


# WoLMIS: a labor market intelligence system for classifying web job vacancies

Roberto Boselli<sup>1,2</sup> · Mirko Cesarini<sup>1,2</sup> · Stefania Marrara<sup>3</sup> ·  
Fabio Mercorio<sup>1,2</sup>  · Mario Mezzanzanica<sup>1,2</sup> · Gabriella Pasi<sup>3</sup> · Marco Viviani<sup>3</sup>

Received: 20 January 2017 / Revised: 8 September 2017 / Accepted: 11 September 2017 /  
Published online: 20 September 2017  
© Springer Science+Business Media, LLC 2017

**Abstract** In the last decades, an increasing number of employers and job seekers have been relying on Web resources to get in touch and to find a job. If appropriately retrieved and analyzed, the huge number of job vacancies available today on on-line job portals can provide detailed and valuable information about the Web Labor Market dynamics and trends. In particular, this information can be useful to all actors, public and private, who play a role in the European Labor Market. This paper presents WoLMIS, a system aimed at collecting and automatically classifying multilingual Web job vacancies with respect to a standard taxonomy of occupations. The proposed system has been developed for the Cedefop European agency, which supports the development of European Vocational Education and Training

---

✉ Fabio Mercorio  
fabio.mercorio@unimib.it

✉ Gabriella Pasi  
pasi@disco.unimib.it

Roberto Boselli  
roberto.boselli@unimib.it

Mirko Cesarini  
mirko.cesarini@unimib.it

Stefania Marrara  
stefania.marrara@disco.unimib.it

Mario Mezzanzanica  
mario.mezzanzanica@unimib.it

Marco Viviani  
marco.viviani@disco.unimib.it

<sup>1</sup> Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan, Italy

<sup>2</sup> CRISP Research Center, University of Milano-Bicocca, Milan, Italy

<sup>3</sup> Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy

(VET) policies and contributes to their implementation. In particular, WoLMIS allows analysts and Labor Market specialists to make sense of Labor Market dynamics and trends of several countries in Europe, by overcoming linguistic boundaries across national borders. A detailed experimental evaluation analysis is also provided for a set of about 2 million job vacancies, collected from a set of UK and Irish Web job sites from June to September 2015.

**Keywords** Labor market intelligence · Text classification · Machine learning · Knowledge discovery · Information systems

## 1 Introduction

In recent years the advent and the increasing popularity of the Web have strongly affected the way in which Labor Market (LM) information diffuses in Europe. On the one side, the development and the spread of dedicated Web-centric services have been growing exponentially, with the consequence of earmarking a significant part of the European labor demand through Web portals and services. On the other side, the informative power of the labor resources available on the Web – as in the case of Web job vacancies – represents a great opportunity for LM specialists, especially for those involved in the design and realization of International Vocational Education and Training (VET) services.

From a technical point of view, reasoning with Web job advertisements needs to address two relevant and distinct tasks: first, data have to be gathered from selected on-line job portals, which rely on different languages, data models, and taxonomies; second, job vacancies are mostly stored as unstructured documents and published as plain text. In this context, job offer writers often use job titles and a lexicon that differ significantly from the ones used in standard classification systems. From an international perspective, these challenges can prevent the effective monitoring and evaluation of LM dynamics and policies across national borders. Furthermore, an in-depth knowledge of the LM domain is required to deal with job vacancies, LM taxonomies, and to understand LM dynamics as well.

### 1.1 The potential impact of labor market data

There is a widely recognized need and a growing interest in the early monitoring and analysis of LM requirements to obtain updated information and up-to-date analyzes on LM dynamics. Indeed, LM analysts increasingly recognize the value of supporting their activity through evidence-based decision-making, as highlighted in the European Commission's Communication "New Skills for New Jobs",<sup>1</sup> and in one of the flagship initiatives of the Europe 2020 strategy, the "Agenda for new skills and jobs".<sup>2</sup> Furthermore, in 2010 the European Commission has published the communication "A new impetus for European Cooperation in Vocational Education and Training (VET) to support the Europe 2020 strategy",<sup>3</sup> aimed at promoting education systems in general, and VET in particular. In 2016 the European Commission has remarked the importance of VET's activities, as they are "valued for fostering job-specific and transversal skills, facilitating the transition into employment

<sup>1</sup>The Commission Communication "New Skills for New Jobs" (COM(2008) 868, 16.12.2008)

<sup>2</sup>The Commission Communication "An Agenda for new skills and jobs: A European contribution toward full employment" (COM(2010) 682, 23.11.2010)

<sup>3</sup><https://goo.gl/Goluxo>

and maintaining and updating the skills of the workforce according to sectoral, regional and local needs”.<sup>4</sup>

The classification of Web job vacancies over a standard taxonomy of occupations fulfills the above requirements, rather than using a country-specific or a proprietary taxonomy. Specifically, it provides to researchers, LM analysts, and policy makers, a *lingua franca* for understanding the LM dynamics over several countries. Indeed, one of the main activities of VET agencies is to prepare people for jobs in the context of a specific trade, occupation, or vocation (traditionally in non-academic scenarios). VET agencies usually gather information about the LM by means of survey-based analyzes that are published by European and national institutes of statistics (LFS 2016). Moreover, to better accomplish their mission, VET agencies aim at collecting and analyzing Web job vacancies to deeply understand the LM dynamics, occupations, and trends: (i) by reducing the time-to-market with respect to the one guaranteed by classical survey-based analyzes (official LM surveys actually require up to one year before being available); (ii) by overcoming the linguistic boundaries through the use of standard classification systems; (iii) by representing the knowledge at a very fine-grained level.

## 1.2 Contribution

In this paper, we present WoLMIS, a system that collects job vacancies from several European Web job sites and classifies them with respect to an international taxonomy of occupations. The system has been designed and developed for the Cedefop agency,<sup>5</sup> within a research tender aimed at investigating the practical usefulness of Web job vacancies in supporting and integrating the agency’s activities.<sup>6</sup> WoLMIS is running on the Cedefop data center since June 2016, gathering and classifying job vacancies from 5 EU countries, namely: United Kingdom, Ireland, Czech Republic, Italy, and Germany.

In particular, WoLMIS has been designed and implemented to perform two distinct and relevant tasks: (i) to collect vacancies by scraping selected Web job boards for the countries involved in the project, dealing with different data models and structures; (ii) to classify each job vacancy with respect to the International Standard Classification of Occupations (ISCO) taxonomy,<sup>7</sup> by performing text classification based on machine learning techniques. ISCO is at present one of the most accurate and standardized taxonomy of labor occupations at the basis of several national and international classification systems. It is worth remarking that classifying job vacancies according to a standardized taxonomy makes vacancy sets comparable (for statistical purposes) even if they are written in different languages. In this paper, the classification architecture is presented focusing on English written vacancies, but the approach used is general, and it can be easily adapted to process several languages.

In this paper, we present and discuss the main characteristics of WoLMIS by describing and motivating the main technical choices at the basis of its development. We also present the WoLMIS interface and we provide some examples and demos showing the nature of

<sup>4</sup>The Commission Communication “A New Skills Agenda for Europe” COM(2016) 381/2, available at <https://goo.gl/Shw7bI>

<sup>5</sup><http://www.cedefop.europa.eu/>

<sup>6</sup>Real-time Labor Market information on skill requirements: feasibility study and working prototype. Cedefop Reference number AO/RPA/VKVET-NSOFRO/Real-time LMI/010/14. Contract notice 2014/S 141-252026 of 15/07/2014

<sup>7</sup><http://www.ilo.org/public/english/bureau/stat/isco/>

LM analyzes that can be performed by the proposed system. The architecture described can be easily extended to several languages, however, to evaluation purposes, in this paper we focus only on the English job offers; more specifically a collection of about 2 million job vacancies was collected from the Web. To date, WoLMIS is the only system able to collect and classify multilingual Web job vacancies over the ISCO taxonomy.

The paper is organized as follows: Section 2 presents some related work in the LM domain and describes some preliminary concepts about text classification based on machine learning. In Section 3 the LM scenario is outlined in detail. Section 4 describes the WoLMIS system, the classification process, and its implementation. Section 5 evaluates the proposed system with respect to several text classifiers. Section 6 describes the system interface and some examples of job market analyzes that can be performed by using WoLMIS. Finally, Section 7 draws the conclusions of the paper and illustrates some further research directions.

## 2 Related work

The work described in this paper is framed within job market analysis, which has been a classic topic in human capital economics that attracted researchers since (Zhu et al. 2016). Job market analysis is an important task for many stakeholders, ranging from governmental agencies, firms, to families and vocational schools.

To better frame our work in the complex and variegated job market world we will introduce two perspectives namely, *micro* and *macro*, inspired by the economic literature. Specifically, *microeconomics* is a branch of economics studying the behavior of single individuals and firms, while *macroeconomics* deals with an economy as a whole (considering the performance, structure, behavior, and decision-making aspects of a collectivity) (Samuelson 1974). In the considered context, we refer to the micro perspective when methodology, application, and service focus on providing value to a single entity (e.g., a firm who must select one candidate among several profiles that applied for a job), while we refer to the macro perspective when broad phenomena are managed as a whole (e.g., by providing statistics about a country, or a cross-national region). The border between the two perspectives is sometimes blurry; nevertheless, they will be useful to disentangle a complex world.

The works in the literature discussed in this section are presented according to distinct areas: job market analysis in the macro perspective (Section 2.1), job market analysis in the micro perspective (Section 2.2), and text classification in general (Section 2.3).

### 2.1 The macro perspective

Traditionally, information on the job market has been gathered through sample-based analyzes performed by European or national institutes of statistics (LFS 2016); however, this procedure suffers the limitations outlined in Section 1. In the last years, public administrations started to explore new ways for obtaining detailed and up-to-date information about the LM, e.g., administrative information collected by public administrations (Mezzanzanica et al. 2012, 2015; Boselli et al. 2014). Unfortunately, administrative data are often collected when people are already hired, and only in countries where public administrations store and manage such information. Therefore, these data do not provide any information about the job demand expressed by employers.

In the last years, classifying and analyzing job vacancies and occupations published on the Web has been a challenging issue. Most of the classification approaches proposed

in the literature are based on different taxonomies (both public and proprietary), and a lot have been performed by experts in a non-automatic way. For instance, Elias and Purcell (2004) the authors study the Graduate LM in the UK using job vacancies classified by experts according to the Standard Occupational Classification (SOC2000) delivered in 2000,<sup>8</sup> without any automated tool.

In the work of Zhu et al. (2016) on-line job vacancies collected from Chinese Web sites (from 2014 to 2015) are analyzed using topic detection techniques to identify recruitment market latent topics and their changes over time. The latter differs from our approach since we are aimed at collecting thorough data about occupation demand and we map the data to the ISCO taxonomy concepts, our approach can be easily extended to encompass other languages so that data derived from different languages can be compared. Conversely, the topic modeling approach used by Zhu et al. (2016) is strongly bound to the text processed, i.e., the detected topics are expressed in the source language, making a cross-language comparison difficult. The approach described by Zhu et al. (2016) can complement the work described in this paper. Our system stores the job vacancies titles and full description texts which can be later used to perform further analysis, like the one done by Zhu et al. (2016).

In the work of Xu et al. (2017) the authors scraped job vacancies from Chinese Web sites and labeled them according to the CGCO (People's Republic of China Grand Classification of Occupations). They built a dataset having 102,581 documents classified into 465 categories. In Xu et al. (2017) the authors have a similar goal with respect to the one described in our work, i.e., creating instruments for analyzing the Web LM in a macro-economic perspective. The main differences between our work and Xu et al. (2017) are: we classify job offers according to the ISCO taxonomy; we deal with alphabetical languages while Xu et al. (2017) the logographic Chinese is used; furthermore, in our work, we built an architecture to deliver information to the final users. It is worth to note that, both SVMs and several types of Neural Networks were used on the dataset described by Xu et al. (2017). Although there is no direct comparison between the SVM and the Neural Network classifiers by Xu et al. (2017) (SVMs were used in the semi-supervised dataset building, while Neural Networks were used on the final dataset), similar performances are reported: SVMs achieved an accuracy rate of 0.891 during a hyper-parameter tuning (on a training-test validation during the dataset building) while a Convolutional Neural Network (CNN) achieves a 0.8895 accuracy (in a 10-fold cross-validation on the final dataset). Although comparing the two values does not make sense (because they have been achieved in different phases of the dataset building), it is interesting to notice that the SVM and the Neural Network classifiers applied to the Chinese language have similar performances, close to the results described in this paper in Section 5.2.

## 2.2 The micro perspective

The automatic extraction of meaningful information from unstructured texts has been mainly devoted to support the e-recruitment process (Lee 2011), e.g., to help human resource departments to identify the most suitable candidate for an open position from a set of applicants or to help a job seeker in identifying the most suitable open positions. For example, the work described by Singh et al. (2010) proposes a system which aims to analyze candidate profiles for jobs, by extracting information from unstructured resumes through the

<sup>8</sup>For more information on SOC2000, the interested reader can refer to SOC2000 (2016).

use of probabilistic information extraction techniques as Conditional Random Fields (Laferty et al. 2001). Differently, by Yi et al. (2007) the authors define Structured Relevance Models (SRM) and describe their use to identify job descriptions and resumes vocabulary, while by Hong et al. (2013) a job recommender system is developed to dynamically update the job applicant profiles by analyzing their historical information and behaviors. Finally, the work described by Poch et al. (2014) illustrates the use of supervised and unsupervised classifiers to match candidates to job vacancies suggesting a ranked list of vacancies to job seekers.

Several works in the literature focused on the information extraction tasks performed on the dataset of computer-related job postings from the Usenet newsgroup `austin.jobs` (Califf 1998). The DiscoTEX system (Nahm and Mooney 2001) has been applied to mine Usenet job postings and resumes. Information extraction techniques – namely the RAPIER (Califf and Mooney 1999) and BWI (Freitag and Kushmerick 2000) systems – have been used to extract text portions from documents to fill fields like *title, state, city, country, area, required years of experience*. Unfortunately, the strings extracted to fill specific data fields can vary substantially across documents even though they refer to the same real-world entity (Mooney and Bunescu 2005). In our context, we can potentially deal with 436 ISCO codes, each one corresponding to a huge set of job titles (much more if several languages are considered), and where the same job title could be related to different ISCO codes depending on the specific job requirements, e.g., the job title “Programmer” may refer either to the ISCO category 2512 (Software Developer) or to 2514 (Applications Programmers) depending on the combination of required tasks, skills, and experiences. This makes the approaches based on information extraction described above unsuitable for ISCO codes occupation classification in the context of producing statistics, or in general in a macro perspective.

Information extraction was also used for processing job vacancies sent by e-mail. In the work of Kessler et al. (2007) an information extraction system was designed for alleviating the workload of an agency specialized in e-recruiting. The agency clerks receive the open position descriptions by e-mails and then publish the vacancy information on different online job boards on behalf of clients. Before the introduction of the information extraction system they had to manually extract information from the e-mails like the *contract, salary, location, reference, duration of mission*, etc., as required by the publishing Web sites.

Concerning firms, their need to automatize Human Resource (HR) department activities is strong; as a consequence, a growing amount of commercial skill-matching products have been developed in the last years, for instance, BurningGlass,<sup>9</sup> Workday,<sup>10</sup> Pluralsight,<sup>11</sup> EmployInsight,<sup>12</sup> and TextKernel.<sup>13</sup> The Google Cloud Job API<sup>14</sup> is a service announced in 2016 for classifying job vacancies and for identifying skills w.r.t a standard taxonomy, i.e., an extension of O\*NET<sup>15</sup>. O\*NET is on its own an extension of the Standard Occupational Classification (SOC) system<sup>16</sup> developed by the U.S. Bureau of Labor Statistics.

<sup>9</sup><http://www.burning-glass.com/>

<sup>10</sup><http://www.workday.com/>

<sup>11</sup><https://www.pluralsight.com/>

<sup>12</sup><http://www.employinsight.com/>

<sup>13</sup><http://www.textkernel.com/>

<sup>14</sup><https://cloud.google.com/jobs-api/>

<sup>15</sup><http://www.onetcenter.org/taxonomy/2010/list.html>

<sup>16</sup>[http://www.bls.gov/soc/major\\_groups.htm](http://www.bls.gov/soc/major_groups.htm)

The service is actually at a *closed alpha* release phase, it is actually advertised as an instrument to improve matching among open positions and applicants. It highlights the added value of reasoning with Web job vacancies using a taxonomy as baseline. To date, the only commercial solution adopting the ISCO taxonomy is Janzz.<sup>17</sup> It is a Web-based platform aimed at matching labor demand and supply in both public and private sectors. It also provides an API-based access to its knowledge base, but it is not aimed at classifying job vacancies.

In the paper by Javed et al. (2016) a multi-label classifier was developed to categorize CareerBuilder<sup>18</sup> job vacancies according to O\*NET.<sup>19</sup> The job titles used in job vacancies may have a high degree of synonymy or semantic closeness, and the granularity of the O\*NET system was considered unsuitable for CareerBuilder search and recommendation activities (Javed et al. 2016). For example, the activity of *software development* is represented in O\*NET by two entities which encompass several job titles such as *Hadoop Engineer*, *.Net Developer*, *Machine Learning Engineer*, and *Java Developer* among others. Both the synonymy and the granularity issues motivated the development of a new taxonomy and a multi-label classification system (Javed et al. 2016). The approach described by Javed et al. (2016) is different from the one proposed in this paper. Our aim is to develop a job occupation classifier system for a macro perspective, while Javed et al. (2016) focus on the micro perspective, i.e., to support customer search and recommendation activities. In the macro perspective it is paramount to reconcile different but equivalent job titles to the same taxonomy concepts, while on the micro perspective, each job title written in a query must match all the related entities. The two problems are quite different and require different solutions.

In the work of Sun et al. (2015) the authors developed a spam classification system for Spanish messages exchanged by LinkedIn users. The goal is to classify messages exchanged by social network users, messages are not necessarily job postings, e.g., they may be unsolicited advertisements. Since a suitable training set is available for the English language but not for the Spanish one, automatic translation and transfer learning are investigated to achieve a suitable classifier for the Spanish messages. Differently, our work focuses on classifying job advertisements.

The works described in this section prove that the interest for effective solutions in the domain of LM is strong and active, but all the proposals presented so far differ from our approach. In fact, our system focuses on the macro perspective, i.e., it classifies job vacancies according to a standard taxonomy to create a (language-independent) knowledge-base for analysis purposes and a system for making available such knowledge-base to a broad set of stakeholders. By means of WoLMIS, Web job vacancies can be analyzed over the geographic dimension ranging from a whole European perspective to the finest granularity of the single vacancy. The results produced by our system can complement sample-based statistics produced by official institutions, which focus on a country-wide scope (or large country subregions), and whose sample-based statistical significance decreases when small territorial environments are investigated (Cesarini et al. 2007).

<sup>17</sup><https://www.janzz.jobs/>

<sup>18</sup><http://www.careerbuilder.com/>

<sup>19</sup>The previously cited extension of the Standard Occupational Classification (SOC) system developed by the U.S. Bureau of Labor Statistics.

## 2.3 Text classification

In the recent literature, *text classification* (TC) has proven to give good results in extracting knowledge from many real-life Web-based data such as, for instance, those gathered by institutional scientific information platforms (Koperwas et al. 2016), or microblogs and other social media platforms (Andrews et al. 2016; Ceci and Malerba 2007; Kanan and Fox 2016; Zubiaga et al. 2015), in many different research areas such as opinion spam detection (Jindal and Liu 2008; Viviani and Pasi 2017) and sentiment analysis (Bifet and Frank 2010; Pang et al. 2002; Perea-Ortega et al. 2013; Vilares et al. 2015). However, text classifiers have not been applied yet to the classification into the ISCO taxonomy of job vacancies written in natural language. Since the system presented in this paper relies on these techniques to assign job vacancies gathered from the Web to the proper label within the ISCO taxonomy, in this section we first provide some basic concepts on text classification.

Text classification has been an active research topic since the early 90s. It has been defined as “the activity of labeling natural language texts with thematic categories from a predefined set” (Sebastiani 2002). Most popular techniques are based on the *machine learning* paradigm, according to which an automatic text classifier is created by using an inductive process able to learn, from a set of pre-classified documents, the characteristics of the categories of interest. The case in which one category must be assigned to each document is called *single-label* classification, while *multi-label* classification is the case when many categories may be assigned to the same document. In our work, a single-label classifier has been used as a component of WoLMIS, as it will be detailed in Section 4.

For decades, constructing a machine learning system required considerable expertise to design the feature extraction phase to transform the raw data into input features for a (machine learning) classifier (LeCun et al. 2015). Word representation before feeding a classifier can be obtained by performing word selection or by replacing words with continuous value representations (e.g., word embeddings (Turian et al. 2010) like word2vec (Mikolov et al. 2013)), or by using a classifier able to discover word representations, or a combination thereof. Deep learning methods are representation-learning methods, i.e., methods that allow a machine to be fed with raw data and to automatically discover the representations needed for classification (LeCun et al. 2015). Several deep learning algorithms have been proposed for text classification based on Convolutional or Recurrent Neural Networks (Sayfullina et al. 2017), namely: fastText (Joulin et al. 2016), Conv-GRNN (Tang et al. 2015), and LSTM-GRNN (Tang et al. 2015). Both (Joulin et al. 2016) and (Tang et al. 2015) recognize SVMs employing the  $n$ -gram frequency representation (the one used in this paper) as the standard for comparison.

In the considered context, vacancy titles are a concise summary of the required occupations, as illustrated in Section 3.1.2, and any attempt to perform word selection on titles was detrimental (excluding stop-word removal and stemming).<sup>20</sup> The contribution of the full descriptions (where word selection might play a more significant role) is minimal as showed in Section 5.2. We decided, therefore, to employ linear SVMs and bag-of-words in the first WoLMIS prototype, considering the additional computational effort that the deep learning algorithms requires, and the fact that they achieve about 5% improvement w.r.t. to the approach based on SVMs and bag-of-words described by Joulin et al. (2016) and Tang

<sup>20</sup>As it will be illustrated in Section 5.2 in Table 4, the 10% of (the most representative) title words are enough to achieve 80% of classification accuracy. Nevertheless, the table shows that the best performances are achieved using all the title words.



et al. (2015) in a sentiment-classification scenario (thus, with few labels as in our case, and with more training example per label).

Since the long-term goal of the project is to build a classification pipeline for each European language, the computational issues have been carefully considered and for the first prototype SVMs were chosen w.r.t. deep learning algorithms. Furthermore, Neural Network algorithms leveraging word embeddings would require an extra task, i.e., to derive word embeddings for each specific language dealt.

The occupation classification pipeline described in this paper focuses on the English language. Nevertheless, the proposed approach is general and can be easily extended to different European languages (i.e., it is enough to select the suitable stop-word set and stemming algorithm). In this way, several pipelines can be easily built (one for each language) where each pipeline output is expressed on the ISCO taxonomy; thus, the results are comparable even if the input documents are written in different languages.

### 3 The labor market scenario

In this section, we provide a short introduction about the influence of Information and Communication Technology (ICT) on the job market in Europe. Moreover, we introduce the concept of LM Intelligence (LMInt) as a way of extracting insights from the job market, and we describe the ISCO taxonomy. Finally, we provide some examples of the considered scenario to discuss some choices on which the development of LMisystem is based.

The development of ICT has strongly changed the way in which the job market operates.<sup>21</sup> The European Commission and each national Public Employment Service (PES) have integrated some Web services and tools into their employment support, placement and information services portals, e.g., the European job mobility Web site EURES.<sup>22</sup>

Due to the huge number of vacancies published in Web job portals in the last years, we can assume that a lot of useful insights on several economic sectors can be discovered by gathering and analyzing these information sources. This leads to the definition of systems and tools for a new research area that we call LM Intelligence.

#### 3.1 Labor market intelligence

Currently, the European Commission is focusing on the concept of Labor Market Information (LMI). This term refers to the information related to the LM, such as skills, competencies, qualifications, and occupations, in addition to all the ICT techniques and services for labor information management, in particular, mobility-related services.

We adopt the term Labor Market Intelligence (LMInt) to refer to the analysis of various forms of job market data and information to the purpose of making them available to organizations for integrating their decision making processes, expand their services, and provide ad-hoc policies. An important step toward the generation of Labor Market Intelligence is the identification of standard taxonomies for occupations and skills, to foster the circulation of information in a multi-language job market like the European one.

<sup>21</sup>The market in which workers find an employment, employers find available workers, and wage rates are determined.

<sup>22</sup><https://ec.europa.eu/eures/public/en/homepage>

### 3.1.1 A description of the ISCO taxonomy

As introduced in Section 1, the system described in this paper employs the International Standard Classification of Occupations (ISCO) taxonomy for classification purposes. ISCO has been developed by the International Labor Organization as a four-level hierarchy; the interested reader can refer to ISCO (2012) for more details. To date, ISCO is one of the most adopted taxonomies in Europe and it is a reference worldwide.

ESCO (a multilingual taxonomy of European Skills, Competences, Qualifications and Occupations) is an ongoing project part of the Europe H2020 strategy and extends ISCO through (i) a further level of fine-grained occupation descriptions, and (ii) a taxonomy of skills, competencies and qualifications.

Once a job vacancy has been classified correctly with respect to the lowest level of the ISCO hierarchy (i.e., the fourth level), one can navigate the ISCO-ESCO cross-linkage to access the list of occupation examples containing a set of skills/competences and qualifications requested by the corresponding occupation. This motivates the use of ISCO as a baseline for the classification process in WoLMIS.

### 3.1.2 An example of job vacancies

In the on-line job market, a *job vacancy* is a document containing two main text fields: a *title* and a *full description*. The title shortly summarizes the job position, while the full description field usually includes the position details and the relevant skills the employee must possess.

Table 1 shows two job vacancies extracted from specialized Web sites. Even if both samples seek for *computer scientists*, the differences in terms of job requirements, skills, and corresponding educational levels are quite strong. The first job vacancy (A) is looking for a software developer, while the second job vacancy (B) is searching for a less skilled candidate, i.e., an ICT technician; indeed, in the latter case, the only requested abilities are to *use* some solutions, and the knowledge of a programming language (optional) that is usually taught in some professional high-schools.

**Table 1** An example of job vacancies. Vacancy titles are in bold followed by full descriptions

<p>(A) <b>Experienced Developer.</b> “Looking to recruit a software developer to join our dynamic R&amp;D team to support work on Windows-based software for current and upcoming products. A flexible, self-starting attitude is essential along with a strong motivation to learn new skills as required. The ideal candidate will have at least two to three years experience working in a fast and dynamic small team. Experience in Python is a key requirements for my client. A degree in computer science / computer engineering is preferable, but other engineering / science graduates will be considered if they have software development experience.” [ISCO 2512: Software Developer]</p>	<p>(B) <b>Application Consultant.</b> “We are seeking for an application consultant that will support our internal SW engineers in the realization of ad-hoc ERP software. The ideal candidate should have (at least) a high level degree and 2+ years experience in using both Sharepoint-MS CRM and Magic. Coding skills on VB are appreciated.” [ISCO 3512: ICT user support technicians]</p>
<p>Workplace: Dublin</p>	<p>Workplace: Inner London</p>
<p>Contract type: Permanent</p>	<p>Contract type: Unlimited Term</p>

### 3.2 The classification of Web job vacancies

In our approach, the task of Web job vacancy classification is treated as a single-label classification problem (see Section 4 for details on the development phase). Each job vacancy is assigned to just one ISCO code over a set of 436 possible codes belonging to the 4<sup>th</sup> level of the taxonomy. Notice that a single-label classification identifies a partition over the classified job vacancies set, and this simplifies the data interpretation and comparison (especially for non-expert users). Conversely, considering multi-label classification, a set of vacancies could be related to an arbitrary number of occupation codes, and this would make the comparison of occupation distributions over different vacancy sets a difficult task.

Using a single-label classification (i.e., one ISCO code per vacancy) would not be suited in the following cases: (i) an advertisement describing several different open positions (e.g., “we are looking for a secretary and a facility manager”); however, the number of such cases is negligible, as evaluated in the source-selection phase by the expert of the ENRLMM network;<sup>23</sup> (ii) some open positions may require people to perform a mix of different occupations (e.g., “we are looking for a plumber who can drive trucks to deliver materials to yards . . .”); this might likely happen in Small and Medium Enterprises (SMEs); nevertheless, in mixed descriptions a *dominant* occupation can be identified in most of the cases. By carefully balancing the added value of data interpretation and comparison, and the frequency of cases (i) and (ii), we opted for a single-label classification.

In addition to this, by exploiting ESCO we can easily complement occupation data with the whole set of skills that each occupation is expected to have as outlined in Section 3.1.1.

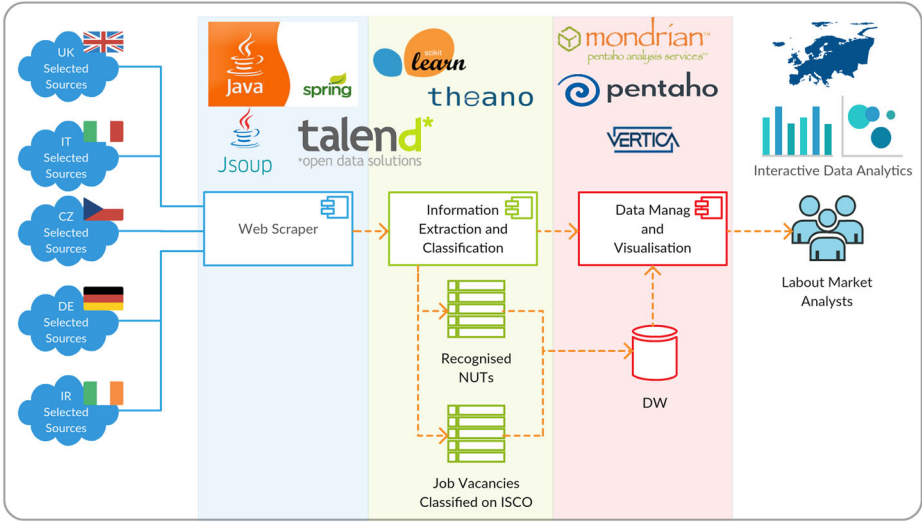
Finally, as illustrated in Table 1, job vacancies often contain information on the workplace and the contract type. These data are usually expressed by using site-specific taxonomies.

## 4 WoLMIS: a system for job vacancies classification

In this section, we describe the WoLMIS system. Figure 1 depicts the system architecture and the data pipeline. The *Web Scraper* module retrieves job vacancies from the selected Web sources in batch mode once per week. The *Information Extraction and Classification* module identifies the portions of text that are useful for the classification task. Then, this module classifies each job vacancy with respect to the ISCO taxonomy, as detailed in Section 5, and loads the outcomes of this step into a data warehouse. This allows OLAP analyzes over three dimensions, i.e., NUT geographical levels, ISCO occupations (4<sup>th</sup> level), and NACE<sup>24</sup> economical sectors. The latter information is available from the Web pages of the scraped job vacancies. Finally, the resulting knowledge is made available to the end user through an on-line interactive dashboard, based on the *Data Mining and Visualization* module.

<sup>23</sup>The European Network on Regional Labor Market Monitoring (ENRLMM 2016).

<sup>24</sup>The European classification system for economical sectors, see [http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Statistical\\_classification\\_of\\_economic\\_activities\\_in\\_the\\_European\\_Community\\_\(NACE\)](http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Statistical_classification_of_economic_activities_in_the_European_Community_(NACE))



**Fig. 1** The WoLMIS workflow and architecture as deployed at the Cedefop Agency

### 4.1 Machine learning in WoLMIS

The WoLMIS system adopts a *supervised learning* approach to implement a *single-label classifier* for identifying ISCO codes from textual job vacancies. As illustrated in Section 2.3, text classification using the machine learning (ML) paradigm is a general inductive process that automatically builds an automatic text classifier by learning, from a set of pre-classified documents, the characteristics of the categories of interest (Sebastiani 2002). The latter is an example of supervised learning, where the algorithm is trained on documents whose class is known. After the learning, the algorithm can classify documents whose label is unknown.

Several machine learning techniques have been selected and implemented in WoLMIS for developing the text classifier, and they have been comparatively evaluated on a data set of job vacancies in English. The implemented techniques are Support Vector Machines (SVMs), in particular SVM Linear (Fan et al. 2008), and SVM RBF Kernel (Müller et al. 2001), Random Forests (RFs) (Breiman 2001), and Artificial Neural Networks (ANNs) (Haykin 1999).

Some characteristics of the employed techniques are presented here below, while their comparative evaluation in classifying the job vacancies in the considered data set is reported in Section 5.

Classifiers based on SVMs try to identify a separation hyperplane in the classifier input space by maximizing the distance among the nearest elements (of a category and of the remaining ones). SVM-based classifiers have good generalization ability; moreover, according to Joachims (1998), they are well suited to the particular characteristics of texts, namely high-dimensional feature spaces, few irrelevant features (dense concept vector), and sparse instance vectors. Their main drawback concerns the difficulty of interpreting the model generated and the SVMs sensitivity to a proper parameter tuning.

Random Forests are an ensemble learning method that operates by constructing a multitude of decision trees at training time. RFs have been successfully employed in a wide range

of applications, and they are fast to train. Breiman in Breiman (2001) has shown that RFs do not overfit, despite the number of trees employed in the combination.

Artificial Neural Networks are classifiers that artificially simulate the behavior of brain neurons. The advantages of Artificial Neural Networks are their robustness to noisy data and their ability to represent linear and non-linear functions of various forms and complexities. Disadvantages include the need of a proper parameter tuning and the difficulty in interpreting the concepts learned by the ANNs. In WoLMIS, Multi-Layer Perceptron (MLP) ANNs have been employed: in particular, a feed-forward Neural Network with a 3-layers configuration (of which 1 hidden layer) has been implemented.

## 4.2 Feature extraction

The feature extraction phase is a critical step in every machine-learning-based classification process. As outlined in Section 3.1.2, a Web job vacancy is composed of a title and a full description. Titles usually shortly summarize the desired occupation profile. Full descriptions generally include further information like the desired skills, the presentation of the company, the legal disclaimers, etc. Though in most cases the title alone provides enough information to correctly identify the most suitable ISCO occupation code for a job vacancy, often additional information from the full text description can be used to improve the classification effectiveness, as we pointed out in Amato et al. (2015a, b), Marrara et al. (2017), Lembo et al. (2015), and Sheth et al. (2017), and as it is confirmed by the results that will be illustrated in Section 5.2.

For the above reasons, two different feature extraction phases were performed, by separately managing titles and full descriptions. For both phases, the text pre-processing includes various steps: (i) HTML tag removal, (ii) tokenization (i.e., sentences are split into separate words and punctuation is removed), (iii) lower case reduction, (iv) stop-words removal (e.g., elimination of common and low informative words as ‘the’, ‘of’, ‘as’, by using a predefined list), and (v) stemming (words are reduced to their stem).

Focusing on job vacancy descriptions, besides the above standard pre-processing steps, a set of standard sentences like:

- “. . . is an equal opportunities employer” or
- “Replying to this advertisement means that you provide us with authorization to add you to our database . . .”

need to be dropped out to keep only the most informative text, thus improving the classification performance. The technique adopted for this purpose was to reduce the text associated with job vacancy descriptions by extracting *word windows* containing ‘sentinel’ *words* and *expressions* identified by domain experts. For example, the sentence: “we are looking for . . .” is very frequently followed by a description of the desired occupation. Extracting word windows was not deemed necessary on titles since their contents usually describe the desired occupations in a concise yet complete way, and dropping words from titles is very likely to cause a loss of information.

The pre-processed titles and word windows extracted from the job vacancy descriptions were finally processed by using a *bag-of-words* representation. We extracted *n*-grams<sup>25</sup> in the form of unigrams, bigrams, 3-grams, and 4-grams frequencies, which have been stored

<sup>25</sup>Generally speaking, an *n*-gram is a set of *n* consecutive words.

in ad-hoc data structures. Textual features represented in this way were used for the comparative evaluation of the machine learning techniques implemented in WoLMIS, described in the following section.

## 5 Evaluation of the classification

This section describes the evaluation of the classifiers implemented in WoLMIS on the data set constituted by the ISCO-labeled English Web job vacancy collection that will be described in Section 5.1. The classification results on this dataset are shown in Section 5.2.

### 5.1 The Web job vacancy collection

A benchmark dataset was prepared by crawling the Web, by selecting a set of vacancies from those collected, and by labeling each item with the most appropriate 4<sup>th</sup> level ISCO code. The benchmark dataset was built with the support of a team of labor experts belonging to the ENRLMM network. The dataset creation process has been defined as follows.

- The team of experts identified the Web sites playing a prominent role in the domestic job market, for each country involved in the project. Considering the English-speaking countries, four Web sites for the UK and three Web sites for Ireland were selected by the domain experts.
- Each selected Web site was crawled at least once per week,<sup>26</sup> from June to September 2015, by identifying and downloading the new job vacancies.
- A set of vacancies were selected and labeled by the domain experts; this reduced dataset was then used to perform training, test, and validation of the machine learning based classifiers.

The scraping phase gathered 2,295,603 job vacancies, of which about 20% were identified as duplicates. This is due to the fact that employers are akin to publish the same job offer on several Web sites at the same time, thus a duplicate detection activity was performed on the basis of both (i) time elapsed since the first publication date, and (ii) the job vacancy content similarity. In this way, the dataset was reduced to 1,806,337 unique *unlabeled* job offers. Furthermore, a language recognition task was performed with the aim of excluding non-English job vacancies. The number of non-English vacancies was negligible (we found some vacancies in Scots Gaelic<sup>27</sup> among the non-English vacancies).

Each Web page containing a job vacancy was processed to extract the title and the full description. From the collected (and duplicate-free) vacancies, a subset was selected and labeled by the domain experts according to the most appropriate ISCO code. This result well describes the real Web job market situation since some occupations are really under-represented or totally absent on Web sites. The interested reader can refer to Beblavý et al. (2016), Carnevale et al. (2014), and Kureková et al. (2015) for further details. As an example, two occupation categories described in ISCO, and not available in the collection

<sup>26</sup>The visiting frequency was tuned for each Web site taking into account: the publishing rate, the average time an advertisement is kept on-line, and suggestions of the Web masters who accepted to collaborate with the project.

<sup>27</sup>Actually, there are some vacancies, mostly looking for language teachers.

of English vacancies we gathered on the Web, are: [9624<sub>ISCO</sub>, Water and firewood collectors],<sup>28</sup> and [7541<sub>ISCO</sub>, Underwater divers].

The labeled vacancies were randomly partitioned into training and test sets (Crowther and Cox 2005). The partition was performed to split the vacancies assigned to a certain occupation code as follows: 75% in the training set and 25% in the test set, respectively 38,509 and 12,813 vacancies. The classifier evaluation described in the following was executed using the labeled vacancies.

## 5.2 Comparative evaluation of the implemented classifiers

The WoLMIS classification system was built using the Scikit-learn framework (Pedregosa et al. 2011) running on a Intel Xeon machine with an Intel Core i7 CPU, 12GB RAM, and Ubuntu 14.04 as operating system. The SVM Linear, SVM RBF, Random Forests, and ANN classifiers were implemented by using the following Scikit-learn APIs: `sklearn.svm.LinearSVC`, `sklearn.svm.SVC`,<sup>29</sup> `sklearn.ensemble.RandomForestClassifier`, and `sklearn.neural_network.MLPClassifier`.

The effectiveness of the implemented classifiers has been evaluated based on the measures of *precision*, *recall*, and *f1-score*, and *accuracy*, defined as follows:

$$\begin{aligned} \textit{precision}_i &: \frac{TP_i}{TP_i + FP_i} & \textit{recall}_i &: \frac{TP_i}{TP_i + FN_i} \\ \textit{f1}_i &: 2 \cdot \frac{\textit{precision}_i \cdot \textit{recall}_i}{\textit{precision}_i + \textit{recall}_i} \end{aligned}$$

where:

- $TP_i$  (True Positives) represents the number of vacancies that have been correctly assigned to a class  $c_i$ ;
- $FP_i$  (False Positives) represents the number of vacancies that have been wrongly assigned to the class  $c_i$ ;
- $TN_i$  (True Negatives) is the number of vacancies that have been correctly not assigned to the class  $c_i$ ;
- $FN_i$  (False Negatives) is the number of vacancies that have been not classified in  $c_i$  even if they represent an instance of the occupation coded by  $c_i$ .

For evaluation purposes, the technique of *macro-averaging*<sup>30</sup> has been applied to summarize the several class values. The measures are computed for each considered class (ISCO code) and are then averaged over the considered classes (Sebastiani 2002).

The *accuracy* measure is the ratio of true positives to the total number of cases examined and is defined as follows:

$$\textit{accuracy} : \frac{\sum_{i=1}^{426} TP_i}{\sum_{i=1}^{426} (TP_i + FN_i)}$$

and is already a summary value, defined overall the 426 ISCO codes.

<sup>28</sup>According to (ISCO 2012), “Water and firewood collectors” gather water and firewood, and transport them on foot or using hand or animal carts.

<sup>29</sup>`sklearn.svm.LinearSVC` is a wrapper around the `liblinear` library (Fan et al. 2008), while `sklearn.svm.SVC` is a wrapper around the `libsvm` library (Chang and Lin 2011).

<sup>30</sup>Also known as weighted averaging.

**Table 2** Classifiers Evaluation using only  $n$ -grams extracted from Web vacancy titles

Only Titles Classifier	Precision	Recall	F1-Score	Accuracy
SVM Linear (Fan et al. 2008)	0.881	0.880	0.879	0.881
SVM RBF Kernel (Müller et al. 2001)	<b>0.884</b>	<b>0.882</b>	<b>0.880</b>	<b>0.882</b>
Random Forest (Breiman 2001)	0.879	0.878	0.875	0.877
Neural Networks (Haykin 1999)	0.853	0.854	0.851	0.854

Precision, Recall, and F1-Score values are the weighted average of the values obtained for each ISCO code. Accuracy is computed considering all classes at once. The classifiers were evaluated on the test set. Bold values indicate the best value among the classifiers

In the proposed approach, the training and test sets for the English language were used in two ways.

- First, they have been employed to tune the parameters of the candidate classifiers, i.e., a grid-search was performed over each classifier parameter space to identify the values maximizing the classification performances. The grid-search was executed using a  $k$ -fold evaluation over the training set (using  $k=10$ ), i.e., the training set was randomly split into  $k$  subsets, for  $k$  times the classifier was trained on  $k - 1$  subsets and evaluated on the remaining one. This operation was executed for each combination of classifier parameters to be evaluated.
- Then, by using the best combination of parameters as returned by the grid search, the classifiers were trained on the training set and evaluated on the test dataset.

Two experiments were performed on the same collection of English vacancies by using two different feature sets. The first set, denoted as  $FS_1$ , includes the  $n$ -gram frequencies extracted from titles; the second set, denoted as  $FS_2$ , includes the  $n$ -gram frequencies computed over both titles and full descriptions.

In the first experiment, which is based on  $FS_1$ , the training phase was performed over a set of 15,987 features ( $n$ -grams, where  $n$  ranges from 1 to 4) extracted from the vacancy titles of the training set; the testing phases were performed using the test set. In this experiment the four classifiers performed in a very similar way, obtaining good results as reported in Table 2; in particular, it emerges that SVM RBF and SVM Linear perform better with respect to the other classifiers, and SVM RBF performs slightly better w.r.t. SVM Linear.

In the second experiment, which considers  $FS_2$ , the training phase was performed on a much larger set of features, i.e., 637,673 different  $n$ -grams, extracted from the vacancy titles and full descriptions of the training set. The results of the second experiment evaluated on the test set considering  $FS_2$  are reported in Table 3.

In particular, the SVM Linear classifier f1-score improved by 3%. Since the number of features is far greater than the number of vacancies, it is not surprising that now the SVM RBF Kernel does not outperform the SVM Linear classifier (Hsu et al. 2003). Furthermore, since the SVM Linear and the (3-layer) Neural Network<sup>31</sup> have similar and good performances, we can conclude that the job vacancy classification problem over the considered job vacancy collection can be suitably addressed by a linear classifier, which is characterized by a reduced training time and good classification performances over the ISCO digit codes.

<sup>31</sup>A 3-layer (of which 1 hidden layer) Neural Network has the ability to properly address linear classification problems (Jain et al. 1996; Lippmann 1987).



**Table 3** Classifiers Evaluation on  $n$ -grams extracted both from titles and full description windows

Titles and Full Description Windows				
Classifier	Precision	Recall	F1-Score	Accuracy
SVM Linear	<b>0.910</b>	<b>0.909</b>	<b>0.908</b>	<b>0.909</b>
SVM RBF Kernel	0.892	0.677	0.750	0.677
Random Forest	0.813	0.814	0.809	0.814
Neural Networks	0.874	0.874	0.872	0.874

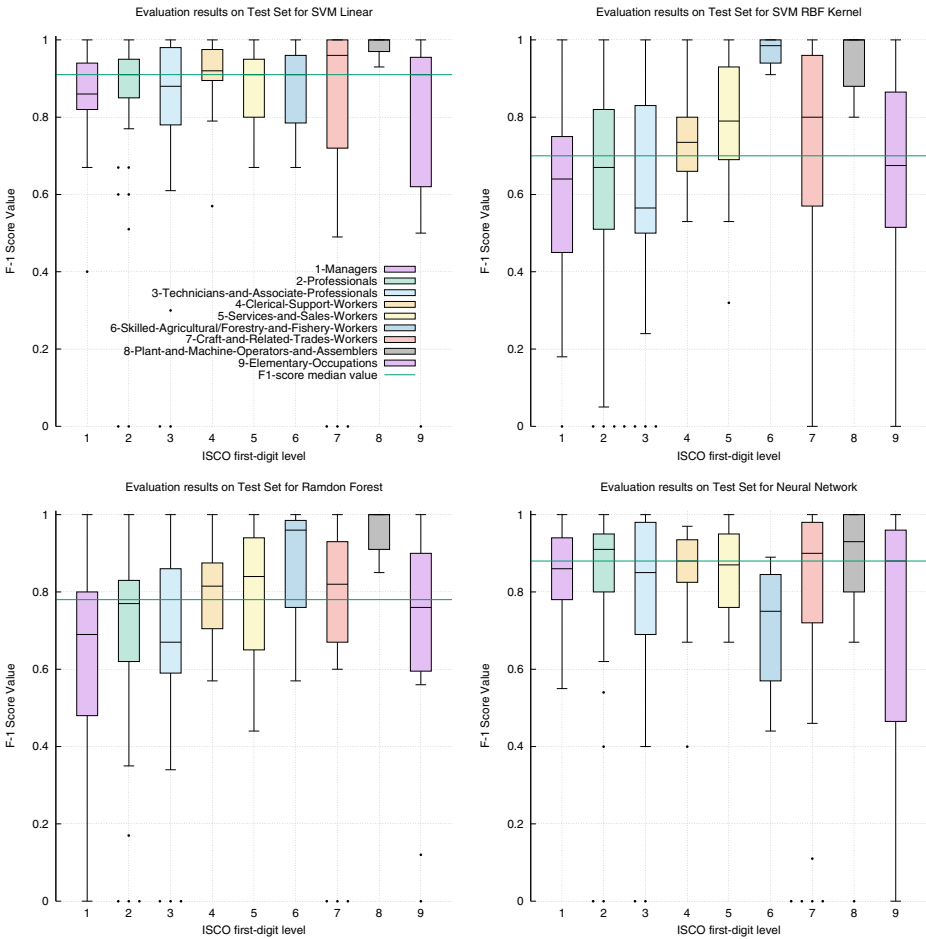
Precision, Recall, and F1-Score values are the weighted average of the values computed for each ISCO code. Accuracy is computed considering all classes at once. The classifiers were evaluated on the test set. Bold values indicate the best value among the classifiers

We also used box plots to evaluate the f1-score measure of each classification algorithm over the first-digit of the ISCO taxonomy, that identifies 9 distinct occupation groups, from 1 up to 9. Box plot is a well-known statistical technique used in exploratory data analysis to visually identify patterns that may otherwise be hidden in a data set by measuring variation changes between different groups of data. In Fig. 2 we report four box-plots, one for each classification algorithm, computed on the  $FS_2$  feature set. Each box-plot shows the distribution of the f1-score value for the respective algorithm over the nine ISCO groups. In this way, the effectiveness of each classification algorithm can be investigated over a specific group of occupations. Considering the f1-score measure, each distribution is partitioned into quartiles as follows: the *box* indicates the positions of the upper and lower quartiles respectively<sup>32</sup>; the *box content* indicates the median value, which is the area between the upper and lower quartiles and consists of 50% of the distribution. The *vertical lines* (also known as *whiskers*) stretch over the extreme of the distribution indicating either minimum and maximum values in the dataset. Finally, *dots* are used to represent upper and lower outliers, namely data items that lie more (less) than 3/2 times the upper (lower) quartile respectively. As it can be observed, SVM Linear still outperforms the other classifiers even considering the overall ISCO distributions at the first level of the hierarchy. Furthermore, SVM Linear obtains a 0.91 median value for the f1-score with respect to the same value obtained by NN (0.88), RFs (0.78), and SVM RBF (0.7).

Furthermore, the impact of a  $\chi^2$  feature selection (Yang and Pedersen 1997) was evaluated against the accuracy of the SVM Linear classifier. The results are illustrated in Table 4. From the results it emerges that 10% of (the most representative) title features are enough to achieve 80% of classification accuracy. Nevertheless, the table also shows that the best performances are achieved using all the title words.

As it emerges from the proposed evaluations, by using machine learning techniques it is possible to achieve good performances in classifying textual job vacancies gathered from the Web with respect to the 4<sup>th</sup> level ISCO taxonomy. Based on the best results obtained, the WoLMIS system that has been deployed at the Cedefop agency for real-time classification of Web job vacancies relies on the SVM Linear classifier employing both titles and full descriptions. To the best of our knowledge, this is the first work focusing on extracting information from Web job vacancies, expressing the data according to the ISCO and ESCO taxonomies for a macro perspective, and making them available to a broad set of stakeholders. This opens the path for developing new products and services in the LM context.

<sup>32</sup>The lower quartile is the 25th percentile while the upper quartile is the 75th percentile.

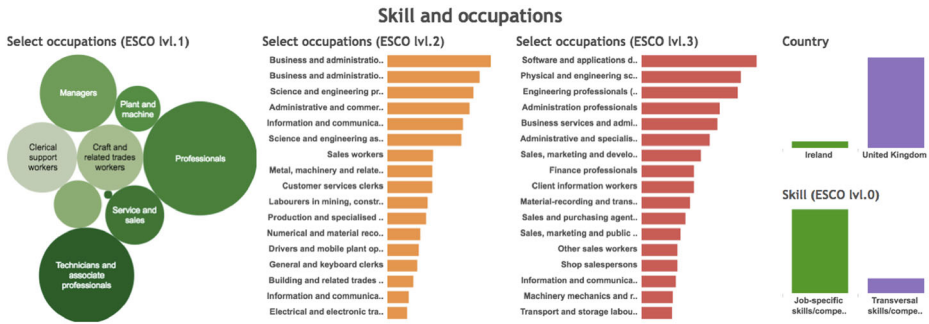


**Fig. 2** Box-plots to evaluate results for the machine-learning algorithms employed

**Table 4** A  $\chi^2$ -based feature selection process was used before feeding data to the SMV Linear classifier

Word Selection Evaluation on Titles	Accuracy
% of selected Title Unigrams	
100%	0.86
75%	0.85
50%	0.85
25%	0.84
10%	0.80

The  $\chi^2$  method identifies the most important features for text classification. The table reports the classification results using several subsets of the input features, from 100% (all features, i.e., no filtering) to the first 10% of the most important features. The classifier considers only unigrams in titles



**Fig. 3** A dashboard example in the WoLMIS Web Interface. The dimensions of circles and bars are related to the number of vacancies per occupation level. In WoLMIS the dashboard is interactive: by clicking on a specific occupation level, the remaining part of the dashboard gets updated consequently (e.g., by selecting an occupation group in a level only the sibling occupations will be showed in the next levels)

## 6 WoLMIS in action: system interface and LMInt analysis

This section presents the WoLMIS interface (Section 6.1) and describes how the job vacancies published on the Web are transformed into data that are useful for decision making analyzes in the LM context (Section 6.2). The data illustrated in this section are the result of the application of the SVM Linear classifier applied to the  $FS_2$  feature set.

### 6.1 The WoLMIS interface

The WoLMIS presentation layer is constituted by a Web-based interface, to facilitate the interaction with a broad set of users. It can display the job market data as reports, dashboards, maps, and OLAP cubes. Predefined dashboards are preferred by non-experienced users, while analysts or users more akin to analyze or investigate data can perform ad-hoc queries, navigate OLAP cubes, and produce customized dashboards and maps for other users. The data presentation layer has been implemented by using Pentaho,<sup>33</sup> an open source solution that supports the creation and the management of reports, data cubes and dashboards. Using the software as a service paradigm, some Service Cloud Computing platforms were added: Carto,<sup>34</sup> to produce dashboards based on geographical maps, and Tableau,<sup>35</sup> to create dashboards based on the narrative visualization paradigm (Segel and Heer 2010). Both are commercial platforms, but we used only their free-of-charge services. A dashboard example is reported in Fig. 3, showing the number of vacancies advertised in the UK and Ireland. They are layered by the different levels of the ISCO hierarchy.

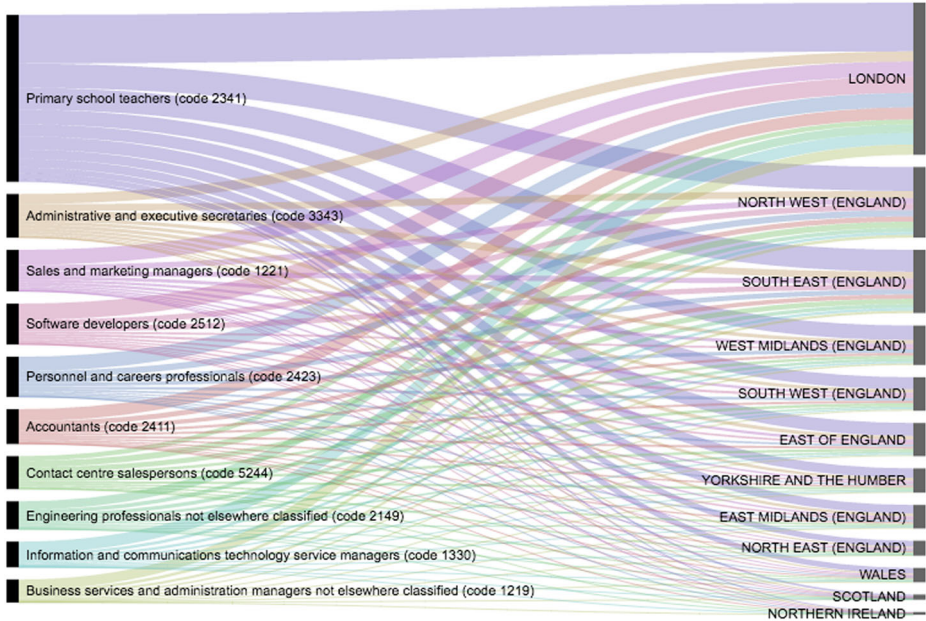
### 6.2 Some examples of WoLMIS LMInt analyzes

In this section, we provide some examples of the fine-grained and short time-to-market analysis that WoLMIS can provide to LM analysts and specialists, with respect to the

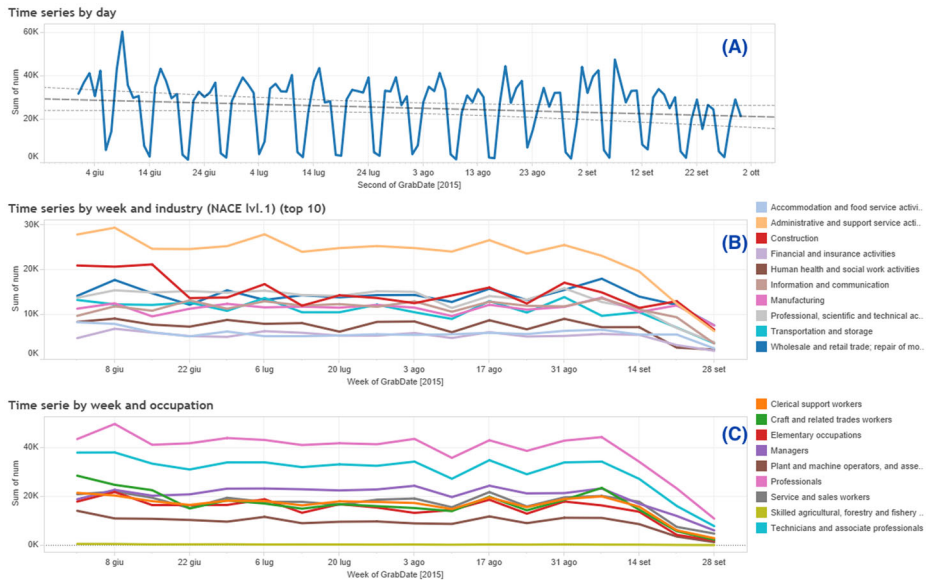
<sup>33</sup><http://www.pentaho.com>

<sup>34</sup><https://carto.com/>

<sup>35</sup><http://www.tableau.com/>



**Fig. 4** The chart illustrating the top 10 occupations required in English speaking Web sites and their most frequent workplace locations are expressed using the level 1 NUTS terms. An interactive version of this figure has been made available on-line at <https://goo.gl/prRKtd>

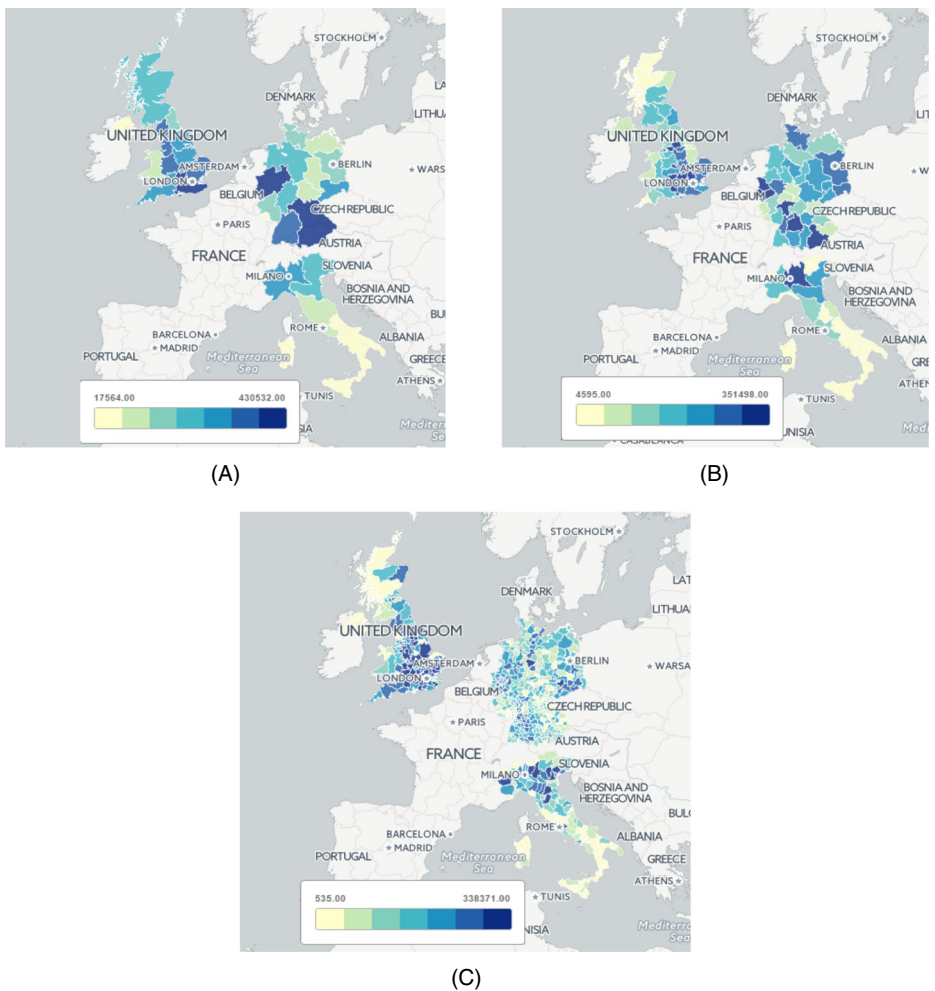


**Fig. 5** An up-to-date snapshot taken from WoLMIS showing the number of job vacancies crawled daily (graph A), the number of job vacancies classified with respect to the economical sector according to NACE standard (graph B) and the ISCO first-level (graph C)

geographical dimension. Figure 4 shows, as an example, the top-10 most requested ISCO occupations and the related most frequent geographical areas where vacancies are located. In this example, the geographical level that has been selected is the 1<sup>st</sup> of the NUTS hierarchy, but more fine-grained areas can be identified.

### 6.2.1 Occupation trends

According to Fig. 4, the most requested occupation in the analyzed countries is [2341<sub>ISCO</sub>, Primary school teachers], which overwhelms the other occupations. Although the result seems a bit surprising, the domain experts acknowledged that “Primary school teachers” open positions concentrate around and during the August summer school holidays, therefore the data are biased by a seasonal effect.



**Fig. 6** WoLMIS coverage at different NUT levels. The amount of occupation increases with the intensity of the blue color for each NUT level, namely NUT1 (a), NUT2 (b) and NUT3 (c)

Exploiting the ISCO hierarchy allows WoLMIS to monitor the LM demands even at a coarser occupational level, by comparing the labor demands over the 5 Countries involved within the project. Indeed, in Fig. 5 we provide an up-to-date snapshot to show Web labor time line since the WoLMIS deployment. Furthermore, ESCO actually provides the translation of ISCO occupation concepts in more than 20 European languages,<sup>36</sup> this was used to add multilingualism to WoLMIS.

### 6.2.2 WoLMIS demo

The geographical distribution of vacancies during the four months in which the WoLMIS system has been tested are shown in Fig. 6. Here it is possible to visualize and evaluate both (i) the geographical areas at different levels of granularity (nation, region, etc.), and (ii) the job vacancies distribution for each area. This information allows to monitor the geographical distribution of job vacancies posted in each EU Country involved in the project at a very fine-grained level. An animated heat-map showing the job vacancy collection progress over time is available at <http://goo.gl/MQUiUn>. Finally, a demo video of the WoLMIS Web interface is available at <http://goo.gl/RRb63S>.

## 7 Conclusions and further research

Web portals publishing job vacancies play a crucial role in the European market of job offers and requests. By gathering and analyzing the job vacancies published on-line, a lot of information describing job market trends can be extracted in a very short time frame with respect to the traditional sample-based surveys used by official institutes of statistics. This allows one to evaluate and to define new policies and services accordingly.

In this paper, we have presented WoLMIS, a system for classifying Web Job vacancies over the International Standard Classification of Occupations (ISCO) taxonomy. WoLMIS is an integral component of a European tender that aims at collecting, processing, classifying, and then analyzing Web job vacancies collected from 5 EU countries to support the Vocational and Educational Training activities of the Cedefop European Agency. Here, we have described how the classification problem, that is the key task of this system, has been addressed by applying machine learning algorithms, and we reported and discussed their classification performances. Focusing on the English vacancies, we have analyzed about 4 months of WoLMIS Web site contents, collecting about 1.8 million unique job vacancies (i.e., excluding duplicates). Furthermore, we have presented the WoLMIS interface, showing some concrete examples of how WoLMIS is actually used in the Cedefop's LM analysis activities.

The architecture described in this paper can be extended to several languages. In fact, even if the information classification pipeline showed in this paper focused on vacancies written in the English language, the approach can be applied to different European languages by selecting different algorithms and datasets.

**Ongoing developments** We are actually working on extracting skills from job descriptions, and on designing a methodology aimed at effectively associating the skill text

<sup>36</sup>For an updated list, see [https://ec.europa.eu/esco/portal/escopedia/ESCO\\_languages](https://ec.europa.eu/esco/portal/escopedia/ESCO_languages)

occurrences with the correct European Skills, Competencies, qualifications, and Occupations (ESCO) skill URI. In this way, information about the skill request frequencies in the on-line LM could be added to the ESCO taxonomy. The combined information on occupation and skill requests can also be used to evaluate existing educational policies and services and to define new ones according to the Web LM expectations. This would represent an important contribution to the EU agency activities. Classifiers based on Convolutional Neural Networks and on deep learning approaches will be evaluated too.

Finally, WoLMIS has been positively evaluated by the Cedefop Agency that funded the project. Then, the same agency has granted the extension of the system to all the 28 EU Countries.

**Acknowledgements** This work was supported by the Cedefop agency as part of the project “Real-time Labor Market information on skill requirements: feasibility study and working prototype”. Cedefop Reference number AO/RPA/VKVET-NSOFRO/Real-time LMI/010/14. Contract notice 2014/S 141-252026 of 15/07/2014.

## References

- Amato, F., Boselli, R., Cesarini, M., Mercorio, F., Mezzanzanica, M., Moscato, V., Persia, F., & Picariello, A. (2015). Challenge: processing web texts for classifying job offers. In *2015 IEEE international conference on semantic computing (ICSC)* (pp. 460–463). <https://doi.org/10.1109/ICOSC.2015.7050852>.
- Amato, F., Boselli, R., Cesarini, M., Mercorio, F., Mezzanzanica, M., Moscato, V., Persia, F., & Picariello, A. (2015). Classification of job advertisements: a case study. In *23rd Italian symposium on advanced database systems, SEBD 2015, Gaeta, Italy, June 14-17, 2015* (pp. 144–151). <http://dblp.uni-trier.de/rec/bib/conf/sebd/AmatoBCMMMPP15>.
- Andrews, S., Gibson, H., Domdouzis, K., & Akhgar, B. (2016). Creating corroborated crisis reports from social media data through formal concept analysis. *Journal of Intelligent Information Systems*, *47*(2), 287–312. <https://doi.org/10.1007/s10844-016-0404-9>.
- Beblavý, M., Fabo, B., & Lenaerts, K. (2016). Skills requirements for the 30 most-frequently advertised occupations in the united states: an analysis based on online vacancy data. Tech. Rep. 132, Centre for European Policy Studies (CEPS). <http://ssrn.com/abstract=2749549>.
- Bifet, A., & Frank, E. (2010). Sentiment knowledge discovery in twitter streaming data. In *International conference on discovery science* (pp. 115). Springer.
- Boselli, R., Cesarini, M., Mercorio, F., & Mezzanzanica, M. (2014). Planning meets data cleansing. In *The 24th international conference on automated planning and scheduling (ICAPS)* (pp. 439–443). <http://www.aaai.org/ocs/index.php/ICAPS/ICAPS14/paper/view/7898>.
- Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. <https://doi.org/10.1023/A:1010933404324>.
- Califf, M.E. (1998). Relational learning techniques for natural language information extraction. Ph.D. thesis University of Texas at Austin.
- Califf, M.E., & Mooney, R.J. (1999). Relational learning of pattern-match rules for information extraction. In *AAAI/IAAI* (pp. 328–334).
- Carnevale, A.P., Jayasundera, T., & Repnikov, D. (2014). Understanding online job ads data: a technical report. Tech. rep., Georgetown University, McCourt School on Public Policy, Center on Education and the Workforce. <https://cew.georgetown.edu/wp-content/uploads/2014/11/OCLM.Tech.Web..pdf>.
- Ceci, M., & Malerba, D. (2007). Classifying web documents in a hierarchy of categories: a comprehensive study. *Journal of Intelligent Information Systems*, *28*(1), 37–78.
- Cesarini, M., Mezzanzanica, M., & Fugini, M. (2007). Analysis-sensitive conversion of administrative data into statistical information systems. *Journal of Cases on Information Technology*, *9*(4), 57–81.
- Chang, C.C., & Lin, C.J. (2011). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *2*(3), 27.

- Crowther, P.S., & Cox, R.J. (2005). A method for optimal division of data sets for use in neural networks. In Khosla, R., Howlett, R.J., & Jain, L.C. (Eds.) *9th International conference on knowledge-based intelligent information and engineering systems, KES 2005, Melbourne, Australia, September 14-16, 2005, Proceedings, Part IV* (pp. 1–7). Berlin: Springer. [https://doi.org/10.1007/11554028\\_1](https://doi.org/10.1007/11554028_1).
- Elias, P., & Purcell, K. (2004). Soc (he): a classification of occupations for studying the graduate labour market. Tech. rep., Institute for Employment Research, University of Warwick, Coventry, UK. <http://www2.warwick.ac.uk/fac/soc/ier/research/completed/7yrs2/rp6.pdf>.
- ENRLMM (2016). The european network on regional labour market monitoring. <http://www.regionallabourmarketmonitoring.net/>. Visited on 2016-11-11.
- Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., & Lin, C.J. (2008). Liblinear: a library for large linear classification. *The Journal of Machine Learning Research*, 9(Aug), 1871–1874.
- Freitag, D., & Kushmerick, N. (2000). Boosted wrapper induction. In *AAAI/IAAI* (pp. 577–583).
- Haykin, S. (1999). *A comprehensive foundation of neural networks*. Upper Saddle River: Prentice Hall.
- Hong, W., Zheng, S., & Wang, H. (2013). Dynamic user profile-based job recommender system. In *2013 8th international conference on computer science & education (ICCSSE)* (pp. 1499–1503). IEEE.
- Hsu, C.W., Chang, C.C., & Lin Chih-Jen, E. (2003). A practical guide to support vector classification. Tech. rep., Department of Computer Science and Information Engineering, National Taiwan University. <https://www.cs.sfu.ca/people/Faculty/teaching/726/spring11/svmguide.pdf>.
- ISCO (2012). International standard classification of Occupations. Visited on 2016-11-11.
- Jain, A.K., Mao, J., & Mohiuddin, K.M. (1996). Artificial neural networks: a tutorial. *IEEE Computer*, 29(3), 31–44.
- Javed, F., McNair, M., Jacob, F., & Zhao, M. (2016). Towards a job title classification system. arXiv:1606.00917.
- Jindal, N., & Liu, B. (2008). Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining* (pp. 219–230): ACM.
- Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In Nédellec, C., & Rouveirol, C. (Eds.) *Machine Learning: ECML-98, Lecture Notes in Computer Science*, (Vol. 1398 pp. 137–142). Berlin: Springer. <https://doi.org/10.1007/BFb0026683>.
- Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv:1607.01759.
- Kanan, T., & Fox, E.A. (2016). Automated arabic text classification with p-stemmer, machine learning, and a tailored news article taxonomy. *JASIST*, 67(11), 2667–2683. <https://doi.org/10.1002/asi.23609>.
- Kessler, R., Torres-Moreno, J.M., & El-Bèze, M. (2007). E-gen: automatic job offer processing system for human resources. In *Mexican international conference on artificial intelligence* (pp. 985–995). Springer.
- Koperwas, J., Skonieczny, Ł., Kozłowski, M., Andruszkiewicz, P., Rybiński, H., & Struk, W. (2016). Intelligent information processing for building university knowledge base. *Journal of Intelligent Information Systems*, 48, 141–163.
- Kureková, L.M., Beblavý, M., & Thum-Thysen, A. (2015). Using online vacancies and web surveys to analyse the labour market: a methodological inquiry. *IZA Journal of Labor Economics*, 4(1), 1–20. <https://doi.org/10.1186/s40172-015-0034-4>.
- Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML* (Vol. 1 pp. 282–289).
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- Lee, I. (2011). Modeling the benefit of e-recruiting process integration. *Decision Support Systems*, 51(1), 230–239.
- Lembo, D., Torlone, R., & Marella, A. (Eds.) (2015). In *23rd Italian symposium on advanced database systems, SEBD 2015, Gaeta, Italy, June 14-17, 2015*. Curran Associates, Inc. ISBN: 978-1-5108-1087-7. <http://dblp.uni-trier.de/rec/bib/conf/sebd/2015>.
- LFS (2016). Labour force survey. <http://ec.europa.eu/eurostat/web/microdata/european-union-labour-force-survey> Visited on 2016-11-11.
- Lippmann, R. (1987). An introduction to computing with neural nets. *IEEE Assp Magazine*, 4(2), 4–22.
- Marrara, S., Pasi, G., Viviani, M., Cesarini, M., Mercurio, F., Mezzanica, M., & Pappagallo, M. (2017). A language modelling approach for discovering novel labour market occupations from the web. In *Proceedings of the international conference on web intelligence, Leipzig, Germany, August 23–26, 2017* (pp. 1026–1034). <http://dblp.uni-trier.de/rec/bib/conf/webi/MarraraPVCMMMP17>, <http://doi.acm.org/10.1145/3106426.3109035>.
- Mezzanica, M., Boselli, R., Cesarini, M., & Mercurio, F. (2012). Data quality sensitivity analysis on aggregate indicators. In Helfert, M., Francalanci, C., & Filipe, J. (Eds.) *Proceedings of the international*



- conference on data technologies and applications, data 2012 (pp. 97–108). INSTICC. <https://doi.org/10.5220/0004040300970108>.
- Mezzanzanica, M., Boselli, R., Cesarini, M., & Mercorio, F. (2015). A model-based evaluation of data quality activities in KDD. *Information Processing & Management*, 51(2), 144–166. <https://doi.org/10.1016/j.ipm.2014.07.007> <http://www.sciencedirect.com/science/article/pii/S0306457314000673>.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).
- Mooney, R.J., & Bunescu, R. (2005). Mining knowledge from text using information extraction. *SIGKDD Explorations Newsletter*, 7(1), 3–10. <https://doi.org/10.1145/1089815.1089817>.
- Müller, K.R., Mika, S., Rätsch, G., Tsuda, K., & Schölkopf, B. (2001). An introduction to Kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2), 181–201.
- Nahm, U.Y., & Mooney, R.J. (2001). Mining soft-matching rules from textual data. In *Proceedings of the 17th international joint conference on artificial intelligence* (Vol. 2 pp. 979984). Morgan Kaufmann Publishers Inc.
- Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on empirical methods in natural language processing* (Vol. 10 pp. 7986). Association for Computational Linguistics.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Perea-Ortega, J.M., Martín-Valdivia, M.T., López, L.A.U., & Martínez-Cámara, E. (2013). Improving polarity classification of bilingual parallel corpora combining machine learning and semantic orientation approaches. *JASIST*, 64(9), 1864–1877. <https://doi.org/10.1002/asi.22884>.
- Poch, M., Bel, N., Espeja, S., & Navío, F. (2014). Ranking job offers for candidates: learning hidden knowledge from big data. In *Language resources and evaluation conference*.
- Samuelson, P.A. (1974). Remembrances of frisch. *European Economic Review*, 5(1), 7–23.
- Sayfullina, L., Malmi, E., Liao, Y., & Jung, A. (2017). Domain adaptation for resume classification using convolutional neural networks. arXiv:1707.05576.
- Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34(1), 1–47.
- Segel, E., & Heer, J. (2010). Narrative visualization: telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6), 1139–1148.
- Sheth, A.P., Ngonga, A., Wang, Y., Chang, E., Slezak D., Franczyk, B., Alt, R., Tao, X., & Unland, R. (Eds.) (2017). In *Proceedings of the international conference on web intelligence, Leipzig, Germany, August 23-26, 2017*. ACM. ISBN:978-1-4503-4951-2.
- Singh, A., Rose, C., Visweswariah, K., Chenthamarakshan, V., & Kambhatla, N. (2010). Prospect: a system for screening candidates for recruitment. In *Proceedings of the 19th ACM international conference on information and knowledge management* (pp. 659–668). ACM.
- SOC2000 (2016). <http://www.ons.gov.uk/ons/guide-method/classifications/archived-standard-classifications/standard-occupational-classification-2000/index.html>. Visited on 2016-11-11.
- Sun, Q., Amin, M., Yan, B., Martell, C., Markman, V., Bhasin, A., & Ye, J. (2015). Transfer learning for bilingual content classification. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2147–2156). ACM.
- Tang, D., Qin, B., & Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. In *EMNLP* (pp. 1422–1432).
- Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 384–394). Association for Computational Linguistics.
- Vilares, D., Alonso, M.A., & Gómez-rodríguez, C. (2015). On the usefulness of lexical and syntactic processing in polarity classification of twitter messages. *JASIST*, 66(9), 1799–1816. <https://doi.org/10.1002/asi.23284>.
- Viviani, M., & Pasi, G. (2017). Credibility in social media: opinions, news, and health information - a survey. WIREs Data Mining and Knowledge Discovery. <https://doi.org/10.1002/widm.1209>.
- Xu, H., Gu, C., Zhou, H., & Zhang, J. (2017). arXiv:1705.06123.
- Yang, Y., & Pedersen, J.O. (1997). A comparative study on feature selection in text categorization. In *ICML*, (Vol. 97 pp. 412–420).

- Yi, X., Allan, J., & Croft, W.B. (2007). Matching resumes and jobs based on relevance models. In *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 809–810). ACM.
- Zhu, C., Zhu, H., Xiong, H., Ding, P., & Xie, F. (2016). Recruitment market trend analysis with sequential latent variable models. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, KDD '16* (pp. 383–392). New York: ACM. <https://doi.org/10.1145/2939672.2939689>.
- Zubiaga, A., Spina, D., Martínez-unanue, R., & Fresno, V. (2015). Real-time classification of twitter trends. *JASIST*, 66(3), 462–473. <https://doi.org/10.1002/asi.23186>.

Reproduced with permission of copyright owner.  
Further reproduction prohibited without permission.