistical source), the same classifier can be used.

5.3.2 Reuse of Data

The possibility of reducing the needs for labelling still further for future occasions, trying to make use of the information already extracted from current websites, should be explored particularly of all the labelling performed to date. This poses some technical and methodological issues:

1. The information of the web pages may be subject to important temporary changes: there may be an appearance over time of new relevant terms, previously not used, or features that are relevant for classification, not present in the first design. This information must be learned from data scraping new web pages.

2. The web page data scraping process generates a dictionary of terms (bag of words) that is specific to the collection of documents used. Data scraping new web pages would generate new bags of words that, in principle, could be incompatible with the original.

3. The labels may be subject to significant change: the same company may substantially change its website, offer electronic sales when it did not do so before, or vice versa, so that the label obtained manually in the first analysis loses its validity.

There are engineering mechanisms to resolve these problems, but they involve the design of software modules that integrate different databases, combining the data labelled manually in the first analysis with new data labelled in a second analysis. It would be hoped that this integration would further reduce the need for relabelling.

CONCLUSIONS

MACHINE LEARNING for IaD enabled processing more than

100,000

BUSINESSES and, BY LABELLING only

0.8%

of the websites, detected the presence of B2C activity with LESS THAN

8%

ERROR RATE.

5.4 Conclusions

Although the definitive evaluation of the viability of using the Internet as a Data source for automatically detecting B2C activity in websites of Spanish businesses requires an evaluation of the results of the analysis by the final users of the application, the results obtained allow certain preliminary conclusions to be drawn:

- Using ML techniques, it is possible to automatically detect the presence of e-commerce with a precision over 92% (that is, 8% error rate).
- The classification error rates are relatively low. Some of these errors are attributable to limitations of the software architecture (such as, for example, it was not possible to extract information from non-textual websites or those in other languages), others were due to websites that are problematic even for manual labelling.
- Comparing the results obtained by ML techniques with those of other years, obtained by non-automated methods, showed a high level of agreement.

- Machine learning did not totally eliminate the need for labelling, but it did significantly reduce this need. The work performed suggested that in successive uses of the software, labelling of less than around 1% of the total websites would be sufficient. For the data used in this analysis, this represented around 1,000 samples.
- The software application allowed labelling to be performed in multiple sessions, which could be interrupted to evaluate classification performance. When performance stabilises, the labelling can be stopped. The work done showed that some 1,500 labels could be sufficient.

The performance obtained does not constitute a limit to the performance of automated learning. By doing additional R&D work, it would be possible to develop technologies for the whole process that would reduce the detector error rates. By way of example, future versions of the detector could include functionality for:

- From the set of pages rejected by the crawler, to discriminate automatically between those that do not have content (and therefore do not have B2C activity) from those that have content that cannot be downloaded. This would reduce the set of 25,965 pages that were not included in the analysis.
- Of the set of pages discarded by the crawler because they redirected to other domains, those domains corresponding to websites for performing B2C on behalf of other companies could be accessed. A listing of these domains would be necessary.
- The crawler could be provided with the capacity to extract information from javascript code. There are tools for doing this, but it would be necessary to include additional intelligence in the crawler to selectively access the options in the javascript code.
- The labelling tool could be enhanced to allow annotation of complementary information (for example, the existence of a restricted access area for customers) and the automated annotation of other data such as the date of labelling or the active learning mechanism used on each page, which could be used to improve the efficiency of automated learning.

6

MATERIALS AND METHODS FOR THE DETECTION AND CHARACTERISATION OF JOB OFFERS AND TRAINING PLANS

6 Materials and methods for the detection and characterisation of job offers and training plans

OBJECTIVE

Analyse the viability of AUTOMATIC DETECTION OF JOB OFFERS and the automatic characterisation of the JOB OFFER and TRAINING PLANS.

The work performed in the second milestone or sub-project of this project has used three main information sources as a starting point:

- A list of URLs of Spanish businesses
- A list URLs of freely accessible job portals.
- Two URLs of official channels for accessing information on professional and university training available in Spain.

The ultimate objective of the sub-project is to analyse the viability of machine learning (ML) techniques for automatically browsing and analysing the data contained in these websites, and for extracting information that would enable the characterisation, also automated, of the curricular offering and labour demand profiles available at the time of the analysis. Likewise, the aim is to evaluate the possibility of comparatively analysing supply and demand, and obtaining alignment measures between labour supply and demand profiles, in such a manner as to enable the identification of labour demand insufficiently covered by the training currently available.

In order to achieve these objectives, it was necessary to complete three interrelated phases:

1. Develop software for capturing, analysing and viewing the data.
2. Apply the software developed to the characterisation of job and curricular offers.
3. Analyse the results obtained to draw conclusions on the viability of ML to perform comparative analysis of curricular offering and labour demand.

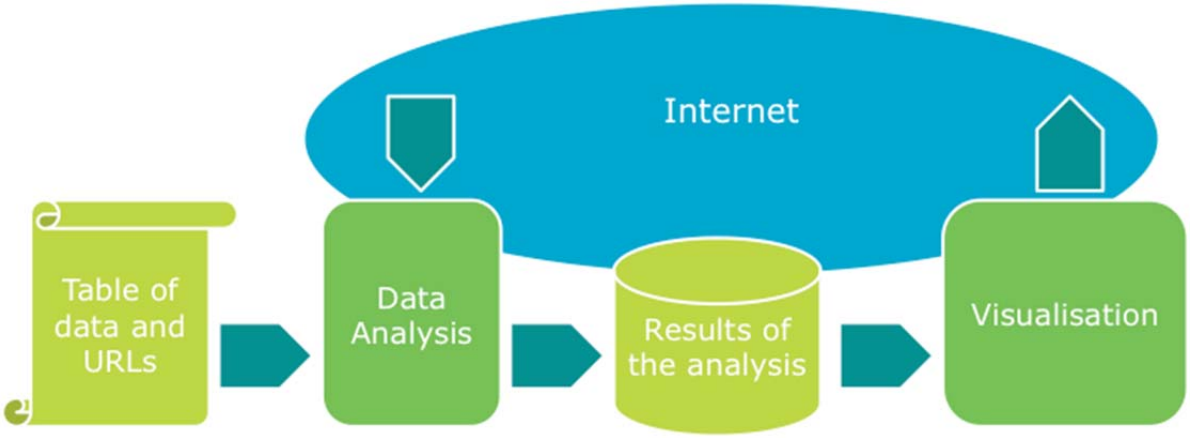
These three phases were highly interrelated: the analysis software was modified in successive iterations as a result of the conclusions of phases 2 and 3 in a process that was cyclical rather than sequential. These iterations in design were fundamental for selecting the ML algorithms and configuring them to optimise performance in the characterisation of labour supply and demand.

In this chapter we will describe the main stages of the labour supply and demand analysis and characterisation process, and the functionality provided by the software tool to carry it out. In subsequent chapters we will describe how the tool was applied in this study and the results obtained.

6.1 Labour supply and demand analysis and characterisation detection process

As indicated above, the software development phase for automated detection of B2C e-commerce requires a first stage of data capture and analysis (automated) and a second phase of visualisation and analysis (by users) of the results of the automated process. This process is illustrated in the following graphic (completely identical to that illustrated by the detection of B2C in Figure 1).

Figure 17. Process of Detecting B2C E-commerce



In order to view the results of the analysis using a conventional web browser, without need for an integrated application, the visualisation module is independent from the analysis application.

For the analysis of both job offers and curricular offering, two alternatives were initially proposed:

1. **Access to portals with centralised information**. The basic idea consists of using web portals that concentrate all the job offers or curricular offering or, at least, a sufficiently significant portion of either. On the one hand, in the case of sources of job offers, there are various portals that publish a significant number of job openings in companies (InfoJobs or Tecnoempleo, for example). The official qualifications offered can be consulted in repositories such as the Spanish Registry of Universities, Centres and Qualifications (RUCT).

    a. Advantages:
        i. Portals concentrate abundant information, sufficient to perform statistical analysis and automatic profiling.
        ii. The information is organised in a relatively uniform format, thereby facilitating analysis.

    b. Drawbacks:
        i. Dependence on a web structure. Whenever job or official qualification portals change their web structure, the analysis software can stop working or give rise to unwanted results.

2. **Direct access to sources of job offers** (corporate websites) **or curricular offering** (university websites and vocational training centres.

    a. Advantages:
        i. Potential access to all the jobs or official qualifications offered.
        ii. Independence from the portal web design.

    b. Drawbacks:
        i. Difficulty to collect information: additional engineering work is required to locate the specific job offer or offers on the website of each company, or the syllabus and course descriptions on the websites of each educational institution (university or vocational training).
        ii. Information heterogeneity: each institution organises and publishes the information with its own format and with very different levels of detail and structure.

Given that the second option requires much more complex engineering work and offers less reliable results, the solution adopted in this project was the following:

1. Perform a comprehensive analysis of supply and demand based on centralised information portals.

2. Explore the viability of automatically detecting the job offer directly on corporate websites. That is, exploratory detection is only performed on the sources of job offers.

The greater information heterogeneity of portals with centralised information offer certain guarantees on the documentary sources used for profiling. This will allow us to determine the potential of the automatic profiling (which is one of the ultimate objectives of this study) and assess future interest in the automatic extraction of information from the websites of companies and educational centres.

Consequently, the data analysis process has two differentiated components.

1. Job offer detection process in corporate websites.

2. Supply and demand profile analysis process, based on specific information sources (employment portals and degree programmes)

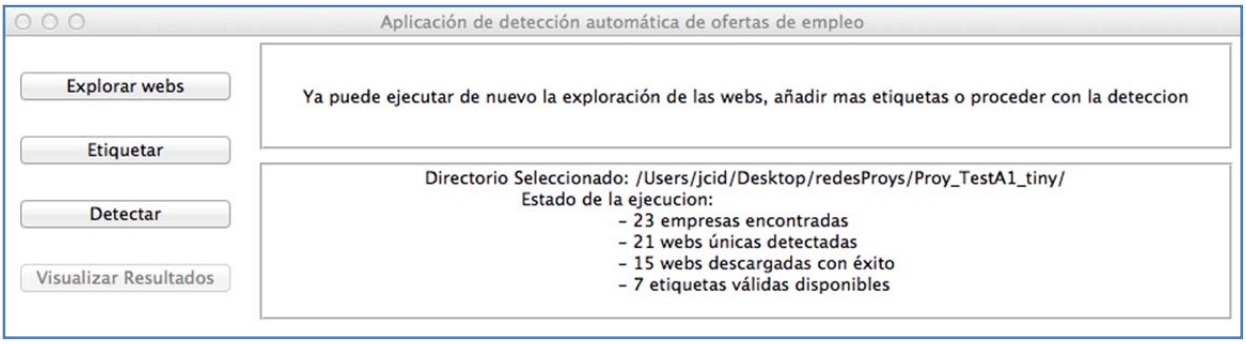
Precisely due to the use of very different information sources, these processes have required the use of differentiated software tools, which were developed independently. For this reason, we describe them separately in the following sections.

We will firstly describe the automated job offer detection process.

6.2 Job offer detection process

An application has been developed to perform the entire automated job offer detection process. The structure and operation of the application is similar to that of the B2C detector described in Chapter 7. On initiating the execution of this application (details of its installation and functioning can be found in the technical appendix of the project), a window opens displaying the following figure.

Figure 18. Main Window of the Data Capture and Analysis Software



The **Archivo** (File) menu allows the user to select the folder containing all the Excel files that contain lists of businesses with the URLs of their websites.

The application allows maintaining a number of different active projects, each with its corresponding listing of businesses.

The process of data capture, analysis and visualisation requires 4 main steps, which correspond to the buttons shown in the window.

6.2.1 PHASE 1: Web Browsing ("Crawling")

After clicking the button **Explorar Webs** (Browse Websites), the application initiates the web crawling or browsing process, data scraping from the websites of businesses and capturing relevant information for detecting job offers. The result of this process is a record of terms that are relevant for the analysis of each of the visited websites.

CRAWLER

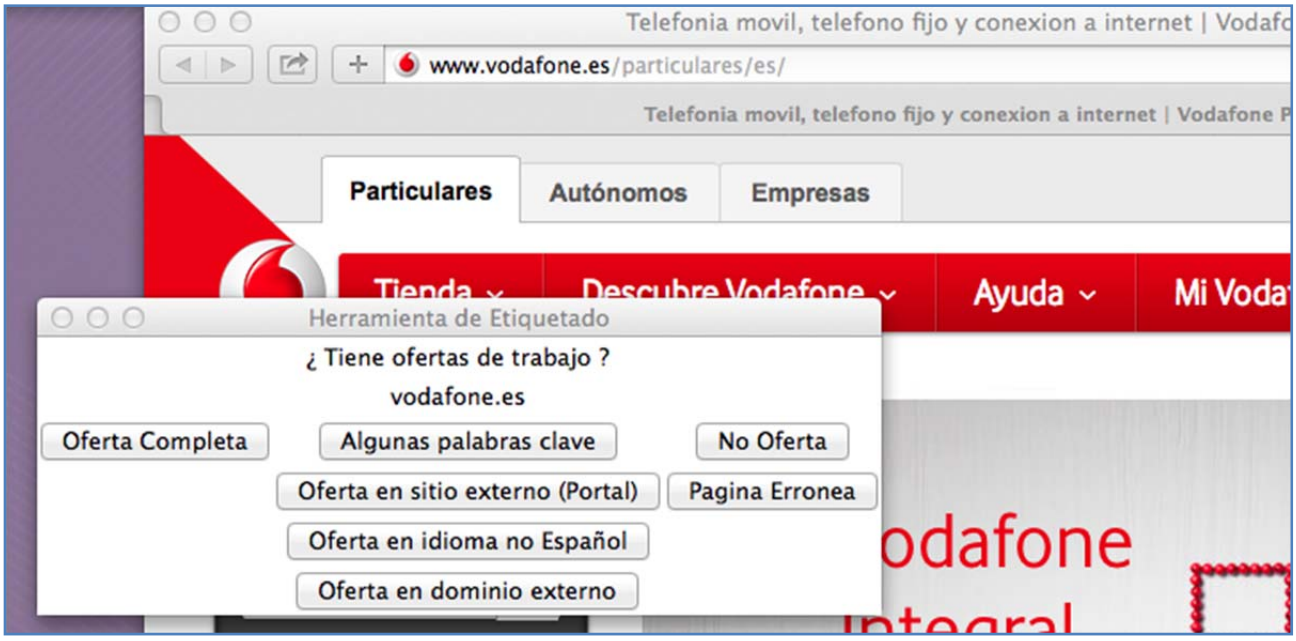
The **INTELLIGENT CRAWLER** only browses web **LINKS** with **EVIDENCE OF JOB OFFERS**

In order to reduce the duration of this process, the crawler uses an intelligent link selection mechanism, similar to that used in detecting B2C activity, which avoids having to browse all the pages of each website, selecting only those links that are likely to contain relevant information for detecting job offers.

6.2.2 PHASE 2: Labelling.

Upon completing the browsing stage, the next step is to label the websites by clicking the **Etiquetar** (Label) button, which will enable us to obtain information on the presence or absence of job offers in a small number of websites, which is essential for ML algorithms. On clicking the labelling button, the application displays the corporate website in the default browser and asks the user about the presence of job offers that can be profiled on the website.

Figure 19. Illustration of Manual Labelling on Mac OS X



Users may freely browse the website until determining the presence or absence of job offers. Multiple labelling is used to detect offers, allowing users to choose between the different options.

1. **Full Offer:** The website has a job offer that can be profiled, i.e. with sufficient detail to be profiled in Spanish.

2. **No offer**: The website does NOT contain any profilable job offers.

3. **Some keywords**: A website appears with terms related to job hunting (for example, links for sending CVs or contact details to potential job seekers in the company) but not containing specific offers that can be profiled. Initially these represent one of the main causes of false positives, as they contain words related to job hunting and therefore also appear in true positives. URLs that publish job offers but to not have any at the time of consultation are also classified here.

4. **Offers on external sites (Portal)**: Users are rerouted to a job portal containing the company's job openings.

5. **Offers in languages other than Spanish**: The offers appear written in a language other than Spanish.

LABELLING

It is possible to apply ***MULTIPLE LABELLING***, recording more accurate information each website's typology

6. **Offers on external domains**: Users are rerouted to a domain other than the main website containing the job offers.

7. **Erroneous Page**: The website does not exist, gives page load errors or reroutes users to a company that sells web domains. Pages "under construction" were classified under No Offer.

Multiple labelling offers information that could be of interest for future analyses. In this project, the main interest is centred on the detection of profilable job offers and therefore the detector will use the "Full Offer" label.

The labelling session is terminated when the user closes the window. It is now that the application saves the labels entered and is ready to apply the ML algorithms.

Labelling can be performed over several sessions. Each session can be interrupted at any time and can be restarted afterwards.

The application uses an intelligent Active Learning mechanism which it chooses to label the websites that contain the most information for the ML algorithm.

6.2.3 PHASE 3: Detection of job offers.

DETECTOR

For each business, it determines a **DECISION** relative to the presence or absence of job offers and a **RELIABILITY** value for this decision.

When the labelling is complete, the ML algorithm can be activated for classification by clicking on the button **Detectar** [Detection].

Following the automated learning process, the job offer detector processes all the relevant information captured by means of the crawling process and determines, for each company, an **Offer Score**: a high score, close to 1, indicates high evidence that the company has job offers, while a low score, close to -1, indicates that the website does not contain job offers.

The final decision of the job offer detector is the result of applying a threshold to all the scores:

- Companies whose Offer Score exceeds the threshold are assigned to the class "**with Offer**"

- Companies whose Offer Score is lower than the threshold are assigned to the class "**without Offer**"

a. Measurement of Detector Performance.

Detector efficiency is measured using the same type of parameters described in Section 3.2.3:

- **TPR** (True Positives Rate): the proportion of companies with job offers that were detected correctly.

- **FPR** (False Positives Rate): the proportion of companies without job offers that were incorrectly assigned to the class "with Offer."

The choice of the threshold value establishes a compromise between a high proportion of companies with detected job offers and a low rate of false positives. Since this compromise may depend on the detector's purposes, the job offer detector pays special attention to the four characteristic threshold values (see Section 3.2.3a, page 16):

- **BEP** (Break Even Point): the threshold value for which the rate of false positives (FPR) and false negatives (companies with job offers that are classified as "without Offer") are equal.
- **TPR**=0.95. Threshold value that guarantees 95% true positives.
- **FPR**=0.05. Threshold value for which the rate of false negatives is 5%.
- **FP**=**FN**. Threshold value for which the total (absolute) number of false positives and false negatives are equal.

Finally, the detector determines a reliability measure on the decision taken, which is calculated as follows:

$$\text{Reliability} = 1 - \frac{|\text{Offer Decision} - \text{Offer Score}|}{2}$$

b. Results of detection

As a result of the classification of all the companies, a file of classification results is generated containing a table in csv format (comma-separated values) combining the classification results with the information contained in the application input data files. Specifically, the table contains the fields indicated below for each company:

1. Data taken from the input files:
    - **Company name**
    - **NIF Code (Tax ID)**
    - **Primary CNAE code (activity classification)**
    - **Web address**
2. Data obtained after the detection of job offers
    - **Manual label** (-1 = "no", 1 = "yes", 0 = "without label", 2 = "partial offer", 3 = "offer on job portal", 4 = "offer on external domain", 5 = "offer not in Spanish", 99 = "erroneous website"). It indicates the presence or absence of profilable job offers obtained during the labelling process.
    - **Website available** (1 = "website downloaded", 0 = "not downloaded"). Indicates businesses excluded from the analysis because their website could not be downloaded.
    - **Offer Score** (value between -1 and 1).

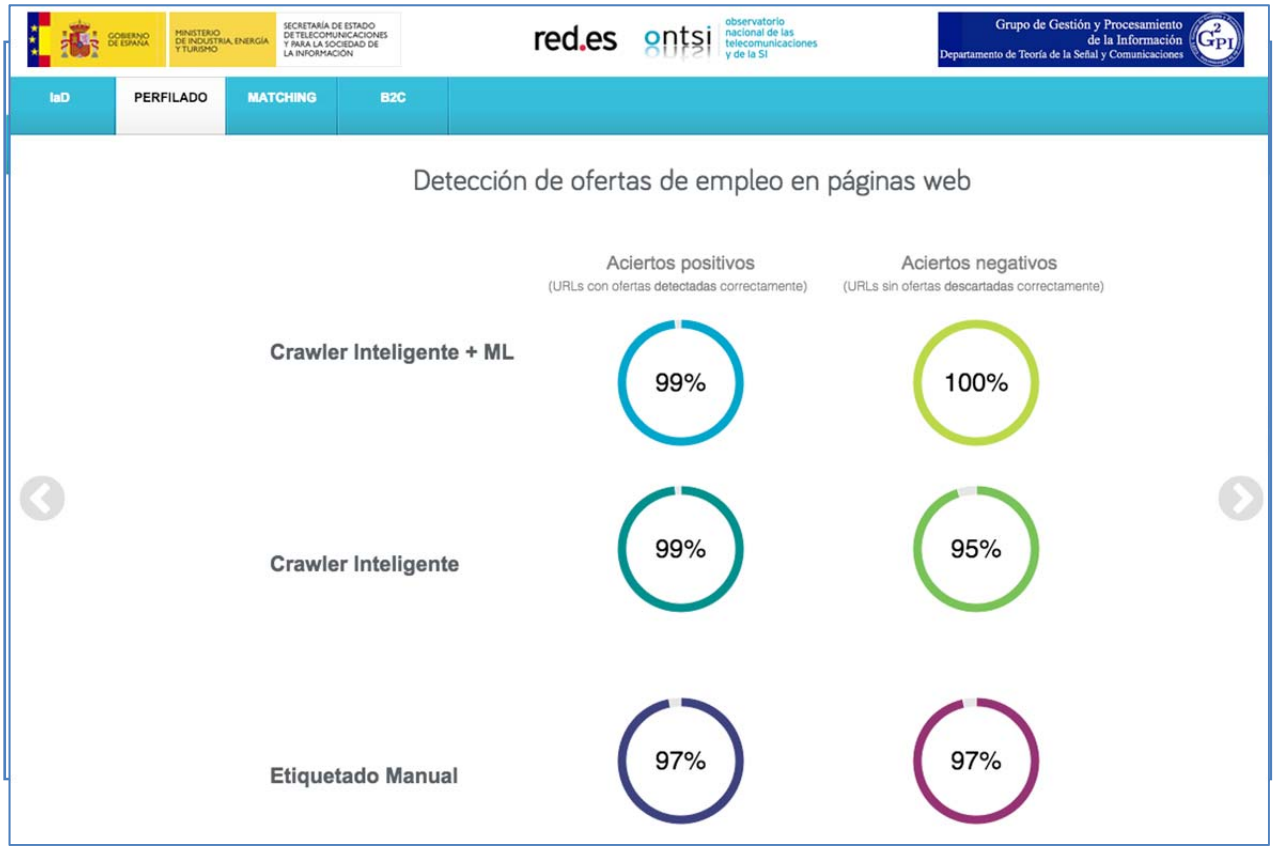
- **Offer Decision in BEP** (-1 = "no", 1 = "yes"). Detector decision when the BEP threshold is applied.
- **BEP Confidence** (value between 0 and 1)
- **Offer Decision in TPR**=0.95 (-1 = "no", 1 = "yes"). Detector decision when TPR=0.95 threshold is applied.
- **TPR confidence**=0.95 (value between 0 and 1)
- **Offer Decision in FPR**=0.05 (-1 = "no", 1 = "yes"). Detector decision when TPR=0.05 threshold is applied.
- **FPR confidence**=0.05 (value between 0 and 1)

6.2.4 PHASE 4: Visualisation

Clicking on the button **Visualizar** (View) in the application leads to a view of the results of the analysis in the web browser. This is independent of the application, so that the viewer can be accessed from any browser by entering the URL.

The display page consists of a web page such as that shown in the following figure, which clearly displays the results of the detection expressed as a percentage of positive hits and negative hits. The results are compared with a manual labelling score based on the frequency of discrepancies between the different individuals performing the labelling, with the accuracy of a detection system based exclusively on the location of offers by means of keywords in the crawling process.

Figure 20. Job offer detection result display page.



In addition to the results of the detection of job offers, it is possible to use the profiling tool described in the following sections to view the profiles of the job offers identified on corporate websites. These results will be discussed in the following chapter.

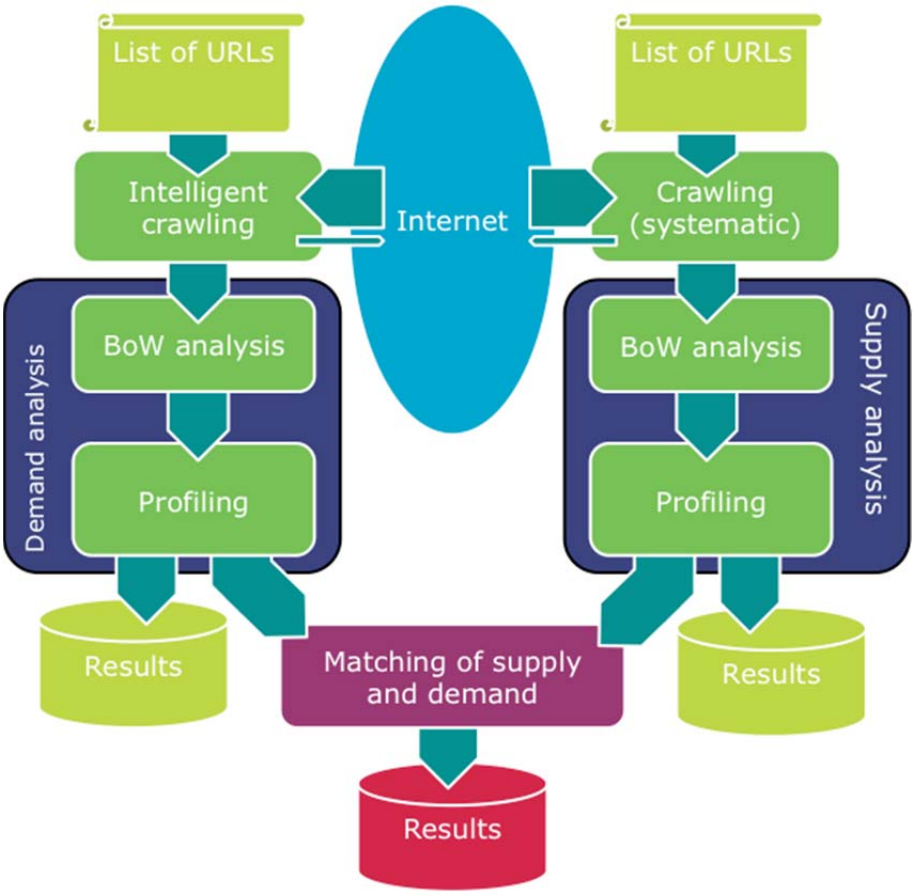
6.3 Supply and demand profile analysis process.

The labour supply and demand characterisation process is described schematically in Figure 21.

The process can be divided into three main tasks: web crawling, analysis (of supply, of demand and joint analysis of supply and demand), matching of supply and demand, and complementary or ancillary tasks.

An application has been developed allowing these tasks to be performed sequentially based on a predefined collection of freely accessible documentary sources. On initiating the execution of this application (details of its installation and functioning can be found in the technical appendix of the project), a window opens displaying the following figure.

Figure 21. Block diagram of the analysis tool



Following is a detailed description of each of these steps:

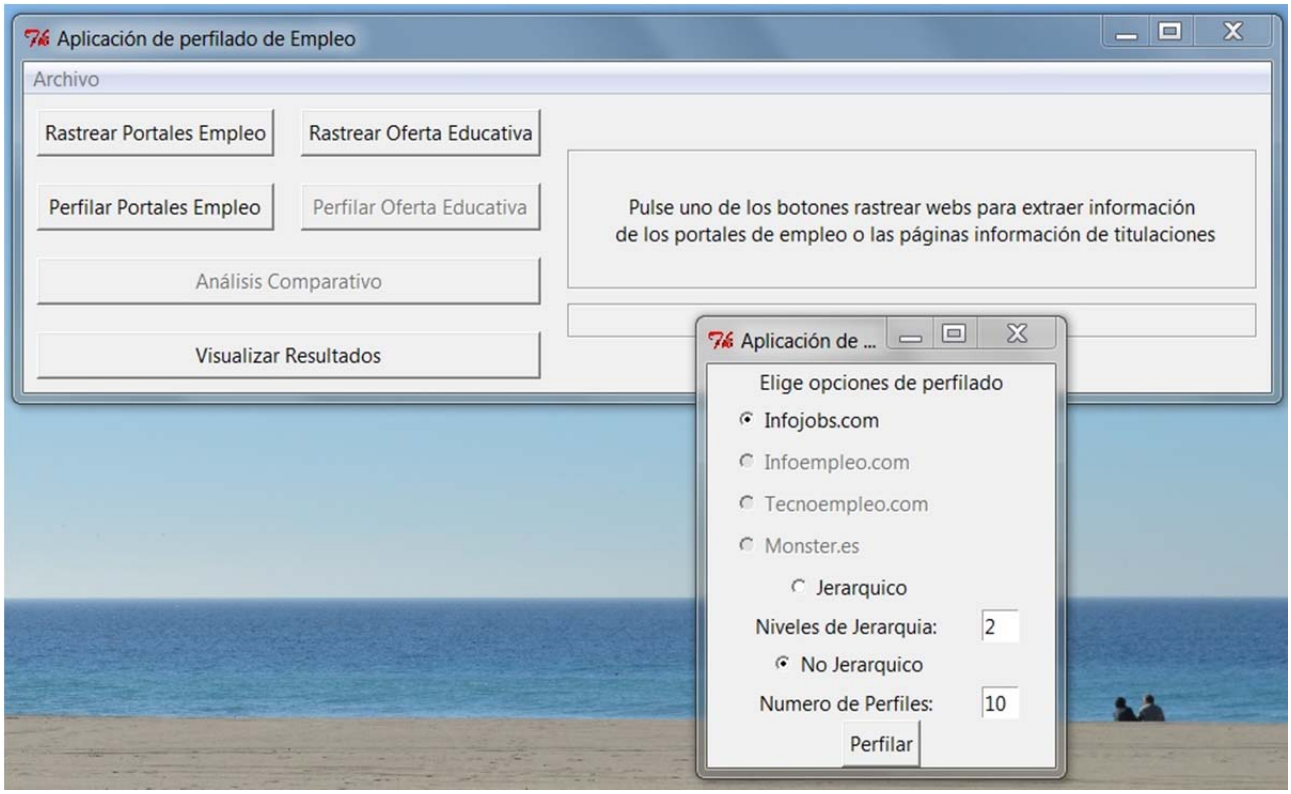
6.3.1 PHASE 1: Crawling

The web crawler accesses websites (job portals or public databases with training offerings) over the Internet and automatically downloads information relevant to profile analysis, without downloading the full offer in any case.

a. Job portal crawling

The use of public job portals is a very attractive option for the study and characterisation of job offers. The crawling functionalities included in the application allow us to download job offers from three different portals: InfoJobs (http://www.InfoJobs.net), Infoempleo (http://www.infoempleo.com) and Tecnoempleo (http://www.tecnoempleo.com). The browsing of offers in Monster (http://www.monster.es) and Ticjob (http://www.ticjob.es) was also evaluated. However, it was observed that the number of job offers obtained for Spanish ICT companies was insufficient for profile analysis in comparison with the three portals evaluated.

Figure 22. Main Window of the Data Capture and Analysis Software



In the selected portals, the crawler captures the content of those fields of the web page of each offer that generally contain relevant information for profiling, and deletes the rest. Specifically, the crawler captures the following information during the crawling process:

- **InfoJobs**: The offers corresponding to the following categories are explored: "IT and Telecommunications", "web design" and "marketing and communication", selecting the information contained in the following fields:
    - Offer name
    - Minimum requirements
    - Desired requirements
    - Labels of required technologies (not available for all offers)

- **Infoempleo**: In this case, the categories browsed are: "Technology and IT", "Telecommunications", "Internet" and "Media, publishing and graphic arts." The following fields are selected from among those available for the different offers:
    - Title
    - Functions
    - Requirements

- **Tecnoempleo**: Due to being a more specific portal than the previous two, the offers of this job portal are limited to technology, although the geographic search area is limited to Spain. Of the accessible information, that contained in the following fields is stored:
    - Title
    - Candidate's profile
    - Functions
    - Job specifications
    - Minimum qualifications
    - Labels of required technologies (not available for all offers)

In addition to the foregoing information, during the crawling process a metadata file is generated in the manner of a table of contents of the downloaded information, which will also be useful for visualising results. Said metadata will be stored in a csv file containing the following information in each line corresponding to a job offer:

- An offer identifier
- The name of the position offered
- The link to the job offer (URL).

b. Browsing of training offers

In order to analyse the curricular offering, users may choose from among two types of syllabi or training plans:

- **University qualifications**, including degrees and master's degrees.
- **Professional qualifications** or Vocational Training studies.

In the first case, the crawler explores the Official Register of Universities, Centres and Qualifications (RUCT, https://www.educacion.gob.es/ruct/home), and, in the second case, the

information made available by the National Qualifications Institute is used (INCUAL, https://www.educacion.gob.es/iceextranet/).

Since in both cases the information will be downloaded from specific portals, advantage is taken of the stable structure of these websites to design specific "crawlers" capable of leveraging the knowledge of the page structure. Therefore, the search is aimed at the site (page, document,...) where the relevant information for profiling the training offer is stored (in particular, those relative to the student's competences and training modules), without need to explore the entire portal content.

Figure 23. RUCT home page.

Internally, of all the available information (different degrees and official qualifications), the crawler selects only those training plans associated with the ICT sector. In order to classify these plans as ICT or non-ICT, the guidelines of various reports from centres such as the INE (Spanish National Statistics Institute), COIT (Spanish Association of Telecommunications Engineers) and ONTSI (Indicators on the Information Society and ICT) (see Section 9.1, where this process is detailed) have been followed.

The crawler obtains the following information for each ICT training plan:

- **Syllabus**: containing the general and specific competences of each university qualification or the list of modules and competences associated with a professional qualification. For each training plan, a text file with the content is generated. Once all the files were downloaded they are processed to create a single file containing the entire body and that will be used to the following phases in which profiles will be obtained.

- A set of **metadata** that are stored in a series of csv files which shall be of use for subsequent processing and, mainly, for visualisation.

    - In the case of university qualifications, the name of the degree, university, branch of knowledge, level and link to the syllabus.

- In the case of professional qualifications, the name of the qualification, the professional family to which it belongs and the level of qualification assigned.

Figure 24. Log in page to the Catalogue of Official Qualifications (INCUAL).

6.3.2 PHASE 2: Extraction of profiles

During the analysis phase itself, the information extracted by the crawler is analysed for subsequent characterisation. Said characterisation is based on the analysis of the words and terms that appear on the website, and can be broken down into three main stages:

a. Pre-processing of the dataset

Firstly, the dataset obtained by the crawler is pre-processed with the objective of adapting the literality of each document (job offer or qualification) to the characteristics of the learning model to be used. The pre-processing tasks carried out can be summarised in the following steps:

- Filtering of documents by language: From among all the job offers downloaded, only those in Spanish are stored.

- Deletion of stopwords, i.e. meaningless words (articles, pronouns, prepositions, etc.) that are not relevant to profiling.

- Detection of pre-selected n-grams. In this context, n-gram is understood to be any chain of words that can be considered a single term (for example, "big data" or "driver's licence"). The list of n-grams to be considered in the analysis of the body may be edited to facilitate the inclusion of new terms. The list included in the application was obtained in a semi-automated manner, using an n-gram detection algorithm followed by a manual inspection in which only those n-grams considered most significant due to providing a clearly differentiated semantic value with respect to its individual components were considered.

- Basic lemmatisation: The pre-processing tool includes a very simple lemmatisation tool that detects words that appear in the dataset in singular and plural, replacing the plurals with their singular counterparts. Additionally, in the case of university qualifications in which the documents typically include more narrative text, lemmatisation substitutes the verbal forms for the infinitive of the verb.

As a result of this pre-processing stage, a new dataset is obtained in which the literality of each document has been substituted for a mere collection of words that have passed all the previous filters. Based on said collection of words, it is not possible to rebuild the original document, but the basic idea is that the frequency of appearance of the terms is a good indicator of the semantic content of the original document, having deleted the semantically irrelevant words.

b. Construction of vocabulary and extraction of bags of words

The full vocabulary for each dataset would be formed by the set of terms that appear in each of its documents. Once the document has been built, a new representation is created for the documents that form each dataset known as "bag of words." Each document is characterised solely by the number of appearances of the dictionary terms, regardless of the order of appearance (i.e. the word sequence in the document). The profile extraction techniques act on the type of representations described below.

c. Learning of profiles

On selecting the profiling tool, the application analyses the pre-processed dataset and identifies the most descriptive profiles thereof. More specifically, a profile is characterised by a collection of words that are likely to be observed, most probably for the documents of said profile. The learning of the profile does not only consist of finding the most representative set of words, but also the specific probabilities with which said words (and all the words of the vocabulary) could be observed for a document that belongs purely to said profile.

During this phase, a set of profiles for the dataset analysed are learned automatically, together with the degree of pertainment of each document to each of the identified profiles. In general, the documents do not belong to a single profile, but rather can be described as a combination of different profiles in different proportions. It is important to mention the random nature of the automated learning tools used, so that different executions will give rise, in general, to slighting different profiles, although the profiles most frequently represented in the dataset usually appear in a nearly systematic manner.

In order to select this option, users must specify the specific dataset they wish to use (job offer or curricular offering) and the number of profiles they wish to obtain (the application establishes 20 profiles as a default option). After a processing time that depends on said number of profiles and size of the dataset, the profiles obtained, the proportion in which said profiles appear in the dataset and the most descriptive terms of each profile are displayed on screen. Likewise, the following files are generated:

- .tsv (tabulated separated values) file, containing the same information displayed on screen, together with the weight assigned to each word of the profile.
- .json files necessary for displaying profiles on a web interface.
- These files contain the numerical data that characterise the model, for subsequent use in the matching tool.

6.3.3 PHASE 3: Matching

In this phase, labour demand and training offering are jointly analysed for the purpose of obtaining different measures for "aligning" the curricular offering (of the Spanish university and vocational training system) with the demand for professionals by companies. The user must select a job offer dataset (InfoJobs/Infoempleo or Tecnoempleo), curricular offering dataset (degrees and master's degrees or vocational training) and the number of profiles to be identified in each of the datasets, in order for the application to estimate the following similarities:

- Similarity between each job offer and each curricular offer.
- Similarity between each job offer and one of the profiles identified with respect to the curricular offering.
- Similarity between each job offer profile and each qualification (university or vocational training)
- Similarity between each pair of job offer and curricular offer profiles.

In the event that the profile model selected is not previously available, the application will obtain it prior to estimating the similarities between documents and profiles.

The tool includes thre different strategies for learning these similarities. In the first strategy, a joint model of both datasets is obtained for the purpose of learning a cross-lingual semantic similarity measure for job offers and qualifications. Once said similarity is established, the tool exploits it to learn the similarities between the job and/or curricular offer profiles. The second strategy limits the search for similarity between documents to the list of terms of greatest relevance to the job offer profiles. Finally, a third strategy has been implemented that constraints the vocabulary to the most relevant terms in the job offer profiles, carrying out an analysis which is based on the frequency of appeareance of such terms in the documents associated to each of the analysed qualifications. The tool is configured to use the third of the aforementioned strategies that provided the best results, although it can opt for using the first strategy changing the default value in the settings file.

On selecting this option of the application, the tool creates the following files/file sets:

- .tsv files containing the following information:
  - For each document of each dataset, the set of documents of the other dataset that are most similar thereto and its degree of similarity with the profiles of the other dataset.
  - For each profile of each dataset, the set of documents of the other dataset that are most directly identifiable with said profile and its degree of similarity with the profiles of the other dataset.

- .json files for visualising the matching results.

- They are files containing numerical data with the similarity matrices between documents, between documents and profiles, and between profiles of both datasets.

6.3.4 PHASE 4: Visualisation

The visualisation tool designed consists of a web page that includes interactive graphics which enable users to visualise the different aspects discussed in the previous modules. Said web page has been developed pursuant to the stylesheet provided by Red.es with regard to the colours to be used and their order of appearance.

Once a profiling model has been generated for a certain dataset, whether relating to job offers or training modules, or even a relationship or matching model between job offers and training, it can be visualised clicking the **Visualizar** (View) button.

On clicking the **Visualizar** (View) button, a web browser opens displaying the home page of the application, on which four tabs appear:

- **IaD**: Home page access tab.

- **Profiling**: Enables analysis of the profiling results, both of job and training plan offers.

- **Matching:** Displays the results of the comparative analysis of supply and demand.

- **B2C:** Enables access to the results of the automatic detection of e-Commerce (see milestone 1 report).

This web tool is independent from the application, so that any browser can be used by entering the URL where the results are located.

Figure 25. Home page of the visualisation tool

In order to analyse supply and demand profiles, we will use the tabs associated with the profiling or matching model studies, due to which we will detail the content of each separately.

a. Visualisation of model profiles

The profiling model view is composed of four views with which users can interact by clicking two side buttons that enable them to move right or left. Visualisation commences with a general view of the profiles where users can select the model to be analysed. By default, users will have a menu showing al the profiling models generated earlier, whether relating to job offers or training modules.

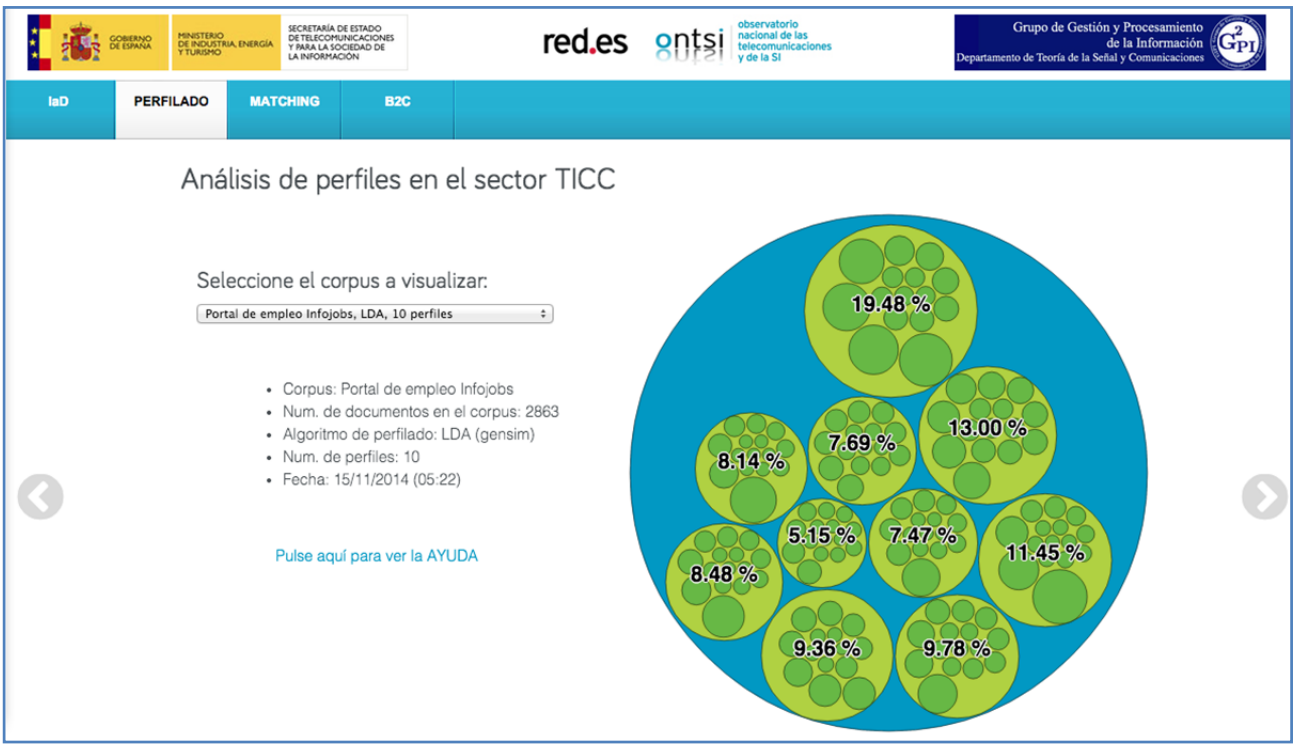
Once a module is selected, the following information will be displayed on screen:

- Dataset: indicating the name of the job portal or whether it is information about professional qualifications or university degrees.

- Number of documents in the dataset.

- Profiling algorithm used: LDA, HLDA...
- Number of profiles created.
- Creation date.

Figure 26. Page one of the profile viewer (module selection)



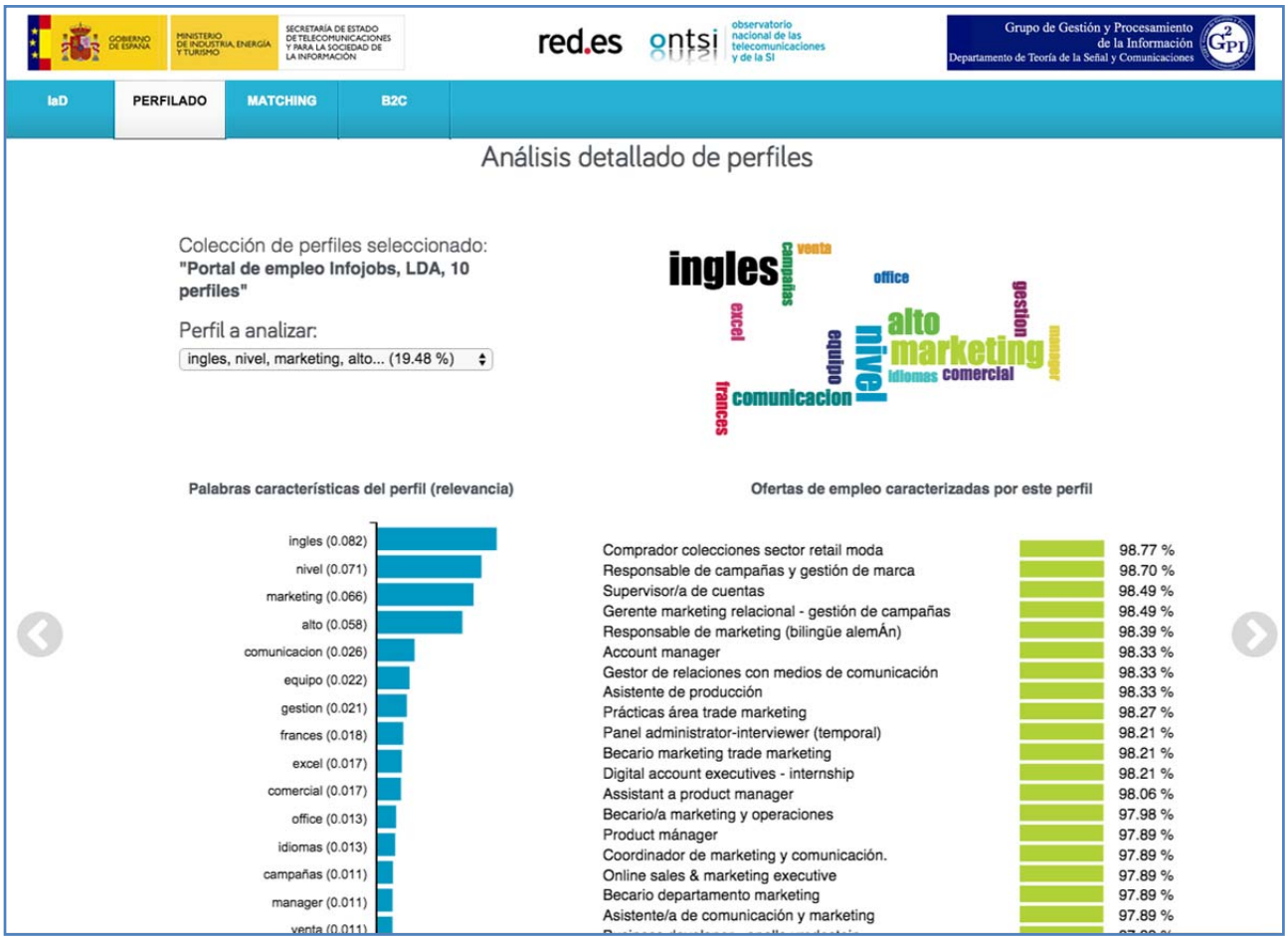
A bubble cloud with the profiles created will be displayed together with this information. Each bubble will contain a profile, indicating the percentage of representation of that profile in the dataset Clicking on the bubble will automatically zoom in on it and the terms or words that comprise the profile will appear.

If we advance towards the page on the right, we will arrive at a new section that allows us to analyse the profiles that comprise the model selected on the previous page in detail. To this end, a drop-down menu appears where we will firstly select a specific profile to be analysed; by default, on starting the application or switching to another model, select the most relevant profile with the greatest weight in the dataset. Once we have selected the profile, the following information will be displayed:

- **Word cloud**: it includes the most important words of the profile, where size is assigned in proportion to their importance.
- **Word list**: it displays the words of the profile, including a side bar whose length is proportional to their relevance. This provides a more quantitative measurement than the previous one for analysing the terms that comprise the profile.

- **Document list**: in this case, a list of the documents that are best characterised with the selected profile is displayed. Additionally, the percentage by which the selected profile represents each of the listed documents is also indicated.

Figure 27. Page two of the profile viewer (profile analysis)



Clicking on the right button again will take us to a third page where we can analyse a specific document. To this end, the page is composed of two elements:

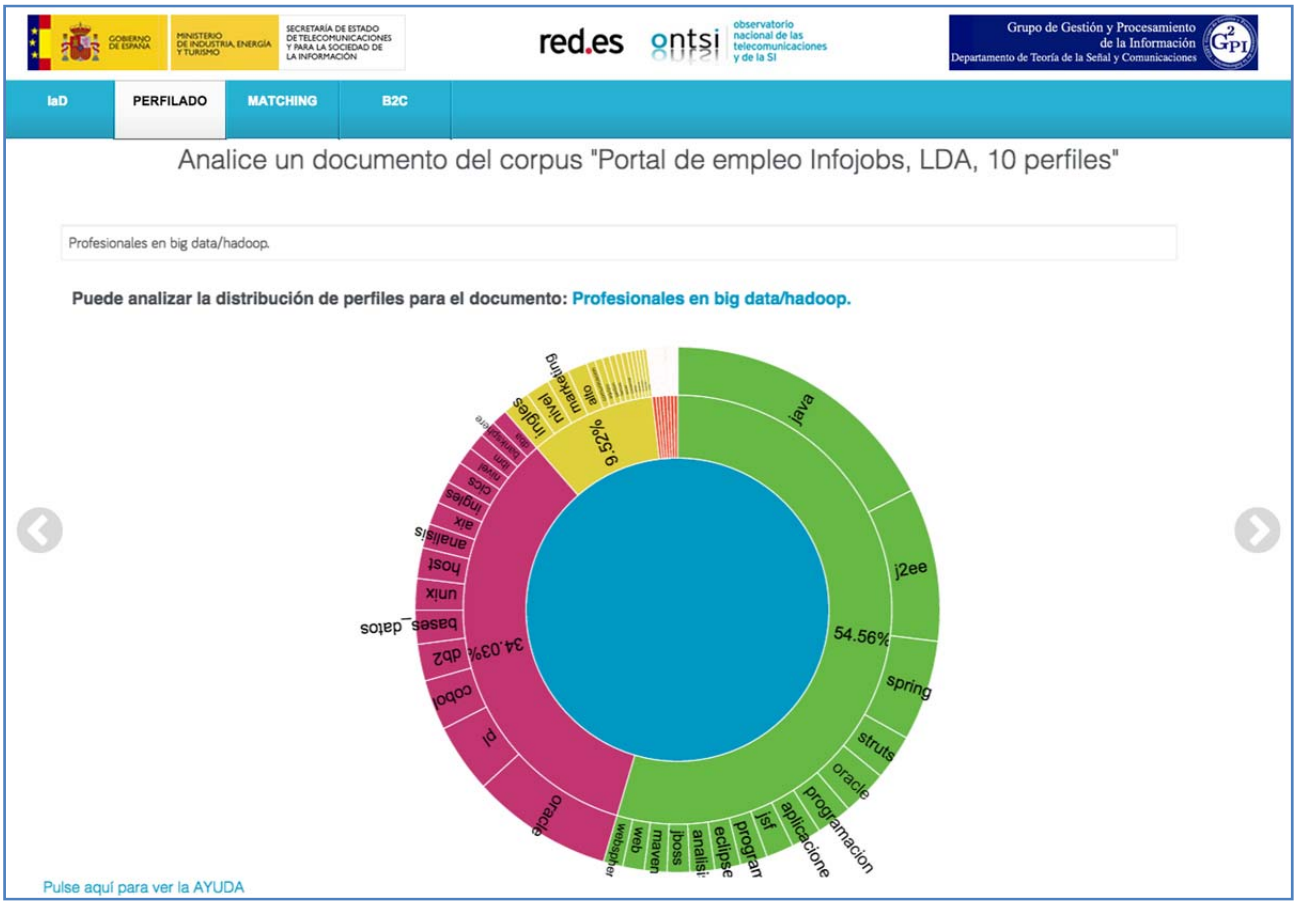
- **Document search engine**: allows us to search for any document of the dataset by name. To help us with our search (rather than presupposing that the user knows the name of the elements that form the dataset), as we type the text in the search field, all the documents whose name contains the typed characters are retrieved. The document is selected by clicking on the desired document from among the options proposed.

- **Distribution of a document across the profiles**: this representation is comprised of a circle with different crowns in which circular sections are assigned to each profile, so that the section is proportional to the participation of each profile in the description of the selected document. The first (inner) crown shows the profiles

with their participation percentages and the second (or outer) crown shows the words that form the profile. Additionally, in order to facilitate visualisation or analysis, users can click one of the profiles to obtain a detailed view of its content and the words that comprise it.

In this last view, if no document is selected, the original distribution of the profiles across the entire dataset is shown.

Figure 28. Page three of the profile viewer (document analysis)



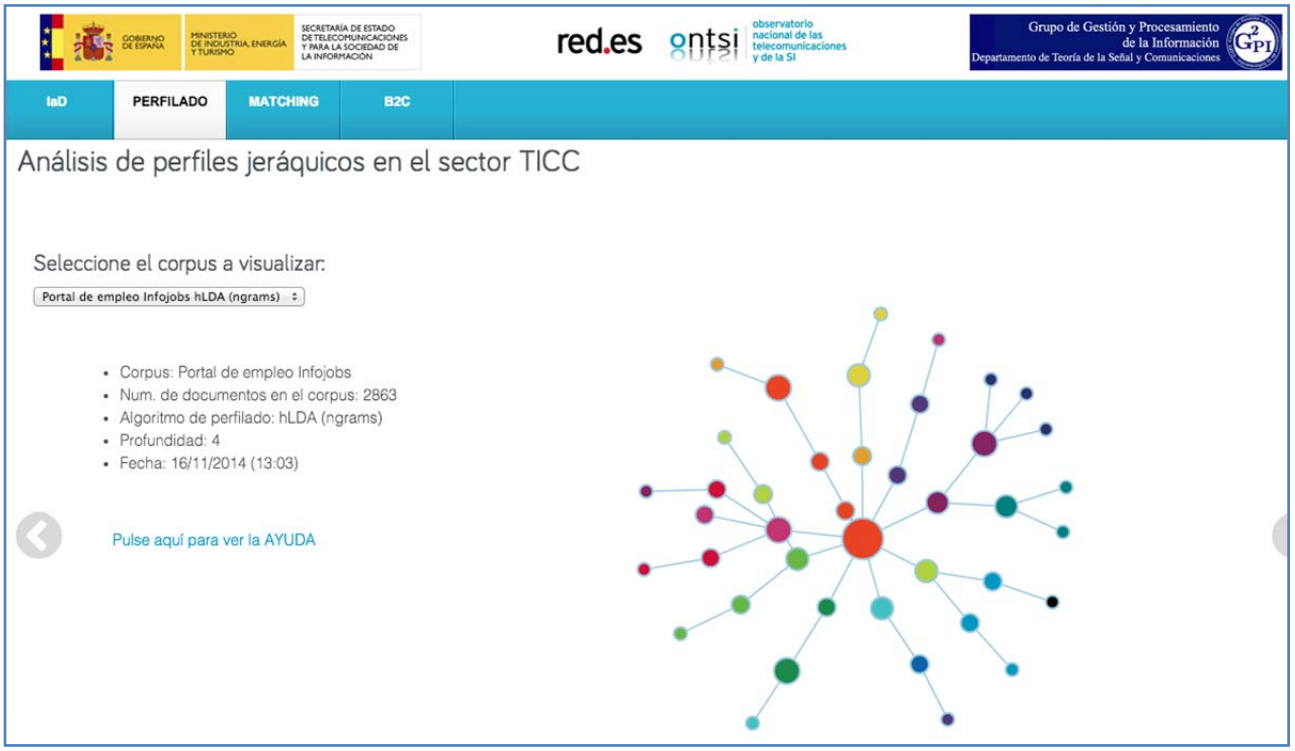
The fourth and last view, which represents the hierarchical profiles and is fully independent from the previous views, it is comprised of two elements:

- A **drop-down menu** on which the user can select the available hierarchical models. On selecting a model, the following information will appear on screen:
  - Dataset: it indicates the name of the job portal or the type of training profile (i.e. training profile of professional qualifications or university degrees).
  - Number of documents in the dataset.
  - Profiling algorithm used
  - Hierarchy levels used
  - Creation date.

- An **interactive tree** (called "code-flower"), where each tree node represents a profile and has a set of branches that associate it with

the parent profile (of the hierarchical model) and with its child profiles. On clicking on a node or profile, a description of the terms that comprise it appears and all the child nodes are grouped together. This enables users to interact with the tree and leave the branches that are of greatest interest.

Figure 29. Page four of the hierarchical profile viewer



b. View for analysing the connection between job offers and training modules.

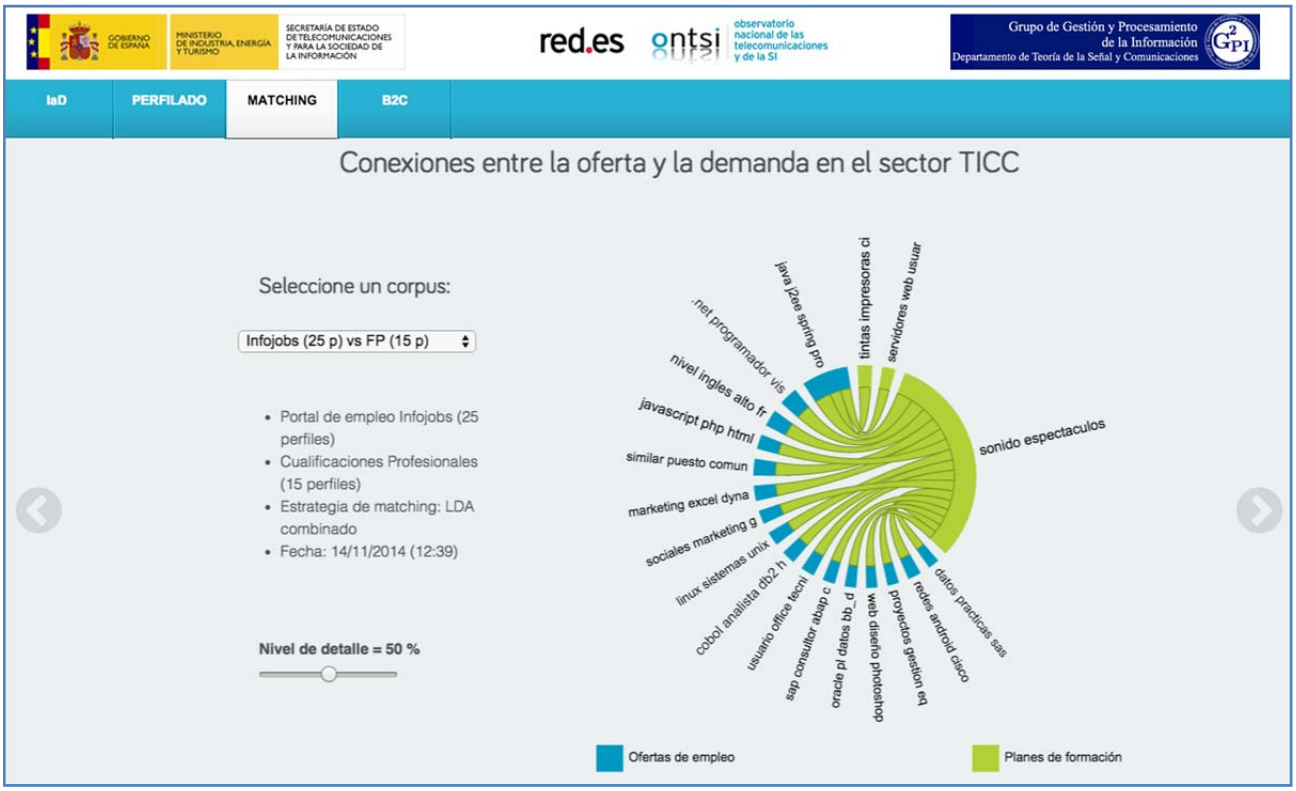
As in the previous case, this view is comprised of three views with which users can interact through two buttons that allow them to move sideways between pages.

Again, in the first or initial page, users can select the model to be analysed through a menu displaying all the models generated. In this case, the model includes a dataset of job offers and a dataset of training offers in which their relationships were analysed. Therefore, in this case, upon selecting a module, the following information will be displayed on the screen:

- Dataset of job offers: indicating the name of the job portal explored and, in brackets, the number of profiles used for modelling thereof.
- Dataset of training offers: indicating whether they are professional qualifications or university degrees and, once again, including the number of profiles used for modelling thereof in brackets.

- Profiling algorithm used: LDA, HLDA...
- Creation date.

Figure 30. Page one of the matching viewer (module selection)

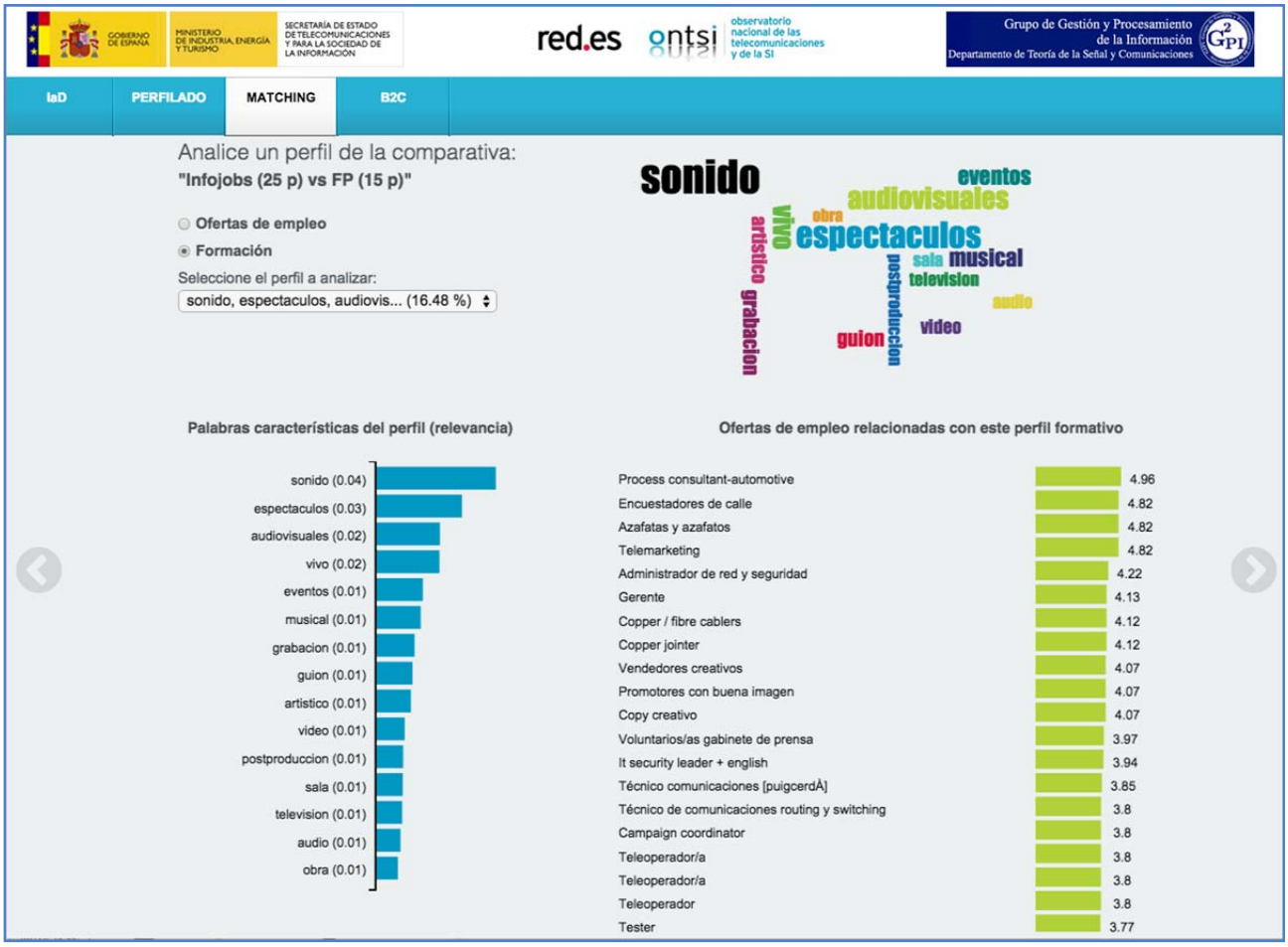


Together with the information about the model, there is a relationship wheel on the right-hand side of the page showing all the profiles of each dataset and a series of arcs that join the related profiles.

Since the matching algorithm obtains a connection value for each pair of crossed profiles, we must select a threshold (relationship level between profiles) below which it would be considered that the connection is insufficient and those two profiles are not related. In order to help the user to control this threshold, a horizontal bar that the user can move to change the threshold has been added. When the user sets the threshold to zero, all the profiles (offers vs training) are interrelated; however, as its value increases, the user can observe how relationships disappear and even profiles that are orphaned (not related to any other) are eliminated from the representation.

If we advance towards the page on the right, we will arrive at a second page that allows us to analyse the profiles that comprise the model selected on the previous page in detail. Since in this case training and job offer profiles are mixed, a first round button will allow us to select which of the two types of profiles we wish to analyse; upon selecting it, a drop-down menu appears in order to specify the profile to be analysed. Upon selecting the profile, the following information is displayed:

Figure 31. Page two of the matching viewer (profile analysis)



- **Word cloud**: as in the case of the profiling module, it includes the most important words within the profile, assigning them a font size proportional to its weight within the profile.

- **Word list**: it represents the same information but with bars proportional to the relevance of each term.

- **Document list**: in this case, as opposed to the profiling page, the documents which are most closely related with the selected profile are listed. Additionally, the level of the relationship of each document with the profile.

Lastly, if we advance towards the third page, we will obtain a view for analysing the relationships between documents. To this end, the page is divided into two vertical sections, one associated to the job offers (left view) and another associated with the training plans (right view). Each section includes two elements:

- **Document search engine**: it allows the search for any document in the dataset by name with enhanced search capabilities (in the same manner as the document search engine used in the profile analysis).

- **List of related documents**: once the document selected in the search is retrieved it displays a list of the ten documents most

related to it. Therefore, if we have searched for a job offer, the training modules or university degrees/master's degrees (according to which is being analysed) most closely related to it are retrieved.

Figure 32. Page three of the matching viewer (analysis of relationships between documents)

Additionally, in order to facilitate the interaction between the search results of each view (portals or training), users can click on one of the listed documents, whereupon the related documents are retrieved and displayed on the complementary side view. That is, if a list of job offers related to a training plan has been obtained, on clicking any offer the list of training plans related to the clicked offer will appear (or will be refreshed).

7

ANALYSIS OF THE DEMAND FOR ICT PROFESSIONALS IN CORPORATE WEBSITES

7 Analysis of the demand for ICT professionals in corporate websites

This chapter describes the main results of the detection and characterisation process of the demand for ICT professionals based on the information published in Spanish corporate websites. The results are based on the application of the job offer detection tool, the functionality of which is described in Section 6.2, and of part of the profiling tool described in Section 6.3.

7.1 Data source

DATASOURCE

8,349

ICT sector COMPANIES.

828

Manually LABELLED COMPANIES.

The starting point of this study consists of a list of 8,349 URLs corresponding to ICT sector companies provided by Red.es. The crawler completed the download of 6,791 of these URLs without errors. Approximately 10% of these URLs are labelled manually to build the classifiers. The result of this manual analysis is as follows:

- 828 URLs analysed, of which:
  - 678 URLs (81.9%) do not have profilable job offers.
  - 25 URLs (3%) contain at least one profilable job offer, i.e. written in Spanish and with a more or less detailed description of the job role, candidacy requirements and working conditions.
  - 15 URLs (1.8%) reroute the user to an InfoJobs type job portal where its job offers are listed.
  - 44 URLs (5.3%) offer some instructions to potential job seekers, such as for example an e-mail address where they can send their CVs and a general message encouraging potential job seekers to contact them, but no profilable offer.
  - 19 URLs (2.3%) have profilable job offers but in a language other than Spanish (English, Danish, Dutch, etc.).
  - 2 URLs (0.2%) have job offers in a domain other than the URL, but without being a job portal.
  - 45 URLs (5.4%) were links to websites other than that of the company in question, either due to having unregistered the domain or due to a downloading error when labelling.

7.2 Purpose of the study.

The ultimate purpose of the analysis of the demand for employment consists of automatically characterising the job offers of Spanish companies or, at least, the offers published on their corporate websites. As mentioned in the introduction to this report, the full automation of this characterisation process requires solving several engineering problems:

1. Detecting the presence of job offers.
2. Finding the locations of the job offers on the Internet.
3. Segmenting the offers (i.e. separate each individual offer in pages including multiple offers).
4. Characterising the job offers of all the companies as a whole.

In this project only the first, second (partially) and fourth problems were addressed. On the one hand, an automatic job offer detection tool (see Section 6.2), intelligent browsing techniques aimed at finding the pages where the job offers are located were applied and, after manually extracting and segmenting the offers detected, the offer was characterised online using an automatic profiling technique.

Therefore, we have excluded the automation of the extraction of each individual job offer as of the URL from this study, due to being a task the difficulty of which falls outside of the objectives of this project. Each URL organises its offers following a specific structure different from the other URLs. Therefore, in some cases the offers are in plain text file in the URL itself, in other cases it links to a pdf or doc file per each offer; there are URLs that include all their offers within the same text, while others use different texts for each offer, etc. To the difficulties involved in the extraction of the offer we must add the low percentage of URLs which have profilable job offers.

Therefore, the first specific purpose of this application is to detect and retrieve those URLs included on the list provided that contain profilable job offers with the help of an automatic classifier To this end, an automatic classifier must be trained to discriminate the URLs containing job offers based on the text contained in each URL.

The approach for solving the problem using machine learning algorithms consists of interpreting the retrieval of web pages with job offers as a binary classification problem: the aim is to classify each website in one of two classes: “there is a profilable job offer” and “there is none”. Henceforth we will call websites labelled manually as containers of profilable job offers “positive class” and websites labelled manually in any of the other classes (without offers, with some keywords, with offers in another language, with offers in external domains or with offers in job portals) "negative class".

The extraction and processing of each specific offer for profiling thereof will be performed in a subsequent task. An off-line profiling prototype will be built based on the URLs labelled as containers of profilable job offers. Off-line profiling is based on manually extracting the contents of the offers found and automatically process them using the profiling module. The objective of

this work is to evaluate the benefits that could be expected from fully automatic profiling in the case that the localisation and segmentation of the individual offers could be automated.

7.3 Test design

The experimental testing of the application will be conducted in the following phases (similar in many aspects to those described for the detection of B2C in sections 4.3, 4.4 and 4.5):

- **Crawling**: the crawler access to 8349 web domains from the company listing with the aim to locate, at each one of them, the web pages that could contain a job offer. Those domains withouth any evidence of a job offer are discarded and classified automatically in the category "Without Offer". After this stage, 7454 domains are classified in this category.

  For the remaining 895 domains, the crawler identifies and downloads the specific web page or pages where some job offer evidence has been found. Those pages will form the input dataset for the ML-based classifier.

- **BoW analysis** (extraction of the bag of words). The texts of the URLs processed are transformed into frequency vectors based on a TF-IDF measure (see Section 4.3.2). This processing gives rise to representations in bags of 796,701 words. Those terms that appear in less than ten websites are eliminated, reducing the size of the bags to 94,763 words.

- **Manual labelling** of 828 URLs. Only 96 of these labels belong to pages with some evidence of a job offer and, consequently, they will be used during the machine learning step. The remaining pages correspond to domains already classified in the category "Without Offer" and will be used for the performance evaluation.

  Labelling is performed on the seven categories described in Section 6.2.2 although, as we are now interested in detecting profilable job offers, we will focus only on Category 1 ("Full Offer"), ignoring the differences between the rest of the categories, which we will group into a single category of "No Full Offer".

- **Classification.** A classifier having the following characteristics is trained:

  - Initial number of words: 94,673.
  - Classification algorithm: Logistic Regression (LogReg)
  - Characteristics selection method (words): Bagging
  - Number of selected characteristics: 1,000.

  This classifier has been chosen after a comparative analysis process of different classification algorithms and characteristics extraction method that we will describe in Section 11.1.

- **Analysis of results.** TPR and FPR will be evaluated for different detection threshold values, graphically represented by means of ROC curves.

- **Profiling** of the job offers found in the URLs belonging to the class "Profilable Offers".

7.4 Results of the Study

The results of this process are shown in Figure 20. The combination of an inteliigent crawling and the machine learning algorithm provides a true positive rate of 98.55 %, and a true negative rate of 100 %, that is, the detector misses only 1 % of the job offers, and it does not make any mistake when it decides that there are no job offers in the company website. Obviously, these estimations are based in the set of labelled webs, which are about 10 % of the total number of webs, but they show that the process of crawling and location preceeding the machine learning is very efficient, in the sense that 95 % of the total number of companies without job offers are correctly discarded.

According to the classifier results, at the time of the study, 522 companies had a job offer in its web site (around 6.25 % of the total).

Lastly, we will analyse the results by applying the profiling tool to the set of 156 offers manually extracted from the 25 websites detected during manual labelling as a positive class.

Table 1. Job profiles obtained from Spanish corporate websites

Topic 0 (34.45%)	Topic 1 (25.25%)	Topic 2 (23.3%)	Topic 3 (10.8%)	Topic 4 (6.27%)
programming	services	partner	csb	attach
java	servers	atsystems	system	careers
atsystems	sql	to consult	German	multiply
javascript	administration	log in	commercial	points
j2ee	unix	directions	offers	view
php	Linux	signup	marketing	disciplines
drupal	vacancies	company	vitae	oral
sql	SAP	position	progress	significant
analysis	computing	summary	send us	outside

The table shows the most relevant terms of the profiles found automatically. It can be observed that offers classified under the profile "Topic 0" correspond to programmer positions, those classified under "Topic 1" to systems administrator positions and "Topic 3" to marketing and sales positions, while offers classified under "Topic 2" and "Topic 4" are not well defined, due to which the most significant words are general words that are not aligned with any of the previously described specific profiles.

In comparison with the results obtained in the job portal analysis, which are shown in the next chapter, the profiles obtained from corporate websites are more vague and difficult to interpret, due mainly to the fact that the

information contained in job offers follow no standard organisation or structure, and the wording styles and contents themselves are more dispersed.

8

ANALYSIS OF DEMAND FOR ICT PROFESSIONALS IN JOB PORTALS

8 Analysis of demand for ICT professionals in job portals

As an alternative to corporate websites, in this chapter will show the results of the characterisation of job offers based on information captured from job portals. The results are based on the application of the profiling tool described in Section 6.3.

8.1 Data Source

DATASOURCE

6,951 JOB OFFERS taken from THREE PORTALS in NOVEMBER 2014.

This analysis is based on data obtained from three job portals, including the offers extracted from each portal on 7 November 2014. The following table summarises the most important properties of the three datasets used in the study. In order to simplify the analysis, certain job offers that were published in a language other than Spanish were discarded. Despite the fact that web design, media and communication categories were considered, it should be noted that the number of offers in categories other than purely technological categories is very low.

Table 2. Most relevant characteristics of the datasets used in the analysis of sought-after job profiles based on job portal data.

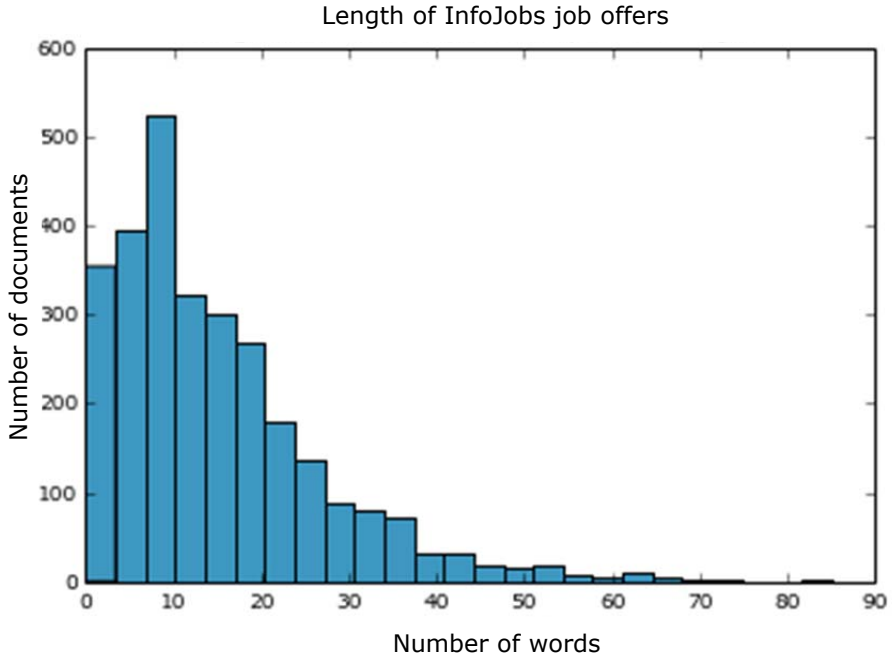
	InfoJobs	Tecnoempleo	Infoempleo
# Crawler offers	3,079	3,380	1,160
# Discarded	216	301	151
# Dataset offers	2,863	3,079	1,009
Categories	IT+Telco Website Design Marketing + Comms	---	Technology Telecommunications Internet Graphic arts
Size of vocabulary	1,211	1,881	823
Average size of the documents (in words)	14.85	46.05	23.67

In order to build the vocabulary of each dataset, all the terms that appear in at least two different offers of the dataset used, but in less than 50% of the documents of said dataset, were preserved. Additionally, the following procedures for building the vocabularies were carried out:

- Manual edition of the stopword files. A large number of terms, of a completely general nature, present in the job offers were included in said list.
- Detection of n-grams. The n-grams considered most relevant to defining the profiles sought were selected from a list created by an algorithm called Turbotopics.

Table 2 also shows the size of the vocabularies used for each of the datasets and the average number of words that comprise each job offer after pre-processing. Figure 33 illustrates the distribution of lengths of the documents that comprise the dataset of the InfoJobs job portal. The distribution of the lengths of the other two datasets shows similar profiles, although with a greater average number of words.

Figure 33. Histogram showing the length of the offers of the InfoJobs job portal



8.2 Purpose of the study and test design

The purpose of the profiling experiments conducted on the datasets of job offers in job portals is to analyse the following aspects:

- Visualise the profiles obtained for each job portal.
- Compare the difference between job portals.
- Study the influence of the number of profiles.
- Analyse the profiles obtained for the dataset obtained from the pooling of all the portals.
- Analyse the results of hierarchical modelling.

For the development of the tests, ten different models were trained for each specific experiment, manually selecting the models included in this report. It should be noted that there are no established procedures for comparing models in an objective and unsupervised manner, since the theoretically most adjusted models are not necessarily aligned with an identification of more useful profiles. In any case, this strategy requires a low level of

supervision with respect to the manual labelling used in other tasks of this project.

8.3 Study results

8.3.1 Selection of profile number

This subsection shows the results of the profiling of job offers from the InfoJobs portal. We analysed the impact of the selection of the number of profiles, which must be performed initially on the type and quality of the profiles obtained.

Table 3, Table 4 and

Table 5 show the profiles obtained for 10, 20 and 30 profiles, respectively. In order to adequately interpret the results, it is important to take into account that the profiling algorithm is not only about grouping together documents with a similar content (in the sense of using the same words), but rather attempting to explain the appearance of all the words of each of the documents of the dataset. Additionally, global alignment is sought, due to which the profiles that are scarcely represented have a smaller influence on the evaluation that the profiling method itself has on the quality of the profiles obtained. For this reason:

- Some of the profiles obtained can be associated to a set of documents explained almost completely on the basis of said profile. However, other profiles may be of a more cross-cutting type and be present in a large number of documents, although in a generally non-dominant manner. A clear example of this is the frequent appearance of a profile containing terms associated to language skills (English, French, German, level, advanced, etc.). Said profile appears in models with 20 profiles (profile 1) and with 30 profiles (profile 1 and 17).

- The selection of the number of profiles implies a certain compromise between the degree of granularity of the analysis (i.e. the accuracy and definition of the profiles that can be obtained) and the appearance of profiles with low semantic content. Therefore it can be observed that, in the case of the model with 10 profiles, all the profiles identified have clear semantic content and reasonable reflect the demand for professionals in the sector. However, said analysis only offers a very global vision of the sector, so that interesting but smaller profiles do not appear at said level of detail. By way of example, profile no.11 of the model with 20 profiles is defined by terms such as "seo", "social_networks", "sem", "community", "google" or "google_analytics". Therefore, it is fairly clear community manager profile that will probably be classified under profile 7 of the model with 10 profiles.

- Although increasing the number of profiles enables the identification of potentially more defined profiles and with less representation in the dataset, the increase in the flexibility of the model also favours the appearance of noisier profiles. This effect is especially evident for

models with more than 20 profiles, as can be visualised in profiles 17, 18, 25, 28 and 29 of the model with 30 profiles (

- Table 5). It is interesting to note in this point that further increasing the number of profiles enables the identification of specific cases of interest. For example, for 50 profiles the two following profiles are obtained (they can be examined in the visualisation tool):
    - Profile 41: "data", "sas", "statistics", "mathematics", "model", "data_mining", "spss"
    - Profile 49: "big_data", "architect", "hadoop"

Table 3. Characterisation of the job offer profiles detected in the InfoJobs portal (10 profiles).

Topic 0 (19.5%)	Topic 1 (13%)	Topic 2 (11.4%)	Topic 3 (9.8%)	Topic 4 (9.4%)
marketing	Linux	java	businessintelligence	javascript
level	Windows	j2ee	power_center	html
French	systems	spring	projects	html5
commercial	security	struts	consultant	php
advanced	networks	oracle	testing	css
communication	vmware	jsf	functional	css3
excel	unix	eclipse	telecommunications	jquery
English	support	maven	sas	web
campaigns	administrator	jboss	processes	design
office	servers	tomcat	business_objects	Ajax
Topic 5 (8.5%)	Topic 6 (8.1%)	Topic 7 (7.7%)	Topic 8 (7.5%)	Topic 9 (5.1%)
oracle	.net	marketing	SAP	sharepoint
PL	asp_net	seo	android	crm
cobol	server	digital	abap	dynamics
db2	visual_studio	online	ios	navision
host	visual_basic	cisco	consultant	red_hat
unix	mvc	manager	liferay	microsoft
aix	programming	sem	tibco	consultant
cics	vb_net	community	SOA	nav
banksphere	entity	social_networks	bpm	cognos
dba	Windows	communication	mobiles	salesforce

We can therefore conclude that the selection of the number of profiles is not a trivial task and requires a certain level of supervision, as it depends on the type of profiles that the user wishes to obtain. It should be noted that the profiling method used always prefers to increase the number of profiles available for optimising its objective criteria, since the models with less profiles would be classified within those that use a greater number of profiles.

As reflected in the tables, the method used enables the measurement of the size of the profiles in the dataset. In this regard, the selection of the number of profiles also influences said estimates, causing said sizes to be smaller, in general, even in the case of very well defined profiles such as profile 2 of Table 3, the weight of which decreases from the 11.3% of the model with 10 profiles to up to 7.02% for the model with 10 profiles. The reason for said decrease is closely related to the nature of the profiling method used.

Table 4. Characterisation of the profiles of the job offers detected in the InfoJobs portal (20 profiles).

Topic 0 (9.1%)	Topic 1 (7.4%)	Topic 2 (7.1%)	Topic 3 (7.05%)	Topic 4 (6.5%)
java	level	.net	team	marketing
j2ee	English	php	communication	advertising
spring	advanced	asp_net	management	online
struts	written	server	organisation	communication
jsf	spoken	programming	orientation	testing
eclipse	German	sharepoint	commercial	English
oracle	engineer	javascript	level	digital
programmers	international	visual_studio	customer	ade
programming	travel	mysql	excel	programmes
maven	advanced	mvc	English	sale
Topic 5 (6%)	Topic 6 (5.1%)	Topic 7 (4.95%)	Topic 8 (4.77%)	Topic 9 (4.6%)
html5	Linux	SAP	Windows	networks
web	unix	abap	vmware	telecommunications
javascript	oracle	management	Linux	cisco
design	systems	projects	server	iP
css3	aix	consultant	systems	engineer
html	dba	project	support	firewall
css	administration	manager	servers	tcp
jquery	shell_script	cloud	red_hat	lan
adobe	backup	bw	microsoft	routers
photoshop	administrator	itil	administrator	dinner
Topic 10 (4.45%)	Topic 11 (4.35%)	Topic 12 (4.3%)	Topic 13 (4%)	Topic 14 (3.9%)
businessintelligence	seo	android	cobol	java
functional	social_networks	ios	db2	web
power_center	manager	mobiles	host	servers
business_objects	sem	telecommunications	cics	websphere
consultant	digital	insurance	banking	applications
services	online	python	banksphere	jboss
tibco	community	sale	risks	services
microstrategy	google	commercial	functional	oracle
server	web	developer	programmers	tomcat
reporting	google_analytics	telephony	partenon	weblogic
Topic 15 (3.85%)	Topic 16 (3.7%)	Topic 17 (3.34%)	Topic 18 (3.19%)	Topic 19 (2.45%)
pl	security	French	microsoft	data
oracle	software	liferay	crm	sas
analysis	management	architecture	dynamics	magento
forms	infrastructures	helpdesk	navision	statistics
reports	design	customer_care	consultant	amb
developer	consultant	computer technician	erp	mathematics
Windows	test	fp	nav	coneixements
model	information	native	salesforce	visual
view	identities	teradata	access	model
xml	tester	technician	office	design

Table 5. Characterisation of the profiles of the job offers detected in the InfoJobs portal (30 profiles).

Topic 0 (7.02%)	Topic 1 (5.6%)	Topic 2 (4.78%)	Topic 3 (4.6%)	Topic 4 (4.4%)	Topic 5 (4.21%)
java	level	.net	unix	marketing	oracle
j2ee	English	server	systems	communication	PL
spring	advanced	sharepoint	Linux	manager	data_bases
struts	international	asp_net	Windows	English	navision
jsf	conversation	visual_studio	administration	ade	dba
eclipse	test	visual_basic	aix	advertising	dynamics
programmers	languages	mvc	vmware	product	nav
maven	oral	microsoft	backup	team	developer
analysis	senior	vb_net	systems_administrator	fluid	forms
junit	travel	entity	administrator	trade	data
Topic 6 (3.99%)	Topic 7 (3.99%)	Topic 8 (3.96%)	Topic 9 (3.52%)	Topic 10 (3.49%)	Topic 11 (3.45%)
user	cobol	engineer	marketing	projects	web
support	db2	telecommunications	online	management	design
microsoft	host	security	digital	team	html
office	banking	systems	seo	manager	html5
level	cics	technician	sem	planning	css
technician	functional	computer technician	campaigns	customer	formatting
computer technician	risks	consultant	social_networks	suppliers	usability
helpdesk	banksphere	infrastructures	google	project_manager	css3
advanced	analysis	identities	advertising	itil	responsive
incidents	partenon	management	management	pmp	ux
Topic 12 (3.13%)	Topic 13 (3.03%)	Topic 14 (3%)	Topic 15 (2.95%)	Topic 16 (2.94%)	Topic 17 (2.1%)
Linux	customer	SAP	businessintelligence	networks	programming
servers	communication	consultant	power_center	cisco	abap
applications	technician	management	business_objects	liferay	webdynpro
weblogic	organisation	bw	microsoft	¡P	fp
websphere	orientation	abap	microstrategy	tcp	iv
red_hat	team	ps	consultant	firewall	dynpro
server	vehicle	projects	etl	routers	programming
jboss	excellent	senior	business	lan	moodle
administrator	driver's_licence	functional	reports	networking	batch
administration	manager	mm	teradata	dinner	bapis
Topic 18 (2.74%)	Topic 19 (2.71%)	Topic 20 (2.63%)	Topic 21 (2.63%)	Topic 22 (2.59%)	Topic 23 (2.57%)
internship	html	applications	French	html5	software
excel	javascript	android	travel	javascript	testing
marketing	css	ios	German	css3	sas
power_point	xml	mobiles	English	jquery	engineer
communication	Ajax	amb	office automation	front	qa
intern	jquery	developer	native	developer	test
image	web	coneixements	customer_care	angular	quality
journalism	jsp	devices	bilingual	js	statistics
research	alfresco	nivell	languages	mvc	design
audiovisual	soap	platforms	translation	wordpress	agile
Topic 24 (2.34%)	Topic 25 (2.31%)	Topic 26 (2.3%)	Topic 27 (2.3%)	Topic 28 (2.09%)	Topic 29 (1.96%)
commercial	domain	crm	php	services	manager
sale	written	business	mysql	integration	basic
telephony	spoken	consultant	big_data	architecture	platforms
work	photoshop	processes	g¡t	tibeo	community
cognos	adobe	internet	hadoop	SOA	magento
teleoperator	office	functional	api	reporting	visual
telemarketing	programmes	insurance	apache	bpm	twitter
telephone	adobejllustrator	salesforce	python	architect	prestashop
webcenter	packet	dynamics	zend	server	facebook
agent	email	consulting	symfony	services	management

8.3.2 Profiles obtained for the different job portals

With the objective of comparing the profiles obtained on using different datasets, we have included the description of profiles for the Infoempleo and Tecnoempleo dataset, and the profiles obtained on jointly considering the three datasets used. In all cases, models with 20 profiles are considered, which in the previous section proved to be a value that strikes a reasonable balance between the relevance and the semantic value of the profiles obtained and controls the number of irrelevant and noisy profiles.

Table 6. Characterisation of the profiles of job offers detected in the InfoJobs portal (20 profiles).

Topic 0 (9.6%)	Topic 1 (8.1%)	Topic 2 (7.25%)	Topic 3 (6.59%)	Topic 4 (5.79%)
java	.net	advanced	telecommunications	php
j2ee	server	English	engineer	developer
spring	sql	level	networks	web
hibernate	web	French	installation	javascript
web	microsoft	international	repair	mysql
struts	sharepoint	German	mobiles	git
jsf	asp.net	computer technician	technician	html
framework	javascript	ms	telephony	framework
oracle	framework	travel	computer technician	magento
jboss	visual_studio	travel	operator	languages
Topic 5 (5.48%)	Topic 6 (5.46%)	Topic 7 (5.28%)	Topic 8 (5.03%)	Topic 9 (5.03%)
management	SAP	banking	programmer	global
security	abap	businessintelligence	analyst	agio
computer technician	crm	oracle	cobol	liferay
systems	bw	power_center	business_objects	project
engineer	consultant	functional	consultancy	ap
telecommunications	mm	business_objects	abap	siebel
business	P¡	sql	printing	administrator
programmes	business_objects	forms	yes	customer
project	sd	videos	machine	remedy
suppliers	srm	pl	technologies	analysis
Topic 10 (4.82%)	Topic 11 (4.61%)	Topic 12 (4.35%)	Topic 13 (3.94%)	Topic 14 (3.65%)
support	servers	web	manager	design
incidents	server	design	functional	data
user	Linux	photoshop	tibco	documentation
care	Windows	figure; graph	banksphere	sw
technician	systems	html	cobol	reports
computer technician	virtualisation	css	project	definition
geographic	sql	graphic_design	partenon	functional
mobility	networks	image	architect	models
helpdesk	support	pages	financial	solution
printer	vmware	formatting	analysis	infrastructure
Topic 15 (3.54%)	Topic 16 (3.49%)	Topic 17 (3.46%)	Topic 18 (2.64%)	Topic 19 (1.86%)
unix	ios	digital	technician	infrastructure
administrator	SAP	marketing	operators	installation
oracle	systems	team	similar	rbs
Linux	android	communication	java	installer
db2	applications	commercial	community	manager
sas	address	organisation	rotary	team
drive	mobiles	online	cable	telecommunications
test	dynamics	seo	manager	stations
aix	functional	strategies	eclipse	sheets
Solaris	mobility	planning	antennas	integration

Overall, we can conclude that the most important profiles appear clearly in the three datasets, although those which have less weight are not clearly identified in the case of the Infoempleo portal, probably due to the lower number of offers of said dataset. If we compare the profiles of InfoJobs and Tecnoempleo, the results seem better in the case of the InfoJobs portal. Although it is difficult to establish the exact causes, we estimate that it could be due to the fact that the crawling of the Tecnoempleo portal establishes a description field that could include terms beyond the description of the technical skills required. The greatest size of the vocabulary and length of the documents reinforce the authenticity of this hypothesis.

Table 7. Characterisation of the profiles of the job offers detected in the Tecnoempleo portal (20 profiles).

Topic 0 (7.27%)	Topic 1 (7.17%)	Topic 2 (6.92%)	Topic 3 (6.8%)	Topic 4 (6.4%)
j2ee	sql	html5	administrator	secure
java	net	javascript	Linux	French
spring	asp	jquery	administration	indra
hibernate	server	css3	Windows	tickets
struts	asp_net	android	unix	advanced
programmer	visual_studio	web	vmware	English
jsp	framework	css	security	level
analyst	web	html	systems	ica
liferay	visual_basic	ios	servers	social
jsf	wpf	Ajax	systems_administrator	internship
Topic 5 (6.08%)	Topic 6 (6.07%)	Topic 7 (5.84%)	Topic 8 (5.44%)	Topic 9 (5.38%)
Windows	web	sql	cobol	security
support	mysql	oracle	commercial	management
helpdesk	developer	pl	banking	consultant
networks	Linux	data_bases	db2	manager
incidents	git	server	functional	processes
operator	backend	unix	host	business
users	developer	dba	cics	director
technical_support	drupal	forms	analyst	risks
installation	zend	llg	banksphere	planning
hardware	framework	administrator	competitive	communication

Topic 10 (4.77%)	Topic 11 (4.31%)	Topic 12 (4.08%)	Topic 13 (3.93%)	Topic 14 (3.89%)
SAP	java	businessintelligence	architect	altran
abap	tibco	crm	j2ee	everis
consultant	hr	power_center	java	innovation
mm	arelance	dynamics	jboss	vass
bw	xml	business_objects	weblogic	consultant
sd	consultancy	consultant	spring	industrial
basis	ejb	microstrategy	soa	future
¡V	spring	navision	websphere	business
P¡	swing	sas	maven	consulting
hcm	information	microsoft	bpm	qualified
Topic 15 (3.83%)	Topic 16 (3.53%)	Topic 17 (3.37%)	Topic 18 (2.7%)	Topic 19 (2.19%)
networks	iso	design	testing	sharepoint
cisco	quality	functional	tester	big_data
¡P	information	applications	definition	python
networking	altia	processes	hp	open
climate	fields	umanis	quality	Linux
alten	infrastructure	erp	test	scripting
tcp	public	headquarters	qa	cloud
juniper	consulting	supervision	quality	source
protocols	exis	band	automation	invoin
competitive	management	exper	jira	powershell

Table 8. Characterisation of the profiles of the job offers detected in the joint modelling of offers and InfoJobs, Infoempleo and Tecnoempleo (20 profiles).

Topic 0 (9.13%)	Topic 1 (7.74%)	Topic 2 (6.86%)	Topic 3 (6.73%)	Topic 4 (6.08%)
java	marketing	javascript	advanced	Linux
j2ee	online	html5	English	Windows
spring	excel	web	level	administrator
struts	digital	jquery	French	unix
jsf	seo	css	written	systems
liferay	communication	html	spoken	vmware
eclipse	social_networks	css3	helpdesk	servers
oracle	advertising	Ajax	support	server
maven	campaigns	developer	German	systems_administrator
jsp	commercial	mysql	languages	virtualisation
Topic 5 (5.83%)	Topic 6 (5.3%)	Topic 7 (5.2%)	Topic 8 (5.7%)	Topic 9 (4.6%)
support	functional	engineer	net	businessintelligence
networks	manager	telecommunications	asp_net	power_center
incidents	banking	java	visual_studio	crm
installation	altran	everis	server	business_objects
systems	banking	indra	visual_basic	sas
management	risks	international	mvc	microstrategy
telecommunications	management	advanced	wpf	etl
users	banksphere	internship	vb	hr
computer technician	projects	qualified	iso	arelance
technical_support	financial	English	wcf	data

Topic 10 (4.1%)	Topic 11 (4.2%)	Topic 12 (4.5%)	Topic 13 (4.5%)	Topic 14 (3.6%)
oracle	SAP	secure	android	security
PL	abap	commercial	ios	networks
weblogic	bw	tibco	design	cisco
websphere	consultant	tickets	mobiles	iP
unix	mm	flexible	applications	communications
forms	sd	sale	web	firewalls
reports	basis	social	testing	networking
data_bases	iv	services	developer	lan
server	ps	competitive	soa	juniper
developer	business_objects	bpm	test	protocols
Topic 15 (3.8%)	Topic 16 (3.3%)	Topic 17 (3.9%)	Topic 18 (2.8%)	Topic 19 (2.6%)
programmers	microsoft	Linux	oracle	architect
analyst	server	python	data_bases	serum
cobol	sharepoint	dynamics	hardware	magento
db2	services	navision	repair	jenkins
host	ms	big_data	dba	rest
cics	reporting	cloud	printers	jira
banking	net	mysql	administrator	api
jcl	Windows	apache	pe	web
mainframe	access	open	microcomputing	test
ap	foundation	perl	laptops	zend

Lastly, the joint analysis of the three datasets does not allow the identification of new data profiles other than those found individually in each dataset, but does show the viability of merging data from different sources of information. It should be noted that said merger was performed directly, without carrying out any uniformisation between the documents from the different portals.

In conclusion, the use of profiling techniques has enabled the identification of certain profiles that reflect the current job offer trends in the ICT sector.

The following table lists the most common profiles that we identified and shows, for each, the most relevant profile or profiles in the different datasets (Table 4, Table 6, Table 7, and

Table 8). Likewise, the most closely related profiles from among those identified in the PAFET VII report (limited to the digital content industry) were included in the ONTSI (Spanish Observatory for Telecommunications and the Information Society) study on the supply and demand of digital content professionals, and in the ESCO classification of the European Union.

Table 9. Summary of the most highly sought-after profiles in job portals: characterisation and relationship with profiles identified in other studies.

Profile Description	Skills (keywords)	Profiles	PAFETVII	ONTSI Study	ESCO Classif.
Expert JAVA programmer	Java / j2ee / spring / struts / jsf / eclipse / maven	InfoJobs (TO) Infoempleo (TO) Tecnoemp. (TO, T13)	Programmer	SW programmer	2512 - SW developers
Community manager	seo / social_networks / manager / sem / community / google / google_analytics	InfoJobs (Til)	Community Manager Content Editor	Community Manager	
Online Marketing Specialist	marketing / advertising / online / communication / commercial / strategies	InfoJobs (T4) Infoempleo (T17)	Online Marketing Specialist Content Editor	Marketing Technician Commercial Engineer	
Web Design Specialist	html / html5 / web / javascript / design / css / css3 / .net / php / asp_net / visual_studio	InfoJobs (T2, T5) Infoempleo (TI, T4) Tecnoemp. (TI, T2)	Graphic/Web Designer UX Specialist Webmaster	Web Technician	2513 - Web and multimedia developers
Digital Content Designer	web / design /photoshop / graphic_design / formatting	Infoempleo (T12)	Content Architect Designer	Graphic Designer Content Editor or Manager	
Systems Administrator	linux / unix / systems / administration / Windows/vmware	InfoJobs (T6, T8) Infoempleo (Til, T15) Tecnoemp. (T3)			2522 - System administrators
Software Support Technician	support / incidents / customer care / mobility	Infoempleo (TIO) Tecnoemp. (T5)			
Project and Product Manager	sap / management / projects / consultant	InfoJobs (T7)	Product Manager		
Database Management Specialist	pl / oracle / analysis / forms	InfoJobs (T15) Tecnoemp. (T7)		Database Administrator	2521 - Database designers and administrators
Test Engineer	testing / tester / quality / qa	Tecnoemp. (T18)	Test Engineer		
COBOL Banking Specialist	cobol / db2 / host /cíes / banking / banksphere / risks	InfoJobs (T13) Tecnoemp. (T8)			
Data Analyst	BusinessIntelligence / Business_objects / power_center / data / sas / statistics	InfoJobs (TIO, T19) Infoempleo (T7) Tecnoemp. (T12)	Data Journalist Specialist *Big Data*		
Mobile Applications Designer	android / ios / mobiles / applications	InfoJobs (T12) Infoempleo (T6)	Programmer Technician Application Developer		2514-Application programmers
ICT Security Specialist / Network Engineer	networks / cisco / firewall / tep / ip / routers / security / management / infrastructure	InfoJobs (T0J16) Infoempleo (T5, T9) Tecnoemp. (T15)		ICT Security Specialist	2523 - Computer network professionals

8.3.3 Subsequent detection of n-grams

The nature of the profile extraction techniques used in this project makes it inadvisable to define a high number of n-grams a priori, as their use as autonomous entities could make it difficult to identify certain similarities between terms. By way of example, the "advanced-english-level" n-gram would be treated in a completely different manner to the term "English" and their interrelationship could only be determined through the coappareance of both in a high number of documents.

In order to detect the highest number of relevant n-grams, an analysis of coappearances of terms in the documents of the dataset was performed subsequently. Said analysis was based on the results of the profiling algorithm, so that the detection of n-grams is specific to each profile. Likewise, said analysis allows the same term to be considered independently and as part of a more complex n-gram.

The visualisation tool enables the selection of the representation of the extracted profiles, including the a-posteriori n-gram detection strategy implemented by the SW "turbotopics". It should be noted that the use of said tool enables the resorting of some of the terms that comprise the profile, in addition to preventing the use of the common term penalisation strategy described in Section 11.2.3.

In summary, and with the aim of illustrating the type of n-grams that said algorithm identifies, below we include the n-grams that were detected for the InfoJobs job portal, on considering the extraction of 10 and 20 profiles. It should be noted that, in general, said n-grams do not usually appear in the most relevant positions of each profile, due to which we limit these lists to those terms that appear among the 15 most descriptive terms of each profile.

- A-posteriori n-grams for LDA Modelling with 10 profiles of the InfoJobs dataset (see also Table 3):
    - *Profile 0 (marketing, level, French…): "advanced English level"*
    - *Profile 3 (business_intelligence, power_center, projects…): "project management"*
    - *Profile 4 (javascript, html5…): "html5 css3", "html css", "javascript jquery"*
    - *Profile 7 (marketing, seo, digital…): "community manager"*
    - *Profile 8 (sap, android…): "consultor sap"*
    - *Profile 9 (sharepoint, crm, dynamics…): "microsoft dynamics"*

- A-posteriori n-grams for LDA Modelling with 20 profiles of the InfoJobs dataset (see also Table 4):
    - *Profile 1 (level, English, advanced…): "spoken written"*
    - *Profile 3 (equipment, communication, management…): "customer oriented"*
    - *Profile 5 (html5, web, javascript…): "html5 css3", "html css", "javascript jquery"*
    - *Profile 7 (sap, abap, management…): "project management", "sap consultant", "project manager"*
    - *Profile 9 (networks, telecommunications, cisco…): "tcp ip", "telecommunications engineer"*
    - *Profile 10 (business_intelligence, functional, power_center…): "reporting services"*
    - *Profile 11 (seo, social_networks, manager…): "community manager", "seo sem"*
    - *Profile 13 (cobol, db2, host…): "cobol cics db2"*
    - *Profile 14 (java, web, servers…): "application servers", "web services"*
    - *Profile 15 (pl, oracle, analysis…): "controller view model"*
    - *Profile 18 (microsoft, crm, dynamics…): "microsoft office", "dynamics nav"*

Other interesting terms which have been detected but do not appear among the 15 most relevant of the foregoing profile models are as follows: "information systems", "google adwords", "incident management", "windows xp", "data consultant", "cloud computing", "web applications", "ruby rails", "web developer", "relational_databases" or "sql server".

8.3.4 Results of the hierarchical model

In this section we display the results of the hierarchical profile extraction model. It should be noted that this strategy varies greatly among the profiles obtained in each execution. As can be observed, in certain cases the hierarchical relationships between profiles seem to make sense, while in other cases said relationships are unclear and could even seem erroneous.

The reasons for these negative results could be the following:

- As mentioned earlier, the datasets are very unbalanced and the frequency of appearance of different terms varies widely. By way of example, the term java appears in a total of 1,946 job offers (considering all the datasets jointly), while big_data appears only in 55 offers. Clearly, this disproportion biases the model towards

offering good results in those profiles that include frequently used terms, even at the expense of worsening the results affecting less present profiles in the dataset.

- In relation to the foregoing, the models used work in a completely non-supervised manner and have the ultimate objective of finding a series of profiles that could credibly have generated the documents of the dataset, which necessarily coincides with the objective of finding a set of profiles with clear semantic meaning. Therefore, the model may opt for merging different profiles scarcely represented in the database or may require the generation of "background" profiles that will enable the generation of some of the terms that appear in the dataset but are not directly associated to any intuitive profile. Said "background" profiles must necessarily be incorporated into the hierarchy of learned profiles.

- Limitations of the technology. The model used requires specifying a priori the maximum depth of the hierarchical tree and said level of depth shall be reached in all the tree branches. At times, this maximum depth causes some of the deepest profiles to scarcely be represented in the dataset. For this reason, these scarcely represented profiles were removed from the list of profiles shown below.

- Lastly, we must reflect on the suitability of a hierarchical model to contain the profile interdependencies. There are cross-cutting skills that are widely sought in nearly all job offers and others that may be required for job offers not related a priori, etc.

In this case, a preliminary analysis of the hierarchical models obtained for the Infoempleo and Tecnoempleo job portals enabled us to quickly conclude on the inferiority of said models with respect to that presented in this report for the InfoJobs portal. To the general limitations of the hierarchical model we must add the following difficulties inherent to each dataset:

- Portal Infoempleo: a low number of job offers

- Tecnoempleo portal: The use of a fairly extensive vocabulary and the inclusion of more detailed and, up to a certain point, less technical descriptions of each job offer could possibly dilute the hierarchy relationships between the technology skills sought. Consequently, the hierarchical model obtained was excessively extensive in terms of the number of identified profiles and the hierarchical relationships obtained were more than doubtful.

For the sake of compactness, only those results corresponding to the InfoJobs dataset have been presented in this report (Table 4.9); the profiles obtained for the other two job portals may be consulted in the visualisation tool developed within this project. The hierarchy of each profile listed in Table 10 is indicated within the profile tree [number 1-3] and the relative importance of said profile in the dataset, calculated as the number of documents of the dataset associated with each profile. Likewise, each profile is described by the collection of most frequent terms.

In the specific case of the InfoJobs portal, it can be observed that the requirement of reaching level three in the hierarchy in all the tree branches is fulfilled on several occasions without actually having to divide the tree

nodes. In fact, there is only one hierarchical structure for three of the four first higher profiles in order of importance. Even so, some of these successive refinements are interesting because they highlight new terms that are directly related to the higher-level profile

Table 10. Hierarchical profiling with a maximum depth of three levels for the InfoJobs dataset

- [1/748] java j2ee javascript web_services programming mvc jquery eclipse liferay analyst
  - [2/387] spring struts jboss tomcat jsp jpa junit java websphere jenkins
    - [3/387] j2ee jsf maven weblogic jee websphere spring ejb swing bpm
  - [2/361] .net visual_basic navision dynamics entity wpf wcf tfs net ado.net
    - [3/361] asp_net visual_studio sharepoint vb_net nav dynamics server webforms ui silverlight

- [1/601] marketing digital communication seo campaigns manager sale commercial community photoshop
  - [2/574] advertising graphic_design excel sem google_analytics business writing programs social_media journalism
    - [3/574] online social_networks internship blogs international public_relations creativity communication similar consumer

- [1/477] linux systems administration networks administrator windows vmware cisco servers systems_administrator
  - [2/274] unix linux aix weblogic cluster shell_script hp solaris hp_ux migration
    - [3/185] red_hat websphere jboss ibm netapp storage boxes enterprise nas murex
    - [3/89] dba 11g rac 10g dataguard 9i tuning rman shell database
  - [2/203] cisco juniper checkpoint wan voip network identities audit lopd security
    - [3/199] security ccna lan wifi iso routing digital cism vlan siem

- [1/271] javascript html5 css3 jquery ios css android html web developer
  - [2/146] design photoshop flash adobe adobe_illustrator 3d rwd web_designer formatting ui
    - [3/146] ux responsive usability bootstrap unity web_formatter web_designer editing graphic adobe_illustrator
  - [2/124] php magento mysql symfony2 linux qlikview php5 symfony django python
    - [3/124] php python zend lamp symfony django yii mysql prestashop stack

- [1/201] helpdesk telephone sale commercial networks care mobiles ftth teleoperator incidents
  - [2/201] office automation telephony huawei telephone desk excellent electronics hardware installation maintenance
    - [3/201] resources labour installation meat repair relations telemarketing vehicle prl antivirus

- [1/199] cobol db2 host banking cics payments risks partenon functional analyst
  - [2/198] banksphere bank manager financial travel business_intelligence vega analysis batch electronic
    - [3/198] tibco datawarehouse risks bank jcl weeks programming_languages programmers monitor dwh

- [1/167] business_intelligence power_center sas microstrategy consultant business_objects salesforce services data cognos
  - [2/167] business_intelligence etl visual statistics apex banking boxes business financial data_mining
    - [3/165] pl teradata foxpro business_objects modelling siebel business actuarial transform ms

- [1/143] sap abap consultant business_objects bw mm ps sd hr reports
  - [2/143] abap webdynpro ps business_intelligence dynpro senior bpc controlling material sap
    - [3/132] consultant hcm bapis solution srm project dynpro oral costs business

- [1/50] testing qa test selenium apache big_data hadoop tester center alm
  - [2/50] mapr available agile pig collections mobile databases cases hive packets
    - [3/47] toad functional quality istqb jmeter git pl hadoop xsd alm

9

ANALYSIS OF ICT TRAINING PROGRAMMES

9 Analysis of ICT training programmes

In this chapter we will show the results of the characterisation of the training offering (information captured from official university degree and vocational training websites). The results are based on the application of the profiling tool described in Section 6.3.

9.1 Data Source

DATASOURCE

UNIVERSITY DEGREES of the Register of Universities, Centres and Qualifications (**RUCT**) and **PROFESSIONAL QUALIFICATIONS** of the Spanish National Institute of Qualifications (**INCUAL**).

The analysis of the curricular offering focuses on two types of training levels:

- **University qualifications**, including degrees and master's degrees. Data will be taken from the Register of Universities, Centres and Qualifications (RUCT, https://www.educacion.gob.es/ruct/home) where most of the[1] syllabi of the different degrees and master's degrees delivered in Spanish universities can be found.

- **Professional qualifications** or Vocational Training studies. In this case, the information available on the portal of the Spanish National Institute of Qualifications shall be used (https://www.educacion.gob.es/iceextranet/).

Additionally, in both cases, we will have to limit the analysis of the curricular offers to those related to ICT, to which end we will follow the indications of the following reports:

- The National Occupational Classification provided by the INE (Spanish National Statistics Institute):
  - http://www.ine.es/jaxi/menu.do?type=pcaxis&path=/t40/cno11&file=inebase

- The ESCO classification (European Skills, Competences, Qualifications and Occupations):
  - http://ec.europa.eu/social/main.jsp?catid=1042&langid=en, https://ec.europa.eu/esco/home

- The "SUPPLY AND DEMAND OF DIGITAL CONTENT PROFESSIONALS" report
  - http://www.ontsi.red.es/ontsi/sites/default/files/anexo_oferta_y_demanda_de_profesionales_de_contenidos_digitales.pdf

- Most sought-after professional profiles in the field of Digital Content in Spain
  - http://www.fti.es/content/pafet-vii-perfiles-profesionales-m%C3%A1s-demandados-%C3%A1mbito-contenidos-digita

[1] We say "most" because, as will explain later, syllabi of some degrees or master's degrees are not available.

- Analysis of the situation and evolution of the knowledge and skills required from ICT professionals in the Electronics, Information Technology and Communications sector (PAFET 3)
    - http://www.coit.es/index.php?op=estudios_215

- "Study of the Professional Profiles and Qualifications Related to the ICT Sector" performed by the ONTSI (Spanish Observatory for Telecommunications and the Information Society) in February 2014.

Given that the information available for each type of studies has its own characteristics and is governed by specific ICT criteria, in the following subsections we will explain the information available in each case, as well as that obtained (or downloaded) after the crawling process.

9.1.1 Data source for the profiling of university qualifications

The qualifications of Degree and Master's Degree, following the classification of the official register of university qualifications, are related to five branches of knowledge: Arts and Humanities, Science, Engineering and Architecture, Legal and Social Sciences and Health Sciences. Within each branch, ICT qualifications have been considered to be those related to the following fields of study:

1. **Arts and Humanities**: Fine Arts, Design, Artistic Production, Editing, Drawing, Radio, Cinema Industry and Multimedia Translation.

2. **Science**: Mathematics, Statistics, Physics and Computing

3. **Engineering and Architecture**: Information Technology, Software, Web, Telecommunications, Systems and Information Technologies, Sound and Image.

4. **Legal and Social Sciences**: Journalism, Audiovisual Communication, Advertising, Information and Documentation, Libraries and Digital Information Services, Consulting and Information Management, Cinema and Television, Digital Information Management, Information Technologies, Design, Creativity for Advertising Communication, Cinema, Television and Interactive Media.

5. **Health Sciences**: in this case there are no ICT-related studies.

After applying the crawling or browsing process on 14 November, we were able to access the syllabus of **636** ICT qualifications, including degrees and master's degrees of the different branches of knowledge. Table 11 details the distribution of these syllabi among the different branches, indicating the total number of qualifications per branch and level (degree (G) or master's degree (M), which of these correspond to ICT and for which the syllabi are available.

DATA SOURCE (UNIVERSITY)

636

ICT Degree and Master's Degree UNIVERSITY QUALIFICATIONS.

Table 11. Distribution of the number of university qualifications (degrees (G) or master's degrees (M) offered in the different branches of knowledge. The number of qualifications belonging to ICT and the number of qualifications whose syllabus is available.

Branch of knowledge	Level	Total qualifications	ICT qualifications	Syllabus
Engineering and Architecture	G	714	257	180
	M	1164	232	130
Science	G	223	75	40
	M	670	104	45
Arts and Humanities	G	403	42	20
	M	728	60	27
Legal and Social Sciences	G	923	178	95
	M	2025	187	99

In the RUCT, the following information is available for each university qualification:

- Description of qualification
  - Basic data: level, name...
  - Distribution of credits: core, electives...
  - University and centre that offer it
- Competences
- Student entrance examination and admission process
- Career plan or syllabus
- Implementation calendar

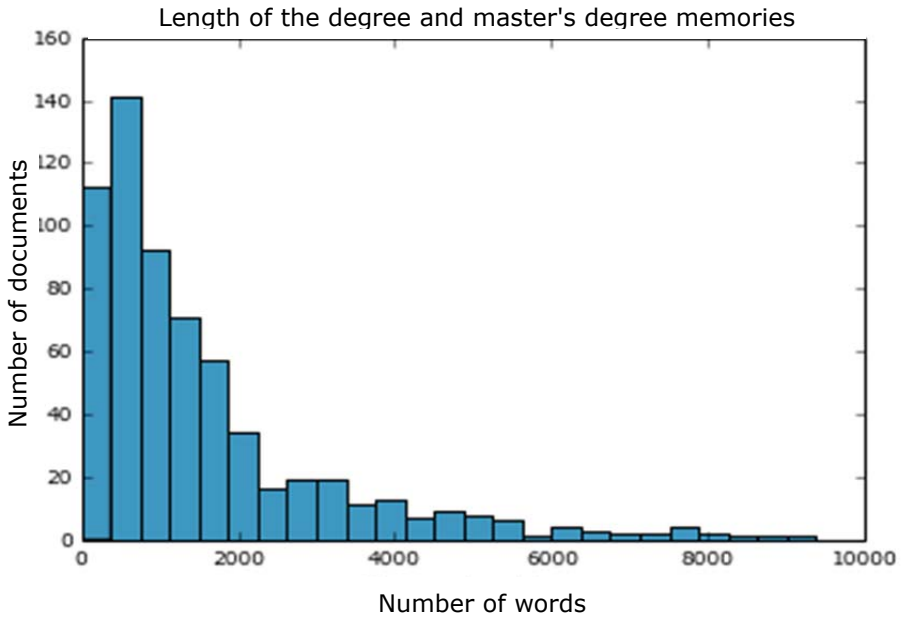
For the study we discarded information that is common among universities and different qualifications, as it does not characterise a specific degree and is therefore irrelevant to the purpose of this study. Therefore, we downloaded the information specific to a university and degree contained in the syllabus where the general and specific competences are included (in many cases broken down by subjects). The specific field of competence was not downloaded; on the one hand, to avoid redundancy with that available in the syllabus and, on the other, because this field only includes common descriptors among universities that do not characterise a specific degree of a university.

Additionally, in order to subsequently analyse this information, the download of the each syllabus was accompanied by information about the degree, university, branch of knowledge and link to the syllabus. It should be noted that, despite, being included in the official register of qualifications, the syllabus of many qualifications is not available, due to which we were obliged

to reduce this study only to those qualifications which could be accessed via this portal.

The analysis of the documents revealed that they have an average length of 1,553 words, with a typical deviation of 1,604 words (see Figure 34, which shows the distribution of the number of words per document).

Figure 34. Histogram of the distribution of the number of words per document in dataset of university qualifications



On generating the dataset with these documents, we obtained a vocabulary of **10,707 terms** or words. Of which, additionally, we observed that only 531 terms overlapped with the InfoJobs vocabulary and 1,261 terms overlapped with the vocabulary of the union of the three job portals.

Lastly, it should be noted that, in order to build the vocabulary, we iterated using the profile obtainment process on five occasions, so as to increase the list of stopwords in each iteration by adding cutoff terms merely in the academic rather than defining sense of the profiles sought (e.g.: University, university, school, schule, Erasmus, Seneca, Socrates, etc.). Additionally, we took advantage of this process to clean some "garbage" terms from the conversion of pdf to text.

9.1.2 Data source for profiling professional qualifications

The web portal of the Spanish National Qualifications Institute (INCUAL) includes a total of 12 professional families, of which the following three were selected (mainly following the recommendations of the ONTSI report) as belonging to the digital content sector:

DATA SOURCE (PROFESSIONAL TRAINING)

72

PROFESSIONAL QUALIFICATIONS.

1. Graphic Arts
2. Image and Sound
3. Information Technology and Communications

Within each of these families, the INCUAL web portal provides a list of the professional qualifications of each family, together with a pdf document including comprehensive information on each qualification. In particular, it contains:

- The general competence.
- The units of competence that comprise it. Additionally, it details its professional outlets and professional context of each unit.
- The professional sphere of the qualification.
- The associated training, indicating the training modules that comprise it. Also, the capabilities and evaluation criteria, contents and capabilities to be acquired by the students to be developed (or completed) in a professional environment are detailed for each training module.

As this information is very well structured for each professional qualification and available for each, the browsing or crawling of this website was focused on downloading this content. It should be noted, however, that each of these documents was accompanied by the name of the qualification, professional family to which it belongs and level of qualification assigned thereto in order to facilitate subsequent study.

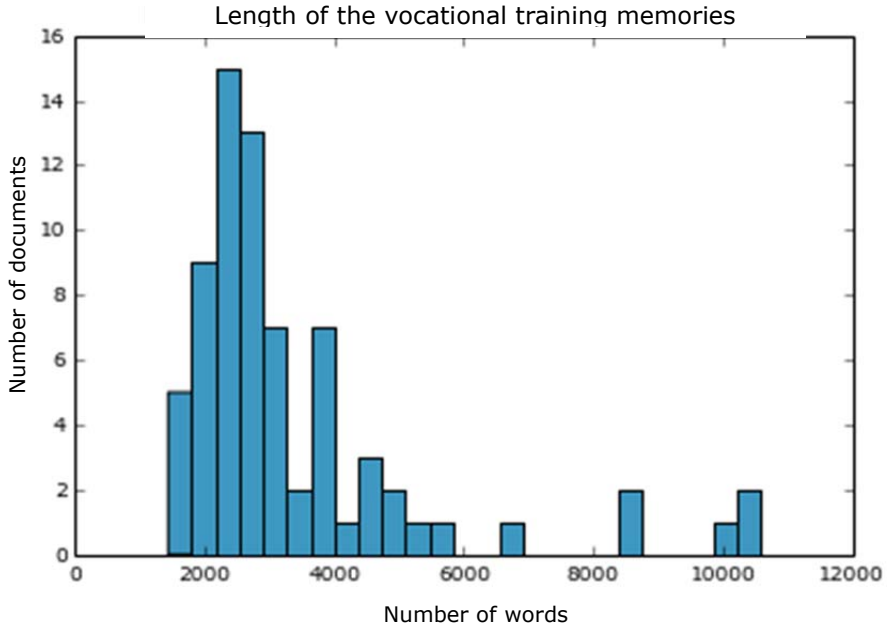
After launching the search process on 14 November 2014, the following were downloaded:

- **72** professional qualifications, distributed as follows:
  - 31 Graphic Arts qualifications,
  - 18 Image and Sound qualifications and
  - 23 Information Technology and Communications qualifications.

In this case, it was observed that, despite working with a smaller number of documents, their length tended to be much greater (there are no excessively short memories), with an average length of 3,402 words and a typical deviation of 1,998.16 words.

shows the distribution of the lengths of the syllabi.

Figure 35. Histogram of the distribution of the number of words per document in the dataset of professional qualifications



On creating the dataset, we obtained a vocabulary of 2,883 terms, of which only 204 overlapped with the vocabulary of InfoJobs, while there are 521 terms in common with the vocabulary of the union of the three job portals.

In this case, as in that of the university qualifications, an iterative profile building process was performed in order to add stopwords, thereby removing general terms from the dataset of little use in the definition of the profiles.

9.2 Purpose of the study and test design

The purpose of the profiling experiments conducted on the dataset is similar to that conducted on the job offers. Once again, the aim is to analyse the following points:

- Visualisation of the profiles obtained in accordance with the number of profiles selected.

- Analysis of the profiles obtained for university qualifications compared to professional qualifications.

- Analysis of the results of hierarchical modelling.

Ten different models were trained for each specific experiment to conduct the tests. Once trained, the models included in this report were manually selected.

9.3 Study results

9.3.1 Selection of the number of profiles

In this subsection we analysed the impact of the selection of the number of profiles, which must be performed a priori, on the profiles obtained. To this end, three models with 10, 15 and 20 profiles were trained per dataset.

If we start by analysing the impact of this parameter on the dataset of university qualifications (see Table 12, Table 13 and Table 14) and take into account that the profiling model does not group together similar documents, but rather its objective is to explain the appearance of all the words of each of the documents of the dataset, we can observe the following:

Table 12. Characterisation of the profiles of the training offering for university qualifications (10 profiles).

Topic 0 (15.2%)	Topic 1 (14.16%)	Topic 2 (12.99%)	Topic 3 (12.64%)	Topic 4 (11.45%)
software	journalism	engineering	sem	audiovisual
computers	marketing	industrial	audiovisual	audiovisual
computing	online	innovation	artistic	marketing
intelligent	advertising	electronics	fine_arts	fiction
security	journalistic	s_coop	multimedia	documentary
operating_systems	journalistic	industrial	mb	advertising
computing	audiovisual	communications	jorge	advertising
intelligence	advertising	technology	artistic	cinema
artificial	telephony	engineering	san	journalism
algorithms	messages	marine	journalism	television
Topic 5 (7.71%)	Topic 6 (7.21%)	Topic 7 (6.93%)	Topic 8 (6.27%)	Topic 9 (5.41%)
communications	biotechnology	audiovisual	industrial	mathematics
circuits	physics	video games	electronics	equations
electronic	biology	cinema	automatic	mathematic
signals	molecular	animation	electric	geometry
telecommunications	genetics	visual	industrial	algebra
telematics	chemistry	photography	automation	calculus
electronics	biotechnological	artistic	chemistry	theorem
signal	cellular	dra	machines	computer
processing	biochemistry	post-production	instrumentation	topology
telematic	animal	image	electronic	numerical

- The profiles obtained belong to different branches of ICT training or specialities, although some of them are characterised by more than

one profile. Therefore, with 10 profiles (Table 12) we can observe the following branches:

- Information Technology: profile 0
- Communication: profiles 1 and 4
- Electronics: profiles 2 and 8
- Artistic production: profile 3
- Telecommunications: profile 5
- Biotechnology, Physics and Chemistry: profile 6
- Audiovisual: profile 7
- Mathematics: profile 9

Table 13. Characterisation of the profiles of the training offering for university qualifications (15 profiles).

Topic 0 (14.9%)	Topic 1 (13.38%)	Topic 2 (10.93%)	Topic 3 (8.79%)	Topic 4 (6.99%)
software	electronics	artistic	audiovisual	mathematics
computers	electronic	marketing	audiovisual	physics
computing	physics	fine_arts	cinema	equations
security	industrial	dra	multimedia	mathematic
intelligent	communications	innovation	fiction	calculus
operating_systems	circuits	artistic	cinema	algebra
computing	mathematics	researcher	post-production	geometry
intelligence	instrumentation	strategic	television	numerical
artificial	automatic	campaigns	indiv	computer
distributed	technology	business	documentary	theorem
Topic 5 (5.71%)	Topic 6 (5.63%)	Topic 7 (5.61%)	Topic 8 (5.24%)	Topic 9 (5.01%)
journalism	industrial	communications	biotechnology	sem
online	electronics	circuits	biology	marketing
chat	physics	telecommunications	molecular	mb
compulsory	industrial	signals	genetics	recovery
method	electric	telematics	biotechnological	journalism
blackboard	mechanics	electronic	cellular	online
videoconference	instrumentation	processing	animal	jorge
blogs	automatic	signal	chemistry	audiovisual
virtual	chemistry	audio	biochemistry	search engines
collaborative	automation	telematic	vegetable	multimedia
Topic 10 (4.47%)	Topic 11 (4.16%)	Topic 12 (3.79%)	Topic 13 (3.7%)	Topic 14 (1.7%)
advertising	online	journalism	uveg	telephony
marketing	marketing	journalistic	free	mailbox
advertising	video games	journalistic	etse	asynchronous
audiovisual	uca.es	audiovisual	continued	synchronous
messages	authority	television	voluntary	wikis
workshop	marine	informative	classroom	moodle
Church	antonio	press	review	blogs
ds	marine	journalist	discussions	glossaries
synthesis	sum	radio	certification	differ
strategic	oe	audiovisual	video games	video

Table 14. Characterisation of the profiles of the training offering for university qualifications (20 profiles).

Topic 0 (14.9%)	Topic 1 (13.38%)	Topic 2 (10.93%)	Topic 3 (8.79%)	Topic 4 (6.99%)
software	electronics	artistic	audiovisual	mathematics
computers	electronic	marketing	audiovisual	physics
computing	physics	fine_arts	cinema	equations
security	industrial	dra	multimedia	mathematic
intelligent	communications	innovation	fiction	calculus
operating_systems	circuits	artistic	cinema	algebra
computing	mathematics	researcher	post-production	geometry
intelligence	instrumentation	strategic	television	numerical
artificial	automatic	campaigns	indiv	computer
distributed	technology	business	documentary	theorem
Topic 5 (5.71%)	Topic 6 (5.63%)	Topic 7 (5.61%)	Topic 8 (5.24%)	Topic 9 (5.01%)
journalism	industrial	communications	biotechnology	sem
online	electronics	circuits	biology	marketing
chat	physics	telecommunications	molecular	mb
compulsory	industrial	signals	genetics	recovery
method	electric	telematics	biotechnological	journalism
blackboard	mechanics	electronic	cellular	online
videoconference	instrumentation	processing	animal	jorge
blogs	automatic	signal	chemistry	audiovisual
virtual	chemistry	audio	biochemistry	search engines
collaborative	automation	telematic	vegetable	multimedia

Topic 10 (4.62%)	Topic 11 (3.82%)	Topic 12 (2.83%)	Topic 13 (2.75%)	Topic 14 (2.36%)
biotechnology	electronics	video games	communications	telephony
biology	industrial	e_mail	computing	mailbox
molecular	electronic	uveg	characteristics	asynchronous
genetics	automatic	compulsory	automatic	synchronous
biotechnological	power	discussions	computers	wikis
cellular	instrumentation	classroom	umh	moodle
chemistry	circuits	complete	automatic	blogs
biochemistry	robotics	continued	cap	glossaries
animal	IM	review	signals	adapted
vegetable	machines	certification	optimisation	video
Topic 15 (2.15%)	Topic 16 (1.9%)	Topic 17 (1.86%)	Topic 18 (1.31%)	Topic 19 (0.86%)
security	sound	marketing	graphic	sem
laser	audio	group	industrial	transparency
comment	electronic	strategic	cuatr	usp
texts	processing	brand	subj	arguments
question	image	mu	physics	multimedia
inter-university	circuits	management	drawing	anthropology
indicators	signals	strategic	workshop	conception
orally	telematics	reputation	biophysics	undertake
plasma	video	branding	manufacturing	audit

- On increasing the number of profiles (Table 13 and Table 14), we can obtain an equivalent classification. Although the branches are characterised by a greater number of profiles, a series of profiles appear (mainly, among the last ones) that are difficult to associate with a specific ICT branch; this is the case of profiles 15 and 19 in the model with 15 profiles or profiles 12, 15, 18 and 19 in model 20. This behaviour is due to the compromise between the degree of granularity of the analysis (i.e. the accuracy and definition of the profiles that can be obtained) and the appearance of profiles with a low semantic content.

- On the contrary, the increase in the number of profiles enables the identification of profiles potentially more defined and with a smaller representation in the dataset. This is the case of profile 14 (in the model with 15 and 20 profiles) the most relevant terms of which are "telephony", "asynchronous", "synchronous" and enable the obtainment of a very specific profile of the field of telecommunications that (model with 10 profiles) was not present before.

If we proceed to analyse the profiles obtained from the dataset of professional qualifications with models 10, 15 and 20 profiles (see Table 15, Table 16 and Table 17), we can draw fairly similar conclusions:

Table 15. Characterisation of the profiles of the training offering for vocational training (10 profiles).

Topic 0 (15.89%)	Topic 1 (14.16%)	Topic 2 (14.1%)	Topic 3 (13.3%)	Topic 4 (12.3%)
microcomputing	multimedia	sound	materials	binding
operating_system	servers	audiovisual	ink	transform
interconnection	web	post-production	cylinder	pre-printing
connection	interactive	engraving	cardboard	industrial
monitoring	pages	video	edition	covers
access	component	musical	stationery	cardboard
local	messaging	television	defects	queues
diagnosis	files	audio	wavy	industry
operating_systems	relational	resonant	inking	manufacturing
malfunctions	queries	cameras	printers	ink
Topic 5 (8.44%)	Topic 6 (7.68%)	Topic 7 (6.18%)	Topic 8 (5.8%)	Topic 9 (2.68%)
stamping	photographic	publisher	shows	packaging
work	digitisation	fixed	live	bottle
illustrations	processing	mobiles	animation	structural
matrices	photographic	wireless	event	complex
sketches	formatting	managers	artistic	container
matrix	films	warehouse	musical	die cut
engraving	vector	work	room	manufacturing
print layouts	developing	extraction	collective	prototypes
artistic	laboratory	repositories	list	split
screens	chemical	author	theatres	dies

- With few profiles (models with 10 profiles, Table 15) it is very easy to assign an ICT speciality or branch to each profile. Thus, for example, we can define the following branches:
  - Computer systems: profile 0
  - Web services, multimedia and databases: profile 1
  - Audiovisual production: profile 2 and 6
  - Printing systems: profiles 3 and 4
  - Digital processing: profile 5
  - Shows and events: profile 8
  - Manufacturing: profile 9

  where, in this case, profile 7 remains between the branches of mobile and wireless communications and editing; given that these two branches are not present among the foregoing, this suggests that a greater granularity is required to define all the branches present in the dataset.

- On increasing the number of profiles (see the model with 15 profiles of Table 16), we can observe that we can define the following branches:
    - Mobile communication systems: profile 0
    - Computer systems: profile 1
    - Web services and multimedia: profile 2
    - Audiovisual production: profile 3 and 6
    - Printing systems: profiles 4, 5, 6, 7, 10 and 14
    - Digital processing: profile 8
    - Editing: profile 12
    - Shows and events: profile 9 and 11
    - Manufacturing: profile 13

  Appearing, precisely, two new branches (mobile communications systems and editing) that were previously merged by profile 7.

- On analysing the model with 20 profiles (Table 17) we can observe how it consists only of 17 profiles, as the three last profiles have a weight of 0.01% and, moreover, are identical. If we take into account that, in this case, we are managing a dataset of 72 documents, it seems evident that 17 profiles are more than enough for their characterisation.

We can therefore conclude that the selection of the number of profiles is not a trivial task and requires a certain degree of supervision. In both cases, after analysing the behaviour of the models obtained with the number of profiles used, it seems that the use of 10 profiles is more than enough to obtain a definition of the different existing training branches and we avoid the appearance of residual profiles that are very difficult to assign to a specific branch.

Table 16. Characterisation of the profiles of the training offering for vocational training (15 profiles).

Topic 0 (13%)	Topic 1 (12.37%)	Topic 2 (8.45%)	Topic 3 (8.42%)	Topic 4 (7.95%)
interconnection	operating_system	multimedia	audiovisual	ink
mobiles	microcomputing	web	cameras	inking
local	operating_systems	servers	post-production	cylinder
fixed	access	animation	television	print
managers	connection	pages	engraving	materials
wireless	servers	video	sound	edition
connection	copies	interactive	video	drying
monitoring	enumerate	room	filming	ink
foreign	monitoring	usability	cinema	rotogravure printing
warehouse	logical	messaging	cinema	pre-printing
Topic 5 (7.28%)	Topic 6 (6.46%)	Topic 7 (6.04%)	Topic 8 (5.68%)	Topic 9 (5.61%)
covers	transform	stamping	photographic	sound
binding	cardboard	work	photographic	musical
materials	wavy	matrices	vector	shows
sheets	bottle	sketches	digitisation	resonant
folded	stationery	matrix	films	mixture
print layouts	packaging	engraving	processing	audio
printers	complex	images	formatting	microphones
stamping	queues	screens	developing	live
hard	manufacturing	stamp	laboratory	engraving
imposition	piling	print layouts	chemical	radio
Topic 10 (5.06%)	Topic 11 (4.58%)	Topic 12 (4.29%)	Topic 13 (2.41%)	Topic 14 (2.4%)
pre-printing	shows	publisher	packaging	binding
transform	live	animation	bottle	industrial
ink	illustrations	work	structural	book
manufacturing	artistic	author	container	covers
preventive	event	rights	prototypes	artistic
calibration	collective	book	manufacturing	containers
emergency	scene	sketches	mockups	spine
suppliers	artistic	formatting	arts	construction
recommendations	list	commission	logistics	hides
queues	theatres	typesetting	cardboard	adhesives

Table 17. Characterisation of the profiles of the training offering for vocational training (20 profiles).

Topic 0 (11.97%)	Topic 1 (11.51%)	Topic 2 (9.47%)	Topic 3 (8.47%)	Topic 4 (6.95%)
interconnection	shows	ink	managers	covers
monitoring	audiovisual	cylinder	operating_system	binding
access	live	inking	servers	materials
connection	artistic	handled	copies	sheets
local	cameras	piling	queries	folded
infrastructure	television	ink	relational	stamping
inventories	event	industry	enumerate	hard
foreign	scene	print	operating_systems	bind
diagnosis	video	materials	files	pre-printing
reports	engraving	palletising	repositories	stitched
Topic 5 (6.77%)	Topic 6 (6.74%)	Topic 7 (6.26%)	Topic 8 (5.32%)	Topic 9 (5.18%)
photographic	servers	transform	multimedia	sound
vector	microcomputing	pre-printing	illustrations	audio
films	web	cardboard	publisher	resonant
photographic	component	manufacturing	interactive	post-production
processing	operating_system	ink	prototypes	microphones
digitisation	operating_systems	industrial	sketches	engraving
formatting	messaging	industry	publications	voicing
laboratory	connection	preventive	screens	resonant
developing	electronics	suppliers	audiovisual	shows
chemical	diagnosis	emergency	author	mixture
Topic 10 (4.29%)	Topic 11 (3.32%)	Topic 12 (3.28%)	Topic 13 (3.25%)	Topic 14 (2.98%)
publisher	packaging	work	wavy	screens
work	bottle	matrices	mobiles	print layouts
stamping	stationery	stamping	fixed	screen printing
engraving	complex	matrix	cardboard	printers
matrix	pages	preservation	wireless	warehouse
sketches	cardboard	sketches	local	stamping
lithographs	manufacturing	formal	trains	screen printing
author	superficial	images	radio	photopolymers
inking	structural	ink	queues	imposition
rights	domestic	drawing	coordinate	separation
Topic 15 (2.46%)	Topic 16 (1.74%)	Topic 17 (0.01%)	Topic 18 (0.01%)	Topic 19 (0.01%)
binding	animation	frame	frame	frame
industrial	musical	relay	relay	relay
book	live	conceal	conceal	conceal
covers	room	retouching	retouching	retouching
artistic	video	taking them	taking them	taking them
containers	radio	bezier	bezier	bezier
spine	sound	producer	producer	producer
hides	emitting	microscopy	microscopy	microscopy
construction	mixture	emulsification	emulsification	emulsification
bind	music	positions	positions	positions

9.3.2 Profiles obtained for the different training plans

This section includes a description of the profiles obtained for both datasets: university training plans (degrees and master's degrees) and professional qualifications. To this end, we will use the profiling obtained previously with a number of 10 profiles as a basis, as we observed in the previous section that this value provides a reasonable balance between semantic value of the profiles obtained and prevented the appearance of irrelevant and noisy profiles.

To perform this analysis, for each profile, the following is included:

- The characteristic words or terms.
- A brief description that will attempt to characterise each profile with one or two terms.
- The branch of knowledge or professional family, according to the type of training plan considered, to which the profile can be associated.
- A list with the training plans that are best characterised with said profile. It should be noted that this document list is usually fairly extensive, due to which a brief summary is presented that includes the most relevant documents. However, it is possible to analyse all of these documents in detail using the visualisation tool. Moreover, in the case of university qualifications, for the sake of brevity only the name of the degree is included, but the name of the degree and the university that offers it can be obtained from the visualisation tool.

Table 18 and

Table **19** show this information for the university qualifications and the professional qualifications, respectively.

In light of the results, we can identify eight ICT specialities in the case of university qualifications, each within a branch of knowledge, and all represented by one or, at most, two profiles of the model obtained. It should also be noted that the most relevant profiles of the model are information technology and communication.

In the case of professional qualifications, a total of nine specialities have been identified, all (except digital processing) associated with a professional family and represented by one or two profiles of the model. Once again, the profiles with weight or relevance within the model are those associated to information technology (in this case, the model has allowed us to divide it in two: (1) computer systems; (2) web, multimedia and databases) and communications (once again, divided in two: (1) audiovisual production; (2) printing systems).

Table 18. Characteristics profiles of the university training offering and university qualifications best represented

Description	Competences (keywords)	Profile no.	Branch of knowledge	University qualifications best represented
Computer Science	Software, computers, computing, security, intelligent, operating_systems, computer, intelligence, artificial, distributed…	0	Engineering and Architecture	D. in Computer Engineering; D. and M.D. in Computer Science; D. in Computer Engineering; M.D. in Data Science and Computer Engineering; M.D. in Distributed and Embedded Systems Software; …
Communication	Journalism, marketing, online, advertising, journalistic, audiovisual, documentary, television, fiction…	1, 4	Legal and Social Sciences	D. and M.D. in Journalism; D. in Advertising and Public Relations; M.D. in Advertising Management; D. in Marketing and Commercial Communication; D. in Audiovisual Communication; …
Electronics	Engineering, innovation, industrial, electronics, industrial, technology, engineering, automatic, electric, automation, machines…	2, 8	Engineering and Architecture	M.D. in ICT Management; M.D. in Telecommunications Engineering; M.D. in Automated Systems and Industrial Electronics; D. in Electronics Engineering; M.D. in Computer Engineering; D. in Industrial and Automated Electronics Engineering; …
Artistic Production	Artistic, fine_arts, image audiovisual, multimedia…	3	Arts and Humanities	D. in Translation and Intercultural Communication; D. in Fine Arts, M.D. in Artistic Production; M.D. in Design Engineering; …
Telecommunica-tions	Communications, circuits, electronic, signal(s), telecommunications, telematics, electronics, processing…	5	Engineering and Architecture	D. and M.D. in Telecommunications Technologies, M.D. in Computer Engineering and Telecommunications; D. in Telematics; D. in Communications Systems, D. in Electronic Systems; ...
Biotechnology, physics, chemistry	Biotechnology, physics, biology, molecular, genetics, chemistry, cell…	6	Science	D. in Biotechnology; D. in Physics; …
Audiovisual	Audiovisual, videogames, cinema, animation, visual, photography, artistic, post-production, image…	7	Arts and Humanities	D. in Cinema; D. in Multimedia and Graphic Design; D. in Design (and audiovisual post-production); D. in Cinematography and Audiovisual Arts; …
Mathematics	Mathematics, equations, geometry, algebra, calculus, theorem, topology, numerical…	9	Science	D. and M.D. in Mathematics. D. in Statistical Techniques; ...

Table 19. Characteristics profiles of the training offering in vocational training and qualifications best represented

Description	Competences (keywords)	Profile No.	V.T.	Professional qualifications best represented
Computer systems	Microcomputing, interconnection, operating_system(s), connection, monitoring, diagnosis, malfunctions…	0	IFC	Database administration and design; network operation; implementation and management of computing elements; operation of voice and data communications systems; operation of microcomputing systems…
Web, multimedia and databases	Multimedia, servers, web, interactive, pages, files, relational, queries…	1	IFC	Programming with object-oriented languages and databases; internet services administration; creation and publication of web pages; database administration; development of multimedia audiovisual products; ...
Audiovisual production	Sound, audiovisual, recording, cameras, video, post-production, television, audio, photographic, digitisation, processing…	2 and 6	IMS	Computer-aided television production; computer-aided direction/production of films and audiovisual works; image laboratory production operations; sound operations; audiovisual montage and post-production…
Printing systems	Ink, printers, defects, edition, stationery, ink, graphic_arts, stamping, pre-printing, binding, industrial, covers, adhesives, stamping, work, artistic, engraving,…	3 and 4	ARG	Flexographic printing, rotogravure printing, screen printing and tampon printing, offset printing,…; ancillary operations in graphics industries; processing and formatting of graphic elements in pre-printing; production management in pre-printing processes, lithography, engraving and stamping techniques, artistic screen printing, stamping…
Digital processing	Photographic, processing, films, digitisation, formatting, laboratory, copies, mockups…	5	IMS-ARG	Image laboratory production operations; photographic production; processing and formatting of graphic elements; graphic product design; …
Shows and events	Show, live, artistic, animation, event, musical, venue…	8	IMS	Computer-aided production of live shows and events; lighting for live shows; musical and visual animation aired live; 2D and 3D animation; ….
Manufacturing	Packaging, containers, structural, complexes, manufacturing, die cutting, container, prototypes, mockups…	9	ARG	Structural design of containers and packaging (paper, cardboard and other graphic media); die cutting; manufacture of complexes, containers, packaging and other paper and cardboard articles; graphic product design; production management of processed paper, cardboard and other graphic media,…
Mobile communications	Mobiles, landlines, cordless, …	7	IFC	Maintenance of first/second level in radiocommunications systems; …
Publishing	Publisher, warehouse, managers, work, book, author, rights, publications…	7	ARG	Computer-aided editing; development of multimedia publishing products; publishing production; …

It should be noted that, despite the fact that the objective of the profiling tool is to explain the appearance of all the words of each of the documents of the dataset, after this analysis the classification or grouping of the different training plans was achieved. This fact or added advantage is due mainly to the supervised selection of the number of profiles (or degree of

granularity of the tool), which prevented the appearance of noisy profiles that could undermine and hinder said grouping.

9.3.3 Hierarchical profiles

This section analyses the results obtained from applying hierarchical profiling to the dataset of university qualification and professional qualifications, respectively. It should be noted that, as in the case of this type of profiling applied to the job portals, the hierarchical relationships between the profiles obtained are in many cases unclear and could even seem erroneous.

Thus, for example, in the case of university qualifications, the first hierarchical level provides 51 profiles, among which we find such well-defined profiles as the following:

- Software, computers, computer, operating_systems, artificial, intelligence, mathematics, security, distributed.
- Physics, experimental, optics, instrumentation, quantum, mechanics, electromagnetism, measurement, waves, mathematics.
- Audiovisual, domain, television, audiovisual, television, radio, economy, journalistic, television, audiences.

But there are also many noisy profiles, such as:

- International, hotel, roches, agreement, education, laureate, management, educational, learning, industrial.
- Dr, videogames, usal, diagnosis, alterations, disorders, garcia, maria, fac, udc.
- San, mb, jorge, journalism, audiovisual, degree, advertising, intercultural, public_relations, complete.

In fact, as we advance along a branch, the number of profiles generated grows exponentially (hindering its analysis) and, what is worse, even if the starting point is a well-defined profile, the semantic component of the profiles is immediately lost. Thus, for example, starting off from the first of the aforementioned profiles and advancing along its first branch to level three, we find completely undefined profiles:

- Software, computers, computing, computer, operating_systems, artificial, intelligence, mathematics, security, distributed
    - Computers, software, computing, electronics, distributed, convalidations, formal, robots, recommendations, computational
        - Intelligence, artificial, installations, video, audit, reasoning, transverse, memories, burgos, graphic
        - Method, exhibit, unob, circuits, cooperative, oriented, skills, instrumental, observation, option
        - Deanery, will form, sport, etsi, expressly, mediate, act, advocate, volunteering, run

- Mathematics, uveg, topology, geometry, mathematical, finance, modelling, algebra, optimisation, discreet
- Christian, theology, instrumental, Cristiana, theology, instrumental, identity, faith, religious, univ, cervantes, acad, integral
- Mechanics, chemistry, electric, agriculture, federico, napoli, delle, marche, Cordoba, universities

If we analyse the profiles obtained for the professional qualifications, the behaviour is completely the opposite; on the first level of the tree there are 21 profiles and this number barely grows on advancing up the tree: 26 profiles of depth 2 and 29 profiles of depth 3. That is, as we go deeper into the tree, there is no division of profiles, due to which the hierarchical modelling could be reduced to that of level 1. If we analyse the type of profiles obtained at that depth, we can see that most of the profiles are well defined, as in the following cases:

- Binding, covers, materials, stamping, folding, hard, edition, defects, sheets, bind
- Audiovisual, television, work, filming, cinema, budget, artistic, sound, recording, television.
- Stamping, engraving, matrix, sketches, lithographs, matrices, inking, stamping, work, ink.
- Transform, binding, industrial, cardboard, manufacturing, ink, glues, industry, adhesives, suppliers
- Sound, shows, audible, recording, audio, microphones, audible, sound, musical

And, although few, some noisy profiles appear, such as:

- managers, warehouse, extraction, repositories, foreign, operating_system, inherent, extract, edited, administration

This unwanted behaviour of the hierarchical models is due, as in the case of the job portals, to several causes. On the one hand, to the generation of "background" profiles arising from the merging of profiles with little weight or from the generation of some of the terms that appear in the dataset but that are not directly associated with any intuitive profile. On the other, the selection of tree depth is critical: in the case of university qualifications, a depth of more than two is more than enough (further disaggregation only gives rise to noisy profiles), while in the case of professional qualifications even a depth or one is sufficient (which suggests that the hierarchical model is not useful in this case). Lastly, we must take into account that the lack of hierarchy on the dataset of professional qualifications is an expected behaviour if we take into account that there are 72 documents and at level one the hierarchical model returns 21 profiles.

10

COMPARATIVE ANALYSIS OF SUPPLY AND DEMAND OF ICT PROFESSIONALS

10 Comparative analysis of the supply and demand of ICT professionals

The most ambitious objective of this project consisted of comparatively analysing the job and training offers using automatic profiling tools. This chapter summarises the main results of the study.

10.1 Data Source

The execution of the matching functionality of the tool requires the joint use of the original dataset and previously trained profile models. Therefore, the data used in this analysis were:

- Dataset of the job portals, described in Section 8.1
- Dataset of the training offering, described in Section 9.1
- Profile models for the datasets of job offers and training offering.

In order to help us to draw conclusions, the analysis has been limited to the datasets that provided the best results in their respective tasks. Specifically, we will focus on the analysis of the alignment between the following datasets and models:

- InfoJobs dataset, together with its model with 20 profiles
- Degree and Master's Degree qualifications dataset, together with its model with 15 profiles.

10.2 Purpose of the study and test design

The application implements two different matching strategies. The primary purpose of this study will be to compare the results of the two strategies and estimate their performance level in general. More specific considerations as to the viability of the matching task will be made in the next chapter.

Both matching strategies are based on establishing a similarity measurement between the two datasets. However, it should be noted that said similarity should be based on the semantics of the documents, due to which it is not possible to directly use distances between the bags of words of both documents. In fact, the profile models are broadly used in information retrieval tasks for obtaining said semantic similarity measures. Specifically, the distance between two documents can be obtained by measuring the distances between the distributions of the documents across the profile spaces. Our implementations are based on said concept and build up the similarity between documents based on the joint profiling of the documents of both datasets. The difference between the two strategies the results of which are being discussed herein lies in the vocabulary used for said joint profiling, having implemented the two following options:

- Strategy 1: Vocabulary is considered to be the intersection of the terms that comprise the vocabularies of the two datasets.

- Strategy 2: Vocabulary is considered to be the set of most relevant terms for the job offer profiles.

Therefore, the second strategy attempts to force the use of a vocabulary as close as possible to the demand for professionals. Once the common profiling has been obtained, the similarity measurement is based on the use of a probabilistic-type divergence known as the Hellinger divergence.

Once we know the similarity between documents, the profile models inherent to the each dataset is used to calculate the cross-lingual similarities between the profiles and documents of different datasets. For example, to calculate the distance between a job profile and a degree memory, we make a weighted average of the similarities between said document and all the job offers, using the degrees of pertainment of the job offers to the profile considered as weightings. Similarly, we can generalise the obtainment of a similarity matrix between crossed profiles.

In addition to estimating all the aforementioned similarities, the application generates rankings of the most relevant documents for each profile, and of the most important profiles for each document. Lastly, in order to analyse the quality of the similarity measures, the following information for each profile of the job dataset is displayed on screen:

- Most relevant words of each profile

- Average number of appearances of said words in the qualifications dataset

- Number of appearances of said words in the qualifications document, selected as the most relevant to said profile.

The results presented in this section are aimed at:

- Comparing the quality of the similarity measures obtained with the two matching strategies

- Comparing the results of using different datasets

10.3 Study results

10.3.1 Analysis of the convenience of restricting the vocabulary

Firstly, we analysed the results obtained on using the two matching strategies implemented in the application. The Table 20 sample, for the ten most important profiles extracted from the InfoJobs dataset, the average number of appearances of the most relevant terms of the profile in the degree and master's degree dataset, and the number of appearances of the first qualifications associated with each profile, according to the ranking

obtained of the similarity measurement on using the two strategies considered.

The results obtained allowed us to draw the following conclusions:

- The number of appearances of the most relevant terms that characterise the job offer is very low on analysing the composition of the qualification memories. In fact, 17 of the 60 terms included in Table 20 do not appear at all in the Degree and Master's Degree dataset. Also, it is precisely some of the most defining terms that do not appear at all (e.g.: j2ee, sprint, struts, css3, asp_net, vmware). Others appear, but with a very low average number of appearances (oracle, php, linux, data_bases, .net).

- On using the first matching strategy, the documents obtained as the most relevant apparently bear no relation to the composition of the data profiles themselves. In fact, in most cases the number of appearances of the most relevant terms is clearly below the average for the full dataset. This is probably due to the fact that said documents overlap with the descriptions of each profile, but for words of little relevance. The very nature of the similarity measurement used allows documents with many words to reach a high alignment estimate. Normalising the frequency of appearance of the words by document length does not solve the problem, as in said case short documents would be highlighted and would appear in the first positions, even if they have a very low number of most relevant terms.

- The results improve clearly on using the second similarity measurement strategy. With said strategy, the frequency of appearance of terms of the selected documents is usually higher than the dataset average and, occasionally, much higher. However, the results may continue to be occasionally poor in certain profiles, particularly in the case of the "web design" profile. We can find a possible explanation to said deficiency on analysing the composition of the documents extracted for profiles seven and nine, which are included in the table ("linux, oracle..." and "windows, systems…"). The very frequent appearance of the term "administration" is probably the cause of the selection of the document selected by strategy 2. It should also be noted that the result may probably be particularly serious, as the term "administration" is possibly used in the selected documents with a semantic meaning very different to the meaning of said term in the profiles that characterise the job offer.

Table 20. Results of the alignment of similarities between the demand for ICT professionals and the training offering. For the ten first profiles obtained for InfoJobs, the average number of appearances of the most relevant terms in the Degree and Master's Degree dataset is indicated, as well as the number of appearances of said terms in the documents selected as being most relevant, using the two matching strategies.

	java	j2ee	spring	struts	programming	oracle
Aver.	0.10	0.00	0.00	0.00	7.67	0.01
Strat. 1	0	0	0	0	0	0
Strat. 2	4	0	0	0	2	0
	English	level	advanced	written	engineer	soken
Aver.	4.53	0.00	0.73	0.00	20.53	0.03
Strat. 1	6	0	0	0	4	0
Strat. 2	9	0	0	0	0	0
	.net	php	programming	server	asp_net	JavaScript
Aver.	0.003	0.03	7.67	0.00	0.00	0.02
Strat. 1	0	0	0	0	0	0
Strat. 2	1	0	1	0	0	0
	team	management	Communica-tion	customer	commercial	Orientation
Aver.	5.09	0.02	0.03	0.65	0.78	2.52
Strat. 1	0	0	0	0	0	0
Strat. 2	35	0	0	1	1	2
	marketing	communica-tion	advertising	English	Online	Testing
Aver.	4.05	0.03	0.01	4.53	1.80	0.01
Strat. 1	0	0	0	6	0	0
Strat. 2	106	0	0	10	18	0
	web	design	html5	javascript	html	css3
Aver.	5.31	17.64	0.00	0.02	0.11	0.00
Strat. 1	0	5	0	0	0	0
Strat. 2	0	5	0	0	0	0
	Linux	oracle	unix	systems	data_bases	administration
Aver.	0.03	0.01	0.03	0.00	0.02	1.78
Strat. 1	0	0	0	0	0	29
Strat. 2	0	0	0	0	0	123
	sap	projects	management	abap	consultant	Manager
Aver.	0.00	0.00	0.02	0.00	0.00	0.39
Strat. 1	0	0	0	0	0	0
Strat. 2	0	0	0	0	0	0
	Windows	systems	server	technician	vmware	administration
Aver.	0.06	0.00	0.00	2.52	0.00	1.78
Strat. 1	0	0	0	0	0	0
Strat. 2	0	0	0	3	0	123
	networks	telecommunica tions	engineer	cisco	technician	IP
Aver.	9.92	1.44	20.53	0.01	2.52	0.34
Strat. 1	4	0	0	0	0	0
Strat. 2	152	1	19	0	1	4

For information purposes, Table 21 includes a description of the documents selected in two especially significant profiles: that of community manager and that of mobile application developer; in both cases the use of the second matching seems particularly adequate, both in comparison to the other strategy and for analysing the very low frequency of appearance of the most relevant terms in the qualifications dataset.

Table 21. Analysis of the results of alignment between similarities between demand for ICT professionals and the training offering. "community manager" and "mobile application developer" profiles

	seo	manager	social_networks	digital	sem	online	community	web	google	management	google_analytics
Aver.	0.08	0.08	0.00	6.19	1.13	1.80	0.15	5.31	0.23	0.02	0.02
Strat. 1	0	0	0	2	0	0	0	0	0	0	0
Strat. 2	6	1	0	107	1	18	5	15	2	0	6
	android	¡os	mobiles	applications	telecom.	developer	sale	insurance	python	commercial	engineer
Aver.	0.04	0.02	1.69	14.27	1.44	0.01	0.16	0.00	0.00	0.78	20.53
Strat. 1	0	0	0	1	0	0	0	0	0	0	0
Strat. 2	3	2	22	11	0	0	0	0	0	0	1

10.3.2 Alignment of the vocational training offering

Table 22 illustrates the behaviour of the similarity estimation techniques on crossing the InfoJobs and vocational training datasets. In this case, the distance between the terms used to characterise the job offers and the descriptions of the vocational training offering is even more evident, as 39 of the 65 terms shown in the table do not appear at all in the Vocational Training Offering dataset. In fact, there are profiles whose five most relevant terms do not appear at all in the Vocational Training Offering dataset.

A possible explanation is that job offers in the ICT sector are mostly aimed at university graduates and vocational training graduates who wish to access these job offers can only apply for very specific profiles (e.g.: network management or web design) or require specific complementary training.

As regards the behaviour of the similarity estimation measures, we found both positive and negative examples. In all cases, the scarce presence of many of the vocabulary terms used in the description of vocational training is probably causing the joint modelling used to estimate the similarities between documents to create different profiles to model the job and training documents, making said similarity estimates unreliable.

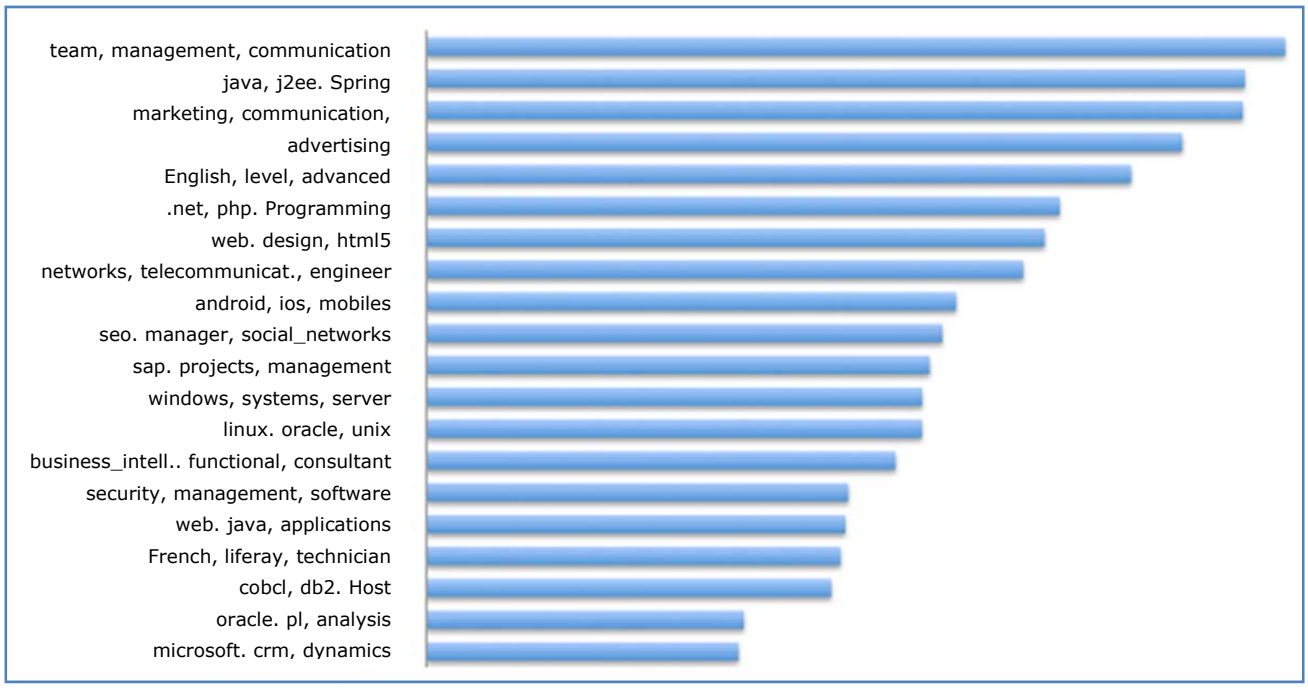
Table 22. Analysis of the results of alignment of similarities between the demand for ICT professionals (based on InfoJobs) and the vocational training offering.

	java	j2ee	spring	struts	programming
Aver. Strat. 2	0.00 0	0.00 0	0.00 0	0.00 0	5.67 2
	English	level	advanced	written	engineer
Aver. Strat. 2	0.00 0	22.35 15	0.13 3	0.00 0	2.49 1
	.net	php	programming	server	asp_net
Aver. Strat. 2	0.00 0	0.00 0	5.67 26	0.00 0	0.00 0
	team	management	communication	customer	commercial
Aver. Strat. 2	0.00 0	18.49 18	17.06 6	7.76 2	0.00 0
	marketing	communication	advertising	English	Online
Aver. Strat. 2	0.18 1	17.06 6	1.14 3	0.00 0	0.00 0
	web	design	html5	javascript	html
Aver. Strat. 2	5.26 125	13.83 41	0.00 0	0.00 0	0.00 0
	Linux	oracle	unix	systems	data_bases
Aver. Strat. 2	0.00 0	0.00 0	0.00 0	65.06 11	0.61 0
	SAP	projects	management	abap	consultant
Aver. Strat. 2	0.00 0	0.00 0	18.49 44	0.00 0	0.00 0
	Windows	systems	server	technician	vmware
Aver. Strat. 2	0.00 0	65.06 234	0.00 0	24.49 7	0.00 0
	networks	telecommunications	engineer	cisco	technician
Aver. Strat. 2	20.78 229	0.26 0	2.49 3	0.00 0	24.49 8
	businessintelligence	functional	consultant	power_center	business_objects
Aver. Strat. 2	0.00 0	0.00 0	0.00 0	0.00 0	0.00 0
	seo	manager	social_networks	digital	sem
Aver. Strat. 2	0.00 0	0.00 0	0.00 0	13.28 9	0.00 0
	android	ios	mobiles	applications	telecommunications
Aver. Strat. 2	0.00 0	0.00 0	3.32 79	12.18 2	0.26 0

10.3.3 Job offer profile rankings

By averaging the estimated similarity between each job profile and each qualification, we can establish a ranking of coverage of the different job offers. shows said ranking. A priori, these results should be regarded with caution since, in addition to the aforementioned limitations in terms of absence of significant terms in the description of the training offer, there seems to exist a high correlation between the size of the profiles and their position in the ranking.

Figure 36. Ranking of coverage of the job offer profiles by training offering for degrees and master's degrees



11

ANALYSIS OF THE RESULTS OF APPLYING ML TECHNIQUES

11 Analysis of the results of applying ML techniques

The main purpose of this project consisted of determining whether ML techniques can constitute a useful tool for characterising labour supply and demand using automatic profiling techniques. In this chapter we explain the process followed to respond to this question in greater detail: firstly, some information is provided on the process for selecting the most adequate ML algorithm for each of the project tasks. Secondly, conclusions on the viability of ML are drawn.

The job offer detector explained in chapter 7 uses and intelligent crawler that filters much of the web domains without job offer, and the final classification is applied to web sites with job offer evidences only. Since the goal of this chapter is analysing the viability of applying ML, the study has been performed without including the intelligent crawling step, in such a way that all web domains take form the dataset analysed by the classifier.

11.1 Viability of ML for analysing the demand for ICT professionals on corporate websites

11.1.1 Efficiency of the automatic classifier.

We will represent the performance of a classifier by a ROC curve (see Section 5.1), which shows the compromise between the False Positive Rate (FPR) and the True Positive Rate (TPR).

In order to estimate this ROC curve, we applied the "Leave-One-Out" (LOO) technique described in Section 5.1.1 Other Methods for Detecting B2C Activity.

We will firstly show the performance of the detector selected for the application interface following a process detailed in the following section. Said detector has the following characteristics:

- Classification algorithm: logistic regression
- Algorithm for extracting features: Bagging.
- Number of features: 1,000 characteristics

The resulting ROC curve for this classifier is shown in Figure 37. For reference purposes, we can observe that in the BEP (diagonal cut) a TPR of 80% and a FPR of 20% is obtained. The best area under the curve (AUC) is 0.8596 (the maximum is 1.0) which gives some idea about the difficulty of detecting the presence of a job offer from the bag of words contained in the web when no selective crawling is applied (for instance, compare this result with the value AUC = 0.9584 obtaiined by the B2C detector (see section 5.1)).

Figure 38 illustrates the evolution of the area under the curve as the classifier is trained with a greater number of samples manually labelled for the same classifier.

Figure 37. ROC curve and AUC of the classifier based on logistic regression.

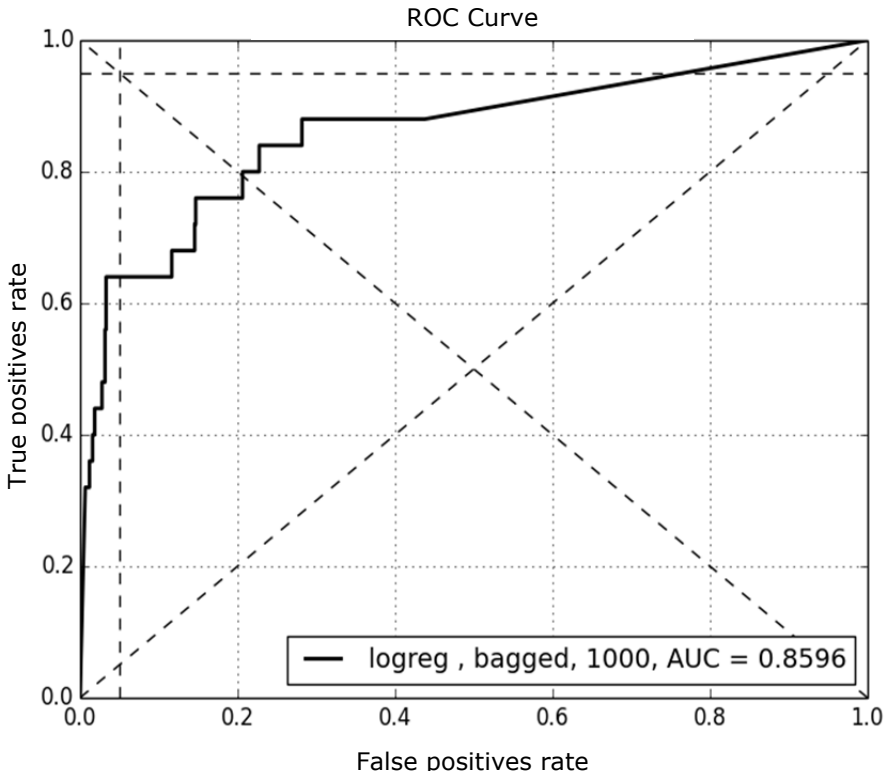
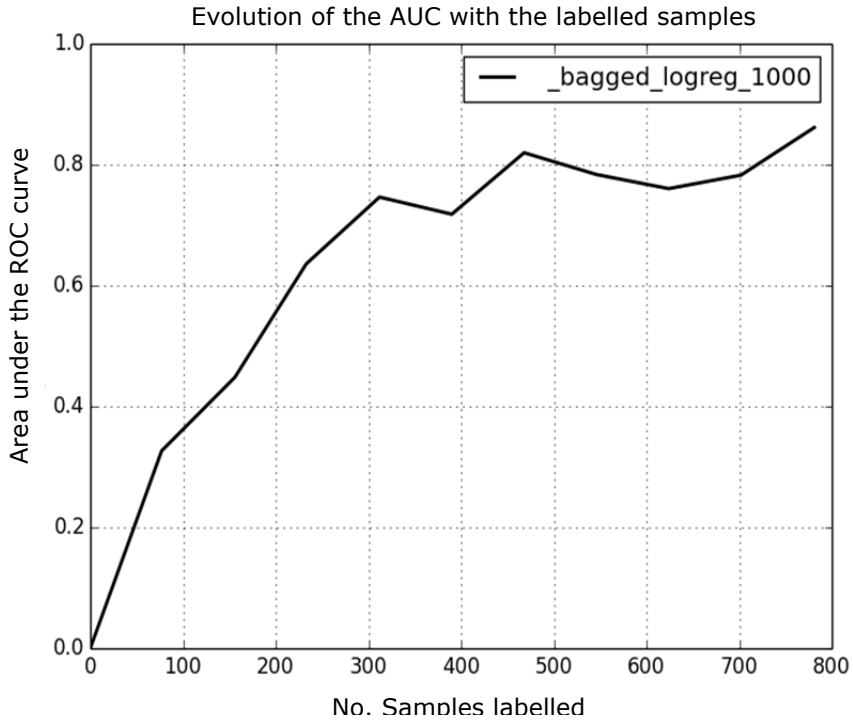


Figure 38. Evolution of the area under the ROC curve of the classifier based on logistic regression in accordance with the available number of manual labels.

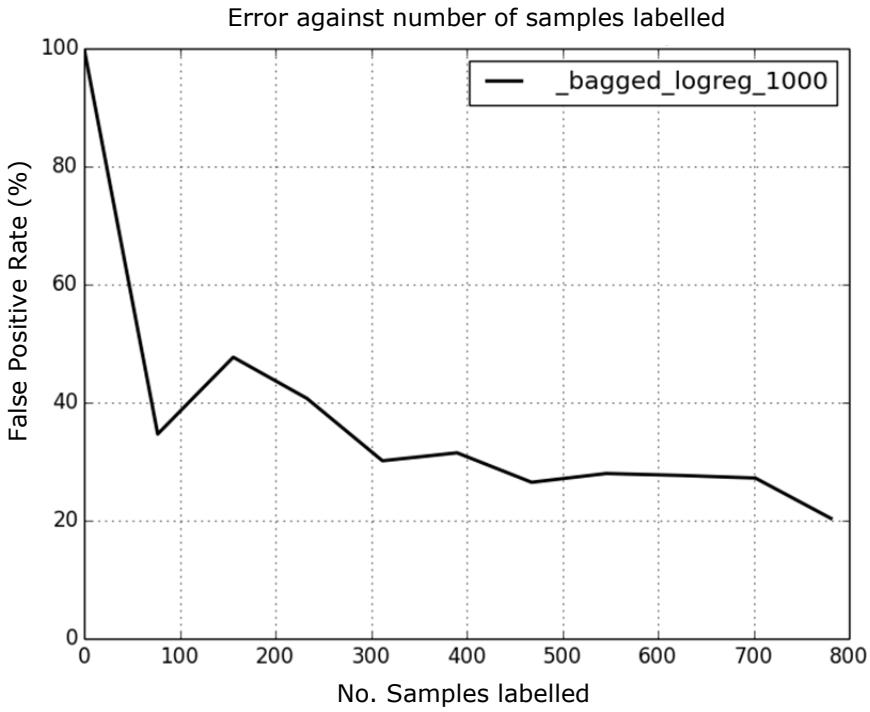


The area under the curve (AUC) grows up to approximately 0.8 with just over 300 labels and continues to oscillate around this value of 0.8 despite increasing the number of manual labels to more than double. This behaviour is due to the fact that, in general, the presence of job offers on a website is independent from the contents of said website. Additionally, the job offers may represent a small fraction of the web content, due to which its impact on the design of a classifier based on content analysis could be very limited.

In any case, the possibility of improving detection does not mitigate a subsequent technological problem, which consists of extracting the content of each job offer (once detected) for performing the profiling. In the fraction of websites that we have labelled manually we have observed that the structure of the layout of the content of the job offers found is far from being remotely homogeneous, due to which we performed a manual exploration of the websites retrieved by the automatic detector to locate and extract the text corresponding to each specific job offer.

The success of this manual extraction depends substantially on the detector having a low false positive rate, i.e. that nearly all the pages it retrieves contain job offers in order to optimise the time of the human operators who would explore these websites detected as potentially containing profilable job offers. The following figure shows the evolution of the false positive rate with the increase in the available number of manual labels.

Figure 39. Evolution of the false positive rate of the classifier based on logistic regression in accordance with the available number of manual labels.



The apparent stagnation in the performance of the classifier on increasing the number of manual labels to over 300 available labels illustrates the fact that this task is difficult to automate due to the extremely varied casuistry of incorporation of job offers into corporate websites and to the fact that these job offers do not represent a very significant part of the website contents.

11.1.2 Selection of the best classifier.

Job offer detection performance differs significantly in accordance with the method used for classifying and selecting characteristics. The selected detection algorithm was the result of a process of exploration of different classification and feature extraction algorithms. The selection made was based on four criteria:

1. Detector performance
2. Processing time
3. Ease of installation
4. Interpretability of the results

The classifier selection and optimisation process of the classifier consisted of the following stages

1. **Crawling** of the URLs with a crawl depth of 2.

2. **BoW analysis** (extraction of the bag of words). The texts of the processed URLs are transformed into vectors of 94,763 words evaluated using the TF-IDF measurement.

3. **Manual labelling** of 828 URLs. Some 100 websites are randomly labelled and a simple active learning algorithm is used to label up to a total of 828.

4. **Classification without word selection**. A classifier is trained without word selection, i.e. using the 94,673 words to separate the Profilable Offer class from the rest. Three different classifiers are used:
    - Support Vector Machine (SVM) with a linear core function; due to the high dimensionality of the input data, the use of more sophisticated cores is not necessary. Moreover, linear cores are the standard procedure in problems defined by means of bags of words.
    - Logistic Regression (LogReg)
    - Decision Trees (DTR)
    - Gaussian Processes (GP) with linear cores for the same reasons as the SVM.

    The results of this stage will serve as a baseline to assess the impact of the characteristics selection stages.

5. **Word selection**. Two word selectors are used: Bagging and Recursive Feature Elimination for choosing a most relevant subset of words that will improve the classification rates.

6. **Analysis of word selection performance**. The area under the ROC curve (AUC) obtained by each selection method for different sizes of word sets is assessed. This study is conducted on the three classification techniques used in the application.

7. **Analysis of performance in accordance with the number of URLs labelled**. This study gives us an idea of the impact of labelling more URLs.

8. **Profiling** of the job offers found in the URLs belonging to the class "Profilable Offers".

Three different types of classifiers have been assessed:

- Support Vector Machine (SVM) with a linear core function; due to the high dimensionality of the input data, the use of more sophisticated cores is not necessary. Moreover, linear cores are the standard procedure in problems defined by means of bags of words.

- Logistic Regression

- Gaussian Processes (GP) with linear cores for the same reasons as the SVM.

The classification module also includes the GP-type classifier (Gaussian Process), as mentioned in the proposal, but the execution of this method was too expensive (up to five days of execution) for such poor performance, due to which it was substituted for another alternative method, that of Classification by means of Decision Trees (DTR).

The ROC curves of Figure 40, Figure 41 and Figure 42 show the performance of the classifiers based on logistic regression, SVM and decision trees, respectively, for different feature selection methods and number of features selected. It can be observed that the best performance was obtained for the combination of a classifier based on logistic regression and the selection method based on bagging, when 1,000 features are used, in which case an AUC equal to 0.8596 is obtained.

It was observed that the rest of the selection methods produced very poor results in this regard. This generally low performance of the classifiers of the presence of job offers based on the global content of the URLs is due to the fact that the portion corresponding to the job offers is small in comparison with the total URL content, due to which its contribution to the design of automatic classifiers is obscured by the other content. This situation is aggravated by the fact that websites belonging to the negative class, specifically websites with partial job offers (not profilable) or websites without available job offers overlap in key terms with the websites belonging to the positive class, greatly hampering their correct detection.

Figure 40. ROC curves for logistic regression and different feature extraction methods and number of characteristics.

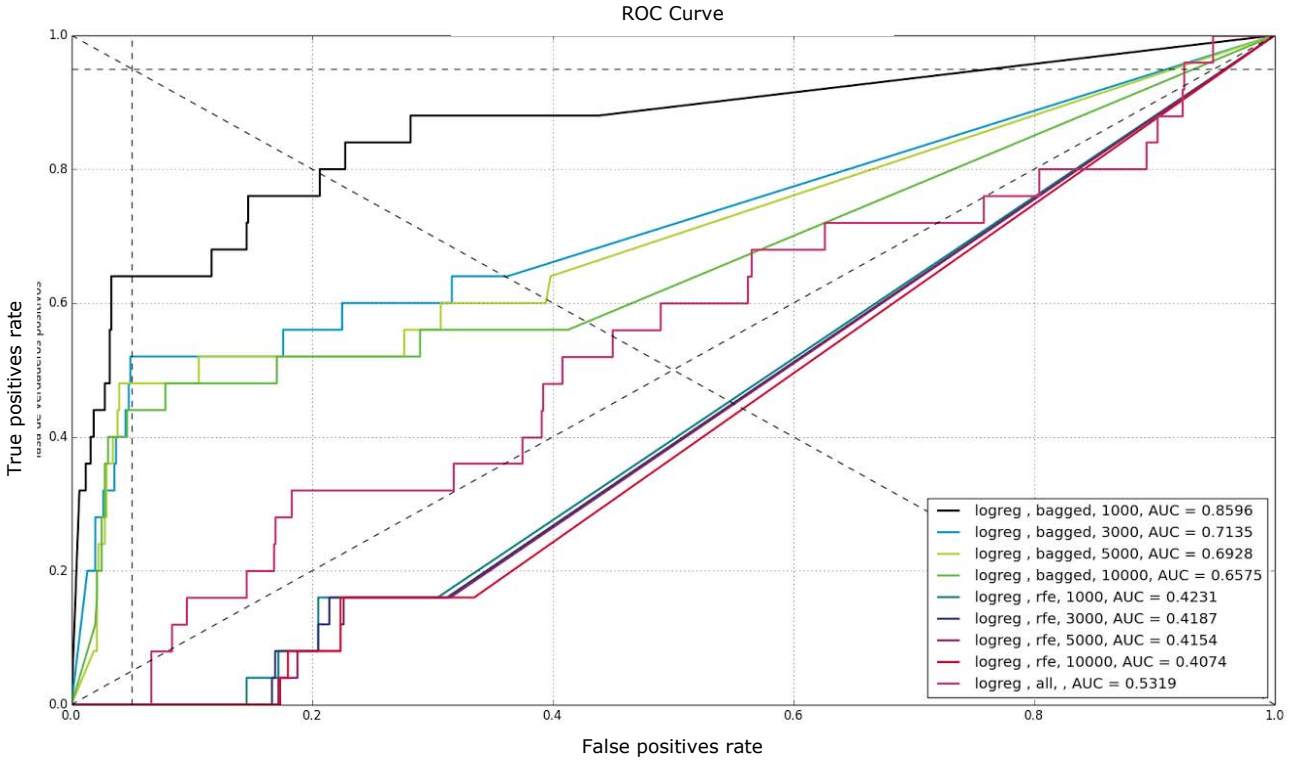


Figure 41. ROC curves for a SMV-type classifier and different feature extraction methods and number of characteristics

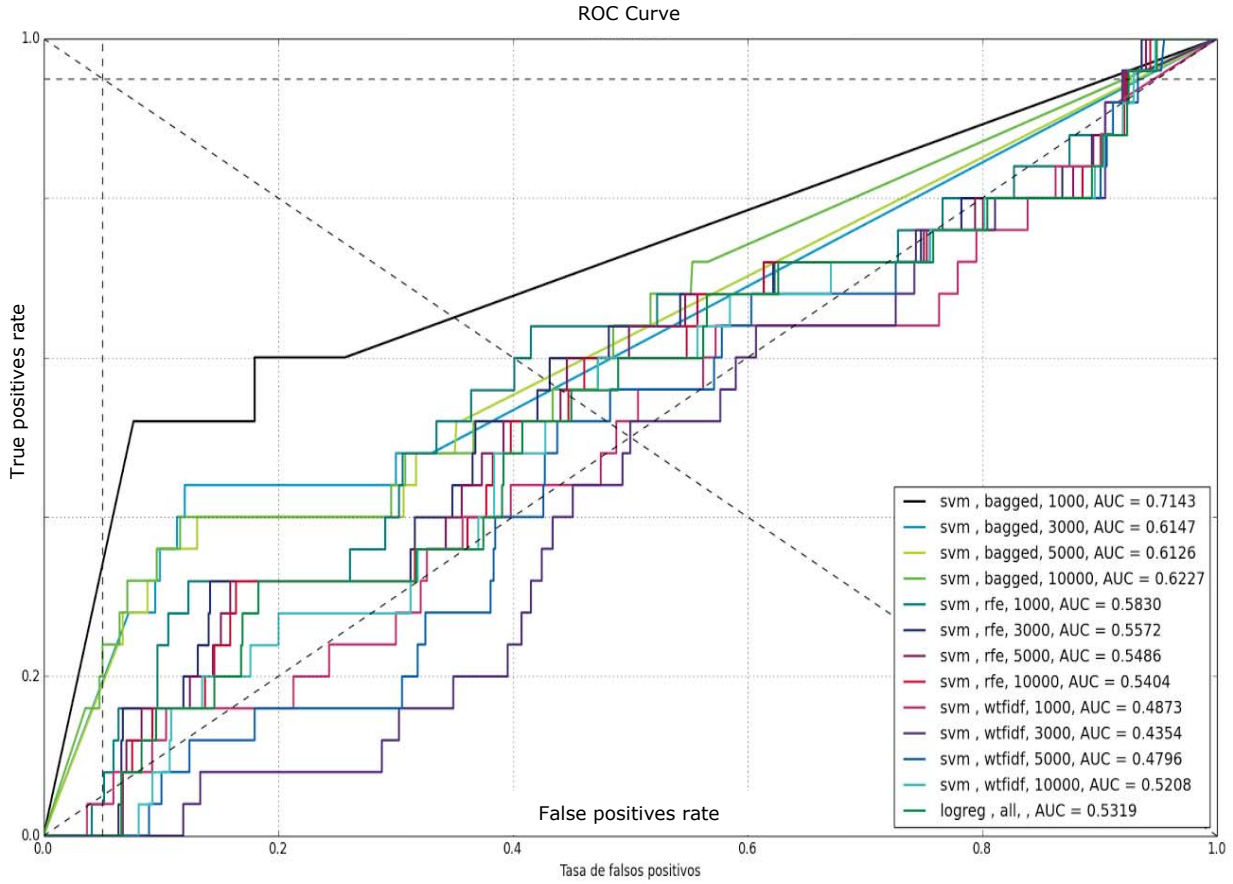
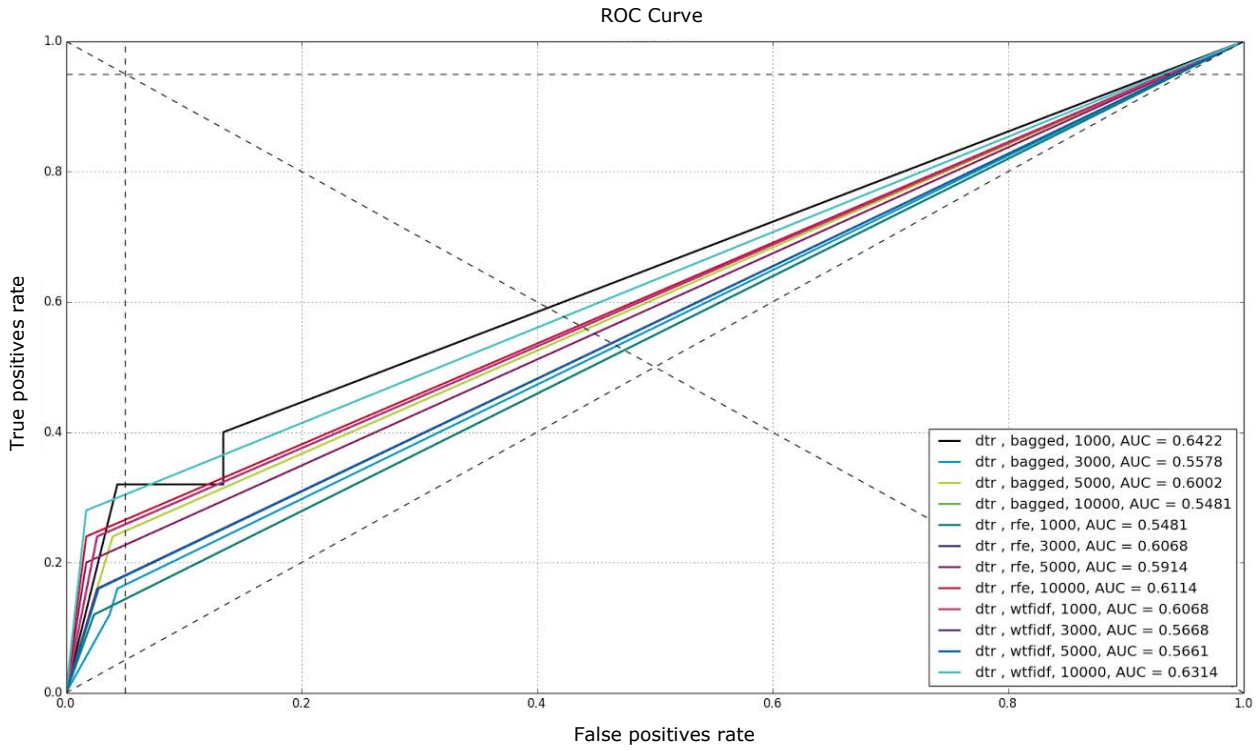


Figure 42. ROC curves for decision trees with different feature extraction methods and number of characteristics



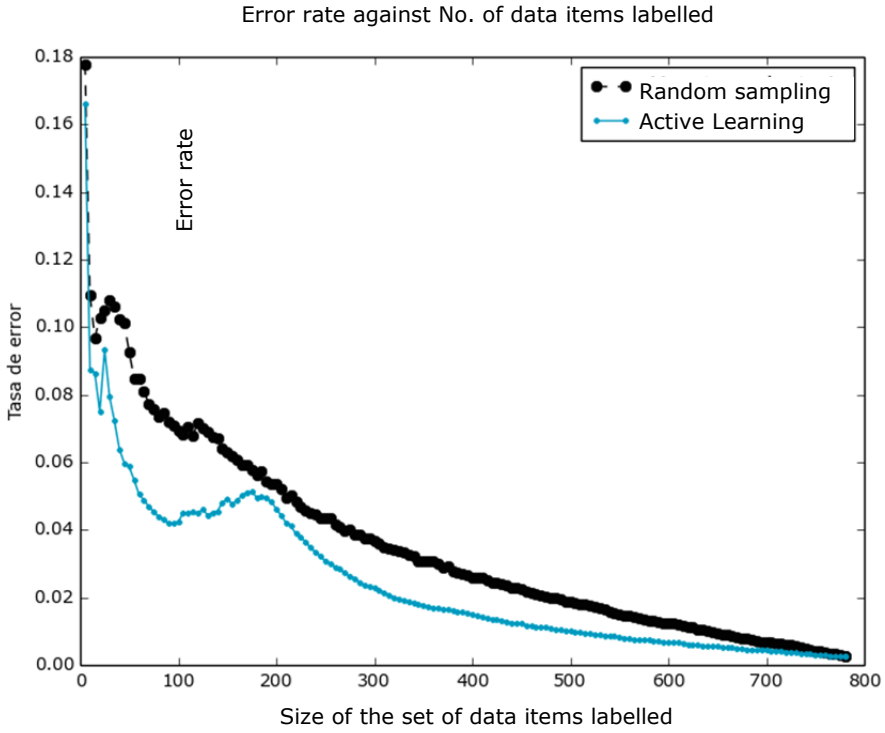
The graphs essentially show that the classifier based on logistic regression, using 1,000 terms chosen by means of the selection of characteristics based on bagging, obtained the best performance. These results are clearly inferior to those of the B2C task due to the difficulties inherent to detecting websites with job offers based on the content of these websites, to which reference has been made throughout this report:

- There are very few examples of corporate websites with presence of profilable job offers
- The manner in which to structure the information of the job offer within the websites of the different companies is very heterogeneous
- Job offers represent a very small portion of the total website content.

11.1.3 Need for labelling and gains for active learning

The following figure shows how the error rate decreased with the number of data items labelled. The tests were performed using logistic regression as the classifier and with vectors of 1000 features. The blue curve shows the evolution of the error rate when applying an active learning (AL) strategy, which falls towards the minimum much more quickly than when using labelling based on random sampling labelling. Thus, almost the same performance was obtained with AL on 600 labels as was obtained with random sampling with 800 labels.

Figure 43. Error rate with and without AL.

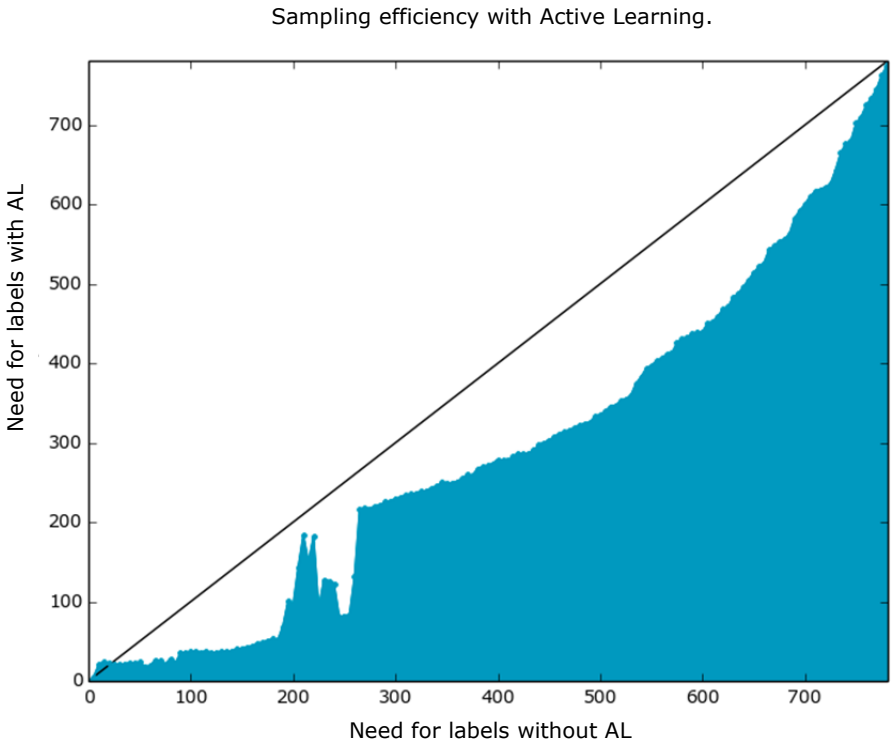


This effect is more evident if this figure is represented directly, comparing labelling based on random sampling with labelling based on active learning. The curve in blue shows, for each value of the number of labels used with random sampling, the number of labels required by the AL algorithm to achieve the same error rate. It can be clearly observed that the AL algorithm reduces labelling needs by approximately 25%.

The work performed shows that it is possible to apply ML techniques to automate the detection of the presence of job offers on Spanish corporate websites. Manual labelling of some 800 of a total of around 8,000 businesses (around 10% of all businesses) was sufficient to perform automated classification with a high level of precision. Moreover, the use of active learning techniques can significantly reduce labelling needs: approximately 600 labels would have been enough (i.e. approximately 7.5% of the total websites).

These figures evidence the data classification potential of ML techniques (in this case websites) that were not used during the design phase. Provided that the database used is of the same type (technically, that it comes from the same statistical source), the same classifier can be used.

Figure 44. Active learning sampling efficiency.



11.1.4Conclusions

The work performed shows that the performance of a job offers detector based exclusively on the use of ML techniques is limited. The main reason is that the content of a corporate website is minimally related to the job offer. The use of an intelligent browsing technique combined with automatic classification seems the most promising method for reducing error rates.

11.2 Viability of ML for obtaining profiles for analysing the demand for ICT professionals in job portals and analysis of ICT training programmes.

In order to better discuss the viability and potential use of automatic profiling tools for analysing demand for ICT professionals based on offers in job portals, and for analysing the ICT training offering, it would be convenient to summarise some of the most important properties of the profiling methods used. The tools are applied on the basis of the following hypotheses:

- Each profile is characterised by a distribution across the terms vocabulary. That is, a profile defines the frequency with which the different terms that comprise the dictionary appear in the documents of said profile.

- Each document is characterised by a combination of profiles. In other words, in general, documents to not belong to a single profile, but rather can be described as a combination of different profiles in different proportions.

The learning techniques used enable the simultaneous identification of the most representative profiles and the degrees of pertainment of all the documents of the dataset to profiles.

It would be easy to learn the profiles if we knew the degree of pertainment of documents to profiles, and to assign documents to profiles if these were available a priori. The difficulty lies in the simultaneous learning of the profiles and compositions of the documents, based solely on the datasets. To this end, different implementations of techniques known as Latent Dirichlet Allocation (LDA) and hierarchical LDA (hLDA) have been used.

The objective of the LDA is that which we have just described: learning without supervision of a set of profiles based on a dataset. As opposed to other pre-existing techniques, LDA is more solidly grounded from a theoretical viewpoint. However, in order to correctly interpret the results of LDA modelling, it should be understood that the real objective pursued by said technique is that of finding a *generative model* of the documents. LDA assumes that each word of each document has been generated by a profile and, therefore, prefers models with profiles that enable the explanation, with the greatest possible authenticity, of the generation of all the words of all the documents of the dataset. Occasionally, this can hinder the identification of profiles scarcely represented in the dataset since, with regard to the model, it may be preferably, for example, to subdivide a large profile into two individual profiles. It also explains the appearance of cross-cutting profiles (such as language profiles, which are partially used by a large number of documents) and even noisy profiles, but that enable the explanation of the residual generation of certain words of the dataset. All of these effects are known and discussed in scientific literature relating to LDA.

In relation to the foregoing, it should also be understood that the optimisation of the objective pursued by the LDA does not necessarily imply that better profiles will be obtained for the user, which hinders the automation of the search for more satisfactory models, and suggests the suitability of certain degree of human supervision in the selection of the parameters of the method, which includes the number of profiles to be extracted or the depth to be used in the hierarchical model. Once again, these characteristics of the LDA are generally accepted and, in fact, are common to the use of any type of non-supervised technique.

Before materialising our conclusions on the viability of the use of profiling tools for analysing the demand for professionals and the training offering in the ICT sector, we consider it convenient to complement the results presented in Chapter 8 with a series of experiments that attempt to illustrate the effects that we have just described. Given that the results obtained in Chapters 8 and 9 allow us to draw very similar conclusions, for the sake of brevity in the report, the detailed study presented below is made on the dataset of job offers; although it can be verified with the visualisation tool that the conclusions obtained are common to those of the dataset of job offers.

11.2.1 Authenticity of the models and selection of the number of profiles

Chapter 8 illustrates the profiles obtained in the InfoJobs dataset for different selections of the number of profiles of the model. Likewise, the profiles obtained for the other datasets on varying the number of profiles can be reviewed using the visualisation tool. The selection of the number of profiles is not an easy task and is subject to a balance between:

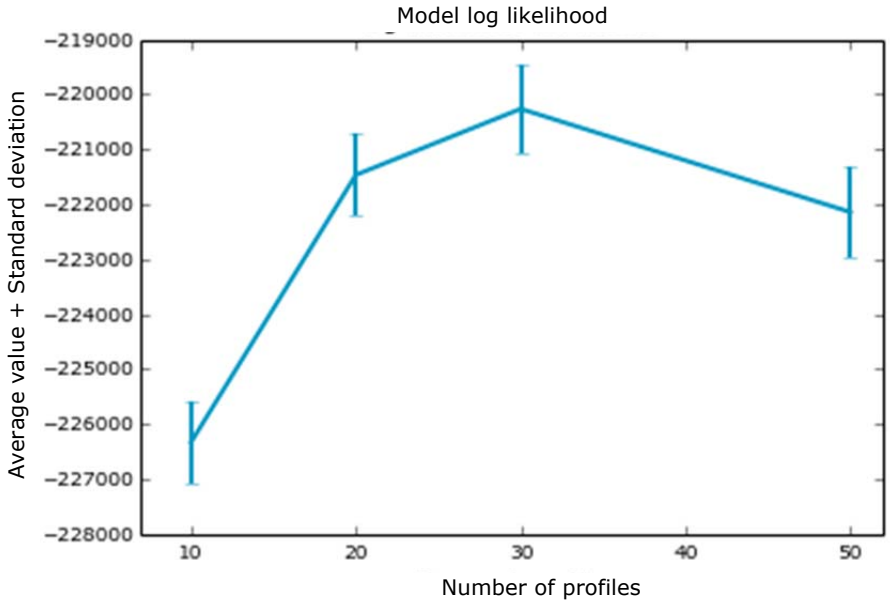
- the obtainment of noisy profiles, but also the capability to identify well-defined profiles but scarcely represented in the dataset (for a high number of profiles), and
- the obtainment of profiles well characterised by skills and technical competences, with no presence of irrelevant profiles, but inability to detect profiles scarcely represented in the dataset, and possibility that the profiles that we would like to detect in an isolated manner appear merged in the model (for a small number of profiles).

In view of the results obtained, in Chapter 8 it was considered that the selection of 20 profiles was reasonable for all the datasets considered, although said selection will ultimately be conditioned by the size of the dataset and by the actual number of topics actually present in the dataset. Therefore, it would be desirable to have an objective criterion for selecting the number of profiles.

An initial possibility for establishing said objective criterion would consist of the use of the log likelihood of the model, which can be understood as (the logarithm of) the probability with which the available dataset for a specific profile model can be observed. illustrates the behaviour of said log authenticity for different values of the number of profiles using the InfoJobs portal dataset. To this end, ten different models have been trained for each selection of the number of profiles, whereupon the figure shows the average value and typical deviations of the log authenticities obtained. The differences between the different embodiments are due to different algorithm initialisations.

We can observe how the maximum value of log authenticity is obtained for 30 profiles, which differs from the selection that we made on directly analysing the profiles obtained. In fact, the profiles selected for visualisation in Chapter 8 in no case coincide with those that obtained greater authenticity for each number of profiles.

Figure 45. Analysis of the authenticity of the model obtained by the LDA model for different selections of the number of profiles of the model (dataset: InfoJobs).



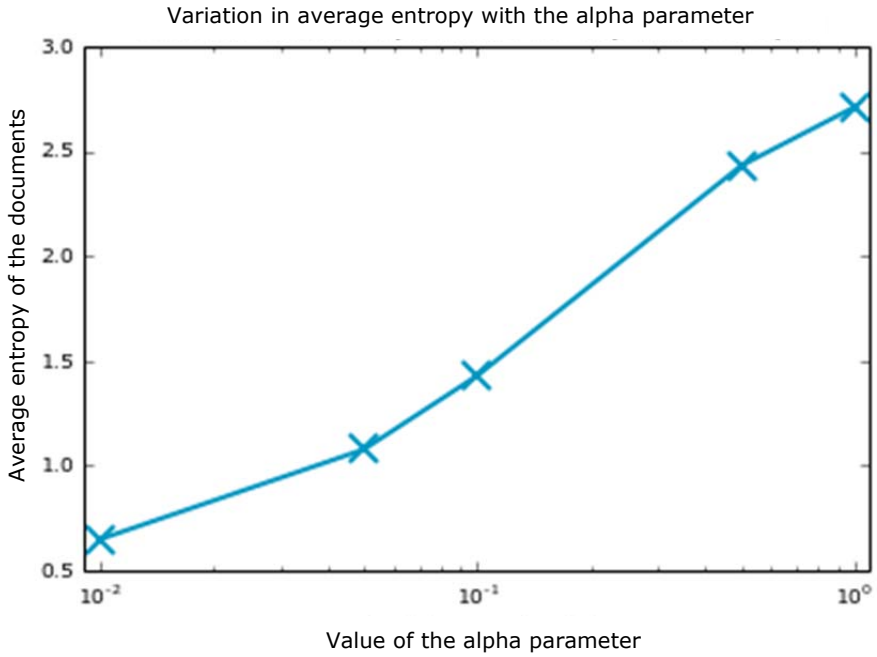
Notwithstanding the foregoing, the use of the model log likelihood could be a support tool for selecting the number of profiles. By way of example, Figure 45 suggests that the use of 10 profiles is clearly insufficient for analysing InfoJobs.

11.2.2 Alignment of the "alpha" parameter a priori

A second parameter that could have a significant influence on the type of profiles obtained, and particularly on the description of the documents as a combination of profiles, is the "alpha" parameter of the LDA algorithm, the Dirichlet distribution parameter that said algorithm uses to generate the decompositions of the documents as combinations of profiles. In general, using a lower value of said parameter gives rise to more dispersed descriptions; i.e. in general, we have ensured that the documents belong more purely to a single profile, while high "alpha" values give rise to descriptions of documents belonging to a higher number of profiles.

To illustrate said effect, the following figure represents the normalised entropy value averaged over all the documents of the InfoJobs portal dataset for different "alpha" values, on training a model with 20 profiles. A null normalised entropy value indicates that the associated document belongs wholly to a single profile, while a unit value is associated with documents with flat distributions across all the profiles of the model.

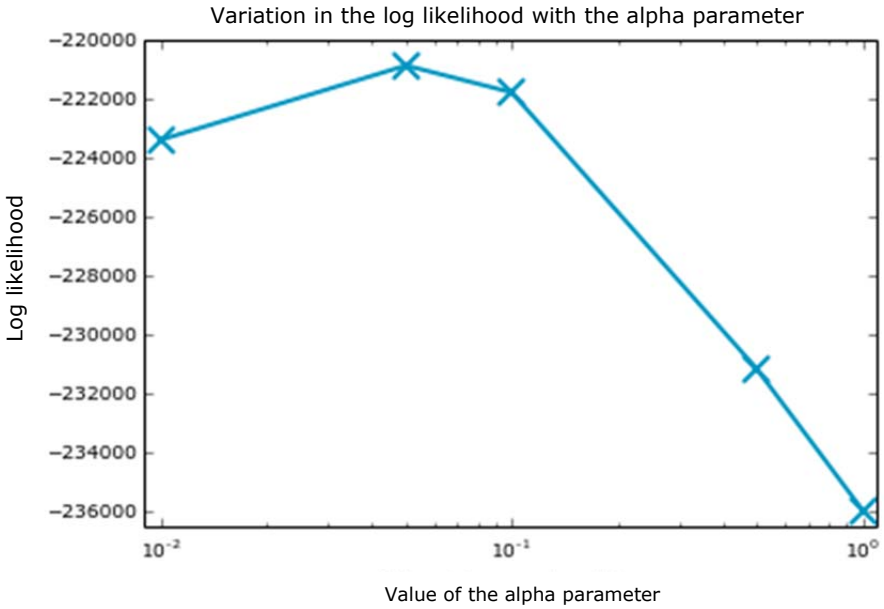
Figure 46. Influence of the free parameter on the normalised average entropy of the documents of the dataset. Low values of said entropy indicate the existence of a large number of documents that are assigned in a practically pure manner to a single LDA profile (dataset: InfoJobs).



It can be observed that the use of low parameter values always gives rise to smaller entropies, which in principle would be a desirable characteristic. Nonetheless, the visualisation of the profiles obtained for different parameter values indicates that said profiles are relatively robust to the selection of any value for "alpha". However, on using extremely low values (alpha = 0.01) a certain deterioration was observed in the detected profiles. The effects are more serious in terms of the estimation of the size of the profiles, given that when using a high "alpha" value the estimated size of the different profiles tends to become equalised.

It is also interesting to observe the influence of the free parameter on the log likelihood of the model, which is shown in Figure 47. We can observe that said likelihood is highly dependent on said value. In practice, there are algorithms for learning an appropriate value for this free parameter, iteratively aligning it with the dataset. Said algorithms have been used in all the results presented in this report, except those included in this Subsection, to adjust the parameter. In the case of InfoJobs, the value selected is close to 0.5 which, as we can observe, coincides with the value that maximises the log likelihood of the model (Figure 47).

Figure 47. Influence of the free parameter on the log likelihood of the model (dataset: InfoJobs).



11.2.3 Visualisation of the results with down-scoring of frequent terms

LDA is based on a model of the production of words in the dataset being analysed. Each profile can be considered as a distribution over the vocabulary; i.e. it is characterised by the probabilities with which said profile generates the different words that comprise the set of terms of the dataset. In order to visualise the profiles learned, the size of the profile is usually represented in the dataset, together with the most frequent words in accordance with said profile. However, it is to be expected that the most common words of the dataset will be highly probable under many of the extracted profiles, which could be inconvenient: it would be preferable to visualise the set of terms most specific to each profile.

An alternative consists of penalising words with high probability of being observed under different profiles. As mentioned earlier, the advantage of said down-scoring is that it highlights the words most specific to each profile; the drawback is that the probabilistic sense is lost, i.e. we only have one word ranking for each profile, but said ranking is not simply based on the frequency of appearance of the terms that comprise the profile.

In order to illustrate the convenience of one representation or the other, the order of the words provided by the model for each profile on analysing the InfoJobs dataset with a model with 20 profiles is shown. The following table lists the terms that would be obtained in five of the profiles applying and without applying the down-scoring technique that we have just reviewed. Although there are no significant differences between the two strategies, it

can be observed that some relatively common terms lose relevance (programming in Topic 0 and Topic 2; web in Topic5), while the most specific terms gain relevance (e.g.> asp_net in Topic 2, vmware in Topic 8).

In general, the term re-weighting strategy has been used throughout this report, although in some specific cases said technique could deteriorate the presentation of specific profiles. In summary, manual inspection is once again required to determine the suitability of one model of profiles with respect to another.

Table 23. Descriptive profiles of the InfoJobs dataset on using the down-scoring strategy (bottom row).

Topic 0	Topic 2	Topic 5	Topic 8	Topic 12
java	.net	web	Windows	android
j2ee	php	design	systems	iOS
spring	programming	html5	server	mobiles
struts	server	javascript	technician	applications
programming	asp_net	html	vmware	telecommunications
oracle	javascript	css3	administration	developer
jsf	sharepoint	css	Linux	sale
programmers	visual_studio	jquery	support	insurance
eclipse	mysql	adobe	administrator	python
analysis	mvc	photoshop	servers	commercial
java	.net	html5	Windows	android
j2ee	php	web	vmware	iOS
spring	asp_net	javascript	Linux	mobiles
struts	server	design	server	telecommunications
jsf	programming	css 3	systems	insurance
eclipse	sharepoint	html	support	python
oracle	javascript	CSS	servers	sale
programmers	visual_studio	jquery	red_hat	commercial
programming	mysql	adobe	microsoft	developer
maven	mvc	photoshop	administrator	telephony

11.2.4Conclusions

The use of unsupervised automatic profiling techniques has enabled the identification of a large number of profiles of interest based on the analysis of the job offers published in job portals. Also, as mentioned in Chapters 4 and 5, it has even been possible to associate the detected profiles in a more or less clear manner with certain well-defined professional or training categories. The use of these machine learning techniques offers clear advantages:

- They operate based on data published on the Internet, due to which the need to perform expensive interview-based studies is eliminated.
- They allow analysis with the desired frequency. With regard to repeating the study presented in this report, the only module that may have to be adapted is the crawling model, in the event that the structure of the web pages of the portals analysed were modified.

- In relation to the possibility of repeating the study, they enable monitoring of the evolution of the demand for professionals in the sector and of the training offering.

- They enable the detection of new profiles, as well as the appearance of new technologies and trends of interest to pre-existing profiles, for example, as a result of the appearance of new terms in the description of certain profiles.

- They offer the possibility of estimating the relative importance of the different profiles observed.

- They allow merging of different information sources, as was observed on analysing the results of a dataset integrated by the merger of the job offers of the three job portals.

Notwithstanding the multiple advantages, it should be noted that the analysis performed in this report required human intervention at certain points thereof and that said intervention was decisive for obtaining relevant profiles. Specifically,

- the lists of stopwords were edited manually

- the lists of n-grams were edited manually

- the selection of models was performed by means of visual inspection of the profiles obtained

- the assignment of profiles to professional categories was performed manually

The results discussed in the previous sections suggest that said partial supervision is essential and cannot be completely eliminated.

Special mention should be made of the use of hierarchical models based on hLDA which did not give satisfactory results in any of the analysed databases, partly due to the datasets themselves, but also to the nature of the model implemented by hLDA.

Other techniques and probabilities that we consider worth exploring are as follows:

- Establishment of a work methodology allowing the replication of study results, reducing dependence on the subjective interpretation of the person conducting the study to the extent possible.

- Maximisation of the feedback tool provided by the experts who analyse the results. Therefore, it would be very interesting to be able to reuse any available labelling of specific profiles or documents to improve the model using supervised LDA.

- Consideration of new hierarchical models, basically different to that implemented by hLDA, that can work with a certain degree of human supervision. Thus, it would be interesting to propose learning schemes in which an expert could mark correct or incorrect profiles, choose those requiring analysis with a greater degree of granularity, etc.

- Lastly, one of the possibilities that we consider to be of greatest interest and potential is the use of dynamic topic modelling (DTM) to analyse the time course of the different profiles. Said model makes it possible to associate a timestamp with the documents and follow the evolution of the profiles over time, for example, by analysing the change in their relative sizes or in the most relevant competences of each profile.

11.3 Viability of ML for the comparative analysis of supply and demand of ICT professionals

11.3.1 Preliminary discussion

Chapter 10 of this report sets out the results of the behaviour of the crossed similarity techniques between documents and/or profiles of the job offer (InfoJobs) and training offering (Degrees and Master's Degrees, and Vocational Training) datasets. Although certain reasonable preliminary results were obtained in the case of the training offering for Degrees and Master's Degrees, an evident difficulty is the absence of important key terms in the vocabularies of the documents of the training datasets. In this section we will endeavour to provide greater evidence of this issue.

Figure 48 represents the percentage of documents of the training offering datasets in which the 146 most relevant terms for profiling the job offer appear (the 10 most relevant terms of the model with 20 profiles were used, eliminating duplication). By way of example, the green curve indicates that of these 146 terms only 38 are present in any of the documents of the vocational training dataset. By way of example, the terms that may be excluded from said shortlist include the following: business_intelligence, unix, oracle, cloud, spring, social_networks, google, android, community, java, etc.

The situation is only slightly better in the case of the Degree and Master's Degree dataset. An advantage of this dataset is that the decay of the curve (represented in blue in Figure 48) occurs at a slower rate. It is precisely these terms, which appear in a relatively small number of documents, which allow said documents to learn specific profiles. The list of terms that appear between 1% and 20% of the documents of this dataset includes the following: marketing, mobiles, servers, tcp, ip, sem, community, html, google, seo, manager, java, administrator, cloud, linux, ios, javascript, mysql, unix, php, business_intelligence.

Figure 48. Percentage of documents of the training offering in which the most relevant terms for profiling job offers appear.

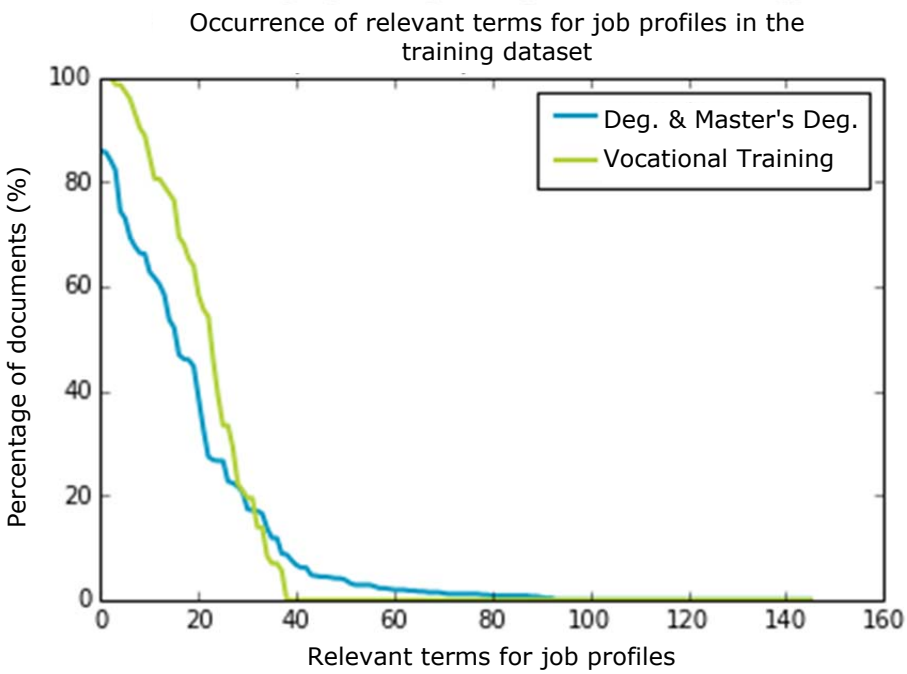
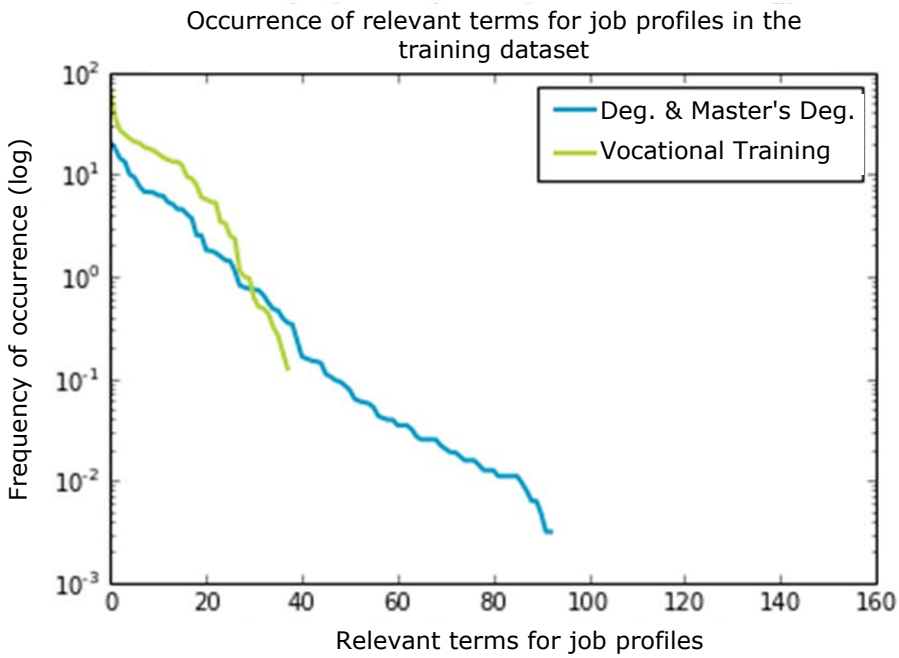


Figure 49 shows the average number of appearances of the different 146 terms over the training offering dataset. It is very interesting to note that the curve corresponding to Degrees and Master's Degrees has a linear aspect over these axes (note the logarithmic scale on the vertical axis). This behaviour (ley-mu) is commonly observed in various scenarios, such as distributions of the number of inbound links in web pages, distribution of the number of links in social networks, etc., and is usually considered an adequate distribution for the search for the structure inherent to the data.

11.3.2Conclusions

The analysis of the results obtained in the matching task and a detailed analysis of the coincidence of the most relevant terms in the job offer and training offering datasets suggest that obtaining high-quality crossed similarity measures between the two datasets is possibly a difficult task. The main difficulty lies in the substantial differences between the terms to be used to enumerate the requirements of a job offer (technical competences) and those used in the qualification verification memories that were accessed to perform this task.

Figure 49. Average number of appearances of the most relevant terms for the profiling of job offers in the training offering dataset.



In our opinion, the main avenue for improving the results would require access to specific documentation on the qualification syllabi and even on the technical competences associated with the different subjects that comprise them. In this regard, universities should probably make an effort to bring the specific technical competencies acquired by the students of their programmes into line with current labour market requirements.

Despite the difficulties of the task undertaken, it must be highlighted that the proposed similarity measures made it possible to achieve certain positive results, particularly in the case of the Degree and Master's Degree training offering dataset, as described in Chapter 10 of this report.

12

FINAL CONCLUSIONS

12 Final conclusions

If the main purpose of this project consisted of determining the viability of ML techniques to detect e-Commerce in Spanish companies and to characterise labour supply and demand, we must draw some final conclusions from the study as a whole, adding the specific conclusions mentioned in preceding chapters and synthesising a response that could act as a guide for future projects.

The work performed has allowed us to confirm that machine learning techniques have extraordinary potential for analysing massive volumes of data from the Internet. In particular, we can conclude that:

- It is possible to automatically detect the presence of e-Commerce on Spanish corporate websites with a high degree of accuracy.
- The combination of an intelligent browsing process with ML algorithms makes it possible to discriminate the presence of job offers on corporate websites with a high degree of accuracy.
- Automatic profiling algorithms enable the identification of general trends in the employment offered by companies as a whole or in the training offering of universities and vocational training centres.
- Despite the differences (in format, vocabulary, structure) of the datasets of job offers and curricular profiles, matching algorithms can represent a powerful tool for performing comparative analysis of the global alignment between the supply and demand of professionals in Spain.

In the course of the study, the work performed has allowed us to learn important lessons with a view to future projects:

- The development of a good system for processing information extracted from Internet data sources requires the efficient combination of several technologies including, most notably, the following:
    1. Web page crawling technologies.
    2. Natural language processing technologies
    3. Machine learning algorithms (ML) and, in general, statistical processing of information
- The quality of the system for IaD will depend to a large extent on the efficient use of all these technologies. The direct application of machine learning algorithms to raw data does not provide good results, due to which it was necessary to:
    - Refine the crawling processing by accessing the information in a selective manner, which makes it possible not only to reduce Internet data capture times but also to reduce the abundance of noise in the data processed by the classification and automatic profiling algorithms.

- Refine the processing of textual information using natural language processing techniques. The scope of the work was limited to treating web pages as "bags of words", but the exploratory work with n-grams already suggests the idea that adequate language processing could have a significant impact on the final results.

- Explore different classification and feature extraction algorithms, taking into account not only classification efficiency, but also other variables related to the scalability of the procedure and how it must address the need arises to process information from hundreds of thousands of URLs.

- The use of all these technologies makes it possible to automatically process large amounts of information, but does not exclude the need for manual intervention:

  - The classification algorithms need the prior labelling of a small set of web pages.

  - The profiles obtained automatically significantly improve with a process for manually supervising results and filtering terms that are not relevant to the analysis.

- This manual intervention process is very important and requires in-depth analysis. The manual intervention must fulfil various important requirements:

  - It must require a minimum of human resources. In this regard, active learning techniques have demonstrated great potential use.

  - It must not require expert knowledge of the technology.

  - It must be rigourous in its application. Manual labelling errors are propagated to the ML algorithms and inadequate filtering of the terms could undermine profiling quality.

- The assessment of the quality of an ML algorithm is not a trivial task.

  - We assessed the quality of automatic detection (of B2C activity or job offers) by comparison with the decisions of the classifier with the results of manual labelling. However, manual labelling is not error free.

  - Comparison with other sources is possible, but also requires playing down discrepancies. By way of example, the comparative analysis of the results of B2C detection with INE sources showed a high degree of correlation, but the differences are not attributable (at least solely) to automatic detection errors, but also to other causes, such as the different nature of the labelling process used by the INE.

  - While there are some statistical measures for assessing the goodness-of-fit of profile model data, the main conclusions of the supply and demand characterisation process are based on subjective assessments arising from the visual inspection of the profiles obtained by the algorithms. It would be

necessary to establish procedures that would make it possible to put the quality of an automatic profiling algorithm in objective terms.

In relation to these aspects, the most important conclusion that can be drawn from this work is that ML techniques represent a powerful tool for the low-cost analysis of massive volumes of data, which does not replace other analysis sources but rather complements them: they provide valuable additional information with potential advantages (the capacity to process massive volumes of data without recurring to sampling techniques, minimisation of human intervention, etc.) that must be leveraged and limitations (basically technological, in the processing of natural language, in the classification algorithm, etc., but also with regard to the need for preliminary engineering processes), the impact of which must be established in each case.

Lastly, it should be understood that the work performed does not represent an end in itself. In a project with a substantial technology component, including software development and application of algorithmic problem solving, application software not only solves the initial problems and provides initial answers, but also paves the way for future technology development. In preceding chapters we already suggested some URLs in which the quality of the results of the projects could be improved in all aspects (improvement in labelling, more efficient processing of the website content, access to JavaScript content, segmentation of job offers on the website, improvements in the hierarchical models, application of dynamic models, etc.). In this connection, the greatest potential would lie in the development of algorithms for the comparative analysis of supply and demand, which probably constitutes the main technological challenge to be addressed within the scope of this project.

13 List of Acronyms

AL: Active Learning

AUC: Area Under the Curve

B2C: Business To Consumer (e-Commerce aimed at consumers)

BEP: Break Even Point

BoW: Bag of Words

CNAE (National Classification of Economic Activities): Spanish national classification of economic activities

INE: Spanish National Statistics Institute

FN: False Negative

FNR: False Negative Rate

FP: False Positive

FPR: False Positive Rate

LOO: Leave-One-Out

ML: Machine Learning

ROC: Receiver Operating Characteristic

TN: True Negative

TP: True Positive

TNR: True Negative Rate

TPR: True Positive Rate

14 Table of Figures